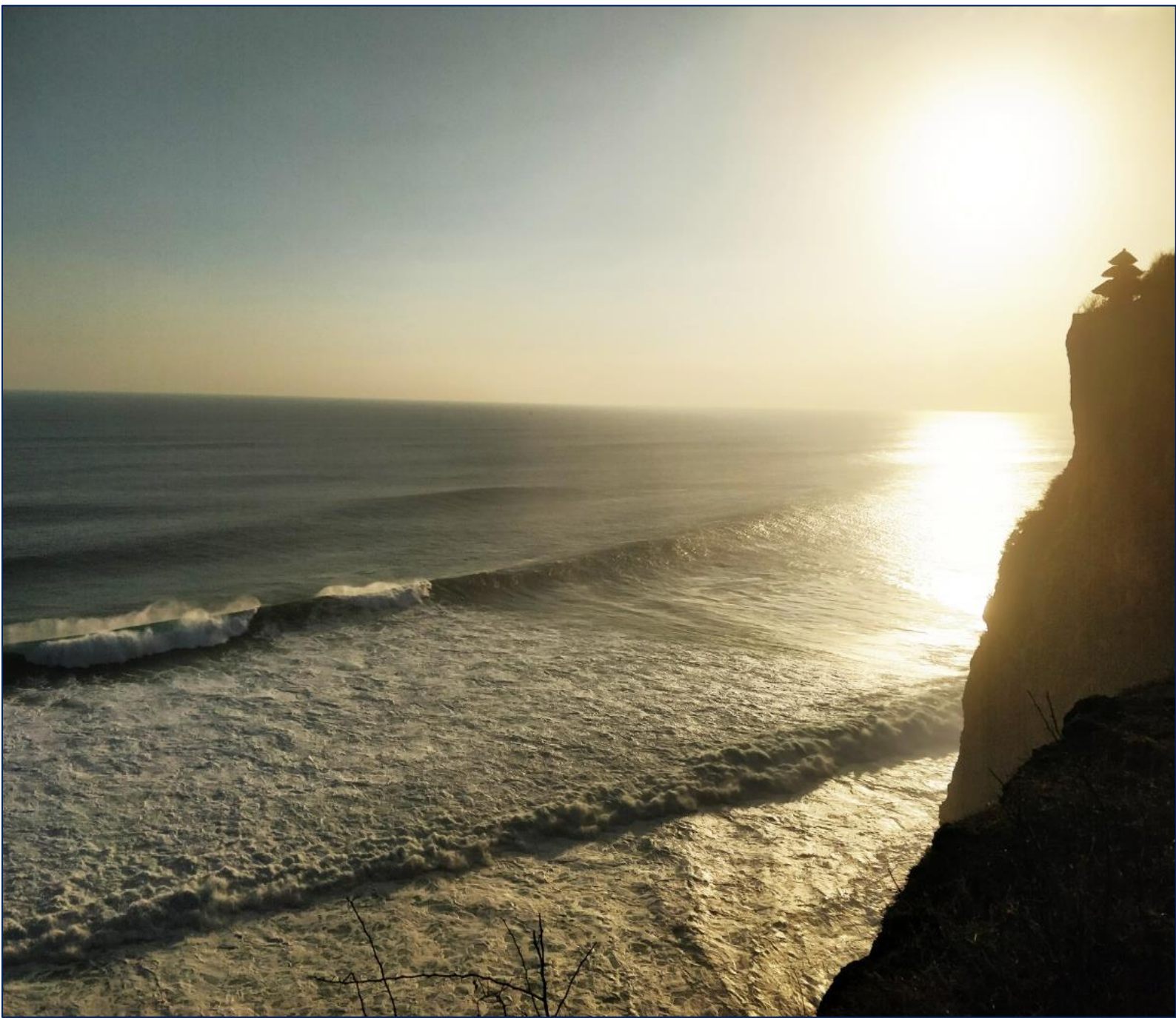
A vine-based approach for defining critical infrastructure loads



Master of *Science* *Thesis*

Written by:

Susana Sellés Valls

TU Delft

Deltares

Van Oord

A vine-based approach for defining critical infrastructure loads

Designing a breakwater in Galveston Bay, Texas

by

S. Sellés Valls

to obtain the degree of Master of Science
at the Delft University of Technology,

Student number:	4592850	
Project duration:	April, 2019 – November, 2019	
Thesis committee:	Dr. O. Morales Nápoles,	TU Delft, supervisor
	Dr. E. Ragno	TU Delft
	Ir. G. Smith	TU Delft
	Ir. E. Moerman	Deltares
	Ir. A. Lioutas	Van Oord

This thesis is confidential and cannot be made public until November, 2019.

An electronic version of this thesis is available at `http://repository.tudelft.nl/`.

TU Delft

Preface

My studies at Delft University of Technology come to an end with this thesis. In Delft, I have finished my last year of BSc in Civil Engineering as an exchange student and my MSc in Civil Engineering following the Coastal Engineering, the Ports and Waterways and Dredging Engineering specializations from the Hydraulic Engineering track. These three years in Delft (and the Netherlands in general) have been a continuous learning process, not only academically but also personally.

Firstly, I would like to acknowledge the funding through "*Beca de la Caixa, para estudios de postgrado en universidades europeas*" and the *TU Delft Faculty Excellence Scholarship 2017-2019*, which made possible my MSc studies at TU Delft.

I would like to thank all the members of my committee for their guidance and support throughout the thesis. You all have stimulated me to work hard and to give my best throughout this period. I would like to express my gratitude to the chairman Oswaldo Morales Nápoles for introducing me to the world of dependence modelling, for giving me the opportunity to discover my passion for probabilistic design (and statistics in general) throughout my studies and for all our conversations and meetings. 'Muchas gracias por todo'. I would like to thank Elisa Ragno for all the time you dedicated reading my work, for providing me with constructive arguments that made my work better and for your insights during the meetings. Your input has been very important for fulfilling the objectives of this thesis. Anestis Lioutas, thank you for all your tips and for sharing your knowledge and experience with me. Thank you for all the time you dedicated to our inspiring brainstorming sessions and for your constant motivation and enthusiasm towards this thesis. I would like to thank Greg Smith for the useful insights and recommendations, and your kind interest in this research from the very beginning. You always helped me to find a balance between theory and practice, and I am very grateful for that. Last but not least, I would like to express my gratitude to Emiel Moerman for your constant help from the very beginning of this research, most specially for all the time you dedicated helping me with ERA5 data and with ORCA. Thank you for our fruitful discussions and for always being available.

Special mention to Anaïs Couasnon who inspired me with her presentation of her thesis work during my first year of MSc. Thank you for sharing your knowledge and your work with me. Thank you for motivating me to learn R, I am very grateful for that. Thanks to Sofia Caires, for her interest in my work and for making the collaboration with Deltares possible.

I would like to thank my friend Len van der Kooij for all our vine-related discussions and your support during this thesis. I would like to thank Joanneke Jacobs, Bas van Wierst, Hassan Khan Niazi and Jochem Roubos for making the thesis experience in the '*Hydraulic hokje*' more special and enjoyable. Special thanks to my best friends: Camila Gaido, Irene Cantoni and Luis Carlos Alfaro Monge. 'Habéis hecho mi vida en Delft mejor en todos los sentidos. Gracias por vuestro apoyo constante durante estos años'.

Finally, and more importantly, I would like to express my deepest gratitude to my parents and my brother: Miguel Sellés Rochina, Susana Valls Miravete and Miquel Sellés Valls. Without you and your constant support (even in distance), I would not be where I am right now. I would like to finish this preface thanking the person whose daily company and support during the past three years has been the most essential to my success: my boyfriend, George Vonhoff. Specially, for all the time you dedicated reading my report (and correcting my English) and listening to all my progress and final presentations.

Susana Sellés Valls
Delft, November 2019

Abstract

The design of offshore and coastal infrastructures, sand nourishment and other 'soft' coastal interventions require the analysis of environmental variables (e.g. wind, waves, rainfall) that can potentially cause the failure of such structures. Processes such as overtopping, beach erosion, and coastal flooding can result from a combined action of two or more physical processes. Traditional infrastructure design practices assume the highest load previously experienced as the design load, regardless of possible interactions between variables (or processes). This may lead to a misrepresentation of critical design loads.

This thesis presents a methodology for defining infrastructure design loads accounting for their interdependence. The methodology is general and is based on regular vines. Vines are graphical tools for defining high dimensional distribution functions through pair-copula construction. With this premise in mind, the main effort was concentrated in formulating a series of steps to integrate several stages of the design: from the processing of raw data up to the choice of design loads for any specific design purpose. The vine-based methodology was applied to the design of a breakwater in Galveston Bay, Texas. This application showed that accounting for the interdependence between design variables provides a more comprehensive description of the physical system acting on the infrastructure. However, the vine-based method is computationally demanding. Hence, the applicability of this methodology should be evaluated on a case by case basis.

In parallel, the possibility to define goodness of fit test for vine-copula based on the concept of tree-equivalent classes is explored. The focus is on model selection strategies based on graphical and statistical properties of the vines. The main motivation to investigate model selection strategies for vines is the considerably large computational time needed to fit all regular vines in more than 6 nodes to the data. In this thesis, a novel algorithm is developed to facilitate the implementation of vines in higher dimensions (vines with more than 6 nodes). This algorithm significantly reduces the computational effort to select a regular vine by allowing the user to test only a subgroup of vines in n-nodes constructed on specific characteristics of the vines in $(n-1)$-nodes.

Contents

List of Figures

List of Tables

1 Introduction

1.1. Motivation

Offshore and coastal infrastructure must be constructed to withstand intense environmental conditions, such as large wind speeds, waves, currents and water levels. Critical loads that can cause failure usually occur when two or more of these events simultaneously reach extreme levels, for example high waves and large storm surge. Hence, the design of offshore and coastal infrastructure, sand nourishment and other 'soft' coastal interventions require the analysis of these extreme environmental variables, usually referred to as *design variables*.

Phenomena such as overtopping, beach erosion, wave loads on structures and coastal flooding (among others) are the result of the combined action of two or more physical processes, which in engineering terms, is referred to as an *event*. Under certain geographic and meteorological conditions there may be dependence between these physical processes, influencing the relative frequency of occurrence for each. Furthermore, the physical characteristics used to describe a single process may also be correlated, for example wave height and period for ocean swell or wind waves. Hence, considering all environmental variables that drive these physical processes and their possible interdependence is essential in the design and risk assessment of offshore and coastal structures.

The design of any structure must consider the extreme loads that may be experienced during a specified length of time, which are used to determine the required geometry, size and material of each required component. Historically this was accomplished using the highest loads previously experienced for a structure or location, whereas modern design methods seek to establish a minimum specified reliability (maximum probability of failure) by taking into account the frequency of a specific loading magnitude. Although methods for quantifying the likelihood of independent design loads are readily available, there is a lack of guidance available for quantifying the likelihood of several dependent loads.

Probabilistic assessment of extreme environmental variables has been widely studied for univariate case, however, this fails to provide a complete assessment for an underlying event characterized by a set of interrelated variables [12]. Univariate extreme models have progressively been extended to bivariate and more generally multivariate cases. Within bivariate cases, the flexibility of copulas has been exploited to investigate the joint occurrence of combined critical sea-state conditions for coastal and ocean engineering in many studies: [15] [5] [43] [21] [48] [71] [12] [13]. Copulas give the possibility to separately model the dependence structure among the random variables and the univariate marginal distributions, without imposing on them any restriction. This is one of the main reasons for the increasing popularity and recent extensive use of copulas. Despite the popularity of univariate and bivariate statistical models, there is a growing interest among researchers and practitioners to quantify the uncertainty associated with multivariate dependent design conditions.

In spite of the advantages copulas offer for the bivariate case, a n-copula cannot simply be used to 'couple' another (n-1)-copula with one variable by setting them as its marginal distributions. This follows from the so-called compatibility problem of multivariate copula constructions that was introduced in [54]. Nevertheless, the literature offer several methods that build multivariate distributions from bivariate copulas: (1) Trivariate setting based on conditional laws, (2) Conditional mixtures, (3) Hierarchical Archimedean copulas and (4) Vine-copulas, among others. The first method uses the concept of conditional distributions to build a trivariate copula

with bivariate copulas, some examples are presented in [10] and [19]. The second method uses similar concepts as the first to define multivariate models with bivariate copulas. The theory and some applications of the second approach are discussed in [66], [20], [35] and [15]. The third method is probably the most popular within the ocean and coastal engineering community. It follows from the concepts on classical hierarchical modelling that were firstly discussed in [29]. Theory on Hierarchical Archimedean copulas is presented in [67]. Some applications of this method are discussed in [41], [15], [72] and more recently, in [44]. These three methods are known by their 'supposed' simplicity. In terms of flexibility, Vine-copulas are better suited to model complex dependence structures [8], such as the ones present in many ocean and coastal systems and processes.

Despite multivariate frequency analysis receiving much attention in present time within the academic community, advanced statistical techniques such as Vine-copulas are slow in being taken up in engineering practice [42]. One of the reasons might be due to the constant focus on the academic approach rather than in the 'end user' necessity and/or perspective. According to [23] there are currently very few applications with (regular) vines within literature, mainly due to the size of the class of regular vines [50]. Consequently, the number of models to choose from is very large.

The concept of return period is widely used in infrastructure design practices to indicate the lifetime of a structure, and consequently to derive design loads. However, the very concept of return period may be confusing and sometimes misleading, in particular, in a multivariate framework. Mainly, the one-to-one relationship between return period and design value that is established in the univariate case is not valid in higher dimensions. For instance, Serinaldi discusses in [69] the misconceptions associated to the notion of return period.

In this research, the suitability of advanced statistical models, Vine-Copulas, to perform multivariate frequency analysis of extreme events is investigated. The concept of multivariate return period is explored to provide guidance for selecting critical design loads (i.e. design values) in a multivariate context.

1.2. Objectives and research questions

The main objective of this thesis is to develop a vine-based methodology for infrastructure design load definition. The methodology aims to be general and applicable to any infrastructure design problem involving multiple design variables. It seeks to integrate several stages of the design: from the processing of raw data up to the choice of design values for any specific design purpose. Besides providing guidance for selecting design values, this research investigates the suitability of vine-copulas to model environmental systems with complex dependence structures. This is investigated for a coastal engineering case study.

In the interest of all these objectives, two main research questions have been prepared:

- How can we use vine-copula models in the design of infrastructure?
- What are the advantages and disadvantages of using the vine-based methodology to derive multivariate design values when compared to the traditional approach where the variables are considered independent?

The two main research questions are directed towards an audience with an engineering background. However, this thesis also aims to reach the statistical community. In the motivation (in 1.1), one of the main issues with vine-copulas was introduced: only the class of regular vines is already very large. In addition, the number of possible models to choose from increases very fast with the number of variables that are being analyzed. Hence, it is vital to develop efficient model selection strategies [18]. In parallel to the main objective, this research aims to investigate goodness of fit for vine-copula models. Building on the work presented in [18], we propose to explore the possibility to define goodness of fit test for vine-copula based on the concept of tree-equivalent classes. In interest of this objective, a secondary research question has been posed:

- Can we identify goodness of fit for vine-copula based on tree-equivalent classes?

1.3. Thesis outline

The research questions investigated in this thesis are directed towards an audience composed of two different backgrounds: engineering and statistics. At the same time, these might have different interests. Engineers might want to dive into the application of vine-copula and its potential link to the actual design process. While statisticians might prefer to focus on the characteristics of the theoretical model and the exploratory work derived from the secondary research question. To cope with these two interests, the thesis focused on four main parts:

- Chapter 2. Literature study
- Chapter 3. A vine-based methodology for infrastructure design
- Chapter 4. Exploratory work on goodness of fit for vine-copula
- Chapter 5. Application: a case study in coastal engineering

A small story has been prepared to help explain the link between these units and to put all the work done in this thesis into context. The story is depicted in figure 1.1.

The answer to the main research question on "*How can we use vine-copula models in the design of infrastructure*" is a *methodology*. The theoretical background on the set of methods and techniques that compose the methodology is presented in Chapter 2. The developed methodology (*vine-based methodology*, hereinafter) is explained step by step in Chapter 3, and it is applied to a coastal engineering case study in Chapter 5. Also in Chapter 5, the results achieved by the developed methodology are compared with the ones achieved when using a more traditional approach. This gives answer to the second main research question posed in 1.2. The answer to the secondary research question on "*goodness of fit for vine-copula based on TEC*" is presented in Chapter 4.

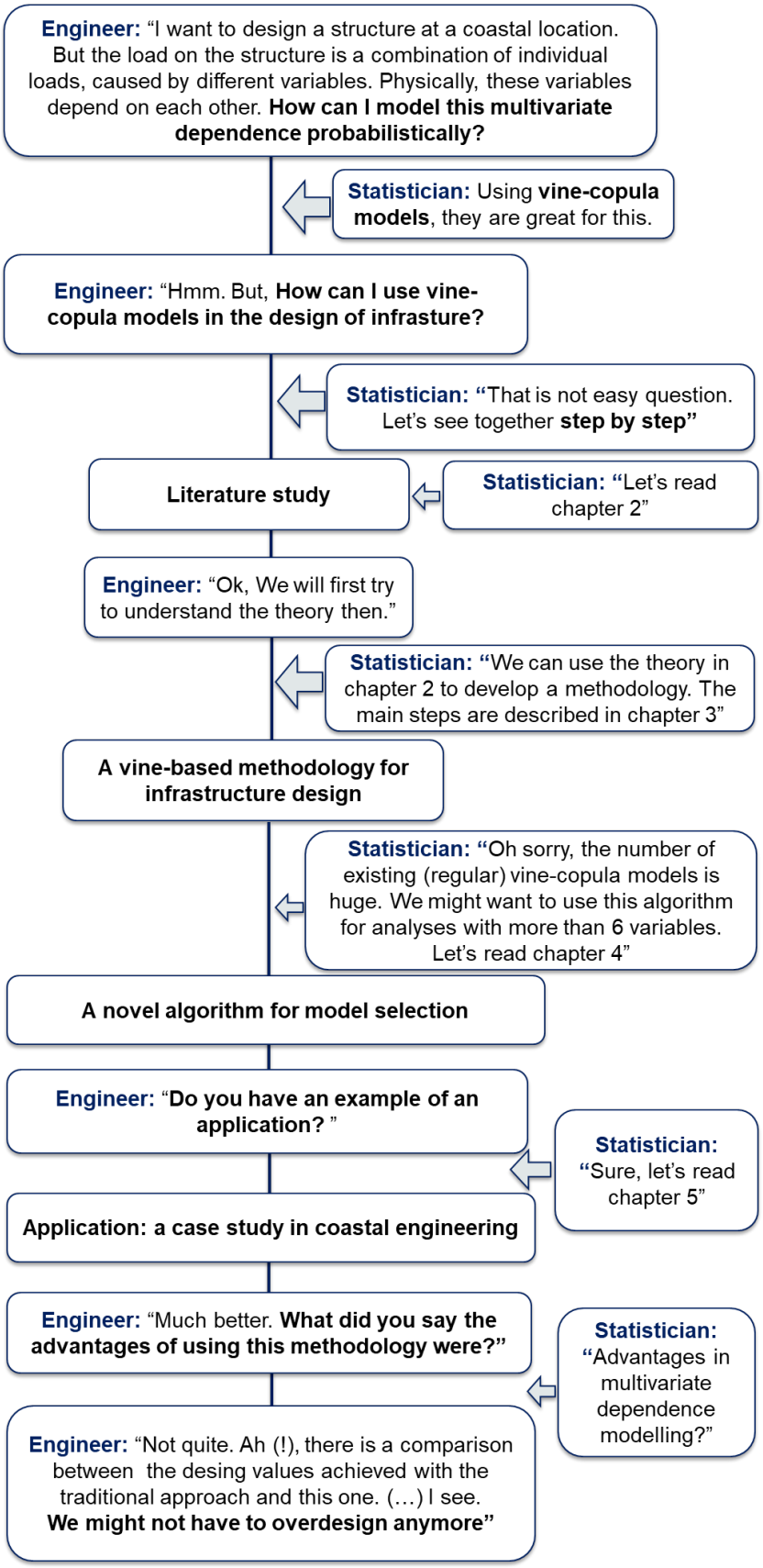


Figure 1.1: A little story that aims to explain in a simple but creative way the research methodology and the overall thesis framework. Source: Author.

2

Literature study

The theoretical background on the set of state-of-the-art methods and techniques that compose the vine-based methodology is presented in the current chapter. This chapter is aimed at readers who are interested in learning more about the theoretical foundations of the methodology that is proposed. For those readers who do not wish to dive into the statistical and mathematical details of the methodology, it is recommended to start with chapter 3.

Figure 2.1 provides an insight on the general procedure to find design loads in a multivariate context. The approach differs depending on whether the design variables are assumed to be dependent or independent. In the following sections, theory on state-of-the-art methods and techniques on the different stages of the procedures in figure 2.1 is presented.

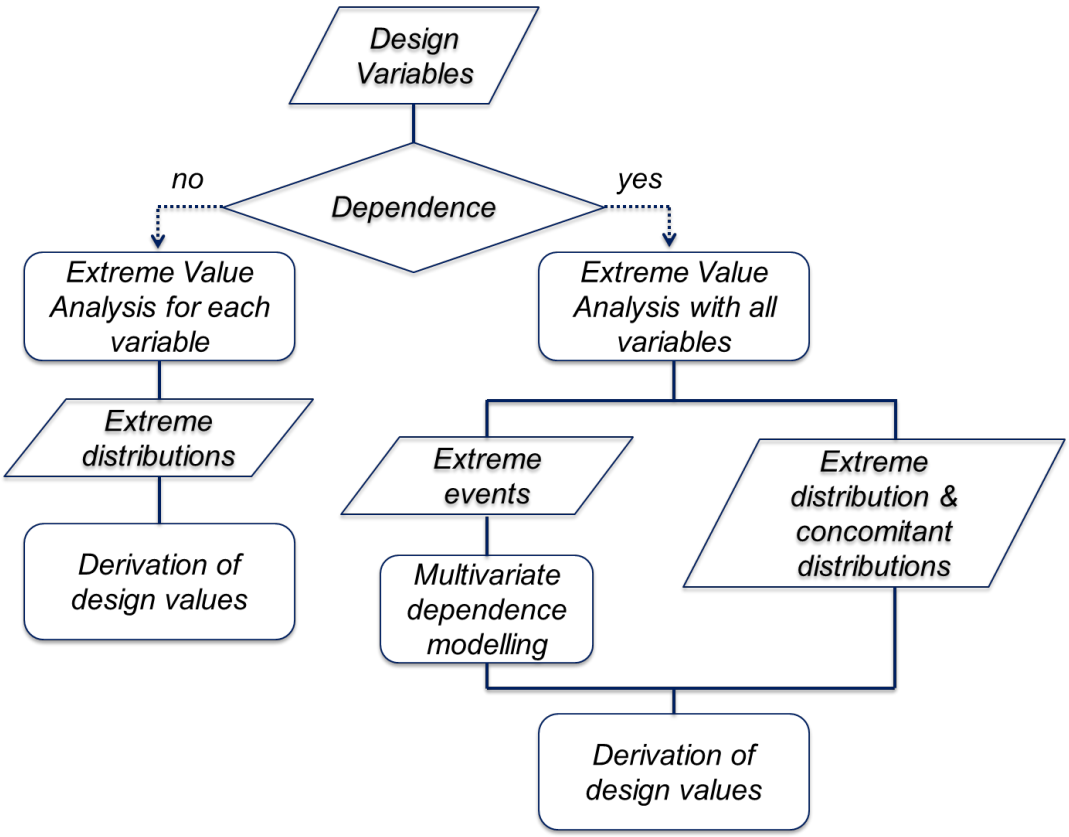


Figure 2.1: Flowchart on the general procedure to find design loads in a multivariate context. Source: Author

2.1. Extreme Value Analysis

An EVA deals with the extreme deviations from the median of probability distributions. It seeks to assess, from an ordered sample of a given random variable, the probability of events that are more extreme than any previously observed. Traditionally, the estimation of extreme environmental events is done fitting a given probability distribution on a sample of historical extreme observations, usually a temporal series of the event-describing variable. The extreme value theory (Pickands 1975) offers a valid theoretical background.

In practice, two different approaches to select extreme observations are widely used : (1) *Block Maxima* (or minima) and (2) *Peak Over Threshold* (POT). In block maxima, the maximum (or minimum) observations over a particular period of time (a month or a year for example) are selected. In contrast, the POT approach needs a predetermined threshold value. An observation is classified as an extreme observation if it exceeds (or is exceeded in the case of minima) the given threshold. One may argue that the main difference between both approaches is their focus. While POT focuses on events (e.g. storms), block maxima focuses on periods of time (e.g. days, weeks, months, years). Examples of both techniques are depicted in figure 2.2. In this thesis, the POT technique is used to sample (univariate) extreme observations.

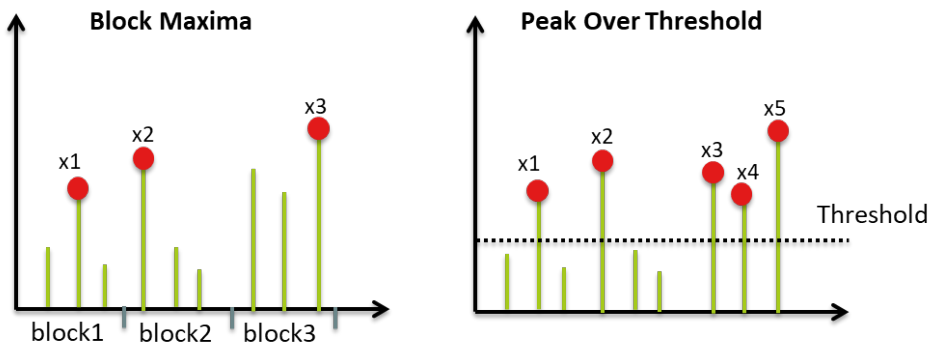


Figure 2.2: On the right, an example of Peak Over Theshold technique. On the left, an example of Block Maxima technique. Source: Author

2.1.1. Peak Over Threshold method

According to the Balkema-de Haan-Pickands Theorem, the values exceeding a given threshold converge through a Generalized Pareto Distribution (GPD) if the original sample is composed by independent and identically distributed random variables [25]. The generalized Pareto distribution has a distribution function defined as [32]:

$$F(X) = 1 - \left(1 - \frac{k}{\sigma}(X - \theta)\right)^{1/k}, k \neq 0 \tag{2.1}$$

where k, σ and θ are the parameters of the distribution and X a random variable.

In the past, the threshold for the physical selection of independent extreme events and the threshold for the statistical sampling of extreme value asymptotically convergent toward GPD were confused [40]. Traditionally, a single threshold was used for sampling data and for meeting the hypothesis of extreme value theory. In [7], the authors introduce a two-step threshold selection framework for over-threshold modeling. This method aims to identify first a 'physical threshold' for the selection of extreme and independent events (called *physical declustering*); and then a 'statistical threshold' through an optimization procedure to satisfy the GPD requirements. At the end, θ should satisfy both physical and statistical requirements.

Physical declustering aims to isolate the maximum observation recorded during a given event (i.e. a storm) with fixed duration. The storms are characterized by a predefined duration that should be longer that the resolution of the time series. On another note, the statistical optimization step is a purely statistical problem for which several methods have been proposed in the literature (see [7] for references).

2.1.2. Multivariate extreme value theory

Extreme value theory in more than one variable introduces additional issues that have to be addressed. In the univariate case, it is straightforward to find the most extreme event of a set of observations simply by taking the

maximum (or minimum) value. In the multivariate case, it is not immediately clear how to find the most extreme pair of joint observations in (for example) a multivariate time series. Suppose that one has observed the values (4,5) at a certain time and the values (7,2) at a later time. There is no universal answer to which of these pairs of observations would be considered the most extreme. The fundamental problem is that there is no natural way to order a set of vectors, as there is with a set of real-valued numbers.

One problem that arises when extending the extreme value theory to the multivariate context is that one must specify what constitutes an *extreme event*. Examples on how to determine extreme events in the bivariate case are presented in [52] and [75]. The main difference between these applications is on how the extreme observations are sampled from the bivariate data, and hence, how extreme events are defined. In [52], the extreme observations are sampled independently for each variable. In [75], the extreme observations are sampled for one variable which is rendered to be the most *dominant*, and the corresponding *concomitant* values observed together with the dominant variable are selected for the remaining variable. In this way, the dependence between the variables is kept in what constitutes the extreme event. One must appeal to dependence modelling to capture such interrelations.

2.2. Bivariate dependence modelling

Before diving into the world of copulas and dependence modelling, one should be acquainted with the concept of dependence (or independence). In statistics, two events are considered *dependent* if the occurrence of one event influences the occurrence of the other event. *Independence* of two events may be described in a similar way. If two events are independent the occurrence of one event does not influence the probability of occurrence of the other. Hence, to calculate the likelihood of two independent events occurring at the same time, one should simply multiply both individual probabilities of occurrence. When these events are dependent, one has to appeal to the art of dependence modelling.

2.2.1. Dependence measures

One way to describe dependence between two random variables is by correlation. Nevertheless, the reader should note that correlation and dependence refer to two different concepts. For example, two variables might be uncorrelated, but not necessarily independent. However, two independent variables always have zero correlation. Kendall's correlation coefficient (τ) is adopted in this thesis as one of the main dependence measures. According to [27], Kendall's correlation coefficient has more attractive properties over other correlation coefficients and seems to be better suited for smaller data sets compared to other correlation coefficients. Kendall's (τ) is defined as follows:

$$\tau = \frac{\mathrm{C-D}}{\binom{n}{2}} \quad \text{with} \quad -1 \leq \tau \leq 1$$

In which:
C number of concordant pairs in a set of observations
D number of disconcordant pairs in a set of observations
n number of observations

(2.2)

For X and Y being bivariate observations, a pair is concordant if the subject ranked higher on X also ranks higher on Y. The pair is discordant if the subject ranking higher on X ranks lower on Y. A positive value of τ indicates a positive correlation. The closest to 1 this value is, the more correlated these variables are. Same logic applies for negatives values of τ. Values of τ close to -1 indicate strong negative correlation.

In short, Kendall's τ gives an indication on the strength of the correlation. However, it does not provide information on the dependence structure between two variables. For that, one needs to appeal to bivariate distributions. Traditional statistical bivariate methods only allow for the individual behavior of the two variables (one dimensional margins) to be characterized by the same parametric family of univariate distributions [26]. For some engineering fields, such as coastal and offshore engineering, this is a big limitation since the variables of interest will likely behave according to different theoretical distributions. This problem was solved with copulas. By considering a copula approach, the dependence structure between a pair of random variables may be specified independently to that of their one dimensional marginal distributions. Thus, copulas are considered in this study to be the most beneficial models to describe bivariate dependence structures.

2.2.2. Copulas

A comprehensive review of copulas is given in [33]. Here we present some of the main concepts regarding copula modelling.

Consider the joint cumulative distribution function H of continuous random variables X and Y. H may be written as:

$$H(x, y) = C(F(x), G(y)) \tag{2.3}$$

for $(x, y) \in R$ and where *F(x)* and *G(y)* represent the one dimensional marginal distributions of X and Y respectively. The function $C : [(0, 1) \times (0, 1)]$ is the unique copula corresponding to H.

A non-parametric version of equation 2.3 (which strictly speaking is not a copula) is given in equation 2.4. There, R_i stands for the rank of X_i among $X_1, ..., X_n$, and S_i stands for the rank of Y_i among $Y_1, ..., Y_n$ and $(u, v) \in (0, 1) \times (0, 1)$. The empirical copula concept is an important concept in itself for example in survival analysis [22].

$$C_n(u, v) = \frac{1}{n} \sum_{1}^{n} \left(\frac{R_i}{n+1} \leqslant u, \frac{S_i}{n+1} \leqslant v \right) \tag{2.4}$$

2.2.3. Tail dependence

Tail dependence is another measure of dependence. In here, it is explained for the bivariate case but it can be generalized for the multivariate case. The notion of tail dependence relates to the amount of dependence in the upper-right quadrant tail or lower-left-quadrant tail of a bivariate distribution. Tail dependence between two continuous random variables is a copula property and hence, the amount of tail dependence is invariant under strictly increasing transformations of these variables [55]. The upper tail dependence coefficient λ_U for two random variables (X, Y) can be expressed as follows:

$$\lambda_U = \lim_{u \to 1} P(X > F_X^{-1}(u) | Y > F_Y^{-1}(u)) = \lim_{u \to 1} P(U > u | V > u) \tag{2.5}$$

where:
$F_X(u)$ and $F_Y(u)$ are the cumulative distributions of X and Y respectively, and u and v the normalized ranks of the aforementioned random variables.

A value of $\lambda_U > 0$ means that it is likely to observe values of U greater than u given that V is greater than u, for u arbitrarily chosen close to 1. The lower tail dependence is defined similarly to equation 2.5, but for the lower quadrant of the joint distribution. To determine whether a ranked sample is characterized by tail dependence without having to compute the limit, the semi-correlations method can be used (refer to [36] for details).

2.3. Multivariate dependence modelling

Before diving into the world of dependence modeling with vines, we shall continue from where we left it: the many benefits of modeling bivariate dependence with copulas. And most importantly, how to go from modeling dependence in 2 dimensions to higher dimensions.

In spite of the advantages copulas offer for the bivariate case, a n-copula cannot simply be used to 'couple' another (n-1)-copula with one variable by setting them as its marginal distributions. This follows from the so-called compatibility problem of multivariate copula constructions that was discussed in [54]. Nevertheless, the literature offer several methods that build multivariate distributions from bivariate copulas: (1) Trivariate setting based on conditional laws, (2) Conditional mixtures, (3) Hierarchical Archimedean copulas and (4) Vine-copulas, among others. The first method uses the concept conditional distributions to build a trivariate copula with bivariate copulas, some examples are discussed in [10] and [19]. The second method uses similar concepts as the first to define multivariate models with bivariate copulas. The theory and applications of this approach are discussed in [66], [20], [35] and [15]. The third method is probably the most popular within the ocean and coastal engineering community. It follows from the concepts on classical hierarchical modelling that was firstly discussed in [29]. Theory on Hierarchical Archimedean copulas is presented in [67]. Some applications of this method are discussed in [41], [15], [72] and more recently, in [44].

Appendix A elaborates on some concepts (and theory) on the first three methods presented in the previous paragraph. These three methods (i.e. (1), (2) and (3) listed above) are known by their 'supposed' simplicity. In

terms of flexibility, vine-copulas are better suited to model complex dependence structures [8], such as the ones present in many ocean and coastal systems and processes. Hence, vine-copulas are used in this thesis to model multivariate dependence structures.

2.3.1. Vines

Multivariate statistical models can be challenging for routine engineering applications, because they tend to require higher modelling complexity. And vines are no exception. Hence, for those readers who are not interested in the mathematical background, it is recommended to start with the application in chapter 5.

Vines are graphical tools for defining high dimensional distribution functions with complex dependence structure between variables. They where first introduced by [62], [6] and [39]. The graphical representation of this statistical model resembles grape vines, hence its name.

As the grape vine, the statistical model is composed by a series of nested trees. A *tree* is defined as an undirected acyclic graph. A tree is acyclic because it can begin in a node and end in another different node. And because there is not a fixed direction, it is also called *undirected* graph. These trees contain information on the dependence structure and this information is kept in the nodes and edges of each tree. The first tree of the vine contains as many nodes as variables in the analysis. Each variable is associated to a node. Hence, the dependence structure between 6 variables can be analyzed with a vine in 6 nodes. The nodes are linked via edges, which are represented by bivariate copulas. An edge can only link two nodes at a time. In higher trees, the edges are represented by conditional bivariate copulas and hence, these trees represent conditional dependencies. For example, the nodes of the second tree contain the same information as the edges in the first tree. And the edges of the second tree represent the conditional dependence between two nodes. The same logic applies for the rest of the trees. This is how vines use (conditional) copulas as building blocks of higher dimensionality dependence.

Formally, a Vine, V, with n-dimensionality is a nested set of connected trees $V = [T_1, \ldots, T_{n-1}]$ where the edges of tree j are the nodes of tree $j+1$, j = 1, . . . , n - 2. A vine in n-nodes is *regular* if all pairs of edges that share a common node in tree j are joined by an edge in tree j +1, with j = 1,..., n - 2. When this condition is violated, the vine is referred to as non-regular vine. An example of regular and irregular vine is presented in figure 2.3.

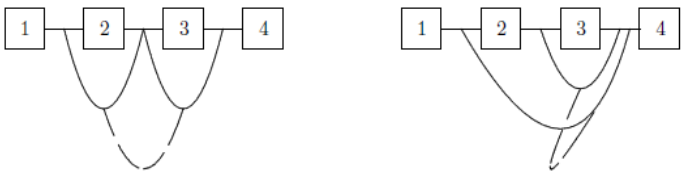


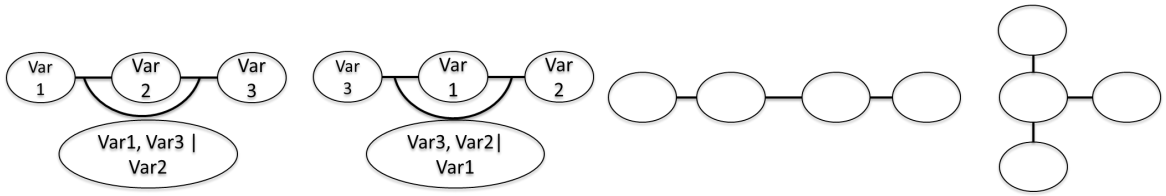
Figure 2.3: A regular vine on the left and a non-regular vine on the right, on four variables. Source: [37]

2.3.2. Regular vines

Regular vines have found application in probability theory and uncertainty analysis [50]. The first regular vine was introduced in [34], with the objective to extend the bivariate extreme-value copula to higher dimensions. Two main ways of constructing regular vines have been treated in the literature [37], being *vine-copulas* and *partial correlation vine representations*. Vine-copula constructions are obtained by assigning a bivariate copula to each edge in the vine. All copulas can be of different type and their parameters can be specified independently from each other. However, copulas will influence each other because the ones specified in a tree will affect the later trees choice of copulas [23]. Similarly, a partial correlation vine representation of a correlation matrix is obtained by assigning a partial correlation coefficient to each edge in the vine. The focus in this thesis is on vine-copulas.

Aas et al. [2] presented two sub-classes of regular vines: canonical vines, C-Vines, and drawable vines, D-vines. D-Vine based model are used in many applications on the contrary to C-Vines, which are less commonly used within literature [23]. C-Vines posse "star" structures in their tree sequence, while D-Vines are represented by "path" structures. Together, they represent the boundaries of all tree-equivalent classes (TEC) of regular

vines. The concept of TEC provides a way of classifying regular vines. According to Definition 2.3.2. in [49], two regular vines are tree-equivalent if they share the same unlabeled trees. A tree is labeled when the nodes are associated to variables or to a conditionalized combination of these in the higher order trees of the regular vine. Figure 2.4a depicts two *labeled* tress. On the contrary, figure 2.4a depicts two *unlabeled* trees. Hence, a TEC comprises different permutations of the same trees structures. An example of two tree-equivalent vines is presented in figure 2.4a. An example of two non-equivalent trees is depicted in figure 2.4b. Finally, figure 2.5 presents a general classification of vine models. This research focuses on regular vines and their TEC classification.



(a) An example of tree-equivalent regular vines

(b) An example of two non-equivalent trees

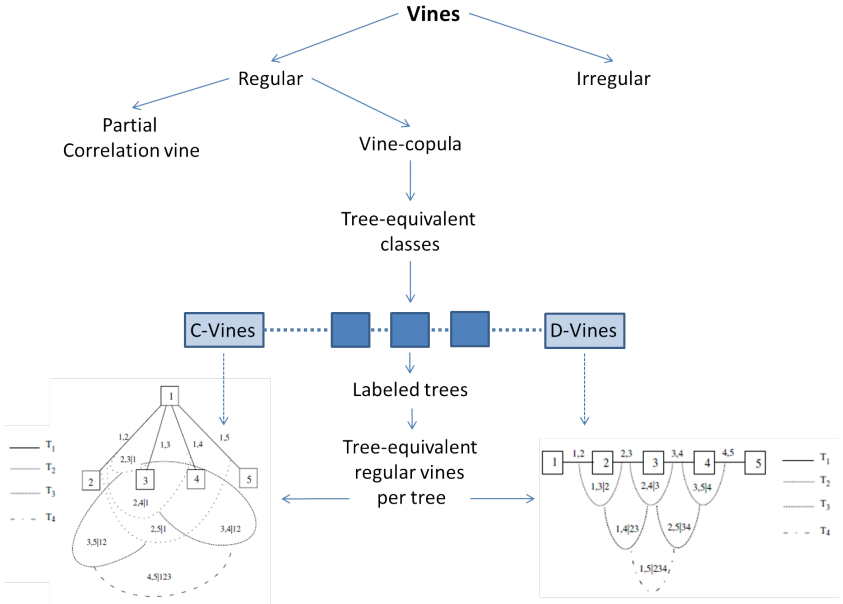


Figure 2.5: A general classification of Vines, with focus on tree-equivalent regular vines. Source: Author. Figures source: [37]

2.3.3. Properties of regular vines

In this subsection, some properties of regular vines are presented. The structure of a regular vine (or a vine in general) is build from *edges* and *nodes*. An edge links two consecutive nodes. In vine-copulas, the edges are represented by unconditional copulas or conditional copulas. Edges that embody bivariate copulas are named with only two indices that are the abbreviation of the two variables. The edges that embody conditional bivariate copulas are represented by a set of indices with the two conditioned variables and the conditioning variables.

In a regular vine, the nodes connected by a given edge in tree T_j are named the constraint set of that edge. When 2 edges are joined by an edge in tree T_{j+1} the intersection of the respective constraint sets forms the *conditioning set*, and the symmetric difference of the constraint sets forms the *conditioned* set. The label of each edge denotes the conditioned and conditioning sets. In tree T_1 the relation between the variables is defined by the rank correlation, and in the subsequent trees this relation is given by the partial correlation between the variables in the conditioned set, given the elements of the conditioning set. Formally, the *complete union* of an edge is a set of all indices that the edge contains. If two nodes a and b are joined by an edge, the *conditioned* and *conditioning* sets of this edge are the symmetric difference and the intersection of the previously define *complete unions* of a

and b, respectively. The conditioned and conditioning sets of all edges of a vine (V) are collected in a so-called *constraint set, CV*, defined as follows:

$$CV = \left[\left((C_{e,a}, C_{e,b}), D_e\right) | e \in E_i, e = (a,b), i = 1, ..., n-1\right] \tag{2.6}$$

In this thesis, the enumeration of nodes of trees in an regular vine is done using their conditioned and conditioning sets, printed before and after 'I', respectively. An example on such notation in a 7-nodes regular vine is presented in figure 2.6.

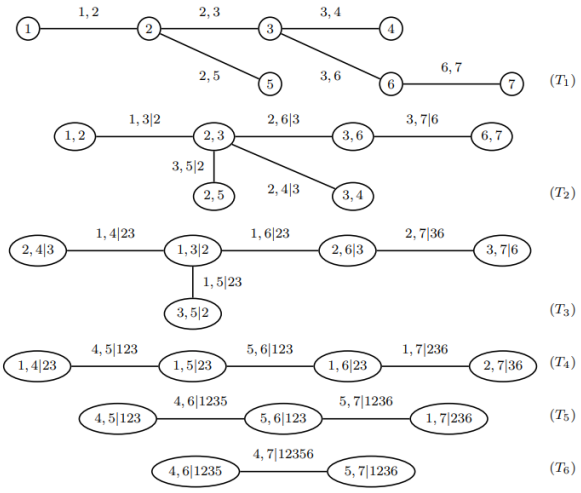


Figure 2.6: An example of regular vine on seven variables. Source: [23].

According to [23], there are currently very few applications of regular vines within literature. The very first application of vines is discussed in [2]. Mainly, the issue with regular vines is the enormous number of possible models to choose from that make them less appealing to practitioners. In view of this issue, Morales-Nápoles proposed in [49] a novel approach in the form of three algorithms for producing and enumerating regular vines. Later on, the importance of a good selection strategy continued to be discussed. Two examples are the papers: [4] and [18].

Formally, the number of regular vines increases very fast with the number (n) of nodes [51]:

$$\binom{n}{2} \times (n-2)! \times 2^{\binom{n-2}{2}} \tag{2.7}$$

As for the number of labeled trees on n nodes:

Theorem 1 *The number of labeled trees on n nodes is* n^{n-2}.

Morales-Nápoles presents in 2.1 the number of unlabeled and labeled trees, vines, regular vines and tree-equivalent vines in 3, 4, 5, 6, 7 and 8 nodes. These quantification are presented in table 2.1.

Nodes	Tree-equivalent classes	Regular Vines
3	1	3
4	2	24
5	5	480
6	22	23040
7	136	2580480

Table 2.1: Number of unlabeled regular vines and tree-equivalent vines in 3, 4, 5, 6 and 7 nodes [51].

According to [6], the density of a regular vine-copula is equal to the product of the conditional and unconditional copulas assigned to the edges.

Theorem 2 *Let $V = (T_1, ..., T_{n-1})$ be a regular vine on n elements. For an edge $e \in E(V)$ with conditioned elements e_1, e_2 and conditioning set D_e, let the conditional copula and copula density be $C_{e1,e2|De}$ and $C_{e1,e2|De}$, respectively. Let the marginal distributions F_i with densities f_i, $i = 1, ..., n$ be given. Then the vine-dependent distribution is uniquely determined, and has a density given by*

$$f_{1\cdots n} = f_1 \cdots f_n \prod_{e \in E(V)} c_{e_1,e_2|D_e}\left(F_{e_1|D_e}, F_{e_2|D_e}\right) \tag{2.8}$$

The cumulative probability function of a vine is computed by integrating its density. The reader should note that is not an easy task. Analytically, this task is very challenging. Numerically, might be easier. This issue is treated in section 3.4 and in the application in chapter 5.

2.3.4. Array representation of regular vines

Considering the numerous amount of regular vines in 6 and 7 nodes, storing their nested set of trees is too expensive. Dissman et al. discuss introduced in [23] a convenient way of representing an regular vine that makes statistical inference algorithms less computationally expensive. His work is based on the research of Morales-Nápoles in [53]. Morales-Nápoles used a lower triangular array to store a regular vine for counting the number of different regular vines. The idea is to store the constraint set of a regular vine in columns of an n-dimensional lower triangular array. By means of a *constraint set* for the array, the information on the lower triangular array can be read more easily.

Definition 1 *(Array constraint set). Let $M=(m_{i,j})_{i,j=1,...n}$ be a lower triangular array. The i-th constraint set for M is*

$$C_M(i) = \left((m_{i,i}, m_{k,i}, D) | k = i+1, ..., n, D = (m_{k+1,i}, ..., m_{n,i})\right) \tag{2.9}$$

for $i = 1, ..., n-1$. If $k = n$ we set $D=\emptyset$. The constraint set for array M is the union $CM = CM(1) \cup ... \cup CM(n-1)$. For the elements of the constraint set $((m_{i,i}, m_{k,i}), D) \in CM$, $(m_{i,i}, m_{k,i})$ is called the conditioned set and D the conditioning set.

Every element of the constraint set is made up of a diagonal entry $m_{i,i}$, an entry in the same column below the diagonal $m_{i,i}$, and all the elements following in that column $(m_{k+1,i}, ..., m_{n,i}), k = i+1, ..., n, i = 1, ..., n$. To illustrate this concept, we build an example, array (A*), with the constraint sets of the regular vine in seven nodes presented in figure 2.6. Focusing on the first column in 2.10, one can visualize the 6-th tree (T_6) edge in figure 2.6 by taking the element $m_{i,i} = 7$ and the element $m_{i+1,i} = 4$ (which become the 'conditioned set') and subsequently, the rest of the elements in that column (5,1,2,3,6) become the conditioning set. According to the definition above, this gives $((7,4),(5,1,2,3,6)) \in CM$ which corresponds to the constraint set of the edge in T_6.

$$A* = \begin{pmatrix} 7 & 0 & 0 & 0 & 0 & 0 & 0 \\ 4 & 4 & 0 & 0 & 0 & 0 & 0 \\ 5 & 6 & 6 & 0 & 0 & 0 & 0 \\ 1 & 5 & 5 & 5 & 0 & 0 & 0 \\ 2 & 1 & 1 & 1 & 1 & 0 & 0 \\ 3 & 2 & 2 & 3 & 3 & 3 & 0 \\ 6 & 3 & 3 & 2 & 2 & 2 & 2 \end{pmatrix} \tag{2.10}$$

2.3.5. Goodness of fit measure

The Akaike Information Criterion (AIC) is a relative goodness of fit measure and it is used in this thesis to select marginal distributions and bivariate copula families in the vines. The AIC is formulated as:

$$AIC = 2k - 2\log L \tag{2.11}$$

where k is the number of parameters and L is the likelihood function estimate of interest.

The advantages of the AIC are its ability to account for both the complexity (number of parameters) and the plausibility (likelihood) of the model. This means that simpler models (with fewer parameters) will be favored. However, AIC provides a goodness of fit measure of one model relative to the others tested without providing information about the absolute quality of the fitting. For this reason it is advised to performed in parallel other goodness of fit tests (e.g, root mean squared error).

2.4. The concept of return period in the derivation of design values

One objective of this thesis is to derive design values accounting for the interdependence between infrastructure loads, hereafter multivariate design values.

The concept of return period is widely used in infrastructure design practices to indicate the lifetime of a structure, and consequently to derive design values. Considering X as the infrastructure load of interest and being X independent from other loads (univariate case), there is a one-to-one relationship between the critical design value, X_{DV}, and return period T:

$$X_{DV} = F^{-1}(\frac{1}{T}) \tag{2.12}$$

where:

$$T = \frac{1}{P(X > X_{DV})} = \frac{1}{1 - P(X \leqslant X_{DV})} \tag{2.13}$$

T is expressed in unit of time, usually years. Hence, P(X>X_{DV}) is the chance that the critical design value is exceeded in the unit time, for example in one year, causing the failure of the infrastructure.

However, climatic variables, which act as loads on infrastructures, are often interdependent. Neglecting this can lead to an over- or an under-estimation of infrastructure loads. For this reason, there is an increasing interest in methods for deriving design values that accounts for dependence. In the following sections, the concept of return period and design values in the multivariate case will be introduced and discussed.

2.4.1. Conditional return period

The concept of conditional return period is often used to define design values of two (or even more) infrastructure loads which are interdependent. Given two loads X and Y and their joint cumulative distribution function $F_{xy}()$, the conditional distribution function $F_{y|x}()$ can be implemented to define the critical design value Y accounting for the information available on the other variable X. The conditional return period is then defined as:

$$T = \frac{1}{(1 - Fy|x)} \tag{2.14}$$

where x is the conditioning design variable of choice corresponding to a given univariate return period T and y is the design value corresponding to the conditioning return period $T_{y|x}$.

It should be noted that this approach does not result in a real bivariate design event with a joint return period, in the strict sense. An example on the conditional approach is discussed by Xu et al. in [73] for the bivariate and trivariate setting.

2.4.2. Bivariate return period

Let us consider the two dependent loads X and Y introduced before and their joint cumulative distribution function $F_{xy}()$. As introduced in section 2.2.2, the joint probability F_{xy} can be expressed in terms of copula function: $F_{xy} = C(F_x, F_y)$, where F_x and F_y are the marginal distributions of the loads X and Y, respectively.

Following the definition of the univariate return period (equation 2.13), the bivariate return period can be defined depending on the system's (infrastructure) failure mode [65]:

$$T_{and} = \frac{1}{\mathrm{P}\{X > x, Y > y\}} = \frac{1}{1 - F_X(x) - F_Y(y) + C\left(F_X(x), F_Y(y)\right)} \tag{2.15}$$

$$T_{or} = \frac{1}{\mathrm{P}\{X > x or Y > y\}} = \frac{1}{1 - C\left(F_X(x), F_Y(y)\right)} \tag{2.16}$$

T_{and}, *And-return period*, considers hazardous the condition in which both variables exceed their design value (failure in parallel) while T_{or}, *Or-return period*, considers hazardous the condition in which at least one of the variables exceeds its design value (failure in series).

However, in a bivariate (and more in general multivariate) context, the one-to-one relationship between return period and design value does not hold anymore. F_{xy} is indeed a surface, which means that there are infinite events, i.e. pairs of (x,y), associated to the same probability (probability isolines or quantile curves), and so to the same return period. Different approaches has been suggested to overcome this limitation [11], [28], [61] and [47]. However, this issue becomes even more difficult in dimensions higher than 2.

2.4.3. Multivariate return period

Salvadori et al. proposes in [64] the theoretical background to extend the bivariate quantile curves to a higher dimensions. However, the multivariate version of a curve is a *hyperplane*. Solving the equations of a hyperplane and backtracking the sets of design values associated with a certain risk level might become challenging and computationally demanding. Thus, we suggest to predefine the design values based on the univariate case and then, calculate their associated multivariate return period (or probability of exceedance) based on a *risk* scenario that represents the system dynamics. These scenarios have been presented in the previous section, 2.4.2.

Following the definition of the bivariate return period (equation 2.13), the multivariate return period can be defined depending on the system's (infrastructure) failure mode [65]:

$$T_{and} = \frac{1}{\mathrm{P}\{F_1(x_1) \geq 1/T_1, F_2(x_2) \geq 1/T_2, ..., F_N(x_N) \geq 1/T_N\}} \tag{2.17}$$

$$T_{or} = \frac{1}{\mathrm{P}\{F_1(x_1) \geq 1/T_1 \text{ or } F_2(x_2) \geq 1/T_2 \text{ or ... or } F_N(x_N) \geq 1/T_N\}} \tag{2.18}$$

where $\{x_1, x_2, ..., x_N\}$ represent the N-design values, $\{T_1, T_2, ..., T_N\}$ the univariate return periods and $\{F(X_1), F(X_2), ..., F(X_N)\}$ the marginal distributions associated to the N- random variables in the analysis.

In the same way as in the bivariate case, these *risk* scenarios should not be compared as they represent two different system dynamics.

3

A vine-based methodology for infrastructure design

The current chapter gives an answer to the main research question on "*How can we use vine-copula models in the design of infrastructure*" from an academic perspective. The answer is in the form of a methodology. The methodology consists of a sequence of *steps* that address in an integrated manner several parts of the engineering design process. The steps are described individually within the sections of the chapter, with the exception of *step 0*. Step 0 is the input to the methodology and is not treated in this thesis. The input is a multivariate time series that contains joint observations of all variables of interest.

3.0.1. Visualization and generic application of the vine-based methodology for infrastructure design

Before going into the details of each step, a generic example on the application of the vine-based methodology is depicted in figure 3.1. This figure presents in a generic manner the main elements needed to define critical design loads (i.e. design values) accounting for their interdependence. The process is summarized below.

1. The analysis starts with the collection of data. The data must be a good representation of the system that the practitioner aims to model and it comes in the form of a multivariate time series. The variables describing this system are part of the physical processes that trigger damage or failure of the structure. Depending on the application, the data might need further processing.

2. When the data is ready, the practitioner might want to analyze the extremes. Normally, infrastructure is designed to withstand extreme conditions. To find these, the practitioner should sample the extreme joint observations from the multivariate data set to have a good representation of the multivariate extremes. In parallel, a distribution is fitted to the dominant and concomitant variables in the extreme sample. This task is depicted with a doted line on the bottom of figure 3.1.

3. The next step is to select a vine-copula to model the extreme sample. In this methodology, all existing regular vine-copulas are fitted to the data. So the best regular vine-copula is selected according to a certain goodness of fit measure.

4. The selected regular vine-copula is used to calculate probabilities of exceedance of extreme scenarios that are of interest to the practitioner. To do so, the practitioner should sample from the regular vine and the size of the resulting sample should be considerably larger than the size of the extreme sample. In addition, the practitioner needs the results of the EVA to make the connection between univariate return period and design values. This link is established via the univariate distributions of the variables. The practitioner should note that these marginal distributions are fitted to the extreme sample and not to the initial time series. Finally, the practitioner can calculate the required multivariate return period or probability of exceedance with the modeled joint observations. The multivariate return period gives information on the overall risk.

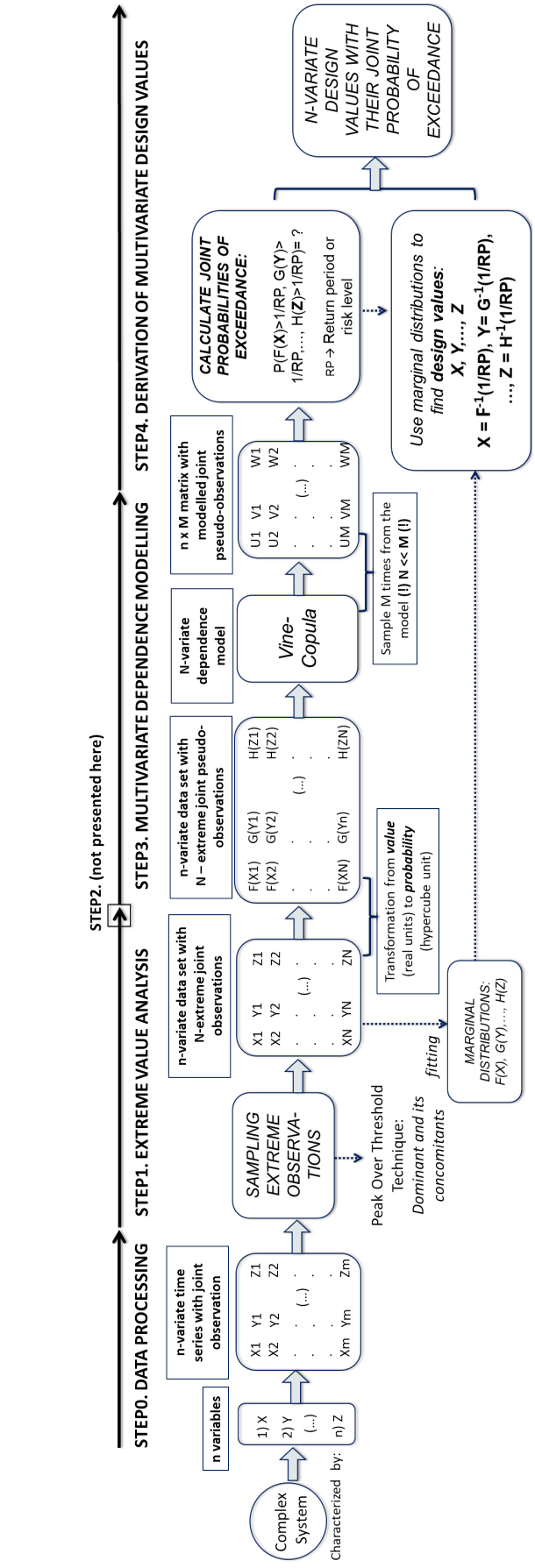


Figure 3.1: Generic example on the application of the vine-based methodology. The figure should be read from left to right. Step 2 is not included in this figure because it is strictly not essential to define critical design loads in a multivariate context. Source: Author

Within offshore and coastal engineering, the basic idea is to decompose the time series of the wave climate in sea states. So one can consider each state independently. In engineering terms, this analysis is called *metocean study* and the decomposing procedure is named *wave splitting*. In this thesis, these analyses are assumed to be included in step 0. Here, we start from step 1 onward. Nevertheless, for the readers interest in wave splitting theory and applications the following papers are recommended: [30], [74], [56] and [57].

3.1. STEP 1. Extreme Value Analysis

One of the goals of this step is to sample multivariate extreme observations from a certain multivariate time series. But usually, the largest values (or extremes) achieved by the variables do not occur together in time. Sampling all the extremes independently will probably result in a meteorological event never experienced. Furthermore, the resulting sample of extreme observations would not be strictly representative of the concept of *joint observations*. This would be consequence of sampling the extremes from different moments in time. This issue is solved by introducing the concept of *extreme event*. By introducing the concept of extreme event, the practitioner can impose the sample to be representative of a set of joint observations and thus, preserve the dependence behavior in the extreme sample.

In literature there is no real consensus on how to define and sample extreme events from multivariate data. In this methodology, we adopt the sampling approach presented in [75] (see section 2.1.2) to define extreme events in the multivariate case. We thus perform POT on one of the variables that is rendered most relevant to the design according to some criteria. We call this variable the *dominant* variable. The variables that are observed together with the dominant variable, during an extreme event, we call the *concomitants*. Consequently, the resulting *extreme sample* is composed of a set of joint observations that contain extreme values from the dominant variable and the maximum values achieved by the concomitant variables during a time window equal to the pre-defined extreme event. The concomitant values are sampled using the block maxima technique with a time-window equal to the duration of an average extreme event. The theory on block maxima is presented in section 2.1).

The next task is to fit a distribution to the dominant and concomitant variables in the extreme sample. The reader should note that the marginal distribution of the concomitant variable might not comply with the asymptotic properties of extreme observations.

3.2. STEP 2. Bivariate dependence modelling

The goal of this step is to gain insight into the physical behavior of the system. Strictly, performing this step is not essential for deriving multivariate design values. That is why it is not depicted in figure 3.1. By performing this step one can ensure that the statistical results are in accordance with what is expected from a physical point of view. If the results of step 2 are not satisfactory, the practitioner can go back to step 0 and start the analysis again.

The bivariate dependence is studied in two manners:

1. *Analyzing correlation coefficients*. All pairwise Kendall's τ estimates are computed from the extreme sample. These estimates give an indication on the correlation strength between each pair. In practice, the correlation coefficient is a measure that quantifies the influence that the underlying physical process exert on each other. For instance, wind generated waves and wind speed should be strongly and positively correlated, and hence, should achieve a large τ estimate (closer to 1 than to 0). On the contrary, swell waves should have a weak dependence with wind speed, and thus, should achieve a low Kendall's τ estimate (closer to 0 than to ± 1). By comparing the statistical results and what would be expected from a physical perspective, the practitioner can gain insight into the quality of the data. If the results are not satisfactory, the practitioner can go back to step 0 and re-do the analysis to obtain better results.

2. *Analyzing the dependency structure with a bivariate distribution*. In here, copulas are selected to model dependency structures for all possible pairs of variables. By looking at the copula's density structure, the practitioner can gain insight into the bivariate behavior of variables. For example, one can analyze tail dependencies. Again, if the practitioner is not satisfied with the statistical results, he/she can go back to step 0 and re-do the analysis to obtain better results.

3.3. STEP 3. Multivariate dependence modelling

The goal of step 3 is to choose a regular vine-copula to model the joint behavior between the variables of interest. In this methodology, all existing regular vine-copulas are fitted to the data. This might become computationally demanding in analyses with more than 6 variables. This issue aims to be solved with the novel algorithm presented in the following chapter (refer to 1).

The software *R* is recommended for its extensive library *VineCopula* [68]. R has the advantage of being an open source, which makes the use of this methodology available to everyone with access to a computer. The code used to perform the analyses is presented in Appendix C. Nevertheless, one must have, know, or be provided with the array representation of all regular vines (i.e. their matrix representation) to perform such analyses. Moreover, these matrices must be in the same format as the one software R requires, which details are explained in section 2.3.4. These matrices were all first calculated in [3].

The fitting of the regular vine-copula to the data is done with the command *RVineCopSelect()*. The AIC value is provided as output of the aforementioned command. This measure of goodness fit can be used to determine the best regular vine which would correspond to model with the lowest AIC. Other goodness of fit tests or selection criteria might also be appropriate. It is important to note that despite the choice of goodness of fit, the true model for the data would probably remain unknown. Thus, the choice of best fit is up to the user.

The selected regular vine must be validated. A simple but effective validation test is to compare a sample drawn from the theoretical model with the original extreme sample. This can be done qualitatively using the command *pairs()*. Another test can be performed by comparing the sum of the pseudo-observation from the theoretical model and the original extreme sample. The sum of the pseudo-observations gives an indication on the density and how spread the observations are. Finally, the practitioner can compare the correlation coefficients computed from modeled observations with the ones computed from the original observations.

3.4. STEP 4. Derivation of multivariate design values

This step is meant to be the link between the pure statistical analysis and the engineering design process, and it is one of the main contributions of this thesis. The objective of this step is to derive the multivariate design values using the results in step 1 (Extreme Value Analysis) and step 3 (multivariate dependence modelling with vine-copulas).

One-to-one relationship between return period (risk level) and design value as it is established in the univariate case is not valid in higher dimensions. It is quite challenging to fix a desired level of risk (or return period) in the multivariate case and back track the sets of design variables associated to that level of risk. This is due to the fact that these sets of design values that result in the exceedance of the risk level form a *hyperplane* in the multivariate setting. So instead of working with this hyperplane, we suggest to impose the design values based on their univariate return periods and calculate their corresponding multivariate return period (T_{and} or T_{or}) according to a certain risk scenario. The choice of which scenario to use should be done a priori, based on physical knowledge of the system analyzed. The probability of exceedance associated with the multivariate return period is calculated with the vine-copula. A simplified visualization of the proposed approach is depicted in figure 3.2. The engineer can adjust the design values to fulfill the requirements of an overall risk profile (depicted with a dashed line in figure 3.2). The information on the risk is given by the multivariate return period, or corresponding probability of exceedance.

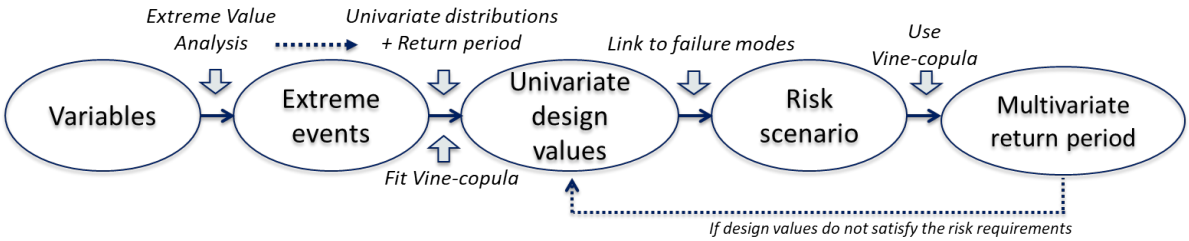


Figure 3.2: Visualization of the main steps to derive design values and determine their multivariate return period. Source: Author

The design values are determined with the univariate distributions ($F(X)$) resulting from step 1. This is depicted in figure 3.2. Given a risk level (p), the practitioner can calculate the design values as follows:

$$X_{DV} = F^{-1}(p) \tag{3.1}$$

being X_{DV} the design value for the random variable X that corresponds to the risk level p

The reader should note that $F(X)$ might not be an extreme distribution if X is one of the concomitant variables. Hence, the risk level (p) in equation 3.1 does not strictly represent a return period (T), as it is understood in engineering practices.

The multivariate return period (T_{and} or T_{or}) in equations 2.17 and 2.18 is associated to an *exceedance probability* of a certain extreme event. The extreme event is the response of the combined action of the design values and it is critical to the structure of interest. The probability of a set of variables exceeding a certain thresholds is usually calculated with the cumulative distribution function (cdf). When using vines, the cdf can be calculated by integrating the density of the vine. The formula of the density is presented in equation 2.8. Solving this integral analytically might become too challenging and not practical. Solving the integral numerically, for analysis with more than 4 or 5 variables, might become computationally demanding.

Instead of solving the integral to calculate the joint probabilities, these are calculated by stochastic simulation. A large-enough sample formed by M-sets of joint pseudo-observations is drawn from the selected regular vine (see figure 3.1). The AND-exceedance probability results from counting how many of these sets exceed all their individual risk levels and dividing that number by "M", the total amount of sets. The OR-exceedance probability results from counting how many of these sets exceed at least one of their individual risk levels and dividing that number by "M", the total amount of sets. The individual risk levels, $p_1, p_2, ..., p_N$, represent the p in equation 3.1.

4

Exploratory work on goodness of fit for vine-copula

The content of this chapter is more statistically oriented. In this chapter, we explore the possibility to define goodness of fit test for vine-copula based on the concept of tree-equivalent classes (TEC, see section 2.3.2). The reader should note that the objective of this chapter is to provide some indications on how the concept of TEC could be used to define goodness of fit test for vine-copula in further research. Consequently, we focus on model selection strategies based on graphical and statistical properties of the vines.

Before diving into the world of vine-copulas and their graphical and statistical properties, a small story has been prepared to put the work in this chapter into context. The story in figure 4.1 is specially recommended for readers whose background is not statistics (or mathematics) but still remain curious about the innovative work presented in this chapter.

The main motivation to investigate model selection strategies for vines is the considerably large computational time needed to fit all regular vines in more than 6 nodes to the data. For example, the computational time needed for a regular laptop to fit all regular vines in 7 nodes (i.e. 2580480) would be around 4 months. This happens because the number of regular vine structures increases very fast with the number of variables or nodes (see equation 2.7).

In this thesis, a novel algorithm is developed to facilitate the implementation of vines in higher dimensions (vines with more than 6 nodes). This algorithm significantly reduces the computational effort to select a regular vine by allowing the user to test only a subgroup of vines in n-nodes constructed on specific characteristics of the vines in $(n-1)$-nodes. However, the choice of the subset is not straightforward.

At the beginning of this chapter, some literature on the state-of-the-art selection strategies for vine-copula is discussed. Following from this work, the algorithm is proposed in the following section. The algorithm is based on a hypothesis (see below) that was discussed in [53]. The approach on how to test the hypothesis is explained next within the chapter. And finally, the results are presented and validated.

4.1. Literature study on selection strategies

The number of regular vines is very large and it increases very fast with the number of nodes (see 2.3.3). Hence, the use of efficient selection strategies has become necessary for all components of a vine specification. The selection strategies are usually presented in the form of sequential algorithms. Sequential selection procedures were developed to mimic the way in which vine models are constructed: using a set of sequential linked trees. These algorithms select a vine-copula model based on some pre-defined criteria. Hence, these are computationally fast because they only fit a subset of vines or even only one vine model that fulfills the criteria.

Two different sequential model selection algorithms for regular vine-copulas are popular in literature: (1) the algorithm by Dissmann et al. in [23] and (2) the algorithm by Kurowicka in [38]. The selection procedure developed in [23] is a top down strategy. Since higher trees capture conditional dependencies, the order of the nodes in the first tree is such so that the strongest pairwise dependencies are well captured. The algorithm

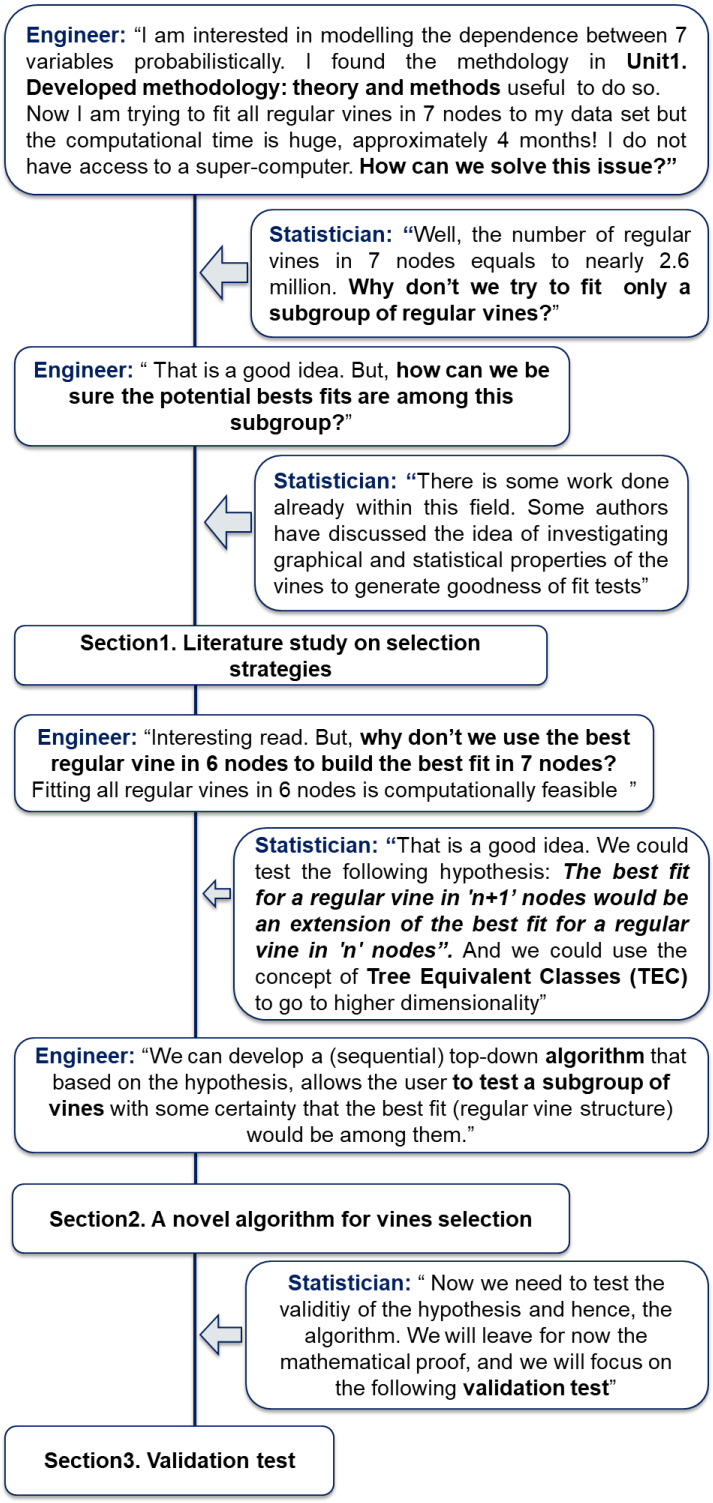


Figure 4.1: A little story that aims to explain in a simple but creative way the work done in this chapter. Source: Author.

presented in [23] calculates for the first tree all pairwise Kendall's τ estimates. Their absolute value is used as edge weight to find a tree that maximizes the sum of edge weights among all possible trees. This is accomplished by applying a maximal spanning tree (MST) algorithm. An example of such is presented in [16]. In the next step, copula families and their parameters are selected for the edges of the top tree by using the smallest AIC. With this information, all pairwise Kendall's τ estimates are computed for edges (in the second tree) that maintain the proximity condition necessary for a regular vine structure. And again, by applying a MST algorithm the

second tree is defined. The corresponding copula families and parameters are chosen again by the smallest AIC approach. These steps are repeated to build the higher trees. According to [18], this approach uses the strongest pairwise conditional dependencies first to specify the complete regular vine.

In contrast, the selection procedure in [38] uses a bottom up approach. It selects the weakest conditional dependencies for the highest trees first. Instead of Kendall's τ estimates, here partial correlations as dependency measure are used. The idea is to select Tree $n-1$ first and then selecting Trees $n-i$ for $i \geq 2$ sequentially and n as the number of variables. Furthermore, if there are several choices for an edge in Tree i, the one with the lowest absolute partial correlation is chosen. This strategy chooses only the tree structure of the vine distribution. The complete algorithm is presented in [38].

In this thesis, the focus is on top-down selection strategies. A basic algorithm that underlies the top-down strategies for regular vines is presented in [18]. The backbone of such algorithm is the availability of appropriate weights that represent a certain characteristic that is important for the user when selecting the regular vine structure. However, the authors do not expect in [18] to identify the best or true regular vine tree structure in general, but rather a reasonable candidate. Mainly because once an edge is selected for a tree, the algorithm does not allow for it to be dropped in a further step.

The choice of weights is usually up to the user. Czado et al. discuss in [18] four different choices. These represent different traits of the bivariate conditional distributions that build up the regular vine distribution. The choices are summarized below:

1. *Absolute Kendall's τ*. The weights chosen in this approach are a measure of dependence. By choosing such weights, the user aims to capture the strongest pairwise dependencies in the data. Kendall's τ is a measure of dependence that captures non linear dependencies and is invariant to monotone transformations of the margins (see section 2.2.1). Because the aim is to select the strongest pairwise dependencies, a tree is defined in such a way that maximizes the sum of the absolute value of τ among all pairs that form the tree.

2. *Akaike Information Criterion*. The weights in here choose edges where the pseudo data are fitted well by the class of pair copula families considered. In Tree 1, the weights represent the lowest AIC values resulted from the fitting of each pair of variables to the chosen copula families and their parameters. In this case, the tree minimizes the sum of AIC. A similar interpretation can be made for higher trees.

3. *Copula goodness of fit p-value*. This approach is very similar to the *Akaike Information Criterion*. It was proposed in [18] to cope with the drawbacks of AIC which mainly, do not allow a quantitative assessment of goodness-of-fit. According to [18], the performance of this method relies on the selection of a pair copula term for the corresponding pair of pseudo-data values.

4. *Copula goodness of fit p-value times absolute Kendall's τ values*. The weights in this approach represent a combination of dependency strength and goodness of fit measure. They are calculated with the product between the Absolute Kendall's τ and the Copula goodness of fit p-value. According to the authors in [18], by applying these weights the effect of parameter estimation error is mitigated while still allowing for copula families that fit the data best.

The *VineCopula* package (in R) provides a function to select a reasonable candidate for the user's data. This function is called: *RVineStructureSelect()*. This function contains in fact the selection strategy that is discussed in the second paragraph of the current section. This allows the user to choose one of the four weight types listed above to select the tree structure and another selection criteria (or goodness of fit) to select pair-copula specifications (i.e. family type and parameters) for the edges of each tree.

These sequential algorithms present reasonable candidates. However, one is never completely sure which model is the true best model for the data without fitting all possible models.

In parallel, Morales-Nápoles briefly discussed another potential selection strategy in [49]. This would also use the graphical and statistical properties of regular vines, and more specifically, the use of the concept of TEC to generate a goodness of fit.

4.2. A novel algorithm for vines selection

Following the work done by Morales-Nápoles in [49], we explore the validity of the following hypothesis to define a model selection strategy in the form of an algorithm. The algorithm aims to facilitate the implementation of vines in higher dimensions (vines with more than 6 nodes).

- *The best fit for a '$n+1$' variables regular vine would be an extension of the best fit of 'n' variables regular vine*

The concept of TEC is used to extend the vine's tree structure in n nodes to $n+1$ nodes. It is assumed that the best regular vine in $n+1$ belongs to a TEC that is an extension of the TEC to which the best regular vine in n nodes belongs to. Thus, the $(n-1)$ *unlabeled* trees of the best regular vine in n nodes are the same as the last $(n-1)$ *unlabeled* trees of the best regular vine in $n+1$ nodes. The best fit is selected according to the AIC goodness of fit measure.

The AIC is a popular goodness of fit measure and is dependent on the log-likelihood function estimate of interest (see equation 2.11). Similarly, the log-likelihood function is computed with the density of the vine-copula. Theorem 2 in section 2.3.3 presents the density of a regular vine-copula. The reader may observe that the density of the vine-copula is a product of the bivariate (conditional) copulas attached to the edges of each tree in the regular vine. For this reason, it is hypothesized that the TEC (or perhaps some other graphical property of the vine) will play a role when assessing goodness of fit when a measure similar to AIC. Intuitively, one may think that the different copulas attached to the edges of regular-vines in the same tree-equivalent class will be 'closer' in some sense than those attached to regular vines in different TEC. The purpose of section 4.2.1 is to initiate exploratory work regarding this hypothesis.

When extending the analysis from n-variables to $(n+1)$-variables, one is actually adding a tree to the regular vine. This tree has one more node than the previous first tree (T1) and hence, this newly added tree must become the new T1 of the regular vine in $(n+1)$-nodes. Extending the TEC might not be enough because the number of labeled trees also increases very fast with the number of nodes (see section 2.3.3). Consequently, testing a subgroup of regular vines (belonging to the extended TEC) without any information on the added tree (T1) is still computationally demanding. Subsequently, we will assume the first *labeled* tree (T1) of the potential best regular vine in $n+1$ nodes is an extension of T1 of the best regular vine in n nodes.

The extension procedures are explored separately in sections 4.2.1 and 4.2.3.

4.2.1. Extension of the TEC

In here, we explore whether the TEC to which the best regular vine in $n+1$ belongs to is an extension of the TEC to which the best regular vine in n nodes belongs to. Due to computational constraints, we can only test whether the TEC of the best fit in 6 nodes is an extension of the TEC of the best fits in 4 and 5 nodes. To do so, all the regular vines that exist in 4, 5 and 6 nodes are fitted to a data set and classified by TEC. This data set is the one used in chapter 5. For this task, we assume that the best fit within all the regular vine class is the regular vine with the lowest AIC.

The AIC values are computed for each regular vine in 4, 5 and 6 nodes and presented in figures: 4.2 for 4 nodes, 4.3 for 5 nodes and 4.4 for 6 nodes. The notation on the tree-equivalent classes (TEC) remains the same as in [49]. With these graphs, one can determine the TEC to which the best regular vine belongs to and also, how each TEC generally performs. The results in figures 4.2, 4.3 and 4.4 suggest that none of the TEC outperforms the rest. The average performance is actually very similar for all TEC in 4, 5 and 6 nodes. The best 10 fits in 4, 5, and 6 nodes are presented in table 4.1. The results in table 4.1 suggest that more than one TEC contain reasonable model candidates for our data.

The best fit in 6 nodes belongs to the TEC *V14* (see table 4.1). According to [49], *V14* can be decomposed as a sum of the trees T13, T7 and T5. The best fit in 5 nodes belongs to the TEC *V8* (see table 4.1). According to [49], *V8* can be decomposed as a sum of the trees T7 and T5. Finally, the best fit in 4 nodes belongs to the TEC *V5* (see table 4.1). According to [49], *V5* can be decomposed as the tree T5. These results suggest the best regular vine in 6 nodes belongs to a TEC that is an extension of the TEC to which the best regular vine in 5 nodes belongs to. And the last, is an extension of the TEC to which the best regular vine in 4 nodes belongs to.

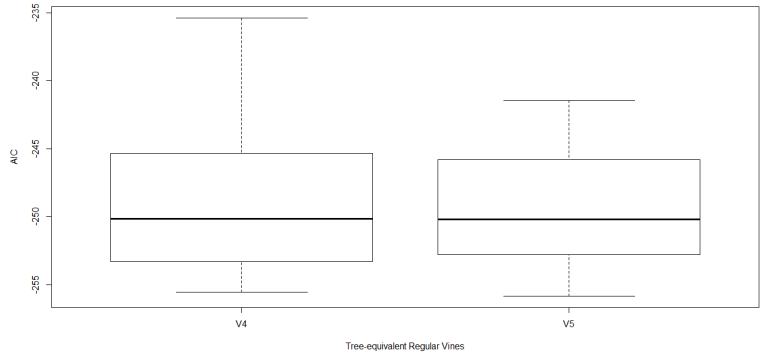


Figure 4.2: Box plot with the AIC values that result from the fitting of each regular vine in 4 nodes, categorized by TEC

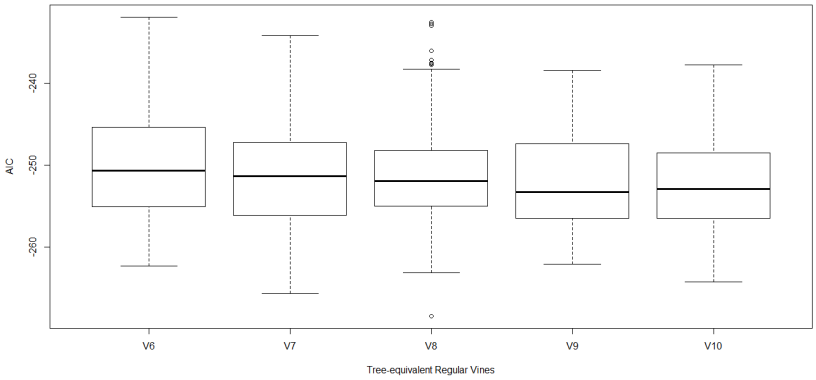


Figure 4.3: Box plot with the AIC values that result from the fitting of each regular vine in 5 nodes, categorized by TEC

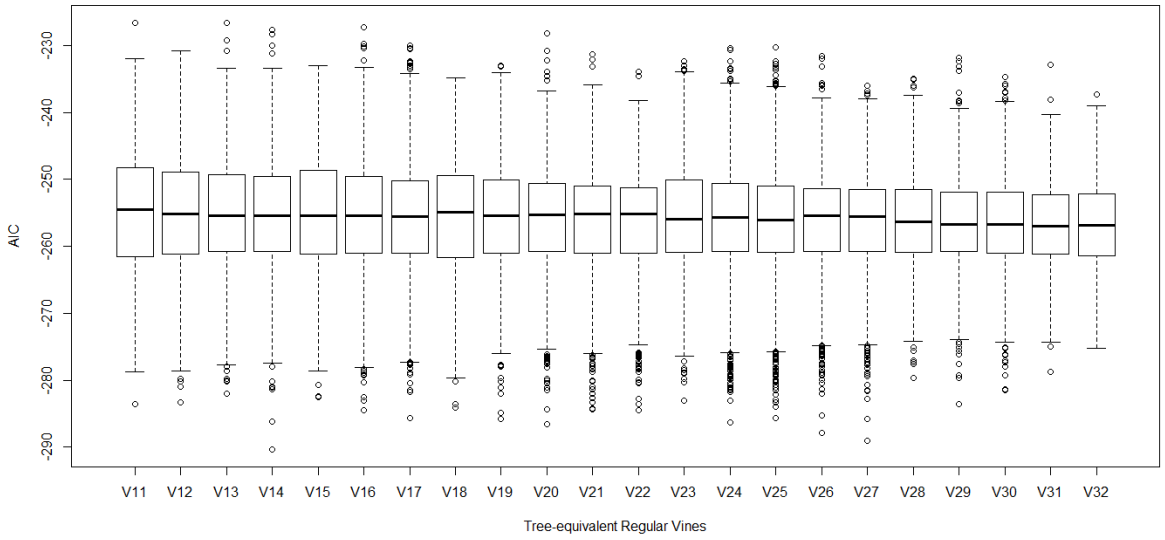


Figure 4.4: Box plot with the AIC values that result from the fitting of each regular vine in 6 nodes, categorized by TEC

General Ranking: 10 BEST REGULAR VINES	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
TEC to which 10 best regular vines in 4 nodes belong to	V5	V4	V4	V4	V5	V5	V5	V5	V4	V4
TEC to which 10 best regular vines in 5 nodes belong to	V8	V7	V7	V10	V10	V7	V8	V10	V8	V8
TEC to which 10 best regular vines in 6 nodes belong to	V14	V27	V26	V20	V24	V14	V27	V19	V25	V17

Table 4.1: The table presents an overview of the best 10 fits (regular vines) in 4, 5 and 6 nodes. For each position within the general ranking the vine-copula's TEC is presented.

4.2.2. Ordering approach

The ordering of the variables in the data set (column wise) plays an important role on the extension concept. The most dominant variables (i.e with the strongest pairwise correlations) should be placed in the first columns of the data set and should be the first to be analyzed and fitted to all regular vines. The less dependent variables (i.e with the weakest pairwise correlations) should be included following a certain order and should be added one at a time when extending the analysis to higher dimensions.

The order is established using a similar approach to the *Kendall's* τ weight method in section 4.1. For the 6-variate data in chapter 5, we calculate all pairwise Kendall's τ estimates. Each variable has 5 associated τ with the 5 remaining variables that form the data set. By summing in absolute value the 5 τ values for the each variable, the *Sum of absolute taus* (see table 4.2) is obtained. The 6 resulting values are ordered in a decreasing manner to obtain the actual order of the variables in the fitting analyses.

Following the order specified in table 4.3, the fitting of the vines was performed in 4 dimensions with the first 4 (most dependent) variables (see notation in section 5.2.3): Tm_{ww}, Ws, Tm_{ts} and Hs_{ts}. For the analysis in 5 dimensions the next (most dependent) variable was added: Hs_{ww} and finally, in 6 dimensions, the least dependent variable, the WL, was added. This is what it is meant by adding the variables by following a certain order.

Kendall's tau	Hs_ww	WL	Ws	Hs_ts	Tm_ww	Tm_ts
tau1	-0.12	-0.12	0.13	0.05	0.09	0.06
tau2	0.13	0.06	0.06	0.04	0.05	0.02
tau3	0.05	0.04	0.2	0.2	0.62	0.33
tau4	0.09	0.05	0.62	0.25	0.25	0.43
tau5	0.06	0.02	0.33	0.43	0.4	0.4
Absolute sum	**0.45**	**0.29**	**1.34**	**0.97**	**1.41**	**1.24**

Table 4.2: Results of the application of the ordering procedure to the variables in the data. The last row of the table contains the *Sum of absolute taus*. tau_1 represents the correlation between the variable in question and the first variable (Hs_{ww}, in the second column). For Hs_{ww}, tau_1 represents the correlation between variable 2 (WL) and itself, Hs_{ww}. Same logic is applied to the remaining taus. Note that the correlations between equal variables (e.g. Hs_{ww} and Hs_{ww}) are not presented in this table.

Order	Variables
1st	Tm_ww
2nd	Ws
3rd	Tm_ts
4th	Hs_ts
5th	Hs_ww
6th	WL

Table 4.3: Table presenting the final order of the variables in the columns of the data set (i.e. the first variable is placed in the first column etc.).

4.2.3. Extension of the first tree

The second part of the extension procedure refers to the extension of the first labeled tree. A tree is labeled when variables and copulas are associated to the nodes and edges of the vine, respectively. To explore the second part of the extension, we study whether the labeled trees of the best vine in 6 nodes are an extension of the labeled trees of the best vines in 4 and 5 nodes. If that were true, this concept could be used to extend T1 of the best vine in n-nodes to find T1 of the potential best fit in a higher dimension. For example, let's assume that a hypothetical *Variable 7* has the strongest correlation with a hypothetical *Variable 3*. Then, one could extend T1 of the best regular vine in 6 nodes by drawing an edge from *Variable 3* to *Variable 7*. The resulting tree would become T1 of the potential best fit in 7 nodes. This example is depicted in figure 4.5.

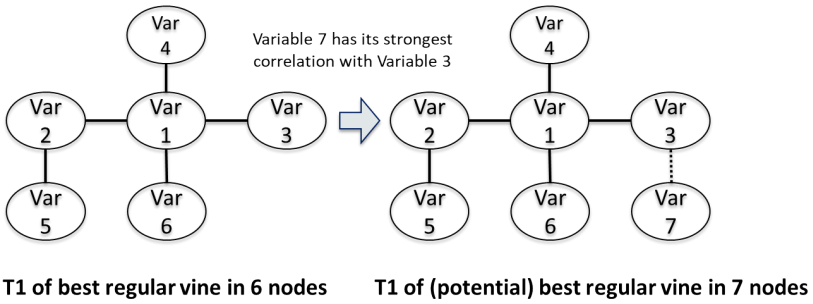


Figure 4.5: Example on the extension of the first labeled tree (T1)

The regular vines in 4, 5 and 6 nodes achieving the lowest AIC were selected in section 4.2.1 as the best models for the data. These vines are depicted in figure 4.6. In this figure, the reader may notice the application of the ordering procedure in table 4.3: the regular vine in 4 nodes contains the 4 most dominant variables, the regular vine in 5 nodes contains these variables plus the following most dominant and finally, the regular vine in 6 nodes contains all the variables and includes the least dominant variable.

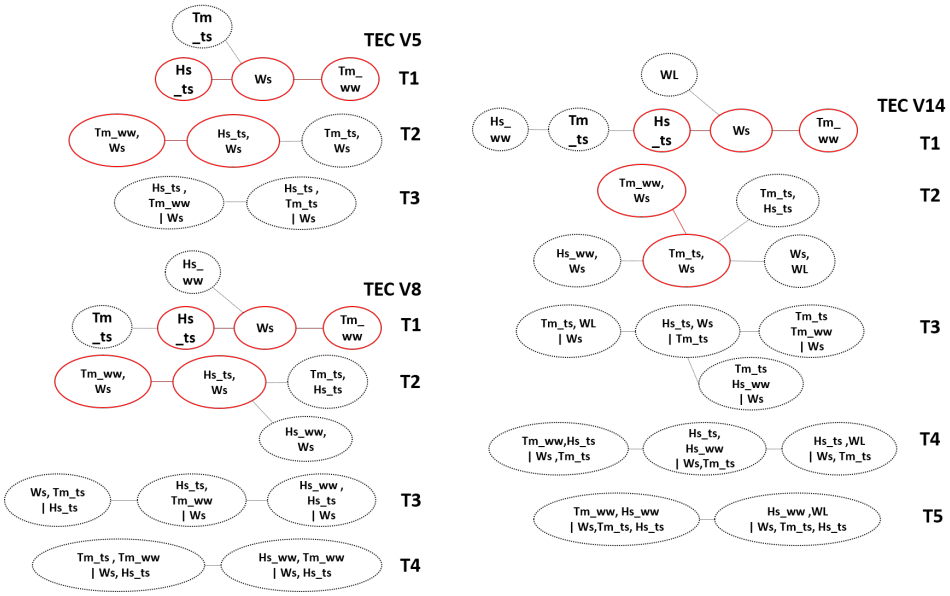


Figure 4.6: Best fits of all regular vines in 4, 5 and 6 nodes for our data according to overall AIC

The tree structures from the best vines in 4, 5 and 6 nodes that conserve in all three vines are depicted in continuous red circles in figure 4.6. The results suggest that the best fits according to the AIC do not preserve the same labeled trees when extending the analysis to higher dimensions. Nevertheless, the ordering procedure in section 4.2.2 does not take into account the goodness of fit of all the pairs of variables to the bivariate copulas. This might be a plausible reason explaining why these results do not support the presented hypothesis.

Instead of choosing the best fit as the regular vine with the lowest AIC, we choose the regular vine with the strongest correlations in its trees as the best fit. This choice is more in line with the ordering procedure established in section 4.2.2. In figure 4.7, the regular vines in 4, 5 and 6 nodes with the strongest correlations in its trees are presented. These regular vines achieve an average AIC when compared to the rest of regular vines. One can see in figure 4.7 that the best vine in 6 nodes is an extension of the labeled trees of the best regular vines in 4 and 5 nodes. In this case, the hypothesis presented at the beginning of this section seems to be valid.

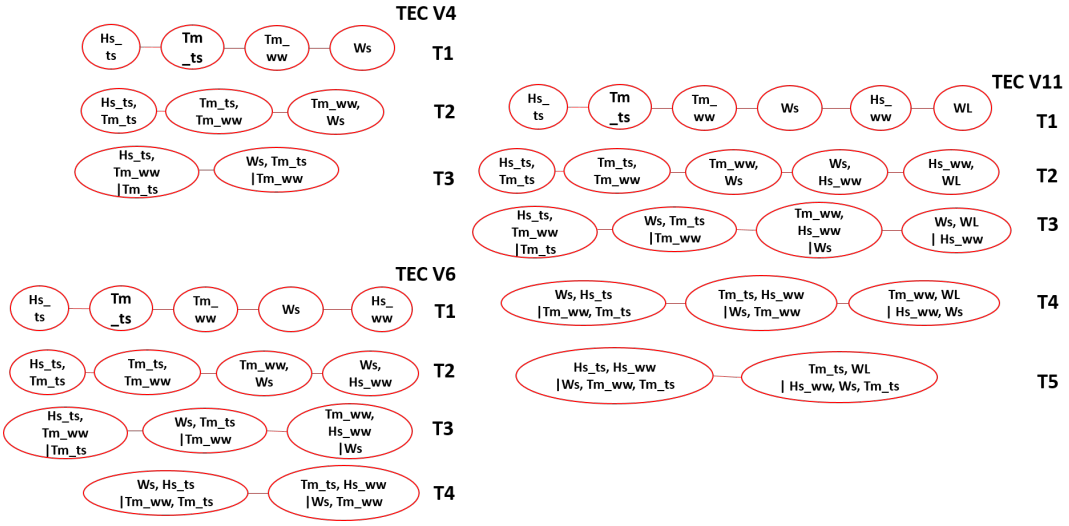


Figure 4.7: Regular vines in 4, 5 and 6 nodes fitted to the data with the strongest correlations in its trees.

The best vine according to the AIC selection criterion differs from the best fit according to strongest correlations' selection criterion. In an engineering application (such as the one in chapter 5), the most important characteristics to model probably are the correlations between the variables of interest. Thus, the best regular vine should be the one that simulates data that is correlated as close to the original data as possible. Subsequently, it is of our interest to determine which of the two regular vines represents best the original correlation coefficients. 10 million samples are simulated from (1) the regular vine in 6 nodes in figure 4.6 and (2) the regular vine in 6 nodes in figure 4.7. The Kendall's τ estimates are calculated for both samples. The absolute error is computed by subtracting the resulting τ's to the 'originally observed' correlation coefficients. The results are depicted in figure 4.8.

The absolute error achieved by the regular vine with lowest AIC is overall the largest. Hence, an engineer might feel more comfortable selecting the regular vines in figure 4.7 as the bests fits rather than the ones with lowest AIC. On the contrary, an statistician might feel more comfortable selecting the regular vine with the lowest AIC. A discussion on what the "best" choice from these two possibilities is, is out of the scope of this thesis.

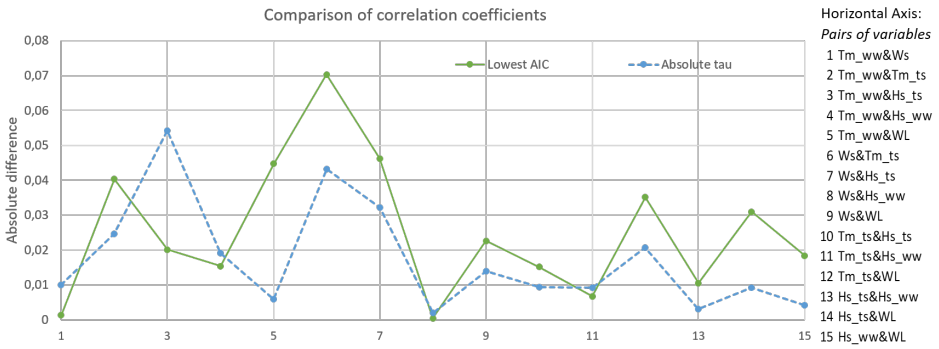


Figure 4.8: Comparison of absolute errors in the prediction of Kendall's τ between regular vine with the lowest AIC of all regular vines and the regular vineselected with the strongest correlations in its tree structures.

4.2.4. Final product

The exploratory work done in the current chapter led to the development of a (sequential) top-down algorithm (presented in 1). The backbone of the algorithm is the hypothesis tested in this chapter: *The best fit for a 'n+1' variables regular vine would be an extension of the best fit for a 'n' variables regular vine*.

The algorithm makes use of the two parts of the extension procedure explained in sections 4.2.1 and 4.2.3. It allows the user to test a subgroup of vines with certainty that a good candidate (regular vine structure) would be among them. Due to the foundations of this algorithm, the selected best fit would probable not have the lowest AIC of all the regular vines class. And probably, the selected best fit would not be the model representing best the correlation coefficients. What the algorithm provides as best model (within all the regular vine class) should be a balance between these two aforementioned criteria. This is elaborated in more detail in section 4.3.

The reader should note that fitting all regular vines with less than 7 nodes to a data set is computationally feasible. Hence, the algorithm is thought for analysis with more than 6 variables.

Data: Data set of N variales with dimension n, i.e. *Md*
Result: Regular vine-copula specification
initialization
begin
- 1: Compute the correlation matrix ("C") for *Md*
- 2: Sum in absolute value all non-diagonal elements in "C(:,i)", for i=1,...,N
- 3: Order the N resulting values from *2:* in a decreasing manner. Keep the old column indices, *i*
- 4: Re-order the columns in *Md* according to the resulting order in *3:*
- 5: Fit all regular vines in 6 nodes to *Md(:,1:6)*
- 6: Determine the best fit, *RV*, according to AIC goodness of fit test
- **for** *N= 6:N* **do**
  - **if** *N=6* **then**
    - a.1: Determine the TEC to which *RV* belongs to and set it to be V_N
    - a.2: Take the unlabeled tree-structure in T1 from *RV* and label it by selecting the combination of variables that maximizes the absolute sum of correlation coefficients in all edges of the pre-defined tree structure
  - **else**
    - b: Decompose V_N in its tree sequence, e.g. V_N = Ta + Tb + Tc (...)
    - c: Extend T1 in *RV* with node N to create a new labeled tree, $T1_N$:
    - Link node N to node *j*, where node *j* represents the variable to which variable N has the strongest correlation with
    - d: Determine the TEC to which the following set of trees belong to: $T1_N$ and the tree sequence resulting in *b:*
    - e: Rewrite V_N with the TEC resulting from *d:*
    - f: Fit a subgroup of regular vines in N nodes that have their first tree equal to $T1_N$ and belong to the V_N
    - g: Determine the best fit from the results in *f:* according to AIC goodness of fit test
    - h: Set *RV* to be the output in *g:*
  - **end**
- **end**
- Print(*RV*)

end

Algorithm 1: Sequential method to select a regular vine model based on the concept of Tree-equivalent classes (TEC)

4.3. Validation test

A mathematical validation of the hypothesis is out of the scope of this thesis. Nevertheless, we propose a test to validate the algorithm 1 presented in section 4.2.4 for our data set. The validation test proposed in here has two parts:

1. Use the algorithm to test a subgroup of regular vines in 6 nodes and to select the potential best fit

2. Compare the performance of the selected regular vine in 1) with the performance of the regular vine with lowest AIC (figure 4.6) and the performance of the regular vine with the strongest correlations in its tree-structures (figure 4.7)

The regular vine in 6 nodes selected using the novel algorithm belongs to the TEC V17 and is depicted in figure 4.9. The TEC V17 is an extension of the TEC V8 and V5. These are the TEC to which the regular vines in 5 and 4 nodes respectively with the lowest AIC belong to. The order of the nodes in the first tree of the selected regular vine maximizes the absolute sum of the correlation coefficients in all edges of T1. The values of these correlation coefficients are depicted in figure 4.9. The algorithm selected a subgroup of 36 regular vines that belong to V17 and comply with the aforementioned criteria in their first tree. Their AIC's are plotted in figure 4.10.

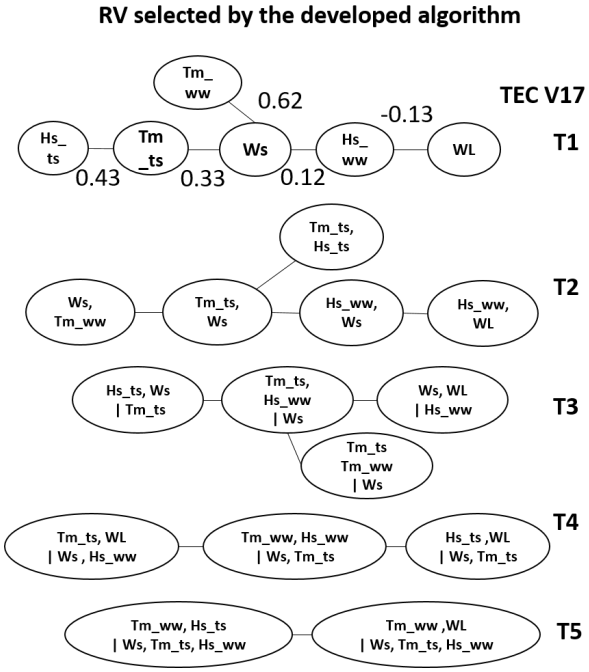


Figure 4.9: Regular vine selected with the developed algorithm for the data set used in chapter 5

The performance of the three regular vines is compared via their AIC values and the sum of the absolute difference between the modeled and original correlation coefficients (for all pairs of variables). The sum of absolute differences gives an indication on how well the dependence structures are represented in the model. These values are presented in table 4.4 for the three regular vines. The absolute difference between simulated and original correlation coefficients is calculated for each pair of variables and the resulting values are depicted in figure 4.11 for the three regular vines. The model achieving the lowest value for the sum of absolute differences predicts the correlation coefficients with the highest accuracy of the three. And the model achieving the lowest AIC is in theory, the model that "looses" the least information from the data. A discussion on what the "best" choice from these three possibilities is, is out of the scope of this thesis.

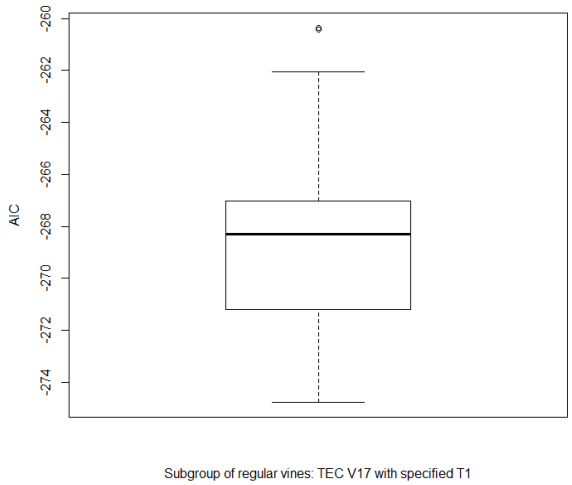


Figure 4.10: Box plot containing all the AIC values that results from the fitting of the chosen subgroup of regular vines to our data

	AIC	Sum of absolute differences of all correlation coefficients
Regular vine with lowest AIC	-290	0.3787
Regular vine with strongest correlations in its tree structures	-255	0.2612
Regular vine selected using novel algortihm	-274	0.3154

Table 4.4: Table presenting the performance of each regular vine according to the AIC goodness of fit and sum of absolute differences of all pairs's correlation coefficients.

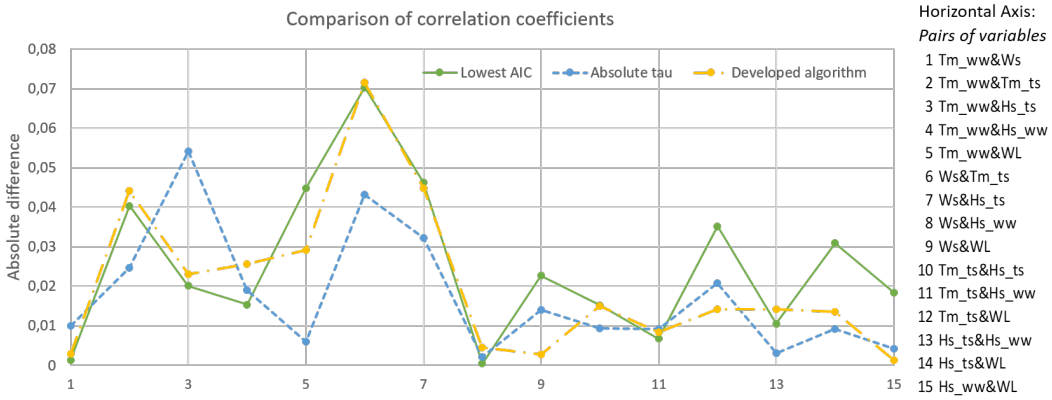


Figure 4.11: Comparison of absolute differences in the prediction of Kendall's τ for regular vine with the lowest AIC (lowest AIC), regular vine with the strongest correlations in its trees (Absolute tau) and regular vine selected by the novel algorithm (Developed algorithm).

In short, the algorithm presented in this chapter seems to select a reasonably good candidate to model the 6-variate data, constructed on specific characteristics of the regular vine in 5 nodes that achieves the lowest AIC. Despite the positive results, the regular vine in 6 nodes that represents best the correlation coefficients pf the design variables is chosen in the following chapter as best fit (see section 5.6). The results presented in this chapter are further discussed in chapter 6 and some conclusions are drawn in chapter 7.

5

Application: a case study in coastal engineering

The current chapter gives an answer to the main research question on "*How can we use vine-copula models in the design of infrastructure*" from a practical perspective. In here, the set of methods and techniques that compose the vine-based methodology for infrastructure design (*vine-based methodology*, hereinafter) are applied to a case study. This chapter aims to bring the work done in this thesis closer to the engineering community's interests.

To show the potential of this methodology, a hypothetical engineering application was presented: the design of a breakwater at the entrance of Galveston Bay, Texas. It is important to point out that the application is meant to be illustrative and it relies on arbitrary design assumptions. By applying the methodology in a potential day-to-day project, we aim to highlight the advantages that this methodology provides when compared to more traditional methods. Not only advantages that have the potential to make the design more economical, but also advantages that provide the engineer with more information on the risk of the overall design. The last point gives the engineer the opportunity to create the optimal design for the situation.

5.0.1. Summary of the vine-based methodology for infrastructure design

Before diving into the application, a short summary of the vine-based methodology is presented below. The main steps and structure of the developed methodology are depicted in figure 5.1.

1. STEP 1: *Extreme Value Analysis*. Normally, infrastructure is designed to withstand extreme conditions. To find these, the practitioner should perform a Peak Over Threshold (POT) on the dominant variable that is assumed most relevant to the design according to some criteria. The variables that are observed together with the dominant variable during an extreme event, the *concomitants*, are sampled using block maxima technique. These sampling procedures lead to the so-called *extreme sample*. The next task is to fit a distribution to the dominant and concomitant variables in the extreme sample.

2. STEP 2: *Bivariate dependence modelling*. The goal of this step is to gain insight into the physical behavior of the system. By performing this step one can ensure that the statistical results are in accordance with what is expected from a physical point of view. The bivariate dependence is studied in two manners: (1) by analyzing correlation coefficients, and (2) by analyzing the dependency structure with bivariate copulas.

3. STEP 3: *Multivariate dependence modelling*. The goal of step 3 is to choose a regular vine-copula to model the joint behavior between the design variables.

4. STEP 4: *Derivation of Multivariate design values*. This step is meant to be the link between the pure statistical analysis and the engineering design process. The objective of this step is to derive the design values using the results in steps 1 and 3. The design variables are imposed based on their univariate return periods and their corresponding multivariate return period or associated probability of exceedance are calculated with the vine-copula selected in step 3. To do so, one needs to select a risk scenario that represents the system dynamics.

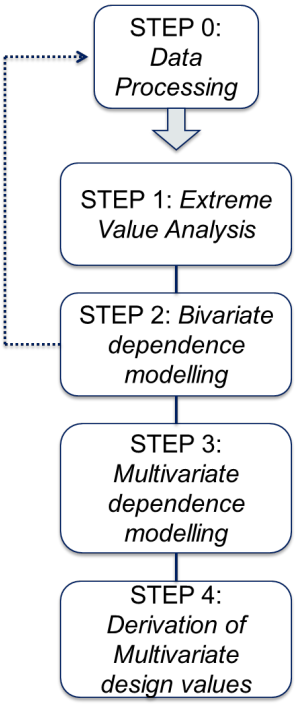


Figure 5.1: Main steps comprising the vine-based methodology. Source: Author

5.1. Case study

Two locations were of interest as case study in this research: (1) Tabasco, México, and (2) Texas, the US. The initial proposal was to perform the multivariate analysis in Tabasco's coastal waters and the Grijalva's river. However, the lack of available data seemed to make it difficult to fulfill some of the research objectives. Therefore, the preferred location is (2) in the area around the Galveston Bay, which is connected to the Gulf of Mexico. Galveston Bay (in figure 5.2) is the second largest estuary in the Gulf of Mexico. It has a surface area of 1600 km^2, is 50 km long, and is 27 km wide. The bathymetry is relatively flat with a mean depth of 3 m, except in the northern entrance (Houston Ship Channel), where a 12 m deep channel is located [24]. The bay has an intertidal range of 0.5 m. The connection to the Gulf of Mexico is via two inlets (southern entrance and northern entrance) and has two major freshwater sources, the San Jacinto and Trinity rivers.

The data that was available for each of the variables of interest is presented in tables B.1 and B.2 in Appendix B, for Tabasco and Texas respectively.

5.1.1. Description of the engineering project

At the area of interest (illustrated in figure 5.3), a considerably long breakwater is located along each side of the shipping channel that provides an access to the Galveston Bay. Supposedly, the breakwater was constructed several years ago and the respective client requested a re-design and possible maintenance of both breakwaters. In order to fulfill the client's requests, the design conditions at the breakwaters location must be determined.

According to [17], the location and characteristics of the Galveston Bay make it prone to the co-occurrence of riverine and coastal floods. The area is exposed to intense rainfall events from local convective storms, large-scale frontal systems, and torrential rainfall brought by tropical cyclones. These meteorological events also have an effect on the sea-variables. Meaning, the coast of interest is exposed to considerably high energetic sea states [63]. The breakwaters are exposed to extreme water levels triggered by the compound effect of riverine and sea variables. Thus, the design conditions must be of multivariate nature and hence, a multivariate frequency analysis is required.

Usually, the design values are thought for specific design criteria. In this application, it is assumed that large wave heights that occur together with large water levels are critical to the stability of the breakwater and

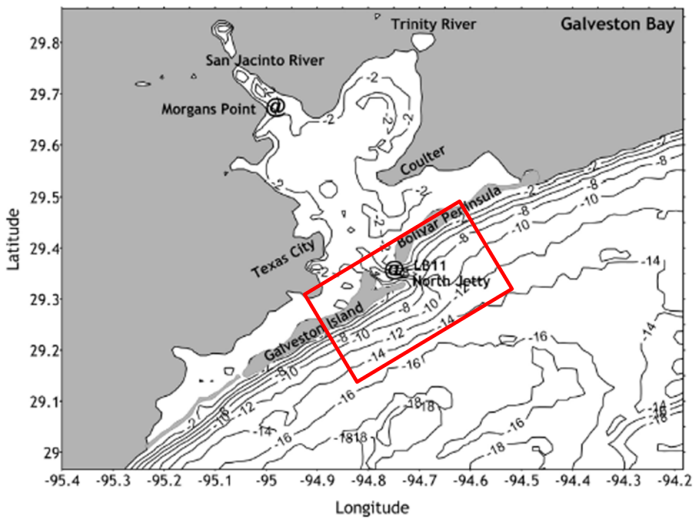


Figure 5.2: Galveston bay, area of interest within Texas and the Gulf of México. Encircled in red, the location of the case study. Source: [63]

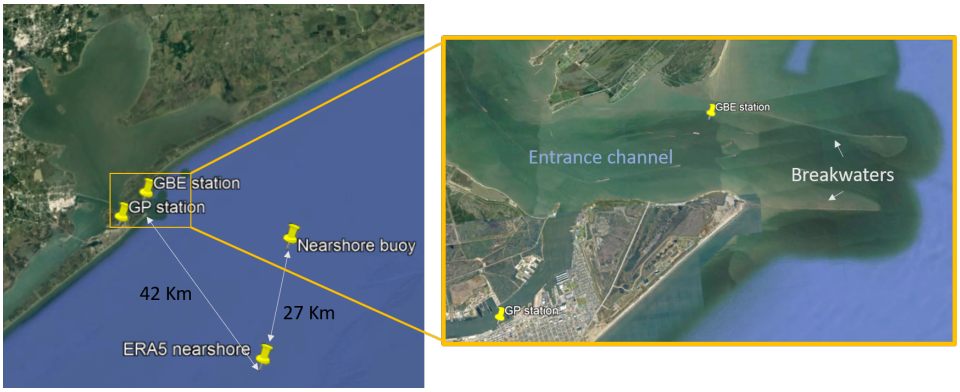


Figure 5.3: This figure illustrates the engineering project accounted for in section 5.1.1. It also presents the distances between the two points of data. Source: Google Earth

to the maximum allowed overtopping discharge. Thus, we render the significant wave height of wind waves the most relevant to the design. It is also in our interest to determine what wind speeds occur simultaneously with combinations of large waves and water levels. At the end of the current chapter, we introduce a simplified application on the design of the breakwater's crest level to illustrate some of the advantages of the vine-based methodology.

5.2. Data collection

5.2.1. Background

Characterization of wave climate at a local scale for offshore and coastal design requires reliable data which is usually not available. Depending on the available budget and richness of wave data in the location, a range of approaches exist for obtaining such data. These include long term and/or short term deployment of wave buoys and subsequent hindcasting to give a sufficiently long record of the local wave climate. Most of the times, wave data must be reconstructed from different sources of information. Generally, one could classify data in base of the source and perceive two groups:

1. Measurements

    - Insitu techniques: the most common in situ instruments are wave buoys and wave poles [31]. Other in situ instruments are inverted echo-sounders, pressure transducers and current meters. These instruments require to be mounted on some structure at the sea location.

- Remote-sensing techniques: The most common remote-sensing technique is radar [31], which is based on actively irradiating the sea surface with electro-magnetic energy and detecting the corresponding reflection.

2. Numerical modeled data

   - Numerical models provide data that is continuous in space and time. Normally, complete information (i.e. wave spectrum) is provided, but often, results in underestimation of wave conditions in enclosed basins [9].

Nowadays, there are several global (atmospheric, among others) reanalysis databases. One of them is ERA5 [14], which is being developed through the Copernicus Climate Change Service (C3S). The ERA5 database is freely available for scientific purposes and includes a wide range of variables. ERA5 data is available through the Climate Data Store (CDS). However some ERA5 datasets do not appear in CDS but are accessible through CDS API (refer to [14]). For example, the ERA5 wave spectra data. The entire data base is separated in 'levels'. The 'level' of interest for this thesis is "*ERA5 hourly data on single levels from 1979 to present*". The *atmospheric* reanalysis has a spatial coverage of 0.25°x0.25°, the reanalysis of *ocean waves* has double the resolution, 0.5°x0.5°. These resolutions represent approximately 27 and 55 kilometers, respectively.

Another public source of simulated data is the eartH2Observe Water Cycle Integrator (WCI) [1]. The WCI portal is an open source project built by Plymouth Marine Laboratory's (PML) Remote Sensing Group. The portal builds on the development of several other EU funded projects, past and present, that PML have involvement in. The WCI data set can be used to obtain river discharge data, with CSIRO model. Unfortunately, the temporal resolution of the measurements (i.e. monthly) is coarser than the rest of the sources and hence, this source is not used in this thesis. For more information, the reader is referred to [1].

On another note, measurements for atmospheric and oceanic variables are being collected world wide by the National Oceanic and Atmospheric Administration (NOAA). This source provides *quasi-2D* wave spectrum, wind data, atmospheric data and water levels. Some of these have been used in this thesis (refer to section 5.2.2).

5.2.2. Data sources

Hourly water levels were downloaded from the National Oceanic and Atmospheric Administration (NOAA) website *(https://tidesandcurrents.noaa.gov)* for station IDs 8771450 (Galveston Pier, *GP* hereinafter) and 8771341 (Galveston Bay Entrance, *GBE* hereinafter). At both stations, one can calculate hourly non-tidal residuals by subtracting the measured water level from the predicted astronomical tide. Usually, non-tidal residuals (water levels) are called 'storm surge' within the coastal engineering community. Regardless, the interest within this thesis is in 'total' still water levels (WL), including the tide and the surge but without waves. One could discuss whether is statistically correct to sample from the 'total' still water level population, including the deterministic component that is the tide. However, the storm surge is a random variable that is summed on top of the tide. Thus, WL can be characterized as a random variable.

The GBE tide station has a limited record length: about 16 years worth of data scattered between 2001 and 2018 excluding 2009 and 2010. The GP location has 113 years of data from 1904 to 2018. In order to obtain the water level at Galveston Entrance Channel (WL_{GP}) site from Galveston Pier 21 (WL_{GBE}) location, the following linear regression model was used:

$$WL_{GBE} = 1.0029 \times WL_{GP} + 0.0038856 \tag{5.1}$$

The linear regression was fitted based on the joint observations ($R^2 = 0.91$) of the simultaneous hourly water levels, the equivalent of 16 years. According to the Two-sample Kolmogorov-Smirnov test discussed in [45] (performed with *kstest2()* function in MATLAB), the two population samples belong to the same continuous distribution. On another note, there is a (water level) data gap between April 1984 and October 1984. These time period is excluded from the analysis, and hence is not be considered for the remaining variables.

Wave data was downloaded from ERA5 reanalysis database (see section 5.2.1). The temporal resolution of the wave data is hourly for the period 1979 to 2018. ERA5 provides wave data that has already been 'splitted' into the two main wave components, *total swell* and *wind generated waves*. The term *total* includes all the swell partitions that occur at the specific location. The variables of interest (which data was downloaded) are the significant wave height, the mean wave period and the wave direction for the total swell and the wind generated

waves, separately. The model is of spectral nature, meaning the variables' values have been calculated from the total sea spectrum.

Atmospheric data (i.e. wind speed) and direction, has also been collected from ERA5 database. Hence, this data has similar properties than the wave data. The only difference is the spatial resolution of these: the grid for wave model double is coarser than for the atmospheric model (see table B.2 in appendix B).

The chosen ERA5 grid point is located approximately 42 kilometers away from the location of interest (see the project location in figure 5.3). The coordinates of the aforementioned data point are 29.0 degrees North and 265.5 degrees East. The bathymetry that surrounds this point is considered to be fairly uniform (refer to figure 5.4). The location has a water depth of around 20 meters and is considered to be transitional (shallow) waters. This assumption is necessary to support the performance of the *Extreme Value Anaysis* (*EVA*) at that location. The 'extreme' waves there must not have reached their breaking point in order to capture them while performing the EVA. If the waves were already broken, the results of the EVA would be underestimating the extreme wave climate at the location of interest. This would result in an underestimation of the design values.

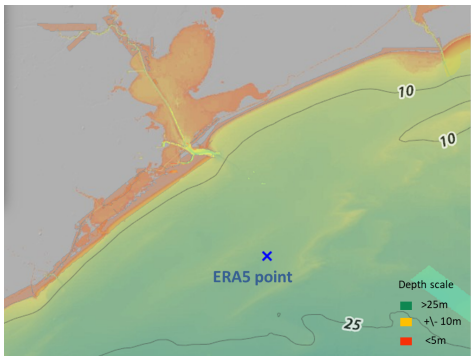


Figure 5.4: Bathymetry of the Galveston bay, the area of interest for this thesis within Texas at the Gulf of México. Source: [70]

5.2.3. Multivariate data set

In this application, the multivariate data set comprises hourly observations of the variables: water level, significant wave height of swell and wind waves, mean wave period of swell and wind waves and wind speed (see figure 5.5). The aforementioned variables are understood as explained in the following list. The variable *water level* (WL) has already been defined in section 5.2.2.

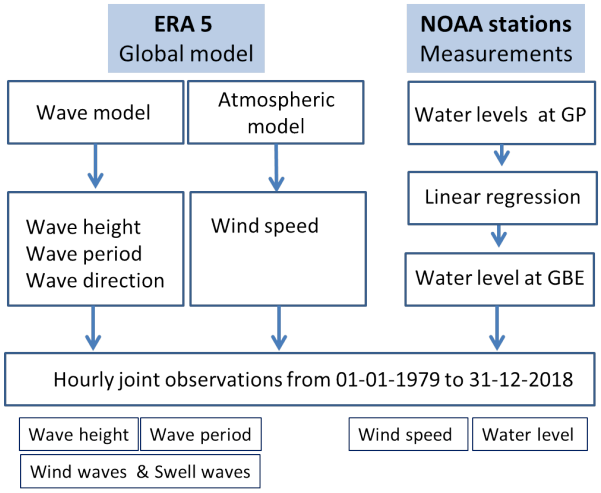


Figure 5.5: This figure presents the used multivariate data set and the variables sources. Source: Author

- *Mean wave period* (Tm). This variable is defined as the average time it takes for two consecutive wave crests, on the surface of the ocean/sea, to pass through a fixed point. The ocean/sea surface wave field

consists of a combination of waves with different heights, lengths and directions (known as the two-dimensional wave spectrum). The mean period is a mean over all frequencies and directions of the two-dimensional wave spectrum for wind waves (Tm_{ww}) and for total swell waves (Tm_{ts}).

- *Significant wave height* (Hs). For wind generated waves, the significant wave height (Hs_{ww}) is defined as four times the square root of the integral over all directions and all frequencies of the wind waves spectrum. The wind waves spectrum is obtained by only considering the components of the two-dimensional wave spectrum that are still under the influence of the local wind. For swell waves, the significant wave height (Hs_{ts}) is four times the square root of the integral over all directions and all frequencies of the total swell spectrum. The total swell spectrum is obtained by only considering the components of the two-dimensional wave spectrum that are not under the influence of the local wind.

- *Wind speed* (Ws). The wind speed is defined as the 10 meters neutral wind speed from the atmospheric model of ERA5, which at the same time is determined from the atmospheric surface stress.

5.3. STEP 0. Data Processing

Step 0 is the input to the methodology and is not extensively treated in here. Mainly because the level of data processing differs per application. In here, the wave components are already decomposed in sea states: swell and wind generated waves (wind waves). The output of step 0 is a multivariate time series that contains joint observations of all variables of interest. An example on the format is presented in figure 5.6 for our application.

Date and time	Water Level (m)	Wind speed (m/s)	Wind-waves wave height (m)	Swell-waves wave height (m)	Wind-waves wave period (s)	Swell-waves wave period (s)
1-1-1979 00:00	0.19	5.56	0.19	0.92	2.04	5.14
1-1-1979 01:00	0.21	9.65	0.99	0.26	3.86	4.22
1-1-1979 02:00	0.18	7.32	0.64	0.25	3.29	3.91
1-1-1979 03:00	0.13	2.52	0.00	0.52	3.86	4.68
1-1-1979 04:00	0.10	12.65	1.97	0.66	5.02	6.79
1-1-1979 05:00	0.11	5.30	0.26	0.32	2.37	4.07
1-1-1979 06:00	0.08	7.53	0.84	1.09	3.82	5.92
1-1-1979 07:00	0.04	8.34	0.64	0.63	3.11	4.96
1-1-1979 08:00	0.05	4.59	0.15	0.62	1.95	5.36
1-1-1979 09:00	0.06	2.64	0.00	0.87	3.86	5.54
1-1-1979 10:00	0.10	2.32	0.00	0.62	3.86	4.72
1-1-1979 11:00	0.02	10.10	1.10	0.42	3.95	5.86
1-1-1979 12:00	-0.09	7.84	0.63	0.49	3.41	4.89
1-1-1979 13:00	-0.19	9.60	1.15	0.69	4.37	6.74
31-12-2018 17:00	-0.13	7.25	0.68	0.47	3.54	4.78
31-12-2018 18:00	-0.129	3.32	0.08	0.60	1.63	4.82
31-12-2018 19:00	-0.078	9.98	1.13	0.36	4.13	5.38
31-12-2018 20:00	0.036	5.84	0.35	0.58	2.65	4.50
31-12-2018 21:00	0.157	0.00	0.00	0.00	0.00	0.00
31-12-2018 22:00	0.318	0.00	0.00	0.00	0.00	0.00
31-12-2018 23:00	0.334	3.57	0.04	0.78	2.90	5.74

Figure 5.6: Example showing part of the multivariate data set analyzed in this application

5.4. STEP 1. Extreme Value Analysis

The significant wave height of wind waves was set to be the *dominant* variable in section 5.1.1. Thus, a POT is performed on the Hs_{ww}. The choice of threshold (or thresholds) is based on the theory presented in section 2.1.1. Subsequently, the physical and statistical thresholds are defined as follows:

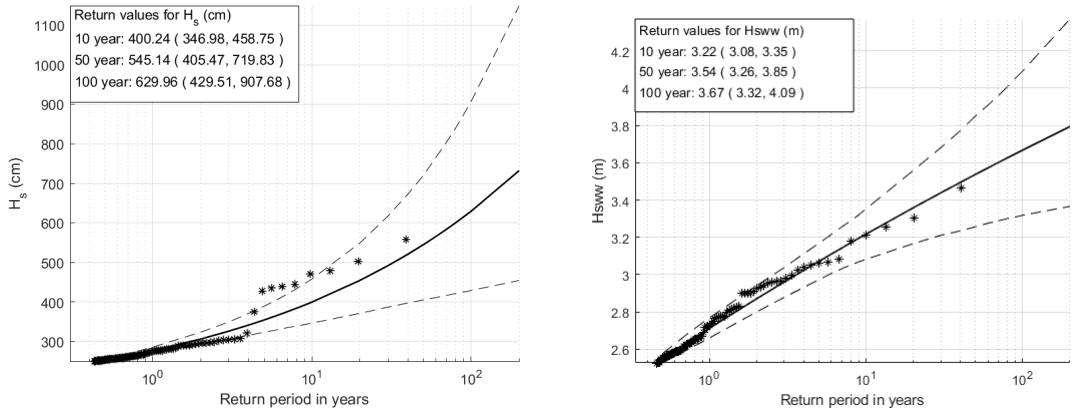
- The physical declustering procedure has been performed so as to obtain a sample of 3 (extreme) storms per year in average, which is a physically sounding number of extreme events per year. The 'physical threshold' (or duration of the event) has been set to 4 days to achieve approximately an average of 3 (extreme) storms per year.

- The statistical threshold has been defined by studying the shape parameter k (see equation 2.1), aiming the sample to converge to a GPD. It has been observed that the best results are achieved with k being negative and as close to zero as possible. Statistically, this value of k implies the distribution to be shaped (slightly)

convexly. Physically, this implies a steady increase rate that converges in a physical boundary. The choice of threshold is 2.45 meters, which together with the 'physical' threshold above, leads to a sample of 128 extreme observations.

Next, a General Pareto Distribution (GPD) is fitted to the resulting sample of extreme wave heights. The fitting results are presented in figure 5.7a. In this figure, the reader can notice something odd in the fitting: it seems there are two different wave systems present in the data and one of these with considerably large extreme waves. These waves are caused by hurricanes and, for the sake of simplicity, the hurricane generated waves are removed from the data. These are assumed to reach wave heights larger than 3.5 meters. The EVA has been performed again for the data without the presence of hurricanes. The resulting wave data without hurricanes is used in the analysis. If the practitioner would prefer to include hurricane data, he/she should treat these values as a separate variable. Such as is done with swell and wind generated waves. In this way, the resulting sample would contain independent data from only one 'type' of meteorological phenomenon.

The results on the POT on Hs_{ww} are presented in figure 5.8. The GPD fit is plotted in figure 5.7b. The fitting results in figure 5.7b are considerably better that the fitting results in figure 5.7a.



(a) GPD representing the best fit for the extreme wave heights in the location of interest including hurricanes. The plot also presents some return values and their 95% confident bounds.

(b) GPD representing the best fit for the extreme wave heights in the location of interest excluding hurricanes. The plot also presents some return values and their 95% confident bounds.

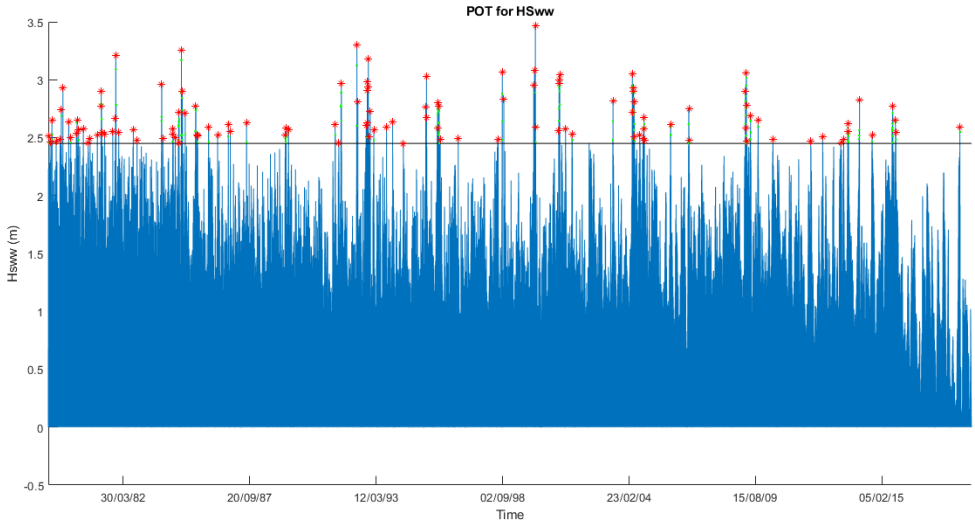


Figure 5.8: POT results on the significant have height of wind waves

The concomitant values are sampled with the block maxima method for a time window (extreme event) of 1 day. This means these concomitant values occur within maximum 1 day from when the extreme wave height of

wind waves occurred. This choice was made so the correlation coefficient values between the dominant variable and the concomitants were maximized. This is further explained in step 2 in section 5.5. The concomitant values have been fitted to several popular distributions. The best fit is considered to be the distribution with the lowest AIC. The fitting results for the concomitant variables are presented in the third column of table 5.1. Two of the concomitant distributions in table 5.1 are extreme value distributions. The samples from the variables WL and Tm_{ts} seem to comply with the asymptotic properties of extreme observations. However, this does not necessarily mean that extreme values from these two variables occur together with extreme wave heights of wind waves. This is further explored in step 2.

In parallel, we perform an EVA to all the variables to determine their extreme distributions. The POT method is applied to the variables, and then, the extreme value distribution is computed. The results of this task are important for step 4 when determining the design values for the *traditional approach*, where the variables are assumed to be independent. The second column in table 5.1 presents the results of EVA performed individually to all variables. That is why the heading of this column refers to *Extreme distribution*.

Variable	Extreme distribution	Concomitant distribution
Hs_ww (m)	Generalized Pareto k= -0.049 sigma= 0.24 theta= 2.45	(Extreme)
WL (m)	Generalized Pareto k= 0.1581 sigma= 0.985 theta= 0.6530	Generalized Pareto k= 0.1413 sigma= 0.1018 theta= 0.6260
WS (m/s)	Generalized Pareto k= 0.2141 sigma= 0.8207 theta= 15.0068	Logistic mu= 10.9638 sigma= 1.3871
Hs_ts (m)	Generalized Pareto k= -0.0241 sigma= 0.2464 theta= 1.7758	Lognormal mu= -0.0204 lambda= 0.3285
Tm_ww (s)	Generalized Pareto k= -0.0431 sigma= 0.5413 theta= 6.1505	Logistic mu= 4.7088 sigma= 0.4
Tm_ts (s)	Generalized Pareto k= -0.1418 sigma= 1.4408 theta= 9.0264	Generalized Extreme Value k= 0.0303 sigma= 0.8521 nu= 6.1379

Table 5.1: This table presents the distributions that were determined to be the best fits to the univariate data. The second column presents the results of EVA performed individually to all variables. The third column presents the best univariate distribution for the concomitant variables.

5.5. STEP 2. Bivariate dependence modelling

In this step, we aim to gain insight into the physical behavior and associated dependencies between variables with Kendall's τ (see section 2.2.1) and bivariate copulas (see section 2.2.2). If the results of step 2 are not satisfactory, the practitioner can go back to step 0 and start the analysis again. This is depicted in figure 5.1. By performing this step one can ensure that the statistical results are in accordance with what is expected from a physical point of view.

5.5.1. Analyzing correlation coefficients

From the extreme data set resulting from step 1, the correlation matrix is calculated. This can be done in R with the command *cor(DATA, method= "kendall")*. The correlation coefficients for our data are presented in figure 5.9. The size of the circles indicate the strength of the correlation. The blue circles represent positive correlation and the red ones negative correlation. This plot gives the practitioner a practical tool to quickly spot the type of correlations present in the system together with their strength.

In the previous step (in section 5.4), we stated that extreme values of WL and Tm_{ts} might not occur together with extreme values of Hs_{ww}. The correlation results in figure 5.9 seem to support that assumption. Specially for WL, because the correlation between WL and Hs_{ww} is slightly negative. Physically, this negative correlation can be explained with the shoaling effect of waves when propagating to shallower waters. The data location has an average depth of around 20 meters and hence, it is considered to be in transitional waters where the largest waves are already shoaling. The shoaling process is related to the water depth: the shallower it is, the more shoaling occurs until the waves break. So in our area of interest, the waves heights are the largest when the water depth is the smallest. Hence, the negative correlation.

Figure 5.9: Correlation matrix for the dominant and concomitant variables

It is also interesting to plot the bivariate observations, in a scatter plot, to further analyze the dependence structures. These joint observations can be plotted conserving the physical units, or can be transformed to 'probability' units and become *pseudo-observations*. Traditional bivariate distributions only allow two random variables that behave according to the same theoretical distribution. Normally, ocean variables behave accordingly to different theoretical distributions. An example of this is presented in the fitting results in table 5.1. By using copulas (and vine-copulas), the dependence structure between a pair of random variables may be specified independently to that of their one dimensional marginal distributions. This is achieved with the transformation to *pseudo-observations*. The transformation can be done empirically via a ranking procedure (see section 2.2.2) and then, the values are normalized from 0 to 1. Or it can be done theoretically by fitting a univariate distribution. The pseudo-observations resulting from the ranking procedure are always distributed uniformly. These are needed as input to the copula and vine-copula models. The dependency structure is more easily spotted with pseudo-observations than real observations. This can be seen when comparing figure 5.10 for pseudo-observations and figure 5.11 for real observations. In these figures the univariate histograms and the Kendall's correlation coefficients are also plotted.

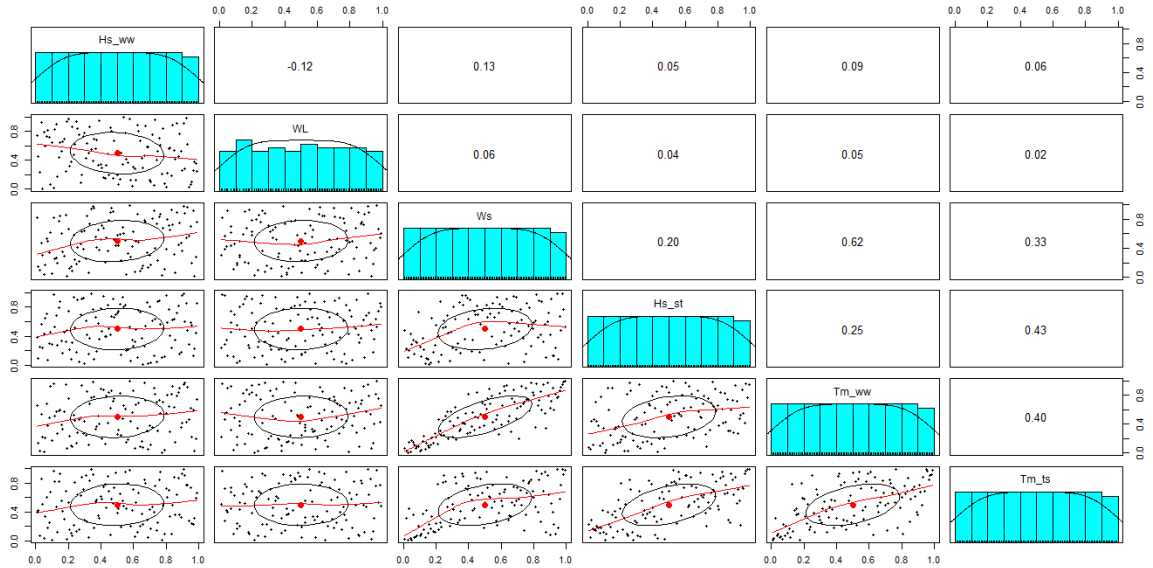


Figure 5.10: Scatter matrix with plotted pseudo-observations for all possible pairs of variables, in the lower triangle. The red lines provide an estimation of the trend and the ellipse depicts areas with the largest mass concentration. The diagonal contains the univariate histograms which for pseudo-observations are uniform. The upper triangle depicts the Kendall's τ for each pair of variables.

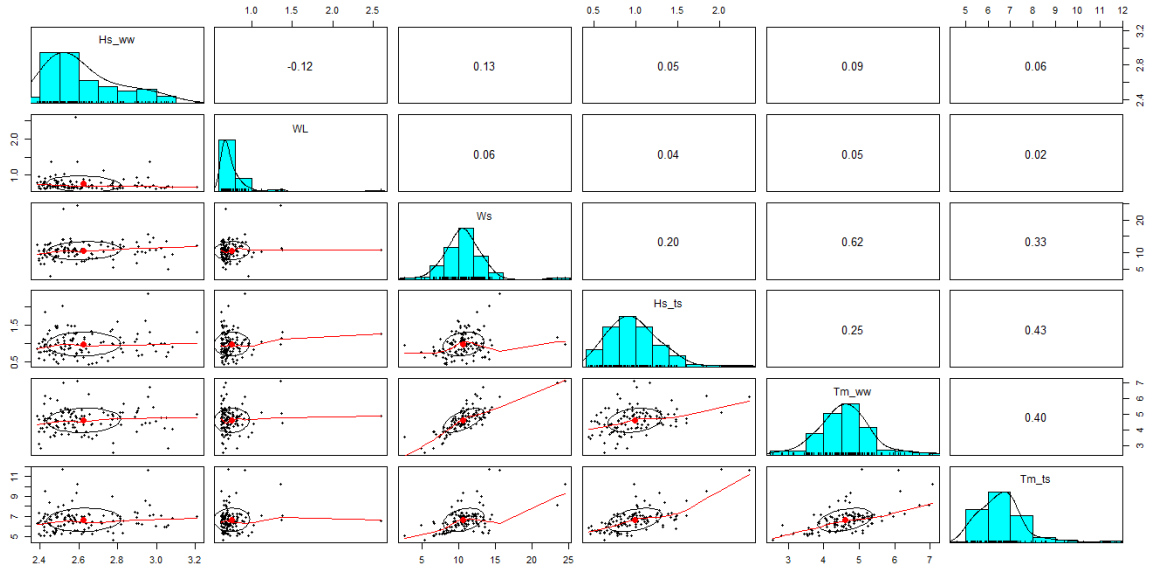


Figure 5.11: Scatter matrix with plotted pseudo-observations for all possible pairs of variables, in the lower triangle. The red lines provide an estimation of the trend and the ellipse depicts areas with the largest mass concentration. The diagonal contains the univariate histograms. The upper triangle depicts the Kendall's τ for each pair of variables.

The largest correlation is achieved between Tm_{ww} and Ws. This statistical result also makes sense from a physical perspective: the wind generates the wind waves. On the contrary, the correlation between Tm_{ts} and Ws and Tm_{ts} is lower than the previous. From a physical perspective, swell waves are not generated by local winds and hence, the relatively low correlation value. However, it is still considerably high considering that this system should be independent from the wind waves. This might suggest that the wave data might not have been splitted completely. Nevertheless, the Gulf of México can be considered an enclosed sea, which means that the swell might be generated from the same meteorological phenomena. In this kind of cases, the practitioner should decide if the statistical results are consistent enough with what the practitioner expects from a physical perspective. If that is not the case, the practitioner should go back to step 0 (data processing) to achieve better final results. For this application, it is assumed the data has been splitted good enough.

5.5.2. Analyzing bivariate dependence structures

The bivariate observations from all pairs of variables are fitted to theoretical bivariate copulas. The fitting has been performed in R with the command *BiCopSelect()*. By plotting the densities of the resulting theoretical copulas (depicted in figure 5.12), the practitioner can identify dependency structures that might be of interest in the analysis. For example, the practitioner can study whether the variables present tail dependence (refer to section 2.2.3).

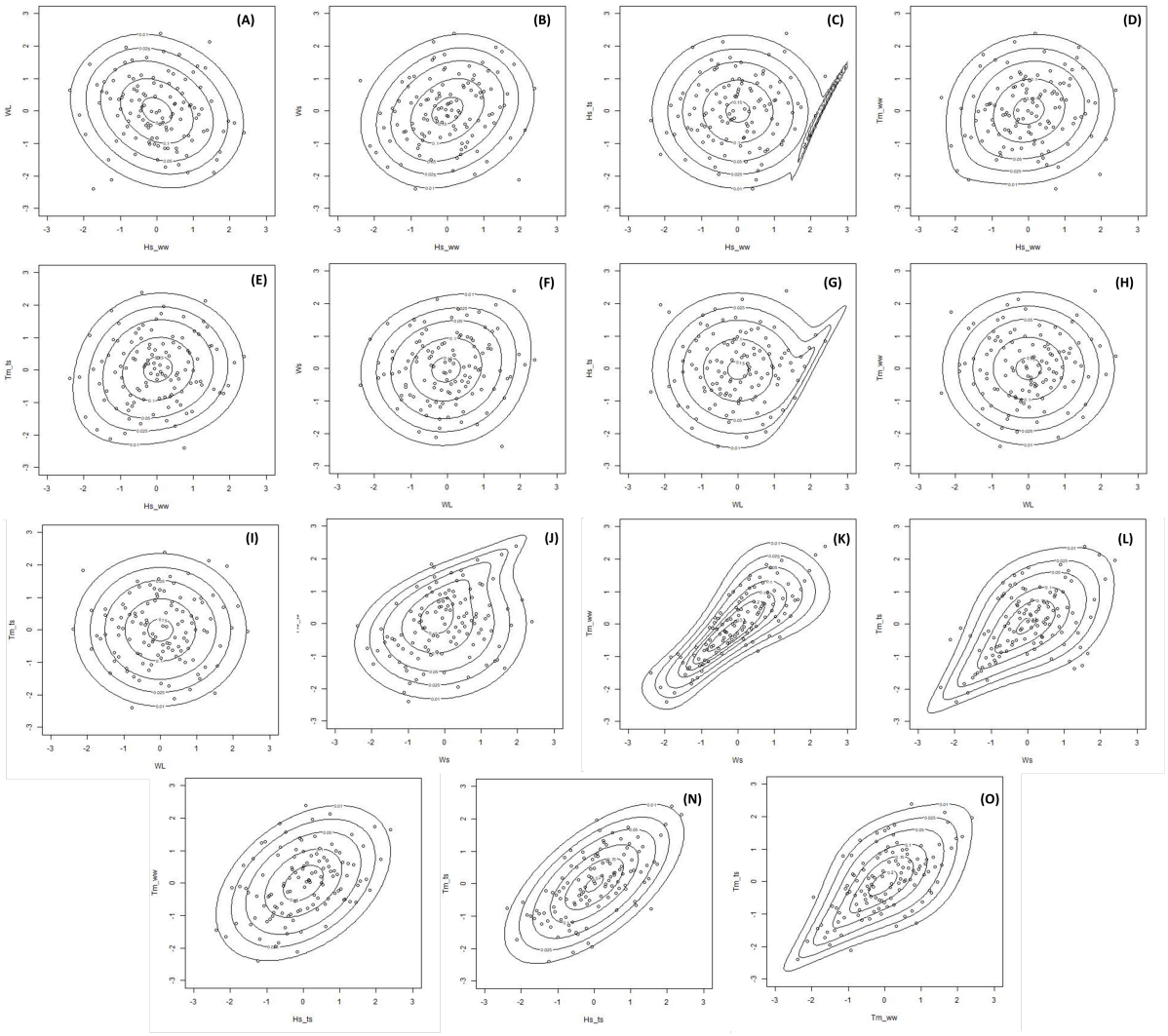


Figure 5.12: Theoretical copula densities for all pairs of observations in the extreme sample. The pseudo-observations are presented in standard normal units.

The dependency structure of copula *C)* (in figure 5.12) presents some asymmetries when values of Hs_{ww} and Hs_{ts} are large. The same occurs with copula *G)*, for WL and Hs_{ts}. This would not have been considered and neither included in a univariate analysis.

The densities of the copulas that in standard normal units resemble circles present variables that have very low correlation values. Examples of such are copulas *H)* and *I)*. These model the density between WL and the wave periods for swell and wind waves, Tm_{ww} and Tm_{ts}. Physically, these variables are not related directly.

On another note, copulas *J)*, *L)* and *O)* (in figure 5.12) present some tail dependence. It is most interesting for copula *J)* which seems to present an upper tail dependence, thus when both variables Hs_{ts} and Ws achieve large values. In the previous section, we detected some anomalies in the correlation results from Ws and Tm_{ts}. However, we assumed the wave data was splitted good enough in swell and wind waves sea states. If the larger waves in Hs_{ts} were wind waves, the spotted upper tail dependence between Hs_{ts} and Ws would make sense from a physical perspective, as large (local) wind speeds generate large wind waves. Thus, these results would mean the largest swell waves are indeed wind waves and would imply the data has not been splitted well enough. Regardless of this results and for the sake of simplicity, we continue to assume the data is splitted well enough for this fictional analysis.

The results presented in this section provide the reader with practical examples on the use of step 2 to check the quality of the data.

5.6. STEP 3. Multivariate dependence modelling

In step 3, we choose a regular vine-copula to model the joint behavior between the variables of interest. In this application, all existing regular vines in 6 nodes are fitted to the data. The approach to do so is explained in section 3.3 and theory on regular vine-copula models is presented in section 2.3.

In section 4.2.4 we present a discussion on which regular vine would be the best to model our data. We present two reasonable choices (depicted in figure 5.13): (1)the regular vine in 6 nodes with the lowest AIC and (2)the regular vine in 6 nodes with the strongest pair-wise correlations in its trees. The latter seemed to predict the correlation coefficients more accurately (refer to figure 4.8), despite achieving a larger AIC. The Kendall's correlation coefficients (τ) are presented for the first tree (T1) in both regular vines. One can see in figure 5.13 that the absolute sum of all τ in T1 is the smallest in the regular vine with the lowest AIC.

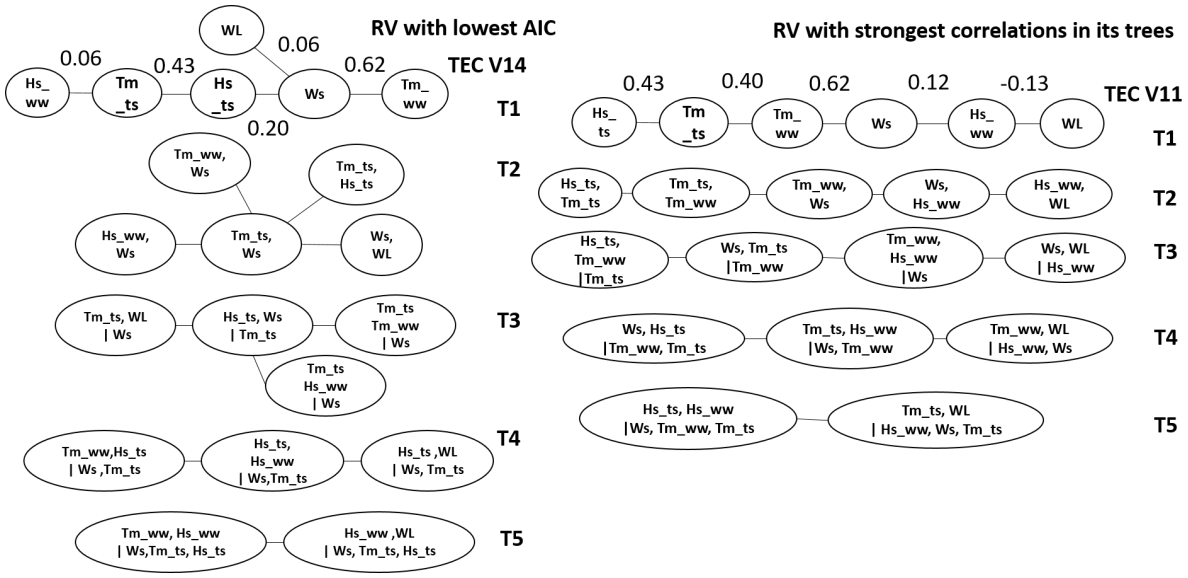


Figure 5.13: On the left, the regular vine (RV) in 6 nodes with the lowest AIC is presented. On the right, the regular vine (RV) in 6 nodes with the strongest pair-wise correlations in its trees is presented. The Kendall's correlation coefficients (τ) are presented for the first tree (T1) in both regular vines.

From a physical perspective, the variable's order in T1 of the regular vine with the strongest correlations (on the right in figure 5.13) seems more logical than the one of the regular vine with the lowest AIC (on the left in

figure 5.13). For example, one might expect that the Ws is directly connected with Hs_{ww} rather than with Hs_{ts}. Moreover, paring Hs_{ww} with WL seems to be physically more logical than paring Hs_{ww} with Tm_{ts}. Specially since the correlation between the first two variables is stronger than for the last pair. Hence, an engineer might feel more comfortable selecting the regular vine with the strongest pair-wise correlations in its trees. On the contrary, an statistician might feel more comfortable selecting the regular vine with the lowest AIC. The real and true best model for the data remains unknown.

For this application, we choose to model our data with the regular vine in 6 nodes with the strongest pair-wise correlations in its trees. This regular vine belongs to the tree-equivalent class V11 and it is a D-Vine. The model is validated according to three tests proposed at the end of section 3.3 and the results are presented in appendix D.

One of the main disadvantages of the methodology is most probably the large computational effort when fitting all the regular vines to the data in order to choose the best model. The computational time needed to fit all existing regular vines in 4,5 and 6 nodes to our data set for a regular laptop is presented in table 5.2. The analyses with 5 and 4 variables are an essential part of chapter 4, hence the resulting best models are presented in there and not in here.

Number of variables	Computational time [s]
4	50.13
5	1541.42
6	186780.44

Table 5.2: This table presents the computational time in seconds that a regular laptop took to fit all the existing regular vines in 4,5 and 6 nodes to the data.

5.7. STEP 4. Derivation of multivariate design values

Step 4 is meant to be the link between the pure statistical analysis and the engineering design process, and it is one of the main contributions of this thesis. The objective of this step is to derive the multivariate design values using the results in step 1 (Extreme Value Analysis) and step 3 (multivariate dependence modelling with vine-copulas). The description of this step is presented in section 3.4 and the underlying theory in section 2.4. In the current section, we focus on applying these concepts to derive the 6 design values for Hs_{ww}, Hs_{ts}, Tm_{ww}, Tm_{ts}, Ws and WL.

The univariate return periods (T_i) of interest are the following: 10, 50, 100, 500 and 1000 years. The reader should note that in practice, these are usually pre-defined by the client. The aforementioned return periods define individual risk levels that are associated with probabilities of exceedance ($1/T_i$): 0.1, 0.02, 0.01, 0.002 and 0.001. The practitioner can calculate the (univariate) design values with equation 3.1. In this equation, we see that the design values depend on the selected marginal distributions. The choice of marginal distribution depends on whether one is designing for independent or dependent variables.

The probabilities of exceedance are computed based on a *risk scenario* that represents the system dynamics. The choice of which scenario to use should be done a priori, based on physical knowledge of the system analyzed. In this application, we calculate probabilities associated to AND and OR scenarios as an example. Engineers might be interested in one or the other, or a mixed scenario. These risk scenarios should not be compared as they represent different system dynamics.

Assuming independence between design variables is a common practice in the ocean and coastal engineering fields. Thus, an approach in which the variables are considered independent is referred in here as *traditional approach*. This approach is depicted in the 'left vertical flow-path' in figure 2.1. On the contrary, the methodology developed in this thesis assumes multivariate dependence and it is referred in here as *vine-based methodology*. This novel approach is depicted in the 'right vertical flow-path' in figure 2.1. To highlight the advantages that the vine-based methodology provides when compared to the traditional method, the design values are derived using the traditional approach and the vine-based methodology. The first is treated in section 5.7.1 and the last in section 5.7.2. A comparison between both approaches is presented in section 5.8. First, the design values derived by both methods are briefly compared. Then, the results of both approaches are applied in a simplified example on probabilistic design of a breakwater's crest level.

5.7.1. Traditional approach

The traditional approach is mainly composed of two steps: (1) EVA and (2) Univariate derivation of design values. The first step is similar to step 1 of the vine-based methodology. However, instead of treating the variables as a system, these are treated separately because are assumed to be independent. Thus, an EVA is performed to each variable individually, and there is no need to define a dominant variable and its concomitants. Coming back to our application, 6 EVA are performed individually to each variable. The resulting extreme distributions are presented in the second column of table 5.1.

For the second step, we go back to equation 3.1 in section 3.4. We already stated that the resulting design values depend on the selected marginal distributions. In this case, these are extreme distributions. With the univariate return periods and the extreme distributions, the univariate design values are easily calculated with equation 3.1.

The five sets of 6 (univariate) design values in tables 5.3 and 5.4 represent extreme events. For a chosen AND-risk scenario, the multivariate return period, T_{and} (see definition in equation 2.17), and the corresponding AND-probability of exceedance of the aforementioned 5 extreme events are presented in table 5.3. For a chosen OR-risk scenario, the multivariate return period, T_{or} (see definition in equation 2.18), and the corresponding OR-probability of exceedance of the aforementioned 5 extreme events are presented in table 5.4.

Return Period (years)	Univariate Exceedance Probability	Hs_ww (m)	WL (m)	Ws (m/s)	Hs_ts (m)	Tm_ww (s)	Tm_ts (s)	AND Exceedance Probability	AND Return Period (years)
10	0.100	3.22	0.93	17.45	2.33	7.337	11.86	1.00E-06	1.00E06
50	0.020	3.54	1.19	20.03	2.70	8.0992	13.35	6.40E-11	1.563E10
100	0.010	3.67	1.32	21.45	2.85	8.4115	13.90	1.00E-12	1.00E12
500	0.002	3.84	1.69	25.68	3.20	9.1016	14.98	6.40E-17	1.563E16
1000	0.001	3.97	1.89	28.00	3.34	9.3844	15.37	1.00E-18	1.00E18

Table 5.3: Table presenting the (univariate) design values for the AND-risk scenario when all variables are considered to be independent and extreme. The five sets of 6 univariate design values represent extreme events and the corresponding T_{and} of these extreme events is presented in the last column.

Return Period (years)	Univariate Exceedance Probability	Hs_ww (m)	WL (m)	Ws (m/s)	Hs_ts (m)	Tm_ww (s)	Tm_ts (s)	OR Exceedance Probability	OR Return Period (years)
10	0.100	3.22	0.93	17.45	2.33	7.337	11.86	4.69E-01	2.13
50	0.020	3.54	1.19	20.03	2.70	8.0992	13.35	1.14E-01	8.76
100	0.010	3.67	1.32	21.45	2.85	8.4115	13.90	5.85E-02	17.1
500	0.002	3.84	1.69	25.68	3.20	9.1016	14.98	1.19E-02	83.76
1000	0.001	3.97	1.89	28.00	3.34	9.3844	15.37	5.99E-03	167.08

Table 5.4: Table presenting the (univariate) design values for the OR-risk scenario when all variables are considered to be independent and extreme. The five sets of 6 univariate design values represent extreme events and the corresponding T_{or} of these extreme events is presented in the last column.

Calculating mutivariate probabilities of exceedance is "easy" when the variables are independent. The AND probability of exceedance is the multiplication of the univariate probabilities of exceedance. The OR probability of exceedance is the sum of the univariate probabilities of exceedance minus their respective intersections, which is equal to 1 minus the AND multivariate cumulative probability. The two scenarios should not be compared because they represent two different system dynamics.

5.7.2. Vine-based methodology

To derive the design values using the vine-based approach, we need the results obtained in steps 1 and 3 of the current chapter. From step 1, we need the marginal distributions of the extreme variable and its concomitants. These are presented in the third column of table 5.1. With the univariate probabilities of exceedance and the marginal distributions, the univariate design values are calculated with equation 3.1.

Calculating multivariate probabilities of exceedance is not an easy task when the variables are dependent. Section 3.4 explains how multivariate probabilities of exceedance are calculated when using the vine-based methodology. Mainly, one needs to sample (as many times as is computationally feasible) from the regular vine chosen in step 3. Then, it becomes a simple counting task. For the AND exceedance probability, for example, one should count how many times all the variables exceed together (at the same time) their individual thresholds, which are equal to the univariate exceedance probabilities. And the resulting number should be divided by the total number of observations sampled from the vine. A similar logic is applied for the OR exceedance probability.

The five sets of 6 (univariate) design values in tables 5.5 and 5.6 represent extreme events. The reader should note only the design values from Hs_{ww}, the dominant variable, are associated with actual return periods (or at least how these are understood within the engineering community). The remaining variables are Hs_{ww}'s concomitants and hence might not comply with the properties of extreme observations.

For a chosen AND-risk scenario, the multivariate return period, T_{and} (see definition in equation 2.17), and the corresponding AND-probability of exceedance of the aforementioned 5 extreme events are presented in table 5.5. For a chosen OR-risk scenario, the multivariate return period, T_{or} (see definition in equation 2.18), and the corresponding OR-probability of exceedance of the aforementioned 5 extreme events are presented in table 5.6.

Univariate Exceedance Probability	**Hs_ww (m)**	**WL (m)**	**Ws (m/s)**	**Hs_ts (m)**	**Tm_ww (s)**	**Tm_ts (s)**	**AND Exceedance Probability**	**AND Return Period (years)**
0.100	3.22	0.90	14.01	1.49	5.59	8.12	1.39E-04	7194
0.020	3.54	1.16	16.36	1.92	6.27	9.67	3.00E-07	3333333
0.010	3.67	1.29	17.34	2.10	6.55	10.34	<1.00E-08	>1.00E08
0.002	3.84	1.64	19.58	2.52	7.19	11.96	<1.00E-08	>1.00E08
0.001	3.97	1.82	20.54	2.70	7.47	12.68	<1.00E-08	>1.00E08

Table 5.5: Table presenting the (univariate) design values when all variables are considered to be dependent, and only Hs_ww extreme. The remaining variables are its concomitants. The five sets of 6 univariate design values represent extreme events and the corresponding T_{and} of these extreme events is presented in the last column.

Univariate Exceedance Probability	**Hs_ww (m)**	**WL (m)**	**Ws (m/s)**	**Hs_ts (m)**	**Tm_ww (s)**	**Tm_ts (s)**	**OR Exceedance Probability**	**OR Return Period (years)**
0.100	3.22	0.90	14.01	1.49	5.59	8.12	3.90E-01	2.56
0.020	3.54	1.16	16.36	1.92	6.27	9.67	9.99E-02	10
0.010	3.67	1.29	17.34	2.10	6.55	10.34	5.22E-02	19.16
0.002	3.84	1.64	19.58	2.52	7.19	11.96	1.10E-02	90.91
0.001	3.97	1.82	20.54	2.70	7.47	12.68	5.50E-03	181.82

Table 5.6: Table presenting the (univariate) design values when all variables are considered to be dependent, and only Hs_ww extreme. The remaining variables are its concomitants. The five sets of 6 univariate design values represent extreme events and the corresponding T_{or} of these extreme events is presented in the last column.

In the two last columns of table 5.5, the reader may notice that 10 million samples are not enough to quantify the exceedance probabilities (and the AND return period) of such large design values. One could roughly estimate the amount of samples needed with the AND-return period in table 5.5, which gives an indication on the frequency of occurrence of these extreme events resulting from the traditional approach. For example, to calculate AND exceedance probabilities when all 6 variables exceed its associated 0.01 percentile, one would need around 10^{12} samples. Sampling such amount of data from a high dimensional regular vine is computationally demanding.

In appendix E, we discuss the magnitude of the error in the probabilities of exceedance when calculating these with the sampling method instead of integrating numerically the regular vine's density.

5.8. Comparison between the outcome of the traditional and the vine-based approaches

5.8.1. Design values

In this section, we compare the outcome of the traditional and the vine-based approaches to highlight some advantages of the last. Firstly, we focus on the already presented design values.

The design values derived with the traditional approach are presented in tables 5.3 and 5.4. The design values derived with the vine-based approach are presented in tables 5.5 and 5.6. The reader may notice the design values resulting from the traditional approach are larger than the ones resulting from the vine-based approach. This is a direct consequence of the difference in the sampling of extreme observations between both approaches.

On another note, the AND exceedance probabilities of the extremes events (i.e. combination of design values) resulting from the traditional approach are smaller than the AND exceedance probabilities resulting from the vine-based approach. For equal AND-Return periods, the set of design values resulting from the traditional approach would be more conservative than the set of design values resulting from the vine-based approach. This might lead to more conservative and possibly, more expensive traditionally-based designs.

In contrast, the OR exceedance probabilities are smaller for the sets of design values obtained with the vine-based approach. This is a direct consequence of taking into account the dependence (or intersection) between the design variables.

5.8.2. Breakwater's crest level application

To characterize all the loads on a breakwater, the engineer focus on several design criteria: maximum overtopping discharge, toe and core stability and maximum wave transmission (among others). These criteria mainly depend on the following environmental variables: wave height, wave period, water level and currents. Thus, it seems essential to study the interdependence between all the design variables despite the fact that not all of them appear together in the same design formulas (or criteria). To illustrate the advantages of accounting for the aforementioned interdependence, we introduce in this section a simplified application linked to the coastal engineering case study: the design of the breakwater's crest level. The remaining design criteria can be addressed in further studies.

The typical design formula to calculate the crest level is a function of three of the analyzed design variables: Hs_{ww}, Hs_{ts} and WL. The crest level is calculated in here in two ways: (1) assuming the aforementioned variables are independent (traditional approach) and (2) assuming that these are dependent (vine-based approach). Despite the fact that the application in here only requires the analysis of 3 random variables, we randomly generate combinations of these variables from the vine-copula in step 3 to calculate the crest level for the dependent case.

Let's assume the main design requirement for the crest level is a maximum overtopping discharge of 50 L/s per meter for the Ultimate Limit State (ULS). Wave-overtopping takes place when waves meet a structure lower than the approximate wave height. During over-topping, water can pass over the structure and the volume of water that passes is usually called overtopping discharge (Q). In figure 5.14, wave-overtopping is depicted.

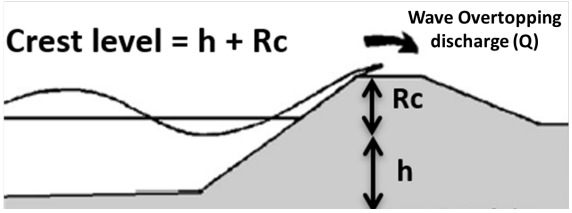


Figure 5.14: Visualization of wave-overtopping and a breakwater's crest level

The crest level is the result of the sum of the the still water level (h) and the free-board (R_c) (see figure 5.14). The still water level is considered a random variable itself and the same variable as WL (which was part of the main analysis performed in this chapter). The free board is the height of the breakwater above the still water

level and hence, the part that is not underwater all the time. To calculate R_c, the EurOtop (European Overtopping manual [59]) recommends the use of the following formula:

$$\frac{q}{\sqrt{gH_{m0}^3}} = 0.1035 \exp\left[\left(-1.35\frac{R_c}{H_{m0}\gamma_f\gamma_\beta}\right)^{1.3}\right] \tag{5.2}$$

where:
$H_{m0} = (Hs_{ww} + Hs_{ts})^{0.5}$ is the total significant wave height and is a random variable, q is the wave-overtopping discharge and is equal to 50 L/s/m, g is the gravitational acceleration and is equal to 9.81, γ_f is a roughness factor which is equal to 0.4 and γ_β is a berm factor which is equal to 1 as we assume the breakwater has no berm.

Manipulating equation 5.2, we obtain an equation for Rc that depends on the significant wave height of wind waves (Hs_{ww}) and swell waves, (Hs_{ts}).

In section 5.7.2, we randomly generated 10 million samples of the design variables from the vine-copula (which included observations for Hs_{ww}, Hs_{ts} and WL). We used these samples to calculate the AND and OR probabilities of exceedance. In here, we calculate the corresponding 10 million outcomes of the *Crest level* with the 10 million modeled observations of Hs_{ww}, Hs_{ts} and WL. Then, we compute the empirical cdf of the breakwater's crest height (i.e. crest level). This procedure is similar to a Monte-Carlo analysis.

To calculate the empirical cdf wit the traditional approach, the 10 million random samples are drawn individually from the univariate distributions of Hs_{ww}, Hs_{ts} and WL resulting from the EVA in step 1 (see table 5.1). With these samples, we calculate the corresponding 10 million outcomes of the *Crest level* and next, we compute the corresponding empirical cdf of the *Crest level* for the independence case.

The aforementioned two cumulative distributions of the *Crest level* are plotted in figure 5.15. This figure presents return periods instead of probabilities of exceedance. The relation between return period and probability of exceedance is presented in equation 2.13. When comparing the two curves in figure 5.15, the traditional approach seems to over-predict the crest height. For the same values of the crest height the exceedance probabilities are smaller, and hence the return periods are larger for the independence case than for the dependence case. These can also be seen with the results presented in table 5.7. Table 5.7 presents a comparison between exceedance probabilities and return periods for crest heights of 9, 10 and 11 meters for the dependence and independence cases.

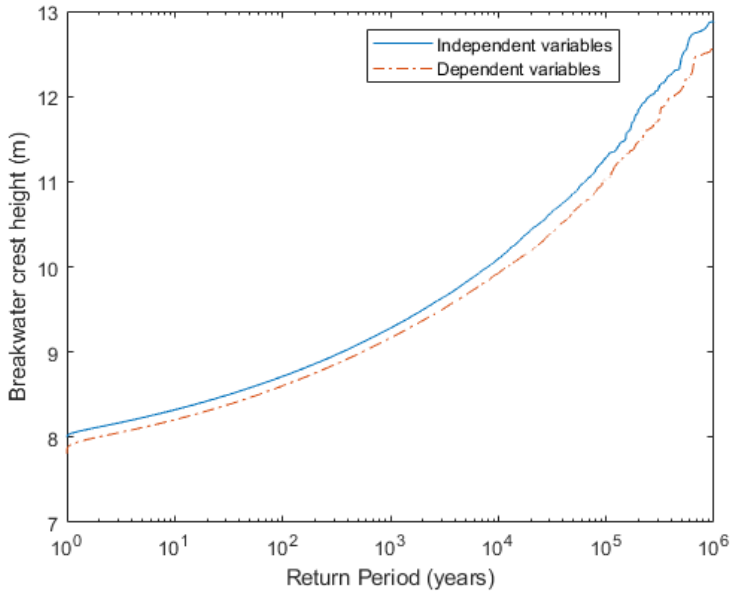


Figure 5.15: Empirical cumulative distribution functions for the breakwater crest height with return periods for (1) assuming the aforementioned variables are independent (traditional approach) and (2) assuming that these variables are dependent (vine-based approach).

Breakwater's Crest Height	Exceedance probability		Return period (years)	
(m)	INDEPENDENT	DEPENDENT	INDEPENDENT	DEPENDENT
9	2.80E-03	1.83E-03	357	548
10	1.24E-04	8.32E-05	8039	12019
11	1.57E-05	1.08E-05	63694	92593

Table 5.7: Breakwater's crest height values and their respective exceedance probabilities and return periods for the dependent and independent case.

Let's assume that the practitioner is designing for a risk level of 0.0001, equivalent to a return period of 10000 years. Considering the results for the dependence case in table 5.7, the practitioner would probably design the crest height smaller than 10 meters. However, if the practitioner would only look at the results achieved by the traditional approach (independent case), he/she would probably pick a value larger than 10 meters. The difference in height might translate to thousand of euros if the breakwater is long enough.

6

Discussion

6.1. Discussion on the vine-based methodology

One objective of the thesis is to define design values in a multivariate case. One problem that arises when extending the extreme value theory to the multivariate context is that one must specify what constitutes an *extreme event*. In the univariate case, this is defined as the most extreme value. Next, the concept of return period is used to define design values. Extending this procedure to find design values in the multivariate case is not straightforward. Two main issues arise:

1. It is not immediately clear how to find the most extreme set of joint observations from a multivariate time series
2. It is quite challenging to fix a desired level of risk in the multivariate case and back track the sets of design variables associated with that level of risk

Regarding the first issue, the fundamental problem is that there is no natural way to order a set of vectors, as there is with a set of real-valued numbers. In this thesis we suggest to define an extreme event following the approach in [75]. The extreme observations are sampled for the most dominant variable and the corresponding concomitant values (observed together with the dominant variable) are selected for the remaining variables. In this way, the dependence between the variables is kept in what constitutes the extreme event. A different approach would be to sample the extreme observations independently for each variable. However, sampling all the extremes independently will probably result in a meteorological event never experienced. This issue is treated in more detail in section 6.2.

The essence of *return period* in current industry-related practices is univariate. In literature, there is no consensus on how to extend this concept to the multivariate framework. The one-to-one relationship between return period – return level in the univariate case is not valid in higher dimensions. The second issue listed above makes reference to this. For this reason, we opted to predefine the design values based on the univariate case. This is a legitimate but simplified way to approach the problem.

The occurrence of environmental extreme event scenarios in a multivariate framework has been addressed trying to determine the probability corresponding to a failure region, considering failure modes with elements in series and in parallel both under independent and dependent circumstances. This probability is associated to what is defined as *multivariate return period*. By adopting a risk scenario, we linked the set of design values to a probability of exceedance and thus, a multivariate return period. Figure 6.1 illustrates the traditional procedure to find design values in the univariate case and the procedure suggested in this thesis to find design values in the multivariate case.

From an engineer's perspective, one could argue what the advantages of imposing the univariate design values are. The main advantage is the benefit of having a one-to-one relationship between the (univariate) return period and design value. Making use of this relationship makes the approach simpler and closer to current industry practices. The added value of this approach when compared to the traditional univariate approach is the computation of the multivariate return period associated to the set of univariate design values. Assuming the

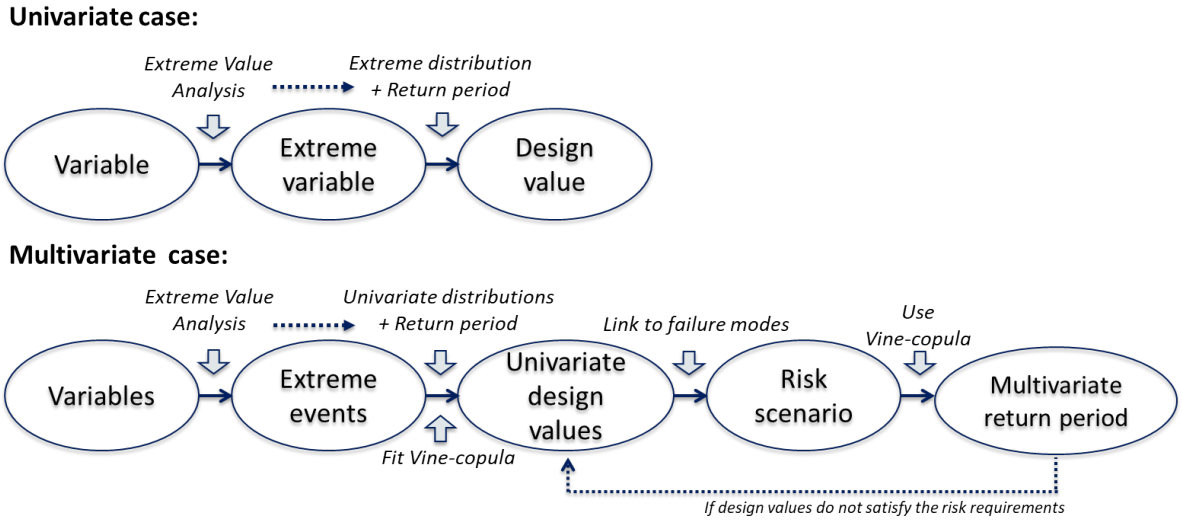


Figure 6.1: Visualization of the proposed extension of design load definition from the univariate to the multivariate case. Source: Author

selected risk scenario represents the failure (or damage) mode of the structure, the information on the overall risk of failure is then given by the multivariate return period. The reader should note that the multivariate return period and probability of exceedance are interchangeable. With this method, the practitioner has more information on the behavior of the system. Thus, it becomes a matter of choice whether it is possible (or not) to modify the design based on experience and knowledge on the structure's behavior.

6.2. Discussion on the case study

In this thesis, a vine-based methodology for infrastructure design load definition is presented. The methodology is general and can be implemented in any design problem involving multiple design variables. To show the potential of this methodology, a hypothetical engineering application was presented: the design of two breakwaters at the entrance of Galveston Bay, Texas. It is important to point out that the application is meant to be illustrative and it relies on arbitrary design assumptions.

The wind and wave data was downloaded from ERA 5 (a global model) at a point 42 Km away from the breakwaters' location. It was assumed that despite the distance the environmental data is representative of the environmental conditions at the breakwater location. One can argue that this is too big of an assumption to make. To solve this issue, the practitioner can transform the resulting design values to the ones that are representative of the extreme environmental conditions at the breakwater location. For example, this can be done with wave-propagation models such as SWAN for wave variables.

Step 0 (data processing) is considered the input to the methodology and has not been treated extensively in this thesis. Mainly because the level of data processing differs per application. Within offshore and coastal engineering, the basic idea is to decompose the time series of the wave climate in sea states, so one can consider each state independently. This analysis is called *metocean study*. In our application, the wave components are already decomposed in sea states: swell and wind generated waves (wind waves). We then assumed the data is already processed and satisfies the requirement for the specific design situation. Nevertheless, an important part of the aforementioned metocean study is to determine the principal and mean wind and wave directions. This has not been treated in this thesis for the sake of simplicity. Nevertheless, the practitioner can select the wind and waves from the principal direction of interest and perform the analysis only with the resulting data. Another option is to split the wind and wave data in directions and perform a frequency analysis per direction. The resulting sets of wind and wave data are called *partitions*, in engineering terms. These 'partitions' should be independent and hence, one should treat them in separate analyses.

The data location has a water depth of around 20 meters and was considered to be transitional (shallow) waters. This was necessary to support the performance of the *Extreme Value Anaysis* (*EVA*) at that location. To correctly infer extreme waves characteristics, the EVA should be performed using a dataset containing observed (or simulated) waves before they reach their breaking point. In this study, it has been assumed that the wave

dataset contains unbroken waves, without performing any additional check. However, it is advised to verify this condition. To avoid this issue, an alternative option could be using wave information from an offshore location.

The multivariate frequency analysis in chapter 5 is performed for extreme events resulting from the combination of extreme wind generating wave heights and its maximum concomitants: wind speed, still water level, swell wave heights, wind waves periods and swell wave periods. Consequently, the wave height of wind waves was assumed the most relevant to the design. The choice of dominant variable is up to the practitioner. One could argue that by choosing one dominant variable could lead to a misrepresentation of the concomitant variables. However, by sampling all the extremes independently will probably result in a meteorological event never experienced. Also, the resulting sample of extreme observations would not be strictly representative of *joint observations*, because the extremes would be sampled from different moments in time. Unless, the practitioner is interested in what happens yearly. This scale might be relevant from a geological point of view where 10's of thousands of years are of interest. To illustrate this issue, we re-sampled the extremes of the 6 analyzed variables independently with the POT technique and we calculated their correlation coefficients. These are illustrated in figure 6.2. In this figure, one can see that some of the correlation coefficients might not make sense from a physical perspective. For instance, the wind speed is negatively correlated with wind wave's height and period. Another example is the considerably large and negative correlation between the water level and the wind wave's period. In addition, the magnitudes of all these correlations are considerably low (close to 0) when compared to the correlation coefficients representing extreme events in figure 6.3. These results support the previous statement: the individually sampled observations are independent.

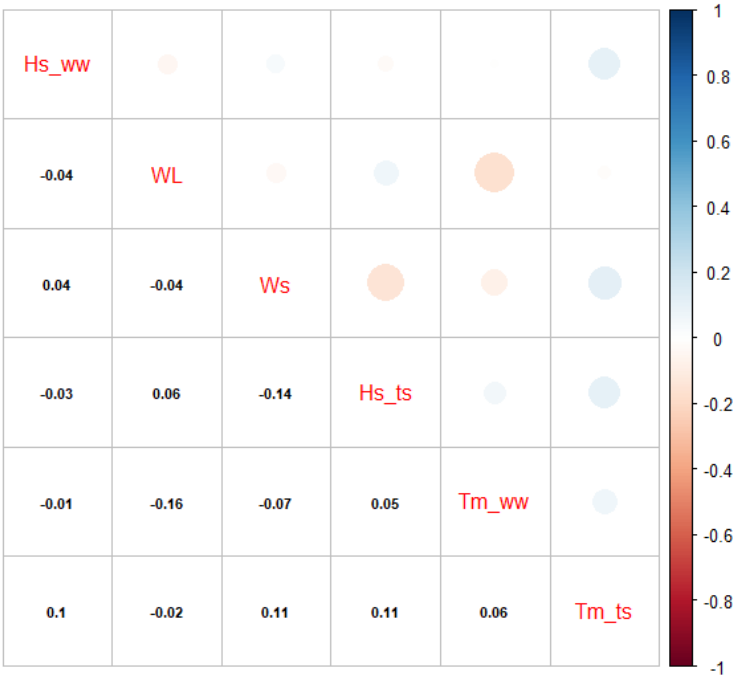


Figure 6.2: Correlation matrix for all variables being dominant, and hence extreme

Figure 6.3: Correlation matrix for the dominant and concomitant variables

In step 3, the best regular vine is selected from all the existing regular vines according to a certain criterion. It is logical to believe that the best regular vine might differ depending on the chosen criterion. In this thesis we propose two selection criteria: (1) lowest AIC and (2) largest correlations represented in the trees structures of the regular vine. In theory, the regular vine selected with (1) contains more information from the data than the rest of the fitted models. However, the results in figure 4.8 suggest that the correlations between variables are better represented by the best vine according to (2). The practitioner should choose the selection criteria with which he/she feels comfortable with.

In step 4, the joint probabilities of exceedance are calculated using a sampling procedure instead of integrating the density function of the regular vine. This saves computational time. Nevertheless, the main drawback might be that probabilities of simultaneous exceedance of several large values are not well represented. This does not apply to the OR exceedance probabilities. 10^{08} samples were not enough to quantify the AND exceedance probabilities of design values associated to percentiles larger than 0.99. We could only estimate that this exceedance probability is lower than 10^{-08}.

6.3. Discussion on the exploratory work on goodness of fit for vine-copula

In this thesis, the possibility to define goodness of fit test for vine-copula based on the concept of tree-equivalent classes (TEC, see section 2.3.2) was explored. The reader should note that the objective was to provide some indications on how the concept of TEC could be used to define goodness of fit test for vine-copula in further research. Consequently, we focused on model selection strategies based on graphical and statistical properties of the vines.

Following the work done by Morales-Nápoles in [49], we explored the validity of the following hypothesis to define a model selection strategy in the form of an algorithm: *The best fit for a 'n+1' variables regular vine would be an extension of the best fit of 'n' variables regular vine*. The algorithm aims to facilitate the implementation of vines in higher dimensions (vines with more than 6 nodes).

The concept of TEC is used to extend the vine's tree structure in n nodes to $n+1$ nodes: the best regular vine in $n+1$ belongs to a TEC that is an extension of the TEC to which the best regular vine in n nodes belongs to. The selection reference (for the best fit) is based on the AIC which requires the trees to be labeled and copulas assigned to each edge in the regular vine. There is no way that one could argue that the AIC is independent from the order of the nodes. A better approach would be to select the best vine according to a more general criteria (or goodness of fit) that involves unlabeled tree structures.

When adding a dimension, one is actually adding another tree to the regular vine. Thus, extending the TEC did not seem enough to extend the regular vine to higher dimensions. Subsequently, we studied whether the first *labeled* tree (T1) of the potential best regular vine in $n+1$ nodes could be an extension of T1 of the best regular vine in n nodes. We saw that the regular vine achieving the lowest AIC was not an extension of the labeled trees of the best vines in 4 and 5 nodes. The way in which we ordered (i.e. column-wise) the design variables in the data set might be a plausible reason to explain why these results do not support the presented hypothesis. The ordering procedure in section 4.2.2 does not take into account the goodness of fit of all the pairs of variables to the bivariate copulas. This was investigated selecting the best fit as the the regular vine with the strongest correlations in its trees. In this case, the hypothesis presented at the beginning of this section seems to be valid.

The algorithm was used to select the best regular vine in 6 nodes for our data set. To validate the results of the algorithm, we determined which of the three regular vines (i.e. (1) Lowest AIC, (2) Strongest correlation coefficients in its trees and (3) Algorithm 1) represented best the original correlation coefficients. The results suggested that the regular vine with the lowest AIC represented the original correlations worst. The regular vine selected with the novel algorithm achieved a lower AIC than the regular vine with the strongest correlations in its trees. However, the last vine represented the original correlations best. In theory, the model achieving the lowest AIC is the model that "loses" the least information from the data. The results of this analysis do not seem to fully support the aforementioned statement.

7

Conclusions and further research

7.1. Conclusions on the vine-based methodology

The main objective of this thesis was to develop a vine-based methodology for infrastructure design load definition. The methodology aims to be general: it can be applied to any infrastructure design problem involving multiple design variables. With this premise in mind, the main effort was concentrated in formulating a series of steps to integrate several stages of the design: from the processing of raw data up to the choice of design values for any specific design purpose. The statistical dependence between random variables was taken into account using a regular vine-copula, where the multivariate dependence structure is modeled using bivariate copulas as building blocks.

The methodology starts with *step 0*: the collection and processing of data. The data comes in the form of a multivariate time series, which needs to be representative of the system that one aims to model. The variables describing this system are part of the physical processes that trigger damage or failure of the structure.

When the data is ready, one should proceed to *step 1*: Extreme Value Analysis. Normally, infrastructure is designed to withstand extreme conditions. To find these, the practitioner should perform a Peak Over Threshold (POT) on the dominant variable that is assumed most relevant to the design according to some criteria. The variables that are observed together with the dominant variable during an extreme event, the *concomitants*, are sampled using block maxima technique. These sampling procedures lead to the so-called *extreme sample*. The next task is to fit a distribution to the dominant and concomitant variables in the extreme sample.

The next step is *step 2*: Bivariate dependence modelling. The goal of this step is to gain insight into the physical behavior of the system. By performing this step one can ensure that the statistical results are in accordance with what is expected from a physical point of view. The bivariate dependence is studied in two manners: (1) by analyzing correlation coefficients, and (2) by analyzing the dependency structure with bivariate copulas.

If the results of step 2 are satisfactory, the practitioner can proceed with *step 3*: Multivariate dependence modelling. The goal of this step is to select a regular vine-copula to model the extreme sample derived from step 1. In this methodology, all existing regular vine-copulas are fitted to the data. The best regular vine is selected according to a specified criteria, and then, the model is validated by stochastic simulation.

The last step in this methodology is *step 4*: Derivation of multivariate design values. This is meant to be the link between the pure statistical analysis and the engineering design process, and it is one of the main contributions of this thesis. The design values are imposed based on their univariate return periods and their corresponding multivariate return period or associated probability of exceedance are calculated with the vine-copula selected in step 3. To do so, one needs to select a risk scenario that represents the system dynamics.

From the work carried out at this point, the research questions can be addressed:

- How can we use vine-copula models in the design of infrastructure?

Vine-copulas are implemented in the vine-based methodology for infrastructure design (vine-based methodology, hereinafter) to model the statistical dependence between design variables. More specifically, they are used to calculate the multivariate exceedance probabilities of certain extreme events, which are defined as a combination of (univariate) design values. These multivariate probabilities of exceedance are calculated by imposing a risk scenario and are interchangeable with their associated multivariate return period. Assuming the selected risk scenario represents the failure (or damage) mode of the structure, the information on the overall risk of failure is then given by the multivariate return period.

The vine-based methodology was tested for a case study in coastal engineering and the results suggested that it is suitable to perform multivariate frequency analysis of extreme events. In the application, the selected regular vine-copula seems to represent the original correlation coefficients varying in the range $\pm 1\%$. The tails and overall mass concentration of the original extreme data set seem to be well represented by the selected model.

In the vine-based methodology, the design values are imposed based on their univariate return periods, and their probability of exceedance (based on a predefined risk scenario) is computed with a sampling procedure instead of solving a 6-dimensional integral to determine cumulative distribution function of the vine-copula. This choice reduces the computational effort considerably. Nevertheless, 10 million samples were not enough to quantify the AND exceedance probabilities of design values associated to percentiles larger than 0.99. It is estimated that around 10^{12} samples are needed to achieve at least one observation where all variables exceed simultaneously values associated to percentiles larger than 0.99. Sampling such amount of data from a high dimensional regular vine is computationally demanding.

- What are the advantages and disadvantages of using the vine-based methodology to derive multivariate design values when compared to the traditional approach where the ariables are considered independent?

Performing a multivariate frequency analysis with the vine-based methodology is theoretically and practically more challenging than performing several univariate frequency analysis under the assumption of independence between design variables.

In our application, the design values derived with the traditional approach are larger than the ones derived with the vine-based approach. For equal exceedance probabilities, the design values of the traditional are more conservative than the ones of the vine-based approach. Consequently, this might translate to more conservative and possibly, more expensive designs when designing with the traditional approach. To illustrate this fact, the results of both approaches are applied in a simplified example on probabilistic design of a breakwater's crest level. The results of the aforementioned application suggest that for equal probabilities of exceedance (or equal return periods), the crest level derived with the traditional approach is more conservative that the one derived with the vine-based approach. In quantitative terms, for the same crest level the return period computed with the traditional approach is at least 1.5 times larger than the one computed with the vine-based approach. This would lead to larger breakwater crest's height for the traditional design than for the vine-based design.

These applications showed that accounting for the interdependence between design variable provides a more comprehensive description of the physical system acting on the infrastructure. This information enables the practitioner to make educated choices when performing the statistical analysis. Thus, it becomes a matter of choice whether it is possible (or not) to modify the design based on experience and knowledge on the structure's behavior. However, the vine-based method is computationally demanding. Hence, the applicability of this methodology should be evaluated on a case by case basis.

7.2. Conclusions on the exploratory work on goodness of fit for vine-copula

The research done in this thesis is exploratory and aims to establish the foundations for further research on goodness of fit for vine-copulas based on the concept of Tree-equivalent classes (TEC). From the work done in this thesis the secondary research question can be addressed:

- Can we identify goodness of fit for vine-copula based on tree-equivalent classes?

Despite the answer not being a resounding *yes*, some advances have been made that bring the community a step closer to answering this question. For instance, it is the first time that all regular vines up to 6 nodes are fitted to a data set and classified by TEC.

In this thesis, the focus was set on model selection strategies based on graphical and statistical properties of the vines. The main motivation to investigate model selection strategies for vines is the considerably large computational time needed to fit all regular vines in more than 6 nodes to the data.

A new algorithm was developed to facilitate the implementation of vines in higher dimensions (vines with more than 6 nodes). This algorithm significantly reduces the computational effort to select a regular vine by allowing the user to test only a subgroup of vines in n-nodes constructed on specific characteristics of the vines in $(n-1)$-nodes. The concept of TEC was used to extend the vine's tree structure in $(n-1)$-nodes to n-nodes.

7.3. Further research

Two different branches for further research are proposed: improving the vine-based methodology and continuing the work on goodness of fit for vine-copula based on tree-equivalent classes.

Regarding the first branch, the methodology presented in this thesis could be improved addressing the following topics:

- Tackling the issue of defining design values based on the multivariate return period. Solving this issue would imply finding a way to solve the equations of a hyperplane to backtrack the sets of design values that are associated to a certain risk level (i.e. multivariate return period). The main issue is that there is no unique solution to this problem. In fact, the space of solutions is infinite.
- Making the methodology more practical so it can be used in the industry. This could be achieved by developing a tool box. This tool box would automatically perform the analysis and produce graphs and/or tables that are easier to interpret than the ones presented in this thesis. One could make use of the coding scripts presented in this report to develop such tool. Another recommendation to make it more practical would be to explore ways to plot multivariate densities and cumulative distributions in 2D. An idea to do so is presented in Appendix F. In short, the entire product would be tied to creativity and originality.
- Quantifying the impact of the vine-based methodology in the design of coastal structures (or other type of infrastructure). In this thesis, the methodology is tested and validated for a case study in coastal engineering. Hence, it is shown how it can be applied in a typical engineering project. Nevertheless, the only step towards quantification was made when the breakwater's crest level was calculated with independent and dependent variables. To study the impact in a broader manner, it is recommended that the vine-based methodology is applied to a range of case studies and for a range of failure mechanisms (or limit state functions). The impact could be quantified in terms of volume of material saved (rock, in the case of a breakwater). The different outcomes could be categorized per design criteria and presented in the form of an *Atlas*. The Atlas would be a tool for engineers to decide whether the increase in quality and the added opportunities for design optimization provided by the vine-based methodology are worth the extra computational time.

Regarding the second branch, the work on goodness of fit for vine-copula based on tree-equivalent classes could be improved addressing the following topics:

- Fitting all regular vines up to the maximum number of nodes possible and categorize them by TEC. This analysis could be done for several data sets. It is recommended that a different and more general goodness of fit measure than the AIC is used to assess the general and individual performance of all TEC.
- Testing the validity of the new algorithm in analysis with different data sets. Compare the performance of the regular vine selected by the algorithm with the regular vine with the lowest AIC and the regular vine with the strongest correlations in its tree-structures. It is recommended that the performance of these is assessed mathematically using the regular vine's density, to aim to provide a more formal conclusion than the one provided in this thesis.

A

Some multivariate models build with copulas

A.1. Trivariate setting based on conditional laws

Chakak and Koehler presented in [10] one of the very first and simplest approaches to develop a trivariate copula by using conditional distributions. The latter technique is based on conditional laws. More specifically, a conditional distribution is used in [10] to describe the probability of observing a variate given that the associate variate is already known. So $F_{X|Y}$ would be the probability of observing X given a known value of Y. Formally, the conditional distribution function and calculation of conditional probabilities of a bivariate copula $C(u, v)$ is defined as [54] [19]:

$$P(U \leq u|V = v) = \frac{\partial}{\partial v} C(u, v) \tag{A.1}$$

$$P(U \leq u|V \leq v) = \frac{C(u, v)}{v} \tag{A.2}$$

$$P(U \geq u|V \geq v) = 1 - \frac{u - C(u, v)}{1 - v} \tag{A.3}$$

and analogous expressions hold for the conditional laws of V given U. Consequently, computing conditional probabilities simply reduces to partial derivatives of suitable copulas. Nevertheless, the resulting trivariate-copula is not uniquely determined and is dependent on the order of which the copulas are combined [15].

A.2. Conditional mixtures

A well established method that uses a similar concept as in [10] to define multivariate models with copulas is the conditional mixtures. This approach is discussed and applied in [66], [20], [35] and [15]. An example of a 3-dimensional family is given in [20] by:

$$F_{XYZ}(x, y, z) = \int_{-\infty}^{y} C_{XZ}(F_{X|Y}(x|t), F_{Z|Y}(z|t))(dt) \tag{A.4}$$

The arguments of the integrand (namely, $F_{X|Y}$ and $F_{Z|Y}$) are conditional distributions and can be written as in equation A.1, in terms of copulas.

A.3. Hierarchical Archimedean copulas

Hierarchical Archimedean copulas are another popular method to construct multivariate structures with copulas. They joint two or more bivariates or higher order copulas by another Archimedean copula. However, the major limitation of hierarchical Archimedean copulas is that not all combinations of joint distribution are modelled uniquely [15]. Archimedean copulas are formally defined in [41] as:

$$C(x_1, ..., x_n) = \phi^{-1}\left(\sum_{i=1}^{d} \phi(F_i(x_i))\right) \tag{A.5}$$

if $\sum_{i=1}^{d} \phi(F_i(x_i))) \leq \phi(0)$ otherwise, the copula is equal to zero. $\phi(x)$ is the generator function and $\phi(x)^{-1}$ the inverse function.

Three classical generator functions are the ones corresponding to the Clayton, Gumbel and Frank copula. The conditional probability property that Archimedean copulas share makes them practical for higher dimensional variables. The reader is referred to [67] for further information on hierarchical Archimedean copulas. These type of multivariate dependence structures are commonly applied in ocean and coastal engineering applications. Some examples are discussed in [41], [15], [72] and more recently, in [44]. Within the trivariate setting, elliptical copulas, in particular the Student t-copula, have proven themselves more than capable of capturing the dependence structure between a set of random variables [58] [60].

B

Data sources

In this Appendix, the sources of data are described in tables B.1 and B.2, for México and U.S respectively.

Location 1: México, Tabasco	**Data**	**Source**	**Description**
Bay of Campeche	Wave climate	Offshore Buoy Depth: 3624 m NOAA - National Data Buoy Center 22°7'25" N 93°56'26" W	Period: 2005-2018 Temporal resolution: hourly Type: wave spectral data (per frequency)
Various offshore locations at the Gulf of México	Wave climate	ERA5: Numerical model (global) Spatial reolution: 0.5°x0.5°	Period: 1979-2019 Temporal resolution: hourly Type: wave spectral variables
Various offshore locations at the Gulf of México	Wind field	ERA5: Numerical model (global) Spatial reolution: 0.5°x0.5°	Period: 1979-2019 Temporal resolution: hourly Type: 10m-wind speed and direction
Various inland locations in Tabasco	Run-off total precipitation	ERA5: Numerical model (global) Spatial resolution: 0.25°x0.25°	Period: 1979-2019 Temporal resolution: hourly
Various locations at the Gulf of México	Mean sea level pressure	ERA5: Numerical model (global) Spatial resolution: 0.25°x0.25°	Period: 1979-2019 Temporal resolution: hourly
Various inland locations in Tabasco	River Discharge	CSIRO: global hydrological model WCI portal	Period: 1979-2012 Temporal resolution: daily Type: point river discharge

Table B.1: Data available at location 1 in Tabasco, México.

Location 2: Texas, the US	**Data**	**Source**	**Description**
Station 42035 (LLNR 1200) GALVESTON,TX	Wave climate	Nearshore Buoy Depth: 16.2 m NOAA - National Data Buoy Center 29°13'54" N 94°24'46" W	Period: 1996-2018 Temporal resolution: hourly Type: wave spectral data (per frequency)
Station 42002 (LLNR 1470) WEST GULF East from Brownsville	Wave climate	Offshore Buoy Depth: 3125.1m NOAA - National Data Buoy Center 29°13'54" N 94°24'46" W	Period: 1996-2018 Temporal resolution: hourly Type: wave spectral data (per frequency)
Station 42019 (LLNR 1285) FREEPORT, TX	Wave climate	Transitional waters Buoy Depth: 82.2 m NOAA - National Data Buoy Center 27°54'22" N 95°21'0" W	Period: 1996-2018 Temporal resolution: hourly Type: wave spectral data (per frequency)
Various offshore locations at the Gulf of México	Wave climate	ERA5: Numerical model (global) Spatial reolution: 0.5°x0.5°	Period: 1979-2019 Temporal resolution: hourly Type: wave spectral variables
Various offshore locations at the Gulf of México	Wind field	ERA5: Numerical model (global) Spatial reolution: 0.5°x0.5°	Period: 1979-2019 Temporal resolution: hourly Type: 10m-wind speed and direction
Various inland locations in Houston	Run-off total precipitation	ERA5: Numerical model (global) Spatial resolution: 0.25°x0.25°	Period: 1979-2019 Temporal resolution: hourly
Various locations at the Gulf of México	Mean sea level pressure	ERA5: Numerical model (global) Spatial resolution: 0.25°x0.25°	Period: 1979-2019 Temporal resolution: hourly
Various inland locations in Houston	River Discharge	CSIRO: global hydrological model WCI portal	Period: 1979-2012 Temporal resolution: daily Type: point river discharge
4 locations within the Buffalo Bayou catchment	River Discharge	4 Gauging station (to be described)	Period: 1980-2016 Temporal resolution: daily Type: point river discharge
Galveston Pier 21 in Galveston Bay	Surge level	Tide station NOAA - National Data Buoy Center 29°18'38" N 94°47'30" W	Period: 1904-2018 Temporal resolution: daily Type: point surge

Table B.2: Data available at the preferable location in Texas, the US.

C

R code for STEP 3

In this Appendix the code that has been used in *STEP 3* is presented. The script is build specifically for multivariate probability analysis with vine-copulas in 6 variables. Nevertheless, the code is proposed as general and it can be used for any amount of variables. For example, when loading the matrices (*matrices-6vines <- read.table("RVine6-clean.txt")*) and the catalog of tree equivalent classes (*catalog-6vines*), the files should be changed before its use.

```
    5pt
############################ C A L C U L A T I N G   V I N E S #############
setwd("C://00THESIS/01Output R")

## 0. DATA ##

EXTREMES_HS <- readMat("EXTREMES_HS.mat")
a <- EXTREMES_HS[1:27,1:7]
b <- EXTREMES_HS[30:57,1:7]
c <- EXTREMES_HS[59:119,1:7]
EXTREMES_HS2 <- rbind(a,b,c)

rank1 <- rank(EXTREMES_HS[["EXTREMES.HS"]][,1], na.last = TRUE, ties.
    method = "average")
rank_1 <- rank1/120
rank2 <- rank(EXTREMES_HS[["EXTREMES.HS"]][,2], na.last = TRUE, ties.
    method = "average")
rank_2 <- rank2/120
rank3 <- rank(EXTREMES_HS[["EXTREMES.HS"]][,3], na.last = TRUE, ties.
    method = "average")
rank_3 <- rank3/120
rank4 <- rank(EXTREMES_HS[["EXTREMES.HS"]][,4], na.last = TRUE, ties.
    method = "average")
rank_4 <- rank4/120
rank5 <- rank(EXTREMES_HS[["EXTREMES.HS"]][,5], na.last = TRUE, ties.
    method = "average")
rank_5 <- rank5/120
rank6 <- rank(EXTREMES_HS[["EXTREMES.HS"]][,6], na.last = TRUE, ties.
    method = "average")
rank_6 <- rank6/120
rank7 <- rank(EXTREMES_HS[["EXTREMES.HS"]][,7], na.last = TRUE, ties.
    method = "average")
rank_7 <- rank7/120
```

```
EXTREMES_HS_ranks <- cbind(Hs_ww = rank_1, WL = rank_2, Ws = rank_3, Hs_ts
    = rank_4, Tm_ww = rank_5, Tm_ts = rank_6, Ro = rank_7)
data6_HS_good2 <- cbind(Tm_ww = rank_5, Ws = rank_3, Tm_ts = rank_6, Hs_ts
    = rank_4, Hs_ww = rank_1, WL = rank_2)
## 1. Correlations to have the correct order

EXTREMES_HS_correlation = cor(EXTREMES_HS_ranks[,1:6], method = c("kendall
    "))
corrplot(EXTREMES_HS_correlation)

data6_HS_ranks<- cbind(EXTREMES_HS_ranks[,1], EXTREMES_HS_ranks[,3:7])
data6_HS_ranks<- EXTREMES_HS_ranks[,1:6]
data6_HS_ranks2<- EXTREMES_HS_ranks2[,1:6]
## BIVARIATE WITH COPULAS ##

# matrix of correlations with histograms in standard normal

data6_HS_stn <- qnorm(data6_HS_ranks)
pairs.panels(data6_HS_stn, method = "kendall")
pairs.panels(data6_HS_ranks2, method = "kendall")

EXTREMES_HS = EXTREMES_HS[["EXTREMES.HS"]]
EXTREMES_HS <- cbind(Hs_ww = EXTREMES_HS[,1] , WL = EXTREMES_HS[,2], Ws =
    EXTREMES_HS[,3], Hs_ts = EXTREMES_HS[,4], Tm_ww = EXTREMES_HS[,5], Tm_
    ts = EXTREMES_HS[,6], Ro = EXTREMES_HS[,7])
pairs.panels(EXTREMES_HS2[,1:6], method = "kendall")


tic()
counting = 0
list_of_copulas<- list()

n = 0
for (i2 in 1:6) {

  for (i3 in i2:6) {

    if (i2 != i3){
      n = n + 1

      counting = counting + 1
      print(counting)

      tic()
      cop <- BiCopSelect(data6_HS_ranks[,i2], data6_HS_ranks[,i3],
          familyset = NA, selectioncrit = "AIC",
                         indeptest = FALSE, level = 0.05, weights = NA,
                             rotations = TRUE,
                         se = FALSE, presel = TRUE, method = "mle")
      toc()

      list_of_copulas[n] <- list(cop)

    }
    else {}
```

```
  }
}

toc()


# plots of copulas and pseudo observations

filename <- paste("Copula_15.jpg", sep="")
jpeg(filename)
contour(list_of_copulas[[15]], xlab="Tm_ww", ylab="Tm_ts")
points(qnorm(data6_HS_ranks[,5]), qnorm(data6_HS_ranks[,6]))
dev.off()


## MULTIVARIATE WITH VINES ##

# analysis with 6 variables
tic()

Rvine_6_HS<- list()
Akaikes_6_HS <- list()
tata = 0
n= 0
for (i in 1:2) {

  #y = i + 10

  n = n + 1
  akaikes_six <- list()
  ssv= 9999
  Good_RV <- list()
  nv=0

  #This for is to order the tree-equivalent classes and 23040 are the
     number of Rvines for 6 nodes

  for (j in 1:23040) {

    #3
    if(catalog_6vines[j,2] == y) {

      p<- 6*j
      q<- p - 5

      mrv<- as.matrix(matrices_6vines[q:p, 1:6])

      #2.1 Starting to select the best Vines
      nv = nv + 1
      tata = tata + 1
      print(tata)

      tic()
      RV<- RVineCopSelect(data6_HS_ranks, familyset = NA, mrv,
```

```
          selectioncrit = "AIC",
                          indeptest = FALSE, level = 0.05, trunclevel = NA
                              , se = FALSE,
                          rotations = TRUE, method = "mle", cores = 4)
        toc()

        # here I save the akaike of EVERY VINE

        akaikes_six[[nv]]<- RV[["AIC"]]

        #This if is to save ONLY the best RVine of each class

        if( ssv > RV[["AIC"]] ) {

          ssv = RV[["AIC"]]
          Good_RV <- RV

        }
        else {}

      } #3
      else {}

      #2
    }

    #I store the best vine of each class and all akaikes for each class;

    #GRV_s <- summary(Good_RV)
    #filename <- paste("RvINE_6_HS_best_V",as.character(y),"BEST_FIT_pertree
        .equiv.class.txt", sep="")
    #capture.output(GRV_s, file = filename)

    # OUTPUT #

    Akaikes_6_HS[[i]]<- akaikes_six
    Rvine_6_HS[[i]]<- list(Good_RV)

    #1
}
toc()

#best vine


data6_HS_ranks2 = cbind(Tm_ww= data6_HS_ranks[,5], Ws= data6_HS_ranks[,3],
     Tm_ts=data6_HS_ranks[,6], Hs_ts= data6_HS_ranks[,4], Hs_ww= data6_HS_
    ranks[,1], WL= data6_HS_ranks[,2])
names(data6_HS_ranks)= c("Tm_ww", "Ws", "Tm_ts", "Hs_ts", "Hs_ww", "WL")

mrv_hs2 <- matrix(c( 6, 0, 0, 0, 0, 0,
                     1, 1, 0, 0, 0, 0,
                     5, 5, 5, 0, 0, 0,
                     4, 4, 4, 2, 0, 0,
                     3, 3, 3, 4, 4, 0,
                     2, 2, 2, 3, 3, 3), 6, 6, byrow = TRUE)
```

```
Best_RV6_AIC<- RVineCopSelect(data6_HS_good2[,1:6], familyset = NA, mrv_
    hs2, selectioncrit = "AIC",
                      indeptest = FALSE, level = 0.05, trunclevel = NA, se =
                          FALSE,
                      rotations = TRUE, method = "mle", cores = 4)



tic()
Best_RV6_aic_aic<- RVineStructureSelect(data6_HS_good2[,1:6], familyset =
    NA, type = 0,
                                        selectioncrit = "AIC", indeptest =
                                            FALSE, level = 0.05,
                                        trunclevel = NA, progress = FALSE,
                                            weights = NA,
                                        treecrit = "AIC", rotations = TRUE, se
                                            = FALSE,
                                        method = "mle", cores = 4)
plottoc()

tic()
Best_RV4<- RVineStructureSelect(data6_HS_good2[,1:4], familyset = NA, type
    = 0,
                                    selectioncrit = "AIC", indeptest = FALSE,
                                        level = 0.05,
                                    trunclevel = NA, progress = FALSE, weights
                                        = NA,
                                    treecrit = "tau", rotations = TRUE, se =
                                        FALSE,
                                    method = "mle", cores = 4)
toc()


summary(Best_RV5)

## 00. First I need to transfor the list into a vector via an easy code

ll <- lengths(Akaikes_6_HS)
Akaikes_6newHS<- list()

for (u in 1:22) {
  v <- c()
  for (rr in 1:ll[u]) {

    v[rr]=Akaikes_6_HS[[u]][[rr]]
  }

  Akaikes_6newHS[[u]]<- v

}

x<- c("V11", "V12", "V13", "V14", "V15", "V16", "V17", "V18", "V19", "V20"
    , "V21", "V22", "V23", "V24",
      "V25", "V26", "V27", "V28", "V29", "V30", "V31", "V32")
names(Akaikes_6newHS)<- x
```

```
boxplot(Akaikes_6newHS, xlab = "Tree-equivalent␣Regular␣Vines",
        ylab = "AIC")
#########################################################

########################## PREPARE   CLASIFICATION
    ##########
vec <- c()
mec <- c()
t = 0

for (i in 1:length(Akaikes5_HS)) {
  y = 5 + i
  tete = 0
  for (j in 1:length(Akaikes5_HS[[i]]) ) {

    t = t + 1
    tete = tete + 1

    vec[t] = Akaikes5_HS[[i]][[j]]
    mec[t] = paste("General_", as.character(t), "_V",as.character(y),"_num
        _",as.character(tete), sep = "")


  }

}

ranking_Akaikes5_HS<- cbind(mec, vec)
ranking_Akaikes5_HS <- ranking_Akaikes5_HS[order(ranking_Akaikes5_HS[,2],
    decreasing = TRUE)]
capture.output(ranking_Akaikes5_HS, file = "Akaikes5_HS.txt")
save(ranking_Akaikes5_HS, file = "RANKED_Akaikes5_HS.RData")


#########################################################

###### SAMPLING FOR MULTIVARIATE DESIGN VALUES ######

#best AIC
samples_RV6_AIC <-RVineSim(10000000, Best_RV6_AIC, U = NULL)
writeMat("samples_RV6_AIC.mat", samples_RV6_AIC = samples_RV6_AIC)
# MY BEST
tic()
samples_RV6_MYBEST <-RVineSim(20000000, Best_RV6, U = NULL)
writeMat("samples_RV6_MYBEST2.mat", samples_RV6_MYBEST = samples_RV6_
    MYBEST)
toc()
#check wich one is best

BEST_AIC_correlation = cor(samples_RV6_AIC, method = c("kendall"))
BEST_my_correlation = cor(samples_RV6_MYBEST, method = c("kendall"))
data_correlation = cor(data6_HS_good2, method = c("kendall"))

differences_AIC = abs(BEST_AIC_correlation - data_correlation)
differences_my = abs(BEST_my_correlation - data_correlation)
A = EXTREMES_HS_correlation[,1:6]
```

```
corrplot.mixed(data_correlation, lower.col = "black", number.cex = .7)
corrplot.mixed(BEST_my_correlation, lower.col = "black", number.cex = .7)
corrplot.mixed(BEST_AIC_correlation, lower.col = "black", number.cex = .7)
corrplot.mixed(EXTREMES_HS_correlation, lower.col = "black", number.cex =
    .7)
```

D

Validation tests for the chosen regular vine

In this appendix, three qualitative tests are performed to validate the selected regular vine-copula model in section 5.6. These are listed below. The results are presented for each test in sections D.1, D.2 and D.3.

- *Mass cocentration*. The sum of the 6 joint pseudo-observations is compared for observed and modeled data. The sum provides a value between 0 and 6. If the value is close to 6, all the pseudo-observations achieve large values. The opposite happens if the sum is close to 0. With this test, we aim to qualitatively compare the original and the modeled mass concentration.
- *Tail dependence*. The joint behaviour between the maximum and the minimum pseudo-observation occurring together are compared for the original and the modeled data. The joint behaviour of the maximum and minimum pseudo-observations gives information on the concept of tail dependence.
- *Bivariate observations*. Qualitative comparison between the behavior of the pairs of variables which were explicitly and not explicitly modeled in the vine copula.

D.1. Mass concentration

Despite the difference in size between the original and modeled samples, the results in figure D.1 suggest that the selected vine-copula model provides a good representation of the original mass concentration. We can see for example, that the largest concentration occurs for values of between 2.5 and 4.5 for both the original and modeled data. On another note, the upper tail seems to be well represented by the model. The upper tail is the most interesting part for the analysis in chapter 5.

D.2. Tail dependence

The results in figure D.2 suggest that neither of the 2 samples exhibit lower or upper tail dependence. Consequently, the model seems to represent the tails of the original data good enough.

D.3. Bivariate observations

Figures D.3a and D.3b present all bivariate observations for all pairs of variables for modeled data and original data, respectively. In the selected vine-copula model, only 5 of all these pairs are directly present in the first tree of the regular vine. Hence, with this test we aim to check whether the selected model represents the remaining pairs good enough. Comparing figures D.3a and D.3b, one can conclude that the modeled data represents well enough the original data. This might be easier to see in the pairs with the largest correlations which are presented in the 4 firsts columns for Tm_{ww}, Ws, Tm_{ts} and Hs_{ts}.

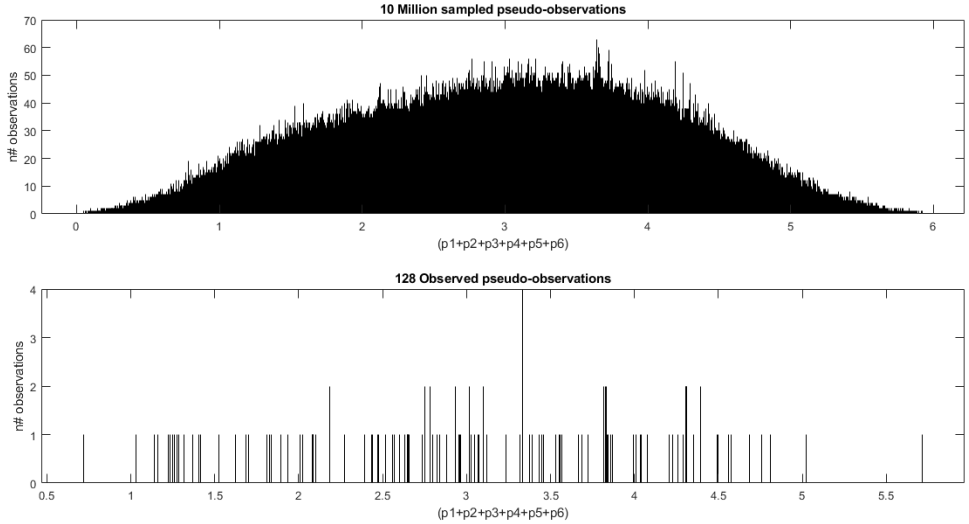


Figure D.1: Figure presenting mass concentration for the sampled data (on the top plot) and original data (on the bottom plot)

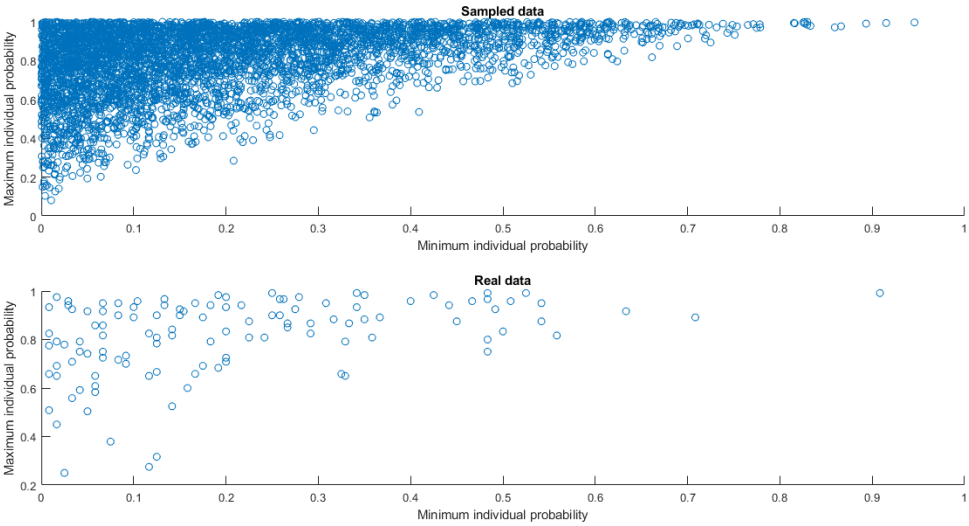
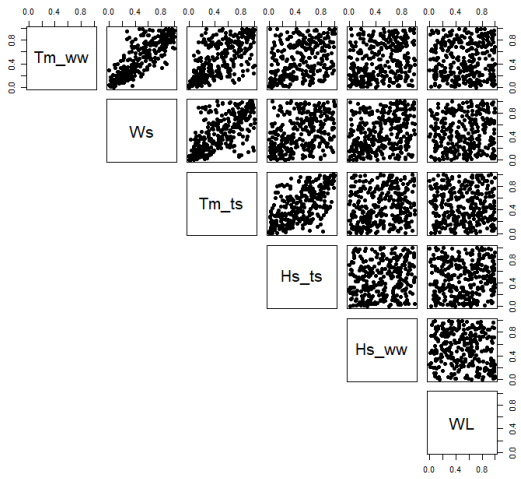
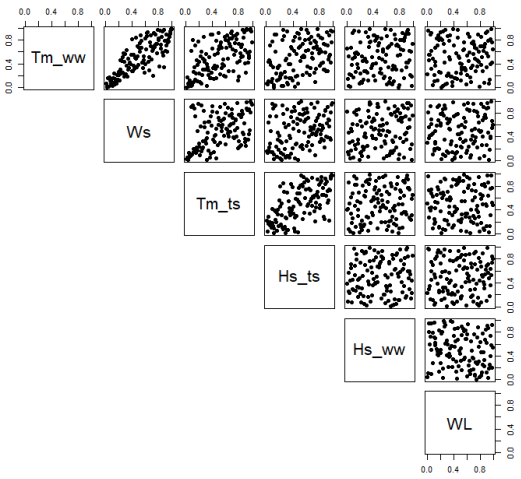


Figure D.2: Figure presenting minimum and maximum pseudo-observation that occur together for the sampled data (on the top plot) and original data (on the bottom plot)



(a) Figure presenting all bivariate pseudo-observations for all pairs of variables for the modeled data



(b) Figure presenting all bivariate pseudo-observations for all pairs of variables for the original data

E

Error estimation in the probabilities of exceedance

As mentioned in section 3.4, the true exceedance probabilities are computed with cumulative distributions functions (cdf) rather than from large modeled samples. In the case of regular vine models, the cdf is obtained by integrating a challenging probability density function (see equation 2.8), which can become computationally demanding when integrating the density numerically and challenging when integrating it analytically.

In here, we propose a hypothetical case with a regular vine only build with Clayton copulas to provide an estimation of the error of the multivariate probabilities presented in tables 5.5 and 5.6. The regular vine tree-structure is the same as for the chosen best vine in section 5.6. This error is provided for analysis with 4, 5 and 6 variables, and consequently, samples from regular vines in 4, 5 and 6 nodes. The AND and OR probabilities obtained by integrating numerically in 6 dimensions are compared to the ones obtained with the sampling procedure for the aforementioned hypothetical case, and the maximum absolute error is calculated. The comparison is made for different number of samples.

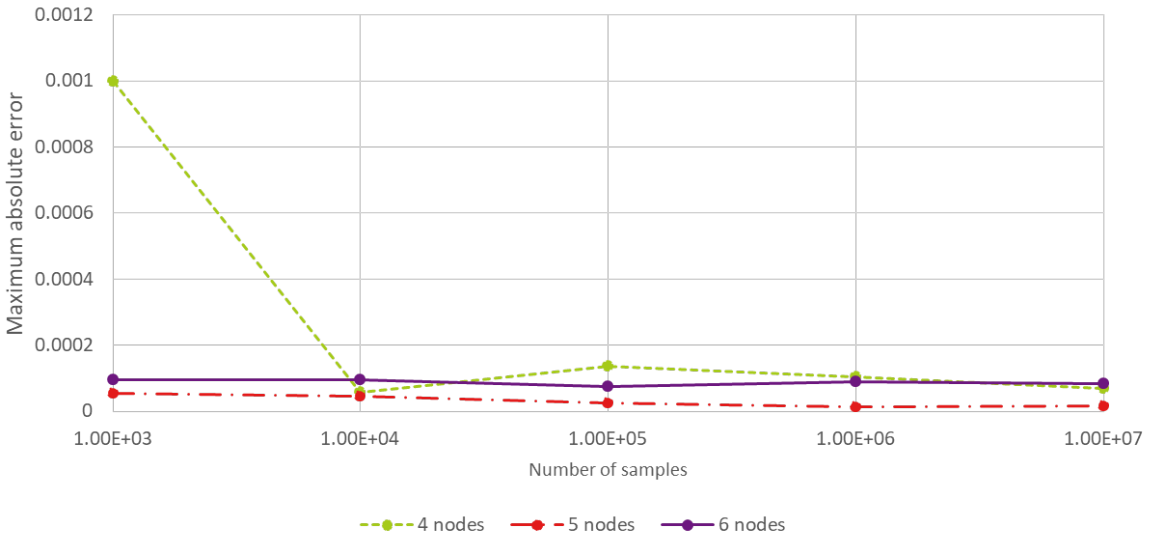


Figure E.1: Maximum absolute error of the multivariate probabilities achieved when using the sample method compared to integrating numerically. The error is presented for different sample sizes (see them in the horizontal axis)

The results in figure E.1 suggest the error is curiously the lowest for all sample sizes in the 5-variate analysis. The maximum error is achieved in the 4-variate analysis with a sample size of 10^3. In our application, the probabilities are calculated with 10^7 samples. The results in figure E.1 suggest that the maximum error in the

exceedance probabilities is of 2.00E-04. This is assumed to be acceptable as we are considering univariate probabilities of 0.001 at the lowest.

F

An approach to plot multivariate distribution functions in 2D

In this appendix, we explore a way to plot multivariate densities and cumulative distributions in 2D or 3D. The reader should note that this appendix is part of the recommendations for further research presented in section 7.3. One of the reasons why advance statistical models are not taken up by the industry is due to their complexity and abstractism. This might be solved by providing graphical tools to illustrate concepts such as multivariate distribution functions, which are difficult to imagine.

The main idea would be to portray all the margins of the variables (i.e. axis) in 2D and use the third dimension to depict the density or the cumulative probability. The most challenging part in this task is to find a way to plot the n-variate (joint) observations in 2D. Because in reality, these 6 values that compose such joint observation would intersect in a hyperplane. In here, we propose to plot these (pseudo-)observations in concentric axes that together would form a semi-circle. This approach is thought to illustrate univariate margins in hypercube units, such as the ones required by copulas and vine-copulas. At the center of the semicircle one would place the value 0. In the radius of the semicircle, one would place the maximum value which in our case is 1. The axes should be uniformly spaced in the semi-circle. Then, one would have to unite the axes by the pseudo-observations to form an area. This method we call the *semi-circle approach*. An example of this is depicted in figure F.1 for 5 variables (i.e. 5 margins). In this figure we show how can one plot 5-variate pseudo-observations. This example can be extended to higher dimensions easily.

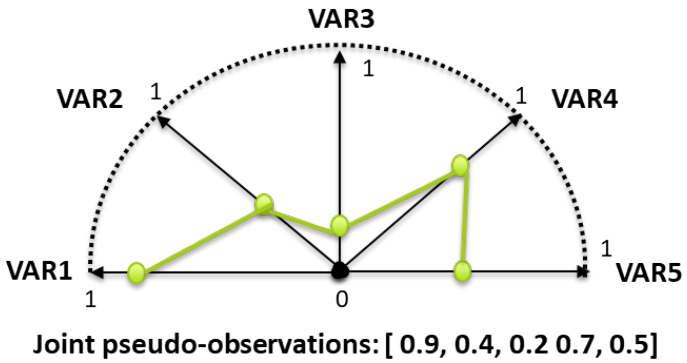


Figure F.1: Example on how to plot 5-variate pseudo-observations in 2D using concentric axes inscribed in a semicircle

Continuing with the example in 5 dimensions, one can calculate the empirical cumulative distribution with a (large enough) sample of modeled data. With the approach proposed in here, this could be done using the area inscribed in the semi-circle by the 5 pseudo-observations. This area should go from 0 to $\pi/2$. So when all pseudo-observations are 0, the area should be 0 and the associated cumulative probability also 0. When all pseudo-observations are 1, the area should be $\pi/2$ and the associated cumulative probability should be 1. The

axis associated with the cumulative probability is then transformed to a scale from 0 to $\pi/2$, or it can also work vice versa. This area allows one to treat the sets of 5 pseudo-observations as one unit. Using this approach, the multivariate cumulative function can be plotted in 3D. For example, if all 5 variables would increase with the same phase and magnitude, their cumulative distribution function plotted using the semi-circle approach would look similar to the one depicted in figure F.2.

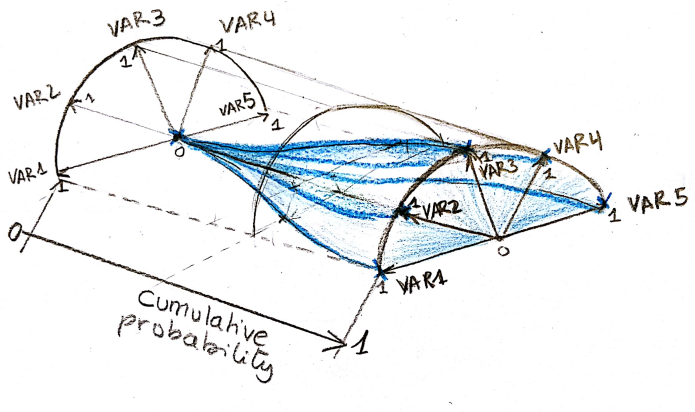


Figure F.2: Example on how a 5-variate cumulative distribution function would look like using concentric axes inscribed in a semicircle

Bibliography

[1] *eartH2Observe Water Cycle Integrator (WCI)*. https://wci.earth2observe.eu/, 2019.

[2] Kjersti Aas, Claudia Czado, Arnoldo Frigessi, and Henrik Bakken. Pair-copula constructions of multiple dependence. *Insurance: Mathematics and economics*, 44(2):182–198, 2009.

[3] M. ajabi Bahaabadi and Morales-Napoles O. Vine copulas and bayesian networks for estimating route travel time distribution. *Working Paper*, 2019.

[4] H. Akaike. Dependence structure and extreme comovements in international equity and bond markets. *Journal of Banking & Finance*, 35(8):1954–1970, 2011.

[5] Guedes Soares C. Antao, E.M. Approximation of bivariate probability density of individual wave steepness and height with copulas. *Coastal Engineering*, 89:45–52, 2014.

[6] Tim Bedford, Roger M Cooke, et al. Vines–a new graphical model for dependent random variables. *The Annals of Statistics*, 30(4):1031–1068, 2002.

[7] Pietro Bernardara, Franck Mazas, Jérôme Weiss, Marc Andreewsky, Xavier Kergadallan, Michel Benoît, and Luc Hamm. On the two step threshold selection for over-threshold modelling. *Coastal Engineering Proceedings*, 1(33):42, 2012.

[8] Eike Christain Brechmann and Claudia Czado. Risk management with high-dimensional vine copulas: An analysis of the euro stoxx 50. *Statistics & Risk Modeling*, 30(4):307–342, 2013.

[9] Luigi Cavaleri and Mauro Sclavo. The calibration of wind and wave model data in the mediterranean sea. *Coastal Engineering*, 53(7):613–627, 2006.

[10] Abderrahmane Chakak and Kenneth J Koehler. A strategy for constructing multivariate distributions. *Communications in Statistics-Simulation and Computation*, 24(3):537–550, 1995.

[11] Fateh Chebana and Taha BMJ Ouarda. Multivariate quantiles in hydrological frequency analysis. *Environmetrics*, 22(1):63–78, 2011.

[12] Ouarda T.B.M.J. Chebana, F. Multivariate quantiles in hydrological frequency analysis. *Environmetrics*, 22:63–78, 2011.

[13] Stansby. P.K. Chini, N. Extreme values of wave overtopping accounting for climate change and sea level rise. *Coastal Engineering*, 65:27–37, 2012.

[14] date of access. Copernicus Climate Change Service Climate Data Store (CDS). *ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate*. https://cds.climate.copernicus.eu/cdsapp/home, 2017.

[15] Stretch Derek.D. Corbella, S. Simulating a multivariate sea storm using archimedean copulas. *Coastal Engineering*, 76:68–78, 2013.

[16] Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, and Clifford Stein. *Introduction to algorithms*. MIT press, 2009.

[17] Anaïs Couasnon, Antonia Sebastian, and Oswaldo Morales-Nápoles. A copula-based bayesian network for modeling compound flood hazard from riverine and coastal interactions at the catchment scale: An application to the houston ship channel, texas. *Water*, 10(9):1190, 2018.

[18] Claudia Czado, Stephan Jeske, and Mathias Hofmann. Selection strategies for regular vine copulae. *Journal de la Société Française de Statistique*, 154(1):174–191, 2013.

[19] Laurens De Haan and Ana Ferreira. *Extreme value theory: an introduction*. Springer Science & Business Media, 2007.

[20] Carlo De Michele, Gianfausto Salvadori, Giuseppe Passoni, and Renata Vezzoli. A multivariate model of sea storms using copulas. *Coastal Engineering*, 54(10):734–751, 2007.

[21] van Gelder P.H.AJ.M. de Waal, D.J. Modelling of extreme wave heights and periods through copulas. *Extremes*, 8:345–356, 2005.

[22] P. Deheuvels. La fonction de dépendance empirique et ses propriétés: Un test non paramétrique d'indépendance. *Bull. Cl. Sci., Acad. R. Belg.,*, 65, 6, 1979.

[23] Jeffrey Dissmann, Eike C Brechmann, Claudia Czado, and Dorota Kurowicka. Selecting and estimating regular vine copulae and application to financial returns. *Computational Statistics & Data Analysis*, 59: 52–69, 2013.

[24] KW Dupuis and A Anis. Observations and modeling of wind waves in a shallow estuary: Galveston bay, texas. *Journal of Waterway, Port, Coastal, and Ocean Engineering*, 139(4):314–325, 2012.

[25] Svenja Fischer and Andreas Schumann. Comparison between classical annual maxima and peak over threshold approach concerning robustness. 2014.

[26] Christian Genest and Anne-Catherine Favre. Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of hydrologic engineering*, 12(4):347–368, 2007.

[27] Jean D Gibbons and Jean D Gibbons Fielden. *Nonparametric measures of association*. Number 91. Sage, 1993.

[28] Benedikt Gräler, Martinus van den Berg, Sander Vandenberghe, Andrea Petroselli, Salvatore Grimaldi, Bernard De Baets, and Niko Verhoest. Multivariate return periods in hydrology: a critical and practical review focusing on synthetic design hydrograph estimation. *Hydrology and Earth System Sciences*, 17(4): 1281–1296, 2013.

[29] Peter J Green, Sylvia Richardson, Nils Lid Hjort, et al. *Highly structured stochastic systems*, volume 27. Oxford University Press on Demand, 2003.

[30] Jeffrey L Hanson and Owen M Phillips. Automated analysis of ocean surface directional wave spectra. *Journal of atmospheric and oceanic technology*, 18(2):277–293, 2001.

[31] Leo H. Holthuijsen. *Waves in Oceanic and Coastal Waters*. 2007. doi: 10.2277/0521860288.

[32] Jonathan RM Hosking and James R Wallis. Parameter and quantile estimation for the generalized pareto distribution. *Technometrics*, 29(3):339–349, 1987.

[33] H Joe. Multivariate models and multivariate dependence concepts. *Monographs on statistics and applied probability, Chapman & Hall*, 73, 1997.

[34] Harry Joe. Multivariate extreme-value distributions with applications to environmental data. *Canadian Journal of Statistics*, 22(1):47–64, 1994.

[35] Harry Joe. *Multivariate models and multivariate dependence concepts*. Chapman and Hall/CRC, 1997.

[36] Harry Joe. *Dependence modeling with copulas*. Chapman and Hall/CRC, 2014.

[37] Harry Joe and Dorota Kurowicka. *Dependence modeling: vine copula handbook*. World Scientific, 2011.

[38] Dorota Kurowicka. *Optimal truncation of vines. Dependence modeling: vine copula handbook*. World Scientific, 2011.

[39] Dorota Kurowicka and Roger Cooke. A parameterization of positive definite matrices in terms of partial correlation vines. *Linear Algebra and its Applications*, 372:225–251, 2003.

[40] M Lang, TBMJ Ouarda, and B Bobée. Towards operational guidelines for over-threshold modeling. *Journal of hydrology*, 225(3-4):103–117, 1999.

[41] F Li, PHAJM Van Gelder, R Ranasinghe, DP Callaghan, and RB Jongejan. Probabilistic modelling of extreme storms along the dutch coast. *Coastal Engineering*, 86:1–13, 2014.

[42] Simmonds D. Reeve D. Li, Y. Quantifying uncertainty in extreme values of design parameters with reampling techniques. *Ocean Engineering*, 35:1029–1038, 2008.

[43] van Gelder P.H.A.J.M. Ranasinghe R. Callarhan D.P. Jongejan R.B. Li, F. Probabilistic modelling of extreme storms along the dutch coast. *Coastal Engineering*, 86:1–13, 2014.

[44] Jue Lin-Ye, Manuel Garcia-Leon, V Gracia, and A Sanchez-Arcilla. A multivariate statistical model of extreme events: An application to the catalan coast. *Coastal Engineering*, 117:138–156, 2016.

[45] Frank J Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.

[46] Franck Mazas and Luc Hamm. A multi-distribution approach to pot methods for determining extreme wave heights. *Coastal Engineering*, 58(5):385–394, 2011.

[47] Franck Mazas and Luc Hamm. An event-based approach for extreme joint probabilities of waves and sea levels. *Coastal Engineering*, 122:44–59, 2017.

[48] Salvadori G. Passoni G . Vezzoli R. Michele, C. A multivariate model of sea storms using copulas. *Coastal Engineering*, 54(10):734–751, 2007.

[49] Oswaldo Morales Nápoles. Bayesian belief nets and vines in aviation safety and other applications. 2010.

[50] Oswaldo Morales-Nápoles. *Counting vines*. World Scientific, 2010.

[51] Oswaldo Morales Napoles, Roger M Cooke, and Dorota Kurowicka. About the number of vines and regular vines on n nodes. 2010.

[52] ID Morton and J Bowers. Extreme value analysis in a multivariate offshore environment. *Applied Ocean Research*, 18(6):303–317, 1996.

[53] O Morales Nápoles. *Bayesian belief nets and vines in aviation safety and other applications*. Delft: TU, 2009.

[54] Roger B Nelsen. *An introduction to copulas*. Springer Science & Business Media, 2007.

[55] Filip Lindskog Paul Embrechts and Alexander McNeil. Modelling dependence with copulas and applications to risk management. *Department of Mathematics ETHZ, www.math.ethz.ch/finance*, 2001.

[56] Petya G Petrova, M Aziz Tayfun, and C Guedes Soares. The effect of third-order nonlinearities on the statistical distributions of wave heights, crests and troughs in bimodal crossing seas. *Journal of Offshore Mechanics and Arctic Engineering*, 135(2), 2013.

[57] PG Petrova and C Guedes Soares. Distributions of nonlinear wave amplitudes and heights from laboratory generated following and crossing bimodal seas. *Natural Hazards and Earth System Sciences*, 14(5):1207–1222, 2014.

[58] Reddy M. Poulomi, G. Probabilistic assessment of flood risks using trivariate copulas. *Theor. Appl. Climatol.*, 111:341–360, 2013.

[59] T Pullen, NWH Allsop, T Bruce, A Kortenhaus, H Schüttrumpf, and JW Van der Meer. Eurotop, european overtopping manual-wave overtopping of sea defences and related structures: assessment manual. *Also published as special volume of Die Küste*, 2007.

[60] D. Simmondsa A. Rabya R. Janea, L. Dalla Valleb. A copula-based approach for the estimation of wave height records through spatial correlation. *Coastal Engineering*, 117:1–18, 2016.

[61] AI Requena, Luis Jesús Mediero Orduña, and Luis Garrote de Marcos. A bivariate return period based on copulas for hydrologic dam design: accounting for reservoir routing in risk estimation. *Hydrology and Earth System Sciences*, 17(8):3023–3038, 2013.

[62] Cooke R.M. Markov and entropy properties of tree and vines-dependent variables. In *In Proceedings of the ASA Section of Bayesian Statistical Science*, pages 517–530. American Statistical Association, Washington, 1997.

[63] David Salas-Monreal, Ayal Anis, and David Alberto Salas-de Leon. Galveston bay dynamics under different wind conditions. *Oceanologia*, 60(2):232–243, 2018.

[64] G Salvadori, GR Tomasicchio, and F D'Alessandro. Practical guidelines for multivariate analysis and design in coastal and off-shore engineering. *Coastal Engineering*, 88:1–14, 2014.

[65] G Salvadori, F Durante, C De Michele, M Bernardi, and L Petrella. A multivariate copula-based framework for dealing with hazard scenarios and failure probabilities. *Water Resources Research*, 52(5):3701–3721, 2016.

[66] Gianfausto Salvadori, Carlo De Michele, Nathabandu T Kottegoda, and Renzo Rosso. *Extremes in nature: an approach using copulas*, volume 56. Springer Science & Business Media, 2007.

[67] Cornelia Savu and Mark Trede. Hierarchies of archimedean copulas. *Quantitative Finance*, 10(3):295–304, 2010.

[68] Ulf Schepsmeier, Jakob Stoeber, Eike Christian Brechmann, Benedikt Graeler, Thomas Nagler, Tobias Erhardt, Carlos Almeida, Aleksey Min, Claudia Czado, Mathias Hofmann, et al. Package 'vinecopula'. *R package version*, 2(5), 2018.

[69] F. Serinaldi. Dismissing return periods! *Stochastic Environmental Research. Risk Assess.*, 29(4):1179–1189, 2015. doi: 10.1007/s00477-014-0916-1.

[70] NOAA Bathymetric Data Viewer. https://maps.ngdc.noaa.gov/viewers/bathymetry/?layers=dem. 2019.

[71] Mudersbach C. Jensen J. Wahl, T. Assessing the hydrodynamic boundary conditions for risk analyses in coastal areas: a stochastic storm surge model. *Nat. Hazards Earth Syst. Sci.*, 11:2925–2939, 2011.

[72] T Wahl, C Mudersbach, and J Jensen. Assessing the hydrodynamic boundary conditions for risk analyses in coastal areas: a multivariate statistical approach based on copula functions. *Natural Hazards and Earth System Science*, 12(2):495–510, 2012.

[73] Changjiang Xu, Jiabo Yin, Shenglian Guo, Zhangjun Liu, and Xingjun Hong. Deriving design flood hydrograph based on conditional distribution: a case study of danjiangkou reservoir in hanjiang basin. *Mathematical Problems in Engineering*, 2016, 2016.

[74] IR Young, LA Verhagen, and ML Banner. A note on the bimodal directional spreading of fetch-limited wind waves. *Journal of Geophysical Research: Oceans*, 100(C1):773–778, 1995.

[75] S Zachary, G Feld, G Ward, and J Wolfram. Multivariate extrapolation in the offshore environment. *Applied Ocean Research*, 20(5):273–295, 1998.