Geoinformatik

DEVELOPING SPATIO-TEMPORAL COPULAS

Inauguraldissertation
zur Erlangung des akademischen Grades
eines Doktors der Naturwissenschaften
durch den Fachbereich Geowissenschaften
der Westfälischen Wilhelms-Universität Münster

Vorgelegt von
Benedikt Gräler
aus Lengerich (Westf.)
im August 2014

Dekan:	Prof. Dr. Hans Kerp
Erstgutachter:	Prof. Dr. Edzer Pebesma
Zweitgutachter:	Prof. Dr. Claudia Czado
Tag der mündlichen Prüfung:	19. Nov 2014

For Lena-Marie and my two daughters.

"Essentially, all models are wrong,
but some are useful."

— George E. P. Box

ACKNOWLEDGEMENTS

Thanks to:

- My supervisor Prof. Dr. Edzer Pebesma for introducing me to geostoatistics, R development and the spatial statistics community.
- Prof. Dr. Claudia Czado for in-depth examinations of vine copulas and scientific invitations to Munich.
- My colleagues at the Institute for Geoinformatics for interesting scientific and non-scientific discussions.
- My co-authors and co-developers for productive and fun collaborations.
- My family for their support and love.

The research presented in this thesis has been funded by the German Research Foundation (DFG) under project PE-1632/4-1.

ABSTRACT

Continuously spreading environmental processes are often only observed at a distinct set of locations and a continuous model is adopted to capture the phenomenon across space. Temperature measurements recorded at weather stations that are used to produce smooth maps is one of the many examples. Hence, an interpolation has to take place to derive values at unobserved locations. Besides deterministic approaches, kriging is a widely used probabilistic interpolation technique. At times, certain properties of the underlying multivariate Gaussian distribution are not desired and a more flexible probabilistic representation of the phenomenon is needed. Copulas have proven to be a useful tool to build up non-Gaussian distributions. Vine copulas allow to flexibly combine bivariate copulas to multivariate copulas leading to distributions of higher dimensions.

This thesis presents a new approach that allows to build vine copulas that are aware of separating distances across space and time. To achieve this, the building blocks of the vine copula are composed out of convex combinations of bivariate copulas. The weight of the convex combination as well as the copulas' parameters are defined by distance over space and time. Different use cases are considered to asses power and quality of this new probabilistic modelling approach. A prototypical implementation is available as R package and as well presented in this thesis. The implementations in R have been made in conjunction with the research to empirically support this new development.

While improvements of the interpolation in terms of cross-validation statistics depend on the application scenario, the obtained confidence bands have desirable properties. This is partly due to the conditional predictive distribution being able to take any shape for each prediction location. Additionally, the freedom to choose any marginal distribution ensures that the confidence intervals are within the range of the modelled distribution. Furthermore, the confidence intervals depend on the predicted value and the layout of the local neighbourhood. Therefore, the spatial vine copula approach is assumed to provide in general a more realistic view of the uncertainties than the kriging variance. While this approach is still in its infancy, its potential to improve especially the modelling of heavily skewed spatial random fields becomes apparent.

ZUSAMMENFASSUNG

Natürliche Prozesse mit kontinuierlicher Ausbreitung werden oft nur an einer Anzahl diskreter Punkte beobachtet. Um den Prozess im Raum zu erfassen, wird dann ein kontinuierliches Modell angenommen. Ein Beispiel sind Temperaturmessungen, bei denen aus Beobachtungen an einigen Wetterstationen kontinuierliche Karten erzeugt werden. Hierbei wird ein Interpolationsverfahren benutzt, das ermöglicht Werte an Orten zu schätzen, an denen keine Beobachtungen gemacht wurden. Neben deterministischen Ansätzen ist Kriging eine weitverbreitete, probabilistische Interpolationsmethode. Jedoch sind die Eigenschaften der zugrunde liegenden multivariaten Gauss'schen Verteilung manchmal nicht zutreffend und ein flexibleres probabilistisches Modell wird benötigt. Copulas haben sich zur Bildung flexibler, nicht-Gauss'scher Verteilungen als sehr nützlich erwiesen. Vine copulas ermöglichen es, verschiedene bivariate Copulas zu einer multivariaten Copula zusammenzusetzen, um mehrdimensionale Verteilungen zu bilden.

Diese Arbeit stellt eine neue Methode vor, die vine copulas entsprechend der Entfernungen im Raum und in der Zeit parametrisiert. Um dies zu erreichen, werden die Bausteine der vine copulas durch Konvexkombinationen aus bivariaten Copulas gebildet. Dabei basieren die Gewichte der Konvexkombination und die Parameter der Copulas selbst auf der Entfernung zwischen den modellierten Orten. Verschiedene Anwendungen werden betrachtet, um die Möglichkeiten des neuen probabilistischen Modells zu erproben. Eine prototypische Implementierung steht in einem R-Paket bereit und wird ebenfalls in dieser Arbeit vorgestellt. Die Implementierung ist parallel zur Forschung erfolgt und hat den Ansatz empirisch ergänzt.

Verbesserungen der Interpolation in Bezug auf die Kreuzvalidierungsstatistiken sind abhängig vom entsprechenden Anwendungsfall. Die zugehörigen Konfidenzbänder bringen jedoch wünschenswerte Eigenschaften mit sich. Dies ist zum Teil darin begründet, dass in der Interpolation die bedingte Verteilungsfunktion eines jeden Ortes jede Form annehmen kann. Des Weiteren stellt die unabhängige Wahl der Randverteilungen sicher, dass die Konfidenzintervalle ausschließlich im Wertebereich der modellierten Verteilung liegen. Darüber hinaus hängen die Konfidenzintervalle neben der Anordnung der benachbarten Orte auch von den vorhergesagten Werten ab. Somit ermöglicht die spatial vine Copula Methode eine realistischere Angabe der Unsicherheiten als die Kriging-Varianz. Auch wenn diese Methode noch sehr jung ist, ist ihr Potenzial, die Modellierung insbesondere stark schiefer räumlicher Zufallsfelder zu verbessern, erkennbar.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

1.1 MOTIVATION

With the aim to better understand the environment we live in, all kinds of environmental variables are recorded on a regular basis at measurement stations spread over space. A large application of these recordings is devoted to weather and addresses for instance temperatures. Obviously, air temperature is not only of interest at a few measurement locations and moments in time, but also as a continuous variable surrounding us. Hence, the observed values need to be interpolated at unobserved locations. This can be achieved in many different ways starting with simple deterministic models such as assigning the value of the nearest measurement station to any unobserved location or calculating a distance weighted average of the observed neighbours.

Often, deterministic models are not sufficient to capture the behaviour of a process well enough and a probabilistic model is adopted. A key feature of most probabilistic spatial models is the *dependence* of measurements on each other and how this dependence changes with varying distance between locations. This dependence and the spread of the values is captured in a multivariate distribution defined over the complete study area or a local subset only. Probably, the interpolation techniques chosen most frequently are the variants of *kriging* [see e.g. 15]. The driving probabilistic model of kriging is a Gaussian Random Field, a multivariate Gaussian distribution with a covariance structure typically parametrized by a variogram. At times, the choice of the Gaussian distribution might be too restrictive and more flexible multivariate distributions are needed. This is where the theory of *copulas* (see e.g. [57]) has proven extremely useful.

Two limitations of the Gaussian multivariate distributions are for instance its elliptical symmetry and the independence of joint extreme events regardless of the modelled covariance. In the field of spatial statistics, the elliptical symmetry implies the same strength of dependence for a pair of high values (e.g. the 95-percentiles) and a pair of low values (e.g. the 5-percentiles) at two spatial locations. This attributes to a smoothing effect of kriging where dips and bumps have the same shape. An example where this asymmetry becomes apparent is an elevation model of a mountainous area. Valleys (the low values) have typically a much smoother shape than the mountain ridges (the high values). Applying the concept of copulas can

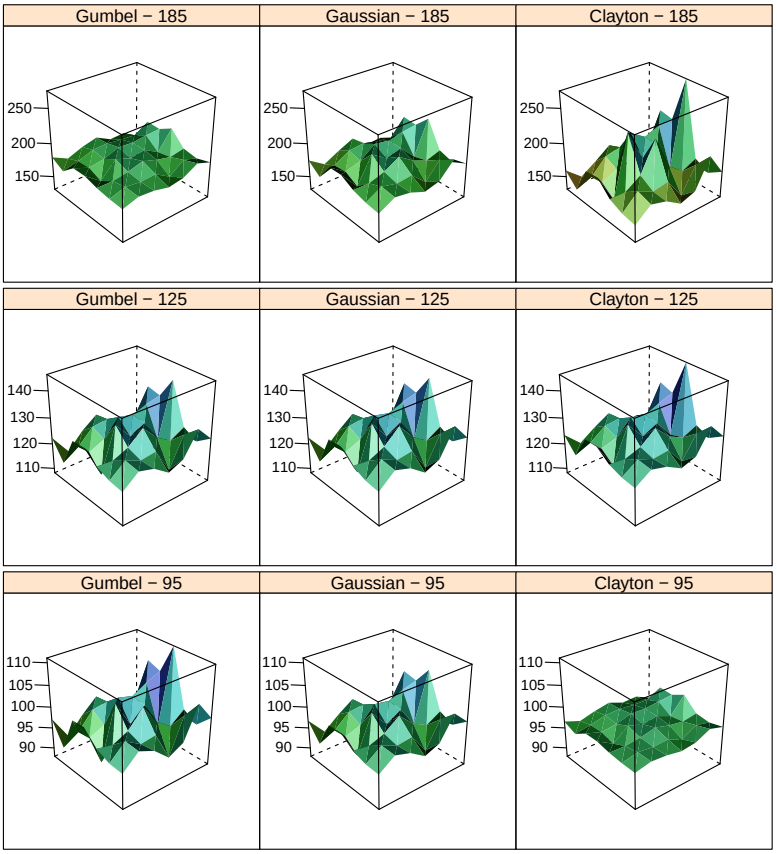

Figure 1.1: The copula's influence on the pattern of a spatial random field. The central cell is fixed to the same value per row (see panel title, corresponding to the 0.95-, 0.5- and 0.05-quantile of the same distribution). The surrounding cells are simulated from the different spatial copulas. The spatial copulas exhibit the same strength of dependence that changes identically with distance from the centre.

overcome this issue by using different copula families than the Gaussian family (e.g. the Gumbel or Clayton families).

Influence of the choice of copula family (e.g. exhibiting an elliptical symmetry) on a spatial random field is illustrated in Figure 1.1. Here, the central location has a fixed value per row (185, 125 and 95 top to bottom) corresponding to the 0.95-, 0.5- and 0.05-quantile of the same distribution. The surrounding cells are simulated conditioned under the central cell (but independently from each other). All random fields (i.e. copulas) exhibit the same correlation (in terms of Kendall's tau) that changes identically with distance from the centre. The Gaussian copula (middle column) allows for some variability in all three cases. The Gumbel copula puts a stronger dependence on larger quantiles and thus allows only for a small variability for high values resulting in a high plateau (top row, left most plot) while low values allow for a larger variability (bottom row, left most plot). The Clayton copula exhibits a stronger dependence for small quantiles

and hence only small variability for the low values resulting in a low plain (bottom row, right most plot) while high values allow for large variability (top row, right most plot). Inspecting the second row of Figure 1.1 reveals that the choice of copula (out of these three families) does only imply vague differences on the middle part of the distribution.

The differences depicted in Figure 1.1 are only due to the dependence structure, the copula, of the spatial random field. The copula's effect on the mean or median value might be moderate, but comes into play when the tails of a distribution are of interest. Therefore, the choice of copula will as well change the confidence band associated with each prediction. These aspects motivate the research on spatial and spatio-temporal copulas.

1.2 OBJECTIVES

The key theme of this thesis is how copulas can contribute to the modelling of temporal, spatial and spatio-temporal phenomena. As the class of copulas is generally wide, the focus is on connecting bivariate copulas through *vines* [8] to higher dimensional *vine copulas* [2]. The following enumeration provides an overview of how the major theme of this thesis is split into building blocks:

1. *How can vine copulas contribute to the modelling of extremes in time series of environmental phenomena?*
2. *How can bivariate spatial/spatio-temporal copulas be connected in a vine copula to model spatial/spatio-temporal random fields?*
3. *How can covariates be included in spatial or spatio-temporal vine copulas?*

These questions explore the power of vine copulas for various commonly found types of environmental data.

1.3 SCOPE

This research focuses on the development of spatial and spatio-temporal vine copulas to model spatial and spatio-temporal random fields. The prototypical software development empirically underpinning this research is centred around a few data sets but is designed to be applicable to a wide set of use cases. Other studies (e.g. [5, 53]) have developed spatial copulas based on single family multivariate copulas, but the central building blocks to the approach presented in this thesis are bivariate spatial copulas that lead to spatial, temporal or spatio-temporal multivariate vine copulas.

1.4 APPROACH

An overview of the central steps that integrate spatial and spatio-temporal bivariate copulas through vine copulas in the probabilistic modelling of spatial and spatio-temporal random fields is given below. Core concepts are briefly introduced and illustrated and the developed spatial and spatio-temporal vine copulas are sketched. Details on this construction and extensions thereof can be found in the subsequent chapters. Each of the chapters 2 to 6 constitutes a separate scientific publication that has already been published or has been submitted for publication.

1.4.1 *Spatial and Spatio-Temporal Random Field*

The theoretical notion of a real-valued spatial random field Z can be given as a continuously indexed set of real-valued random variables $Z : \Omega \times \mathcal{S} \to \mathbb{R}$. Where Ω is some probability space and $\mathcal{S}$, the spatial index set, describes the region of interest. A random variable at location $s \in \mathcal{S}$ is denoted by $Z(s)$. Even though the random field is continuously indexed, it is typically evaluated only at a set of discrete locations $s_1, \ldots, s_n \in \mathcal{S}$. Such a discrete representation can then be modelled by a n-dimensional multivariate distribution capturing the dependence between locations. In the case of a real-valued spatio-temporal random field, we refer analogously to $Z : \Omega \times \mathcal{S} \times \mathcal{T} \to \mathbb{R}$ where $\mathcal{T}$ is a temporal index set describing the time interval of interest. Mostly, $\mathcal{T}$ is perceived as continuous but observations are typically made only at a discrete set of timestamps. The random variable at location $(s, t) \in \mathcal{S} \times \mathcal{T}$ is denoted by $Z(s, t)$.

1.4.2 *Copulas*

The theory of *copulas* has its origin in the theorem by Sklar [85]. The theorem's power lies in the split of multivariate distributions H of any dimension $d \geqslant 2$ into their marginal distributions $F_1, \ldots, F_d$ and the *copula* C combining these into the multivariate cumulative distribution H:

$$H(x_1, \ldots, x_d) = C\big(F_1(x_1), \ldots, F_d(x_d)\big)$$

This allows for a huge flexibility in the design of multivariate distributions, as countless combinations of marginal distributions and suitable copulas will lead to well defined multivariate distribution functions. Due to the probability integral transform, all $F_i(x_i)$ will be uniformly distributed. Hence, a copula C is defined on the unit hypercube $[0, 1]^d$. To reduce notation, we set $d = 2$ and look at bivariate copulas $C : [0, 1]^2 \to [0, 1]$ only. A copula can be imagined as

multivariate cumulative distribution function due to its requirements of

$$C(u_1, 0) = 0 = C(0, u_2) \text{ and } C(u_1, 1) = u_1 \text{ and } C(1, u_2) = u_2$$

for any $u_1, u_2 \in [0, 1]$. Additionally, for every u_1, u_2, v_1, $v_2 \in [0, 1]$ with $u_1 \leqslant v_1$ and $u_2 \leqslant v_2$ a copula needs to fulfil

$$C(v_1, v_2) - C(u_1, v_2) - C(v_1, u_2) + C(u_1, u_2) \geqslant 0$$

(see for instance Nelsen [57] for further details). A *copula's density* c can be interpreted as the strength of dependence. Independence of two random variables is reflected by a copula density $c(u_1, u_2) \equiv 1$ and the corresponding copula $\Pi(u_1, u_2) := u_1 u_2$ commonly referred to as the *product copula*. A frequently used copula is the *Gaussian copula*

$$C_N(u_1, u_2) := \Phi_\rho\big(\Phi^{-1}(u_1), \Phi^{-1}(u_2)\big)$$

with well defined correlation matrix $\rho \in [-1, 1]^{2\times 2}$ of a standard normal bivariate Gaussian distribution Φ_ρ and standard normal univariate Gaussian distributions Φ. The class of elliptical copulas is accompanied by the *Student t copula* C_t that follows the same construction principle as the Gaussian copula. A wide class of bivariate families being all symmetric is formed by the *Archimedean copulas* that follow the construction principle

$$C(u_1, u_2) := \phi^{[-1]}\big(\phi(u_1) + \phi(u_2)\big)$$

with $\varphi : [0, 1] \to [0, \infty]$ being a continuous, strictly decreasing function with $\varphi(1) = 0$ and $\varphi^{[-1]}$ its pseudo-inverse. Prominent members of this class are the *Clayton* C_C, *Frank* C_F and *Gumbel copulas* C_G (see Nelsen [57, Chapter 4] for further details and more families). Several copula families have multivariate extensions but some lack flexibility as they only allow for a single parameter. Furthermore, a multivariate distribution might exhibit different pairwise dependence structures across its margins and thus a single family might not be sufficient to capture this behaviour. These limitation can be overcome with *vine copulas*.

1.4.3 *Vine copulas*

The construction principle of *vine copulas* has first been published by Aas et al. [2] as the *pair-copula construction* based on work by Bedford and Cooke [8] that gave rise to the current naming of *vine copulas*. The general idea of the pair-copula construction is that multivariate copulas can be approximated with a cascade of bivariate copula building blocks. While a vine copula *is* a multivariate copula, they might only *approximate* the target copula because not all copulas can be re-build

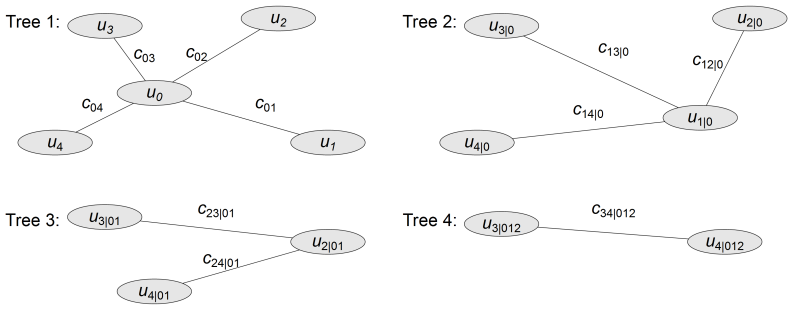

Figure 1.2: A five dimensional *canonical vine* with its four trees and ten bivariate copulas that leads to the copula density in (1.1). The conditional variables $u_{k|0\ldots j-1}$ can be calculated from the previous bivariate copulas as in (1.2).

as a vine copula as has been discussed by Hobæk Haff, Aas, and Frigessi [44].

In general the decomposition of the multivariate copula into bivariate building blocks is not unique and many different regular vines exist denoting the decompositions. Hence, the choice of the actual vine might affect the goodness of fit to the multivariate target copula. In the spatial and spatio-temporal case where we use a neighbourhood to define the decomposition, the canonical vine where all initial dependencies are with respect to the central location is a natural choice. A five-dimensional canonical vine is shown in Figure 1.2. The density of the respective pair-copula construction reads:

$$\begin{aligned} c(u_0,\ldots,u_4) &= c_{01}(u_0,u_1)\cdot c_{02}(u_0,u_2)\cdot c_{03}(u_0,u_3)\cdot c_{04}(u_0,u_4) \\ &\quad \cdot c_{12|0}(u_{1|0},u_{2|0})\cdot c_{13|0}(u_{1|0},u_{3|0})\cdot c_{14|0}(u_{1|0},u_{4|0}) \\ &\quad \cdot c_{23|01}(u_{2|01},u_{3|01})\cdot c_{24|01}(u_{2|01},u_{4|01}) \\ &\quad \cdot c_{34|012}(u_{3|012},u_{4|012}) \end{aligned} \tag{1.1}$$

where

$$u_{k|0\ldots j} = \frac{\partial C_{jk|0\ldots j-1}(u_{j|0\ldots j-1},u_{k|0\ldots j-1})}{\partial u_{j|0\ldots j-1}} \tag{1.2}$$

with $0 \leqslant j < k \leqslant 5$. This procedure can be extended to higher dimensions in a natural manner.

1.4.4 *Spatial and spatio-temporal vine copulas*

The motivation to use spatial or spatio-temporal vine copulas is to model non-Gaussian spatial or spatio-temporal random fields Z, where the non-Gaussianity not only refers to marginal distributions, but also to the dependence structure between locations. Every d-dimensional neighbourhood of a random field Z can be imagined as d-variate distribution. The bivariate building blocks of the vine cop-

ula describing these neighbourhoods need to capture changing correlations across space and time (assuming an isotropic and stationary spatial random field for now), in order to flexibly adapt to the different arrangements of neighbourhoods. Therefore, we introduce bivariate spatial/spatio-temporal copulas that are convex combinations of common bivariate copula families parametrized by distance. These blocks are then used in the vine copula leading to distance-aware spatial/spatio-temporal vine copulas. This new concept allows for a very flexible modelling of strength and shape of the dependence structure between locations in space and time. Adding marginals to this copula yields a local probabilistic model of the random field that allows for prediction at unobserved locations surrounded by confidence intervals and simulation of the random field.

1.5 OUTLINE

The temporal application of vine copulas to an annual rainstorm maxima time series and the implications of different definitions of multivariate return periods is described in the following chapter. A first step towards spatial vine copulas has been presented at the Spatial Statistics Conference 2011 and the corresponding extended abstract can be found in Chapter 3. This single spatial tree vine copula has been extended to a single tree spatio-temporal vine copula as described in Chapter 4. The manuscript enclosed in Chapter 5 describes a single tree spatio-temporal vine copula that additionally allows for a co-variate to model spatio-temporal random fields. The extension of the single tree to the multiple tree spatial vine copulas leads to an improvement of the modelling capabilities as presented in the paper corresponding to Chapter 6. A synthesis and discussion of the research conducted in this thesis is given in Chapter 7 and conclusions are drawn in Chapter 8.

2

COPULAS FOR MODELLING EXTREMES IN TIME SERIES

This chapter consists of the work published with the title *Multivariate return periods in hydrology: a critical and practical review focusing on synthetic design hydrograph estimation* by Gräler et al. [34]. Some references that have been *in press* at the publication of the original paper have been updated in this thesis. My core contributions to this paper are the description of the copula related methodology and its implementation. Further contributions of mine are major parts of the discussion of the results and of the conclusions.

ABSTRACT

Most of the hydrological and hydraulic studies refer to the notion of a return period to quantify design variables. When dealing with multiple design variables, the well-known univariate statistical analysis is no longer satisfactory and several issues challenge the practitioner. How should one incorporate the dependence between variables? How should a multivariate return period be defined and applied in order to yield a proper design event? In this study, an overview of the state-of-the-art for estimating multivariate design events is given and the different approaches are compared. The construction of multivariate distribution functions is done through the use of copulas, given their practicality in multivariate frequency analyses and their ability to model numerous types of dependence structures in a flexible way. A synthetic case study is used to generate a large data set of simulated discharges that is used for illustrating the effect of different modelling choices on the design events. Based on different uni- and multivariate approaches, the design hydrograph characteristics of a three-dimensional phenomenon composed of annual maximum peak discharge, its volume and duration are derived. These approaches are based on regression analysis, bivariate conditional distributions, bivariate joint distributions and Kendall distribution functions, highlighting theoretical and practical issues of multivariate frequency analysis. Also an ensemble-based approach is presented. For a given design return period, the approach chosen clearly affects the calculated design event and much attention should be given to the choice of the approach used as this depends on the real world problem at hand.

2.1 INTRODUCTION

A very important objective of hydrological studies is to provide design variables for diverse engineering projects. Recently, there is an increasing interest in, and need for, simultaneously considering multiple design variables, which are likely to be associated with each other. In hydrology and hydraulics, several applications including sewer systems, dams and flood risk mapping require the selection of storm or hydrograph attributes with a predefined return period.

Standard hydrological design approaches are mostly based on well-established univariate frequency analysis methods. Notwithstanding this, approaches to describe hydrological phenomena involving multiple variables have recently been proposed, aiding the practitioners to estimate multivariate return periods. In literature, as will be described later on, several approaches have evolved over the years. However, it is not clear how these compare to each other and which one is appropriate for a given application.

Recent developments in statistical hydrology have shown the great potential of copulas for the construction of multivariate cumulative distribution functions (CDFs) and for carrying out a multivariate frequency analysis [24, 66, 67, 69, 72, 26, 71, 90]. Copulas are functions that combine several univariate marginal cumulative distribution functions into their joint cumulative distribution function. As such, copulas describe the dependence structure between random variables and allow for the calculation of joint probabilities, independently of the marginal behaviour of the involved variables. For more theoretical details, we refer to Sklar [85] and Nelsen [57]. Several studies have been dedicated to the frequency analysis of multivariate hydrological phenomena such as storms and floods, often within the context of design. However, limited applications have been developed with more than two variables [91, 62, 51, 49, 27, 81, 96, 41, 39]. For a complete and continuously updated list of papers about copula applications in hydrology see the website of the International Commission on Statistical Hydrology of International Association of Hydrological Sciences[1].

Multivariate frequency analysis is becoming more and more widespread and several papers provide insight into generalizations of the univariate case and into new definitions of the multivariate return period (see e.g. Salvadori, De Michele, and Durante [71], Salvadori and De Michele [67], and Shiau [84] and Yue and Rasmussen [95]). Since some of the proposed approaches are in contradiction and others are introduced within specific contexts, there exists a need to clarify the definitions provided so far and to highlight their differences. This study is devoted to this issue and compares a set of different

1 Available at `www.stahy.org`.

approaches on a large simulated data set, allowing to illustrate the implications of different modelling choices.

In this paper, the construction of multivariate distribution functions based on vine copulas (also referred to as pair-copulas by Aas et al. [2]) is first briefly introduced (Section 2.2.2) followed by an overview of several approaches commonly used to estimate multivariate design events based upon different definitions of joint return periods (JRP) (Section 2.3). Subsequently, a synthetic case study addressing the selection of a design hydrograph is presented, which will serve as a test case for evaluating the different approaches. Section 2.4 provides all details on the practical context of this case study. Then, in Section 2.5, extreme discharge events are selected and their most important variables such as annual maximum peak discharge, its volume and duration are analysed, as they form the basis of the analysis. Section 2.6 deals with evaluating the performance and differences between the investigated approaches in quantifying design hydrograph characteristics and highlights important issues for practitioners concerned with multivariate frequency analyses in hydrology. Finally, conclusions are drawn in Section 2.7.

2.2 CONSTRUCTING MULTIVARIATE COPULAS

2.2.1 *Choice of construction method*

Most of the copula-based research in hydrology addresses the application of two-dimensional copulas, for which several fitting and evaluation criteria are becoming more and more widespread. In contrast, the use of multi-dimensional copulas remains a more challenging task. Only a few hydrological studies address this issue and almost always face severe (practical) drawbacks of the available high-dimensional copula families. Most work has been done in the trivariate analysis of rainfall [96, 51, 70, 39], floods [81, 27] and droughts [50, 86, 93].

Recently, a flexible construction method for high-dimensional copulas, based on the mixing of (conditional) two-dimensional copulas, has been introduced and has been shown to have a large potential for hydrological applications. In literature, this construction is known as the vine copula (or pair-copula) construction [56, 2, 1, 44]. The underlying theory for the vine copula construction is described in Bedford and Cooke [7, 8]. This construction method originates from work presented by Joe [47] on which also the method of 'conditional mixtures', as applied by De Michele et al. [18], is based. In this paper, the vine copula method will be used to construct the three-dimensional copula for peak discharge Q_p, duration D and volume V_p. The construction and fitting is discussed in the next section.

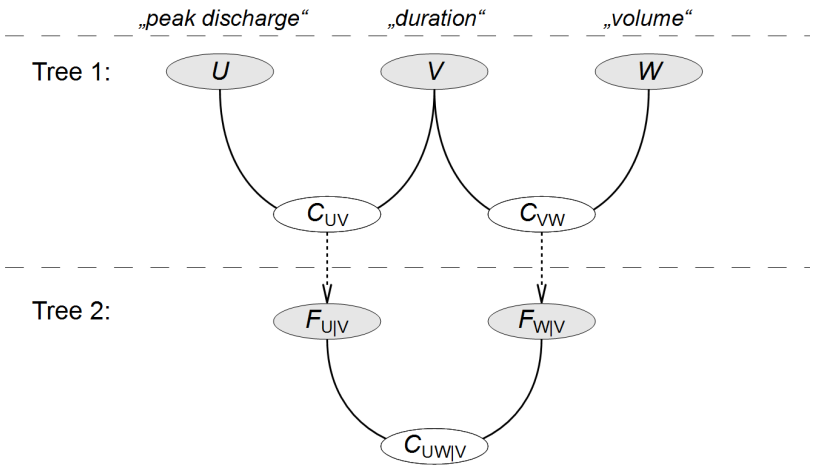

Figure 2.1: Hierarchical nesting of bivariate copulas in the construction of a 3D vine copula.

2.2.2 *Construction of a 3D vine copula*

In this paper, the focus will be on a three-dimensional vine copula joining the three marginal distributions of three random variables X, Y and Z. In general, the approach can be extended to any number of dimensions, although limitations may be introduced by the computational power and available data. In the following, we assume that the samples of all three variables have each been transformed using the following rank-order-transformation S in order to obtain the marginal empirical distribution functions:

$$S(x) := \frac{\text{rank}(x)}{n+1},$$

where n denotes the number of observations for the given variable. We denote the transformed variables by U, V and W so that all three variables are now approximately uniformly distributed on $[0, 1]$.

The basic idea of vine copulas is to construct high-dimensional copulas based on a stagewise mixing of (conditional) bivariate copulas. This corresponds to decomposing the full density function into a product of low-dimensional density functions. At the base of the construction all relevant pairwise dependences are modelled with bivariate copulas. If all mutual dependences are with respect to the same variable, the construction is called a canonical vine (C-vine). If all mutual dependences are considered one after the other, *i.e.* the first with the second one, the third with the fourth one, etc., this is called a D-vine. C- and D-vines are special cases of regular vines, the latter being all possible pairwise decompositions. In the three-dimensional case there is no difference between a C- or a D-vine, only the ordering of variables can be changed.

Figure 2.1 illustrates the construction of a three-dimensional vine copula. In the first tree, three variables U, V, W are given, and their pairwise dependences are captured by the bivariate copulas C_{UV} and C_{VW}. These bivariate copulas can be conditioned under the variable V through partial differentiation [2]. This conditioning is indicated by dashed arrows in Figure 2.1 and results in the conditional cumulative distribution functions $F_{U|V}$ and $F_{W|V}$ (see Eq. (2.1)).

$$F_{U|V}(u|v) = \frac{\partial C_{UV}(u,v)}{\partial v} \quad \text{and} \quad F_{W|V}(w|v) = \frac{\partial C_{VW}(v,w)}{\partial v}. \tag{2.1}$$

In the second tree, the conditional CDF values are calculated for all triplets (u,v,w) in the sample. These 'conditioned observations', which are again approximately uniformly distributed on $[0,1]$, are then used to fit another bivariate copula $C_{UW|V}$. The full density function c_{UVW} of the three-dimensional copula is thus given by:

$$\begin{aligned} &c_{UVW}(u,v,w) \\ &= c_{UW|V}\big(F_{U|V}(u|v), F_{W|V}(w|v)\big) \cdot c_{UV}(u,v) \cdot c_{VW}(v,w). \end{aligned} \tag{2.2}$$

It should be noted that the choice of the conditioning variable (*i.e.* V) is not unique and different choices might lead to different results. In general, different vine copula decompositions differently approximate the underlying multivariate distribution [44]. In this paper, the ordering of variables is based on the two bivariate copulas C_{UV} and C_{VW} that fitted best considering the investigated copula families. The bivariate marginal distribution of C_{UW} is only implicitly modelled through the conditional joint distribution.

Thus, in order to derive the building blocks of the three-dimensional copula, three bivariate copulas C_{UV}, C_{VW} and $C_{UW|V}$ need to be fitted. This is done stage-wise and one can choose any of the available methods in literature. Here, each bivariate copula is fitted by means of the maximum likelihood method, considering different copula families. The best fit is determined by the highest log-likelihood value (see Section 2.5.3).

Several goodness-of-fit tests can be considered to validate the fitted bivariate copulas. In this paper, the chosen goodness-of-fit test is the $A7$ approach appearing in Berg [9] and originating from Panchenko [58]. The advantage of this approach is that it estimates the distance between the two multivariate distribution functions without the need of any explicit dimension reduction, *i.e.* it is directly based on a comparison of observed pseudo-observations and simulated pseudo-observations under the null hypothesis. A simulation approach is taken to obtain the distribution of this test statistic under the null hypothesis. The original procedure as proposed by Berg [9] is slightly altered in this paper as the test statistic of the hypothesis is averaged over

the same number of simulations that are conducted during the simulation. A p-value estimate is derived from the fraction of test statistics exceeding this mean test statistic.

Combining the bivariate copulas as in Eq. (2.2) and substituting the marginal distribution functions F_X, F_Y and F_Z yields the three-dimensional distribution function of (X, Y, Z). Let f_X, f_Y and f_Z denote the marginal density functions and define $u := F_X^{-1}(x)$, $v := F_Y^{-1}(y)$ and $w := F_Z^{-1}(z)$. The full density function f_{XYZ} of the distribution for any triplet (x, y, z) is then given by:

$$f_{XYZ}(x, y, z) := c_{UW|V}\big(F_{U|V}(u|v), F_{W|V}(w|v)\big) \cdot c_{UV}(u, v) \cdot c_{VW}(v, w) \cdot f_X(x) \cdot f_Y(y) \cdot f_Z(z).$$

The estimations in this paper have been done using R [64], a free software environment for statistical computing, and the package *spcopula*[2] building on the packages *copula* [55] and *CDVine* [13]. The R-scripts are available upon request from the authors. A demo related to this paper is available in the spcopula package.

2.3 ESTIMATING DESIGN EVENTS: DEFINITIONS AND METHODS

In literature and in practice, several approaches to estimate multivariate design events for a given design return period exist. The following sections provide a short overview of the most popular approaches, focusing on how a multivariate design event for a given return period could be calculated. In the specific case of multivariate joint return periods (JRP), typically a set of possible design events is found. In order to be able to assess the differences among the described approaches, we select the most probable of all possible design events. An ensemble-based design approach, in contrast to a single design event, will also be presented.

It is important to note that we present different classes of approaches: univariate (Sections 2.3.1 and 2.3.2), bivariate (Sections 2.3.3 and 2.3.4) and trivariate approaches (Section 2.3.4). In all cases, multivariate design events are provided, however, in the first case the procedure is based on the concept of a univariate return period while in the second and third case the procedure is based on the concept of a bivariate and trivariate joint return period, respectively. This premise is pivotal since statistically these classes are incomparable due to the different intrinsic nature of the return period concepts. However, it is important to illustrate the differences in design events that stem from these modelling choices.

2 under development, available at r-forge: `http://r-forge.r-project.org/projects/spcopula`

2.3.1 *Design events derived from a regression analysis*

A first approach is based on a univariate frequency analysis (denoted by REG). First, the driving variable X, *i.e.* the variable with a prominent role in the design, is chosen. Then a design return period T_{REG} is fixed, and given the marginal cumulative distribution of the design variable $F_X(x)$ the corresponding design quantile x_{REG} (equal to the design quantile of the univariate approach x_{UNI}) is sought, based on Eq. (2.3), with μ_T the mean interarrival time [years]. In the case of annual maxima, μ_T equals 1 year. Then, based on a linear regression of X with the other design variable Y, the second design value y_{REG} is obtained. This approach has been applied, among others, by Serinaldi and Grimaldi [82]:

$$T_{REG} = \frac{\mu_T}{1 - F_X(x_{REG})} \quad \Leftrightarrow \quad x_{REG} = F_X^{-1}\left(1 - \frac{\mu_T}{T_{REG}}\right) \tag{2.3}$$

and some regression function f_{REG} modelling Y in terms of X. Thus, $y_{REG} := f_{REG}(x_{REG})$ is the predicted value based on the regression model for a given quantile x_{REG} of the independent variable X. As previously mentioned, this approach does not provide an estimate following a joint return period definition. The motivation behind this approach is to provide a simple, but statistically sound method when one can select a dominant driving variable in the practical application and only a small data set is available hindering a deeper analysis.

2.3.2 *Design events derived from a bivariate conditional distribution*

A second approach (denoted by MAR) consists of conditioning the bivariate cumulative distribution function (CDF) $F_{XY}(x, y)$ on the univariate marginal design quantile $x_{MAR} = x_{UNI}$ corresponding to the chosen univariate design return period T_{UNI}. The resulting (univariate) conditional CDF $F_{Y|X}(y|x = x_{UNI})$ can then be used to calculate the value y_{MAR} for the conditional univariate design return period T_{MAR}.

Advantage will be taken of the bivariate copula $C_{XY}(x, y)$ to perform the calculation. With $u_{MAR} = F_X(x_{MAR})$ and $v_{MAR} = F_Y(y_{MAR})$ the procedure can be expressed as follows. We can rewrite the initial definition

$$T_{MAR} = \frac{\mu_T}{1 - F_{Y|X}(y|x = x_{MAR})}$$

in terms of a copula with $U := F_X(X)$ and $V := F_Y(Y)$ as

$$\begin{aligned} T_{MAR} &= \frac{\mu_T}{1 - \frac{\partial C_{UV}(u, v_{MAR})}{\partial u}\Big|_{u_{MAR} := 1 - \frac{\mu_T}{T_{UNI}}}} \\ &= \frac{\mu_T}{1 - C_{V|U=u_{MAR}}(v_{MAR})} \\ \Leftrightarrow \quad v_{MAR} &= C^{-1}_{V|U=u_{MAR}}(1 - \frac{\mu_T}{T_{MAR}}). \end{aligned}$$

Inverse transformation yields:

$$y_{MAR} = F_Y^{-1}(v_{MAR})$$

It should be noted that this approach does not result in a real bivariate design event having a joint return period in the strict sense as well as the afore described regression based approach. The bivariate distribution is conditioned for the quantile of interest to the practitioner (corresponding with a univariate return period). This conditioned distribution is then used to obtain the other quantile, again based on the principles of a univariate return period. Therefore, the two obtained design quantiles x_{MAR} and y_{MAR} should not be considered as a real joint design event. Furthermore, one should keep in mind that the regression approach predicts the expected value for Y given a certain quantile of X, while the conditional approach estimates the quantile of Y conditioned under the quantile of X. Thus, both approaches cannot directly be compared from a probabilistic point of view but are commonly found in literature and are therefore included.

2.3.3 *Design events derived from a bivariate joint distribution*

Instead of using a conditional CDF, a widely used approach to calculate a bivariate return period can be followed which exploits the full bivariate CDF $F_{XY}(x, y)$. This can easily be expressed by means of a bivariate copula $C_{UV}(u, v)$ with $U := F_X(X)$ and $V := F_Y(Y)$ as before. We refer to this approach as OR as it corresponds to the probability of $P[X > x \vee Y > y]$ following the notation introduced by Vandenberghe et al. [90]:

$$\begin{aligned} T_{OR} &= \frac{\mu_T}{1 - F_{XY}(x_{OR}, y_{OR})} \\ &= \frac{\mu_T}{1 - C_{UV}\big(F_X(x_{OR}), F_Y(y_{OR})\big)} \\ &= \frac{\mu_T}{1 - C_{UV}(u_{OR}, v_{OR})}. \end{aligned}$$

This approach is in fact an intuitive extension of the definition of a univariate return period. All couples (u_{OR}, v_{OR}) that are at the same probability level $t_{OR} = C_{UV}(u_{OR}, v_{OR})$ of the copula will have the

same bivariate return period T_{OR}. For a given design return period, the corresponding level t_{OR} can easily be calculated, the most likely design point (u_{OR}, v_{OR}) of all possible events at this level can be obtained by selecting the point with the largest joint probability density:

$$(u_{OR}, v_{OR}) = \underset{C_{UV}(u,v)=t_{OR}}{\operatorname{argmax}} f_{XY}\left(F_X^{-1}(u), F_Y^{-1}(v)\right). \tag{2.4}$$

The corresponding design values x_{OR} and y_{OR} are easily calculated through the inverse CDFs:

$$x_{OR} = F_X^{-1}(u_{OR}) \quad \text{and} \quad y_{OR} = F_Y^{-1}(v_{OR}).$$

Once the joint density along the level curve is derived, one may consider different alternative approaches. Instead of the most likely event, one may calculate the expected value of the conditional distribution or calculate quantiles for given probabilities that might lead to a design approach incorporating more than a single design event. To limit the number of approaches, we will focus on the most-likely event only, as e.g. used by Salvadori and De Michele [68].

2.3.4 *Design events derived from a copula's Kendall distribution function*

Another definition of the bivariate return period is given by Salvadori and De Michele [67] and Salvadori [66] and Salvadori [72]. Recently, the concept of this bivariate secondary return period was extended to a complete multidimensional setting by Salvadori, De Michele, and Durante [71], called 'Kendall return period' (denoted by KEN). This return period corresponds to the mean interarrival time of events more critical than the design event, the so-called 'super-critical' or 'dangerous' events. The super-critical events are potential threats to the structure and will appear more rarely than the given design return period. This partitioning of the probability distribution into a super-critical and non-critical region is based on the Kendall distribution function K_C. This function is a univariate representation of multivariate information as it is the CDF of the copula's level curves: $K_C(t) = \mathbb{P}\{C(u,v) \leqslant t\}$. It allows for the calculation of the probability that a random point (u, v) in the unit square has a smaller (or larger) copula value than a given critical probability level t_{KEN}. The ability of the Kendall function to project a multidimensional distribution to a univariate one is similarly exploited by Kao and Govindaraju [50] in the context of a joint deficit index for droughts.

The use of the Kendall distribution function to define the probability measure for calculating a JRP is advocated by Salvadori, De Michele, and Durante [71] as it is a theoretically sound multivariate approach sharing the notion of a critical layer, defined through the

cumulative distribution function, with the univariate approach. The definition of the return period in both the univariate and in the multivariate Kendall approach is characterized by making a distinction between super-critical and non-critical events based on a critical cumulative probability level. The only way to extend this to a multivariate context is by using the Kendall distribution function. Probability measures that are constructed differently always entail events that will have a joint cumulative distribution function value that is larger or smaller than the critical probability level, and thus fail in subdividing the space between super-critical and non-critical events with respect to the joint cumulative distribution function. Following this avenue, any critical probability level t_{KEN} uniquely corresponds to a subdivision of the space into super-critical and non-critical regions. This is different from the OR-case mentioned before, where in general different choices of critical events from the same critical probability level t_{OR} subdivide the space differently. From a return period point of view, the copula approach refers to super-critical events where at least one of the margins is larger than the design event, but the joint cumulative probability may be lower than the designated level yielding a shorter return period. On the other hand, the Kendall-based approach ensures that all super-critical events have a longer return period than the limit value, while some non-critical events might have larger marginal values than any selected design event.

For any given copula of any dimension, the Kendall distribution function can be calculated either analytically (e.g. for Archimedean copulas) or estimated numerically, and can thus be used to calculate the Kendall joint return period. Until now, only a very limited number of studies actually applied this kind of return period (e.g. Vandenberghe et al. [91]). In the following sections, the procedure for the two- and three-dimensional cases is outlined.

Two-dimensional Kendall joint return period

After choosing the design return period T_{KEN2}, the corresponding probability level t_{KEN2} of the copula can be calculated by means of the inverse of the two-dimensional Kendall distribution function (Eq. (2.5)). In 2D, this corresponds to finding an isoline on the copula.

$$\begin{aligned} & T_{KEN2} = \frac{\mu_T}{1 - K_C(t_{KEN2})} \\ \Leftrightarrow \quad & K_C(t_{KEN2}) = 1 - \frac{\mu_T}{T_{KEN2}} \\ \Leftrightarrow \quad & t_{KEN2} = K_C^{-1}\left(1 - \frac{\mu_T}{T_{KEN2}}\right) \end{aligned} \tag{2.5}$$

When no analytical expression for K_C is available, the inverse can be calculated numerically based on an extensive simulation algorithm,

described in [71]. Once t_{KEN2} is known, the most likely design event in the unit square (u_{KEN2}, v_{KEN2}) is selected on the corresponding isoline in an analogous way as described by Eq. (2.4). Through the use of the inverse of the marginal CDFs the corresponding design event (x_{KEN2}, y_{KEN2}) is found.

Three-dimensional Kendall joint return period

In three dimensions, the corresponding probability level t_{KEN3} should be found again in the same way as in Eq. (2.5). To calculate the inverse of the function K_C, one might need to rely on a numerical method as for instance described by Salvadori, De Michele, and Durante [71]. However, in contrast to the two-dimensional case, the probability level t_{KEN3} corresponds to an isosurface, *i.e.* all triplets (u, v, w) on this surface have the same copula value t_{KEN3}. Generally, for a n-dimensional copula a isohypersurface of dimension n-1 exists that contains all n-dimensional points with the same copula level t_{KENn}. A single design event $(u_{KEN3}, v_{KEN3}, w_{KEN3})$ should again be selected on this isosurface. Therefore the point $(u_{KEN3}, v_{KEN3}, w_{KEN3})$ with the highest joint likelihood is selected yielding the most likely event. In fact this is the three-dimensional extension of the approach given in Eq. (2.4), *i.e.*:

$$\begin{aligned}&(u_{KEN3}, v_{KEN3}, w_{KEN3})\\ &= \underset{C_{UVW}(u,v,w)=t_{KEN3}}{\operatorname{argmax}} f_{XYZ}\left(F_X^{-1}(u), F_Y^{-1}(v), F_Z^{-1}(w)\right).\end{aligned} \tag{2.6}$$

2.3.5 *Theoretical comparison of JRP definitions*

The above defined JRPs (T_{OR} and T_{KEN}) do not provide answers to the same problem statement. Therefore, one has to carefully consider the practical implications of the selected approach on the probability of interest. Vandenberghe et al. [90] mentioned the inequality $T_{OR} \leqslant T_{AND}$ which can be extended to:

$$T_{OR} \leqslant T_{KEN} \leqslant T_{AND} \tag{2.7}$$

where T_{AND} refers to the exceedance probability of $P[X > x \wedge Y > y]$. The OR, AND and KEN JRPs can, in terms of 2-dimensional copulas, be graphically interpreted on the unit square. The different return periods T_{OR}, T_{KEN2} and T_{AND} for a fixed design event (u, v) can then, in every case, be expressed by $1/\big(1 - \text{area}(\text{safe events})\big)$. This is shown in Figure 2.2 where the areas represented by the different approaches for a given design event (u, v) are indicated alongside with the copula level curve $C(u, v)$. It can be seen that the OR definition only declares all events in the lower-left rectangle as safe. The KEN approach declares the top-left and lower-right curved areas ("KEN") as safe as well, and they are added to the lower-left rectangle yielding a larger

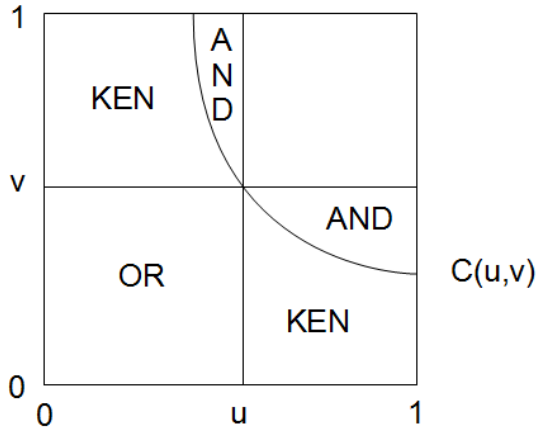

Figure 2.2: Graphical representation of the different JRP definitions in terms of a copula (2-dimensional case).

return period for the same design event (u, v). Lastly, the AND case adds the top-left and lower-right rectangles, resulting in the largest return period. Note that these inequalities hold only within the same dimensionality of a problem.

2.3.6 *Ensembles of design events*

From Sections 2.3.3 and 2.3.4 it should be clear that for a design event characterized by several variables, one has to select an event out of a range of events which all share the same JRP. The selection of merely one event sensibly reduces the amount of information that can be obtained by the multivariate approach chosen. Volpi and Fiori [92] present an approach to select a subset of the critical level to reflect the variability within the set of critical events. We follow a similar path and define a conditional distribution along the level curve to obtain a sample of the possible design events. The importance of an ensemble-based approach has already been stressed by Salvadori, De Michele, and Durante [71]. Vandenberghe et al. [91] provided a first attempt to benefit from the richness of an ensemble of critical values in a practical context.

Consider first the bivariate case, in which the JRP approaches based on copulas (OR, Section 2.3.3) and based on the Kendall distribution function (KEN2, Section 2.3.4) result in the finding of a contour level t_{OR} and t_{KEN2} on which all pairs (u, v) have the same respective JRP. Instead of using Eq. (2.4) to select the most likely point, the function f_{XY} over the t-isoline could be used as a univariate weight function out of which an ensemble of pairs could be sampled. In general, a rescaling is necessary to ensure that f_{XY} integrates to 1 and yields a probability density function (PDF) [71]. Generally, not all pairs (u, v) on the t-isoline have the same likelihood, *i.e.* pairs on the edges are

less likely than pairs closer to the centre of the isoline. In this way, sampling according to f_{XY} makes more sense from a practical point of view than uniformly sampling over the isoline (as done by Vandenberghe et al. [91]).

Eventually, one will end up with an ensemble of (u_i, v_i)-pairs (with i ranging from 1 to N, the ensemble size). By means of the inverse marginal CDFs, these pairs are easily transformed to real values. This ensemble could then be used to run simulations from which the variability of specific design variables (e.g. thickness or height of a dam) can be assessed. This approach needs additional analysis as it will yield several design vectors going beyond the standard notion of a single design event. As an example, one could route an ensemble of 1000 pairs of peak discharge and volume through a dam model and consider the water height in the reservoir. Using just one design event, only one water height is obtained. However, using the ensemble, information on the range and likelihood of possible water heights for the given design return period is obtained, making it possible to incorporate the variability within the design variables stemming from multiple design events along the critical level.

In the trivariate case (see Section 2.3.4) no isoline is obtained but an isosurface. Similar to the two-dimensional case, the full weight function over this isosurface could be rescaled to a bivariate probability density function out of which an ensemble of triplets could be sampled. The higher the dimensionality of the design problem, the more advantageous the ensemble approach becomes: in three dimensions more information is lost than in two dimensions by selecting just one design event. The drawback of the ensemble approach is the increasing need for run time when higher dimensions are considered.

2.4 DIFFERENCES AMONG MULTIVARIATE DESIGN EVENTS IN THE SYNTHETIC DESIGN HYDROGRAPH APPLICATION

2.4.1 *Experimental set-up*

In order to illustrate differences among estimated design events by the approaches described in the previous sections, a simulation experiment is set up and analysed with respect to the synthetic design hydrograph (SDH) attributes. The SDH is defined as a hydrograph with an assigned return period (uni- or multivariate), which can be characterized by random variables such as the flood peak Q_p, the volume V_p and the duration D. Specifically, given an observed or simulated runoff time series from which a set of extreme hydrographs is selected, one can determine the SDH shape in several ways (see Serinaldi and Grimaldi [82] and references therein). In a two-dimensional set-up, two hydrograph parameters (peak-volume, peak-duration or volume-duration) should be fixed, while the third one is

obtained from the chosen hydrograph shape distribution. In a three-dimensional set-up, the three characteristic parameters are obtained jointly.

In most common hydrological applications the interest is in the flood peak (Q_p) and volume (V_p). Consequently, the two-dimensional analyses in this paper focus on these variables. However, as described in Section 2.3, there are several approaches that lead to the design values for Q_p and V_p, including a three-dimensional approach. Applying the proposed approaches to the same data set allows to compare the different underlying definitions and implications of the model selection. However, in a practical context one is typically tied to a specific frequency analysis that corresponds to the unique design characteristics.

The case study proposed in this paper consists of applying a continuous simulation model on a small, ungauged basin for which 500 years of synthetic direct runoff time series at a 5 min resolution are simulated. From this series, the 500 maximum annual peaks are selected together with their corresponding hydrograph (identified as the continuous sequence of non-zero direct discharge values including the annual peak). Note that as direct discharge is considered, a zero discharge value does not imply a dry river. Consequently, 500 (Q_p, D, V_p) triplets are available to which the described approaches estimating design events are applied. By considering a real case study, the obtained differences and hence the implications of a modelling choice can be evaluated in a practical context. In order to simulate the 500 years runoff time series, the COSMO4SUB model, described in the following section, is applied.

2.4.2 *The COSMO4SUB framework*

The synthetic data set on which the previously described approaches are applied is obtained through the use of the COSMO4SUB framework [37, 38]. COSMO4SUB is a continuous model which allows the simulation of synthetic direct runoff time series using minimal input information from rainfall data and digital terrain support. Specifically, the watershed digital elevation model (DEM) with a standard resolution used in hydrological modelling, the soil use and type, daily (preferably at least 30 years long) and sub-daily (preferably at least 5 years long) rainfall observations are the only data necessary to run the model. COSMO4SUB includes three modules: a rainfall time series simulator, a rainfall excess scheme and a geomorphological rainfall-runoff model. Next, the general principles are explained and in Section 2.5.1 specific details of the calibration are presented.

The first module is based on a single-site copula-based daily rainfall generator [79] and on the continuous-in-scale universal multifractal model [76] for disaggregating the daily rainfall to the desired time

scale (up to 5 min). The parameters included in this first module (six for each month for the daily rainfall simulator and three for the disaggregation model) are calibrated on the basis of the available rainfall observations (at two different scales).

The second module is related to the rainfall excess step. A new mixed Green Ampt-Curve Number (CN4GA Curve Number for Green Ampt) procedure was recently proposed [36] and included in the present version of the COSMO4SUB framework. The key concept is to use the initial abstraction (*i.e.* all the losses due to initial saturation, filling terrain gaps, interception, etc.) and the total SCS-CN excess rainfall volume to estimate the effective saturated hydraulic conductivity and the ponding time of the Green-Ampt model. Consequently, the CN4GA approach tries to appropriately distribute the volume estimated by the SCS-CN method over time. This module is characterized by five parameters (specified in Section 2.5.1) which are empirically assigned using the soil use and soil type map information. In addition, the event separation time (T_s) is included in this module since the continuous implementation of the SCS-CN method requires to fix a no-rain time interval for which the cumulative gross and excess precipitation can be reset to zero. As shown in Grimaldi, Petroselli, and Serinaldi [37, 38], this parameter has a limited influence on the final results and the value can be arbitrarily assigned in the range of 12-36 hours.

The third module allows to carry out a continuous convolution of the rainfall excess for obtaining the direct runoff time series through an advanced version of the Width Function Instantaneous Unit Hydrograph (WFIUH). The adopted model, named WFIUH-1par [40, 35], identifies the watershed IUH through the topographic information and needs only one parameter that can be quantified referring to the watershed concentration time (T_c), estimated using empirical equations. Following the application of the three described modules a continuous runoff scenario is obtained from which maximum annual hydrographs in terms of their peak discharge are selected. It is important to note that the variables duration and volume in the selected triplets do not necessarily reflect annual maxima.

2.5 DATA AND MATERIALS

This study is based on simulated data and a statistical model is fitted to this data set. This way, a data set of sufficient size to compare the various approaches presented in this paper is obtained.

2.5.1 *Model set-up*

In order to provide a realistic scenario that can be used to evaluate the previously described approaches, the COSMO4SUB model was

applied on the Torbido River, a small tributary of the Tiber River located in central Italy (watershed area: 61.67 km^2). Basin elevations range from 85 m to 625 m, the average slope is 22% and the maximum distance between divide and outlet is 25.8 km. The watershed DEM at a 20 m spatial resolution was provided by the Italian Geographic Military Institute [46], while land cover was extracted from the CORINE database [22].

Observed rainfall data, useful for calibrating the two-stage rainfall simulator parameters, are available from the Castel Cellesi rain gauge station for a period of 49 years at a daily time scale and for a period of 10 years at a 5-min resolution [79, 78]. For a description and evaluation of the 500-years rainfall synthetic time series, we refer to Grimaldi, Petroselli, and Serinaldi [38, 37].

2.5.2 *Annual extreme discharge events*

Once the 500-years synthetic direct runoff time series is determined, as described in Section 2.4.1, the 500 maximum annual peak discharge events are selected and characterized by their peak discharge Q_p, duration D and volume V_p. For only six years the model provides a zero direct runoff, which is reasonable considering the limited size of the watershed. These values are excluded in the following analyses.

All approaches rely on the marginal distribution functions of Q_p, D and V_p that need to be fitted in the first place. As the peak discharge variable consists of annual extreme values selected from the simulated 500 year discharge series and the other two variables are closely correlated (but not necessarily annual maxima), the fit of several extreme value distributions is considered, *i.e.* the exponential, the Weibull and the Generalised Extreme Value (GEV) distribution functions. These distributions are, respectively, a one, two and three parameter distribution, allowing for various degrees of model complexity. Furthermore, the GEV distribution generally encompasses three different distributions, namely the Fréchet, the (reversed) Weibull and Gumbel distributions either directly, or through a transformation as in the case of the Weibull distribution which corresponds to a reversed Weibull distribution. These different distribution types each represent a different kind of tail behaviour, namely a light tail (Gumbel), a heavy tail (Fréchet) and a bounded upper tail (Weibull). These behaviours can be separated based on the shape parameter ξ of the GEV. Furthermore, the Weibull distribution is fitted separately as well, as it only corresponds to a GEV distribution after transformation. Finally, most extreme value distributions are of the exponential type, and cannot deal with an offset, *i.e.* when the smallest value of the variable in the CDF is larger than zero. However, as a result of censoring the zeros, the smallest value of the variables tends to be significantly higher than zero, leading to poor fits of the CDF. Therefore, a location

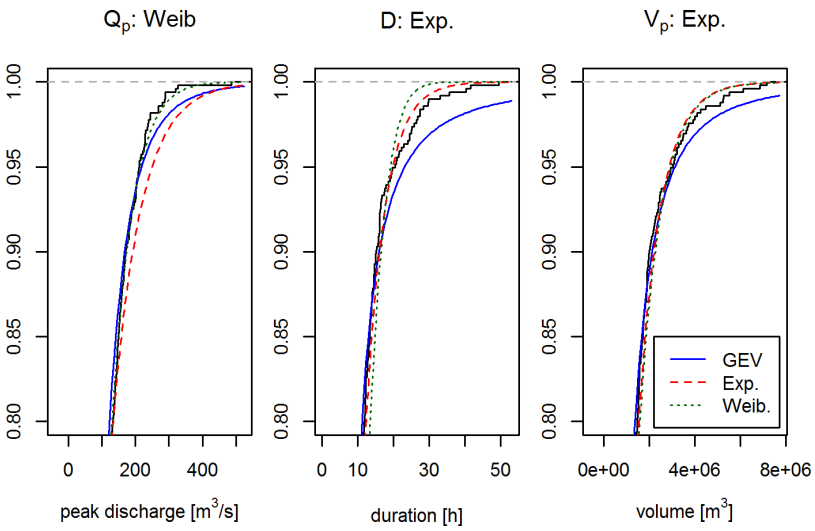

Figure 2.3: The various cumulative distribution functions together with the empirical cumulative distribution function for the three variables. The best fitting distribution is denoted in the title of each graph.

parameter has been introduced in the distributions to ensure a proper fit in the tails.

Table 2.1: The values of the AIC for the various distributions of the respective variables.

	GEV	exponential	Weibull
Q_p	5370	5360	5326
D	2610	2646	2928
V_p	14641	14599	14601

A first test to ascertain the appropriate distribution for the three marginal variables is to display the empirical CDFs together with the directly fitted distribution. This is shown in Figure 2.3, in which only the upper tail of the CDF is shown, *i.e.* the interval $[0.80, 1]$ as the focus is on the extremes. It can immediately be seen that not all the distributions fit these tails equally well. This is corroborated by the Akaike Information Criterion (AIC) computed for all different models, shown in Table 2.1, as well as the log-likelihood of each model (not shown). Based on these criteria and Figure 2.3, we select the Weibull distribution for Q_p and the exponential distribution for V_p. Seemingly, the GEV provides the overall best fit according to the AIC, despite the poor representation of the upper tail (see Figure 2.3). As the focal point of this study is set around a return period of ten years addressing the top 10 % of the CDF, we chose to select the exponential distribution because of its better fit in this region. More in-depth testing through Q-Q plots (not shown here) indicates that this is indeed a better approximation of the distribution. Further investigation

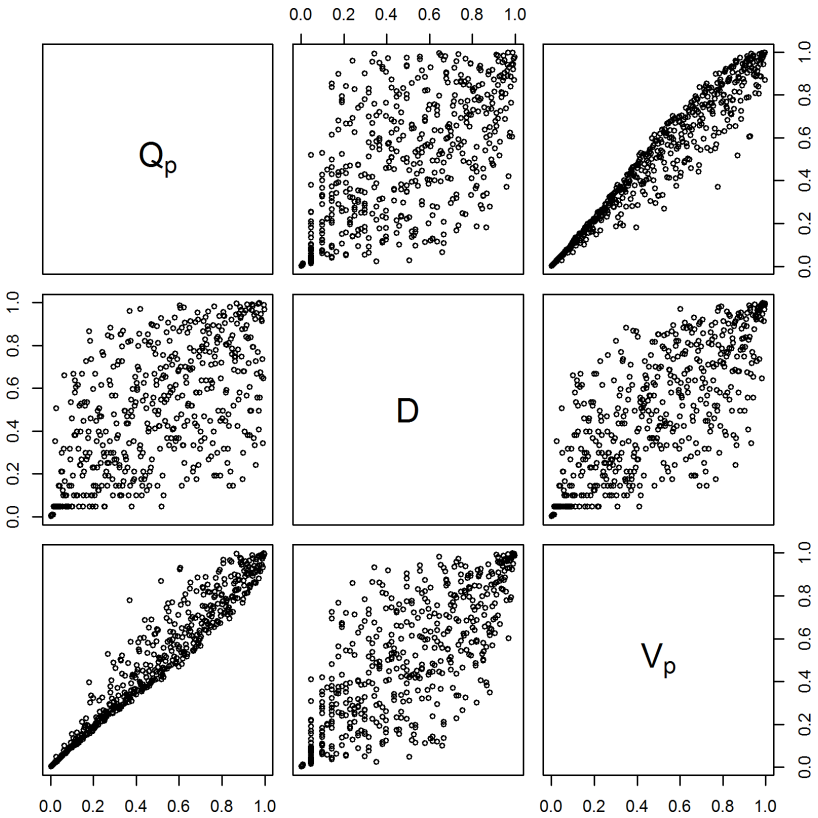

Figure 2.4: Normalized rank scatter plots for all pairs of variables. Kendall's tau is 0.85 for (Q_p, V_p), 0.42 for (Q_p, D) and 0.54 for (V_p, D).

of additional distribution families and combinations of these might improve the fit of the marginals, but is out of scope of this paper. Nevertheless, a good fit of the marginal distributions is key to the practical application. Hence, the following models are selected:

- Q_p: Weibull (Anderson-Darling $p = 0.59$),
- D: Exponential (Anderson-Darling $p = 0.08$),
- V_p: Exponential (Anderson-Darling $p = 0.18$).

Here, the Anderson-Darling test was used to determine whether the samples were significantly different from the fitted distributions. It should be understood that a consistency in marginal distribution functions across the different approaches is far more important for comparison reasons than a perfect fit, considering the underlying data are simulations.

To analyse the association between the variables, which will be modelled by means of copulas, Kendall's tau is calculated and normalized rank scatterplots are evaluated for each pair of variables (Figure 2.4). Evidently, there are strong positive associations. Also, some ties are present, especially for D, which have been assigned with their mean rank in the transformation. The next section deals with the modelling of these associations.

2.5.3 *Fitting of the 2D and 3D copulas*

As described in Section 2.2.2, we used maximum likelihood estimation to fit a copula from each investigated family for every pair of variables and selected the best fitting one by the highest log-likelihood value. The copula families investigated include Normal, Student, Gumbel, Frank, Clayton, BB1, BB6, BB7, BB8 and the survival copulas of the 4 latter ones (details on all these families can be found in Nelsen [57] and Joe [47]). Table 2.2 gives an overview of the parameters and goodness-of-fit results. The p-values are estimated from 1000 iterations each.

Table 2.2: An overview of the fitted bivariate copulas in the 2D copula-based and 3D vine copula-based approach.

	pairs of var.	id	τ_K [-]	copula fam.	param.		p-val.
2D	$Q_p \sim V_p$	13	0.85	BB7	2.24	14.10	0.69
3D	$Q_p \sim D$	12	0.42	surv. BB7	2.05	0.35	0.74
	$D \sim V_p$	23	0.54	surv. BB7	2.25	1.09	0.75
	$(Q_p \sim V_p)\vert D$	13|2	0.83	Student	0.96	2.00	0.66

The following approaches in the two-dimensional case make use of the fitted BB7 copula C_{13} which models the dependence between Q_p and V_p. It should be noted that this copula is not able to represent the boundary effect present in the rank-scatter plot (Figure 2.4). To the authors' knowledge, no copula is available in literature that would be able to model such a boundary effect. As the BB7 copula family belongs to the class of Archimedean copulas, its Kendall distribution function can easily be obtained analytically.

For the three-dimensional case, the three fitted bivariate copulas C_{12}, C_{23} and $C_{13|2}$ are then composed into the three-dimensional vine copula as given in Eq. (2.2). For comparison purposes, three-dimensional copula fits for the three parameter Gaussian copula and the one parameter Clayton, Frank and Gumbel copulas have also been performed. The log-likelihood shows a 10% increase for the fitted vine copula (1047) with respect to the Gaussian one (935), while the three one-parameter Archimedean copulas have far smaller values (432 – 532). Thus, the vine copula yields the best fit within this set of copula families in terms of the log-likelihood. As no closed form exists for the cumulative distribution function of this vine copula, a numerical evaluation based on a sample of 100.000 points was carried out in order to be able to calculate the (inverse of the) Kendall distribution function.

However, the singularity appearing in Figure 2.4 for the pair (Q_p, V_p) is neglected in both, the bivariate and the vine copula (as well as in the other considered copulas). In the bivariate case, no copula family with such a limited support could be found while the vine

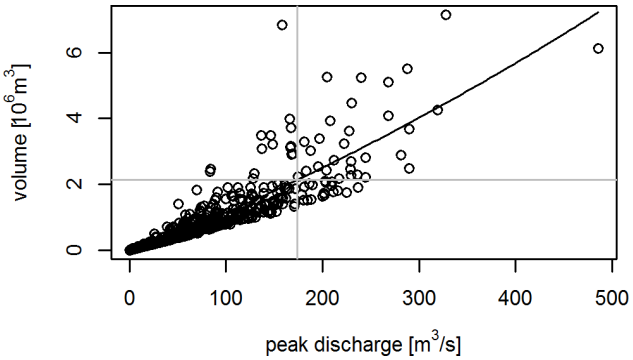

Figure 2.5: Illustration of the derivation of the design quantiles based on the regression approach.

copula's decomposition is based on the bivariate copulas C_{12} and C_{23} addressing the pair (Q_p, V_p) only through the conditional joint distribution. Thus, all investigated copula families would in general sample unrealistic point pairs (Q_p, V_p) beyond the border appearing in the scatter plot. A discussion on this singularity and its underlying process can be found in Serinaldi [80].

2.6 RESULTS AND DISCUSSION

2.6.1 *Calculation of single design events*

In this section, the design values for the SDH with a design return period of 10 years are calculated based on the 2D and 3D approaches presented in Section 2.3. The triplet (Q_p, D, V_p) is considered for which the following transformations hold:

$$U = F_{Q_p}(Q_p),\ V = F_D(D) \text{ and } W = F_{V_p}(V_p)$$

As a reference, the univariate case is analysed first. Based on the inverse of the CDFs F_{Q_p}, F_D and F_{V_p}, at a probability level of

$$1 - \frac{\mu_T}{T_{UNI}} = 1 - \frac{1}{T_{UNI}} = 0.9$$

the design values $q_{p,UNI} = 174\ m^3/s$, $v_{p,UNI} = 2.21 \cdot 10^6 m^3$ and $d_{UNI} = 16.02\ h$ are obtained. In the following, Table 2.3 and Figure 2.6 provide a way to compare these and all further estimated design events. In order to be able to compare design events with the data, the simulated pairs (Q_p, V_p) are visualized as grey dots in Figure 2.6 that summarizes all described approaches.

First the two-dimensional case is considered, in which the focus is on the couple (Q_p, V_p). In the regression-based approach (REG, Section 2.3.1) the starting point is the univariately derived quantile $q_{p,UNI}$, being usually the driving variable in many hydrological applications (see Eq. (2.3)). Based on a regression between Q_p and V_p, as

Table 2.3: Overview of the calculated design event for $T = 10$ years, based on several approaches. The values are rounded to address the limited numerical precision and ease comparison.

Approach	subs.	t [-]	K_C [-]	u_T [-]	v_T [-]	w_T [-]	$q_{p,T}$ [m^3/s]	d_T [h]	$v_{p,T}$ [$10^6 m^3$]
univariate	UNI	×	×	0.9	0.9	0.9	174	16.02	2.21
lin.regr.	REG	×	×	0.9	×	0.892	174	×	2.14
cond. cop.	MAR	0.9	×	0.9	×	0.952	174	×	2.92
copula 2D	OR	0.9	×	0.927	×	0.925	192	×	2.49
K_C-2D	KEN2	0.836	0.9	0.877	×	0.875	161	×	1.99
K_C-3D	KEN3	0.730	0.9	0.844	0.820	0.851	147	12.90	1.83

shown in Figure 2.5, the design volume $v_{p,REG}$ is easily estimated as $2.14 \cdot 10^6 m^3$. This volume is lower than the one obtained by a purely univariate analysis partly due to the different definition based on the expectation instead of a quantile.

The second two-dimensional approach is based on the conditional copula (MAR, Section 2.3.2). The conditioning of the bivariate copula C_{uw} (denoted as C_{13} in Section 2.5.3) for $u_{UNI} = 0.9$ results in the function $C_{w|u}(w, u = 0.9)$. The value of $w_{MAR} = 0.9521$ corresponds with a probability level of 0.9. By means of the inverse $F^{-1}_{V_p}(w_{MAR})$, the design volume $v_{p,MAR}$ is calculated as $2.92 \cdot 10^6 m^3$, which is considerably larger than the former design volumes.

The true joint return period approaches based on the bivariate copula C_{uw} (OR, Section 2.3.3) is the third 2D approach. For $T_{OR} = 10$ years, the corresponding copula level t_{OR} equals 0.9 and corresponds to an isoline. Using the marginal CDFs for Q_p and V_p Eq. (2.4) can be solved to find the point (u_{OR}, w_{OR}) with the highest joint likelihood, *i.e.* $(u_{OR}, w_{OR}) = (0.927, 0.925)$. Using the inverse CDFs the design event is obtained: $(q_{p,OR}, v_{p,OR})$ = (192 m^3/s, $2.49 \cdot 10^6 m^3$). Both the design peak discharge Q_p and the volume V_p are larger than what is obtained in the univariate case.

The last two-dimensional approach is the one in which the JRP is calculated using the Kendall distribution function (KEN2, see Section 2.3.4). Here, the focus is on the inverse of K_C for a probability level of 0.9: $t_{KEN2} = K_C^{-1}(0.9)$. The Kendall distribution function of the bivariate copula C_{uw}, allows to calculate the t_{KEN2}-level corresponding to a cumulative probability of 0.9, *i.e.* $t_{KEN2} = 0.836$. This level is smaller than the one obtained in the former copula-based JRP approach. Again, Eq. (2.4) can be solved to obtain the most likely design event $(u_{KEN2}, w_{KEN2}) = (0.877, 0.875)$. Transformation to the real domain by means of the inverse CDFs results in the design event $(q_{p,KEN2}, v_{p,KEN2})$ = (161 m^3/s, $1.99 \cdot 10^6 m^3$).

Besides the estimation of two-dimensional design events, also one approach for estimating a three-dimensional design event is present-

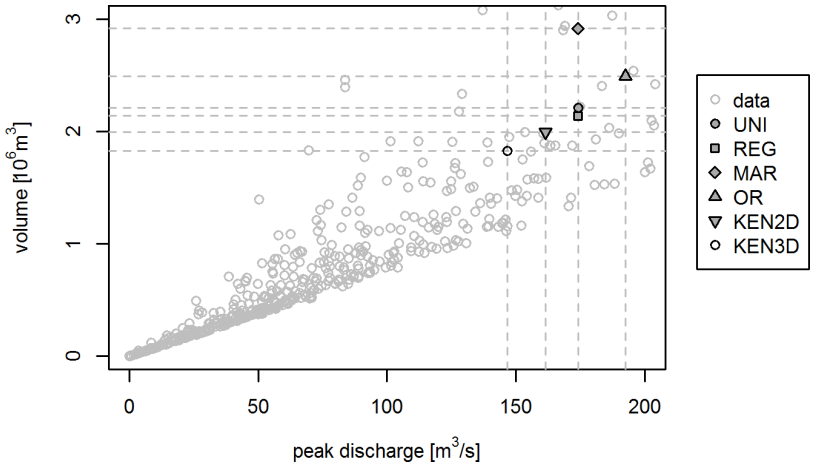

Figure 2.6: An overview of the different design values for a design return period of 10 years obtained with the different definitions. Note that only a subset is shown and the data points exceed both axes.

ed in Section 2.3.4 together with the fitted three-dimensional vine copula (see Section 2.5.3). The three-dimensional vine copula is used for simulating 100 000 triplets (u, v, w) as a basis for the numerical inversion of the Kendall distribution function. Here, the probability level of 0.9 corresponds to a t_{KEN3}-level of 0.730 on the three-dimensional vine copula. In contrast to the two-dimensional approaches, the t_{KEN3}-level corresponds to a surface. Using the marginal CDFs in combination with Eq. (2.6), the most likely point on this surface is found as $(u_{KEN3}, v_{KEN3}, w_{KEN3}) = (0.844, 0.820, 0.851)$. Using the inverse CDFs this results in the design event ($q_{p,KEN3}$, d_{KEN3}, $v_{p,KEN3}$) = (147 m^3/s, 12.90 h, 1.83 $\cdot 10^6 m^3$). Note that the Kendall distribution function is a univariate representation of multivariate information and that its form is different in the two-dimensional and three-dimensional cases.

2.6.2 *Obtaining an ensemble of design events*

The preceding analyses resulted in a single design event, however, as stated in Section 2.3.6 the generation of an ensemble would be preferable. For example, consider the approach where the JRP is based on the Kendall distribution function in the two-dimensional case. The t_{KEN2}-level was found to be 0.836 for a 2D Kendall-based JRP of 10 years (see Table 2.3). Figure 2.7 shows this t_{KEN2}-level and the t_{OR}-level of 0.9, together with the earlier identified most likely design events (u_{KEN2}, w_{KEN2}) and (u_{OR}, w_{OR}) along with a sample of size 500 each. Obviously, along this contour the occurrence of several other events is possible. The sampling across these contours according to the likelihood function results in ensembles of events all having a copula-based and 2D Kendall-based JRP equal to 10 years, respec-

tively. All sampled events clearly lie on a contour, corresponding with the t_{OR}-level and t_{KEN2}-level. According to the grey-scale, the highest density of design events is sampled around the most likely realization whereas less design events are sampled on the two outer limits of each contour.

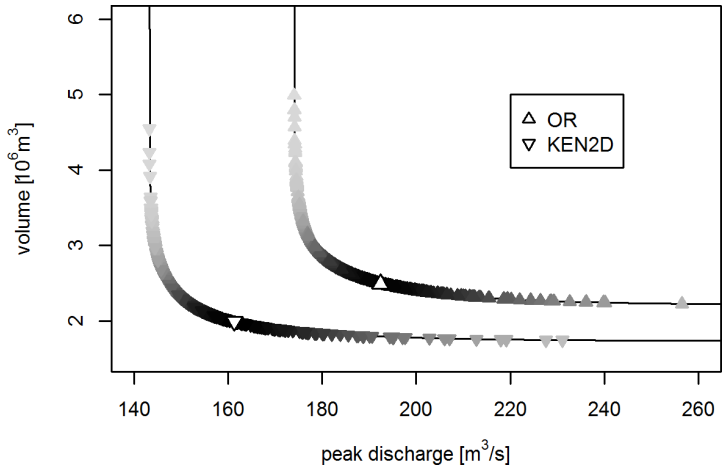

Figure 2.7: An ensemble of 500 (q_p, v_p) pairs that all have a copula-based and 2D Kendall-based JRP of 10 years respectively. The density of the ensemble is given in grey-scale: the brighter the grey, the less events sampled. The most likely event is also indicated.

The density of the ensembles across these contours could be projected (and normalized) on both the Q_p and V_p axis, resulting in univariate PDFs for Q_p and V_p underlying the ensembles. These are shown in Figures 2.8 and 2.9. The most likely design events are naturally situated at the maximum of these PDFs. In general, these conditional distributions do not have to be bounded and extremely large events might possess a positive likelihood.

These PDFs hold a lot of information on the design events. For example, 90% of all design events with a 2D Kendall-based JRP equal to 10 years have a peak discharge in the range of [150 m^3/s, 238 m^3/s] and a volume in [183 $\cdot 10^6 m^3$, 326 $\cdot 10^6 m^3$]. Note from Figure 2.7 that lower volumes occur together with higher peak discharge values and

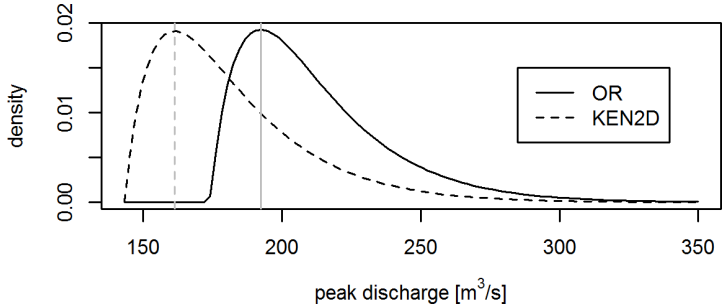

Figure 2.8: PDF of Q_p in the OR and KEN2D ensembles. The most likely design discharge values are indicated by vertical lines.

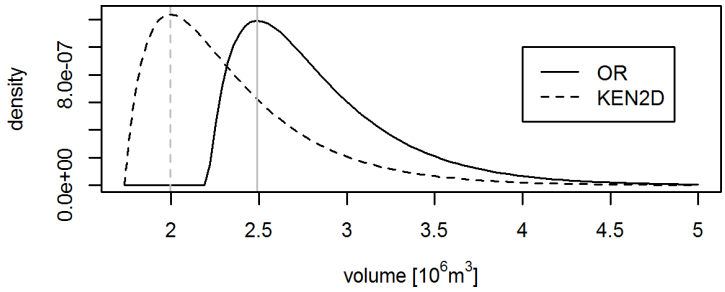

Figure 2.9: PDF of V_p in the OR and KEN2D ensembles. The most likely design volumes are indicated by vertical lines.

vice versa. As briefly mentioned in Section 2.3.6, the ensemble of design events can also be used to calculate another design variable, such as the water height in a reservoir, for which again a PDF of possible design values can easily be obtained. However, this exercise is beyond the scope of this paper.

2.6.3 *Uncertainty in the design event estimation*

In the previous section and Section 2.3.6, we discussed that for a given return period, different design events can be selected due to the multivariate nature of the design problem and that the designer can make use of this variability for parametrising hydraulic structures. From the shown analysis, it is also clear that the presented approaches provide different design event estimates. Yet, these approaches are also prone to uncertainty because of the fact that the copula or the model used to select the design event is fitted to a (small) number of extreme events in an observed time series. Variations in the time series might lead to different model parameters and hence result in alternative design events. The question can thus be posed whether the different approaches generate statistically different design events if one accounts for the uncertainty due to fitting of the probabilistic model. To answer this question, the uncertainty has to be addressed resulting in confidence bands. As no closed form exists, a common approach is to run simulations. In each simulation step, we sampled 494 pairs (the same number as originally observed pairs) from our fitted bivariate distribution and re-estimated the copula and marginal parameters. From the newly obtained probabilistic model, all approaches provided the most likely design event estimate (resulting in the scattered estimates shown as squares and triangles in Figure 2.10). For each approach, the 0.025-quantile and 0.975-quantile design events out of all simulated ones are selected in terms of their return period definition for the null-hypotheses model. These quantiles describe the border of the 95 % confidence band denoted by the corresponding copula t-level curve. The results are summarized in Figure 2.10. As the inner 95 % of points of each cloud do not intersect

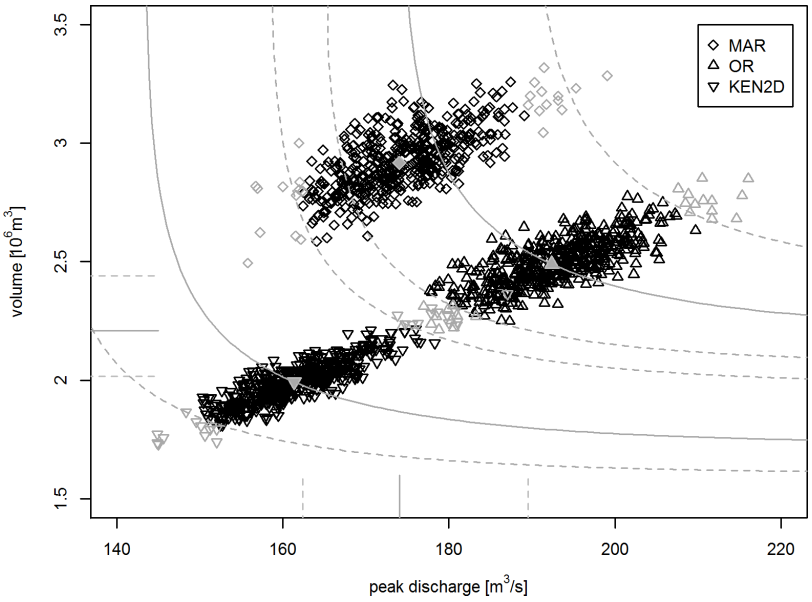

Figure 2.10: 500 simulations of different approaches to obtain a single design event with a design return period of 10 years. Straight line segments indicate the univariate approaches and curved lines represent the bivariate approaches with dashed lines approximating the ensemble confidence band estimates for their corresponding point cloud each. The estimated design events are denoted as filled shapes. The open grey shapes represent the most and least extreme 2.5% for each approach.

between the three approaches, it is evident that the predicted design events are significantly different. However, projecting the multivariate design events to their univariate margins, all confidence intervals intersect except for the univariate predicted volume and the design volume based on the conditional copula (MAR). Addressing the additional uncertainty of the estimates due to the selection of merely a single design event, the dashed curved lines in Figure 2.10 provide an approximation. They limit the region where 95 % of the simulated t-level curves fall. Ensembles of design events would then be drawn along these level curves. Thus, the copula-based (OR) and the Kendall-based (KEN2D) approach provide significantly different design event ensembles.

2.6.4 *Some practical considerations*

Table 2.3, Figure 2.6 and Figure 2.10 clearly demonstrate that the choice of the estimation approach influences the design event values. This evidently is something the practitioner should consider when designing a hydraulic structure, e.g. a dam, based on a specific design hydrograph, as it directly influences the safety and the cost of the structure to be built.

With respect to the univariate design quantile $q_{p,UNI}$, only the approach using the copula-based JRP provides a larger quantile, whereas for the approaches using the Kendall-based JRP a smaller quantile is found. The other approaches use the univariate quantile as a starting point, resulting in identical quantiles. Considering the design quantiles $v_{p,\cdot}$, the MAR and OR approaches yield a larger quantile than the univariate quantile, while the REG and Kendall-based approaches yield smaller values.

The three true JRP approaches use information of the full bivariate (copula-based and 2D Kendall-based approach) or trivariate (3D Kendall-based approach) distribution function. The Kendall-based approaches have the advantage of using a mathematically consistent way of defining the probability of extremes or dangerous events relying on the CDF as in the univariate approach, unlike the JRP approach based on the copula solely. For a full discussion, please refer to Salvadori, De Michele, and Durante [71]. However, there is no universal choice of an appropriate approach to all real world problems. The most important is to address the problem from a probabilistic point of view and to be aware of the practical implications of the approach chosen (outlined in this paper). It is also evident, but not necessarily the case, that the more variables are included (2D vs. 3D), the smaller the design quantiles become. Salvadori and De Michele [68] discuss this *dimensionality paradox* and provide a theoretical explanation for it.

Furthermore, the issue of selecting just one design event out of a range of events all having the same joint return period (*i.e.* on an isoline or isosurface of the copula) could be seen as a drawback of the multivariate approaches available in literature, as the most likely event does not necessarily correspond to the most severe one for a given hydraulic structure. However, there is the full potential to set a step aside from this 'one-event-design' approach to a full ensemble-based design approach. Therefore this paper includes an approach for the generation of a design ensemble. It is clear that the ensemble approach provides a lot more information on the possible outcome of design events. The proposed ensemble-based approach entails the most likely design event, but furthermore provides a clear idea on the probability that other events (but all having the same JRP) will occur. Checking these ensembles against the desired design of the hydraulic structure will illustrate the real threat to the structure. It therefore provides a way of assessing some uncertainty of the design variables associated with the selection of a single design event. If a single design event is sought, the pure copula-based approach (Section 2.3.3) has the advantage to guarantee that only a fraction of $1/T_{OR}$ events exceeds the margins of any of the possible estimated design events.

It should also be noted that the fitting of the copulas (bivariate, trivariate or multivariate) is a very important part of the design event

estimation. If the practitioner is not acquainted with this initial aspect of design studies, it is very easy to make wrong choices. Naturally, the multidimensional approaches require a larger data set in order to produce robust parameter estimates. Thus, the length of the time series and the amount of missing data have to be considered before an approach is selected. The authors of this work believe that the vine copula approach is the way to go for constructing flexible multivariate distribution functions, as it enables to use more widely spread bivariate copulas as building blocks for more complex multivariate distribution functions. Of course, a good balance between the number of variables considered and the (numerical) complexity of the vine copula should be sought, keeping in mind that all this also affects the eventual design. Further studies are necessary to assess the sensitivity of the JRP analysis to sample size and sample selection.

In general, the Kendall-based approach can be applied to any copula and can be used for both large (e.g. floods) and small (e.g. droughts) extremes. However, one should also be aware of the fact that the approaches in this study are only applied to variables that are positively associated and with a focus on extremes in terms of large values. In all other cases, adaptations should be made in order to operate in the right 'area' of the copula. Further applications of the copula-based and Kendall-based JRP approaches in other case studies should provide more insight on this in the near future. Serinaldi [77] highlighted the potentially misleading notion of return periods and suggests to report return periods alongside with annual exceedance probabilities as done e.g. by Theiling and Burant [88]. This aspect should be included in further studies as well.

2.7 CONCLUSIONS

The aim of this study was to provide an overview of state-of-the-art approaches to estimate design events for a given return period and to discuss their differences in a practical application. Therefore, a synthetic case study focusing on the estimation of design parameters for a synthetic design hydrograph (SDH) was considered. As they are the most important SDH variables, the peak discharge Q_p, its duration D and volume V_p were chosen.

In first instance, a review of several approaches yielding design events available in recent literature was provided focusing on how to apply these. As multiple variables were considered in the different return period approaches, an important aspect is (the modelling of) the dependence between variables. In this context, the potential and the use of copulas for the construction of multivariate distribution functions was stressed and illustrated. On the one hand a bivariate copula of (Q_p, V_p) was fitted. On the other hand, also the fitting of the

trivariate copula of (Q_p, D, V_p) was elaborated in a comprehensive way by means of the vine copula approach.

Eventually, design events for a 10-year joint return period were obtained considering a 2D regression based, a 2D conditional copula-based, a 2D copula-based, a 2D Kendall-based and a 3D Kendall-based approach. The traditional 1D return period definition is considered as a reference for comparison purposes. Differences in design quantiles were discussed while also the theoretical appropriateness was explained. This paper warns practitioners for blind use of just one available design event estimation approach, and stresses the importance of good copula fitting and the effect on the eventual design event outcome. A simulation study showed that the investigated approaches yield statistically different design events. Thus, the predictions are not only different following the theoretical inequality (Eq. (2.7)) but do withstand the variability due to uncertainties associated with the probabilistic model fitted to the data. Based on the available literature and the case study in this paper, the copula-based and Kendall-based JRP approaches are valuable multivariate extensions of the univariate approaches. However, their applicability always depends on the availability of data and the probabilistic nature of the actual real world problem. For constructing multivariate copulas, the vine copula method is advised.

Further (joint) research efforts should focus on a shift from one-design-event approaches to ensemble-design-event approaches, enabling to incorporate the variability in the design event selection. A first valuable approach to this ensemble-based design was provided in this paper. The ultimate goal should be the elaboration of a useful and understandable framework for multivariate frequency analyses, with clear guidelines to practitioners.

From a practical perspective, it is impossible to provide a general suggestion for an appropriate approach to estimate multivariate design events applicable to a vast set of design exercises. Firstly, as previously described, the available approaches are different from a statistical point of view. Until now many applications are based on the concept of univariate return periods, as the concept of multivariate return periods has a different meaning and is potentially less conservative. Secondly, the best approach, in our opinion, is related to the hydraulic structure to be designed. Different design exercises might be critical to single variables, which should then be selected as the driving component in the data selection and in the modelling process. If one is for example interested in the hydrograph volume for designing a reservoir for flood regulation, it is essential to understand whether there is a predominant driving variable. Specifically, if the reservoir is regulated by a levee, the volume design is related to a specific discharge value. In this case, the bivariate conditional distribution could be preferred, since the discharge analysis is performed

with a standard univariate approach and the volume return period is estimated conditioned on the discharge design value. In similar practical problems, the regression analysis could be preferred when the data availability does not justify a richer statistical model application. On the contrary, for instance, when the analyst is estimating the extension of flood inundation for which both peak discharge and volume could play a similar role, a joint return period approach could be appealing. Indeed, an ensemble of equally rare scenarios (i.e. having the same return period) could be used to assess the variability of the obtained flood maps due to the selection of a single design event. Also in this case, it should be kept in mind that the univariate n-year of return period is different to the bivariate and trivariate n-year return period. Even though it is to be expected that including more variables improves the modelling of the process, one should keep in mind the drastically increasing need of data to fit such models.

Acknowledgements

The critical and useful comments of two anonymous reviewers, as well as G. Salvadori greatly improved this work, for which the authors are very grateful. The interested reader is referred to the public review process of this paper at HESSD for a discussion on JRP going beyond this publication. The research of the first author is partially funded by the German Research Foundation (DFG) under project PE 1632/4-1. S. Vandenberghe was a doctoral research fellow of the Research Foundation Flanders (FWO). The Special Research Fund of Ghent University financially enabled the research of M. J. van den Berg.

3

SINGLE TREE SPATIAL VINE COPULAS

This chapter consists of the work published in proceedings of the first *Spatial Statistics Conference* held 2011 in Enschede, The Netherlands. The original work is entitled *The pair-copula construction for spatial data: a new approach to model spatial dependency* and has been written by Gräler and Pebesma [33]. References have been updated to be consistent throughout this thesis.

ABSTRACT

Copulas are a flexible tool to model dependence of random variables. They cover the range from perfect negative to positive dependence, include the independent case and incorporate asymmetric dependence as well as the widely used Gaussian dependence structure. The pair-copula construction for multivariate copulas exploits the ease of bivariate copulas and suggests a decomposition of a multivariate copula into a set of bivariate ones. We successfully adapted this approach for spatial data and developed a powerful spatial pair-copula based interpolation method.

3.1 INTRODUCTION

The common approach to describe dependence merely by correlation measures or covariance functions reduces the dependence structure of two random variables to a single measure and thus introduces strong simplifications. This approach might be too simplistic especially for complex dependence structures. Such complexity can often be found in natural phenomena. Therefore, it can be desirable to model the full multivariate distribution of the observed process.

The set of possible distributions is vast and their estimation might be cumbersome. The theory of copulas offers a flexible tool to build multivariate distributions where the dependence structures are modelled detached from the marginal distributions. While the bivariate case is quite well understood, the estimation of higher dimensional copulas however is still an elaborate procedure. Many different families of bivariate copulas have been developed and discussed in various fields of applied statistics. Some of these families can easily be extended to higher dimensions, but most of them lack this possibility. The pair-copula construction (PCC) described by Aas et al. [2] follows the concept of vine copulas: a multivariate copula is decomposed into a set of bivariate copulas. This approach offers a powerful tool to de-

scribe complex dependence structures and exploits the ease of using bivariate copulas.

We picked up the idea of the flexible PCC and adapted it to the spatial framework. Bárdossy [5], Bárdossy and Li [6] and Kazianka and Pilz [52] have based their modelling approaches on copulas as well but they use a comparatively small set of families. Our proposed procedure utilizes the full flexibility of the PCC and allows for asymmetric dependencies.

In the next section we give details on the underlying theory of copulas and our proposed procedure. Section 3.3 illustrates an application of the pair-copula interpolation for the Meuse data set [10]. In the last section we discuss the potential of the proposed pair-copula interpolation.

3.2 THEORY AND PROCEDURE

Copulas $C_n : [0,1]^n \to [0,1]$ can be understood as joint cumulative distribution functions of some random variable with on $[0,1]$ uniform distributed margins. Following Sklar's Theorem (see [57]) any n-variate distribution $H(x_1, \ldots, x_n)$ can be decomposed into its margins $F_1(x_1), \ldots, F_n(x_n)$ and its n-dimensional copula C_n by:

$$H(x_1, \ldots, x_n) = C_n(F_1(x_1), \ldots, F_n(x_n))$$

The transformation of the margins by their cumulative distribution functions guarantees on $[0,1]$ uniformly distributed random variables. Detaching the specific margins from the dependence structure allows us to handle the dependence of all multivariate distributions in a common way. The estimation of the copula is then solely based on the transformed margins. Instead of looking only at the copula itself we are interested in the strength of dependence reflected by the copula's density $c : [0,1]^n \to [0,\infty)$.

Bivariate copulas are quite well understood and can be estimated easily using maximum likelihood or moment based estimators. The inversion of Kendall's tau and Spearman's rho are two estimators that ground on the correlation measurements of Kendall and Spearman. Any bivariate copula can be used to calculate these correlation measures and the inversion of this relationship serves as estimator. This method is advantageous when a couple of copulas have to be estimated as the correlation measure has to be calculated only once for all families based on the data. The estimates can then often be calculated using a functional relationship specific for each family. Many different families of bivariate copulas have been developed and studied. Some of them can easily be extended for higher dimensions like the Archimedean or Gaussian copulas but they usually lack flexibility. The multivariate Gaussian copulas only allow for a Gaussian dependence structure between all margins and the higher dimensional

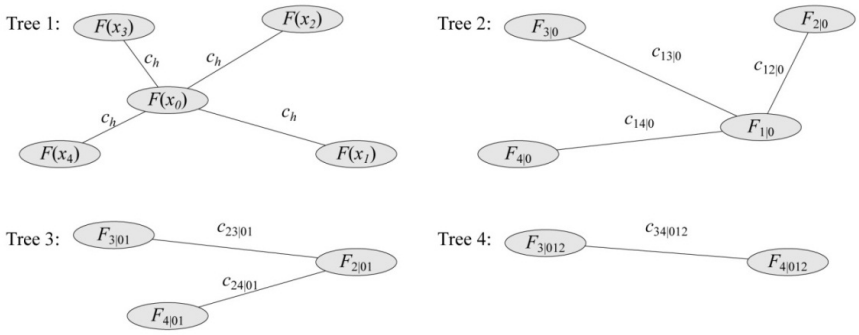

Figure 3.1: Decomposition of a 5 dimensional spatial random process following a canonical vine structure. The arguments of the conditional cumulative distribution functions are dropped.

Archimedean copulas use only a single generator function as in the bivariate case.

The pair-copula construction as described by Aas et al. [2] provides a procedure to further decompose a n-dimensional copula C_n into a set of $n(n-1)/2$ bivariate copulas. This approach allows to combine different families of copulas for different pairs of margins and higher order dependencies. The choice of decomposition is not unique in the sense that different decompositions will give different approximations of the full copula. Our approach will be based on the so called canonical vine. Its structure is given in Figure 1. Another simplification is made as the copulas may change for different values of the conditional variable and this influence is only represented in the conditional cumulative distribution functions. The limitations and usefulness of this simplification are discussed by Hobæk Haff, Aas, and Frigessi [44].

The density of the full pair-copula is the product of all bivariate copula densities following the decomposition structure. The conditional cumulative distribution functions occurring in the scheme can be calculated by partial derivatives of the copulas involved as follows. We denote the set of indices of the conditioning variables by v and the set of indices excluding j by $v-j$:

$$F_{i|v}(x_i) = \frac{\partial C_{ij|v-j}\left(F_{i|v-j}(x_i), F_{j|v-j}(x_j)\right)}{\partial F_{j|v-j}(x_j)}$$

The canonical vine puts the highest emphasis on a single variable that builds the root of the decomposition as shown in Figure 3.1. We consider the unobserved location surrounded by its nearest neighbours to be the central element of dependence in our procedure and adapt the scheme of the canonical vine for the spatial pair-copula used for interpolation.

As typical in geostatistics, we assume a stationary and isotropic spatial random field. The neighbourhood based random process

$H(x_0, x_1, \ldots, x_k)$ that we model describes the distribution of some variable of interest for a single location and its k nearest neighbours. The assumption of stationarity allows us to use the same marginal cumulative distribution function F for all locations. Exploiting as well isotropy, the influence of the neighbours is controlled only by the separating distance and can be described with a distance dependent copula $C_h(u, v)$. This spatial copula has to approach the upper Fréchet-Hoeffding bound $M(u, v) = \min(u, v)$ (mimicking perfect positive dependence) when the distance tends to zero and the product copula $\Pi(u, v) = uv$ (describing independence) when the distance tends to the range up to which data are spatially correlated. Instead of limiting a spatial copula $C_h(u, v)$ to a single family, we use a convex combination of copulas for different distances $h_1, \ldots, h_l$ and introduce M for zero separation and Π for the maximum range h_l. The spatial copula is given with $\lambda_i := (h_i - h)/(h_i - h_{i-1})$ by:

$$C_h(u, v) = \begin{cases} \lambda_1 M(u, v) + (1 - \lambda_1) C_{1,h}(u, v) & , \ 0 \leqslant h \leqslant h_1 \\ \vdots & \\ \lambda_i C_{i-1,h}(u, v) + (1 - \lambda_i) C_{i,h}(u, v) & , \ h_{i-1} \leqslant h \leqslant h_i \\ \vdots & \\ \lambda_l C_{l-1,h}(u, v) + (1 - \lambda_l) C_{l,h}(u, v) & , \ h_{l-1} \leqslant h \leqslant h_l \end{cases} \tag{3.1}$$

The parameters of each copula $C_{\cdot,h}$ involved in the convex combination may depend on the distance as well. Incorporating the functional relationship provided by the inversion of Kendall's tau or Spearman's rho allows for a common parametrisation of the parameter spaces. We use this spatial copula C_h in the first tree of the pair-copula C_{k+1} and adjust its parameter for each neighbour according to the separating distance. This approach strengthens the influence of the spatial information and we call this copula a spatial pair-copula. The copula families in the remaining trees of the canonical vine are chosen with respect to the best approximation of the natural random process.

The interpolation procedure grounds on the conditional density of the pair-copula C_{k+1} that is given by:

$$c_{k+1}(u_0 | u_1, \ldots, u_k) = \frac{c_{k+1}(u_0, u_1, \ldots, u_k)}{c_k(u_1, \ldots, u_k)}$$

As the pair-copula density c_{k+1} only approximates the full copula density, it is often difficult to estimate the lower dimensional denominator as a marginal distribution. However, it can always be calculated as the integral of c_{k+1} over the first parameter from 0 to 1 fixing the remaining parameters.

The random variable $\hat{Z}(s_0)$ at an unobserved location s_0 follows the distribution $H(x_0|x_1,\ldots,x_k)$ conditioned under the observed values of the k nearest neighbours $x_1,\ldots,x_k$. This conditional distribution can be expressed in terms of the conditional pair-copula. Point estimates can for instance be obtained by calculating the mean or the median:

$$\hat{Z}_{mean}(s_0) = \int_0^1 F^{-1}(u)\cdot c_{k+1}\big(u|F(x_1),\ldots,F(x_k)\big)du$$
$$\hat{Z}_{median}(s_0) = F^{-1}\big(C_{k+1}^{-1}\big(0.5|F(x_1),\ldots,F(x_k)\big)\big)$$

Furthermore, the conditional density of the copula can as well be used to generate several realizations at the unknown locations of the random field or to derive confidence intervals. This allows for a comprehensive uncertainty analysis of the estimates.

3.3 APPLICATION

We applied our pair-copula based interpolation to the Meuse river bank sample data set provided with the R package sp [61]. The dataset includes measurements of four heavy metals for 155 locations along with other secondary parameters. We aimed at interpolating the measurements of zinc over a regular grid along the river bank. The local neighbourhood consisted of the four nearest neighbours. The maximum distance at which the dependence between two locations was significant turned out to be approximately 555 m.

The estimation of the pair-copula involves several steps. At first, we estimated the distribution function F of the spatial random process incorporating classical tools. The best fit could be achieved for a generalized extreme value distribution (location = 246, scale = 147, shape = 0.71). Then, the data is transformed into a uniform distributed random variable using the cumulative distribution function. The following steps of the estimation procedure will only incorporate the transformed data.

The spatial copula was estimated using a lag based approach. For a set of distance classes one copula per class was estimated and assigned to the mean of these distances. This set of copulas was used as representatives for the distance dependent convex combination of copulas introduced in (3.1). The relationship of Kendall's correlation measure and the distance turned out to be a linear one for the first seven lags and was approximately zero for the remaining lags. This function was used to estimate Kendall's tau that serves as input for the inverse tau estimator for different distances. We compared the following 6 copula families: Clayton, Frank, Gumbel, Gaussian, an asymmetric one and a family of copulas with cubic-quadratic-sections (the definitions of the last two ones are given in the appendix 3.4). The Gumbel copula achieved the highest log-likelihood for the first 4 lag

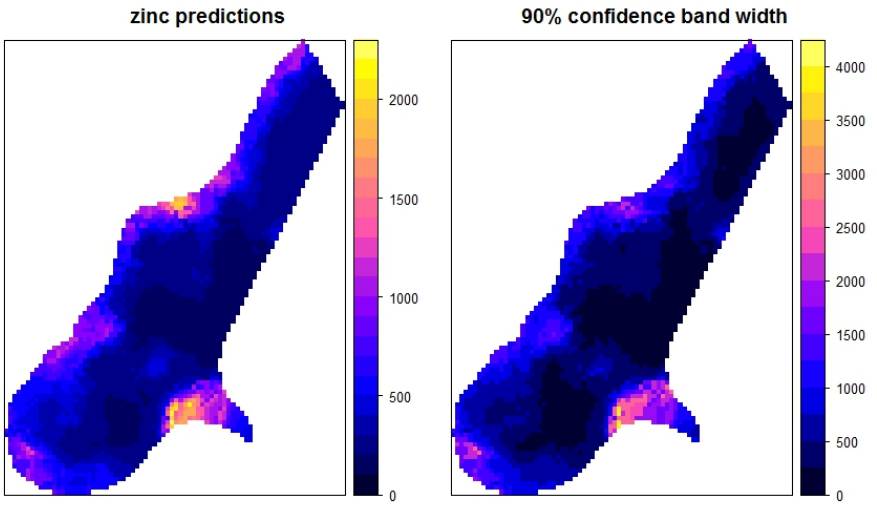

Figure 3.2: The Meuse river bank interpolated based on a pair-copula for a local neighborhood of 4 locations (left) and the width of its 90 % confidence band at any location (right).

classes while the Clayton copula was used for the remaining three. The pair-copula was estimated based on the data arranged in 5-tupels associating one observation to the observations of its 4 nearest neighbours. The four copulas in the first tree were already given by the spatial copula and could be used to calculate the conditional distribution functions $F_{(\cdot|0)}$ depending on the separating distance of the actual pair involved. The copulas for the higher order dependencies were as well estimated using the inversion of Kendall's tau. In cases where this estimate was not unique an additional maximum likelihood estimation on the reduced set of parameters was performed. The choice of the copula family was made by comparing the log-likelihoods of the best estimates across different families. In all but one cases the Gumbel family achieved the best fit to the data. The remaining fit originated from the family with cubic-quadratic-sections. The interpolation was carried out based on the expected value where the denominator of the conditional copula density was numerically evaluated. The interpolated grid and the width of its 90 % confidence interval are shown Figure 3.2. The large magnitude of some of the confidence widths is mainly due to the marginal extreme value distribution.

3.4 DISCUSSION AND CONCLUSION

We compared the log-likelihood of the spatial pair-copula with fits of the 5-dimensional Gaussian, Clayton, Frank and Gumbel families to assess the quality of fit. In all cases, the spatial pair-copula achieves a 1.3 to 1.5 times higher value. Providing the wrong distances to the spatial copula in the first tree reduces the log-likelihood as well. Thus,

the spatial pair-copula is a good fit within the set of copulas considered.

A successive cross-validation was carried out for all measurement locations. The root mean square error (RMSE) turned out to be slightly higher for the spatial pair-copula interpolation than for ordinary kriging, but the bias could be reduced by a factor of 2. Further cross-validations were carried out using a random subsample to estimate the spatial pair-copula and the variogram. For some subsamples, the pair-copula based approach could reduce the RMSE compared to kriging.

Inspecting different scatter plots during the estimation process of the pair-copula revealed some dependence structures that could only partly be described by the set of copulas investigated. Thus, some more flexible families or the development of more suitable bivariate copulas may improve the fit of the spatial pair-copula and the interpolation.

The spatial pair-copula strengthens the influence of the separating distance of the neighbours compared to other naïve copula based approaches. It already sufficiently captures the dependence structure of a local neighbourhood to compete with ordinary kriging. The provided conditional distribution for each interpolation location allows for a sophisticated uncertainty analysis of the estimates.

APPENDIX

The asymmetric family originates from the example 3.16 in [57] and is given by

$$C_{ab}(u,v) = uv + uv(1-u)(1-v)\big((a-b)v(1-u)+b\big)$$

The symmetric version is given by

$$C_{ab}(u,v) = uv\big(1-b(1-u)(1-v)+(b-a)(1-v)^2(1-u)^2\big)$$

In both cases the parameters are bounded by $-1 \leqslant b \leqslant 1$ and $(b-3-\sqrt{9+6b-3b^2})/2 \leqslant a \leqslant 1$.

4

SINGLE TREE SPATIO-TEMPORAL VINE COPULAS

This chapter has initially been published by Gräler and Pebesma [32] as extended abstract for the *Geostats Conference 2012* that was held in Oslo. The original work is entitled *Modelling Dependence in Space and Time with Vine Copulas*. References are updated to meet with the Bibliography of this thesis and some typesetting changes have been made.

ABSTRACT

We utilize the concept of *Vine Copulas* to build multi-dimensional copulas out of bivariate ones, as bivariate copulas are quite well understood and easy to estimate. The basis of our multidimensional copula is a *bivariate spatio-temporal copula* varying over space and time. The *spatio-temporal vine copula* models the underlying spatio-temporal random field for local neighbourhoods in a fully probabilistic manner.

Focusing on the interpolation of spatially under-sampled but temporally rich random fields, we apply this newly developed approach to a large data set of daily mean PM_{10} measurements over Europe during 2005. A cross-validation study is conducted to asses the power and quality of this approach.

4.1 INTRODUCTION

Copulas are capable of modelling any kind of dependence between random variables detached from their margins. The ability to capture the dependencies of extreme values made them popular in finance. Extreme values can also be found in many spatial datasets and their non-Gaussian dependence structures can easily be captured with copulas. Exploiting copulas potentially improves the interpolation of skewed and heavy tailed data.

The concept of *Vine Copulas* allows us to build multi-dimensional copulas out of bivariate ones. As bivariate copulas are quite well understood and easy to estimate, vine copulas are a promising tool to model multivariate distributions. The basis of our multidimensional copula is a bivariate spatio-temporal copula varying over space and time. The vine copula is fitted to local neighbourhoods and used to derive estimates from the neighbourhood's multivariate distribution.

Focusing on the interpolation of spatially under-sampled but temporally rich random fields, we apply this newly developed approach to a large data set of daily mean PM_{10} measurements over Europe

during 2005. To asses the goodness of this model, we conduct a cross validation on the PM_{10} measurements predicting the stations mean, median and 95 %-quantile.

In the following, we will give a brief introduction to copulas and extend this concept to spatial and spatio-temporal bivariate copulas and later to spatio-temporal vine copulas. In Section 3, the new approach is applied to a one year series of daily PM_{10} measurements in Europe followed by a discussion in Section 4. In the closing section, we conclude and point to further directions of this work.

4.2 COPULAS

Copulas are a probabilistic tool that allow to model altering dependencies across the full range of multivariate distributions. Following Sklar's theorem (see e.g. [57], as well for a detailed introduction), any d-variate distribution H can be decomposed into its marginal cumulative distribution functions $F_1, \ldots, F_d$ and its copula C by:

$$H(x_1, \ldots, x_d) = C\big(F_1(x_1), \ldots, F_d(x_d)\big)$$

The copula C can be seen as a d-variate distribution function over the hyper unit-cube $[0,1]^d$. Following the above decomposition, allows to build a vast set of multivariate distributions out of desired margins and a dedicated dependence structure.

Unfortunately, as flexibility increases with the dimension, so does the effort to estimate an appropriate copula. Quite many copula families have been discussed for the bivariate case, of which only few can easily be extended to the multivariate case without loosing the necessary flexibility. One possible approximation of multivariate copulas is obtained by *vine copulas* [2, 7, 44]. Vine copulas decompose a multivariate copula into a set of (conditional) bivariate ones. Any of these bivariate building blocks can be modelled by the best suitable copula without any restriction. This is advantageous (1) as it allows for a huge degree of flexibility and (2) as established estimation routines for the bivariate case can be used. The complete d-dimensional density of this copula is given as the product of all involved $\frac{1}{2}d(d-1)$ bivariate copulas and corresponding conditional cumulative distribution functions.

Naturally, the decomposition of a multivariate copula is not unique and a different ordering of the variables might lead to a different estimate. The two basic concepts of decomposition are called canonical vines (C-vines) and D-vines [2] where in the first approach, the first variable is used as conditioning variable for the following ones, and in the latter approach, the conditioning is done sequentially. More general decompositions are referred to as regular vines (R-vines). In this work, we will build on the canonical vine structure exemplary

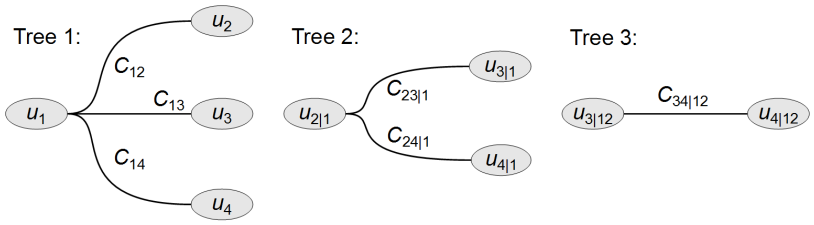

Figure 4.1: Structure of a 4-dimensional C-vine. The conditioned variables $u_{i,\nu} := F_{i|\nu}(i|\nu)$, with $i \in \{2,3,4\}$ and $\nu \in \big\{\{1\},\{1,2\}\big\}$ can be derived from the bivariate copulas of the preceding tree as illustrated in Equation (4.1).

depicted in Figure 4.1 for a 4-dimensional copula. The full density of this C-vine copula is given by:

$$\begin{aligned} c(u_1,\ldots,u_4) = \; & c_{34|12}(u_{3|12},u_{4|12}) \\ & \cdot c_{23|1}(u_{2|1},u_{3|1}) \cdot c_{24|1}(u_{2|1},u_{4|1}) \\ & \cdot c_{12}(u_1,u_2) \cdot c_{13}(u_1,u_3) \cdot c_{14}(u_1,u_4) \end{aligned}$$

The conditioned variables $u_{i|\nu}$, $i \in \{2,3,4\}$ and $\nu \in \big\{\{1\},\{1,2\}\big\}$, are derived through the copulas in the preceding tree (e.g. from tree 1):

$$u_{i|1} := F_{i|1}(u_i|u_1) = \left.\frac{\partial C_{1i}(u_1,u_i)}{\partial u_1}\right|_{u_1}, \; i \in \{2,3,4\} \tag{4.1}$$

A similar equation holds for the higher order trees.

4.2.1 *Spatial and Spatio-Temporal Bivariate Copulas*

In the domain of geosciences, one typically deals with spatially or spatio-temporally spread data. The locations of measurements in space and time can usually be used to derive relationships of the variables. In the following, we will assume a stationary and isotropic spatial (or spatio-temporal) random field. That is, we assume for any location $s \in \mathcal{R}$ in our spatial (spatio-temporal) region $\mathcal{R}$ the random field Z to take the same random variable $X = Z(s)$ and the dependence between two random variables $X_1 := Z(s_1)$ and $X_2 := Z(s_2)$ is a function of the separating Euclidean distance $h := \|s_1 - s_2\|$ only.

Considering the task of interpolating a spatial random field, for instance, the locations (or distances between locations) are used to derive the covariance matrix for the kriging predictor. We will follow a similar avenue and define a *spatial bivariate copula* as a bivariate copula taking the distance $h \in \mathbb{R}_{\geqslant 0}$ as parameter with the property that for $h \to \infty$ the copula tends to the product copula $\Pi(u,v) = uv$ denoting independence. Typically, the spatial bivariate copula will tend to the upper Fréchet-Hoeffding bound $M(u,v) := \min(u,v)$ denoting perfect positive dependence as h approaches 0. However, due to missing information on the very short distance variation of the

phenomenon, this bound does not have to be reached (similar as the nugget effect in kriging). In this work, the spatial copula is given as a convex linear combination of bivariate copulas where the mixing parameter function $\lambda : \mathbb{R}_{\geqslant 0} \to [0, 1]$ and the copulas depend on the separating distance h of two locations s_1, s_2:

$$C_h(u_1, u_2) := \lambda(h) \cdot C_i(u_1, u_2) + \big(1 - \lambda(h)\big) \cdot C_j(u_1, u_2), \ (i, j) := I(h)$$

Where I denotes a set of paired indicators separating the spatial range r_S of the model into a set of disjoint intervals (lags) and $I(h)$ provides the one pair of indexes (i, j) with respect to the distance h and corresponding copulas C_i and C_j. The copulas C_i and C_j denote the boundary conditions. Any distance larger than the spatial range r_S is modelled with the product copula Π. Due to the convex combination of copulas, the spatial bivariate copula will again be a copula for any distance h.

We define a *spatio-temporal bivariate copula* as a bivariate copula taking two parameters, the spatial and temporal separating distances h and t fading towards the product copula Π if h or t tend to infinity. This could for instance be realized as a convex combination of spatial bivariate copulas. For now, we consider only discrete points in time. Typically, this corresponds to the temporal resolution of measurements or aggregates thereof. Thus, the spatio-temporal bivariate copula can be defined as a set of spatial bivariate copulas indexed by the temporal gaps $1, \ldots, r_T$ investigated:

$$C_{h,t}(u_1, u_2) := \begin{cases} C_h^1(u_1, u_2) & , t = 1 \\ \vdots & \vdots \\ C_h^{r_T}(u_1, u_2) & , t = r_T \end{cases}$$

4.2.2 *Spatio-Temporal Vine Copulas*

A *spatio-temporal vine copula* models a neighbourhood of a spatio-temporal random field of size $k + 1$. This neighbourhood is composed of one central location and its k-neighbours in space and time. The first tree of the vine is realized by spatio-temporal bivariate copulas, reflecting the fact that the dependence structure changes over space and time. The remainder of the vine, i.e. the vine of the variables conditioned under the value of the central location, is modelled as some k-dimensional R-vine (a C-vine in our case). The structure of the first tree of a spatio-temporal vine copula is illustrated in Figure 4.2. Every curved connection is modelled by the same spatio-temporal copula $C_{h,t}$ but with different spatial and temporal distances h and t derived from the spatio-temporal locations involved. Combining the multivariate copula with the margins, which are assumed to be stationary, yields a full multivariate distribution of the neighbourhood of

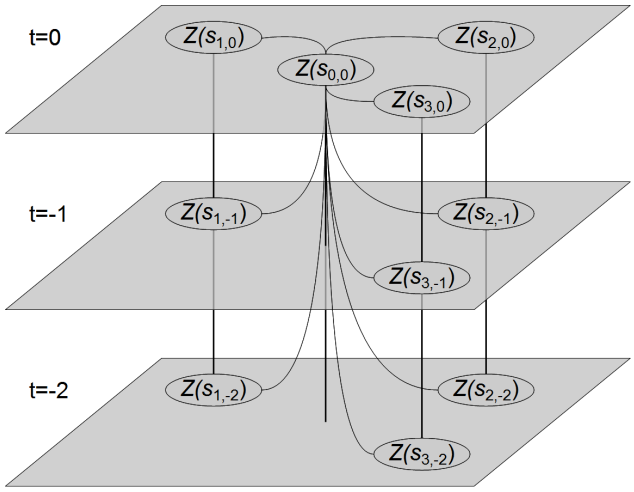

Figure 4.2: The first tree of a spatio-temporal vine copula for a neighbourhood of size 10 with 3 neighbours in space and 3 instances in time (the same moment, one and two time instances before indicated with 0,-1 and -2 respectively). Every curved connection is modelled by the same spatio-temporal copula $C_{h,t}$ but with different spatial and temporal distances h and t derived from the indicated locations $s_{i,j}, i \in \{0,1,2,3\}, j \in \{0,-1,-2\}$. The remaining trees follow a 9-dimensional C-vine.

the spatio-temporal random field. This distribution can then be used to simulate, predict or analyse the observed phenomenon. Quantile predictions can be made for any fraction $p \in (0,1)$ through Equation (4.2). The expected value of the conditioned distribution can be calculated by Equation (4.3). The two predictors are given by

$$\widehat{Z}_p(s_0) = F^{-1}\Big(C^{-1}\big(p|F(Z(s_1)),\ldots,F(Z(s_k))\big)\Big) \tag{4.2}$$

$$\widehat{Z}_m(s_0) = \int_{[0,1]} F^{-1}(u)\, c\big(u|F(Z(s_1)),\ldots,F(Z(s_k))\big)\, du \tag{4.3}$$

where F denotes the cumulative distribution function of the stationary random field and $s_1, \ldots, s_k$ are the spatio-temporal neighbours of s_0. Even though not explicitly stated, the copula C and its density c depend on the spatial and temporal distances between s_0 and its k-neighbours through the spatio-temporal bivariate copula in the first tree.

4.3 APPLICATION TO DAILY PM_{10} CONCENTRATIONS

The vine copula method is applied to a sample data set of daily mean PM_{10} measurements across Europe, using rural background stations from 2005. The data is publicly available through the AirBase[1]

1 http://www.eea.europa.eu/themes/air/airbase

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
$\Delta = 0$	t	F	t	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	Q	Q	Q	F
$\Delta = -1$	F	F	F	F	F	F	F	F	F	F	F	F	F	A	A	A	A	F	F	F	F	A	A	A	A	A
$\Delta = -2$	F	F	F	F	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	Q	Q	Q	Q	Q

Table 4.1: The copula families with the highest log-likelihood values for the first 26 spatial lag classes corresponding to distances up to 1000 km and three time instances. Abbreviations are as follows: t =Student, F = Frank, A = asymmetric, Q = cubic-quadratic-sections. The vertical lines indicate 100 km, 250 km, 500 km breaks

database hosted by the European Environmental Agency (EEA). Typically, the marginal distributions are unknown and have to be estimated as well. In order to not affect the copula by this estimation, we use rank-order transformed observations (i.e. $u_i := \text{rank}(x_i)/(n+1)$, where n is the length of the sample). To infer on the spatial dependencies, a set of 40 spatial lag classes is derived. Separate lag classes are filled with rank transformed data pairs from the same day, the one and two preceding days resulting in a set of 120 spatio-temporal lag classes. For every spatio-temporal lag, the best fitting copula from several copula families (elliptical, Archimedean, copulas with cubic-quadratic sections including an asymmetric one) is selected based on their log-likelihood values. These estimated copulas are then combined in a spatio-temporal bivariate copula as described in Section 4.2.1. As to be expected, the spatial bivariate copula fitted to rank transformed pairs measured the same day does not show any asymmetric dependencies. However, the asymmetric copula family is preferred over the other investigated families for some lags with pairs one day apart and dominates the convex linear combination for data pairs two days apart.

To build multiple samples of the stationary neighbourhoods, the data was arranged as spatial neighbourhoods and a random sample of 90 days for every station was included in the following analysis to reduce unwanted autocorrelation effects. This data is grouped in spatio-temporal neighbourhoods building the basis of the 10-dimensional vine copula. The fitted spatio-temporal bivariate copula is used in the first tree (see Figure 4.1) to derive the 9 dimensional data set conditioned under the one central location $s_{0,0}$ (see Figure 4.2). The remaining trees consist of 36 bivariate copulas and are iteratively estimated based on their maximum log-likelihood values. To assess the quality of our fit, we calculated the overall log-likelihood and compared it against simpler approaches. The log-likelihood value of our spatio-temporal vine copula (72709) is about 35 % larger than the fit of a Gaussian copula (53305) which included 45 covariance parameters. Furthermore, the Gaussian copula does not allow for asymmet-

Table 4.2: Cross validation results for the expected value and median estimates following the vine copula approach and the best performing method in [31] for comparison.

	expec. value	median	metric cov. kriging
root mean sq. er.	11.2	12.08	9.84
bias	-0.73	1.94	-0.24
mean abs. er.	6.95	6.87	5.66

ric dependencies opposed to the vine copula including asymmetric copulas.

To extend the validation of the fit beyond the log-likelihoods, we perform a cross validation leaving the full time series of one station out after another and predicting the expected value (Equation (4.3)), median and 95 % quantile (Equation (4.2) for $p = 0.5$ and $p = 0.95$) for every day during the year based on the conditional distribution from the three spatial neighbours and their three temporal instances. Thus, the cross-validation relies purely on spatial and spatio-temporal dependencies. To fully estimate the desired indicators, the marginal distribution has to be fitted. The best fit is achieved for a generalized extreme value distribution $\mathrm{GEV}(\mu, \sigma, \xi)$ with its parameters location, scale and shape set to $\mu = 13.94$, $\sigma = 8.54$ and $\xi = 0.20$ respectively (following the notation of the R-package `evd` [87]).

The full study is performed in the statistical computing environment and language `R` [64] using the package `spcopula`[2] that connects and extends the packages `spacetime` [59], `copula` [94, 55] and `CDVine` [74]. The `R`-scripts are available on request from the authors.

4.4 DISCUSSION

In the following, we will compare the copula approach with a spatio-temporal interpolation procedure described in the recent ETC/ACM Technical Paper 2011/10 [31]. The method therein performing best, applies residual kriging assuming a metric spatio-temporal covariance model (where 1 day $\approx$ 120 km) following a log-transformation of the original measurements and detrending by a linear regression with altitude and daily EMEP model predictions (Further details can be found in [31].).

Cross validation indicators for the predictions based on the conditional expected value and the median following the vine copula approach are shown in Table 4.2 a long with the values of the best performing method from [31]. Based on these numbers, no improve-

2 avaialble from r-forge:
`http://r-forge.r-project.org/projects/spcopula/`

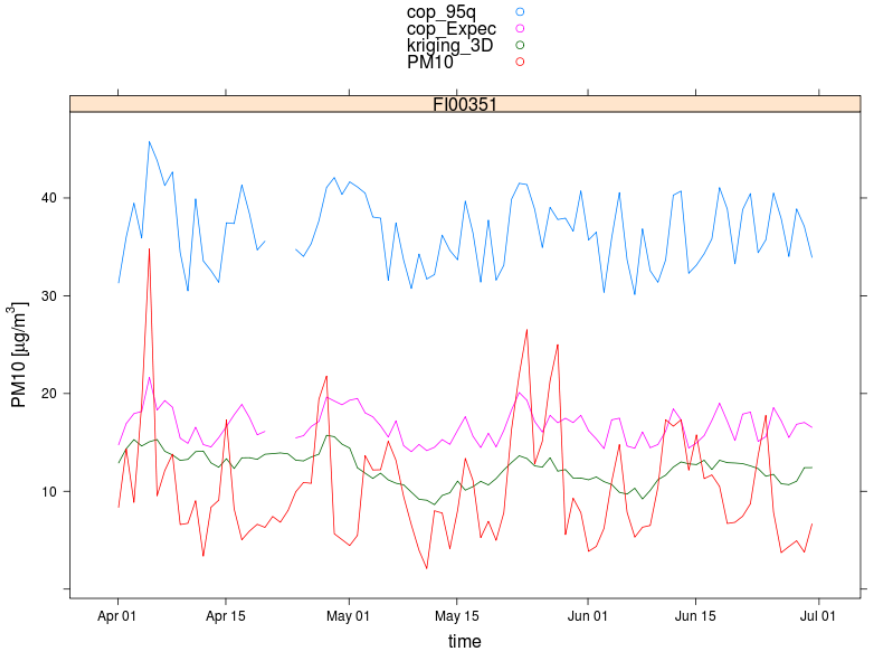

Figure 4.3: A time series plot of a Finish station showing the 95%-quantile of the copula prediction (cop_95q, blue), the conditional expected value estimate (cop_Expec, magenta), the metric covariance kriging predictions (kriging_3D, green) and the original observed PM_{10} concentrations (PM10, red).

ment could be achieved. However, the errors of the estimates based on the conditional expected value are of the same order of magnitude. It has to be noted that the neighbourhoods of the metric covariance model relying on the 100 nearest neighbours in a metric spatio-temporal space differs from the one underlying the vine copula approach building on the three nearest neighbours in space and three instances in time. The overall reproduction of the data set by the copula interpolation is rather good. The predicted median and 95 %-quantile are almost precisely exceeded by 50 % and 5 % of the original observations respectively. Looking into the predictions of single stations reveals cases where the copula approach outperforms kriging, but also versa. Two extreme scenarios are discussed in the following.

In Figure 4.3, a time series plot of a Finnish station roughly 600 km apart from any other station is shown. In our application, the copula prediction (drawn in magenta) is far above the real observations (drawn in red) while the metric covariance kriging prediction (drawn in green) seems to represent the mean process. An explanation might be given by the fact that the metric kriging model relies on the nearest stations in time and space. In this specific neighbourhood and a temporal scaling of 1 day$\approx$120 km, the nearest neighbours are roughly dominated by a factor of 10 through temporal instances. Thus, the predictions are similar to a temporal moving window av-

erage of a very few spatial neighbours. In the copula approach, we always rely on three nearest spatial neighbours and three instances in time. For larger distances, the conditioning influence of these neighbours is rather weak and the prediction value tends towards the expected value of the marginal distribution (21.0 $\mu g/m^3$ in our case) as the conditional density approximates a uniform distribution.

Another extreme case is shown in Figure 4.4. Here, the vine copula approach outperforms the 3D kriging approach for instance with respect to the station-wise root mean squared error with 4.0 $\mu g/m^3$ opposed to 5.4 $\mu g/m^3$. Even though both predictors follow the shape of the observations, the kriging estimate is often considerably above the observed concentrations. As this German measurement station is situated within a rather dense and dominating network, it closely follows the marginal distribution and the copula seems to well capture the spatio-temporal dependencies.

A potential advantage of the copula approach is the ease and flexibility with which one can predict quantiles of the distribution. In general, the conditional distributions at unobserved locations derived from the vine copula are not restricted to any specific distribution opposed to the kriging approach where every location is assumed to follow a Gaussian distribution. The blue lines in Figure 4.3 and Figure 4.4 showing the 95 %-quantile of the copula prediction are estimated with the same general Equation (4.2) as the median. Deriving quantiles for the complete modelling approach in [31] including log-transformations and detrending would require a simulation procedure.

4.5 CONCLUSION AND OUTLOOK

This paper reports on a early stage attempt to model spatio-temporal dependencies with copulas, and exposes cases where predictions based on copulas are worse than following a residual kriging approach as well as where the flexibility of the copula's dependence structures seems beneficial. The fully probabilistic nature of copulas allows to predict different kinds of statistical values. Mean estimates can immediately be given alongside with quantile estimates. However, further research will be necessary to fully identify the strengths and weaknesses of the vine copula approach in comparison to kriging. The presented study includes effects of log-transformations and detrending (altitude, EMEP model predictions) in the metric covariance kirging procedure. Thus, the kriging approach relies on more information and their effect on the predictions is hard to assess.

As illustrated in Figure 4.3, the assumption that the PM_{10} concentrations across Europe follow the same distribution, i.e. that the random process has the same mean at any location, seem not very well supported by the data. Further work will be needed to address the

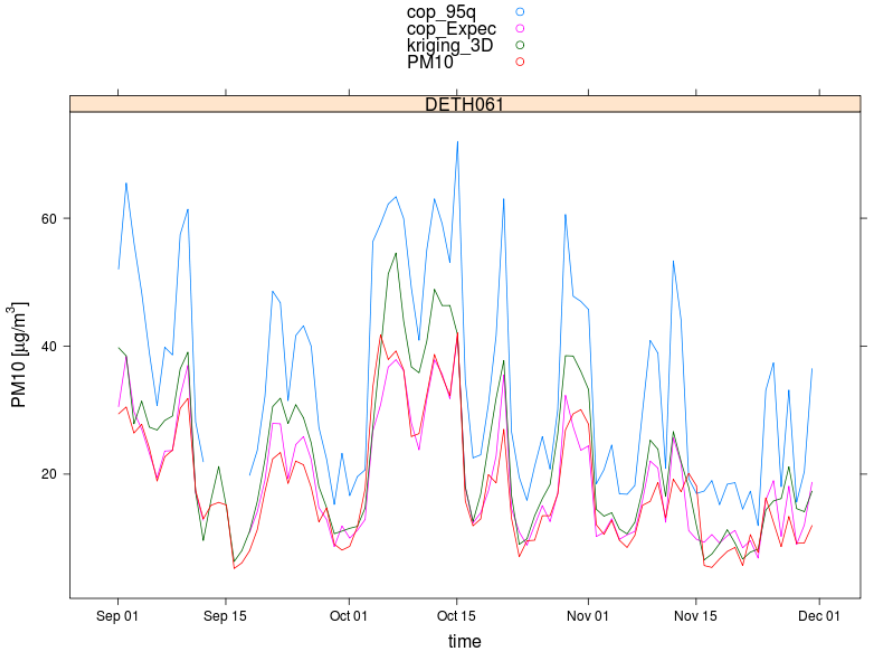

Figure 4.4: A time series plot of a German station showing the 95%-quantile of the copula prediction (cop_95q, blue), the conditional expected value estimate (cop_Expec, magenta), the metric covariance kriging predictions (kriging_3D, green) and the original observed PM_{10} concentrations (PM10, red).

issue of non-stationarity. Even though building higher dimensional distributions as in the presented approach remains a challenge, theoretical and software tools evolve to tackle these issues. Further developments will ease the estimation and application of spatial and spatio-temporal copulas.

5

SPATIO-TEMPORAL COVARIATE VINE COPULAS

This chapter originates from my manuscript entitled *spcopula: Modelling Spatial and Spatio-Temporal Dependence with Copulas in R* that has been submitted for publication. An overview of the spcopula package is provided alongside with an extension of the spatio-temporal vine copulas using covariates.

ABSTRACT

The spcopula R-package provides tools to model spatial and spatio-temporal phenomena with *spatial* and *spatio-temporal vine copulas*. Copulas allow us to flexibly build multivariate distributions with mixed margins where the copula describes the multivariate dependence structure coupling the margins. In classical geostatistics, a multivariate Gaussian distribution is typically assumed and dependence is summarized in a covariance matrix implying limitations like elliptical symmetry in the strength of dependence. Copulas allow for dependence structures beyond the Gaussian one, being for instance asymmetric. We developed the spatio-temporal vine copulas such that the bivariate copula families in the lower trees may change with distance across space and time allowing not only for a varying strength of dependence but also for a changing dependence structure. These spatio-temporal distributions are used to predict values at unobserved locations, assess risk, or run simulations. Based on the concept of vine copulas, the spcopula package provides a large set of multivariate distributions. As bivariate spatial copulas do not have any probabilistic restrictions, the spatial vine copula is a powerful approach for modelling skewed or heavy tailed data with complex and potentially asymmetric dependence structures in the spatial and spatio-temporal domain.

5.1 INTRODUCTION

Interpolation of spatial random fields is a common task in geostatistics. Simple approaches like inverse distance weighted predictions or the well known kriging procedures have routinely been applied for many years. However, when the underlying assumptions (i.e. a multivariate Gaussian distribution, possibly after transformation) of these approaches are hard to be fulfilled, alternatives are needed. Copulas have been used in some applications in the domain of spatial statistics. Bárdossy [5] was one of the first to apply copulas in a geosta-

tistical context. Some recent advances incorporating copulas in this field have for instance been published by Kazianka and Pilz [52] and Kazianka and Pilz [53], Bárdossy [3], Bárdossy and Pegram [4] or Bárdossy and Li [6]. They use a comparatively small set of copula families to model spatial processes.

The spatio-temporal domain rises in interest since several years, and several extensions of the spatial approaches to spatio-temporal ones have been developed (see e.g. Cressie and Wikle [16]). A major challenge with extending spatial kriging to spatio-temporal kriging is to build and fit well defined spatio-temporal covariance functions. The approach presented here differs from the classical geostatistical approaches by using *spatio-temporal vine copulas* to build a spatio-temporal distribution that does not rely on the Gaussian assumption, nor involves a covariance matrix. It extends the spatial vine copula [30] to the spatio-temporal context. Similar to co-kriging, we will introduce a spatio-temporal vine copula approach incorporating covariates.

The advantage of the *spatio-temporal vine copula* is its flexibility in the selection of copula families through *bivariate spatio-temporal copulas*. Bivariate spatio-temporal copulas are a convex combination of different copula families that are parameterised by spatial and temporal distance (Equation 5.1 in Section 5.2). This changing dependence structure allows for instance to preserve spatial symmetry within each time step while allowing for a directional effect across time. The introduction of a *bivariate spatial copula* into a vine copula for interpolation has been described by Gräler [30]. The bivariate spatial or spatio-temporal copulas are combined into a *vine copula* (also known as *pair-copula construction* [2, 8]) for a local spatial or spatio-temporal neighbourhood. A first approach to extend the spatial to the spatio-temporal approach has been presented in Gräler and Pebesma [32]. This paper describes a more flexible spatio-temporal neighbourhood structure and the introduction of a covariate to improve the prediction. Adding marginal distributions to the spatial or spatio-temporal vine copula yields a full multivariate distribution describing a local distribution of the observed phenomenon.

The spcopula R-package provides functions and classes to model spatial and spatio-temporal phenomena by vine copulas. Different tools have been implemented to fit spatial and spatio-temporal vine copulas to a data set, to interpolate the random field, and to predict quantiles from it. The package extends the copula R-package [55, 94] and provides additional copula families. Wrapper classes following the copula design to the copula families available in VineCopula [75] that have been implemented in spcopula are now directly available in VineCopula. The functionality for non-spatial vine copulas relies on VineCopula. For handling spatial and spatio-temporal data, spcopula builds on the R-packages sp [61] and spacetime [59]. A more de-

package	spcopula reuses and extends	spcopula adds to the functionality
copula	• S4-class definition `copula` • methods `fitCopula`, `dCopula`, `pCopula`, `rCopula`	• bivariate copula families `asCopula`, `cqsCopula` and `tawn3pCopula` • empirical and analytical tail dependence functions `empTailDepFun` and `tailDepFun` • partial derivatives via methods `dduCopula` and `ddvCopula` • inverse of partial derivatives via methods `invdduCopula` and `invddvCopula` • inverse of bivariate copulas for a given u or v via methods `qCopula_u` and `qCopula_v`
VineCopula	• function `BiCopSelect` • S4-class wrapper `vineCopula`	
sp	• abstract S4-class definition `Spatial` • function `spDists`	• nearest spatial neighbour calculation via function `getNeighbours`
spacetime	• abstract S4-class definition `ST`	• nearest spatio-temporal neighbour calculation via function `getStNeighbours`

Table 5.1: Overview of core dependencies and contributions of spcopula.

tailed overview of core dependencies and contributions of spcopula is provided in Table 5.1.

The remainder of this paper is organized as follows. The theoretical background of copulas, bivariate spatial copulas, bivariate spatio-temporal copulas and vine copulas are addressed in the next section. A strategy to estimate spatio-temporal vine copulas is given in Section 5.3. Section 5.4 discusses different applications of the multivariate distribution such as the possibility to predict values at unobserved locations or simulate from the spatial or spatio-temporal random field. An application is illustrated in Section 5.5, where daily mean PM_{10} concentrations (particulate matter less than 10 μm in diameter) across Europe observed throughout the year 2005 are interpolated including an additional covariate. Results are discussed in Section 5.6. Conclusions are drawn in Section 5.7.

5.2 SPATIO-TEMPORAL VINE COPULAS

In the following, we assume a spatio-temporal random field $Z : \Omega \times \mathcal{S} \times \mathcal{T} \to \mathbb{R}$ defined over some spatial domain $\mathcal{S}$ and temporal domain $\mathcal{T}$ of interest and an underlying probability space Ω. Typically, a sample $\mathbf{Z} = \big(z(s_1, t_1), \dots, z(s_n, t_n)\big)$ has been observed at a set of distinct spatio-temporal locations $(s_1, t_1), \dots, (s_n, t_n) \in \mathcal{S} \times \mathcal{T}$ where in general some spatial locations have been sampled at multiple time instances. Often, one is interested in modelling Z from the sample $\mathbf{Z}$ in order to predict at unobserved locations in space and time or simulate from the distribution. The spatio-temporal random field might as well be accompanied by a co-variate Y leading to a bivariate spatio-temporal random field: $(Z, Y) : \Omega \times \mathcal{S} \times \mathcal{T} \to \mathbb{R}^2$.

Copulas describe the dependence between the margins of multivariate distributions. Sklar [85] proved that any n-variate distribution H can be split into its margins $F_1, \dots, F_n$ and the copula C which couples the margins with a given dependence structure: $H(x_1, \dots, x_n) = C_n\big(F_1(x_1), \dots, F_n(x_n)\big)$. A copula can be imagined as a multivariate cumulative density distribution with uniform margins where its density reflects the strength of dependence between the margins. Many different parametric copula families exist allowing for very different dependence structures including certain symmetry properties but as well asymmetric, directional influences. In a bivariate symmetric case, the strength of dependence, i.e. the copula density (denoted by c), of a pair $(u, v) \in [0, 1]^2$ does not depend on the order, i.e. $c(u, v) = c(v, u)$, $\forall\ (u, v) \in [0, 1]^2$ while this does not hold for an asymmetric copula. Thinking of u and v as cumulative distribution values of two consecutive time steps, this allows to model a sudden rise in value with a different strength of dependence than a sudden drop. For further details we refer to the introductory book by Nelsen [57].

Sklar's Theorem is true for any dimension $d \geqslant 2$, but we will at first only consider bivariate copulas $C : [0, 1]^2 \to [0, 1]$. The density of a copula expresses the strength of dependence which changes over the range of the marginal distributions. The only copula exhibiting a constant strength of dependence across its margins is the product copula Π describing independence. Commonly, strength of dependence in a bivariate setting is measured by the correlation (or covariance) between two random variables and a Gaussian distribution is typically assumed implicitly. As a Gaussian distribution can be decomposed into a Gaussian copula with Gaussian margins, the Gaussian dependence structure is imposed which is elliptically symmetric (following the notion of elliptical contours of the bivariate Gaussian distribution). Hence, by only investigating the correlation of two variables, potential deviations of dependence from the Gaussian elliptical model are neglected. Different copulas might reflect samples having identical correlation, but different dependence structure (Figure 5.1). The same

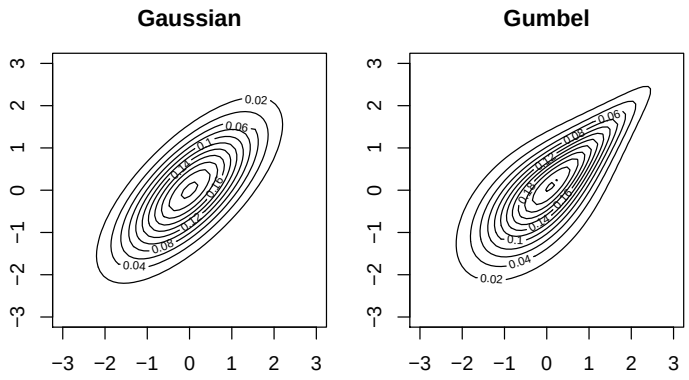

Figure 5.1: Copula densities of the Gaussian and Gumbel copulas. Both copulas are shown with standard normal distributed margins and a Kendall's tau correlation of 0.5.

applies to the spatial and spatio-temporal domain where kriging implicitly assumes a Gaussian dependence structure. However, looking into different data sets and investigating pairwise scatter plots reveals non-Gaussian dependence structures. Especially the correlation structure of pairs spanning across time may exhibit an asymmetric (directional) dependence. Such structures can be captured by copulas.

Copulas allow to model the dependence structure of a multivariate distribution disjoint from the marginals. This introduces a huge flexibility and eases the estimation at the same time. As the analytically known multivariate copula families are rather limited in their flexibility, we use vine copulas that allow to flexibly build multivariate copulas by any combination of bivariate copulas. For a successful model it is important to obtain good fits of both, margins and copula. The fitting of the margins can be carried out with any approach available in the literature. For the subsequent development of spatio-temporal vine copulas, we assume to have marginal distributions $F_{q,r}$ of Z and $G_{q,r}$ of Y for any location $(s_q, t_r) \in \mathcal{S} \times \mathcal{T}$.

We briefly introduce *bivariate spatial copulas* as in Gräler [30] by incorporating distance as the only parameter but utilizing the flexibility of many bivariate copula families. For pairs of locations we assume that the separation distance of these is the driving parameter determining the dependence. Hence, pairs of locations very close to each other are likely to exhibit a dependence structure close to perfect dependence where noise might reduce the strength of dependence to some degree (analogous to the nugget effect in kriging). For large distances, the pairs will tend to be independent and are modelled by the product copula Π. The approaches by Bárdossy [3] and Kazianka and Pilz [53] allow only for a single multivariate copula family. The bivariate spatial copula $c_h(u, v)$ recalled here is designed as a convex combination of bivariate copulas (in terms of their densities) that

is not limited to a single family (Equation 5.1). Hence, we allow not only for a varying strength of dependence but also for a dependence structure changing with distance:

$$c_h(u,v) := \begin{cases} c_{1,h}(u,v) & , 0 \leqslant h < l_1 \\ (1-\lambda_2)c_{1,h}(u,v) + \lambda_2 c_{2,h}(u,v) & , l_1 \leqslant h < l_2 \\ \vdots & \vdots \\ (1-\lambda_k)c_{k-1,h}(u,v) + \lambda_k \cdot 1 & , l_{k-1} \leqslant h < l_k \\ 1 & , l_k \leqslant h \end{cases} \tag{5.1}$$

where $\lambda_j := \frac{h-l_{j-1}}{l_j-l_{j-1}}$ is a linearly interpolated weight, h denotes the separating distance and $l_1, \ldots, l_k$ denote chosen distances of spatial bins (e.g. midpoint or mean distance of all point pairs involved in the estimation). The parameters of the copulas $c_{i,h}$ in the convex combination may as well depend on the distance h. With the help of the marginal CDF or a rank order transformation, the arguments u and v are the values of the pairs of locations transformed to $[0,1]$. Inspecting Equation 5.1 reveals that different choices of bins will in general yield different approximations to the underlying spatial dependence structure. This binning faces the same balancing issue as a classical variogram estimation where many bins allow for a flexible model but too few observations per bin and conversely few but well filled bins reduce the flexibility. It is important to maintain enough pairs per bin to allow for a sensible copula family selection throughout the estimation process.

The temporal extension of the bivariate spatial copula is yet another convex combination of bivariate spatial copulas c_h^Δ at different time lags Δ. In the case where one does not want to predict the spatio-temporal random field between time steps, the bivariate spatio-temporal copula can be reduced to a piecewise defined copula where the temporal lag between both spatio-temporal locations selects the bivariate spatial copula to be used.

Concentrating on a local neighbourhood of d spatio-temporal neighbours (Figure 5.2), we now model the pair-wise dependence between locations through a bivariate spatio-temporal copula. However, these copulas need to be joined to benefit from the full d-dimensional distribution of the neighbourhood. A technique to combine bivariate copulas into multivariate copulas has been introduced by Aas et al. [2] building on work from Bedford and Cooke [8]. This approach has first been introduced as the *pair-copula construction* and the resulting copulas are now known as *vine copulas* in the literature.

Vine copulas approximate multivariate copulas through bivariate building blocks (Figure 5.3). The joint density is obtained as the product of all bivariate copula densities involved. In the special case of spatio-temporal vine copulas with an additional covariate $Y : \mathcal{S} \times \mathcal{T} \to \mathbb{R}$, we model the first tree by bivariate spatio-temporal

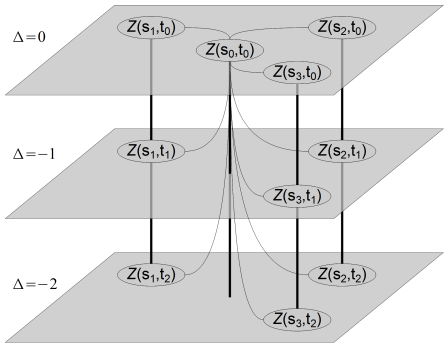

Figure 5.2: A spatio-temporal neighbourhood including the three spatially closest neighbours at three different time lags.

copulas c_h^Δ and c_{ZY}, the copula describing the dependence between Z and Y. The remaining upper trees are modeled by $c_{j,j+i|0,\dots,j-1}$ fixed over space and time:

$$\begin{aligned} &c_h^\Delta(u_0, v_0, u_1, \dots, u_d) \\ =&c_{ZY}(u_0, v_0) \cdot \prod_{i=1}^{d} c^\Delta_{h(0,i)}(u_0, u_i) \cdot \prod_{i=1}^{d} c_{Y,i|0}(u_{Y|0}, u_{i|0}) \\ &\cdot \prod_{j=1}^{d-1} \prod_{i=1}^{d-j} c_{j,j+i|Y,0,\dots,j-1}(u_{j|Y,0,\dots,j-1}, u_{j+i|Y,0,\dots,j-1}) \end{aligned} \tag{5.2}$$

where $v_0 = G_0\big(Y(s_0,t_0)\big)$ with G_0 being the co-variate's Y marginal cumulative distribution function at (s_0,t_0), $u_i = F_i\big(Z(s_q,t_r)\big)$ for $0 \leqslant i \leqslant d$ with (s_q,t_r) denoting the i-th closest neighbour of (s_0,t_0) with marginal cumulative distribution function $F_i = F_{q,r}$. For the spatio-temporally fixed upper part of the vine it is

$$\begin{aligned} u_{Y|0} &= F_{Y|0}(v_0|u_0) = \frac{\partial C_{Z,Y}(u_0, v_0)}{\partial u_0} \\ u_{i|0} &= F_{i|0}(u_i|u_0) = \frac{\partial C^\Delta_{h(0,i)}(u_0, u_i)}{\partial u_0} \end{aligned} \tag{5.3}$$

$$\begin{aligned} u_{j+i|Y,0,\dots,j-1} &= F_{j+i|Y,0,\dots j-1}(u_{j+i}|v_0, u_0, \dots, u_{j-1}) \\ &= \frac{\partial C_{j-1,j+i|Y,0,\dots j-2}(u_{j-1|Y,0,\dots j-2}, u_{j+i|Y,0,\dots j-2})}{\partial u_{j-1|Y,0,\dots j-2}} \end{aligned}$$

for $1 \leqslant j < d$ and $0 \leqslant i \leqslant d-j$.

In general, different decompositions of a multivariate copula exist, refereed to as regular vines, but in the spatial or spatio-temporal interpolation where a central element is naturally identified, we use a canonical vine where all initial dependencies are with respect to the central location. In the spatio-temporal tree (first tree in Figure 5.3)

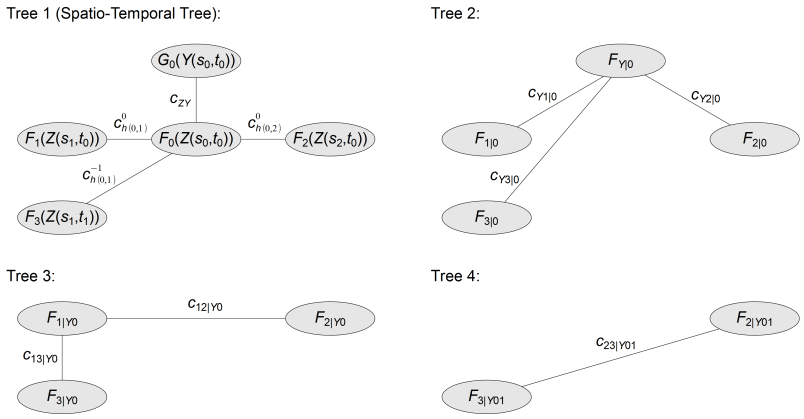

Figure 5.3: Graphical representation of a 5-dimensional local spatio-temporal vine copula with covariate Y, reoccurring location s_1 at the current and one preceding time slice and location s_2 at the current time slice. The notation follows the one introduced in Equation 5.2.

of the spatio-temporal vine, all edges connecting different spatio-temporal neighbours are modelled through a spatio-temporal copula $c^{\Delta}_{h(0,q)}$ parametrized by the spatial distance $h(0,q)$ and temporal lag $\Delta = t_0 - t_r$ between the data locations (s_0,t_0) and (s_q,t_r). The edge connecting the central location with its co-located covariate $Y(s_0,t_0)$ is represented by the best fitting bivariate copula c_{ZY}. In general, the dependency between the variable of interest and its covariate might as well change over space and time. All consecutive upper trees are modelled through spatio-temporally constant copulas. The upper vine structure does not impose any restriction on the bivariate copulas involved and are kept fixed no matter how the neighbourhood might be organized. The conditional distribution functions given in the above equations can immediately be obtained as partial derivatives of the already modelled copulas.

To achieve a full distribution describing the local behaviour of the spatio-temporal random field Z, margins need to be joined with the spatio-temporal vine copula. Depending on the properties of the phenomenon to be modelled, a single margin for all locations (in case the random field can be assumed to be stationary) or several margins incorporating some trend that is based for example on location, elevation or additional covariates might be used. The density of the full distribution is obtained by multiplying the copula's density with the marginal densities and the variables are mapped to the copula scale through the marginal cumulative distribution functions G_0 for the covariate Y and F_0 for Z both at (s_0,t_0) and $F_i = F_{q,r}$ of Z with (s_q,t_r) being the i-th neighbour of (s_0,t_0):

$$
\begin{aligned}
& f^{\Delta}_{\mathbf{h}}(z_0, v_0, z_1, \ldots, z_d) \qquad (5.4)\\
= \; & g_0(y_0) \cdot \prod_{i=0}^{d} f_i(z_i) \cdot c^{\Delta}_{\mathbf{h}}\big(F_0(z_0), G_0(y_0), F_1(z_1), \ldots, F_d(z_d)\big)
\end{aligned}
$$

where the z_i are representations of the random field Z at (s_i, t_i), the i-th neighbour of (s_0, t_0).

5.3 SPATIO-TEMPORAL VINE COPULA ESTIMATION

We introduce an estimation procedure for the spatio-temporal vine copula that borrows ideas from classical geostatistical approaches. To estimate the bivariate spatio-temporal copula, the data is grouped into spatial bins for each temporal lag. Kendall's tau correlation measure is marginally independent and thus represents the correlation at the copula level. This makes it very useful in the application of copulas and some one-parameter copula families even exhibit a one-to-one relationship between Kendall's tau and their parameter. The correlogram, using Kendall's tau, is calculated for the binned data. For each bin, several copula families are fitted to the transformed data (using a rank-order transformation or the fitted cumulative distribution functions of the margins) and the best fitting family (based on e.g. likelihood, AIC or BIC) is selected. Restricting the set of copula families to those exhibiting a direct link between Kendall's tau and their parameter, functions might be fitted to the afore obtained correlograms. These functions then relate separating distance through Kendall's tau to a parameter estimate for the copulas involved in the convex combination for each temporal lag. This way, the bivariate spatio-temporal copula will exactly reproduce Kendall's tau for any distance as modelled through the function from the correlogram. In case several best fitting families cannot be parametrized through Kendall's tau, one representative fit for each bin is obtained and combined as given in Equation 5.1. Using these static representations in the convex combination of copulas produces Kendall's tau values as a piecewise linear interpolation of the values obtained in the correlogram.

For further processing, the data needs to be re-arranged in spatio-temporal neighbourhoods of central locations and their spatio-temporally closest neighbours. Typically, the complex dependence structure over space and time does not relate to an easily obtained Euclidean distance measure in $\mathcal{S} \times \mathcal{T}$. To select the most correlated neighbours, a considerably larger spatio-temporal block neighbourhood (Figure 5.2) as the target dimension of the spatio-temporal vine copula is investigated and the d neighbours having the strongest correlation using the fitted correlogram functions are selected. This introduces some additional flexibility compared to the approach described by Gräler and Pebesma [32] as the neighbourhood does not depend on a fixed

spatio-temporal block size and missing values may easily be replaced by the next strongest correlated location. These neighbourhoods generate a $d+1$-dimensional dataset with approximately uniform on $[0,1]$ distributed margins. The bivariate spatio-temporal copula c_h^Δ on the first tree can now be used to derive the conditional sample of dimension d (conditioned to the value at the central location (s_0, t_0)). The spatio-temporally conditioned data is combined with data conditioned on the covariate and used for the remainder of the vine (Figure 5.3). The spatio-temporally fixed vine copula estimation proceeds sequentially by using the best fitting copula per bivariate pair (details are provided in Aas et al. [2], Czado, Schepsmeier, and Min [17] and Dissmann et al. [19]).

The joint copula density $c_{\mathbf{h}}^\Delta$ can then be obtained through Equation 5.2 where the first sequence of products reflects the spatio-temporal tree. The remaining spatio-temporally constant trees appear in the second and third product sequences. Fitting the marginal distributions, following generally any approach available in the literature, yields a full distribution through Equation 5.4 describing the local behaviour of the spatio-temporal random field $Z \times Y$.

5.4 PREDICTION OF THE SPATIO-TEMPORAL RANDOM FIELD

The local representation of the random field Z can be used for different purposes. A typical task is prediction of the modelled phenomenon at unobserved locations in space and time. To produce such predictions from a local neighbourhood, every unobserved location needs to be grouped with its d closest, i.e. strongest correlated, observed neighbours. Conditioning the $d+1$-dimensional copula $c_{\mathbf{h}}^\Delta$ on the observed values, yields a 1-dimensional distribution of the phenomenon. This conditional distribution can then be used to calculate the expected value (5.5), median or any other desired quantile (5.6) denoting for instance confidence intervals. At any location $s_0 \in \mathcal{S}$, the predictors for the mean value $\widehat{Z}_m$ and quantile values $\widehat{Z}_p$ for any $p \in (0,1)$ are:

$$\begin{aligned}\widehat{Z}_m(s_0) &= \int_{\mathbb{R}} z \cdot f_{\mathbf{h}}^\Delta(z|y_0, z_1, \dots, z_d)\, dz \\ &= \int_{[0,1]} F_0^{-1}(u)\, c_{\mathbf{h}}^\Delta\big(u|v_0, u_1, \dots, u_d\big)\, du \qquad (5.5)\end{aligned}$$

$$\widehat{Z}_p(s_0) = F_0^{-1}\big(C_{\mathbf{h}}^{\Delta\,-1}(p|v_0, u_1, \dots, u_d)\big) \qquad (5.6)$$

where $u_i = F_i(z_i) = F_i\big(Z(s_i, t_i)\big)$ for $1 \leqslant i \leqslant d$ and $v_0 = G_0(y_0)$ as before. The equality for $\widehat{Z}_m$ is based on a probability integral transform. An advantage of this approach is that the conditional distribution describing the random field at the unobserved location may take any form. This is different from kriging, where every predictive distribution is again a normal distribution. This richer flexibility has the

potential to provide more realistic uncertainty estimates. Another advantage that is immediate from Equation 5.5 and Equation 5.6 is that the only information on the marginals needed is their quantile function. This allows for instance to use approximations derived from the empirical cumulative distribution function without the knowledge of any explicitly known form of the family's density. However, the empirical cumulative distribution function is typically limited to the domain defined by the smallest and largest observation.

5.5 APPLICATION

The following calculations have been made using R 3.0.2 [65] and can be reproduced with the spcopula package. The demo `stCoVarVineCop` runs the estimation of the spatio-temporal vine copula as described below (on a 2 month subset of the full dataset).

Data

The dataset used in this application was obtained from the openly accessible AirBase[1], an European air quality data base maintained by the European Environmental Agency (EEA). We consider daily mean rural background PM_{10} concentrations (particulate matter smaller than 10 µm in diameter) in $\mu g/m^3$ across Europe for the entire year 2005. The data consists of 194 rural background stations with some missing observations at random points in time. A histogram of the skewed distribution is depicted in Figure 5.4. Preliminary results of this study excluding covariates and station wise marginal distributions were presented by Gräler and Pebesma [32] and the same data set has been analysed by Gräler, Gerharz, and Pebesma [31] using spatio-temporal approaches based on kriging. As a covariate, we included daily mean PM_{10} concentrations derived from model driven estimates by the European Monitoring and Evaluation Programme (EMEP) [23]. The 50 km EMEP Polar Stereographic grid is converted and projected to match the 10 km gridded interpolation domain in the standard EEA ETRS89-LAEA5210 projection.

Marginal distributions F_s *and* G_s

We fit marginal distributions for each location $s \in \mathcal{S}$ based on the time series leading to margins F_s and G_s for daily mean PM_{10} measurements and EMEP model estimates respectively. This leads to a slightly less general set-up as introduced earlier where the marginal distributions might as well change over time. Using the evd R-package [87], a generalized extreme value distribution is automatically fitted to each station's time series of PM_{10} and the EMEP model estimates. The assumptions of a single marginal distribution describing all stations

1 available from `http://acm.eionet.europa.eu/databases/airbase/`

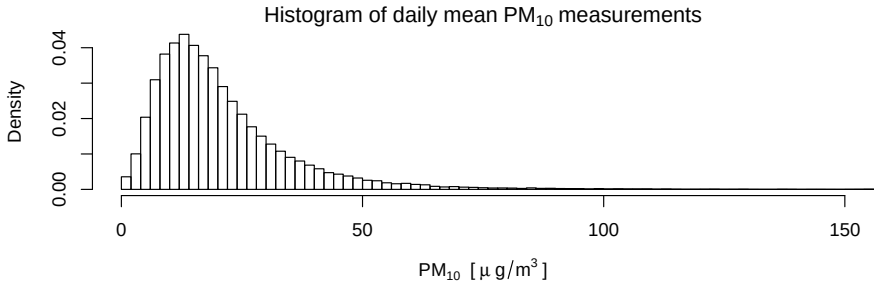

Figure 5.4: Histogram of the daily mean PM_{10} rural background concentrations across Europe during the year 2005. 60 observations extend beyond the plot up to approximately 400 —g/m^3.

could not be verified for either variable, as too many stations rejected these distributions in a Kolmogorov-Smirnov test. Obviously, these marginal distributions can only directly be fitted where we observed data. For an interpolation scenario, the marginals need to be extended towards prediction locations as well. Assuming that the marginal distributions change rather smoothly over space, we use two different models based on spatial proximity. One relies on a linear model incorporating the locations' coordinates and altitude followed by an inverse distance weighted interpolation of the residuals and the second one uses only inverse distance weighted means of the local neighbourhood's marginal parameters as estimates. Both approaches only use the spatially closest 9 locations for the inverse distance weighted means. Even though the spatio-temporal field we are modelling is not assumed to be stationary, we assume that the local dependence structure is the same all over Europe and can sufficiently be described by a station's 9 strongest correlated neighbours from up to 4 preceding time steps. In the following code snippets, we assume `EU_RB_2005` to be a spatio-temporal full data frame (class `STFDF` from spacetime [59]) that holds the already station-wise marginal transformed data and model estimates denoted as `marPM10` and `marEMEP` respectively.

Covariate copula C_{ZY}

Starting with the covariate copula c_{ZY}, we investigate the correlation structure of the daily mean PM_{10} measurements and EMEP model estimates over time. Figure 5.5 illustrates how the strength of correlation and the dependence structure changes throughout the year 2005. We use weekly correlations (red line in Figure 5.5) averaging out a great deal of variation, but largely maintaining the changes over time. The copula families are selected among the elliptical Gaussian and Student and the Archimedean Clayton, Gumbel, Frank [57] and Joe [47] copula families as indicated by the black line segments in Figure 5.5. The marginal independence of Kendall's tau ensures that this strength of dependence is the same for the marginal transformed as well as the raw data. This temporally changing covariate copula needs

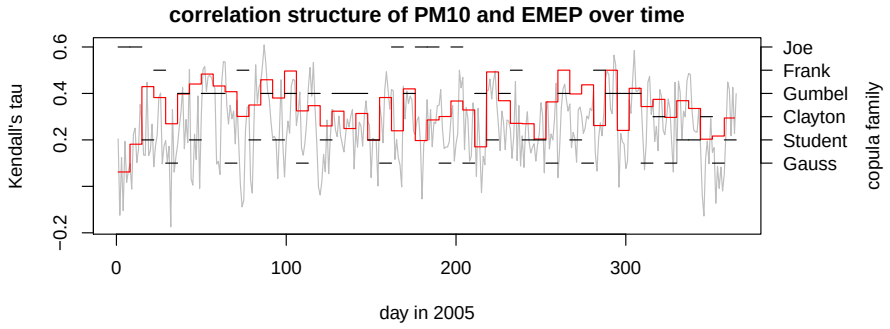

Figure 5.5: Correlation structure of daily mean PM_{10} measurements and EMEP predictions over time. The grey line indicates daily Kendall's tau values while the red step function describes weekly empirical Kendall's tau values. The black line segments denote the copula family best describing the dependence structure during each week for the given empirical correlation (without any particular ordering).

to be encoded as function taking the current spatio-temporal indices and returning a copula object:

```
R> library(spcopula)
R> coVarCop <- function(stInd) {
+    week <- min(ceiling(stInd[2]/7), 52)
+    copulaFromFamilyIndex(weekCop[[week]]$family,
+                          weekCop[[week]]$par,
+                          weekCop[[week]]$par2)
+  }
```

Spatio-temporal bivariate copula

For the estimation of the spatio-temporal bivariate copula, we follow the suggested procedure from Section 5.3 and start by grouping the data into spatial bins for five temporal lags (i.e. the same and first to fourth preceding day). For each spatio-temporal lag, the mean distance of all involved pairs and their Kendall's tau are calculated by:

```
R> stBins <- calcBins(EU_RB_2005, "marPM10",
+                     nbins=40, tlags=-(0:4))
```

The resulting object `stBins` is of type `list` with entries `meanDists` denoting the mean spatial distance of spatio-temporal bins, `lagCor` holding the correlation values per spatial bin and temporal lag and `lags` providing spatial and temporal indices to access the data from the underlying `STFDF`. Correlation functions (here five polynomials of degree three each) can directly be fitted to the above output and are joint in a spatio-temporal dependence function (`stDepFun`) through:

```
R> stDepFun <- fitCorFun(stBins, rep(3, 5), tlags=-(0:4))
```

Figure 5.6 illustrates the empirical values of Kendall's tau per spatial bin and temporal lag alongside with the corresponding polynomial fits. These polynomials describe how the strength of dependence

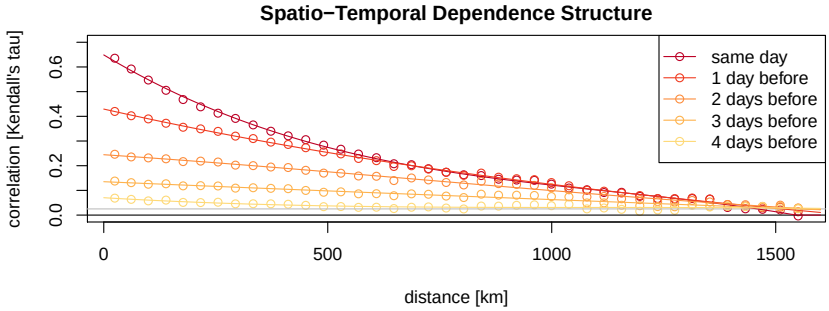

Figure 5.6: Empirical and modelled values of Kendall's tau for the bivariate spatio-temporal copula over five temporal lags.

changes with spatial and temporal distance. Following the estimation procedure, we need to investigate how the dependence structure, i.e. the copula families change for different spatial bins and temporal lags. The polynomials describing Kendall's tau in terms of distance are used to derive the parameter of a set of copula families and their log-likelihood is calculated per spatial bin and temporal lag. In case the relationship between Kendall's tau and the copula parameters is not unique, the parameter is optimised based on a log-likelihood approach restricted under the desired value of Kendall's tau. The function `loglikByCopulasStLags` calculates the log-likelihoods per spatial bin and temporal lag additionally returning the evaluated copula:

```
R> families <- c(normalCopula(0), tCopula(0),
+                claytonCopula(0), frankCopula(1),
+                gumbelCopula(1), joeBiCopula())

R> loglikTau <- loglikByCopulasStLags(stBins, EU_RB_2005,
+                                      families, stDepFun)
```

Copula families considered for the bivariate spatio-temporal copula (`families`) include the elliptical Gaussian and Student copulas, the Archimedean Clayton, Frank, Gumbel [57] and Joe [47] copulas. The best fitting copula family is selected based on the highest log-likelihood.

```
R> bestFitTau <- lapply(loglikTau,
+                       function(x) {
+                         apply(apply(x$loglik, 1, rank),
+                               2, which.max)
+                       })
```

In this application, the copula families change rather little and the Gumbel copula family (compare Figure 5.1) dominates the dependence structure. The spatio-temporal bivariate copula configuration is listed in Table 5.2. Using the earlier fitted polynomials and this selection of copula families, the convex combination of copulas (Equation 5.1 and the following paragraph) can now be composed to a spatio-temporal bivariate copula. The selected copula fits `listCops` and the representative distances `listDists` are provided as lists with

	spatial lag														
ID	1	2	3	5	6	7	22	23	25	26	27	28	29	30	33
mean dist. [km]	25	61	99	177	216	255	843	881	961	999	1038	1079	1117	1156	1274
$\Delta = 0$	t	G	t	...	t	G		...		G	F	N	F	N	...
$\Delta = -1$	G			...			G	F	...		F	N	F	N	N
$\Delta = -2$	G				...				G	N			...		
$\Delta = -3$	G					...						G	N	...	
$\Delta = -4$	G	...		G	J		...			J	G	G	N	...	

Table 5.2: Spatio-temporal bivariate copula family configuration for the first 33 spatial lags as suggested by the highest log-likelihoods. Δ indicates the time lag and the copula families are abbreviated as follows: N = Gaussian, t = Student, C = Clayton, F = Frank, G = Gumbel and J = Joe.

one entry for each temporal lag. Each of this entries contains a list of copulas in spatially ascending order:

```
R> distSelect <- function(x) {
+    stBins$meanDists[sort(unique(c(which(diff(x)!=0),
+                                   which(diff(x)!=0)+1,
+                                   1,40)))]
+  }
R> listDists <- lapply(bestFitTau, distSelect)

R> famSelect <- function(x) {
+    families[x[sort(unique(c(which(diff(x)!=0),
+                             which(diff(x)!=0)+1, 1,40)))]]
+  }
R> listCops <- lapply(bestFitTau, famSelect)
```

As the corresponding Kendall's tau value used to tune the bivariate copula's parameter is calculated for each spatio-temporal distance, the above components and the spatio-temporal dependence function define the bivariate spatio-temporal copula `stBiCop`:

```
R> stBiCop <- stCopula(components = listCops,
+                      distances = listDists,
+                      tlags=-c(0:4), stDepFun=stDepFun)
```

Joining vine copula

For the further processing, the data needs to be rearranged in local neighbourhoods. In our application, we are interested in the nine strongest correlated neighbours. Typically, there is no easily identified metric selecting these and we start with a larger static neighbourhood structure following Figure 5.2. The function `reduceNeighbours` selects the n strongest correlated neighbours as given in the third argument. To accomplish this, the spatio-temporal distances of the larger cubic neighbourhood are passed to the spatio-temporal dependence function `stDepFun` and the neighbours with the highest values are selected:

```
R> stNeigh <- getStNeighbours(EU_RB_2005, var="marPM10",
+                              spSize=9, tlags=-(0:4),
+                              timeSteps=90, min.dist=10)
R> stRedNeigh <- reduceNeighbours(stNeigh, stDepFun, 9)
```

This approach allows to drop missing values and to select the next best (i.e. next strongest correlated) available neighbour. The argument `spSize` denotes the number of locations at the current level including the central location. By setting `timeSteps` to 90, every location will only be used at 90 randomly assigned time steps as a centre of the neighbourhood. This way, only temporal chunks of five consecutive days are used reducing unwanted autocorrelation in the time series but reflecting the local temporal structure of the phenomenon. The argument `min.dist` ensures that any pair of neighbours has a minimum spatial separation distance (here 10 m).

The conditioning on the covariate EMEP and the evaluation of the spatio-temporal bivariate copula on the neighbours can be done separately. The spatio-temporal bivariate copula returns pseudo observations that are the values of the neighbours conditioned on the value of the central location ($u_{i|0}$ from Equation 5.3) incorporating the spatio-temporal distances between these by:

```
R> condData <- dropStTree(stRedNeigh, EU_RB_2005, stBiCop)
```

In a loop, the weekly varying copulas as depicted in Figure 5.5 and encoded as `coVarCop` are used to relate the marginal transformed variable PM_{10} with its covariate EMEP. In order to select the appropriate copula, the spatio-temporal indices need to be retrieved from the larger neighbourhood:

```
R> condCoVa <- condCovariate(stNeigh, coVarCop)
```

Binding this conditioned column `condCoVa` with the conditioned data from the spatio-temporal bivariate copula `condData` yields the data set for the upper vine estimation through the generic function `fitCopula`:

```
R> secTreeData <- cbind(condCoVa, as.matrix(condData@data))
R> vineFit <- fitCopula(vineCopula(10L), secTreeData)
```

Following the initial definition of the function `fitCopula`, the copula family that should be fitted needs to be provided as copula object. The call of the constructor `vineCopula` with an integer as argument (denoting the dimension) generates a canonical vine with independence copulas. The fitting routine sequentially selects the best fitting copula from a versatile set of copulas (see the documentation of VineCopula). The final spatio-temporal covariate vine copula is defined by:

```
R> stCVVC <- stCoVarVineCopula(coVarCop, stBiCop,
+                               vineFit@copula)
```

The cross-validation carried out to assess the goodness of fit of this spatio-temporal vine copula drops the complete time series of each

location in turn and predicts this time series based on the neighbouring data. In this application, we predict the expected value as given in Equation 5.5 of the time series for each moment in time. Prediction for a spatio-temporal target geometry `targetGeom` from the spatio-temporal vine copula is obtained by:

```
R> predNeigh <- getStNeighbours(EU_RB_2005, targetGeom,
+                                spSize=9, tlags=-(0:4),
+                                var="marPM10",
+                                coVar="marEMEP",
+                                prediction=TRUE,
+                                min.dist=10)
R> predNeigh <- reduceNeighbours(predNeigh, stDepFun, 9)
R> stVinePred <- stCopPredict(predNeigh, stCVVC,
+                              list(q=qFun),
+                              method="expectation")
```

Where the argument `method` selects from the prediction methods `expectation` and `quantile` as given in Equation 5.5 and Equation 5.6 respectively. The default quantile is the median, but any fraction between 0 and 1 can be assigned to an argument `p`. This can also be used for simulation purposes, as simulated values can be obtained through uniform distributed fractions assigned to `p`.

5.6 RESULTS AND DISCUSSION

Performing a cross-validation by leaving one complete station time series out in turn, the performance of this approach is assessed. Table 5.3 lists the results of a cross-validation using the expectation predictors for the newly presented spatio-temporal covariate vine copula approach (STCV) for different margins, a STCV solely composed out of Gaussian copulas and the simpler spatio-temporal copula presented in Gräler and Pebesma [32]. Additionally, results of a cross-validation for an approach based on spatio-temporal metric residual kriging with an underlying log-linear regression of the same data set but incorporating 100 nearest neighbours [31] is presented. In the case where the marginals refer to *local GEV*, the station-wise estimates are used. This is only possible in the case of a cross-validation as in a typical prediction or simulation application this extra knowledge about the margins is not available. This can be seen as a very good model of the marginals across space as the distributions of the margins are still not known, but extra data is used to estimate them. Hence, in a realistic cross-validation scenario, an additional model on the marginal distributions' parameters needs to be used. Here, we fitted two models for each of the three parameters. One using a linear model including coordinates and the station's altitude and performing a inverse distance weighted interpolation on the residuals of the spatially closest 9 neighbours (denoted *lm+IDW*) and another using only an inverse distance weighted interpolation of the spatially clos-

dependence model	margin	RMSE	MAE	ME	COR
STCV $\widehat{Z}_m$	local GEV	8.53	4.61	-0.05	0.84
Gaussian STCV $\widehat{Z}_m$	local GEV	8.65	4.59	0.08	0.83
STCV $\widehat{Z}_m$	lm+IDW GEV	10.12	5.79	0.17	0.76
STCV $\widehat{Z}_m$	IDW GEV	10.82	6.26	0.14	0.72
metric res. kriging	log linear reg.	10.67	6.16	0.47	0.74
sp.-temp. vine $\widehat{Z}_m$	global GEV	11.20	6.95	-0.73	NA

Table 5.3: Cross-validation results comparing the expectation spatio-temporal estimators for the newly presented spatio-temporal covariate vine copula with different margins, the spatio-temporal vine copula as in Gräler and Pebesma [32] and results from an earlier study by Gräler, Gerharz, and Pebesma [31] using kriging based approaches assuming a metric spatio-temporal covariance structure of the residuals of log-linear regression.

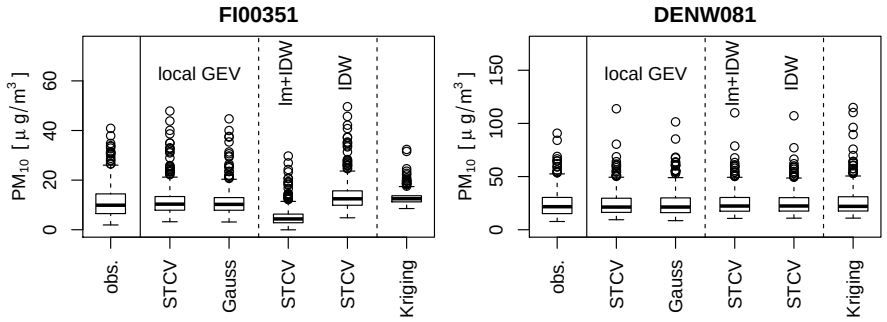

Figure 5.7: Boxplots comparing the different predictors with the observed values at two stations. The Finish station (left) is an isolated location while the German station (right) is situated in a rather dense network.

est 9 neighbours (denoted *IDW*). The major benefit in Table 5.3 is due to the additional knowledge on the marginal distribution functions. A smaller improvement could be made for the STCV using the lm+IDW marginal distributions but these statistics are within the same order of magnitude as the earlier presented spatio-temporal vine copula or the kriging predictor. These results underline how important it is to obtain good fits of both, the copula and the marginal distributions.

Besides looking at pure cross-validation statistics, it is important to consider the reproduction of the full distribution. Figure 5.7 shows boxplots of the observed versus modelled time series at two exemplary stations in Finland (FI00351) and Germany (DENW081). The Finish station is far away from any other station while the German station is situated in a rather dense network. Looking at many more plots of this kind for several stations (not shown) reveals that the copula based methods are rather close to each other and represent the original data quite well. Within the copula approaches, the full

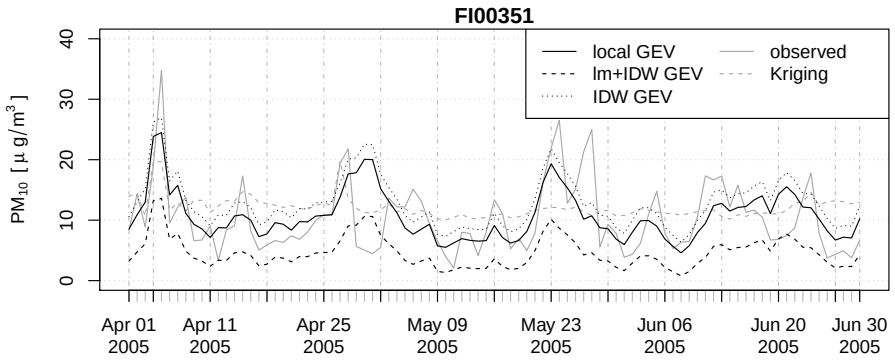

Figure 5.8: A subset of the time series at a Finish station comparing different predictors.

cross-validations *lm+IDW* and *IDW* have the largest deviations. The kriging based approach turns out to predict too large as well as too small ranges of values depending on the single station with a notable shift of the median at some stations. The issue of failing to reproduce the time series at isolated locations detected in the first application of a simpler spatio-temporal vine copula [32] could be overcome with the local GEV margins and considerably improved with *lm+IDW* and *IDW*. In Figure 5.8 the same temporal subset is plotted as in Figure 3 of Gräler and Pebesma [32] but the copula approaches are now able to follow the time series more closely than the kriging based predictor.

An advantageous feature of the copula approaches is their ability to provide potentially more reliable uncertainty assessments. Different from kriging, the prediction variance does not only depend on the spatio-temporal configuration of the locations, but also on the predicted value. Due to the nature of kriging, every conditional distribution is again a normal distribution. This is different for the copula approaches where the predictive density can take any form and the fitted marginal distribution functions ensure a reasonable range of possible values. Figure 5.9 illustrates the predictive densities at two different days at location DENW081. Note that besides the position also the shape of the prediction CDFs changes.

Even though simulating from a spatio-temporal random field modelled by a spatio-temporal covariate vine copula has not been shown in this application, it is readily done by not predicting for a constant cumulative distribution value but a random p in $\hat{Z}_p(s,t)$ for each location (s,t). In a conditional simulation, it is a modellers choice to which degree already sampled and closer locations are preferred over conditioning but further apart locations in the local neighbourhood. Modelling the spatio-temporal random field only locally requires a simulation along a random path. To avoid clustering effects along this path, simulating might start on a regular coarse subset of the target locations and subsequently continue on finer regular subsets [28].

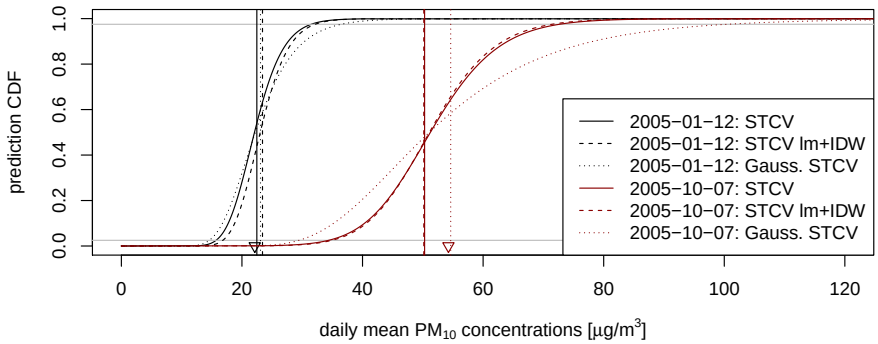

Figure 5.9: Prediction CDFs for station DENW081 at two different days in 2005. The vertical lines denote the predicted while the symbols on the x-axis indicate the observed value. The two horizontal lines denote the cumulative probabilities 0.025 and 0.975 easing the assessment of the 95%-confidence intervals.

Computational cost of the vine copula approaches might be a burden where results need to be generated very fast. The prediction of approximately 70000 values in the presented cross-validation takes a bit more than a day on a common laptop. No efforts have been made to speed up the execution for example by making use of parallel evaluation.

5.7 CONCLUSIONS

The spcopula package allows the modelling of spatial and spatio-temporal random fields by vine copulas. An earlier study [30] demonstrated the potential of spatial vine copulas for heavily skewed spatial random fields. The newly presented spatio-temporal covariate vine copula improves the interpolation of particulate matter compared to earlier spatio-temporal vine copulas [31] and linear geostatistical approaches. Nevertheless, the current bottleneck of the presented application is the flexible modelling of marginal distributions. This became evident by the presented cross-validation using extra knowledge about the margins.

The bivariate spatio-temporal copulas on the first tree mainly incorporate a Gumbel copula indicating a stronger dependence in the tails, a feature that is not available in a Gaussian set-up. Each spatio-temporal location has its own individual conditional distribution that is used for prediction. Different from kriging where each predictive distribution is again Gaussian, the conditional distributions of the STCV can take any form, potentially providing more realistic uncertainty estimates. Simulating from the modelled random field is possible and has been implemented in the presented package. A disjoint modelling of margins and dependence structure introduces a large flexibility to define a random field's distribution. Nevertheless, it is

very important to obtain good models for both components for a successful application.

6

MULTIPLE TREE SPATIAL VINE COPULAS

This chapter contains the paper entitled *Modelling Skewed Spatial Random Fields through the Spatial Vine Copula*. It has been published by Gräler [30] in the journal *Spatial Statistics*. Some typesetting changes have been made including updated references to meet the bibliography of this thesis.

ABSTRACT

Studying phenomena that follow a skewed distribution and entail an extremal behaviour is important in many disciplines. How to describe and model the dependence of skewed spatial random fields is still a challenging question. Especially when one is interested in interpolating a sample from a spatial random field that exhibits extreme events, classical geostatistical tools like kriging relying on the Gaussian assumption fail in reproducing the extremes. Originating from the multivariate extreme value theory partly driven by financial mathematics, copulas emerged in recent years being capable of describing different kinds of joint tail behaviours beyond the Gaussian realm. In this paper *spatial vine copulas* are introduced that are parametrized by distance and allow to include extremal behaviour of a spatial random field. The newly introduced distributions are fitted to the widely studied emergency and routine scenario data set from the spatial interpolation comparison 2004 (SIC2004). The presented spatial vine copula ranks within the top 5 approaches and is superior to all approaches in terms of the mean absolute error.

6.1 INTRODUCTION

Interpolation of spatial random fields is a common task in geostatistics. Simple approaches like inverse distance weighted predictions or the well known kriging procedures have routinely been applied for many years. However, when the underlying assumptions (i.e. Gaussianity) of these approaches are hard to be fulfilled, alternatives are needed. Copulas have been used in different but few applications in the domain of spatial statistics. Bárdossy [5] was one of the first who applied copulas in a geostatistical context. Some recent advances incorporating copulas in this field have for instance been published by Kazianka and Pilz [52] and Kazianka and Pilz [53], Bárdossy [3], Bárdossy and Pegram [4] or Bárdossy and Li [6]. They use a comparatively small set of copula families to model spatial processes. Copulas

describing the dependence structure of extremes can for instance be found in Grimaldi and Serinaldi [39], Salvadori and De Michele [68], Salvadori, De Michele, and Durante [71] or Kao and Govindaraju [50]. These applications typically investigate multivariate extremes without addressing spatial dependence.

The set of methods to model spatial data including extremes is diverse. The different approaches go beyond the field of geostatistics [e.g. 25] and incorporate techniques such as neural networks [e.g. 89] or support vector machines [e.g. 63] as presented in the spatial interpolation comparison 2004 [SIC2004: 20]. Typically studied spatial phenomena exhibiting extremes are for example radioactive radiation, as in SIC2004, rainfall data [42] or air quality indicators [45].

The advantage of the *spatial vine copula* approach presented in this paper is its flexibility in the selection of appropriate copula families through bivariate spatial copulas. Schepsmeier [73] suggests an approach where the tree structure of the vine is derived through spatial distances, but the copula families do not change with distance. Another approach modelling several air-quality indicators across a set of stations is briefly introduced by Brechmann [12] using a hierarchical Kendall copula.

The introduction of a bivariate spatial copula into a vine copula for interpolation has been described by Gräler and Pebesma [33] and is extended in this paper. Convex combinations of bivariate copulas parametrized by distance are combined in a *vine copula* [also known as *pair-copula construction*: 2, 8] for a local neighbourhood. Adding marginal distributions to the spatial vine copula yields a full multivariate distribution describing a local spatially dependent distribution of the observed phenomenon.

In the following, we will assume a spatial random field Z with $Z : \Omega \times \mathcal{S} \to \mathbb{R}$ defined over some spatial domain of interest $\mathcal{S}$ and probability space Ω. Typically, a sample $\mathbf{Z} = \big(z(s_1), \dots, z(s_n)\big)$ has been observed at a set of distinct locations $s_1, \dots, s_n \in \mathcal{S}$. Often, one is interested in modelling Z from the sample $\mathbf{Z}$ in order to predict $Z(s_0)$ at unobserved locations $s_0 \in \mathcal{S}$ or to simulate the spatial random field.

The remainder of this paper is organized as follows. The theoretical background of copulas, bivariate spatial copulas and vine copulas yielding the spatial vine copulas, which are the driving probabilistic tool in the applications, are addressed in the following section. A strategy to estimate a spatial vine copula is illustrated in Section 6.3. Section 6.4 discusses different uses of the multivariate distribution such as the possibility to predict values at unobserved locations or simulate from the spatial random field. An application is illustrated in Section 6.5 where we use the emergency and routine scenario data sets from the SIC2004 [20]. Results are discussed in Section 6.6. Conclusions are drawn in Section 6.7.

6.2 SPATIAL VINE COPULAS

Copulas describe the dependence between the margins of multivariate distributions. Sklar [85] proofed that any multivariate distribution H can be split into its margins $F_1, \dots, F_n$ and the copula C which couples the margins with a given dependence structure: $H(x_1, \dots, x_n) = C\big(F_1(x_1), \dots, F_n(x_n)\big)$. Many different families exist allowing for very different dependence structures. A copula can be imagined as a multivariate cumulative distribution function on the unit (hyper-) cube with uniform margins where its density reflects the strength of dependence between the margins. For further details we refer to the introductory book by Nelsen [57].

Sklar's Theorem is true for any dimension $d \geqslant 2$, but we will at first only consider bivariate copulas $C : [0,1]^2 \to [0,1]$. The density of a copula (denoted as c) expresses the strength of dependence which changes over the range of the marginal distributions. The only copula exhibiting a constant strength of dependence across its margins is the product copula Π describing independence. Commonly, strength of dependence in a bivariate setting is measured as a single correlation (or covariance) between two random variables and a Gaussian distribution is often implicitly assumed. As a Gaussian distribution can be decomposed into a Gaussian copula with Gaussian margins, the Gaussian dependence structure is implicitly imposed which is elliptically symmetric (following the notion of elliptical contours of the bivariate Gaussian distribution). Hence, by only investigating the correlation of two variables, one completely neglects the variation in the distribution of the strength of dependence over the range of the variables. Naturally, different copulas might reflect samples of an identical correlation, but their density might show a different pattern. The same applies to the spatial domain where kriging implicitly assumes a Gaussian dependence structure. However, looking into different data sets and investigating pairwise scatter plots reveals non-Gaussian dependence structures. These structures can be captured with copulas.

Incorporating distance as the only parameter but utilizing the flexibility of many bivariate copula families, we introduce *bivariate spatial copulas*. For pairs of locations in a local neighbourhood we assume that the separation distance of these is the driving parameter determining the dependence. Hence, pairs of locations very close to each other are likely to exhibit a dependence structure close to perfect dependence where noise might reduce the strength of dependence to some degree (analogous to the nugget effect in classical kriging). For large distances, the pairs will tend to be independent and are modelled by the product copula Π. The approaches by Bárdossy [3] and Kazianka and Pilz [53] allow only for a single multivariate copula family. The bivariate spatial copula $c_h(u, v)$ described here is de-

signed as a convex combination of bivariate copulas (in terms of their densities) that is not limited to a single family (see Eq. (6.1)). Hence, we allow not only for a varying strength of dependence but also for a changing dependence structure with distance:

$$c_h(u,v) := \begin{cases} c_h^{(1)}(u,v) & , 0 \leqslant h < l_1 \\ (1-\lambda_2)c_h^{(1)}(u,v) + \lambda_2 c_h^{(2)}(u,v) & , l_1 \leqslant h < l_2 \\ \vdots & \vdots \\ (1-\lambda_k)c_h^{(k-1)}(u,v) + \lambda_k \cdot 1 & , l_{k-1} \leqslant h < l_k \\ 1 & , l_k \leqslant h \end{cases} \tag{6.1}$$

where $\lambda_j := \frac{h-l_{j-1}}{l_j-l_{j-1}}$, h denotes the spatial separating distance of a pair of locations and $l_1, \ldots, l_k$ denote the representative distances of the spatial bins (e.g. midpoint or mean distance of all involved point pairs during the estimation). The parameters of the copulas $c_h^{(i)}$ in the convex combination may as well depend on the distance h. This allows for a smoothly changing strength of dependence and complete parametrization by distance. The arguments u and v are the values of the modelled pairs of locations transformed to the unit interval $(0,1)$ with the help of the marginal cumulative distribution functions or a rank order transformation. Inspecting Eq. (6.1) reveals that different choices of bins will in general yield different approximations to the underlying spatial dependence structure. The choice of the binning typically has to balance the two aspects of too little flexibility using few but well filled bins and too few observed pairs per bin using many bins achieving a high flexibility. Its important to ensure a reasonable number of data pairs per bin allowing for a sensible copula estimation.

Concentrating on a local neighbourhood of d neighbours, we now model the pair-wise dependence between locations through a bivariate spatial copula. However, these copulas need to be joined to benefit from the full d-dimensional distribution of the neighbourhood. A technique to combine bivariate copulas into multivariate copulas has been introduced by Aas et al. [2] building on work from Bedford and Cooke [8]. This approach has first been introduced as the *pair-copula construction* and the resulting copulas are now known as *vine copulas* in the literature.

Vine copulas allow to approximate multivariate copulas through bivariate building blocks (see Figure 6.1). The joint density is then obtained as the product of all involved bivariate copula densities. In the general case of spatial vine copulas, where we model the trees up to a certain level $1 \leqslant l \leqslant d$ through bivariate spatial copulas $c_{j,h(j,\cdot)}$ and the remaining ones $c_{j,j+i|0,\ldots,j-1}$ with a fixed parameter and family, we obtain:

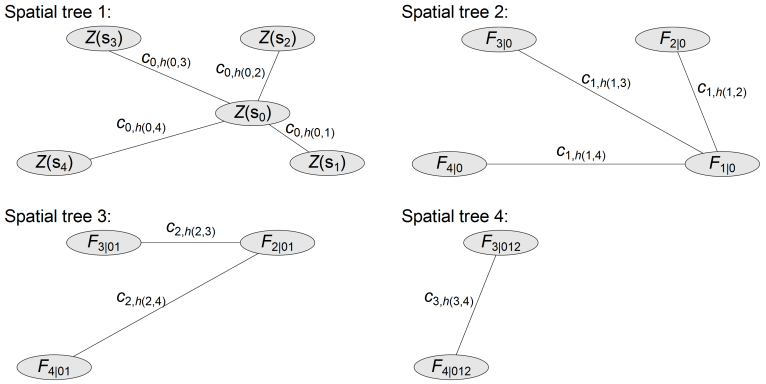

Figure 6.1: Graphical representation of a pure spatial vine copula of dimension 5. All trees are spatial trees capturing the dependence between the central location s_0 and its 4 nearest neighbours in ascending order $s_1, \dots, s_4$. The nodes in the vine represent spatial locations and the length of the connecting edge $h(j-1,k)$ in each tree j represents the distance between locations s_{j-i} and s_k parametrizing the bivariate spatial copula $c_{j-1,h(j-1,k)}$.

$$
\begin{aligned}
&c_{\mathbf{h}}(u_0, \dots, u_d) \\
&= \prod_{i=1}^{d} c_{0,h(0,i)}(u_0, u_i) \cdot \prod_{j=1}^{l-1} \prod_{i=1}^{d-j} c_{j,h(j,j+i)}(u_{j|0,\dots,j-1}, u_{j+i|0,\dots,j-1}) \\
&\cdot \prod_{j=l}^{d-1} \prod_{i=1}^{d-j} c_{j,j+i|0,\dots,j-1}(u_{j|0,\dots,j-1}, u_{j+i|0,\dots,j-1}) \qquad (6.2)
\end{aligned}
$$

where $u_i = F_i\big(Z(s_i)\big)$ for $0 \leqslant i \leqslant d$,

$$
\begin{aligned}
u_{j+i|0,\dots,j-1} &= F_{j-1,h(j-1,j+i)}(u_{j+i}|u_0, \dots, u_{j-1}) \\
&= \frac{\partial C_{j-1,h(j-1,j+i)}(u_{j-1|0,\dots,j-2}, u_{j+i|0,\dots,j-2})}{\partial u_{j-1|0,\dots,j-2}}
\end{aligned}
$$

with $1 \leqslant j < l$, $0 \leqslant i \leqslant d-j$ and for the non-spatially varying upper part of the vine

$$
\begin{aligned}
u_{j+i|0,\dots,j-1} &= F_{j+i|0,\dots,j-1}(u_{j+i}|u_0, \dots, u_{j-1}) \\
&= \frac{\partial C_{j-1,j+i|0,\dots,j-2}(u_{j-1|0,\dots,j-2}, u_{j+i|0,\dots,j-2})}{\partial u_{j-1|0,\dots,j-2}}.
\end{aligned}
$$

with $l \leqslant j < d$ and $0 \leqslant i \leqslant d-j$.

In general, different decompositions of a multivariate copula exist, referred to as regular vines, but in the spatial interpolation where a central element is naturally identified, we use a canonical vine where all initial dependencies are with respect to the central location. In each spatial tree $0 \leqslant j < l$ of the spatial vine (see Figure 6.1), all edges are modelled through a spatial copula $c_{j,h(j,k)}$ parametrized by

the spatial distance between the (conditioned) data pairs of the conditioning location s_j and a member of the neighbourhood s_k. Once a level is approached where the influence of the spatial distance vanishes, the consecutive upper trees might be modelled through spatially constant copulas. This spatially fixed upper vine structure does not impose any restriction on the bivariate copulas involved and are kept fixed no matter how the neighbourhood might be spatially organized. The conditional distribution functions involved in the above equations can immediately be obtained as partial derivatives of the already modelled copulas $C_{j-1,j+i|0,\dots,j-2}$.

To achieve a full distribution describing the local behaviour of the spatial random field Z, margins need to be fitted and joined with the spatial vine copula. Depending on the properties of the phenomenon to be modelled, one might use a single margin for all locations (in case the random field can be assumed to be stationary) or several margins incorporating some trend that is based for example on location, elevation or additional covariates. The density of the full distribution is obtained by multiplying the copula's density with the marginal densities and the variables are mapped to the copula scale through the cumulative distribution functions of the margins $F_0, \dots, F_d$:

$$f_{\mathbf{h}}(z_0, \dots, z_d) = \prod_{i=0}^{d} f_i(z_i) \cdot c_{\mathbf{h}}\big(F_0(z_0), \dots, F_d(z_d)\big) \tag{6.3}$$

where the z_i are representations of the random field $Z(s_i)$. Even though copulas allow to separately model the dependence structure and the margins of a distribution, a successful application requires good fits of both components.

6.3 SPATIAL VINE COPULA ESTIMATION

In the following, we introduce an estimation procedure for the spatial vine copula that borrows ideas from classical geostatistical approaches. A flow chart illustrating the estimation procedure of a spatial vine copula is shown in Figure 6.2. To estimate the first bivariate spatial copula, all spatial data is grouped into bins pairwise according to their spatial separation distance. Kendall's tau correlation measure is marginal independent and thus represents the correlation at the copula level. This makes it very useful in the application of copulas and some one-parameter copula families exhibit a one-to-one relationship between Kendall's tau and their parameter. The correlogram, using Kendall's tau, is calculated for the binned data. For each bin several copula families are fitted to the rank-order transformed data and the best fitting family (based on e.g. likelihood, AIC or BIC) is selected. When one restricts the set of copula families to those exhibiting a direct link between Kendall's tau and their parameter, one

might fit a function to the afore obtained empirical correlogram. Thus, the separating distance is used two-fold providing through Kendall's tau a parameter estimate for the copulas involved in the convex combination and tuning the weight λ. This way, the bivariate spatial copula will exactly reproduce Kendall's tau for any distance as modelled through the function from the correlogram. In case several best fitting families cannot be parametrized through Kendall's tau, one representative fit for each bin is obtained with a fixed parameter and combined as given in Eq. (6.1). Using these static representatives in the convex combination of copulas produces Kendall's tau values as a piecewise linear interpolation of the empirical values obtained in the correlogram. The reproduction of Kendall's tau through the bivariate spatial copula c_h can be seen from the fact that a copula's Kendall's tau value relies on the double integral of the copula [57, Theorem 5.1.3]. For any distance h, this integration of a convex combination results in a convex combination of two Kendall's tau values. This pair is either identical, in case the one-to-one relationship between Kendall's tau and the parameter is utilised or equals the corresponding bins' values resulting in the piecewise linear interpolation.

For further processing, the data needs to be grouped in neighbourhoods of central locations and their closest $\hat{d}$ neighbours. The size of these neighbourhoods depends on the dimension d of the spatial vine copula sought and the number of spatial trees. Iteratively, data pairs are selected from the neighbourhoods based on their spatial distance in the corresponding tree and re-arranged in spatial bins. In the case of $d = \hat{d}$ this reduces the pairs of locations for the last tree to the length of the sample that might be too short for a flexible binning. To increase the number of data pairs in the estimation of the consecutive spatial trees, it is beneficial to use a neighbourhood extending beyond the dimension of the spatial vine copula ($d \leqslant \hat{d}$) thus adding additional location pairs to each binning step for all trees. A rank-order transformation of these neighbourhoods generates a $\hat{d}+1$-dimensional dataset with uniform margins distributed on $(0,1)$.

The bivariate spatial copula $c_{0,h(0,\cdot)}$ on the first tree can now be used to derive the conditional sample of dimension $\hat{d}$ (conditioned to the value at the central location s_0) to which the remainder of the spatial vine is fitted (see Figure 6.1 and Figure 6.2). This conditional sample is again grouped into bins according to the spatial distance of the involved pairs (s_1, s_2), (s_1, s_3), $\ldots$, (s_1, s_d) and a second spatial copula $c_{1,h(1,\cdot)}$ is estimated. Conditioning the neighbourhood through $c_{1,h(1,\cdot)}$ on the values of the closest neighbours s_1 reduces its dimension again by 1. The procedure of rearranging the data into bins, estimating the next tree's spatial copula and conditioning the neighbourhood once more can be repeated until the desired level of spatial copulas is reached. Given that the vine needs to be completed, the (repeatedly) spatially conditioned neighbourhood is used as ini-

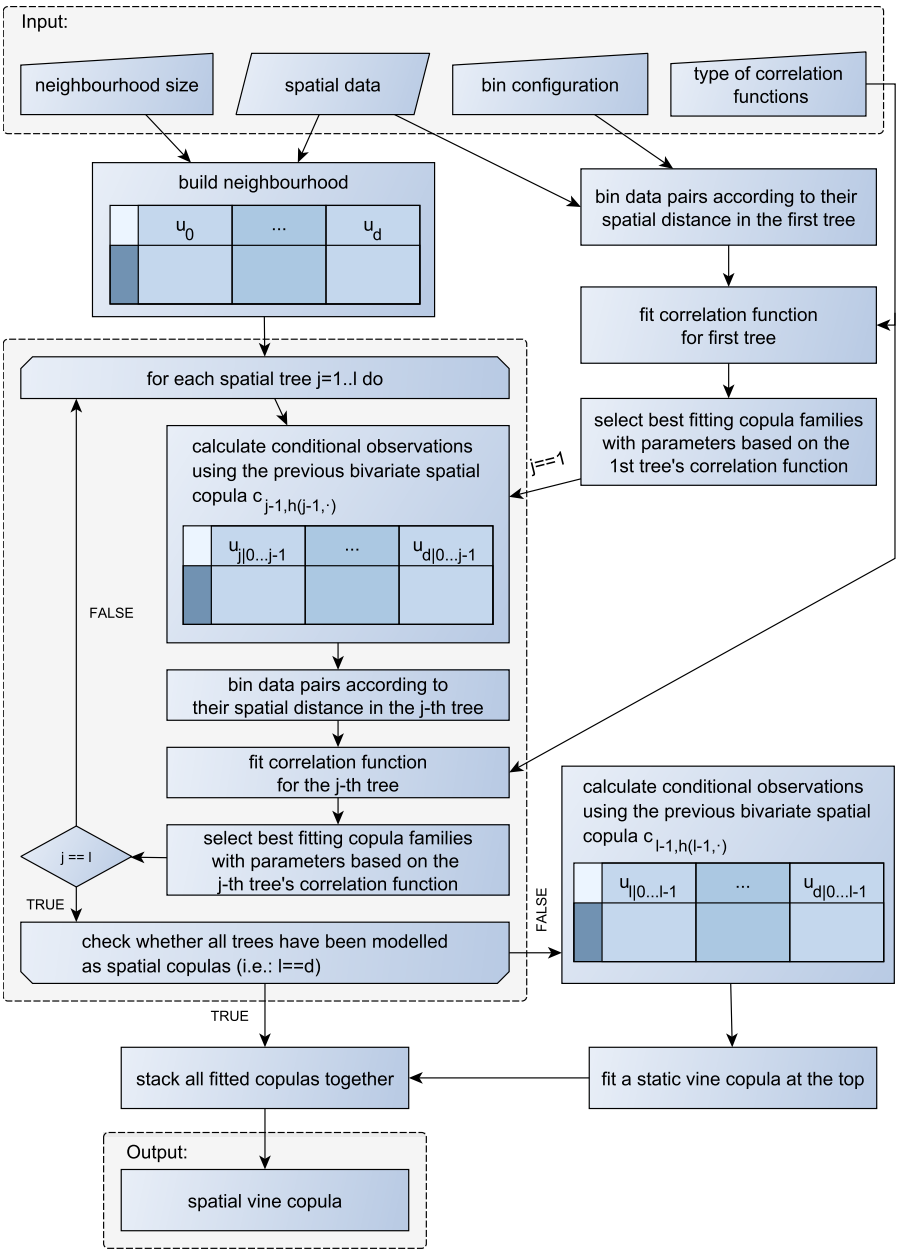

Figure 6.2: Flow chart of the estimation of a spatial vine copula.

tial data set for the spatially fixed upper canonical vine. The vine copula estimation proceeds sequentially by using the best fitting copula per bivariate pair (details are provided in Aas et al. [2], Czado, Schepsmeier, and Min [17] and Dissmann et al. [19]).

The joint copula density $c_{\mathbf{h}}$ can then be obtained through Eq. (6.2) where the first product reflects the first spatial tree. The remaining spatial trees are represented in the second product and the spatially constant trees appear in the third product. Fitting the marginal distributions, following generally any approach available in the literature, yields a full distribution through Eq. (6.3) describing the local behaviour of the random field Z.

6.4 PREDICTION AND SIMULATION OF THE SPATIAL RANDOM FIELD

The local representation of the random field Z can be used for different purposes. A typical task is prediction of the modelled phenomenon at unobserved locations. To produce such predictions from a local neighbourhood, every unobserved location needs to be grouped with its d nearest observed neighbours. Conditioning the $d+1$-dimensional copula $c_{\mathbf{h}}$ on the observed values, yields a 1-dimensional distribution of the phenomenon. This conditional distribution can then be used to calculate the expected value (see Eq. (6.4)), median or any other desired quantile (see Eq. (6.5)) denoting for instance confidence intervals. The predictors are given as:

$$\begin{aligned}\widehat{Z}_m(s_0) &= \int_{\mathbb{R}} z \cdot f_{\mathbf{h}}(z|z_1,\dots,z_d)\,dz \\ &= \int_{[0,1]} F_0^{-1}(u)\; c_{\mathbf{h}}\big(u|u_1,\dots,u_d\big)\,du \qquad (6.4)\\ \widehat{Z}_p(s_0) &= F_0^{-1}\big(C_{\mathbf{h}}^{-1}(p|u_1,\dots,u_d)\big) \qquad (6.5)\end{aligned}$$

where $u_i = F_i(z_i) = F_i\big(Z(s_i)\big)$ for $1 \leqslant i \leqslant d$ as before and $p \in (0,1)$ the desired fraction (e.g $p = 0.5$ to obtain the median). The equality for $\widehat{Z}_m$ is based on a probability integral transform. An advantage of this approach is that the conditional distribution describing the random field at the unobserved location may take any form. This is different from kriging, where every predictive distribution is again a normal distribution. This richer flexibility is supposed to provide better uncertainty estimates. Another advantage that is immediate from Eq. (6.4) and Eq. (6.5) is that the only information on the marginals needed is their quantile function. This allows for instance to use approximations derived from the empirical cumulative distribution function without the knowledge of any explicitly known form of the family's density. However, the empirical cumulative distribution func-

tion is typically limited to the domain defined by the smallest and largest observation.

For simulation purposes based on a local distribution only, we suggest a sequential simulation algorithm proceeding along a random path [48, 29]. At first, d+1 locations of the target geometry are selected. For these initial locations a complete sample is drawn from the spatial vine copula. In the following, further locations are randomly selected one by one and the univariate conditional copula density $c_h(u|u_1,\dots,u_d)$ based on the d nearest neighbours is obtained following the notation of Eq. (6.4). The estimate is then drawn from this conditional distribution. Inserting this spatial sample on copula scale into the marginal quantile functions yields a simulation on the original scale. Repeated iterations of this procedure produce several realisations of the modelled spatial random field Z. Conditional simulations can be obtained by introducing the observed values as additional samples that might be included in the conditioning neighbourhoods. This can be seen as starting the simulation at a point where a couple of simulations have already be drawn. A modellers choice is to decide to which degree conditioning variables are preferred over spatially closer already simulated values.

6.5 APPLICATION

As the advantage of this new approach is presumed to lie in the modelling of skewed spatial random fields exhibiting extremes that do not follow a Gaussian dependence structure, we will apply it to the emergency scenario data set from the Spatial Interpolation Comparison in 2004 [SIC2004: 20]. This simulated data set was generated from a dispersion process mimicking an accidental release of radioactivity. Following the idea of the spatial interpolation comparison, we apply the spatial vine interpolation procedures as well to the routine data set that only reports low background values. Additionally, we will draw simulations from the modelled spatial random field using the spatial vine copula. All calculations were made using R 3.0.2 [65] and can be reproduced by using the publicly available spcopula[1] R-package.

Interpolation of the Emergency Scenario data set

The data consists of 1008 simulated values of which 200 are provided for model fitting and 808 are held back for prediction validation. The area roughly extends to 350 km in east-west and to 700 km in north-south direction. Figure 6.3 shows a histogram of the emergency scenario training data set. The routine data reports only small background radiations up to 153 nSv/h that as well compose the majority of the emergency scenario. A more detailed description of the data

1 available from r-forge.r-project.org/projects/spcopula/

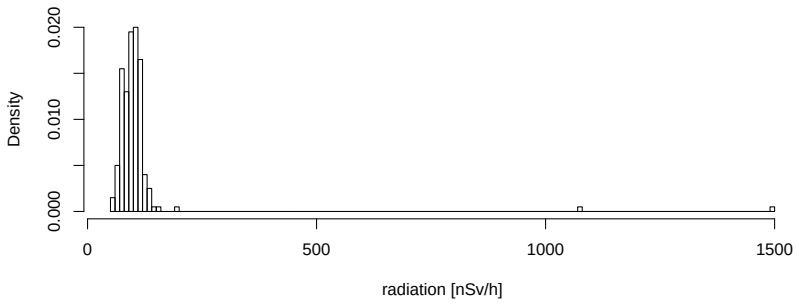

Figure 6.3: Histogram of the emergency scenario training data set showing relative frequency. Note that the two columns above 1000 nSv/h only represent a single value each.

set can be found in Dubois and Galmarini [20]. The data set is freely available and can for instance be obtained through the R-package gstat [60].

In a preliminary step, a simple inverse distance weighted interpolation suggests to fit a trend surface to the background data. A linear regression using only the non-extreme data points on the (squared) coordinates x, y and y^2 provides a reasonable model with an adjusted R^2 of 0.41 and coefficients being significant at a level of 0.9. The data set consisting of this model's residuals will be used subsequently.

In the following, we follow the repeated binning, spatial copula estimation and conditioning routine as described in Section 6.3 and illustrated in Figure 6.2. The set of families investigated includes the elliptical Gaussian and Student copulas, the Archimedean Clayton, Frank, Gumbel [57] and Joe [47] copulas and the survival versions of the latter ones as well as a copula exhibiting cubic and quadratic sections (CQS copula) representing only weaker dependencies [57, Example 3.16]. All these copula families exhibit a positive dependence and can be parametrized by Kendall's tau. In case of the two-parameter CQS copula family, the second parameter appears to be constant over space and is fixed at its mean value. Figure 6.4 shows the graphical representation of the 4 bivariate spatial copulas for the emergency scenario and the routine scenario data sets. The estimation of $c_{0,h(0,\cdot)}$ starts with the entire data set being grouped into spatial bins with bounds at 0 km, 20 km, 30 km, ..., 100 km maintaining at least 100 pairs of locations per bin. The estimation of the consecutive bivariate spatial copulas $c_{1,h(1,\cdot)}$, $c_{2,h(2,\cdot)}$ and $c_{3,h(3,\cdot)}$ to be used in the 2^{nd}, 3^{rd} and 4^{th} spatial tree are initially based on the 9-dimensional neighbourhoods around the central locations. Thus, each step relies on 200 tuples of (conditional) data being repeatedly rearranged into bins with equally spaced boundaries and filled with roughly 170 pairs of locations each. For each tree, a copula of every family is fitted for each bin with the help of the one-to-one relationship to Kendall's tau and its log-likelihood is evaluated. The family with the highest log-likelihood per bin is selected and associated with the mean distance of all pairs

of locations of that bin. The set of families and distances determines a bivariate spatial copula per tree. Extending the neighbourhood to $1 + 9$ locations increases the number of conditional pairs that can be used to estimate the bivariate spatial copulas at higher trees. Thus a 5-dimensional pure spatial vine copula can be estimated. Additionally, static vine copulas are fitted to the conditioned 5-dimensional neighbourhood using only the first spatial tree and to the 10-dimensional neighbourhood using all 4 spatial trees. Thus, we fit three different spatial vine copulas addressing the models behaviour in terms of neighbourhood size and spatial truncation level l. To illustrate the difference to the Gaussian dependence structure, a spatial Gaussian copula is fitted based on the correlation function of the first tree. Using only the first tree's correlation function is possible as the spatial Gaussian vine is already completely defined through the spatial correlation matrix given by this unconditioned correlation function [14].

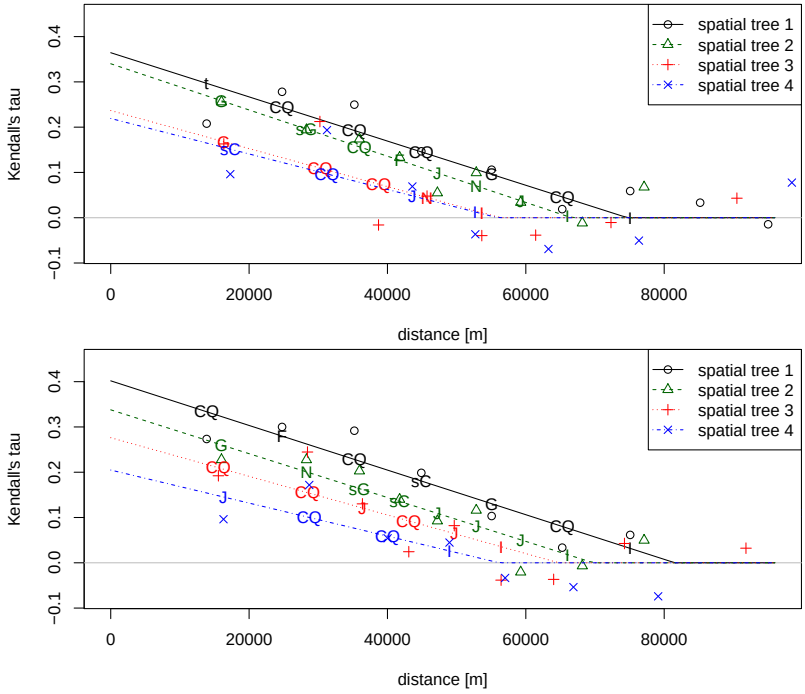

Figure 6.4: Bivariate spatial copulas used in the spatial vine copulas for the emergency scenario (upper panel) and the routine scenario (lower panel). Lines describe the modelled correlation function while symbols represent the empirical values from the binning. Letters denote the chosen copula family: Gaussian (N), student (t), Clayton (C), Frank (F), Gumbel (G), Joe (J), survival Clayton (sC), survival Gumbel (sG) survival Joe (sJ) and one copula based on cubic-quadratic sections (CQ). The product copula (I) representing independence is used for any distance larger than its first appearance.

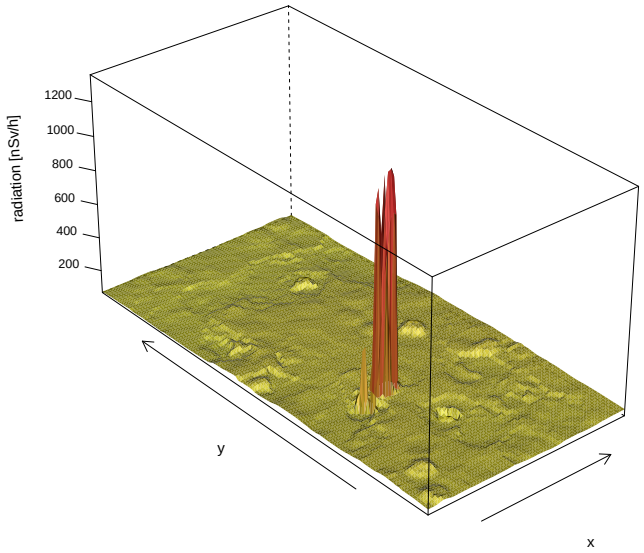

Figure 6.5: Surface plot of the median predictor of the pure spatial vine copula with empirical marginal distribution function on a grid spanning the study area.

A few model assumptions are made during the estimation process of the different bivariate spatial copulas. The functions relating spatial distance with the value of Kendall's tau were assumed to be linear and constantly 0 once they hit the x-axis. However, any other function describing this relationship is in general possible. As a consequence, all values of Kendall's tau are non-negative and only copula families being capable of representing non-negative correlations and having a one-to-one relationship between Kendall's tau and their parameter have been considered. The copula being defined through its cubic-quadratic sections is in general a two parameter family. However, the second parameter remained mainly constant over the spatial domain and was fixed at the mean value for each bivariate spatial copula. This leads to the desired relationship between Kendall's tau and the first parameter for this copula as well.

For the comparison study, we compose the above described different spatial vine copulas each for the routine and the emergency scenario. In summary, two are designed for a neighbourhood of 1+4 locations where one uses only 1 spatial tree and the other is composed completely out of spatial trees (denoted "pure"). For a neighbourhood of 1+9 locations, we investigate the performance for a spatial vine copula using 4 spatial trees (the same as for "pure") followed by a 6-dimensional spatially fixed canonical vine copula at the top. All approaches use the same two types of marginal quantile functions. One type of margin is purely empirical and defined as piecewise linear approximation of the inverse of the empirical cumulative distribution function extended for the ranges of the 99.5%-percentile and 0.5%-percentile to the top and bottom respectively. The second type

consists of parametric distributions where in the routine scenario a Gaussian distribution and in the emergency scenario a convex combination of three uniform distributions and a generalized extreme value distribution is used (denoted "PoT"). An interpolation of a grid spanning the study area using the median predictor of the pure spatial vine copula and the empirical marginal distribution function is shown in Figure 6.5.

For each of the spatial vine copulas and the spatial Gaussian copula, we perform the prediction at the 808 locations by predicting the median following Eq. (6.5). For comparison against classical geostatistical approaches, we perform residual kriging after the trend surface has been subtracted and trans-Gaussian kriging on the original data using the log-transform. The calculations are done using gstat [60] while the variograms are fitted using routines provided by the automap package [43]. To evaluate the performance of the procedures, the mean-absolute error (MAE), root-mean-squared error (RMSE), mean error (ME) and correlation (COR) are calculated between the predicted values and the provided simulated concentrations. All results for the emergency scenario and the routine scenario are summarized in Table 6.1.

Table 6.1: Results for the emergency (upper half) and routine (lower half) scenario. The values of the approach denoted "Kazianka" are taken from Kazianka and Pilz [54] where the same data set has been studied. Note that a comparison of log-likelihoods is only possible for the same neighbourhood size. The best performance is indicated in bold.

copula/approach	dim.	margin	log-lik.	MAE	RMSE	ME	COR
pure	5	emp.	118	**14.6**	68.0	-6.2	0.59
4 spatial trees	10	emp.	201	15.1	68.7	-5.6	0.58
1 spatial tree	5	emp.	114	16.0	74.7	-5.1	0.49
Gaussian	5	emp.	70	16.2	80.6	-7.7	0.32
pure	5	PoT	118	16.3	78.8	-8.0	0.42
4 spatial trees	10	PoT	201	16.7	79.1	-8.0	0.39
1 spatial tree	5	PoT	114	16.4	78.3	-7.7	0.42
Gaussian	5	PoT	70	16.2	80.6	-7.8	0.31
Kazianka	15	GEV		16.2	**65.9**	-2.6	**0.71**
TG log-kriging	200			20.8	78.2	**-2.1**	0.39
resid. Kriging	200			21.1	75.6	5.2	0.43
pure	5	emp.	116	**9.2**	12.7	-1.2	0.78
4 spatial trees	10	emp.	211	9.5	13.1	-0.8	0.76
1 spatial tree	5	emp.	114	9.3	12.8	-1.1	0.78
Gaussian	5	emp.	82	**9.2**	12.6	-1.3	0.78
pure	5	Gauss.	116	**9.2**	12.7	-1.3	0.78
4 spatial trees	10	Gauss.	211	9.6	13.2	-0.8	0.75
1 spatial tree	5	Gauss.	114	9.4	12.9	-1.2	0.77
Gaussian	5	Gauss.	82	**9.2**	12.6	-1.3	0.78
TG log-kriging	200			**9.2**	**12.5**	-1.3	**0.79**
resid. Kriging	200			9.3	12.9	**-0.4**	0.76

Simulation of the emergency scenario data set

As described in Section 6.4, we simulate from the spatial vine copula using a sequential simulation along a random path. As the small neighbourhood of only four neighbouring locations might lead to unwanted results when clusters emerge from the random path, we adopted a multiple grid strategy [28]. The target resolution of the grid is approached step-wise starting with a coarse representation and adding finer grids once all grid points have been simulated. One realisation conditioned on the 200 emergency scenario measurements is drawn from the 5-dimensional pure spatial vine and shown in Figure 6.6.

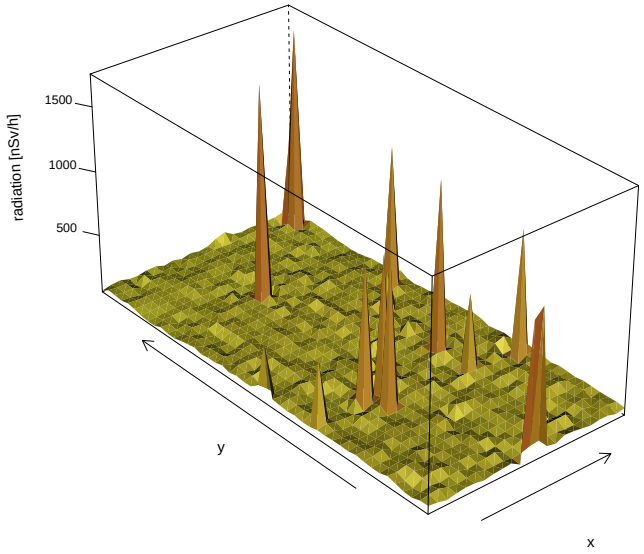

Figure 6.6: Surface plot of a conditional simulation of the pure spatial vine copula on a coarse grid spanning the study area.

6.6 RESULTS AND DISCUSSION

Besides the cross validation results presented in Table 6.1, it is interesting to observe how well the overall distribution is represented. This is illustrated on log-scale in Figure 6.7 where a box-plot for each method is shown along with the provided 808 data points for validation in the emergency scenario. It is immediate that the trans-Gaussian kriging procedure fails to reproduce the target distribution while residual kriging performs a bit better. Just as kriging, the spatial Gaussian copula does not produce any extreme estimates and merely represents the background radiations. The spatial vine copula approaches using the empirical marginal distribution function produce too few extreme values, but within the correct range of the extremes. The predictions based on the theoretical marginal distribution re-produce a moderate heavy right tail, but fail to capture the large extremes. Counting

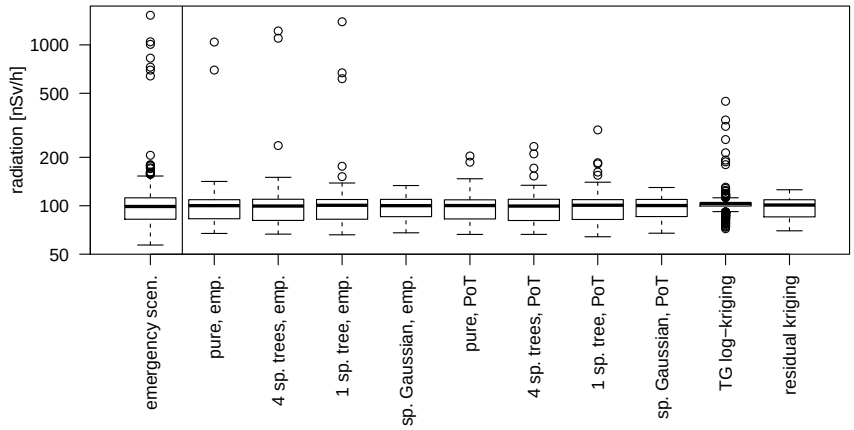

Figure 6.7: Box-plots (on log-scale) illustrating the overall distribution of the predicted values for the different interpolation approaches along with the provided data set in the emergency application. The y-axis is shown in log-scale. The approaches are ordered as in Table 6.1.

the number of values below the predicted median yields a fraction of about 0.5 for all different copulas supporting this approach. In the routine data case, the box-plots (not shown) are less heterogeneous and all approaches capture the overall marginal distribution rather well.

The same split as in the box-plots (Figure 6.7) between the marginal distributions is apparent from Table 6.1. Within each group of margins, the spatial Gaussian copula performs slightly worse than the spatial vine copulas. Using the empirical margin that better represents the overall distribution, this difference is pronounced. Hence, the Gaussian copula family fails in capturing the dependencies of this heavily skewed spatial random field. It does not give fractions large enough to produce extreme quantiles even though the marginal distribution would allow for it. However, the spatial vine copulas only slightly improve the prediction if the marginal distribution function is not capable of reproducing the sample. This stresses the importance of both, a good match in dependence structure and in the marginal distribution function. According to the MAE in Table 6.1, using more spatial trees improves the prediction. Further improvement of the predictions might be possible for different choices of marginal distributions. An obvious limitation of the empirical distribution function is that the largest value of the 200 records of training data is smaller than the maximum of the 808 target values. Hence, the true extreme values will never be captured using this limited margin. Deeper insights in the process might lead to better marginal distribution functions. However, it will remain challenging to estimate a heavily skewed marginal distribution with only 2 extreme values out of 200 observations.

The best approach reported in Dubois and Galmarini [21, Table 4] by Timonin and Savelieva [89] is based on techniques including neural networks and achieves in the emergency scenario an MAE of 14.9 that is slightly worse than the best spatial vine copula prediction. However, our approaches could not meet the other indicators. Except for the ME, our best performing approach ranks at least within the top 5 of the listed SIC2004 participants and the copula interpolation approach described by Kazianka and Pilz [54]. Taking as well the performance in the routine scenario into account, the median predictor using the pure spatial vine copula with empirical marginal distribution function slightly outperforms the approach by Timonin and Savelieva [89]. However, all approaches are rather close to each other in the routine scenario and hardly distinguishable from their performance indicators.

Another important aspect is the ability of this approach to flexibly represent the uncertainty in form of a conditional distribution. Different than for the kriging predictor, the conditional distribution may take any form of probability distribution and not only the Gaussian one. In Figure 6.8 the cumulative distribution functions (cdf) of the conditional distributions based on the median predictor of the pure spatial vine copula with empirical marginal distribution function and the residual kriging predictor are compared for two stations in the emergency scenario. It is immediate that the confidence bands in both scenarios differ strongly. The ranges of a confidence band may be both, larger or smaller, for the spatial vine copula than for the kriging approach across the data set. Due to the flexibility to choose any marginal distribution, the copula based approach allows to ensure that its confidence bands do not include any unreasonable values (e.g. negative radiation or concentration) opposed to classical kriging confidence bands that are always symmetric around the predicted mean. This can as well be seen from Figure 6.8 where both conditional distributions based on the kriging predictor show (some) probability for negative values. The kriging variance is known to be only dependent on the spatial configuration of the locations. This is different from the spatial vine copula approach where the separating distance and the magnitude of the values influence the conditional predictive distribution. Both uncertainty estimates illustrate how uncertain the model is about its prediction, but neglect uncertainties associated to the model selection and parameter estimation.

In the neighbourhood building step for fitting and prediction, the neighbours are selected by distance only. Given a prominent influence acting on the spatial random field introducing anisotropic dependencies between locations, a more complex neighbourhood selection is likely to be advantageous. Hence, one might want to arrange the neighbourhoods for a 5-dimensional spatial vine copula in such a way that the closest neighbours are selected per (rotated or sheared) quad-

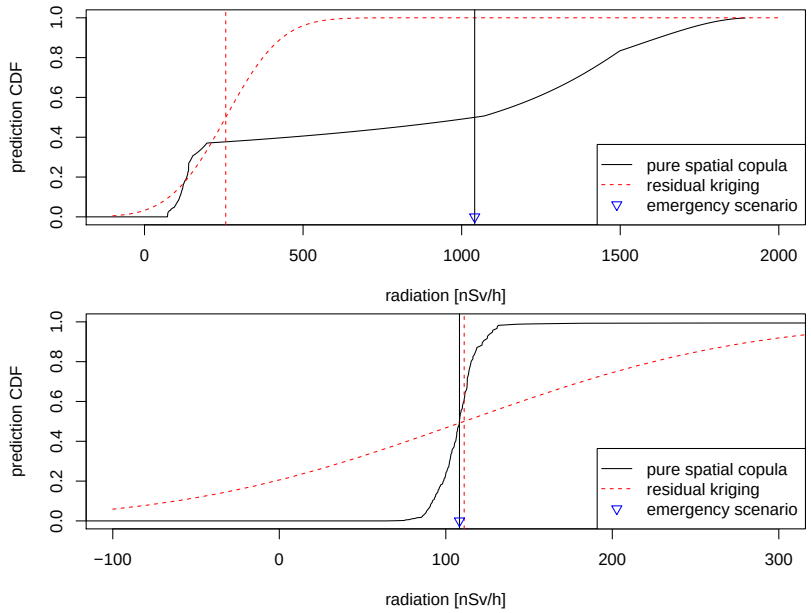

Figure 6.8: The predictive cumulative distribution functions for the median predictor of the pure spatial vine copula with empirical marginal distribution function and the residual kriging predictor at an extreme location (upper panel) and at a random background station (lower panel).

rant centred at s_0 and let the i^{th} neighbour always be selected from the i^{th} quadrant. For higher dimensional spatial vine copulas one might split the plane into multiple sectors or select multiple neighbours per quadrant/sector.

The spatial vine copula used in this paper and illustrated in Figure 6.1 is only based on spatial dependencies of a single variable. However, additional covariates might be introduced through additional edges in the first tree. These additional edges would then be separately modelled by the best fitting bivariate copula. The sequential structure of the vine would then grow including additional trees. Besides including spatial trees only up to a certain level, one could as well truncate the upper vine copula at a certain level using a truncated vine [11]. This would speed up the computational time during the estimation phase as well as the prediction phase and might be helpful in modelling the dependence of larger neighbourhoods.

Spatial vine copulas that are not only composed out of elliptical copulas (i.e. Gaussian and t copula) introduce a side effect worth noting. In the case of linear geostatistics, the correlation between every pair of locations in a neighbourhood is given by the correlation function (or e.g. variogram model). In a spatial vine copula, only the explicitly modelled edges from the first tree typically follow the first tree's correlation function. All other pairs of locations exhibit correlations that are implicitly modelled through the upper trees and their correlation functions. No closed form exists that describes the rela-

tionship between the correlations in the upper vine trees and the correlation matrix of the neighbourhood except for the elliptical copula families [14]. This leads to the effect that in the local approximation of the spatial random field a pair that does not model any edge of the first spatial trees in two different neighbourhoods might receive different correlation values. This represents the design of a canonical vine with the focus on a single vertex and the contributions of the surrounding vertices to it. The effect of this property on the spatial random field needs to be further investigated.

Methods based on copulas typically increase the computational cost compared to standard approaches such as kriging. Modelling the spatial random field only locally reduces the computational burden and produces predictions outperforming the classical approaches using all available data for the investigated data set. In the presented application, the estimation of the copulas based on 200 locations as well as the prediction at the 808 unobserved locations is done in a few minutes on a common laptop.

The one realisation from the spatial vine copula shown in Figure 6.6 shows many more extremes than the one used to condition the simulation. Repeated simulations show that the additional extremes jump over the study area while the true extreme in the lower left quadrant is frequently reproduced. The rather weak spatial dependence and comparatively small spatial neighbourhood leads for some locations to an almost independent sampling from the marginal distribution. This explains the additional extremes scattered across the study area.

6.7 CONCLUSION

The presented spatial vine copula approach extends that of Gräler and Pebesma [33] by using more than one spatial tree at the foundation of the vine. The additional spatial trees add valuable information on the dependence of the higher order neighbours leading to an improved model of the skewed spatial random field. However, further research is needed to develop strategies to select copula families and to fit functions modelling Kendall's tau in terms of separating distance. Additional model constraints need to be explored to improve the model describing the spatial random field. Negative correlations found in bins were considered to be due to noise in the data in this study. Nevertheless, it needs to be investigated in which scenarios negative correlations may improve the model and how they relate to non-elliptical copulas producing inconsistent spatial correlations for only implicitly modelled pairs of locations.

The introduced spatial vine copula achieves good results in the prediction validation in comparison to the other methods applied to the emergency and routine scenario data sets (Table 6.1). Hence, the spatial vine copula only predicts extremes where the data is heavily

skewed and reproduces the marginal distribution very well compared to residual kriging (Figure 6.7). A simpler copula based approach relying on a spatial Gaussian copula is shown to fail in capturing the dependencies of this heavily skewed spatial random field that are successfully modelled by the spatial vine copula. The illustrated flexibility of the conditional marginal distributions describing the uncertainty adds to the value of this new approach.

A spatio-temporal extension of the spatial vine copula with a single spatio-temporal tree has been presented in Gräler and Pebesma [32]. The extensions made to the spatial case in this paper are assumed to be applicable to the spatio-temporal setting, to improve the interpolation of spatio-temporal data. The directional property of the temporal domain is likely to introduce asymmetric dependencies, which can easily be modelled by asymmetric copulas in the bivariate spatio-temporal copulas. This is an advantageous feature of the copula approach in modelling spatio-temporal data.

Acknowledgements

The helpful feedback provided by two anonymous reviewers has thankfully be included. This research has been funded by the German Research Foundation (DFG) under project PE 1632/4-1.

7

SYNTHESIS

This chapter summarizes the results obtained in chapters 2 to 6 with respect to the objectives raised in Section 1.2. A general discussion of the developed methods follows in the sequel.

7.1 SUMMARIZED RESULTS

How can vine copulas contribute to the modelling of extremes in time series of environmental phenomena?

The study conducted in Chapter 2 illustrates how the choice of a copula model affects the representation of the observed phenomenon. Besides different copula models, the implications of different multivariate return period definitions in the multivariate space are studied. A simulation study for a 10-year return period on the 3-variate annual maxima rainfall time series shows that the theoretically motivated differences are significant. Vine copulas allow to flexibly compose the 3-dimensional distribution out of two survival BB7 copulas and a student copula. The selected marginal distributions are the Weibull and twice the Exponential distribution. This composed distribution allows the simultaneous calculation of annual maximum peak discharge and associated duration and volume characterising an annual extreme rainfall event with a multivariate 10-year OR return period. The components of this event differ considerably from the respective values derived from the lower dimensional approaches. Complete and flexible probabilistic models of multivariate phenomenon as the model obtained in the presented study are a contribution of vine copulas to the modelling of extremes in time series.

How can bivariate spatial/spatio-temporal copulas be connected in a vine copula to model spatial/spatio-temporal random fields?

The answer to this question stretches over several chapters. The single tree spatial vine copula has first been presented in Chapter 3 where the first tree of a vine copula is composed out of a convex combination of bivariate copulas parametrized by distance. This initial design of a spatial vine copula is applied to zinc concentrations along the river bank of a stretch of the Meuse river. A local neighbourhood approach of four neighbours (hence a five-dimensional spatial vine copula) is used to interpolate a regular grid. Based on a leave-one-out cross validation, the root mean squared error appeared to be slightly

larger than for an ordinary kriging approach, but the bias could be reduced. However, a comparison of the log-likelihood values for multivariate single family copulas and the spatial vine copula indicate a better fit of the spatial vine copula.

The concept of a single tree spatial vine copula is extended to a single tree spatio-temporal vine copula in Chapter 4. The extension to the spatio-temporal domain is obtained by fitting spatial bivariate copulas to different time lags and combining these in a convex manner controlled by temporal distances. This spatio-temporal vine copula is applied to an European air quality data set where rural daily mean PM_{10} concentrations are given for the entire year 2005. A static neighbourhood selection is based on the three spatially closest neighbours over three time instances. The cross-validation statistics are of the same order of magnitude as for a metric spatio-temporal kriging approach from an earlier study [31]. A closer look at individual measurement stations reveals that the spatio-temporal vine copula approach improves the interpolation for certain stations, but has considerable drawbacks compared to the spatio-temporal kriging approach at other stations.

A further extension of the single tree spatial vine copulas has been developed in Chapter 6 where multiple trees of the vine are equipped with bivariate spatial copulas. The spatial distances incorporated for the higher order trees are revealed from the distances between the locations of the conditioned pairs (i.e. the copula $c_{1h(1,2)}(u_{1|0}, u_{2|0})$ is parameterized by the distance between locations s_1 and s_2, compare Figure 6.1). The multiple tree spatial vine copula is applied to a widely studied simulated data set of nuclear radiation that exhibits a high peak surrounded by small background measurements. Following the same rules as the original interpolation comparison carried out with this data set, the best spatial vine copula approach scored within the top five of all approaches. The research conducted in Chapter 6 reveals that the number of spatially enabled trees in the spatial vine copula plays a more important role for the prediction quality than the number of neighbours (compare Table 6.1). Additionally, attention is drawn to the fact that the marginal distribution function has a strong influence on the overall model. Interpolating only the background values shows hardly any differences between the approaches. A striking feature of the spatial vine copula is the conditional prediction distribution that may vary for each location in its shape and is not bound to any distribution family. Hence, confidence intervals depend on the values of the local neighbourhood and their dependence structure. Values limiting the confidence interval are only those possibly taken by the marginal distribution function. This supports the assumption that the spatial vine copulas provide a more realistic picture of associated uncertainties than for instance the kriging variance.

How can covariates be included in spatial or spatio-temporal vine copulas?

The lack of flexibility in the neighbourhood selection motivated the research conducted in Chapter 5 along with the aim to include covariates in the vine copula approach. Earlier results suggest that a flexible marginal distribution improves the overall model. These aspects have been the focus of Chapter 5 and the extended spatio-temporal copula has again been applied to the European air quality data set. Different from the first application, the marginal distribution now has varying parameters across the study area based on an inverse distance weighting of the parameters at the measurement stations solely and in conjunction with a linear model. An improvement could as well be made in the neighbourhood selection that is now based on the most correlated 9 neighbours across space and time. The cross validation statistics show that the varying marginal distributions improve the prediction. The strongest improvement could be achieved with the more complex model including a linear model. However, an even better understanding of the marginal distribution would again considerably improve the prediction. This highlights once more the importance of both, a good model of dependence and of the marginal distributions. It has been shown that covariates can easily be integrated as additional edges in the first tree allowing for an even richer model of the spatio-temporal vine copula.

7.2 GENERAL DISCUSSION

While vine copulas are in general very flexible, in Chapter 2, an obvious boundary condition exists (compare Figure 2.4), that could not be captured with the used copula families. This limitation affects the quality of the model. A trial to design a copula tailored to this data set has been put forward, but its copula properties could not yet be verified. During the open review process of the corresponding paper, a reviewer briefly illustrated a different approach to accomplish models reproducing the observed boundary. However, this approach was beyond the scope of the paper at that time and has now been published by the reviewer [83]. Nevertheless, the vine copula could be used to build a reasonable model of the annual extreme triples of peak discharge, duration and volume of the studied area.

The design of the bivariate spatial copulas as convex combinations of copulas parametrized by distance has proven useful throughout the different applications. However, the selection of copula families depends on the layout of the spatial and temporal binning. The shape of the correlation function remains rather unaltered with changing bins. The switch between families is not too surprising, as they form non-disjoint subsets of the possible dependence structure. For instance, all of the studied copulas do contain the independence copula Π

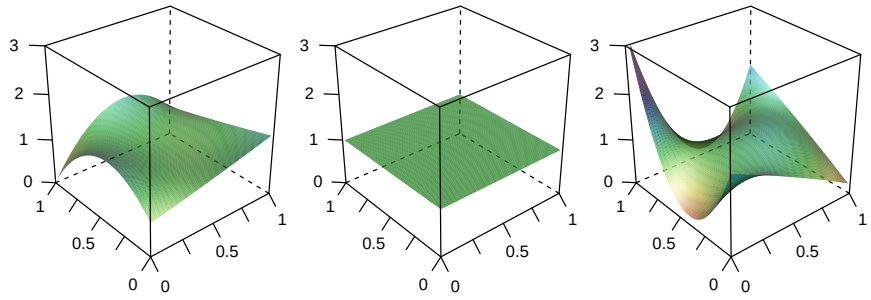

Figure 7.1: Different members of the asymmetric copula family all having a Kendall's tau value of zero but differently shaped densities. The second plot corresponds to the density of the independence copula Π.

(at least as a limit case). Therefore, the copula selection for larger distances with rather weak correlation is less stable in some cases. However, the single tree spatio-temporal vine from Chapter 4 (see Table 4.1) is in favour of an asymmetric copula for larger spatio-temporal distances. This copula family only models rather small values of Kendall's tau, but does have a dependence structure that is considerably different from the dependence structure of the independence copula even for a correlation of zero. Figure 7.1 shows three different members of the asymmetric copula family all having a Kendall's tau value of zero. Hence, it might be beneficial to fit a copula family even for weak correlations instead of rigorously preassigning a single family. Obviously, it will depend on the spatial layout of the locations and the size of the local neighbourhood of each application scenario whether large distances are at all considered for an interpolation.

A precondition that is essential for the application of the highly flexible spatial and spatio-temporal vine copulas is a sufficiently sized data set. Otherwise, a model that reflects artefacts of the sample is likely obtained and resulting in overfitting. In terms of the correlation function that relates distances with strength of dependence, a set of well suited functions should further be explored. In general, these functions can be chosen without any limitations, but assumptions on the process might lead to a reduced set of possibilities being more robust against overfitting. The motivation for the well defined variogram functions in kriging is different, as the covariance matrices for the underlying multivariate normal distribution need to be valid. However, it might be beneficial to explore and build a set of plausible correlation functions that ease the application of spatial and spatio-temporal vine copulas, but might be flexibly extended as needed.

In the current applications, the correlation functions are constrained to the range $[0, 1]$ that might in general be extended to $[-1, 1]$ if required by the process (e.g. sea level where high tides are negatively correlated with low tides a couple of hours later). The extension of

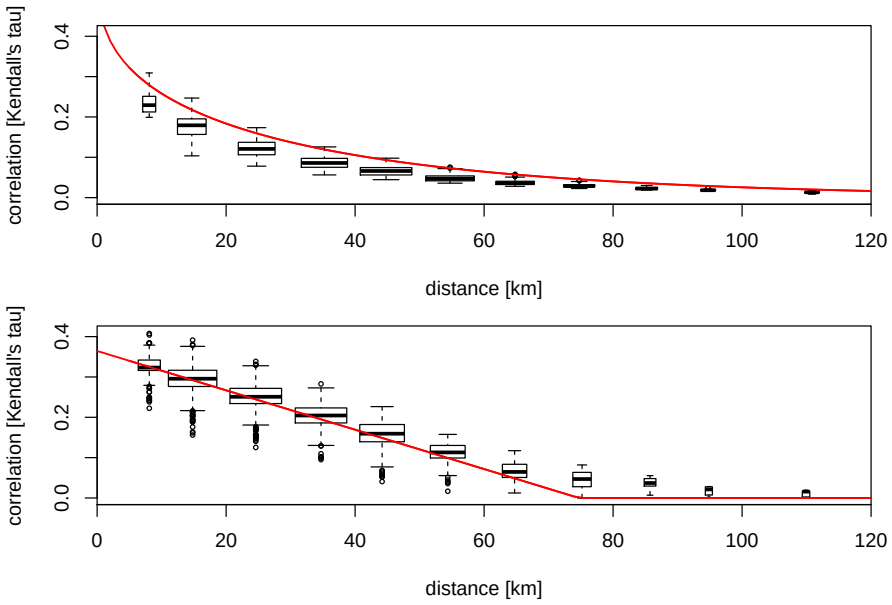

Figure 7.2: Top: Analytically evaluated correlations of the second tree of a pure Gaussian spatial vine copula together with the modelled correlation of the first spatial tree (red curve).
Bottom: Spread of the implicitly modelled correlation between locations for each neighbourhood. The underlying spatial vine copula is the multiple tree spatial vine from Chapter 6 (red curve) but with Gaussian copulas only.
Correlations have been binned every 10 km into boxplots to improve readability. The width of the boxes indicate the number of pairs that fall in the corresponding spatial bin.

the spatial bivariate copulas from the first tree to the upper trees (Chapter 6) might benefit from a correlation function that does not only take into account the distance of the conditioned pair, but also previous distances. The top plot of Figure 7.2 shows that per spatial distance the actual correlations of the conditioned observations lie within a range of values. The shown correlations have been derived from a spatial vine copula with only Gaussian copulas using a Matern covariance function. For this type of model, it is possible to analytically evaluate the correlations for pairs of locations throughout the vine [14]. Extending the correlation curve to a surface might tackle this issue on higher order trees and improve the model, but would yet need more data. Linear models composed out of powers of the direct distance between the conditioned pair and one of the distances between pairs of the first tree explain the second trees correlation up to an adjusted R^2 of 0.9 in the above simulation study.

The above simulation started of with a well defined variogram that ensures a well formed covariance matrix of each local neighbourhood. In particular, distance completely determines the pair-wise correlation between locations. A second simulation study starts with a multiple tree spatial vine copula using only Gaussian copulas. The condi-

tional correlations are then transformed back to a covariance matrix following again Cooke, Joe, and Aas [14]. The covariance matrix is by construction well defined for each neighbourhood, but plotting correlations against distances reveals a non functional relationship and considerable spread (Figure 7.2 bottom plot). This is partly attributed to the design of vines where the focus is on the influence of the neighbours on the central locations, modelling the correlation between the neighbours only implicitly.

An issue that has not been addressed in this thesis is that of formal goodness of fit testing. In some applications, the fit has been assessed based on the likelihood of the fitted copula model. The prediction quality is evaluated using cross-validation techniques. A goodness of fit test could for instance be accomplished by bootstrap techniques but further research would be needed to identify a useful test statistics to compare different simulated random fields. A different approach focusing on the spatial component could be based on a bootstrap per bin, where the goodness of fit boils down to the bivariate copula case. Nevertheless, the full vine structure would be neglected in this case. Additional research will be necessary to further investigate how the model fit can be assessed in greater detail.

Due to the larger complexity and flexibility, the computational cost is considerably higher than for the classical kriging approaches. The source of these costs lies in numerical evaluation of one-dimensional integrals during the prediction phase and at times computationally costly copula densities. No formal evaluation of the algorithmic complexity has been carried out, but rough timings were taken during the different applications. The calculations have been made on common personal computers without any optimisation towards parallel execution. The largest application in this thesis, predicting the about 70000 values for the cross-validation carried out in Chapter 5, takes about a day.

8

CONCLUSIONS

This thesis presents a new approach to model spatial and spatio-temporal random fields through bivariate spatial/spatio-temporal copulas that construct spatial/spatio-temporal vine copulas. In this construction, the spatial bivariate copulas are given by a spatially weighted convex combination of spatially varying bivariate copulas. The motivation behind this approach is to allow for dependencies in random fields that cannot be captured with approaches relying on a Gaussian distribution. Nevertheless, the copula based model allows as well for Gaussian dependencies requiring no decision beforehand whether the random field follows a Gaussian or non-Gaussian multivariate distribution. The selection of different copula families for different distances allows even to have mixed models of Gaussian and non-Gaussian dependencies.

The improvement in terms of cross-validation statistics of the spatial vine copula in comparison to kriging ranges from moderate to considerable. A clear improvement can be seen in the notion of uncertainties. The flexible and non-parametric conditional distribution functions used for interpolating or simulating the field allow for a richer probabilistic model of associated uncertainties. The flexibility in choosing the marginal distribution functions makes spatial vine copulas very appealing for skewed data sets. However, flexible margins alone are not capable of reproducing extreme values in the prediction and a non-Gaussian dependence structure is required.

The newly developed approach has not only been studied in a conceptional manner, but also been applied to different data sets. These data sets suggest mainly non-Gaussian dependencies that were captured rather well by different copulas. Nevertheless, the overall prediction quality depends considerably on the marginal distributions. This stresses the importance of both, a good model of the dependencies and a good fit of the marginal distributions.

While the development of this approach has mainly been driven by spatial and spatio-temporal phenomena, it is in general applicable to different distance measures as well. Instead of using spatial distance as a parametrization, co-variates could be used as well. A version that only uses temporal distances or even points in time would allow for a similar concept in the field of time series modelling.

BIBLIOGRAPHY

[1] K. Aas and D. Berg. 'Models for construction of multivariate dependence–a comparison study'. In: *The European Journal of Finance* 15.7-8 (2009), pp. 639–659 (cit. on p. 11).

[2] K. Aas, C. Czado, A. Frigessi, and H. Bakken. 'Pair-copula constructions of multiple dependence'. In: *Insurance: Mathematics and Economics* 44 (2009), pp. 182–198. DOI: 10.1016/j.insmatheco.2007.02.001 (cit. on pp. 3, 5, 11, 13, 39, 41, 48, 58, 62, 66, 80, 82, 87).

[3] A. Bárdossy. 'Interpolation of groundwater quality parameters with some values below the detection limit'. In: *Hydrology and Earth System Sciences* 15.9 (2011), pp. 2763–2775. DOI: 10.5194/hess-15-2763-2011 (cit. on pp. 58, 61, 79, 81).

[4] A. Bárdossy and G. G. S. Pegram. 'Copula based multisite model for daily precipitation simulation'. In: *Hydrology and Earth System Sciences* 13.12 (2009), pp. 2299–2314. DOI: 10.5194/hess-13-2299-2009 (cit. on pp. 58, 79).

[5] A. Bárdossy. 'Copula-based geostatistical models for groundwater quality parameters'. In: *Water Resources Research* 42.11 (2006), W11416. DOI: 10.1029/2005WR004754 (cit. on pp. 3, 40, 57, 79).

[6] A. Bárdossy and J. Li. 'Geostatistical interpolation using copulas'. In: *Water Resources Research* 44.7 (W07412 2008), W07412. DOI: 10.1029/2007WR006115 (cit. on pp. 40, 58, 79).

[7] T. Bedford and R. M. Cooke. 'Probability Density Decomposition for Conditionally Dependent Random Variables Modeled by Vines'. In: *Annals of Mathematics and Artificial Intelligence* 32.1-4 (2001), pp. 245–268. DOI: 10.1023/A:1016725902970 (cit. on pp. 11, 48).

[8] T. Bedford and R. M. Cooke. 'Vines – a New Graphical Model for Dependent Random Variables'. In: *The Annals of Statistics* 30.4 (2002), pp. 1031–1068. DOI: 10.1214/aos/1031689016 (cit. on pp. 3, 5, 11, 58, 62, 80, 82).

[9] D. Berg. 'Copula goodness-of-fit testing: an overview and power comparison'. In: *European Journal of Finance* 15.7-8 (2009), pp. 675 –701. URL: http://ideas.repec.org/a/taf/eurjfi/v15y2009i7-8p675-701.html (cit. on p. 13).

[10] R. S. Bivand, E. J. Pebesma, and V. Gómez-Rubio. *Applied Spatial Data Analysis with R*. Use R! New York: Springer, 2008, p. 378 (cit. on p. 40).

[11] E. C. Brechmann, C. Czado, and K. Aas. 'Truncated regular vines in high dimensions with application to financial data'. In: *Canadian Journal of Statistics* 40.1 (2012), pp. 68–85. ISSN: 1708-945X. DOI: 10.1002/cjs.10141 (cit. on p. 96).

[12] E. Brechmann. 'Air pollution modeling at different stations using a hierarchical copula construction'. In: *Abstracts of the Spatial copula day - 21.02.2013, Technical University of Munich*. 2013, pp. 3–4. URL: http://www-m4.ma.tum.de/allgemeines/veranstaltungen/spatial-copula-day/ (cit. on p. 80).

[13] E. C. Brechmann and U. Schepsmeier. 'Modeling Dependence with C- and D-Vine Copulas: The R Package CDVine'. In: *Journal of Statistical Software* 52.3 (Feb. 2013), pp. 1–27. URL: http://www.jstatsoft.org/v52/i03 (cit. on p. 14).

[14] R. M. Cooke, H. Joe, and K. Aas. 'Vines Arise'. In: *Dependence Modelling*. Ed. by D. Kurowicka and H. Joe. World Scientific Publishing Co, 2011. Chap. 3, pp. 37–72 (cit. on pp. 90, 97, 103, 104).

[15] N. Cressie. *Statistics for Spatial Data, Revised Edition*. John Wiley & Sons, 1993 (cit. on p. 1).

[16] N. Cressie and C. K. Wikle. *Statistics for spatio-temporal data*. John Wiley & Sons, 2011 (cit. on p. 58).

[17] C Czado, U Schepsmeier, and A Min. 'Maximum likelihood estimation of mixed C-vines with application to exchange rates'. In: *Statistical Modelling* 12.3 (2012), pp. 229–255. DOI: 10.1177/1471082X1101200302 (cit. on pp. 66, 87).

[18] C. De Michele, G. Salvadori, G. Passoni, and R. Vezzoli. 'A multivariate model of sea storms using copulas'. In: *Coastal Engineering* 54.10 (2007), pp. 734–751. DOI: 10.1016/j.coastaleng.2007.05.007 (cit. on p. 11).

[19] J. F. Dissmann, E. C. Brechmann, C Czado, and D Kurowicka. 'Selecting and estimating regular vine copulae and application to financial returns'. In: *Computational Statistics & Data Analysis* 59.1 (2013), pp. 52 –69. DOI: 10.1016/j.csda.2012.08.010 (cit. on pp. 66, 87).

[20] G Dubois and S Galmarini. 'Introduction to the Spatial Interpolation Comparison (SIC) 2004 Exercise and Presentation of the Datasets'. In: *Applied GIS* 1.2 (2005), pp. 09–1–09–11. DOI: 10.2104/ag050009 (cit. on pp. 80, 88, 89).

[21] G. Dubois and S. Galmarini. 'Spatial Interpolation Comparison (SIC) 2004: introduction to the exercise and overview of results'. In: *automatic mapping algorithms for routine and emergency monitoring data - spatial interpolation comparison 2004*. Ed. by G. Dubois. Office for Official Publication of the European Communities,

2005. URL: http://www.ai-geostats.org/pub/AI_GEOSTATS/EventsSIC2004/EUR_SIC_2004_online.pdf (cit. on p. 95).

[22] EEA. *CORINE (Coordination of Information on Environment) Database, a key database for European integrated environmental assessment*. Tech. rep. Programme of the European Commission, 2000 (cit. on p. 24).

[23] EMEP. *Transboundary particulate matter in Europe Status report 4/2007*. Kjeller: Norwegian Institute for Air Research (EMEP Report 4/2007), 2007. URL: http://www.nilu.no/projects/ccc/reports/emep4-2007.pdf (cit. on p. 67).

[24] A. C. Favre, S. E. Adlouni, L. Perreault, N. Thiemonge, and B. Bobee. 'Multivariate hydrological frequency analysis using copulas'. In: *Water Resources Research* 40.01101 (2004), pp. 10–1029 (cit. on p. 10).

[25] B. Fournier and R. Furrer. 'Automatic mapping in the presence of substitutive errors: A robust kriging approach'. In: *Applied GIS* 1.2 (2005), pp. 12–01–12–16. DOI: 10.2104/ag050012 (cit. on p. 80).

[26] C. Genest and A. Favre. 'Everything You Always Wanted to Know about Copula Modeling but Were Afraid to Ask'. In: *Journal of Hydrologic Engineering* 12.4 (2007), pp. 347–368. DOI: 10.1061/(ASCE)1084-0699(2007)12:4(347) (cit. on p. 10).

[27] C. Genest, A. C. Favre, J. Beliveau, and C. Jacques. 'Metaelliptical copulas and their use in frequency analysis of multivariate hydrological data'. In: *Water Resour. Res* 43.9 (2007), W09401. DOI: 10.1029/2006WR005275 (cit. on pp. 10, 11).

[28] J. Gómez-Hernández. 'A Stochastic Approach to the Simulation of Block Conductivity Fields Conditioned upon Data Measured at a Smaller Scale'. PhD thesis. Stanford, CA: Stanford University, 1991 (cit. on pp. 75, 93).

[29] J. Gómez-Hernández and A. Journel. 'Joint Sequential Simulation of MultiGaussian Fields'. In: *Geostatistics Tróia '92*. Ed. by A. Soares. Vol. 5. Quantitative Geology and Geostatistics. Springer Netherlands, 1993, pp. 85 –94. ISBN: 978-0-7923-2157-6. DOI: 10.1007/978-94-011-1739-5_8 (cit. on p. 88).

[30] B. Gräler. 'Modelling skewed spatial random fields through the spatial vine copula'. In: *Spatial Statistics* (2014). available online, in press. ISSN: 2211-6753. DOI: 10.1016/j.spasta.2014.01.001 (cit. on pp. 58, 61, 76, 79).

[31] B. Gräler, L. E. Gerharz, and E. J. Pebesma. *Spatio-temporal analysis and interpolation of PM10 measurements in Europe*. Tech. rep. ETC/ACM, 2012. URL: http://acm.eionet.europa.eu/reports/ETCACM_TP_2011_10_spatio-temp_AQinterpolation (cit. on pp. 53, 55, 67, 73, 74, 76, 100).

[32] B. Gräler and E. J. Pebesma. 'Modelling Dependence in Space and Time with Vine Copulas'. In: *Expanded Abstract Collection from Ninth International Geostatistics Congress, Oslo, Norway June 11 – 15, 2012*. International Geostatistics Congress, 2012. URL: http://geostats2012.nr.no/1742830.html (cit. on pp. 47, 58, 65, 67, 73–75, 98).

[33] B. Gräler and E. J. Pebesma. 'The pair-copula construction for spatial data: a new approach to model spatial dependency'. In: *Procedia Environmental Sciences* 7 (2011), pp. 206–211. ISSN: 1878-0296. DOI: 10.1016/j.proenv.2011.07.036 (cit. on pp. 39, 80, 97).

[34] B. Gräler, M. van den Berg, S. Vandenberghe, A. Petroselli, S. Grimaldi, B. D. Baets, and N. Verhoest. 'Multivariate return periods in hydrology: a critical and practical review focusing on synthetic design hydrograph estimation'. In: *Hydrology and Earth System Sciences* 17.4 (2013), pp. 1281–1296. DOI: 10.5194/hess-17-1281-2013 (cit. on p. 9).

[35] S. Grimaldi, A. Petroselli, and F Nardi. 'A parsimonious geomorphological unit hydrograph for rainfall–runoff modelling in small ungauged basins'. In: *Hydrological Sciences Journal* 57.1 (2012), pp. 73–83. DOI: 10.1080/02626667.2011.636045 (cit. on p. 23).

[36] S. Grimaldi, A. Petroselli, and N. Romano. 'Green-Ampt Curve-Number mixed procedure as an empirical tool for rainfall–runoff modelling in small and ungauged basins'. In: *Hydrological Processes* 27.8 (2013), pp. 1253–1264 (cit. on p. 23).

[37] S. Grimaldi, A. Petroselli, and F. Serinaldi. 'A continuous simulation model for design-hydrograph estimation in small and ungauged watersheds'. In: *Hydrological Sciences Journal* 57.6 (2012), pp. 1035–1051 (cit. on pp. 22–24).

[38] S. Grimaldi, A. Petroselli, and F. Serinaldi. 'Design hydrograph estimation in small and ungauged watersheds: continuous simulation method versus event-based approach'. In: *Hydrological Processes* 26.20 (2012), pp. 3124–3134 (cit. on pp. 22–24).

[39] S Grimaldi and F Serinaldi. 'Design hydrograph analysis with 3-copula function'. In: *Hydrological Sciences Journal* 51.2 (2006), pp. 223–238. DOI: 10.1623/hysj.51.2.223 (cit. on pp. 10, 11, 80).

[40] S. Grimaldi, A. Petroselli, G. Alonso, and F. Nardi. 'Flow time estimation with spatially variable hillslope velocity in ungauged basins'. In: *Advances in Water Resources* 33.10 (2010), pp. 1216–1223 (cit. on p. 23).

[41] S. Grimaldi and F. Serinaldi. 'Asymmetric copula in multivariate flood frequency analysis'. In: *Advances in Water Resources* 29.8 (2006), pp. 1155–1167. ISSN: 0309-1708. DOI: 10.1016/j.advwatres.2005.09.005 (cit. on p. 10).

[42] U. Haberlandt. 'Geostatistical interpolation of hourly precipitation from rain gauges and radar for a large-scale extreme rainfall event'. In: *Journal of Hydrology* 332.1 – 2 (2007), pp. 144 –157. ISSN: 0022-1694. DOI: 10.1016/j.jhydrol.2006.06.028 (cit. on p. 80).

[43] P. Hiemstra, E. Pebesma, C. Twenhöfel, and G. Heuvelink. 'Real-time automatic interpolation of ambient gamma dose rates from the Dutch Radioactivity Monitoring Network'. In: *Computers & Geosciences* 35.8 (2009), pp. 1711 –1721. DOI: 10.1016/j.cageo.2008.10.011 (cit. on p. 92).

[44] I. Hobæk Haff, K. Aas, and A. Frigessi. 'On the simplified pair-copula construction – Simply useful or too simplistic?' In: *Journal of Multivariate Analysis* 101.5 (2010), pp. 1296–1310. DOI: 10.1016/j.jmva.2009.12.001 (cit. on pp. 6, 11, 13, 41, 48).

[45] J. Horálek, B. Denby, P. de Smet, F. de Leeuw, P. Kurfürst, R. Swart, and T. van Noije. *Spatial mapping of air quality for European scale assessment*. Tech. rep. ETC/ACC, 2007. URL: http://acm.eionet.europa.eu/reports/ETCACC_TechnPaper_2006_6_Spat_AQ (cit. on p. 80).

[46] IGMI. *Raster (Matrix) numerical DEM of Italy*. 2003. URL: http://www.igmi.org/pdf/info_matrix2003.pdf (cit. on p. 24).

[47] H Joe. *Multivariate Models and Dependence Concepts*. Chapman and Hall, 1997 (cit. on pp. 11, 27, 68, 70, 89).

[48] A. G. Journel and C. J. Huijbregts. *Mining Geostatistics*. Academic Press London, 1978 (cit. on p. 88).

[49] S.-C. Kao and R. S. Govindaraju. 'A bivariate frequency analysis of extreme rainfall with implications for design'. In: *Journal of Geophysical Research: Atmospheres* 112.D13 (2007), pp. 1 –15. ISSN: 2156-2202. DOI: 10.1029/2007JD008522 (cit. on p. 10).

[50] S.-C. Kao and R. S. Govindaraju. 'A copula-based joint deficit index for droughts'. In: *Journal of Hydrology* 380.1–2 (2010), pp. 121–134. ISSN: 0022-1694. DOI: 10.1016/j.jhydrol.2009.10.029 (cit. on pp. 11, 17, 80).

[51] S.-C. Kao and R. S. Govindaraju. 'Trivariate statistical analysis of extreme rainfall events via the Plackett family of copulas'. In: *Water Resources Research* 44.2 (2008) (cit. on pp. 10, 11).

[52] H. Kazianka and J. Pilz. 'Bayesian spatial modeling and interpolation using copulas'. In: *Computers & Geosciences* 37.3 (2011), pp. 310–319. ISSN: 0098-3004. DOI: 10.1016/j.cageo.2010.06.005 (cit. on pp. 40, 58, 79).

[53] H. Kazianka and J. Pilz. 'Copula-based geostatistical modeling of continuous and discrete data including covariates'. In: *Stochastic Environmental Research and Risk Assessment* 24.5 (2010), pp. 661–673. DOI: 10.1007/s00477-009-0353-8 (cit. on pp. 3, 58, 61, 79, 81).

[54] H. Kazianka and J. Pilz. 'Spatial interpolation using copula-based geostatistical models'. In: *geoENV VII - Geostatistics for Environmental Applications*. Ed. by P. Atkinson and C. Lloyd. Springer, 2010, pp. 307–320 (cit. on pp. 92, 95).

[55] I. Kojadinovic and J. Yan. 'Modeling Multivariate Distributions with Continuous Margins Using the copula R Package'. In: *Journal of Statistical Software* 34.9 (2010), pp. 1–20. URL: http://www.jstatsoft.org/v34/i09 (cit. on pp. 14, 53, 58).

[56] D. Kurowicka and R. M. Cooke. 'Sampling algorithms for generating joint uniform distributions using the vine-copula method'. In: *Computational statistics & data analysis* 51.6 (2007), pp. 2889–2906 (cit. on p. 11).

[57] R. B. Nelsen. *An Introduction to Copulas*. Second. New York: Springer Science+Buisness, 2006 (cit. on pp. 1, 5, 10, 27, 40, 45, 48, 60, 68, 70, 81, 85, 89).

[58] V. Panchenko. 'Goodness-of-fit test for copulas'. In: *Physica A: Statistical Mechanics and its Applications* 355.1 (2005), pp. 176–182 (cit. on p. 13).

[59] E. Pebesma. 'spacetime: Spatio-Temporal Data in R'. In: *Journal of Statistical Software* 51.7 (2012), pp. 1–30. URL: http://www.jstatsoft.org/v51/i07/ (cit. on pp. 53, 58, 68).

[60] E. J. Pebesma. 'Multivariable geostatistics in S: the gstat package'. In: *Computers & Geosciences* 30 (2004), pp. 683–691. DOI: 10.1016/j.cageo.2004.03.012 (cit. on pp. 89, 92).

[61] E. J. Pebesma and R. S. Bivand. 'Classes and methods for spatial data in R'. In: *R News* 5.2 (2005), pp. 9–13. URL: http://CRAN.R-project.org/doc/Rnews/ (cit. on pp. 43, 58).

[62] M. A. S. Pinya, H. Madsen, and D. Rosbjerg. 'Assessment of the risk of inland flooding in a tidal sluice regulated catchment using multi-variate statistical techniques'. In: *Physics and Chemistry of the Earth, Parts A/B/C* 34.10 (2009), pp. 662–669 (cit. on p. 10).

[63] A Pozdnoukhov. 'Support vector regression for automated robust spatial mapping of natural radioactivity'. In: *Applied GIS* 1.2 (2005), pp. 21–01–21–10. DOI: 10.2104/ag050021 (cit. on p. 80).

[64] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2012. URL: http://www.R-project.org/ (cit. on pp. 14, 53).

[65] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2013. URL: http://www.R-project.org/ (cit. on pp. 67, 88).

[66] G Salvadori. 'Bivariate return periods via 2-copulas'. In: *Statistical Methodology* 1.1 (2004), pp. 129–144 (cit. on pp. 10, 17).

[67] G Salvadori and C De Michele. 'Frequency analysis via copulas: Theoretical aspects and applications to hydrological events'. In: *Water Resources Research* 40.12 (2004) (cit. on pp. 10, 17).

[68] G Salvadori and C. De Michele. 'Multivariate Extreme Value Methods'. In: *Extremes in a Changing Climate*. Ed. by A. AghaKouchak, D. Easterling, K. Hsu, S. Schubert, and S. Sorooshian. Springer, 2013. DOI: 10.1007/978-94-007-4479-0_5 (cit. on pp. 17, 34, 80).

[69] G Salvadori and C De Michele. 'On the use of copulas in hydrology: theory and practice'. In: *Journal of Hydrologic Engineering* 12.4 (2007), pp. 369–380 (cit. on p. 10).

[70] G Salvadori and C De Michele. 'Statistical characterization of temporal structure of storms'. In: *Advances in Water Resources* 29.6 (2006), pp. 827–842 (cit. on p. 11).

[71] G. Salvadori, C. De Michele, and F. Durante. 'On the return period and design in a multivariate framework'. In: *Hydrology and Earth System Sciences* 15.11 (2011), pp. 3293–3305. DOI: 10.5194/hess-15-3293-2011 (cit. on pp. 10, 17, 19, 20, 34, 80).

[72] G. Salvadori. *Extremes in nature: an approach using copulas*. Vol. 56. Springer, 2007 (cit. on pp. 10, 17).

[73] U. Schepsmeier. 'Spatial R-vine copula models'. In: *Abstracts of the Spatial copula day - 21.02.2013, Technical University of Munich*. Technische Universität München. 2013, pp. 2–3. URL: http://www-m4.ma.tum.de/allgemeines/veranstaltungen/spatial-copula-day/ (cit. on p. 80).

[74] U. Schepsmeier and E. Brechmann. *CDVine: Statistical inference of C- and D-vine copulas*. R package version 1.1-5. 2011. URL: http://CRAN.R-project.org/package=CDVine (cit. on p. 53).

[75] U. Schepsmeier, J. Stoeber, and E. C. Brechmann. *VineCopula: Statistical inference of vine copulas*. R package version 1.1-2. 2013 (cit. on p. 58).

[76] D. Schertzer and S. Lovejoy. 'Physical modeling and analysis of rain and clouds by anisotropic scaling multiplicative processes'. In: *Journal of Geophysical Research: Atmospheres (1984–2012)* 92.D8 (1987), pp. 9693–9714 (cit. on p. 22).

[77] F Serinaldi. 'Interactive comment on "Joint return periods in hydrology: a critical and practical review focusing on synthetic design hydrograph estimation" by S. Vandenberghe et al'. In: *Hydrol. Earth Syst. Sci. Discuss* 9 (2012), pp. C3721–C3721 (cit. on p. 35).

[78] F Serinaldi. 'Multifractality, imperfect scaling and hydrological properties of rainfall time series simulated by continuous universal multifractal and discrete random cascade models'. In: *Nonlinear Processes in Geophysics* 17.6 (2010), pp. 697–714. DOI: 10.5194/npg-17-697-2010 (cit. on p. 24).

[79] F. Serinaldi. 'A multisite daily rainfall generator driven by bivariate copula-based mixed distributions'. In: *Journal of Geophysical Research: Atmospheres (1984–2012)* 114.D10103 (2009). DOI: 10.1029/2008JD011258 (cit. on pp. 22, 24).

[80] F. Serinaldi. 'Closure to "Synthetic Design Hydrographs Based on Distribution Functions with Finite Support" by Francesco Serinaldi and Salvatore Grimaldi'. In: *Journal of Hydrologic Engineering* 18.1 (2012), pp. 126–129 (cit. on p. 28).

[81] F. Serinaldi and S. Grimaldi. 'Fully nested 3-copula: procedure and application on hydrological data'. In: *Journal of Hydrologic Engineering* 12.4 (2007), pp. 420–430. DOI: 10.1061/(ASCE)1084-0699(2007)12:4(420) (cit. on pp. 10, 11).

[82] F. Serinaldi and S. Grimaldi. 'Synthetic design hydrographs based on distribution functions with finite support'. In: *Journal of Hydrologic Engineering* 16.5 (2010), pp. 434–446. DOI: 10.1061/(ASCE)HE.1943-5584.0000339 (cit. on pp. 15, 21).

[83] F. Serinaldi and C. G. Kilsby. 'The intrinsic dependence structure of peak, volume, duration, and average intensity of hyetographs and hydrographs'. In: *Water Resources Research* 49.6 (2013), pp. 3423–3442. ISSN: 1944-7973. DOI: 10.1002/wrcr.20221. URL: http://dx.doi.org/10.1002/wrcr.20221 (cit. on p. 101).

[84] J. Shiau. 'Return period of bivariate distributed extreme hydrological events'. In: *Stochastic environmental research and risk assessment* 17.1-2 (2003), pp. 42–57. DOI: 10.1007/s00477-003-0125-9 (cit. on p. 10).

[85] A. Sklar. 'Fonctions de répartition á n dimensions et leurs marges'. In: *Publ. Inst. Statist. Univ. Paris* 8 (1959), pp. 229 –231 (cit. on pp. 4, 10, 60, 81).

[86] S. Song and V. P. Singh. 'Meta-elliptical copulas for drought frequency analysis of periodic hydrologic data'. In: *Stochastic Environmental Research and Risk Assessment* 24.3 (2010), pp. 425–444. DOI: 10.1007/s00477-009-0331-1 (cit. on p. 11).

[87] A. G. Stephenson. 'evd: Extreme Value Distributions'. In: *R News* 2.2 (2002), pp. 31–32. URL: http://CRAN.R-project.org/doc/Rnews/ (cit. on pp. 53, 67).

[88] C. Theiling and J. Burant. 'Flood inundation mapping for integrated floodplain management: Upper Mississippi River system'. In: *River Research and Applications* 29.8 (2013), pp. 961–978. DOI: 10.1002/rra.2583 (cit. on p. 35).

[89] V. Timonin and E. Savelieva. 'Spatial prediction of radioactivity using general regression neural network'. In: *Applied GIS* 1.2 (2005), pp. 19–01–19–14. DOI: 10.2104/ag050019 (cit. on pp. 80, 95).

[90] S. Vandenberghe, N. E. C. Verhoest, C. Onof, and B. De Baets. 'A comparative copula-based bivariate frequency analysis of observed and simulated storm events: A case study on Bartlett-Lewis modeled rainfall'. In: *Water Resources Research* 47.7 (2011), n/a–n/a. DOI: 10.1029/2009WR008388 (cit. on pp. 10, 16, 19).

[91] S. Vandenberghe, N. E. C. Verhoest, E. Buyse, and B. De Baets. 'A stochastic design rainfall generator based on copulas and mass curves'. In: *Hydrology and Earth System Sciences* 14.12 (2010), pp. 2429–2442. DOI: 10.5194/hess-14-2429-2010 (cit. on pp. 10, 18, 20, 21).

[92] E Volpi and A Fiori. 'Design event selection in bivariate hydrological frequency analysis'. In: *Hydrological Sciences Journal* 57.8 (2012), pp. 1506–1515. DOI: 10.1080/02626667.2012.726357 (cit. on p. 20).

[93] G. Wong, M. F. Lambert, M. Leonard, and A. V. Metcalfe. 'Drought analysis using trivariate copulas conditional on climatic states'. In: *Journal of Hydrologic Engineering* 15.2 (2010), pp. 129–141. DOI: 10.1061/(ASCE)HE.1943-5584.0000169 (cit. on p. 11).

[94] J. Yan. 'Enjoy the Joy of Copulas: With a Package copula'. In: *Journal of Statistical Software* 21.4 (Oct. 2007), pp. 1–21. ISSN: 1548-7660. URL: http://www.jstatsoft.org/v21/i04 (cit. on pp. 53, 58).

[95] S. Yue and P. Rasmussen. 'Bivariate frequency analysis: discussion of some useful concepts in hydrological application'. In: *Hydrological Processes* 16.14 (2002), pp. 2881–2898. DOI: 10.1002/hyp.1185 (cit. on p. 10).

[96] L Zhang and V. P. Singh. 'Gumbel-Hougaard copula for trivariate rainfall frequency analysis'. In: *Journal of Hydrologic Engineering* 12.4 (2007), pp. 409–419. DOI: 10.1061/(ASCE)1084-0699(2007)12:4(409) (cit. on pp. 10, 11).

A

LIST OF PUBLICATIONS

A.1 JOURNAL ARTICLE

1. B. Gräler. 'Modelling skewed spatial random fields through the spatial vine copula'. In: *Spatial Statistics* (2014). available online, in press. ISSN: 2211-6753. DOI: 10.1016/j.spasta.2014.01.001.

2. B. Gräler, M. van den Berg, S. Vandenberghe, A. Petroselli, S. Grimaldi, B. D. Baets, and N. Verhoest. 'Multivariate return periods in hydrology: a critical and practical review focusing on synthetic design hydrograph estimation'. In: *Hydrology and Earth System Sciences* 17.4 (2013), pp. 1281–1296. DOI: 10.5194/hess-17-1281-2013.

3. T. Kauppinen, G. M. de Espindola, J. Jones, A. Sánchez, B. Gräler, and T. Bartoschek. 'Linked Brazilian Amazon Rainforest Data'. In: *Semantic Web Journal* 5.2 (2014), pp. 151 –155. URL: http://www.semantic-web-journal.net/content/linked-brazilian-amazon-rainforest-data-0.

4. M. Kilibarda, T. Hengl, G. B. M. Heuvelink, B. Gräler, E. Pebesma, M. Perčec Tadić, and B. Bajat. 'Spatio-temporal interpolation of daily temperatures for global land areas at 1 km resolution'. In: *Journal of Geophysical Research: Atmospheres* (2014). available online, in press. ISSN: 2169-8996. DOI: 10.1002/2013JD020803.

A.2 CONFERENCE ARTICLE

1. L. E. Gerharz, B. Gräler, and E. Pebesma. 'Disaggregating gridded air quality data for individual exposure modelling'. In: *Procedia Environmental Sciences* 7 (2011). <ce:title>Spatial Statistics 2011: Mapping Global Change</ce:title>, pp. 146 –151. ISSN: 1878-0296. DOI: 10.1016/j.proenv.2011.07.026.

2. B. Gräler, H. Kazianka, and G. M. de Espindola. 'Copulas, a novel approach to model spatio-temporal dependence'. In: *GIScience for environmental change - symposium proceedings*. Ed. by K. Henneböhl, L. Vinhas, E. Pebesma, and G. Câmara. Vol. 40. ifgiPrints. Campos de Jordão (São Paulo), Brazil: IOSPress & AKA Verlag, Nov. 2010, pp. 49–54. ISBN: 9783898386449. URL: http://ifgi.uni-muenster.de/~b_grae02/publications/GEOChange_GraelerKaziankaDeEspindola.pdf.

3. B. Gräler and E. J. Pebesma. 'Modelling Dependence in Space and Time with Vine Copulas'. In: *Expanded Abstract Collection from Ninth International Geostatistics Congress, Oslo, Norway June 11 – 15, 2012*. International Geostatistics Congress, 2012. URL: http://geostats2012.nr.no/1742830.html.

4. B. Gräler and E. J. Pebesma. 'The pair-copula construction for spatial data: a new approach to model spatial dependency'. In: *Procedia Environmental Sciences* 7 (2011), pp. 206–211. ISSN: 1878-0296. DOI: 10.1016/j.proenv.2011.07.036.

A.3 CONFERENCE ABSTRACTS

1. B. Gräler. 'An Application of Vine Copulas in the Spatio-Temporal Domain'. In: *Spatial Copula Day*. Technische Universität München, Garching-Hochbrück, Germany, 2013. URL: http://www-m4.ma.tum.de/fileadmin/w00bdb/www/veranstaltungen/Spatial_Copula_Day/Abstracts.pdf.

2. B. Gräler. 'Modeling Extremes with the Spatial Vine Copula'. In: *Spatail Statistics 2013*. Columbus, Ohio, USA, 2013. URL: http://ifgi.uni-muenster.de/~b_grae02/publications/SpatStat2013_extremes.pdf.

3. B. Gräler. 'Modelling Spatial Phenomena and Joint Return Periods with Copulas using R: the spcopula Package'. In: *European Geosciences Union (EGU) General Assembly 2013*. Vienna, Austria, 2013. URL: http://ifgi.uni-muenster.de/~b_grae02/publications/EGU2013_Graeler.pdf.

4. B. Gräler. 'The spcopula R-package: Modelling Spatial and Spatio-Temporal dependence with copulas'. In: *Spatail Statistics 2013*. Columbus, Ohio, USA, 2013. URL: http://ifgi.uni-muenster.de/~b_grae02/publications/SpatStat2013_extremes.pdf.

5. B. Gräler. 'Vine copulas for spatial interpolation'. In: *4th Workshop on Vine Copula Distributions and Applications*. Technische Universität München, Garching-Hochbrück, Germany, 2011. URL: http://www-m4.ma.tum.de/lect-conf/vinesworkshop/index.html.

6. B. Gräler and C. Stasch. 'Flexible Representation of Spatio-Temporal Random Fields in the Model Web'. In: *European Geosciences Union (EGU) General Assembly 2012*. Vienna, Austria, 2012. URL: http://ifgi.uni-muenster.de/~b_grae02/publications/EGU2012_Graeler_Stasch.pdf.

7. B. Gräler, S. Vandenberghe, M. J. van den Berg, S. Grimaldi, A. Petroselli, B. D. Baets, and N. E. C. Verhoest. 'Multivariate Return Periods based on Vine Copulas'. In: *European Geosciences Union (EGU) General Assembly 2012*. Vienna, Austria, 2012. URL: http://ifgi.uni-muenster.de/~b_grae02/publications/EGU2012_Graeler_et_al.pdf.

A.4 TECHNICAL REPORTS

1. L. Gerharz, B. Gräler, and E. J. Pebesma. *Measurement artefacts and inhomogeneity detection*. Tech. rep. ETC/AM, 2011. URL: http://acm.eionet.europa.eu/reports/ETCACM_TP_2011_8_artefacts_inhom_detection.

2. B. Gräler, L. E. Gerharz, and E. J. Pebesma. *Spatio-temporal analysis and interpolation of PM10 measurements in Europe*. Tech. rep. ETC/ACM, 2012 (Erratum: March 2013). URL: http://acm.eionet.europa.eu/reports/ETCACM_TP_2011_10_spatio-temp_AQinterpolation.

3. B. Gräler, M. Rehr, L. Gerharz, and E. J. Pebesma. *Spatio-temporal analysis and interpolation of PM10 measurements in Europe for 2009*. Tech. rep. ETC/AM, 2013. URL: http://acm.eionet.europa.eu/reports/ETCACM_2012_8_spatio-temp_PM10analyses.

B

SOFTWARE CONTRIBUTIONS

In the course of my research, I contributed to several R-packages. These implementations backed-up my research and allowed me to evaluate and illustrate my new developments. A detailed overview of contributions is available at my Open Hub profile.

SPCOPULA The spcopula package combines the copula, VineCopula, sp and spacetime packages and enables the user to model spatial or saptio-temporal random fields with spatial vine copulas. This package summarises all approaches described in this thesis. See as well Chapter 5 and Table 5.1.

VINECOPULA Some adoptions to efficiently link the copula and VineCopula packages in the spcopula package were necessary. Initially in spcopula implemented S4-class wrappers to the copula families available in VineCopula are now available from VineCopula directly. This allows to use these additional families seamlessly with the copula package.

GSTAT Extensions of the spatial kriging capabilities of the gstat package have been extended to the spatio-temporal case. I contributed largely to this effort.

SPACETIME In the course of using the spacetime package, I made a couple of changes and a few additions.

COPULATHEQUE For teaching purposes, I composed an interactive exploration tool for bivariate copulas named copulatheque.

LEBENSLAUF

- removed from the digital version -