Goodness-of-Fit Testing for Copula-based models with Applications in Atmospheric Science

by

Albert Rapp

A Thesis Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Master of Science

in Mathematics

at

The University of Wisconsin–Milwaukee

May 2017

Abstract

Goodness-of-Fit Testing for Copula-based models with Applications in Atmospheric Science

by

Albert Rapp

The University of Wisconsin–Milwaukee, 2017
Under the Supervision of Professor Vincent E. Larson

Every elementary probability course discusses how to construct joint distribution functions of independent random variables but joint distribution functions of dependent random variables are usually omitted. Obviously, the reason is that things are not as simple as in the independent case. In this matter, so-called copulas can be an elegant tool to investigate dependency structures other than independence.

A copula is a convenient function which links the marginal distributions of random variables to their joint distribution. The beauty here is that one can use suitable copulas to model any desired dependence structure between any set of random variables without even knowing their marginal distributions.

In the end, using copulas for modeling comes down to figuring out which copula is suitable given a set of observations. One way to investigate this is based on goodness-of-fit tests which are specifically designed for copulas.

Ultimately, this thesis gives an introduction into the necessary theory of copulas and their goodness-of-fit tests in order to use them to compare popular models for cloud overlap in atmospheric science.

To my loving parents

who support me in my personal

and academic endeavors.

Table of Contents

List of Figures

List of Tables

Acknowledgements

First, I would like to thank Dr. Vincent E. Larson for agreeing to work with me on this master thesis and giving me the opportunity to work on this project.

Second, I want to thank Dr. Jay Beder and Dr. Richard Stockbridge for being part of my thesis committee and for assisting me throughout my thesis. I am grateful for the valuable and detailed feedback the two of you provided for me.

Third, I want to thank William Langhoff for agreeing to read my thesis and listening to my theories and explanations which was especially helpful for my implementations in Matlab.

Finally, I want to thank everybody in the TA office for making the struggles of being a graduate student bearable.

Rule of math: If it is easy, you're doing it wrong.

– *Everyone who ever tried to solve a math problem*

Chapter 1

Introduction

A copula is a convenient function which links the marginal distributions of random variables to their joint distribution. The beauty here is that one can use suitable copulas to model any desired dependence structure between any set of random variables without even knowing their marginal distributions. In the end, using copulas for modeling comes down to figuring out which copula is suitable given a set of observations. One way to investigate this is based on goodness-of-fit tests which are specifically designed for copulas.

Ultimately, this thesis gives an introduction into the necessary theory of copulas and their goodness-of-fit tests in order to use them to compare popular models for cloud overlap in atmospheric science. More precisely, this thesis focuses on establishing a testing methodology for analyzing the dependence of the amount of cloud ice in two adjacent cloud layers.

In the pursuit to of that goal, chapter 2 outlines the fundamental definitions and theorems needed for understanding copulas and their goodness-of-fit tests. Subsequently, this is followed by chapter 3 which introduces the most prominent models this thesis is going to investigate. Then, chapter 4 uses these models and the goodness-of-fit tests on data to establish which model fits the data best. Finally, chapter 5 concludes the thesis with a summary of the findings.

Chapter 2

Copula Theory

This introduction to copulas follows the book "An Introduction to Copulas" by Roger Nelson [1] closely but only summarizes the key elements which are important for this thesis. It is worth noting that Nelson focuses on bivariate copulas which is all that is needed here. However, Nelson also offers explanations as to how to extend copulas to higher dimensions. The theory of Goodness-of-Fit testing for copulas, which follows after the introduction to copulas, is based on the corresponding paper by Genest et al. [2].

2.1 Introduction

2.1.1 Preliminaries

This section is designed to introduce the most important terminology to deal with copulas.

Definition 2.1. (H-Volume of a Set) For two nonempty sets $S_1, S_2 \subset \mathbb{R}$, a mapping $H : S_1 \times S_2 \to \bar{\mathbb{R}}$ and a set $B = [x_1, x_2] \times [y_1, y_2] \subset S_1 \times S_2$, one defines the *H-Volume of B* as

$$V_H(B) = H(x_2, y_2) - H(x_1, y_2) - H(x_2, y_1) + H(x_1, y_1).$$

Definition 2.2. (H-Measure) The mapping H in definition 2.1 is called 2-increasing if $V_H(B) \geqslant 0$ for all rectangles B whose vertices lie in the domain of H. Also, if H is 2-increasing, then $V_H(B)$ is called H-measure of B.

Definition 2.3. (Groundedness) Suppose for two nonempty sets $S_1, S_2 \subset \mathbb{R}$ and a mapping $H : S_1 \times S_2 \to \bar{\mathbb{R}}$ there exist minima $a_1 = \min(S_1)$ and $a_2 = \min(S_2)$. Then H is said

to be *grounded* if

$$H(x, a_2) = 0 = H(a_1, y) \text{ for all } (x, y) \in S_1 \times S_2.$$

Definition 2.4. (Margins) Suppose for two nonempty sets $S_1, S_2 \subset \mathbb{R}$ and a mapping $H : S_1 \times S_2 \to \bar{\mathbb{R}}$ there exist maxima $b_1 = \max(S_1)$ and $b_2 = \max(S_2)$, then H is said to have *margins* F and G which are defined by

$$F : S_1 \to \mathbb{R}, F(x) = H(x, b_2) \text{ for all } x \in S_1$$

$$G : S_2 \to \mathbb{R}, G(x) = H(b_1, y) \text{ for all } y \in S_2.$$

Given this new terminology, it is possible to describe a joint distribution function H without any probabilistic notions:

Definition 2.5. (Joint Distribution Function) A joint distribution function is a function H with domain $\bar{\mathbb{R}}^2$ such that

1. H is 2-increasing
2. $H(x, -\infty) = 0 = H(-\infty, y)$
3. $H(\infty, \infty) = 1$

Thus, H is grounded with margins F and G which are distribution functions as well.

2.1.2 Copulas and Sklar's Theorem

Now that the necessary terminology is properly introduced, it is time to focus on the essentials of copulas.

Definition 2.6. (Copula) A copula is a function $C(u, v) : [0, 1]^2 \to [0, 1]$ which is grounded and 2-increasing such that

1. $C(u, 1) = u$

2. $C(1, v) = v$

for all $u, v \in [0, 1]$.

So, one way to think about a copula is to think of it as a two-dimensional distribution function with support $[0, 1]^2$ and uniform margins. However, this is not why copulas are so interesting. The real power of copulas becomes apparent once one introduces Sklar's theorem.

Theorem 2.7. (Sklar, [3])

1. If H is a joint distribution function with margins F and G, then there exists a copula C such that $H(x, y) = C\big(F(x), G(y)\big)$ for all $x, y \in \bar{\mathbb{R}}$. Additionally, if F and G are continuous, then C is unique.

2. On the other hand, if C is a copula and F and G are distribution functions, then $H(x, y) = C(F(x), G(y))$ is a joint distribution function with margins F and G

Sklar's theorem shows that copulas can be used to couple a joint distribution function to its univariate margins. Thus, copulas can be used to model dependencies between random variables by taking their respective distribution functions (which are often known) and plugging them into a copula. This way, one gets a joint distribution and needs to assess if this joint distribution fits real world observations. In the continuous case, there is a unique copula that fits the real joint distribution. Consequently, it comes down to figuring out which copula to use.

With this in mind, one very fundamental fact from probability theory can be expressed in terms of copulas:

Theorem 2.8. (Independence Copula) If X and Y are continuous random variables, then X and Y are independent if and only if the copula that couples the marginal distributions, F and G respectively, and the joint distribution H is given by the independence

copula $C(u, v) = uv$, i.e.

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

One very last piece that is needed in this thesis deals with strictly increasing transformations of random variables

Theorem 2.9. (Invariance under strictly increasing Transformation) Suppose X and Y are continuous random variables with associated copula C_{XY}, i.e. C_{XY} links the marginal distributions of X and Y to their joint distribution.
Then, for strictly increasing functions α, β the copula that links $\alpha(X)$ and $\beta(Y)$ is given by C_{XY}, i.e. the copula is invariant under strictly increasing transformation.

2.2 Goodness-of-Fit Testing

As already mentioned, the main purpose of this thesis is to analyze a data set and figure out which copula-based model fits this data best. The models that are being used will be thoroughly introduced in the next chapter. Here, the focus is on establishing a general testing methodology that allows one to infer which model fits the data best.

In their paper about goodness-of-fit testing, Genest et al. [2] propose a so-called "blanket test" based on a sample of independent and identically distributed realizations of d-dimensional random vectors $\mathbf{X}_i = (X_{i1}, \ldots, X_{id})$ $(i = 1, \ldots, n)$, where n is the sample size. They describe the meaning of "blanket test" as not needing any prior "parameter tuning or other strategies". In this thesis, only two-dimensional data will be analysed and as such all subsequent formulas in this section are specifically rewritten for the case $d = 2$.

The first step in dealing with data requires a strictly increasing transformation of the data by using the empirical marginal distribution function

$$\hat{F}_j(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_{ij} \leqslant t) \quad (j = 1, 2)$$

to transform the observations into

$$U_{ij} = \frac{n}{n+1}\hat{F}_j(X_{ij}) \quad (j = 1, 2).$$

It is worth pointing out that the underlying copula is not affected by this transformation as established in theorem 2.9. Also, Genest et al. describe the factor $\frac{n}{n+1}$ to be only of technical nature.

With the transformed observations, also called pseudo-observations, one can compute the associated empirical copula which is given by

$$C_n(u_1, u_2) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left(U_{i1} \leqslant u_1, U_{i2} \leqslant u_2\right) \text{ for } u_1, u_2 \in [0, 1].$$

So far, there were no hypotheses mentioned which are supposed to be tested. This is due to the fact that this wasn't needed because one has to do the previous steps in any case as the empirical copula is an estimator for the underlying copula as proven by Fermanian et al. [4]. However, it becomes apparent that there is a need for a hypothesis to move forward. After all, for testing purposes the empirical copula needs to be "compared" to some copula to establish how close the empirical copula comes to the underlying copula. So, the hypothesis that will be tested here is $H_0 : \ C \in \mathcal{C}_0$, where C represents the underlying copula and $\mathcal{C}_0$ represents a parametric family of copulas.

Now that there is a hypothesis which assumes a specific parametric copula family, one can use the this hypothesis to estimate the family's parameter(s) θ_n based on the pseudo-observations $\mathbf{U}_1, \ldots, \mathbf{U}_n$. The corresponding copula will be denoted by C_{θ_n}. As both C_{θ_n} and C_n are two-dimensional functions with support $[0, 1]^2$, one can compute the so-called empirical process

$$\mathbb{C}_n(u_1, u_2) = \sqrt{n}\big(C_n(u_1, u_2) - C_{\theta_n}(u_1, u_2)\big) \text{ for } u_1, u_2 \in [0, 1].$$

The empirical process now serves as a tool to compute a statistic for testing purposes. Genest et al. mention two statistics which can be used here, namely versions of the Cramér-

von-Mises and Kolmogorov-Smirnov statistics

$$S_n = \int_{[0,1]^2} \mathbb{C}_n^2(u_1, u_2)\ dC_n(u_1, u_2) \text{ and } T_n = \sup_{u_1,u_2\in[0,1]} |\mathbb{C}_n(u_1, u_2)|.$$

Genest et al. continue by stating that the limiting distributions of these statistics depend on the null hypothesis and unknown parameter θ which makes it impossible to find values to compare these statistics to. To overcome this problem, they offer a parametric bootstrap procedure to compute p-values in order to use these statistics for testing purposes. Also, their findings show that the statistic S_n is to be preferred over the statistic T_n.

Since the parametric bootstrap will be needed later for the data analysis, it is worth establishing the procedure here. This is especially useful since many of the copulas which are of interest in this thesis will need Monte Carlo approximations because the copula cannot be described by analytical expressions:

1. Based on the pseudo-observations $\mathbf{U}_1, \ldots, \mathbf{U}_n$, compute the empirical copula C_n and estimate the parameter θ_n under the null hypothesis. (Chapter 4 states all null hypotheses which will be considered as part of this thesis. Each parameter estimation is implemented as described in chapter 3.)

2. Choose $m \geqslant n$ and

   (i) Generate a random sample $\mathbf{U}_1^*, \ldots, \mathbf{U}_m^*$ from the distribution C_{θ_n}.

   (ii) Approximate C_{θ_n} with the empirical copula based on $\mathbf{U}_1^*, \ldots, \mathbf{U}_m^*$, i.e. compute

   $$B_m^*(u_1, u_2) = \frac{1}{m} \sum_{i=1}^{n} \mathbb{1}\left(U_{i1}^* \leqslant u_1, U_{i2}^* \leqslant u_2\right) \text{ for } u_1, u_2 \in [0, 1].$$

   (iii) Approximate S_n by

   $$S_n = \sum_{i=1}^{n} \left\{C_n(U_{i1}, U_{i2}) - B_m^*(U_{i1}, U_{i2})\right\}^2$$

3. For some large integer N, repeat for every $k \in \{1, \ldots, N\}$

   (i) Generate a random sample $\mathbf{Y}_1^*, \ldots, \mathbf{Y}_n^*$ from the distribution C_{θ_n} and compute

their pseudo-observations $\mathbf{U}_1^*, \ldots, \mathbf{U}_n^*$

(ii) Based on $\mathbf{U}_1^*, \ldots, \mathbf{U}_n^*$, compute the empirical copula, i.e. compute

$$C_n^*(u_1, u_2) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1} \left(U_{i1}^* \leqslant u_1, U_{i2}^* \leqslant u_2 \right) \text{ for } u_1, u_2 \in [0, 1].$$

and estimate θ_n^*.

(iii) Choose $m \geqslant n$ and

(a) Generate a random sample $\mathbf{Y}_1^{**}, \ldots, \mathbf{Y}_m^{**}$ from the distribution $C_{\theta_n^*}$.

(b) Approximate $C_{\theta_n^*}$ with the empirical copula based on $\mathbf{Y}_1^{**}, \ldots, \mathbf{Y}_m^{**}$, i.e. compute

$$B_m^{**}(u_1, u_2) = \frac{1}{m} \sum_{i=1}^{n} \mathbb{1} \left(Y_{i1}^{**} \leqslant u_1, Y_{i2}^{**} \leqslant u_2 \right) \text{ for } u_1, u_2 \in [0, 1].$$

(c) Now, let

$$S_{n,k}^* = \sum_{i=1}^{n} \{ C_n^*(U_{i1}^*, U_{i2}^*) - B_n^{**}(U_{i1}^*, U_{i2}^*) \}^2$$

4. An approximate p-value is given by

$$\frac{1}{N} \sum_{k=1}^{N} \mathbb{1}(S_{n,k}^* > S_n)$$

How this is used for testing purposes will be elaborated further in chapter 4. Also, it is worth noting that Genest and Rémillard [5] established the validity of this parametric bootstrap.

Chapter 3

Introduction of Copula-based Models

A main objective of this thesis is to compare multiple models to a given data set. A first look at this data is given by figure 3.1. Each model uses different copulas to model dependencies between hydrometeors of two adjacent cloud layers. All of these models won't fit the data perfectly as these models are idealized but are chosen due to computational simplicity. As such, every parameter estimation (if the model has a parameter) is designed to take this mismatch of reality and model into account.

Further, it is worth pointing out that a lot of these copulas are described in the atmospheric science literature solely by means of simulation as they often do not have an analytical form. However, for computational purposes (as in this thesis) the description by simulation is sufficient. Also, the bootstrap which was introduced in 2.2 can be modified for copulas which have an analytical form such that bootstrap does not rely on simulations as intensely as it does now. However, for the purpose of applying the exact same test procedure to each copula, the bootstrap was not modified for copulas with analytical formulas.

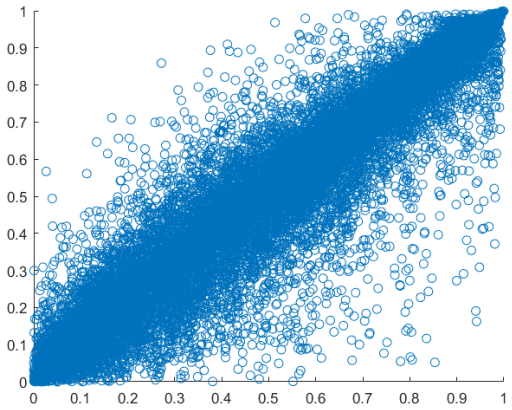

Figure 3.1: Pseudo-observations generated from cloud ice data at heights 10.75 km and 11 km

3.1 Independence and Identity Copula

The first two models are based on the independence copula $C^I(u, v)$ and on the identity copula C^{ID}, which are described by Ovchinnikov et al. [6]. These two models are very basic but also extremely contrasting. The model using the independence copula asserts that there is no dependency between two random variables of interest whereas the other model is based on the other extreme, namely that the two random variables show perfect positive correlated. In the atmospheric science literature, this copula is often referred to as maximum copula but this thesis will refer to it as identity copula. Both models are illustrated in figure 3.2.

Also, it is worth noting that these models do not require any parameter estimations. Further, both copulas have analytical forms which are given by

$$C^I(u, v) = uv \text{ and } C^{ID}(u, v) = \min(u, v) \text{ for } u, v \in [0, 1].$$

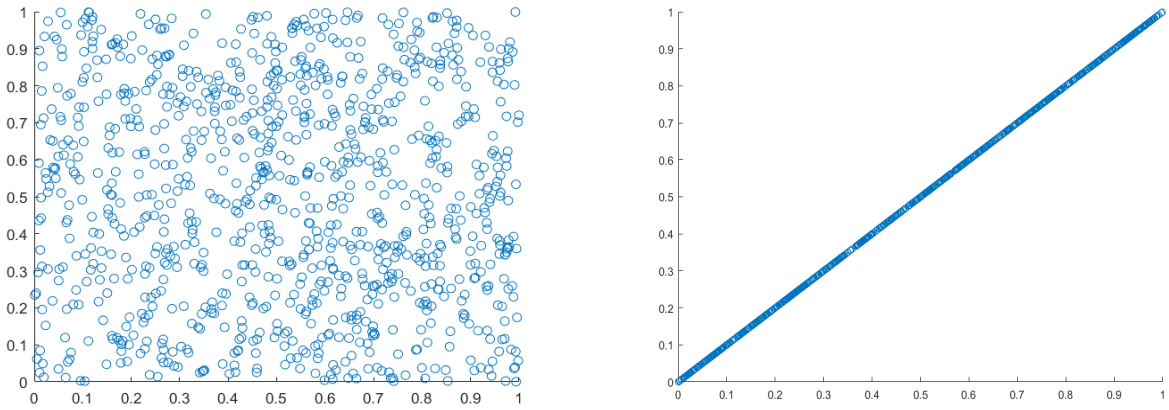

Figure 3.2: Scatter plot of 1000 realizations of independence copula (left) and identity copula (right)

3.2 Independence combined with Identity Copula

Since the previous two copulas were quite extreme, it might be a good idea to mix them. Räisänen et al. [7] propose using a convex combination of the independence copula $C^I(u, v)$

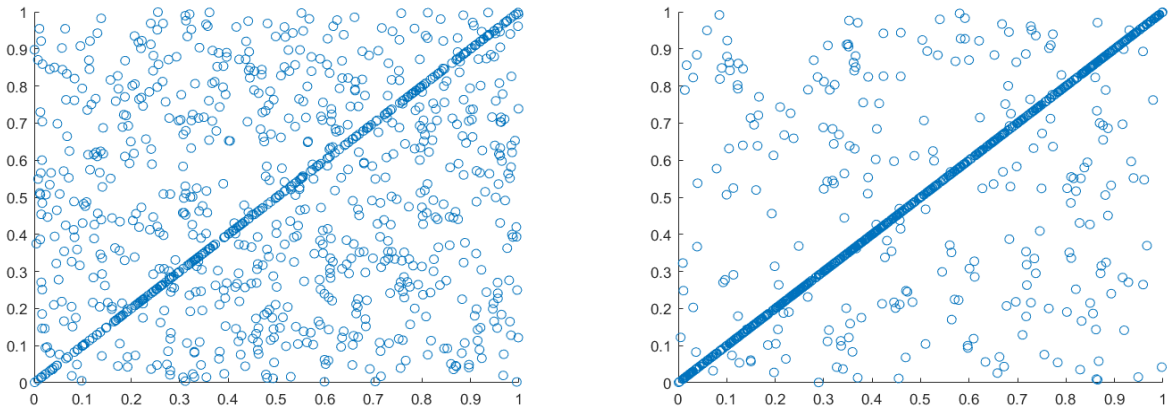

Figure 3.3: Scatter plot of 1000 realizations of the the mixing of identity and independence copula with parameter $\alpha = 0.25$ (left) and $\alpha = 0.75$ (right)

and identity copula $C^{ID}(u, v)$ to generate a new copula

$$C_\alpha(u, v) = \alpha C^{ID}(u, v) + (1 - \alpha)C^I(u, v) \text{ for } u, v \in [0, 1]$$

where $\alpha \in [0, 1]$ can be seen as a parameter of this copula describing the probability for realizations to be on the diagonal. 1000 realizations of this copula with different values for the parameter α can be seen in figure 3.3. This model is still fairly simple and from observations (u_i, v_i) $(i = 1, \ldots, n)$, the parameter can be intuitively estimated by

$$\hat{\alpha} = \frac{\#\{(u, v)|u = v\}}{n}.$$

3.3 The Diagonal Band Copula

The so-called diagonal band copula will be the main focus of the data analysis in this thesis. Therefore, it is vital to introduce it properly. In figure 3.4, one can see a scatter plot of 10000 realizations of this copula and by looking at it, it becomes clear how this copula ended up with the name diagonal band copula.

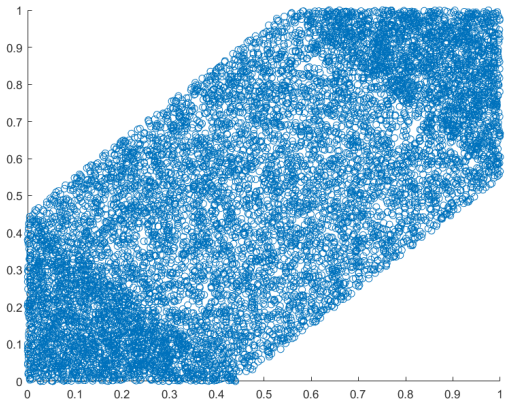

Figure 3.4: Scatter plot of 10000 realizations of a band diagonal copula with half-width 0.45

To create such a realization Larson and Schanen [8] introduce an easy approach which consists of just a few simple steps:

1. Generate N realizations of a standard uniform random variable $U_1, \ldots, U_N$.

2. For each realization U_i $(i = 1, \ldots, N)$, generate one realization of a random variable $\varepsilon \sim U(-\delta, \delta)$ where $\delta \in [0, 1]$ is the half-width of the diagonal band copula. Now, set $V_i = U_i + \varepsilon$ $(i = 1, \ldots, N)$.

3. One quickly realizes that (U_i, V_i) is not bound to the unit square $[0, 1]^2$ and therefore cannot be a copula. It is possible to overcome this problem by reflecting V_i along the line $V = 0$ or $V = 1$. Consequently, one computes

$$\tilde{V}_i = \begin{cases} -V_i & , V_i < 0 \\ 1 - V_i & , V_i > 1 \end{cases}$$

The copula's computational simplicity allows it to be used for practical purposes. However, in order to answer practical questions like how to estimate the band width from data or how well a certain diagonal band fits data, it is advisable to derive some mathematical properties first.

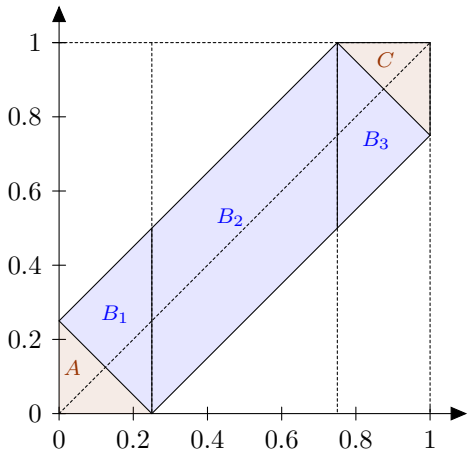

Figure 3.5: Support of a diagonal band copula with half-width $\delta = 0.25$.

3.3.1 The Density of the Diagonal Band Copula

To get a better understanding of the diagonal band copula, it is a good idea to have a look at its support. Figure 3.5 shows the support of a diagonal band copula with half-width $\delta = 0.25$ and the figure also shows how the support can be split into the following 5 areas:

$$A = \big\{(u,v)|u \in [0,\delta], v \in [0,\delta-u]\big\}$$

$$B_1 = \big\{(u,v)|u \in [0,\delta], v \in [\delta-u,\delta+u]\big\}$$

$$B_2 = \big\{(u,v)|u \in (\delta,1-\delta), v \in [u-\delta,u+\delta]\big\}$$

$$B_3 = \big\{(u,v)|u \in [1-\delta,1], v \in [u-\delta,2-\delta-u]\big\}$$

$$C = \big\{(u,v)|u \in [1-\delta,1], v \in [2-\delta-u,1]\big\}$$

To derive a density for the band diagonal copula it is reasonable to assume that the density is constant within A, B_1, B_2, B_3 and C, respectively. Further, one may assume that the density is constant within $B = B_1 \cup B_2 \cup B_3$. Similarly, the density's value within A is the same as within C but is twice the value within B_1 due to the reflection in the last step of the previous algorithm. Consequently, finding the copula's density is possible by finding a

constant $c \geqslant 0$ such that

$$c_\delta(u,v) = \begin{cases} 2c & ,(u,v) \in A \text{ or } (u,v) \in C \\ c & ,(u,v) \in B \end{cases} \quad \text{and}$$

$$\int_{[0,1]^2} c_\delta(u,v) d(u,v) = 1.$$

By simple calculations one easily finds that $c = \frac{1}{2\delta}$ and gets

$$c_\delta(u,v) = \begin{cases} \frac{1}{\delta} & ,(u,v) \in A \text{ or } (u,v) \in C \\ \frac{1}{2\delta} & ,(u,v) \in B \end{cases} \quad \text{if } \delta > 0.$$

Interestingly, $\delta = 1$ reduces this to $c_1(u,v) = 1$ for $(u,v) \in [0,1]^2$ which implies that this is (in some way) a generalization of the independence copula. Further, for $\delta = 0$ it is not possible to derive a density because its support would be given by $D = \left\{(u,v) \in [0,1]^2 \mid u = v\right\}$ and one would have to find a constant $c \geqslant 0$ such that $\int_D c \; d(u,v) = 1$. However, one can see that if $V = U$ the dependence between the two random variables can be described by a diagonal band copula with half-width 0. Consequently, the diagonal band copula can be seen as a generalization of the identity copula too.

3.3.2 Estimating the band half-width

To estimate a copula's parameter(s) from a data sample (X_{i1}, X_{i2}) $(i = 1, \ldots, n)$ Genest et al. [2] propose maximizing the log pseudo-likelihood function $\ell(\delta)$ which is given by

$$\ell(\delta) = \sum_{i=1}^{n} \log \left\{ c_\delta \left(\hat{F}_1(X_{i1}), \hat{F}_2(X_{i2}) \right) \right\}$$

where

$$\hat{F}_j(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_{ij} \leqslant t) \quad (j = 1, 2)$$

are the empirical marginal distribution functions. However, in this case using this approach isn't feasible as the density's support is dependent on the parameter δ. This leads to the fact that this method can only be used to determine when all data points are within a band with half-width δ.

Looking at the data in figure 3.1, it becomes apparent that the pseudo-observations are highly concentrated around the line from $(0, 0)$ to $(1, 1)$. In fact, Larson et al. [6] found that the amount of pseudo-observations increases towards this line. Of course, this implies that the band diagonal model won't fit the data perfectly but, as already stated in the beginning of this chapter, the computational ease of this model makes it attractive to use anyways. So, the goal here is to find a diagonal band with half-width δ such that the band is as narrow as possible but also includes significantly many observations. Consequently, to figure out which half-width δ is suitable in order for a "significant" amount of the data points to be within the band, a different approach needs to be found. A (very unintuitive) approach is given by the following algorithm:

1. For $\delta \in [0, 1]$ count how many data points fall within the area $A \cup B_2 \cup C$ and define $L(\delta)$ to be equal to this number.

2. Find a local maximum $L_{\max}$ of $L(\delta)$ and define $\hat{\delta}$ such that $L(\hat{\delta}) = L_{\max}$.

As δ increases, the areas of A and C increase, whereas the area of B_2 increases at first but decreases once the underlying half-width δ is reached. Consequently, this algorithm tries to find the moment for which the increase in the area of $A \cup C$ does not compensate for the decrease in the area of B_2. In other words, the algorithms determines when an increase in the band's half-width doesn't allow for a significant increase in the amount of data points within the band.

In figure 3.6 one can see that the graphs of $L(\delta)$ look substantially different for different underlying half-widths. If the real half-width is less than 0.5, then the graph looks as in the figure on the left, whereas the graph looks as in the figure on the right if the real half-width

is greater than or equal to 0.5. Consequently, $\hat{\delta}$ in the above algorithm delivers an estimate for the real half-width δ if δ is less than 0.5 but not if δ is greater than or equal to 0.5.

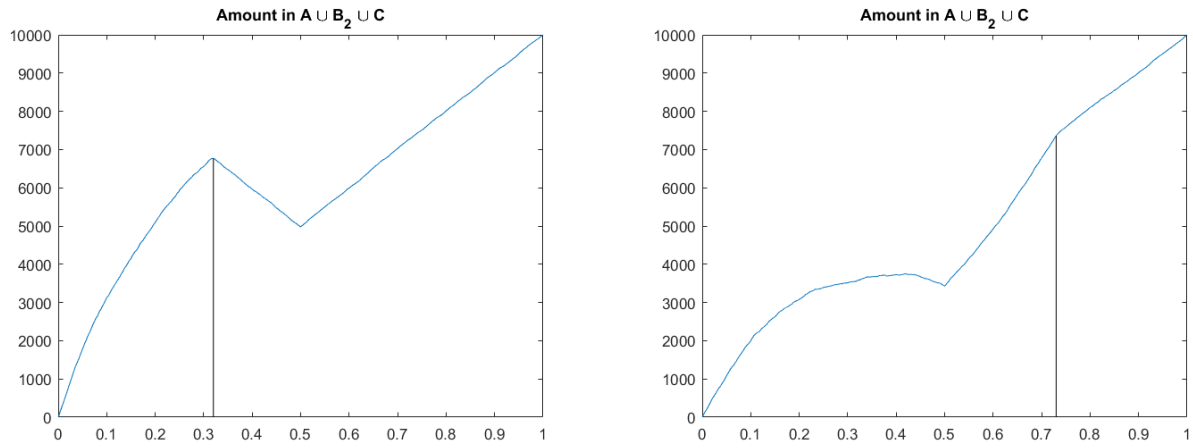

Figure 3.6: $L(\delta)$ computed for 10000 realizations of a diagonal band copula with a half-width of 0.32 (left) and a half-width of 0.73 (right)

To account for this, the above algorithm needs to be changed in a couple of ways. First, the algorithm needs to determine if the underlying half-width is greater than or equal to 0.5 or not. In the latter case, the approach using a local maximum yields good results and in the former case, the algorithm needs to find a "significant" change in the slope of $L(\delta)$ for $\delta \in [0.5, 1]$. In this thesis, the detection of the slope change is implemented by a combination of linear interpolation and minimal distance between interpolation and data. Concretely, the adjusted algorithm looks like this:

1. For $\delta \in [0, 1]$ count how many data points fall within the area $A \cup B_2 \cup C$ and define $L(\delta)$ to be equal to this number.

2. Find a local maximum $L_{\max}$ of $L(\delta)$ and check if $L(L_{\max})$ is greater than $\frac{n}{2}$ where n is the number of observations.

   (a) If $L(L_{\max}) > \frac{n}{2}$, then $L(\delta)$ looks like the left graph in figure 3.6. Consequently, define $\hat{\delta}$ such that $L(\hat{\delta}) = L_{\max}$.

(b) If $L(L_{\max}) \leqslant \frac{n}{2}$, then $L(\delta)$ looks like the graph on the right side of figure 3.6. Therefore, δ is greater than or equal to 0.5. Now,

- For every $\delta \in [0.5, 1]$, define function $y_\delta(x)$ by linearly interpolating the three points $\big(0.5,\ L(0.5)\big), \big(\delta, n \cdot \delta\big),\ \big(1, n\big)$.
- Compute distance between interpolation and data by calculating

$$d(\delta) = \int_{0.5}^{1} \big(L(x) - y_\delta(x)\big)^2 \, dx$$

and choose

$$\hat{\delta} = \underset{\delta \in [0.5,1]}{\arg\min}\ d(\delta).$$

.

To test this algorithm's accuracy, a simulation study was implemented which simulates 1000 runs of n realizations of a diagonal band copula and estimates its parameter afterwards. The mean of the estimated parameters is then compared to the real parameter. Figure 3.7 shows the results of this study for $n = 1000$ and $n = 2000$. Evidently, the algorithm delivers good results if the underlying half-width is less than 0.5. Around half-widths of 0.5, the algorithms shows significant estimation errors and the error for higher half-widths is lower than the error around half-widths of 0.5 but still increases with increasing half-width. It is also worth noting that the increase in the amount of data decreased the error around 0.5 but did not significantly affect the rest.

A look at the data in figure in figure 3.1 shows that the data can be described by a diagonal band copula with an half-width less than 0.5. For future purposes it might be interesting to see how the detection of the slope change can be improved by means other than linear interpolation. However, for the purposes of this thesis, the algorithm is sufficient.

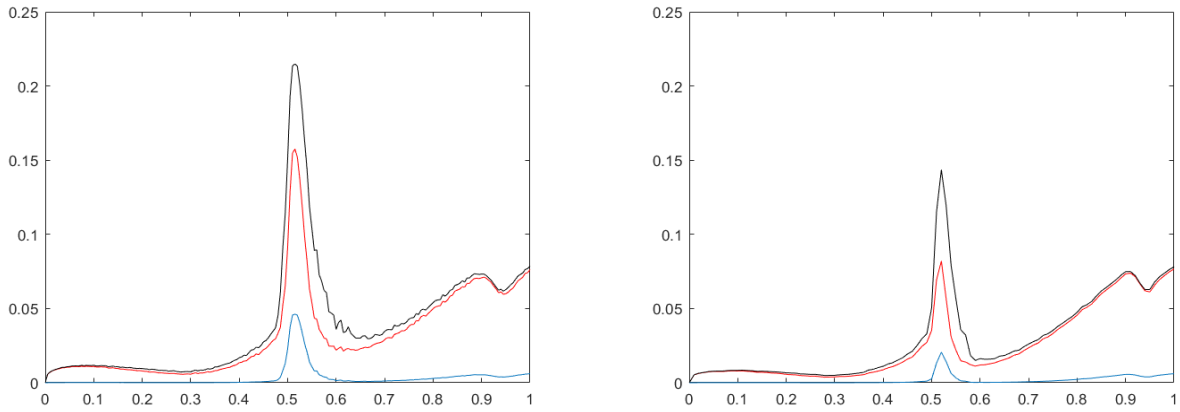

Figure 3.7: Simulation study to test algorithm's accuracy with the number of observations being $n = 1000$ (on the left) and $n = 2000$ (on the right). The graphs show mean absolute error (MAE, in blue), mean squared error (MSE, in red) and its square root (in black).

3.4 Diagonal band combined with Identity Copula

As already mentioned, the concentration of observations increases towards the line from $(0, 0)$ to $(1, 1)$. Consequently, Vincent Larson suggested (during regular meetings as part of this thesis) that it might be beneficial to combine the diagonal band copula with the identity copula to account for this phenomena. Again, one can use a convex combination to combine identity copula $C^{ID}(u, v)$ and diagonal band copula $C_\delta(u, v)$:

$$C_{\alpha,\delta}(u, v) = \alpha C^{ID}(u, v) + (1 - \alpha)C_\delta(u, v) \text{ for } u, v \in [0, 1]$$

where $\alpha \in [0, 1]$ is the parameter describing the probability for realizations to be on the line from $(0, 0)$ to $(1, 1)$ and $\delta \in [0, 1]$ is the parameter describing the diagonal band's half-width. The effect of using different α is depicted in figure 3.8.

Estimation of the two parameters α and δ is straightforward and can be done separately from each other by using the estimators given in sections 3.2 and 3.3.

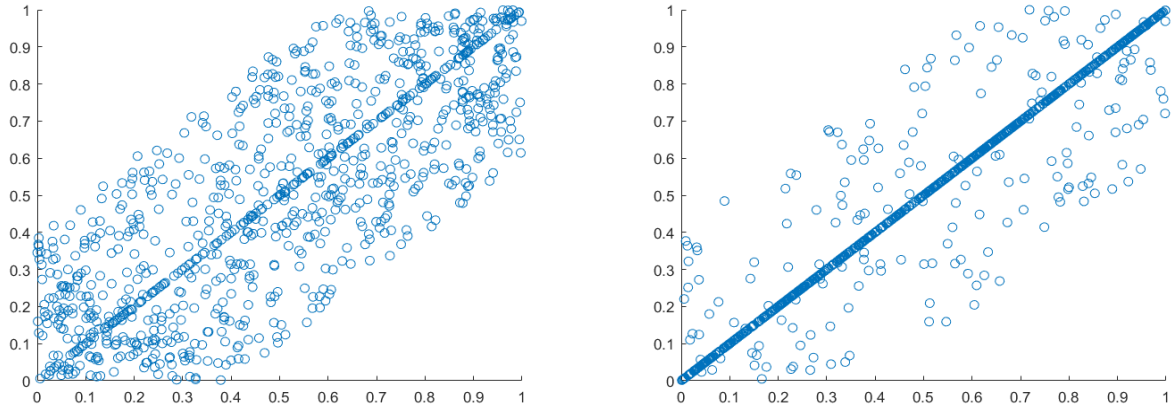

Figure 3.8: 1000 realizations of the combination of identity and diagonal band copula for $\delta = 0.4$ and $\alpha = 0.2$ (left) or $\alpha = 0.8$ (right)

Chapter 4

Data Analysis

4.1 Testing Methodology

As part of this analysis, the approach described in section 2.2 is applied to different null hypotheses which each correspond to one of the models described in chapter 3. This means that the following null hypotheses will be tested:

$$H_0^{(1)} : C \in \mathcal{C}_1 = \{C^{ID}\}$$

$$H_0^{(2)} : C \in \mathcal{C}_2 = \{C^{I}\}$$

$$H_0^{(3)} : C \in \mathcal{C}_3 = \{C_\alpha | \alpha \in (0,1)\}$$

$$H_0^{(4)} : C \in \mathcal{C}_4 = \{C_\delta | \delta \in (0,1)\}$$

$$H_0^{(5)} : C \in \mathcal{C}_5 = \{C_{\alpha,\delta} | \alpha \in (0,1), \delta \in (0,1)\}$$

In fact, each hypothesis will be tested multiple times and will be rejected each time if the approximated p-value (as described in section 2.2) has a value of 0.05 or less. Depending on how many times the hypotheses are tested we get a percentage of how many times each hypothesis was rejected.

Given the fact that each test procedure relies on generating random samples, even if the data's underlying copula were in fact described by one of the null hypotheses, then the rejection percentage still would not be zero percent. However, as already stated multiple times, all of the models won't fit perfectly and as such, a deviation of the rejection percentage from zero is to be expected. Consequently, an intuitive benchmark to judge which of the models describes the data best is to compare the rejection percentages to figure out which hypothesis is rejected least often.

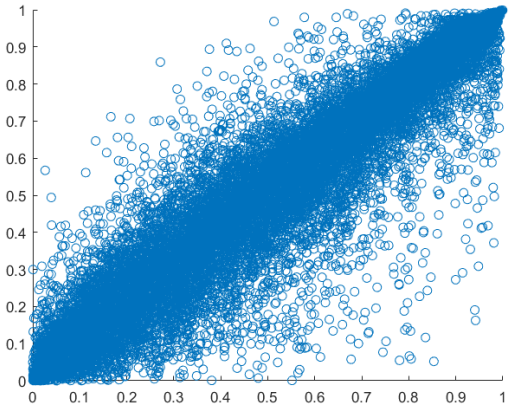

Figure 4.1: Pseudo-observations generated from cloud ice data at heights 10.75 km and 11 km

As part of this thesis, the models are tested on the data which was already illustrated in figure 3.1. For convenience this figure is restated here in figure 4.1. This data set consists of around 16000 data points and unfortunately it is computationally impossible to test all 16000 data points simultaneously. So, instead the tests are performed using random samples from the data of size 50. As the time it takes to run these tests increase more than quadratically with increasing sample size, the data size of 50 was chosen to ensure that as many data points can be tested within a reasonable amount of time.

Once 50 data points are randomly sampled from the data, the hypotheses are tested 200 times each and the p-value of each test is computed using the parametric bootstrap described in section 2.2 while using $N = 200$ repetitions as part of the bootstrap. Here, one should point out that all of these numbers are rather low due to the absence of sufficient computing resources and are mainly chosen this way to ensure the availability of results within the given time frame of this thesis.

As Genest et al. [2] point out, the lack of computing resources is especially unfortunate here because there are no closed formulas for the some of copulas this thesis looks at and one has to rely on simulating a lot of samples as part of the bootstrap. Nevertheless, the described test was run 260 times and the rejection rates after each test gave birth to the box plot in figure 4.2 and the average rejection rates in table

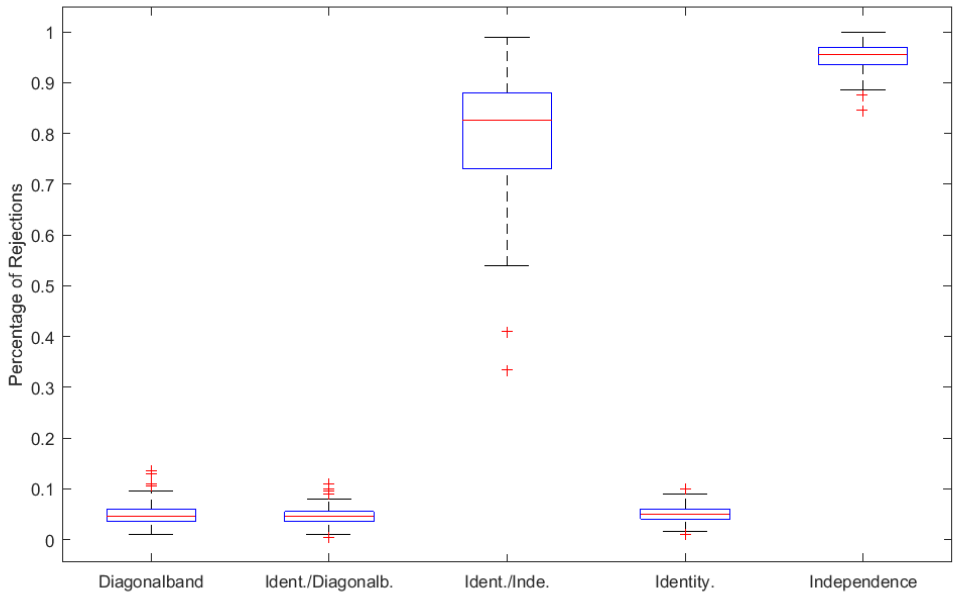

Figure 4.2: Box plot of the rejection rates of each model for 260 test runs

Model	Average rejection rate	Average p-value
Independence	94.86%	0.0103
Identity	5.01%	0.4650
Independence / Ident.	80.11%	0.0445
Diagonal Band	4.91%	0.4849
Diagonal Band / Ident.	4.73%	0.4916

Table 4.1: Average rejection rates and average p-values

4.2 Test Results and Model Improvement

Looking at the results in figure 4.2, one clearly notices that the independence copula is the worst choice among the models. Intuitively, this isn't much of a surprise as figure 4.1 clearly shows a concentration of the pseudo-observations around the main diagonal and one could have immediately concluded that no dependency between the cloud ice of adjacent cloud

layers is unlikely.

Similarly, it doesn't come as a surprise that mixing the independence copula with the identity copula improved the model by lowering its average rejection rate and increasing the average p-value as seen in table 4.1. Further, judging by figure 4.2 and table 4.1, it seems like mixing a copula with the identity copula improves the fit for this data since it improved not only the independence copula but also the diagonal band copula. Specifically, the mixture of identity and diagonal band copulas has a lower average rejection rate, a higher average p-value than the diagonal band copula and in figure 4.2, one can see that adding the identity copula also lowered the variation among rejection rates.

Consequently, one could argue, given the results, that the mixture of diagonal band and identity copula is the best fit (among these choices) for this data. Again, these results need to taken with a grain of salt due to the low sample size and number of repetitions. However, a couple of test runs with a higher sample size didn't seem to alter the general implications of the above calculations. For the future and with more computational power, it would be interesting to see if these results still hold true when sample size and amount of repetitions are increased.

Finally, for the sake of some form of sanity checks, it is worth to have a look at the empirical copulas of the different models and the data to see if there is indeed a difference which the results account for. Specifically, it might be of interest to understand why the diagonal band copula seems to be so much better than the mixture of independence and identity copula. After all, one might say that they follow the same idea by saying that it is important to incorporate the main diagonal in the model and allowing some deviation from this line.

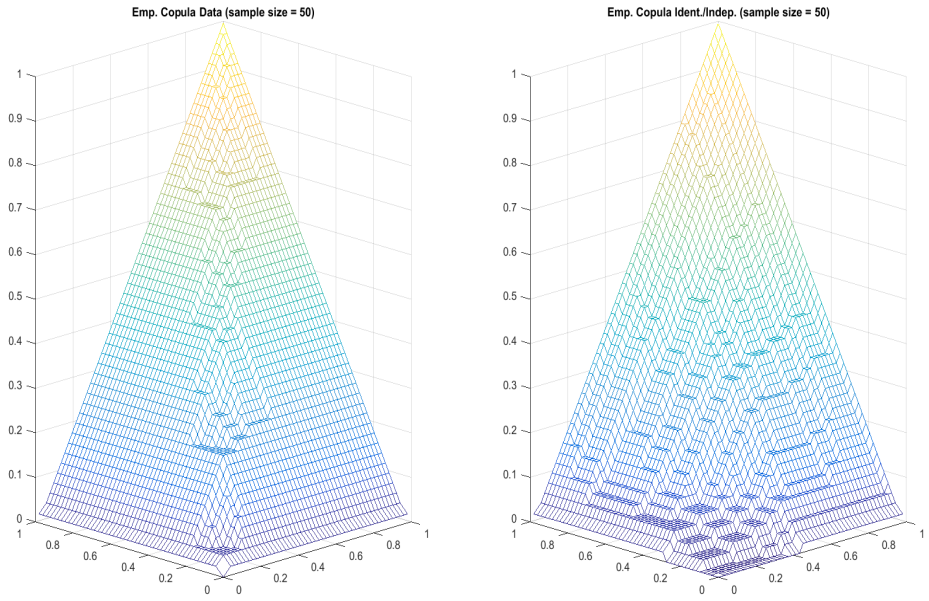

Figure 4.3: Empirical Copula observed from data (left) and from identity / independence copula (right) for sample size 50

To understand this one has to see two things. First, the estimation of the parameter α of the identity / independence copula, as introduced in section 3.1, yields very low values for α. Consequently, this model offers only a marginal deviation from the independence copula itself. As observable in figure 4.2, this has effects on the rejection rates but is still not enough to significantly improve the model.

Second, one should look at the empirical copulas. For that reason, figure 4.3 shows the empirical copula observed from data next to the empirical copula derived from simulations of the identity / independence. Obviously, one should do the same using the diagonal band copula which is shown in figure 4.4. Looking at these two figures, one can already see substantial differences and notice that the empirical copula of the diagonal band copula looks more similar to the data than the empirical copula of the independence / identity copula.

Here one has to compute the empirical copula only once (as opposed to multiple times as part of hypotheses testing), one can even compare empirical copulas for larger sample sizes.

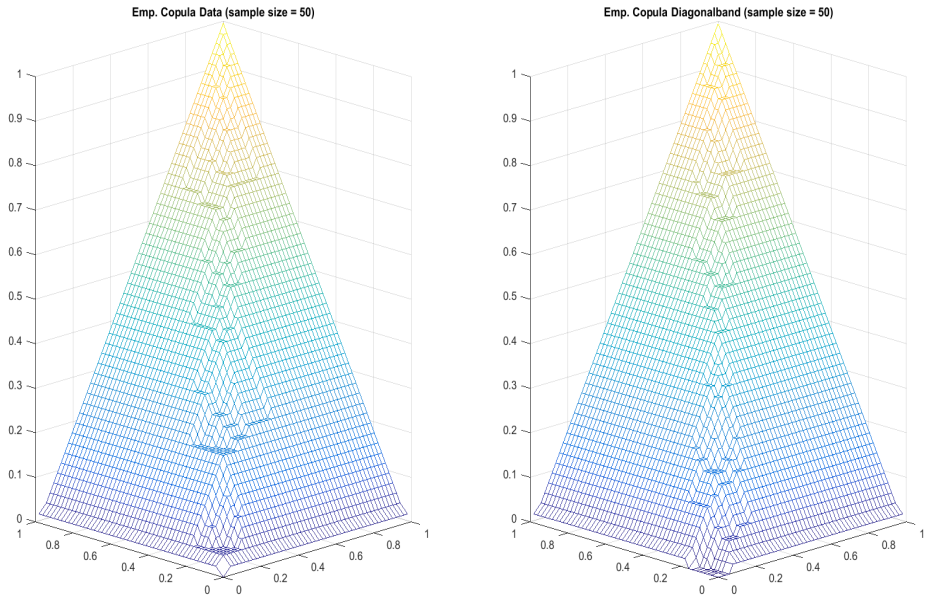

Figure 4.4: Empirical Copula observed from data (left) and from diagonal band copula (right) for sample size 50

However, one needs to change the angle from which to look at this because one wouldn't be able to clearly see differences otherwise. For this purpose, figure 4.5 shows the empirical copulas for different sample sizes by looking at them from the side. Consequently, it becomes apparent that the empirical copula derived from data looks more like that of the diagonal band copula even for larger sample sizes.

As already discussed, the underlying cause for the poor performance of this copula is the low estimate of α. So, this begs the question if the idea of combining the independence and identity copula is a hopeless endeavor altogether or if the issues can be fixed by using higher values of α. Obviously, for $\alpha = 1$ this model simply reduces to the identity copula (which has shown to have reasonable rejection rates) but it might be possible to find a higher value for α which is below 1 to allow for deviations from the main diagonal and, more importantly, make this model a better fit.

A different approach to estimate α is introduced by Hogan and Illingworth [9] who assert

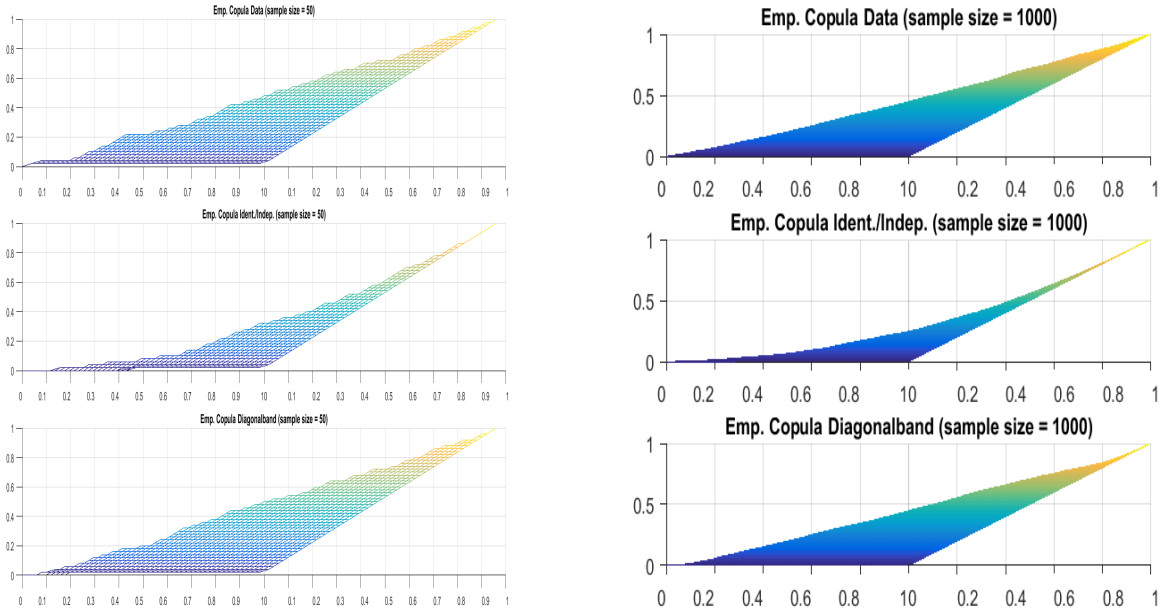

Figure 4.5: Empirical Copula observed from data (top), from independence / identity copula (middle) and from diagonal band copula (bottom) for sample sizes 50 (left) and 1000 (right)

that α can be estimated using

$$\hat{\alpha} = \exp\left(-\frac{\Delta z}{\Delta z_0}\right)$$

where Δz is the level separation given by the data and Δz_0 corresponds to a so-called decorrelation distance which needs to be estimated.

Ovchinnikov et al. [6] establish that it is feasible to reduce the estimation of Δz_0 to solving

$$R = \exp\left(-\frac{\Delta z}{\Delta z_0}\right)$$

where R is the rank-based correlation coefficient of the data. Consequently, another estimate of α is given by

$$\hat{\alpha} = R.$$

For the complete data set, this approach delivers $\hat{\alpha} = 0.9309$. The results of this new

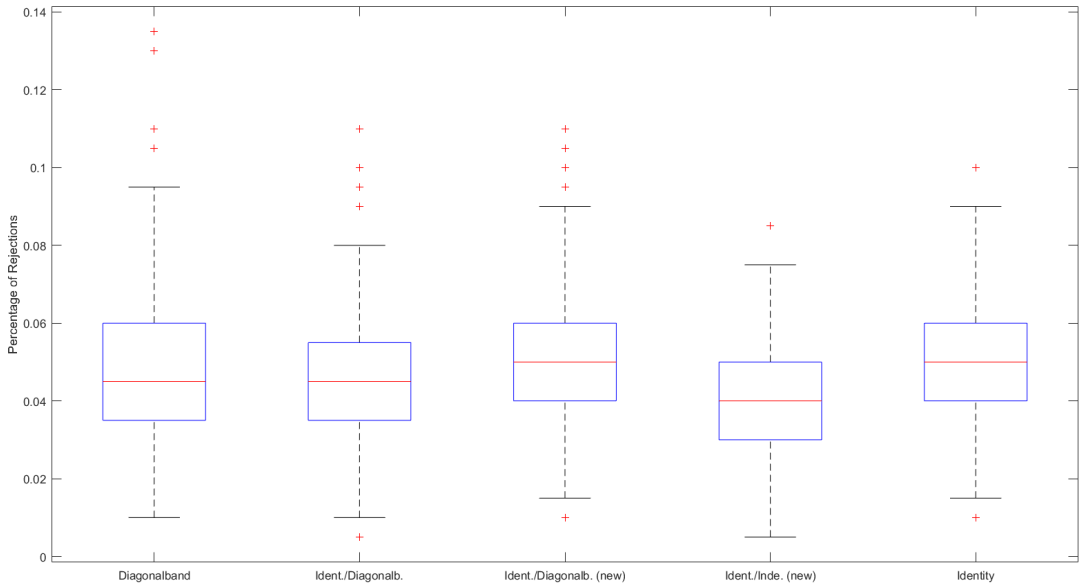

Figure 4.6: Updated box plot using only the relevant copulas and the alternative approach for α

approach are depicted in figure 4.6 and table 4.2. One has to point out that the tests using the new approach for α were not run with the same random samples as the previous tests. Therefore, a variation in their rejection rates and in their p-values might partly be due to effects relating to different random samples. However, the strong improvement of the mixture of identity and independence copula is mainly a result of estimating α differently.

Using the same (simple) benchmark as before, one would have to argue that, given these new results, the best fit is given by the mixture of independence and identity model using the alternative estimation of α. However, one should mention that it might be beneficial to take other considerations into account when ranking the performance of these models. This is nicely demonstrated by the rejection rates of the identity copula.

Looking at the low average rejection rate of the identity copula one might be tempted to say that it is a nice fit to the data. Intuitively, this cannot be correct, so obviously one should find out what is happening here. The most promising theory here is that the test statistic S_n is not sensitive enough to the degenerative nature of the identity copula as it

Model	Average rejection rate	Average p-value
Independence	94.86%	0.0103
Identity	5.01%	0.4650
Independence / Ident.	80.11%	0.0445
Independence / Ident. (new)	3.89%	0.4945
Diagonal Band	4.91%	0.4849
Diagonal Band / Ident.	4.73%	0.4916
Diagonal Band / Ident. (new)	5.05%	0.4659

Table 4.2: Updated average rejection rates and average p-values

only sums the squared errors of empirical copulas. Figure 4.7 compares different simulations of the independence, identity and diagonal band copula to a random sample of size 50. The simulations of the diagonal band copula show that deviations of simulations and data are normal, however due to the nature of the data, the points of the random sample are more likely drawn to be close to the main diagonal. Consequently, the sum of deviations when using the identity copula is not much different than the sum of deviations when using the diagonal band copula. Thus, the tests show similar rejection rates for these two. So, a different test statistic which is more sensitive to this might me more helpful here. In fact, the Kolmogorov-Smirnov statistic might actually be beneficial because it uses the maximum deviation.

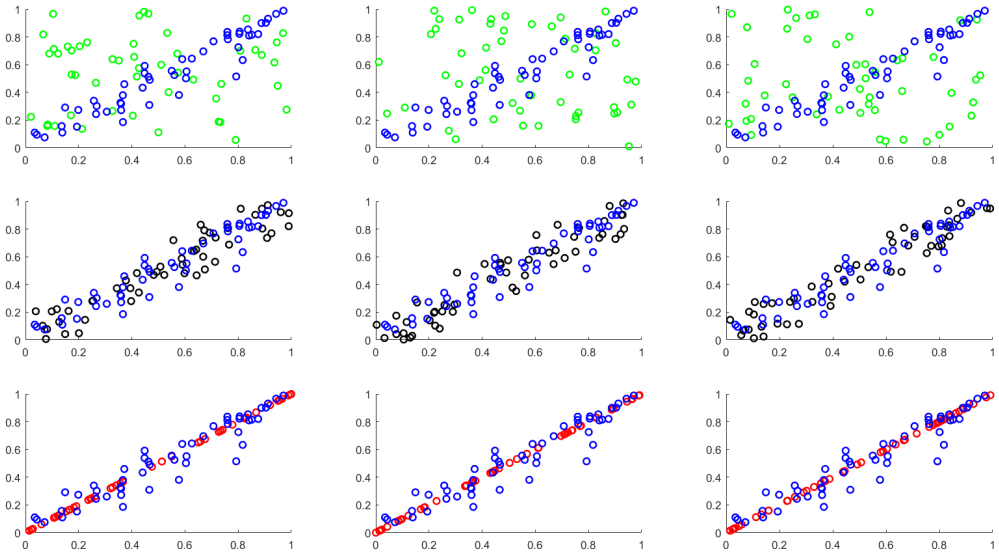

Figure 4.7: Comparing different simulations (independence in green, diagonal band in black and identity in red) to one data sample (in blue)

Chapter 5

Summary

Introducing copulas and their goodness-of-fit tests, this thesis established a testing methodology for comparing different models of dependency between quantities of interest. In this matter, copulas have proven themselves as a valuable tool because they made it possible to make inferences on the dependence structure without having to look at the marginal distributions.

In fact, this thesis showed that the established methods could be applied to a data set from the field of atmospheric science by offering copulas as a means to model cloud overlap. As part of this application, this thesis used the data at hand to compare prominent models which are frequently used to model cloud overlap. The results from that comparison showed that goodness-of-fit tests are able to offer some insight into how well a model fits the data but it also showed that the rejection rates of said tests cannot tell the whole truth and should not be the only consideration when determining which model to use.

One should point out that using copula-based goodness-of-fit tests also came with some complications. First, it turned out to be extremely computationally expensive to test copulas which don't have a closed analytical form. Second, estimating the copula's parameter is a vital step in the test and showed that different methods of estimation can lead to (vastly) different results.

Nevertheless, copulas model dependencies in a very elegant way and if one is solely interested in the dependence structure between quantities, copulas are a worthwhile tool to investigate that despite the computational intensity for testing purposes.

Bibliography

[1] Roger B. Nelson. *An Introduction to Copulas*. Springer Series in Statistics. Springer New York, 2006.

[2] Christian Genest, Bruno Rémillard, David Beaudoin. "Goodness-of-fit tests for copulas: A review and a power study". In: *Insurance: Mathematics and Economics* 44, Issue 2 (April 2009), pp. 199–213.

[3] Abe Sklar. "Fonctions de répartition à n dimensions et leurs marges". In: *Publ. Inst. Statist. Univ. Paris* 8 (1959), pp. 229–231.

[4] Jean-David Fermanian, Dragan Radulovic, Marten Wegkamp. "Weak convergence of empirical copula". In: *Bernoulli* 10, Number 5 (2004), pp. 847–860.

[5] Christian Genest, Bruno Rémillard. "Validity of the parametric bootstrap for goodness-of-fit testing in semiparametric models". In: *Annales de l'Institut Henri Poincaré - Probabilités et Statistiques* 44, No. 6 (2008), pp. 1096–1127.

[6] Mikhail Ovchinnikov, Kyo-Sun Sunny Lim, Vincent E. Larson, May Wong, Katherine Thayer-Calder, Steven J. Ghan. "Vertical Overlap of probability density functions of cloud and precipitation hydrometeors". In: *Journal of Geophysical Research: Atmospheres* 121, Issue 21 (November 2016).

[7] Petri Räisänen, Howard W. Barker, Marat F. Khairoutdinov, Jiangnan Li, David A. Randall. "Stochastic generation of subgrid-scale cloudy columns for large-scale models". In: *Quarterly journal of the royal meteorological society* 130, Issue 601 (July 2004), pp. 2047–2067.

[8] V. E. Larson, D.P. Schanen. "The Subgrid Importance Latin Hypercube Sampler (SILHS): a multivariate subcolumn generator". In: *Geosci. Model Dev., 6, 1813-1829, doi:10.5194/gmd-6-1813-2013* (2013).

[9] Robin J. Hogan, Anthony J. Illingworth. "Deriving cloud overlap statistics from radar". In: *Quarterly journal of the royal meteorological society* 126 (2000), pp. 2903–2909.

[10] Robert Kopocinski. "Simulating dependent random variables using copulas. Applications to Finance and Insurance." In: *Master's thesis at Wroclaw University of Technology* (2007).

-

Appendix: Matlab-Codes

Simulation Routines

Listing 1: Simulation of diagonal band copula

```
function [U1,U2]=DiagonalBand(N,delta_U)
%Variables: delta_U = band half-width
%           N = number of simulations
%           U1 = realizations of U(0,1)
%           U2 = realizations of U(-delta_U,delta_U)


%Simulate the Realizations
U1=rand(N,1);
U2=U1+unifrnd(-delta_U,delta_U,N,1);

%Mirror those that are outside of unit square
upper_end=(U2>1)*2-(U2>1).*U2;
lower_end=-(U2<0).*U2;
U2=lower_end+(U2>=0 & U2<=1).*U2+upper_end;
end
```

Listing 2: Simulation of identity / diagonal band copula

```
function [U1,U2] = MaxDiagonal(N,delta_U,alpha)
%Variables: delta_U = band half-width
%           N = number of simulations
%           alpha= probability to use maximum
%           U1 = realizations of U(0,1)
%           U2 = result of mixing

%Simulate the Realizations
U1=rand(N,1);
%Create dummy to determine for each case if we use max or rand
dummy=rand(N,1);
U2=(dummy>alpha).*(U1+unifrnd(-delta_U,delta_U,N,1))+...
(dummy<=alpha).*U1;

%Mirror those that are outside of unit square
upper_end=(U2>1)*2-(U2>1).*U2;
lower_end=-(U2<0).*U2;
U2=lower_end+(U2>=0 & U2<=1).*U2+upper_end;
end
```

Listing 3: Simulation of identity / independence copula

```
function [U1,U2]=MaxRand(N,alpha)
%Variables: N  = Number of Simulations
%           U1 = N realizations of U(0,1)
%           U2 = N realizations of Mixing U1(rand) and max
%           alpha = probability to use max approach

%Create first uniformly random realizations
U1=rand(N,1);
%Create dummy to determine for each case if we use max or rand
dummy=rand(N,1);
%Mix max and rand according to dummy variable
U2=(dummy<=alpha).*U1+(dummy>alpha).*rand(N,1);
end
```

Auxillary Routines

Listing 4: Compute pseudo observations

```
function [U1,U2]=ComputePseudoObservations(data1,data2)
%Computes pseudoobservations of the copula

[f1,x1]=ecdf(data1);
[f2,x2]=ecdf(data2);

n=length(data1);
U1=zeros(n,1);
U2=zeros(n,1);

for i=1:n
        U1(i)=n/(n+1)*f1(find(x1==data1(i),1));
        U2(i)=n/(n+1)*f2(find(x2==data2(i),1));
end
end
```

The following code was published online by Robert Kopocinski [10].

Listing 5: Computation of empirical copula

```
function ecop = ecopula(x)
%ECOPULA Empirical copula based on sample X.
%ECOP = ECOPULA(X) returns bivariate empirical copula.
   Extension to n dimensional empirical copula is
   straightforward.
%Written by Robert Kopocinski, Wroclaw University of Technology
   , for Master Thesis: "Simulating dependent random variables
   using copulas. Applications to Finance and Insurance".Date:
   2007/05/12
%
%Reference:
%[1]  Durrleman, V. and Nikeghbali, A. and Roncalli, T. (2000)
   Copulas approximation and          new families, Groupe de
   Recherche Operationnelle Credit Lyonnais

[m n] = size(x);
y = sort(x);
ecop=zeros(m,m);

for i=1:m
for j=1:m
        ecop(i,j) = sum( (x(:,1)<=y(i,1)).*(x(:,2)<=y(j,2)) )/m
           ;
end
end
```

Listing 6: Evaluation of empirical copula

```
function [C_val] = EvaluateEcop(C,x,y,x_i,y_j)
%Evaluates Empirical Copula C which has grids x & y at the
   points x_i, y_j
x=sort(x);
y=sort(y);
x_tmp=find(x<=x_i,1,'last');
y_tmp=find(y<=y_j,1,'last');
if isempty(x_tmp) || isempty(y_tmp)
        C_val=0;
else
        C_val=C(x_tmp,y_tmp);
end
end
```

Listing 7: Approximate test statistic

```
function S= ApproxS(C_n,C1,C2,B_n,B1,B2)
%Approximatex S_n by finding maximal absolute difference of emp
   . copulas. C_n and B_n evaluated at (C1,C2)
S=0;
n=length(C1);
for i=1:n
        C=EvaluateEcop(C_n,C1,C2,C1(i),C2(i));
        B=EvaluateEcop(B_n,B1,B2,C1(i),C2(i));
        S=S+(C-B)^2;
end
end
```

Estimation Routines

Listing 8: Estimation of alpha

```
function [alpha] = EstimateAlpha(data1,data2)
%Use small delta to detect small deviations from diagonal
eps=0.0005;
n=length(data1);
alpha=sum(data1<=data2+eps & data1>=data2-eps)/n;

end
```

Listing 9: Estimation of delta

```
function [delta] = EstimateDelta(data1,data2)
%Estimates half-width delta by calculating number of
   oberservations
%in (A|B2|C) and either using local maximum or linear
   Interpolation

n=length(data1);
[U1, U2]=ComputePseudoObservations(data1,data2);
u=(n+1)/n*U1;
v=(n+1)/n*U2;

delta=0;
n_max=0;
delta_step=0.001;
deltas=0:delta_step:1;
numb_in_band=zeros(1,length(deltas));
```

```
15
16 for i=1:length(deltas)
17         theta_tmp=deltas(i);
18         A=(u >= 0 & u<=theta_tmp) .* (v >=0 & v <= theta_tmp-u)
              ;
19         C=(u >= 1-theta_tmp & u<=1) .* (v >=2-theta_tmp-u & v
              <=1);
20         B2=(u > theta_tmp & u<1-theta_tmp) .* (v >=u-theta_tmp
              & v <= theta_tmp+u);
21         numb_in_band(i)=sum(A | B2| C);
22
23         if numb_in_band(i)>n_max  && theta_tmp<=0.5
24                 n_max=numb_in_band(i);
25                 first_max=theta_tmp;
26         end
27 end
28
29 if numb_in_band(deltas==first_max)<0.49*n
30         %limit data to thetas >= x_cutoff (maybe not 0.5 to
              consider bad data)
31         x_cutoff=0.5;
32         numb_in_band=numb_in_band(deltas>=x_cutoff);
33         deltas=deltas(deltas>=x_cutoff);
34
35         %Arbitrary high number to start with
36         diff_min=9999999999;
37
```

```
38          %Interpolate linearly and find minimal deviation
39          diff_tmp=zeros(length(deltas),1);
40          for i=1:length(deltas)
41                  xi= [x_cutoff deltas(i) 1];
42                  yi1=linspace(numb_in_band(1),numb_in_band(i),
                       length(x_cutoff:delta_step:xi(2)));
43                  yi2=linspace(deltas(i)*n,numb_in_band(end),
                       length(xi(2)+delta_step:delta_step:xi(3)));
44                  yi=[yi1,yi2];
45                  diff_tmp(i)=sum((numb_in_band-yi).^2);
46                  if diff_tmp(i) < diff_min
47                          diff_min=diff_tmp(i);
48                          delta=xi(2);
49                  end
50          end
51  else
52  delta=first_max;
53  end
54  end
```

Test Routines

All test routines are basically the same. Every routine just needs to be adjusted to the corresponding model and the associated parameter estimation. For the sake of avoiding repetition and the fact that all necessary changes are obvious, only the routine for the diagonal band copula is shown.

Listing 10: Test routine for diagonal band copula

```
function [p_sum, count_rejections]=TestData1(fileID,U1,U2,
   N_simulations,data_length,N_pvalue)
%Step 1
delta=EstimateDelta(U1,U2);
C_n=ecopula([U1 U2]);
%
   ------------------------------------------------------------------


count_rejections=0;
p_sum=0;
tic
for simulation_run=1:N_simulations

        %Step 2 in Bootstrap
        [V1, V2]=DiagonalBand(data_length,delta);
        B_star=ecopula([V1 V2]);
        S=ApproxS(C_n,U1,U2,B_star,V1,V2);

        %Step 3 in Bootstrap
        count_pvalue=0;
```

```
18          for k=1:N_pvalue
19                  %Step 3a
20                  [Y1_star,Y2_star]=DiagonalBand(data_length,
                        delta);
21                  %Step 3b without estimation part
22                  [U1_star,U2_star]=ComputePseudoObservations(
                        Y1_star,Y2_star);
23                  C_star=ecopula([U1_star,U2_star]);
24                  delta_star=EstimateDelta(U1_star,U2_star);
25                  %Step3c
26                  [Y1_d_star,Y2_d_star]=DiagonalBand(data_length,
                        delta_star);
27                  B_d_star=ecopula([Y1_d_star,Y2_d_star]);
28                  S_star=ApproxS(C_star,U1_star,U2_star,B_d_star,
                        Y1_d_star,Y2_d_star);
29                  if S_star > S
30                          count_pvalue=count_pvalue+1;
31                  end
32          end
33          p_value=count_pvalue/N_pvalue;
34          p_sum=p_sum+p_value;
35          %Reject Nullhypothesis if p-value < 0.05
36          if p_value<0.05
37                  count_rejections=count_rejections+1;
38          end
39  end
40  time=toc;
```

```
41 fprintf(fileID,'Used␣model:␣Diagonalband␣with␣parameter␣%1.2f␣\
     n',delta);
42 fprintf(fileID,'Samplesize:␣%i␣\n',data_length);
43 fprintf(fileID,'Number␣of␣simulations␣per␣model:␣%i␣\n',
     N_simulations);
44 fprintf(fileID,'Number␣of␣simulations␣per␣p-value:␣%i␣\n',
     N_pvalue);
45 fprintf(fileID,'Percentage␣of␣rejections:␣%1.2f␣\n',
     count_rejections/N_simulations);
46 fprintf(fileID,'Averarge␣p_value:␣%1.2f␣\n',p_sum/N_simulations
     );
47 fprintf(fileID,'Time:␣%i␣minutes␣and␣%i␣seconds␣\n',floor(time
     /60),mod(floor(time),60));
48 end
```