On the relation between cloud top temperature and precipitation.

Daniëlle Ilona Post

Technische Universiteit Delft

TU Delft

On the relation between cloud top temperature and precipitation.

by

Daniëlle Ilona Post

to obtain the degree of Bachelor of Science
at the Delft University of Technology,
to be defended publicly on Tuesday July 11, 2019 at 12:30 PM.

Student number:	4448626	
Project duration:	February 18, 2019 – July 11, 2019	
Thesis committee:	Prof. dr. ir. A. W. Heemink,	TU Delft, supervisor
	Dr. S. Lu,	TU Delft, supervisor
	Dr. J. -J. Cai,	TU Delft
	Dr. B. van den Dries,	TU Delft

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

TU Delft

Preface

This bachelor's thesis has been an educational experience and hereby completes my mathematical career. As I aim to broaden my academic path, my studies will continue in another field from now. I want to thank everyone close to me that has supported and inspired me during the course of this project. Special thanks are in place for my parents who have supported me endlessly and indispensably throughout the whole of my studies.

For this thesis in particular I want to thank prof.dr.ir. A.W. Heemink and dr. S. Lu for supervising my work and the process we went through. Moreover, I want to thank dr. J.-J. Cai and dr. B. van den Dries for being part of the committee for this thesis.

Daniëlle Ilona Post
Delft, July 2019

Abstract

This thesis provides insights into the relation betwen two vital elements within the physical precipitation system: *cloud top temperature* (CTT) and *precipitation*. It investigates gauge data from ground stations in Germany and satellite data on CTT, targeted at the same stations in Germany. Cloud and rainfall characteristics are considered and both simple and advanced mathematical tools are covered.
Statistical methods show that the data present is not stationary. Stationary data is required for applying time series analysis, either in the time or frequency domain. The key step in this research is dividing the big dataset into smaller subsets based on the provided characteristics: elevation, region and time. For these subsets the Pearson correlation and Spearman's rank correlation coefficient show that significant correlations can be found. Lastly, this thesis looks into the possibilities for setting thresholds for cloud top temperature in relation to the precipitation. These thresholds should be able to tell whether a day can be considered 'dry' or 'wet'. Due to the incoherent data, these thresholds are not very optimal. Many circumstances need to be considered in order to make these thresholds fit specific situations.

Contents

List of Figures

List of Tables

1 Introduction

1.1. Weather data

In the Netherlands, almost every part of our lives is in some way related to the weather. We depend on forecasts for travelling, working and planning activities. Also the infrastructure of our lives, our food and other supplies are depending on weather related issues.
The number of weather stations on our planet is decreasing [22]. Weather stations are essential for forecasts, modelling and climate-, hydrology- and weather related research. Lack of proper data indirectly affects agriculture, the ability to give out proper severe weather warnings, water resources management and infrastructure.
Due to this decline and due to an increase in extreme weather, weather analysis, research and forecast is complicated and the accuracy might decrease with fewer available data. Placing more stations and maintaining old ones could reverse this trend. High costs and time consuming maintenance will be required for modern technological models. In sub-Saharan Africa this is not at all a priority. Therefore, the scarcity of stations, data and its results will continue.

Apart from the general forecasts, there is more potential in the availability of numerous and accurate data.
In Eastern Africa, about 80% of the citizens depend on agriculture for their livelihood [4]. With better and more inclusive forecasts, their seeding and harvesting plans can be optimised for a more efficient flow of resources, work and money. Moreover, (global) climate models will improve, insurances can be set up, irrigation can be improved and more insights can be provided for all weather-service related industries.
The presence of weather stations and data also creates training possibilities for local students and scientists for future continuity. Satellite data is often combined with ground data for even more insights and possibilities.

1.2. Trans-African Hydro-Meteorological Observatory

In 2014 the Trans-African Hydro-Meteorological Observatory (TAHMO) was set up. TAHMO is a non-governmental organisation in the Netherlands run by participants from Delft University of Technology and Oregon State University. The organisation aims to place 20,000 frugal weather stations across sub-Saharan Africa "to understand better Africa's environment through participatory sensing, scientific modelling and education" (R.Hut & N. van de Giesen, 2010 [7]). This potential network of 20,000 stations will become "the largest network of common sensor systems in the world" (N. van de Giesen, R. Hut and J. Selker, 2014 [22]).
Within TAHMO there is research being done and various collaborations are set up with other weather institutes, insurance organisations and several big companies such as IBM.

1.3. Research

Plenty of research is carried out based on data that is available on a greater scale right now: satellite data. Even the areas that weather stations will never reach can be observed by satellites. The observations provide a lot of data on various weather, air and cloud characteristics. When there is accurate knowledge on the possibilities of predictions based on satellite data, providing forecast, doing research and building models for

possibly every place on earth will improve.
Within the broad subject of weather, precipitation is a vital meteorological variable that is essential in many of the mentioned social economic activities. However, there are many difficulties in the accurate measuring of precipitation.

Within the faculty of Civil Engineering (CITG), TU Delft, there are multiple researchers who are looking into these issues. CITG (dr. Sha Lu) is currently studying the relationship between the time series of gauge precipitation and the time series of satellite-based cloud top temperature (CTT) at different time scales [18]. Cloud characteristics relate to precipitation frequencies and magnitudes. To validate relations, models and ideas, Lu's study is based on two regions: Germany and Tanzania. With a surface area 2.5 times smaller then Tanzania, Germany has about 123 times more weather stations, namely 1965 stations as opposed to 16 stations in Tanzania in total [5].

In Germany there is a high density of ground weather stations. At the same time there is a lot of satellite data available (targeted at the locations of these weather stations). These ground and satellite time series should be able to provide us with relations between the different characteristics. Once this is validated we can try and apply it to Tanzania where similar satellite data is available while ground data is scarce.
There are 941 Climate Data Center (CDC) weather stations located throughout the whole of Germany, provided by the Deutscher WetterDienst (DWD). They provide time series from the years 2000-2015. The precipitation is measured over an interval from 6 am to 6 pm and longitude, latitude, elevation and date are given. Gridded satellite data from the years 2000-2015 is provided by CLARA-A2.
CITG (dr. Sha Lu) worked on the pre-processing of the datasets, monotonic and linear correlation between the CTT and precipitation time series, on non-linear indirect correlations and on a weather generator (a rainfall model). She extracted the time series including the cloud characteristics and other values corresponding to the CDC stations.

CITG (dr. Sha Lu) has, more specifically, been working on the possible relations between precipitation and cloud top temperature. Cloud top height, cloud top pressure, cloud phase, liquid water path and ice water path have also been explored. Correlations for these time series between CTT and precipitation have been studied by Lu. The time series are available on different time scales. At monthly, dekadal (10-day interval) and pentadal (5-day interval) scale the CTT and precipitation can be linearly correlated over Tanzania. However, at a daily scale the correlation seems to be strongly non-linear and indirect. Also, these correlations are not as high in Germany.

1.4. Objectives

The first objective for this research is to study the possible relationship and correlation among CTT and precipitation. Different statistical methods can be used to analyse the time series, both separately and joined together. The advanced theories for time series and spectral analysis can possibly provide more details on the relation between these two essential features. Time series will sometimes be referred to as 'TS'. Chapter 2 and 3 provide preliminary information on physical weather characteristics and mathematical tools respectively. Chapter 4 and 5 elaborate on the approach and use of this knowledge.

Secondly, the CTT will be studied on itself with the aim to find convenient thresholds. When do we consider a day to be 'rainy' or 'wet' depending on both the gauge measurements and the CTT? First of all, the data has to be observed thoroughly for exact thresholds. Afterwards, different thresholds can be estimated based on accuracy or significance requirements. Real life consequences of the rain need to be considered, especially in case of the sub-Saharan agricultural life. In chapter 6 the approach of this and the additional results are described. In chapter 7 the research will be concluded.

2

Precipitation preliminaries

For proper analysis of the time series, knowledge of the physical elements is vital. This chapter will describe hydrometeors and the research that follows from this knowledge. All particles in our atmosphere that consist of water in any form, are said to be hydrometeors [14].

2.1. Hydrometeors

There are different states for these particles: suspended, falling (precipitation) or being blown by the wind. Water particles found on the ground such as snow are not hydrometeors.

2.1.1. Clouds

Clouds are a fundamental part of our global climate and of all weather processes. They are responsible for the transport and change of heat, for all forms of precipitation and they influence solar and infrared radiation in complex ways.[10]
Clouds, hydrometeors consisting of liquid water or ice particles suspended in the atmosphere, appear in many different forms and at about 3 different heights within the troposphere. The different forms are divided into 10 types.[3][19][14]

The highest level clouds are found at 5 to 12 kilometres altitude and come in three different forms. Generally they are thin, fuzzy and white. However, when the sun is set low, they can show other colours through these rays of sunlight. Due to the height of these clouds, they can only consist of ice crystals.
Cirrus clouds (figure 2.1a) are the most transparent form of clouds. The ice crystals within are separated, which creates the transparency and detachment that the cirrus clouds are known for. Cirrus clouds appear in separate transparent bands and layers that hardly affect the sunlight going through. Near the horizon they turn up yellowish because of the distance from the sunlight. And during sunset and sunrise, cirrus clouds are responsible for magnificent colours. Clouds of this type never produce any precipitation but can be a warning for a potential storm.
Cirrocumulus clouds (figure 2.1b) have a similar appearance to cirrus clouds. The main (visual) difference is the way they appear in smaller patches then cirrus do. Cirrocumulus form the same bands and layers but are divided into many small patches. These clouds are not often seen; but both cirrus and cirrostratus clouds can transform into these them.
Clouds that look like cirrus but instead fully cover the sky are known as *cirrostratus* clouds (figure 2.1c). They are dense like cirrocumulus but instead form one smooth layer. Cirrostratus clouds appear as white sheet covering the whole globe. Some sunlight will still pass through it. These clouds can indicate that rain or even snow is coming.

The middle level of clouds are found at an altitude of 2 to 7 kilometres and come in three different forms. These clouds generally consist of water droplets. However, when temperatures get low enough, they will become ice crystals. *Altocumulus* clouds (figure 2.2a) will often appear together with other forms of clouds and with layers at different heights. The altocumulus clouds can be distinguished by the more dense layers or rolls of patches. They are not found in a weather front but do produce precipitation sometimes.

(a) Cirrus (b) Cirrocumulus (c) Cirrostratus

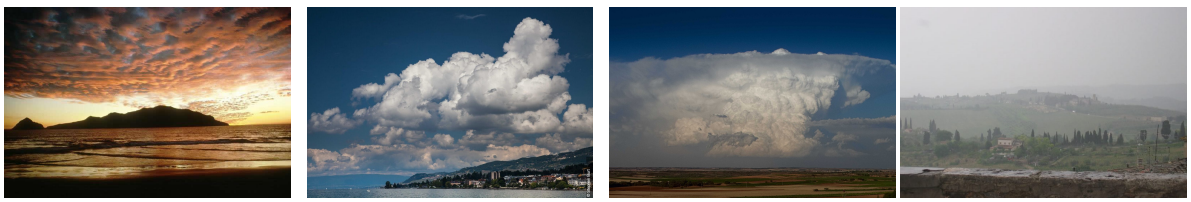
Figure 2.1: High level clouds (Sources:(a)(c)[14],(b)[6])

(a) Altocumulus (b) Altostratus (c) Nimbostratus

Figure 2.2: Mid level clouds(Sources:(a)[6],(b)&(c)[14])

Altostratus clouds (figure 2.2b) are a more dense and more grey version of cirrostratus clouds. Also, they bring light forms of precipitation as they are formed in a front. The sun can still be seen as a small shiny circle but as the precipitation intensifies, the clouds may transform into *nimbostratus* clouds (figure 2.2c). Nimbostratus clouds occur at the lowest level of the middle level clouds, at 2 kilometres height. However, during the precipitation they might lower more and disturb visibility.

The lowest level of clouds can rise up to 2 kilometres and come in four main forms. These clouds, just like the middle level clouds, consist solely of water drops. These lower clouds generally look more soft, silky and dense. *Stratocumulus* clouds (figure 2.3a) come in various forms but can generally be described as big dense separate patches in rolls and layers. Rays of sunlight appear in between but will never shine straight through them. Stratocumulus clouds can form light precipitation.
Cumulus clouds (figure 2.3b) are a bigger, more separate form of the stratocumulus clouds. Also, they appear brighter and more white due to the sunlight that is able to reach different sides of these clouds. They will mostly be found on bright days with blue skies.
Cumulonimbus clouds (figure 2.3c) occur in vertical form. From the ground, (dark) grey dense but often smooth and flat masses of clouds can be seen. Higher up they will tower into the sky. These clouds bring darkness and precipitation which might be very intense, including hail, thunder and heavy winds.
Lastly, *stratus* clouds (figure 2.3d) are similar to the earlier described nimbo- and altostratus clouds. However, they are more dense and will sometimes even touch the ground, which we call fog. They often produce lighter forms of precipitation but can also carry snow and ice.



(a) Stratocumulus (b) Cumulus (c) Cumulonimbus (d) Stratus

Figure 2.3: Low level clouds(Sources:(a)&(d)[6],(b)&(c)[14])

2.1.2. Precipitation

The WMO Cloud Atlas [14] defined precipitation as "a hydrometeor consisting of a fall of an ensemble of particles. The forms of precipitation are: rain, drizzle, snow, snow grains, snow pellets, diamond dust, hail and ice pellets".
This research is based on gauge data. The millimetres of water measured by the gauge can be a result of any form of precipitation. Precipitation, specifically rain, can appear in 3 main forms: frontal rain, orographic rain and convective rain.[12][14]

Frontal rain is created by two different masses of air. Tropical air, which is warm and less dense, meets a mass of polar air which is cold and very dense. The warm mass lifts over the cold and forms what we call a front: the boundary of two of such air masses. In this front the warm air cools down so that what was water vapour before, will now fall down as raindrops. At this front is where the heaviest rainfall occurs. Frontal rain showers can last for many hours. This rain is very common in the UK, the Netherlands and parts of Germany.

Orographic rain arises near mountainous land close to the coast. When a mass of air carries a lot of moisture and is forced to move upwards over high grounds, clouds are produced. Here precipitation occurs. Wind from the sea often contains a lot of moisture and is therefore often the cause of orographic rain.
When orographic and frontal rain meet, the following happens: rain falling from high frontal clouds, or 'seeder' clouds, arrives at the lower moist air, and take those droplets with them down from this 'feeder' cloud. As a result the amount of rain can be increased ten times over these high grounds.

Convective clouds are formed by bubbles of floating air that rise upwards, cool down and condense into the clouds described in the previous paragraph, such as cumulus and cumulonimbus. This occurs when the air heats up from below by land or sea. It therefore takes place above the land during summer and above the sea during winter. From these convective clouds *convective rain* is produced. Convective rain showers often last between 20 and 60 minutes and change rapidly in intensity. The surface area of these showers will be small since the cumulonimbus clouds are narrow. The phenomenon 'sunshine and showers', described by the alteration of short dry and wet periods, takes place here. Convective rain is very common in tropical areas.

2.2. Cloud properties

Precipitation is the main weather characteristic measured on the ground that will be useful for this research. Alongside the gauge measurements satellite data will be used. Satellites can measure many other features, which will be summed up in the following paragraphs.[14][19]

Cloud top temperature (CTT) is defined as the temperature that is measured by satellites at the upper surface of a cloud.

2.2.1. Cloud water content properties

There are different properties that concern the water content of clouds. Clouds can consist of water and ice particles. The *liquid water path* (LWP) or *ice water path* (IWP) measures the amount of water content in liquid or ice form, respectively, in a column within a cloud or the atmosphere. LWP can be defined as the integral over the liquid water content. The LWP in a lower level cloud would generally measure 20 up to 80 gm^{-2}.
Both the albedo of a cloud and the absorption of radiation depend on the LWP. This is based on the relation between LWP and the cloud optical depth.
Another water content property is the *water vapour column density*, which is the density of water vapour in a column with an infinitesimal base.

2.2.2. Cloud optical properties

Due to the composition of our atmosphere and the characteristics of moist and other elements in our atmosphere, not all wavelengths of the electromagnetic spectrum can reach the Earth. Those that can, together make up the *atmospheric window*.
A surface where solar radiation reaches can be characterised by its *albedo*. Surface albedo is the ratio of the reflected flux density to the incoming flux density. This depends on the properties of the surface and its position relative to the sun.

The *(cloud) optical depth* or *thickness* defines the degree of radiation that passes through a cloud. *Brightness temperature* measures the the radiance of microwave radiation from the Earth's atmosphere. All these optical properties are influenced by the form and phase of the particles within clouds (water, snow or ice).

2.3. Remote sensing of clouds

The focus of this bachelor's thesis will be on the possible relations between precipitation and cloud top temperature specifically.

2.3.1. Existing methods

There exist many projects that use various methods for rainfall estimation, cloud classifications, weather forecasts and other purposes. All of them aim to use the available satellite data to determine the physical relations between various features and characteristics.
The methods are mainly based on either the emission or the scattering of radiation to space from our atmosphere and the clouds or ground within this atmosphere[10]. The emission-based methods can extract information from both infrared (IR) and microwave (MW) radiation. The scattering-based methods use the scattering of microwave radiation as the source for information.

Emission-based methods[10]
The emission-based methods have limited information for measurements: only information from the upper layers of clouds can be retrieved.

The absorption and emission of IR radiation by ice crystals of certain sizes and at certain wavelengths in the atmospheric IR spectrum can be used to determine the optical thickness of cirrus clouds and the size of their particles. This method is referred to as the split-window method. The difference in the brightness temperature of the radiance of a specified channel is related to the difference in the optical depth in specified wavelengths which is then proportional with the size of the ice particle.

The emission of MW radiation of clouds and water vapour at certain frequencies is useful for the estimation of the liquid water path of clouds. However, this method has been solely applied over oceans and seas until now. By measuring at frequencies near the water vapour absorption line (22 GHz) and at a higher frequency (e.g. 35 GHz) and by retrieving the MW brightness temperature and optical depth the water vapour column density and LWP can be derived.
These measurements are not possible yet above the land; the MW emission from the ground is too variable and too big to detect the MW emission from the clouds with the current instruments.
Many algorithms have been developed to derive precipitation values from MW emission using atmospheric models and radiative transfer equations.

Scattering-based methods[10]
Scattering-based methods generally use the scattering of sunlight and the scattering of MW radiation.
An especially advanced way to retrieve optical properties of clouds is based on the reflection of sunlight. A widely developed method is the bi-spectral reflectance method (BSR). Basically, this method relates the reflection of sunlight to the cloud optical depth and single scatter albedo (the capacity of reflection) by observing layers of clouds consisting of different particle sizes and optical depths.

Rain clouds can consist of ice particles. The high frequencies of MW radiation (> 50 GHz) can be used to observe the scattering effects by these ice particles. At different frequencies with known particle sizes there is a relation between the brightness temperature caused by the scattering and the IWP of those clouds. From these relations information on the precipitation over land can be derived. However, they vary a lot over different regions. Even higher frequencies are used for snow algorithms based on the same information.

All these methods aim to relate radiation to optical and water content properties of clouds and of precipitation. In many of these methods uncertainties arise due to assumptions that are made for micro physical properties of the atmosphere, clouds and of precipitation. According to Greame L. Stephens and Christian D. Kummerow (2007) [10], models might be too simple for the complex relations they describe and are therefore neglecting accurate errors and uncertainties.

3

Mathematical preliminaries

Data analysis can be performed by the use of various statistical tools. For determining possible relationships among CTT and precipitation there are multiple methods. To decide which way to go, the present data must be studied accurately.

3.1. Statistical tools[17][15]

Probability theory is essential in statistics. Basic probability tools are of great relevance.
Let Ω be sample space of all possible events. The *probability* P is a set function that assigns a number $P(A)$ to every possible event A in Ω, such that:

- if $A \in \Omega, P(A) \geq 0$
- $P(\Omega) = 1$ (3.1)
- if for $A_1, A_2, A_3, .. : A_i \cap A_j = \emptyset,$ for $i \neq j,$ then for the finite union $P(A_1 \cup A_2 \cup ... \cup A_i) = P(A_1) + P(A_2) + ... + P(A_i)$ and for the countably infinite union $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

The *conditional probability* of an event A_1 given that event A_2 has occurred can be defined as follows:

$$P(A_1|A_2) = \frac{P(A_1 \cap A_2)}{P(A_2)} \qquad P(A_2) > 0. \tag{3.2}$$

The *probability mass function* (PMF) of a discrete random variable X is a function $P(X = x) = p(x)$ such that:

- $p(x) > 0$ if $X \in \Omega$ (3.3)
- $\sum p(x) = 1$

If $p(x)$ 3.3 is the PMF of the discrete random variable X, then the *expected value* of X, denoted by $E(X)$, is

$$E(X) = \sum_i x_i \, p(x_i) \tag{3.4}$$

provided that $\sum_i |x_i| p(x_i) < \infty$.
The expected values of X, $E(X)$, can also be referred to as the *mean* of X and is often denoted by μ.
If X is a random variable with expected value $E(X)$, the *variance* of X, is defined as

$$Var(X) = E\{[X - E(X)]^2\} \tag{3.5}$$

provided that the expectation exists. The *standard deviation* of X is the square root of the variance.
For large data sets this will be referred to as the *sample mean* $\bar{x}$, *sample variance* σ^2 and *sample standard*

deviation σ instead:

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{3.6}$$

$$\sigma^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + ... + (x_n - \bar{x})^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{3.7}$$

$$\sigma = \sqrt{\sigma^2}. \tag{3.8}$$

Lastly, if X and Y are two random variables with means σ_X and σ_Y. Then the *covariance* of X and Y, denoted by $Cov(X,Y)$ or σ_{XY}, is defined as

$$Cov(X,Y) = \sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)]. \tag{3.9}$$

Which for large data sets will instead be defined as

$$\sigma_{XY} = \sum_{(x,y)\,\epsilon\Omega}\sum (x - \mu_X)(y - \mu_Y) f(x,y). \tag{3.10}$$

Correlation

The correlation coefficient measures the correlation, that is, the strength of a possible statistical relationship, between two variables. There are several types of correlation.
The Pearson correlation coefficient measures the linear correlation. The (Pearson) *sample correlation coefficient* r_{xy}, given there are n pairs of data $\{(x_1, y_1), ..., (x_n, y_n)\}$, can be defined as follows:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}. \tag{3.11}$$

The coefficient takes a dimensionless value between -1 and 1. Both -1 and 1 describe a strong linear correlation, either negative or positive, respectively. If the coefficient takes on 0 there is no linear correlation between the two variables. Interpreting the values between 0 and 1 or -1 highly depends on the context, size and purpose of the data set.

Spearman's rank correlation coefficient

Another type of correlation coefficient is based on the ranking of the variables: ordering the variables based on the size of their values. *Spearman's rank correlation coefficient*, for sample size n and converted x_i and y_i to ranks $rg(x_i)$ and $rg(y_i)$, can be defined as follows:

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{3.12}$$

with $d_i = rg(X_i) - rg(Y_i)$, the difference between the two ranks of each observation. It measures the correlation between the ranks of the variables. The value of the coefficient shows whether the relationship between the two variables can be described by a monotonic function. From a coefficient of 1 we can derive that $x_i - x_j$ and $y_i - y_j$ always have the same sign, and from a coefficient of -1 always the opposite sign. The benefit of the Spearman's rank coefficient over the Pearson correlation coefficient is that the first is less sensitive to outliers.

Likelihood ratio and testing hypotheses

Let there be a parameter θ and an event X. The PMF $p(x)$ depends on parameter θ. The probability that X takes on value x, given θ is $P(X = x|\theta)$ or $P_\theta(X = x)$. Let's now say θ is unknown. We observe that X takes on value x. The *likelihood (function)* $L(\theta|X = x)$ is a function of θ, given that $X = x$.
Let's now take 2 different parameters θ_0 and θ_1 and an event X. The *likelihood ratio* can then be defined as

$$\frac{P(X = x|\theta_0)}{P(X = x|\theta_1)} = \frac{P_{\theta_0}(X = x)}{P_{\theta_1}(X = x)}. \tag{3.13}$$

Lets now define two complementary hypotheses H_0 and H_1 according to the two parameters θ_0 and θ_1.The

prior probabilities for each parameter before observations are made are $P(H_0)$ and $P(H_1)$. After observing $X = x$, the posterior probabilities will turn out to be $P(H_0|x)$ and $P(H_1|x)$ respectively. And

$$P(H_0|x) = \frac{P(H_0, x)}{P(x)} = \frac{P(x|H_0)P(H_0)}{P(x)}. \tag{3.14}$$

As a result, the ratio of the posterior probability is the product of the ratio of prior probability and the likelihood ratio.

$$\frac{P(H_0|x)}{P(H_1|x)} = \frac{P(H_0)}{P(H_1)} \frac{P(x|H_0)}{P(x|H_1)}. \tag{3.15}$$

Based on the prior probability and the likelihood ratio, the parameter with the highest posterior probability will be chosen. H_0 would be chosen if:

$$\frac{P(H_0|x)}{P(H_1|x)} = \frac{P(H_0)}{P(H_1)} \frac{P(x|H_0)}{P(x|H_1)} > 1 \tag{3.16}$$

$$\frac{P(x|H_0)}{P(x|H_1)} > c. \tag{3.17}$$

H_0 would be accepted if the likelihood ratio is greater than c, where c depends upon the prior probability. There are now two options for an error:

Type I Reject H_0 while it was true

Type II Accept H_0 while it was false.

3.2. Regression analysis

Regression analysis originates from astronomical observations in the early 1800's and the term regression was invented by geneticist Sir Francis Galton (1822-1911). The most common and ancient form is the method of least squares.

A regression model is based on *independent* (or *predictor*) variables x_i, the *dependent* (or *response*) variables y_i and unknown parameters b_i. It can be defined as follows, with vectors $\mathbf{x}, \mathbf{y}, \mathbf{b}$:

$$\mathbf{y} \approx \mathbf{f}(\mathbf{x}, \mathbf{b}) \tag{3.18}$$

An essential assumption is $n > k$, that is; the amount of data points n is strictly larger than the amount of unknown parameters b_i ($i = 1, ..k$), so that there is enough information in the data to determine or estimate the unknown parameters. To be specific; we say that the model is overdetermined. There are multiple solutions for b_i. With the method of least squares we are able to estimate the best fit for the parameters.

3.2.1. Linear regression

The predicting value x_i, the response variable y_i and the predicted response value $\tilde{y}_i$ are defined. The *intercept* is b_0: the expected value of $\tilde{y}$ when $\mathbf{x} = 0$, and b_1 is the *slope*.

$$\tilde{y}_i = b_0 + b_1 x_i \tag{3.19}$$

with prediction, or *residual* error

$$e_i = y_i - \tilde{y}_i \tag{3.20}$$

Least squares

The idea of the least squares method is to minimise the residual errors. The line with the lowest error, is the best fit. More specifically, the residual sum of squares will be minimised:

$$Q = \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2 = \sum_{i=1}^{n} (e_i)^2, \tag{3.21}$$

and from there b_0 and b_1 can be estimated (with $\bar{x}$ and $\bar{y}$ the mean of the predictor and response variables, respectively).

$$b_0 = \bar{y} - b_1 \bar{x} \tag{3.22}$$

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{3.23}$$

Residuals

For determining the best fit of a regression line, the plot of residuals can be used. If the plot of the residuals against the predictor values show complete randomness, so no relation between the residuals and the x values, the linear regression model can be confirmed.

Coefficient of determination

The coefficient of determination, R^2, often referred to as R-squared can be defined as the proportion of variation that the predictor values account for. R-squared can formally be defined as $1 - \frac{\text{residual sum of squares}}{\text{total sum of squares}}$

$$R^2 = 1 - \frac{\sum_i (e_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{3.24}$$

It takes on values between 0 and 1. If it takes on value 1, it characterises a perfect regression line. All data points will be found exactly on the line in that case. If it takes on value 0, the regression line is horizontal. In that case, we say that the predictor values x account for none of the variation in y [15]. The size of the data set and the goal of the regression will have to be considered for deciding whether the R-squared value is significant.

Correlation and regression

The correlation coefficient r is related to the R-squared value; $r = \sqrt{R^2}$ if the estimated slope > 0 and $r = -\sqrt{R^2}$ if the estimated slope < 0.

Polynomial regression

Polynomial regression is very similar to linear regression, it is even considered to be linear regression since the regression coefficients $b_0, b_1, .., b_h$ are linear. Polynomial regression is defined as

$$\tilde{y}_i = b + 0 + b_1 x_i + b_2 x_i^2 + ... + b_h x_i^h \tag{3.25}$$

where h is the degree of the polynomial.

Significance tests

Once calculations and estimations have been done, values need to be tested for their significance. In statistics there are many tests for this.
Let's assume that a correlation coefficient r has been found between 0 and 1. To test the significance a null hypothesis and an alternative hypothesis are settled respectively:

H_0 r is **not** significantly different from 0

H_a r is significantly different from 0.

For testing this significance the p-value, a well known statistical value for validating an hypothesis, is often used. In this case, if the null hypothesis is rejected based on the p-value, there can be concluded that r is statistically significant. Prior to the testing there should be decided on a significance level α. Generally α takes on values like 0.05 or 0.01. The p-value then represents the probability that the extreme event happens while the null hypothesis is true. If the p value is found to be lower then the preset significance level α, the null-hypothesis will be rejected. In that case r is significantly different from 0 and thus statistically significant.

3.3. Time series analysis

Time series can be quite a challenge. They take many different forms and they can be analysed by various approaches and theories. In various fields, time series can be detected. Generally the analysis of these series can be divided into 2 different approaches. The first is based on the time domain. Various methods built upon the time domain of these series. The general analysis of this type will be explained in the next paragraph. The second approach is based on the frequency domain. The according methods are based on the Laplace or Fourier transformations of the time domain into the frequency domain. One of the well-known theories; Spectral Analysis, will be investigated in the subsequent paragraph.

3.3.1. Time series analysis on the time domain

From the late twenties of the previous century mathematicians have been looking into time series analysis [21]. Forecasting had been done before, but since time series are essential in forecasting, this analysis became a priority. In the process of building these theories the book by George Box and Gwilym Jenkins (1970) was essential. They can both be considered pioneers in this field of mathematics.
Multiple tools are of the essence for analysing the precipitation and CTT time series of Germany.

For this composition we will look into the theories and tools that might be of use for weather time series based on the theory described in 'Time Series Analysis: Forecasting and Control (Fith Edition)' by George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel and Greta M. Ljung (2015) [2] and the online course 'STAT510' from Penn State University Department of Statistics [15]. Mostly the theory has been developed for eventual forecasting. However the base for this is finding different forms of relations within and between the series on which the forecast can be based. Basically the theory is based on comparing the value at time t with a known value after a lag of time $t+l$ so that the unknown values after a longer time can be predicted. In our time series the values are discretely distributed at an equispaced distance, that is, one time step corresponds to one day everywhere.

A *time series* x_t is a set of observations generated sequentially over time, with $t = 0,1,2,..$ a predetermined step in time. The time series is *discrete* since the set is discrete. Although the process observed by the time series is *continuous*, the measuring at different moments in time makes it discrete.
For a time series to be (weakly) stationary, the following characteristics need to be independent of time:

- Mean: $E(x_t)$
- Variance: $\sigma(x_t)$
- Covariance $Cov(x_t, x_{t-1})$
- Correlation $Corr(x_t, x_{t-1})$.

Many of the functions and tools within time series analysis require the series to be stationary. When a time series is not stationary,the n-th difference of the time series can be used instead. The *first difference* of a series is defined as $y_t = x_t - x_{t-1}$. Taking the difference might create a stationary series. Based on these stationary series several time domain models can be applied.
The *autocorrelation function* (ACF) is the correlation of two values x_t and x_{t-h} from one time series with a time difference, or lag, of $h = 1,2,3,..$. The ACF is defined as follows:

$$\frac{Cov(x_t, x_{t-h})}{\sigma(x_t)} \tag{3.26}$$

since the time series is stationary. The ACF can identify possible structure within the time series. The *partial autocorrelation function* (PACF) gives the partial correlation which is a conditional correlation of an autocorrelation. It is the autocorrelation of x_t and x_{t-h}, while taking into account the dependence of x_t on x_{t-1} up to x_{t-h+1}. It is defined:

$$\frac{Cov(x_t, x_{t-h}|x_{t-1}, x_{t-h+1})}{\sqrt{Var(x_t|x_{t-1}, x_{t-h+1})Var(x_{t-h}|x_{t-1}, x_{t-h+1})}}. \tag{3.27}$$

ARIMA

An *autoregressive integrated moving average* (ARIMA) is a statistical model that aims to predict future values and understand present data. It consists of different features.
An *autoregressive* (AR) model is a linear model for predicting a value at time t by using the value at an earlier time $t-h$. The order of an AR model depends on the amount of previous values used for the prediction. The AR(p) model can be defined as:

$$x_t = \phi_0 + \phi_1 x_{t-1} + \phi_2 x_{t-2} + ... + \phi_p x_{t-p} + \omega_t \tag{3.28}$$

where ω_t represents the errors and $\phi_i \in \mathbb{R}$. The AR(1) model is similar to the earlier described (ordinary) least squares regression. However, the main difference is that the x values are random here. For positive values of

ϕ_1 the ACF is often reducing gradually for increasing time lags; either only positive or alternating the x-axis.
A *moving average* (MA) model of order q is based on the (past) error values instead of the values themselves, and can be defined as follows:

$$x_t = \mu + \omega_t + \theta_1\omega_{t-1} + \theta_2\omega_{t-2} + ... + \theta_q\omega_{t-q}. \tag{3.29}$$

For all lags greater than q in the ACF of MA(q), the auto correlation is 0.
An AR and MA model combined is defined as an ARMA model. If differencing is also included, we talk about an ARIMA model. The orders of the models are given as follows: ARIMA(p,d,q), where p is the order of AR, q the order of MA and d the order of the difference function.

Approach

For applying these models to time series there are different steps to take. First of all, the time series should be stationary. Secondly, differencing might be necessary when seasonality appears. It can also be used in case of an obvious upward or downward trend in the raw data. Time series that show an upward curve could be transformed by a logarithm or square root before further analysis.
Once this has been settled, the ACF and PACF should be retrieved to recognise patterns or recurrences. Based on different characteristics of both the ACF and PACF, the orders of the model can be guessed and estimated by, amongst other tests, maximum likelihood estimation.

3.3.2. Spectral analysis [9]

Time series are, obviously, based on an order in time. However, it might be useful in some situations to observe the series not in the time, but in the frequency domain. That is, referring to the values in the series, with respect to its frequency. The time series can be converted to the frequency domain by various transforms. Two well-known are the *Laplace* and *Fourier* transforms. These can be defined as follows for a function $f(t)$:

$$\text{[Laplace]} \qquad F(s) = \int_0^\infty f(t)\exp\{-st\}dt \qquad \forall\, t \geq 0 \text{ and } s \in \mathbb{C} \tag{3.30}$$

$$\text{[Fourier]} \qquad F(\omega) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^\infty f(t)\exp\{-i\omega t\}dt \qquad \text{with } \omega \in \mathbb{R} \tag{3.31}$$

The series can also be converted back by the inverse of the transforms.
Where earlier, the auto correlation function was described, in spectral analysis there is the spectral density function. The auto covariance $\gamma(h) = Cov(x_t, x_{t-h})$, from the auto correlation function, and the spectral density function ($f(\omega)$) are Fourier transform pairs:

$$f(\omega) = \int_{-\infty}^\infty \gamma(h)\exp\{-2\pi i\omega h\}dh \tag{3.32}$$

$$\gamma(h) = \int_{-\infty}^\infty \exp\{2\pi i\omega h\}f(\omega)d\omega \tag{3.33}$$

A plot of the spectral density function against the frequencies is called the *frequency spectrum*.

3.3.3. Comparison

On mathematical level, the auto correlation and spectral density function are equivalent. Although both the approaches in the time and frequency domain contain very similar actions and calculations, there are reasons to prefer one or the other, based on the context of the data. For example, spectral analysis might be more fit for larger data sets, but requires more regular data.

4

Analysing the time series

Before applying mathematical theory and tools, it is essential to observe the data. The scatter plot in figure 4.1 shows how the data points (CTT, precipitation) are distributed. Precipitation is bell-shaped distributed and generally skewed to the right. Days with no precipitation measured have similar temperatures and reach even further in highs and lows.
The ragne of CTT in this data is $[188.6, 307.7]$ Kelvin. In general the precipitation values lie within the domain $[0, 80]$ mm/day. However, by taking into account the outliers, the range is $[0, 312]$ mm/day. From this graph it can easily be concluded that there is no obvious linear regression to be found.

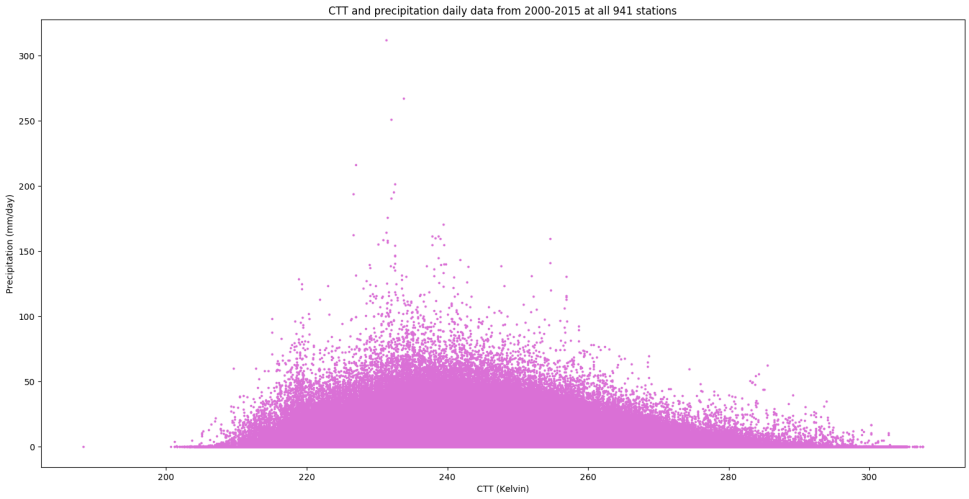
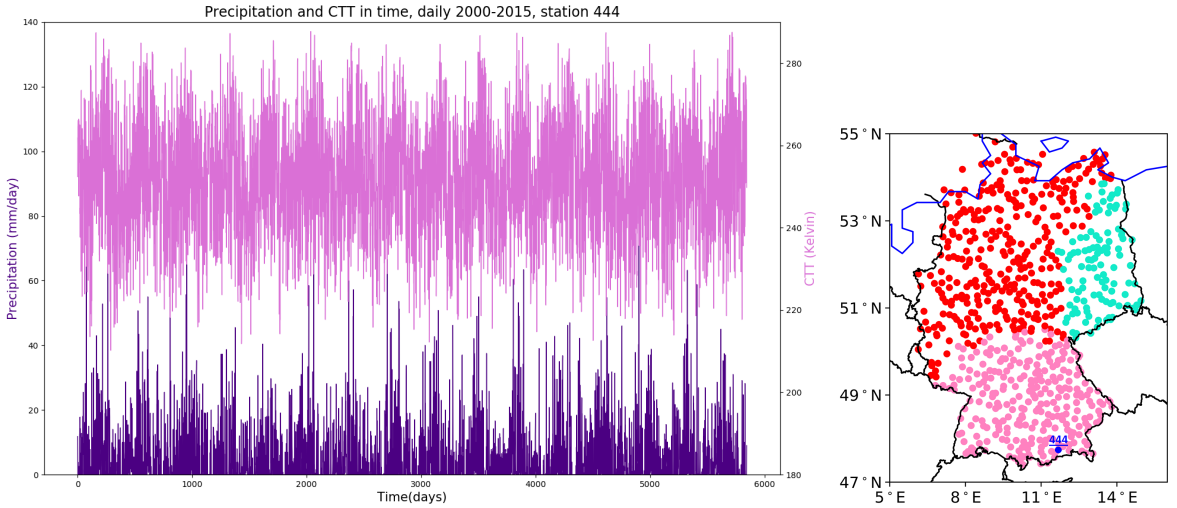


Figure 4.1: Scatter plot of the CTT and precipitation of all 941 stations. Every data point corresponds to a day in the years 2000-2015.

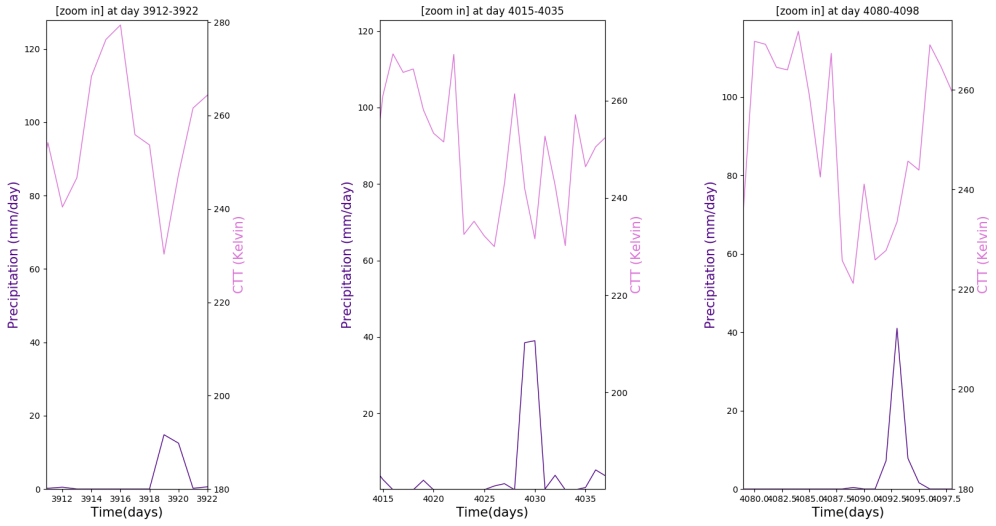


(a) Plot (b) Station 444

Figure 4.2: Precipitation and CTT plotted against time for station 444 in the years 2000-2015 (5844 days).

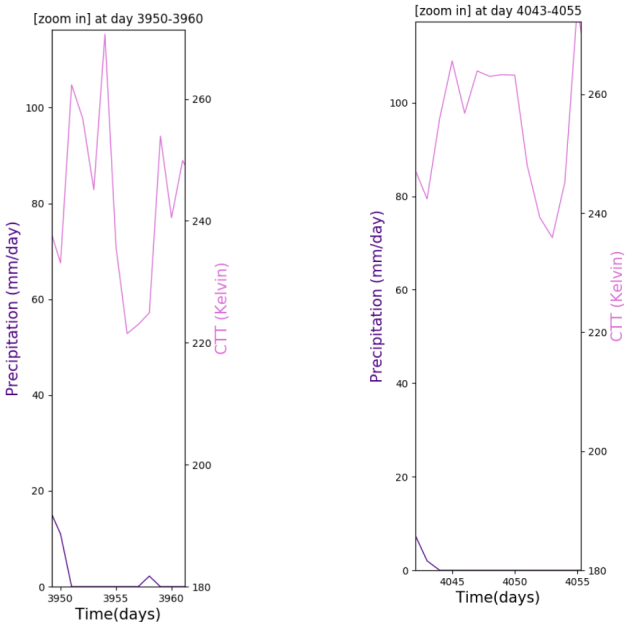
4.1. CTT and precipitation time series

In figure 4.3a the two time series are visualised separately, for an arbitrarily chosen station (444). By inspecting the plots by hand some reoccurring patterns could come up. A change, usually a drop, in temperature right before heavy or long-term rainfall would be expected.
Small parts of the figure show some reoccurring events. In figure 4.3a, 4.3b and 4.3c there are peaks of rain, $18-40$ mm/day, after a significant drop in temperature: a drop of about 40 to 50 degrees Kelvin over a period of only a few days. Although this phenomenon takes place multiple times throughout these 16 years, there are many situations that show opposite behaviour as well. Figure 4.4a and 4.4b show a similar drop in temperature but almost no rain taking place in the following days. As a consequence there is nothing to be recognised as a clear pattern or behaviour.



(a) Zoomed in at day 3912-3922 (b) Zoomed in at day 4015-4035 (c) Zoomed in at day 4080-4098

Figure 4.3: Small subsets of the line plot in figure 4.2a



(a) Zoomed in at day 3950-3960 (b) Zoomed in at day 4043-4055

Figure 4.4: Small subsets of the line plot in figure 4.2a

4.2. Problems

After the rough inspection, the data will be observed more accurately. One of the first requirements mentioned in chapter 3 is that the time series observed should be stationary, as a start for further investigation. This has to be examined for the data present. Histograms show the probability distributions of the data and the rolling mean and variances show the behaviour of the series over time.
Figure 4.5a shows the distribution of the CTT data, which can be assumed to be Gaussian. Figure 4.5b and 4.5c show the rolling mean and variance for the CTT time series, respectively. Both the figures show a big variation. The mean and variance are thus not (nearly) constant over time: the data seems to be non-stationary.

(a) Histogram

(b) Rolling mean: the mean from different subsets of the TS

(c) Rolling variance: the variance from different subsets of the TS

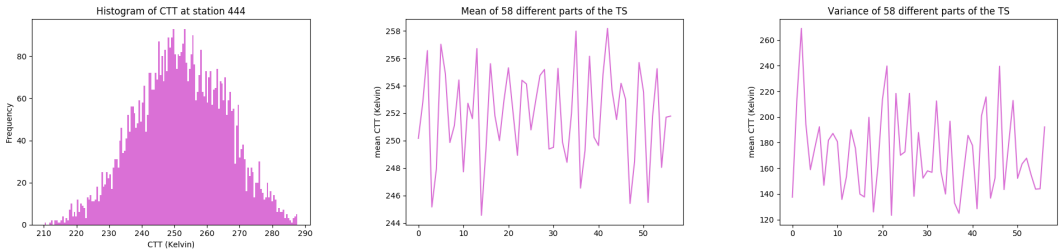
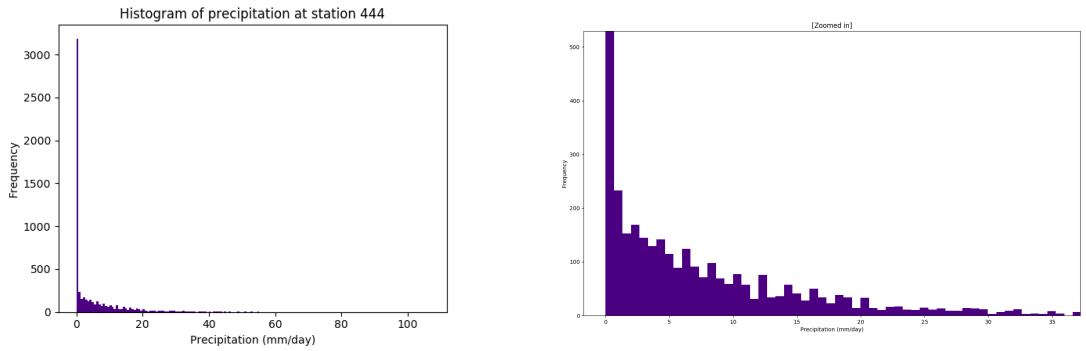


Figure 4.5: Statistical visualisations of the CTT times series at station 444

The same graphs are generated for the precipitation time series. The distribution of the precipitation time series (figure 4.6) is not normal (Gaussian). But figures 4.7a and 4.7b show similar behaviour as the rolling mean and variance for CTT. Transforming the series by logarithm or square root is of no help either. From these graphs we can now conclude that the time series are not stationary. This obstructs further use of the advanced time series analysis as described in the previous chapter. Determining the ACF and PACF does not make sense for time series that are not (weakly) stationary [16].



(a) Histogram

(b) Histogram, zoomed in

Figure 4.6: Histogram of the precipitation times series at station 444

(a) Rolling mean: the mean from different subsets of the TS

(b) Rolling variance: the variance from different subsets of the TS

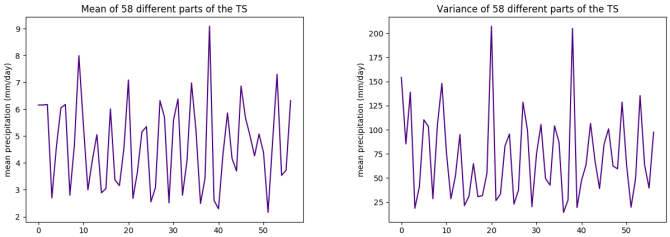


Figure 4.7: Statistical visualisations of the precipitation times series at station 444

4.3. Research approach

From observing this scattered data we can conclude that direct time series analysis is not suitable and not relevant at this time. For applying the given tools the data must satisfy some of the requirements. Once appropriate data becomes available we might be able to investigate further.

The various characteristics within the data, that are now invisible, might be causing the scattering. The next step is dividing this data in a convenient way. After dividing or gathering pieces of the time series according to their characteristics we hope to create stationary time series or find regression to describe the possible relationship among CTT and precipitation. Also, more basic statistic tools can be considered now.

If this approach indeed results in stationary data, some of the tools described in the time series analysis paragraph may be applied in the future.

5

Adjusting the time series

The data that is available and required for this research consists of multiple characteristics. The series are made up of daily CTT and precipitation measurements in 941 stations scattered over Germany. Previous chapters showed that the advanced mathematical time series analysis can not be applied on the available data. On that account, more basic well-known statistical regression tools can be explored. On top of that, adjustments to the time series can possibly create data that is more manageable.

5.1. Correlation coefficients

For observing the data and generating the analysis in this research, Python was the main provider. Furthermore, Python is equipped with packages for regression analysis. Numpy, Matplotlib and Scikit Learn provide all the tools required for statistical calculations in Python. For both the precipitation and the CTT time series, basic statistical values as described in chapter 3 can be measured: Pearson's correlation coefficient, Spearman's rank correlation coefficient and the characteristics of various degrees of regression.
For doing this, all of the stations will be concatenated into one long data set. The two series, one for CTT and one for precipitation, will then be used in the functions provided by the Python packages.
By using the these two time series (covering all of the stations and all of the years) as input for the correlation coefficient and the rank correlation coefficient, the values in table 5.1 were returned. From now 'CC' will be used to address the correlation coefficient and 'SPR' will be used to address Spearman's rank correlation coefficient.

CC	SPR
-0.31	-0.43

Table 5.1: The calculated values (CC, SPR) for the data of all days in 2000-2015 in all stations.

Before conducting more research on these values and their significance, similar values for various situations will be investigated in the next section.

5.2. Adjusting the series based on its characteristics

Apart from the CTT and precipitation values, the time series provide us with geographical quantities. Sorting the time series according to their geographical characteristics might reduce the scattering of the data.

First of all, Germany can be divided into regions. In countries that stretch far like Germany and Tanzania, big differences in the weather system appear due to geographical characteristics. Through these differences in the weather system, various forms, sorts, frequencies and magnitudes of rain are formed. Although it is impossible to take into account all these characteristics right now, 3 regions were extracted from the time series by CITG based on dissimilarities in the precipitation time series, computed by the use of the dynamic time warping method. We will not look into this specific method here.

Secondly, the elevation of every station is given. Although a separation based on height might overlap with

that of the regions, it is worth looking into a division between high and low stations since the type of rain is directly influenced by the elevation of land.
Over the centuries there have been various definitions for mountains. According to the 'Mountain Watch Report, 2002' by UNEP [1] we consider an area mountainous whenever:

- there is an elevation of 2500 metres
- there is an elevation of 1500 metres with a slope of at least 2 degrees
- there is an elevation of 1000 metres with a slope of at least 5 degrees
- or there is an elevation of 300 metres with a range of at least 7 kilometres.

Another (obvious) rule for dividing the data is time. A higher chance for regression or stationary features when comparing similar months or seasons is to be expected. However, we should not forget that the climate, and thus the weather, has been changing over the years.

Lastly, as explained in chapter 2, there are three types of rain. In Germany we mostly observe frontal rain and sometimes convective rain. Meanwhile in Tanzania there is mostly convective rain. Orographic rain might appear near the Tanzanian mountains and the sea. Eventually we want to relate the findings over Germany to Tanzania. Since the frontal and convective rain differ in their processes, it would be of value to be able to separate the time series based on these types of rain.
This might be possible if we are able to distinguish them based on specific fluctuations in temperature or specific duration of rain periods. For these characteristics, more detailed information on the physical process of these types of rain is essential and reaches further than the possibilities within this research.

5.3. Approach

Subsets of the original time series were made by concatenating separate, specific pieces out of the big data set, for each case differently. The way these separations are carried out, facilitates the possibility of calculating the coefficients for these cases in a similar way as has been done before.
On top of the correlation coefficients that were mentioned, three degrees for regression are observed for every case as well: $d = 1, 2, 3$. For these regression models, the R-squared values are calculated. The next paragraphs elaborate on the different cases and show the results that were found.

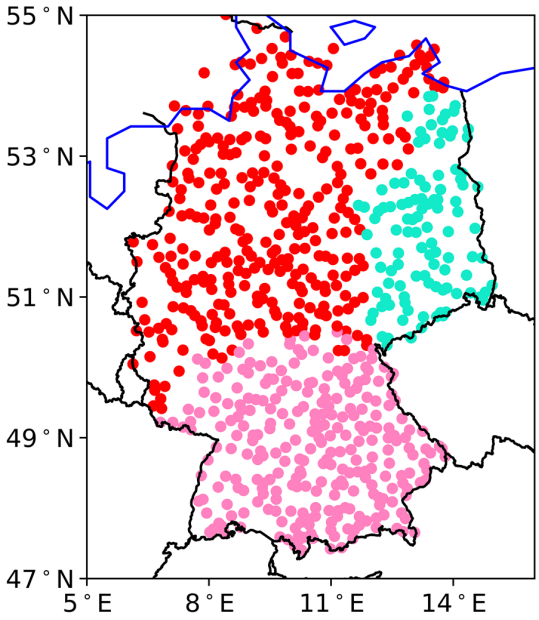
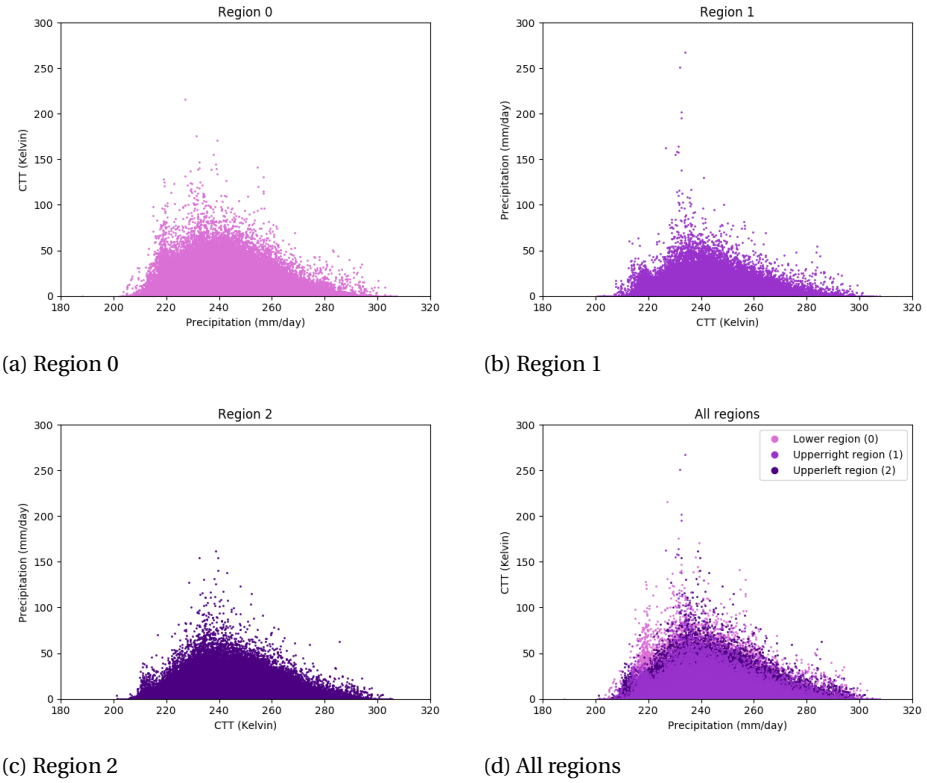


Figure 5.1: Map of the 3 regions of CDC weather stations. The 277 pink dots represent region 0, the 113 cyan dots represent region 1 and the 345 red dots together form region 2.

5.3.1. Regions

Figure 5.1 shows how the regions have been arranged by CITG. Region 0 is the lower region with 277 stations, region 1 is the upper right region with 133 stations and the upper left ensemble is region 2 with 345 stations. Not all of the 914 stations were labelled for a region. In total, 735 stations are covered here.
Dividing the time series into these three regions shows minimal changes in the scatter plots (figure 5.2) as compared to what we have seen in the previous chapter (figure 4.1). The similarities imply that the mean and



(a) Region 0

(b) Region 1

(c) Region 2

(d) All regions

Figure 5.2: Scatter plots of the CTT and precipitation of all the stations per region. Every data point corresponds to a day in the years 2000-2015.

variance will not become stationary by this specific division. The correlation coefficients among CTT and precipitation in these regions and the R-squared values for three degrees of regression are listed in table 5.2. The table emphasises the minimal differences that were mentioned before.

Region	CC	SPR	$R^2, d = 1$	$d = 2$	$d = 3$
0	-0.31	-0.45	0.10	0.10	0.10
1	-0.28	-0.41	0.08	0.08	0.08
2	-0.31	-0.43	0.10	0.10	0.10

Table 5.2: The calculated values (CC, SPR, R^2) for all days in 2000-2015, separated per region.

Compared to the values in table 5.1 not a lot has changed. Only region 2 depicts even less of a correlation and the Spearman's rank correlation coefficient of region 0 might be interesting for the significance research in the next section since it stands out from the others.
The R-squared values are very low.

5.3.2. Elevation

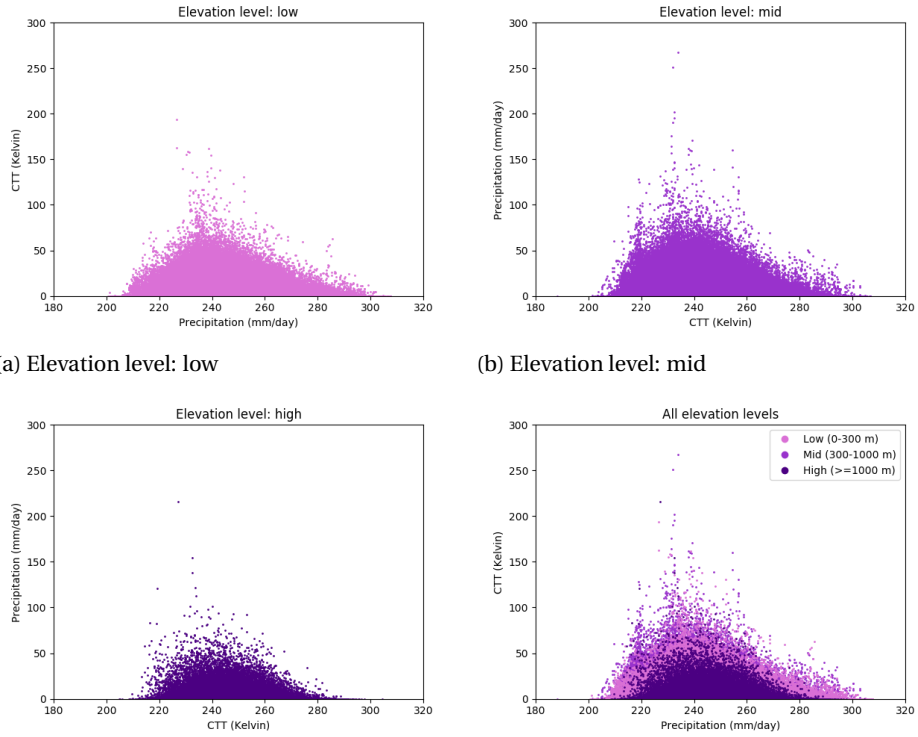
Based on the elevation levels described in section 5.2 and the heights of the different stations, the data was divided into three groups of altitude:

Low The lowest elevation level represents all stations which measure from 0 up to 300 metres. This group consists of 517 stations.

Mid The middle elevation level represents all stations which measure 300 up to a thousand metres high. This group contains 416 stations.

High The highest elevation level represents all stations which measure a thousand metres or higher. Only 8 stations are found in this category.

There is no data available on the slope of the area near the stations. Because this the slopes mentioned in section 5.2 are neglected.
The scatter plots of the low and middle elevation levels show minimal changes. Only the scatter plot of the higher level of stations reveals some different behaviour. However, this category consists of only 8 stations.



(a) Elevation level: low

(b) Elevation level: mid

(c) Elevation level: high

(d) All elevation levels

Figure 5.3: Scatter plots of the CTT and precipitation of all the stations per elevation level. Every data point corresponds to a day in the years 2000-2015.

The correlation coefficients among CTT and precipitation, and the R-squared values for three degrees of regression are listed in table 5.3. Again we can easily derive that there are no big differences from the original values for the same coefficients and the R-squared values are low.

Elevation	CC	SPR	$R^2, d = 1$	$d = 2$	$d = 3$
Low	-0.30	-0.42	0.09	0.09	0.10
Mid	-0.31	-0.44	0.10	0.10	0.10
High	-0.31	-0.43	0.10	0.10	0.11

Table 5.3: The calculated values (CC, SPR, R^2) for the data of all days in 2000-2015, separated per elevation level.

5.3.3. Time

The time series present can be separated based on time in various ways. We would expect that months and seasons act similarly in different years and stations. For every month the data of 941 stations during all the 16 years was gathered into one time series, both for CTT and precipitation. The correlation coefficients among CTT and precipitation, and the R-squared values for three degrees of regression are listed in table 5.4.

Month	CC	SPR	$R^2, d = 1$	$d = 2$	$d = 3$
Jan	-0.39	-0.46	0.15	0.15	0.15
Feb	-0.34	-0.45	0.12	0.12	0.13
Mar	-0.32	-0.43	0.10	0.10	0.11
Apr	-0.28	-0.40	0.08	0.08	0.09
May	-0.32	-0.43	0.10	0.10	0.11
Jun	-0.33	-0.45	0.11	0.11	0.12
Jul	-0.34	-0.50	0.12	0.12	0.13
Aug	-0.34	-0.50	0.12	0.13	0.13
Sep	-0.34	-0.46	0.12	0.12	0.12
Oct	-0.28	-0.39	0.08	0.08	0.08
Nov	-0.29	-0.35	0.08	0.09	0.09
Dec	-0.34	-0.44	0.12	0.12	0.13

Table 5.4: The calculated values (CC, SPR, R^2) for the data of all days in 2000-2015, separated per region.

Some months display more interesting results: January is particularly prominent in its correlation coefficient and slightly increased R-squared values. Both July and August stand out from the other months in the rank correlation coefficient.

The seasons have been categorised as follows:

Winter December, January, February

Spring March, April, May

Summer June, July, August

Autumn September, October, November

For every season the data of the according 3 months of all 941 stations during all the 16 years was gathered into one time series. The correlation coefficients among CTT and precipitation, and the R-squared values for three degrees of regression are listed in table 5.5.

Season	CC	SPR	$R^2, d = 1$	$d = 2$	$d = 3$
Winter	-0.36	-0.45	0.13	0.13	0.13
Spring	-0.30	-0.42	0.09	0.09	0.10
Summer	-0.34	-0.48	0.11	0.12	0.12
Autumn	-0.30	-0.40	0.09	0.09	0.09

Table 5.5: The calculated values (CC, SPR, R^2) for the data of all days in 2000-2015, separated per region.

Among the seasons, the correlation coefficient of winter is slightly higher than others, and summer shows a higher Spearman's rank correlation coefficient.

5.3.4. Combining characteristics

The greatest distinction would be expected from a combination of multiple of the explored characteristics, since these create the smallest, but still significantly big, data sets.
The first distinction has been for the combination of single months and single regions. This case has been worked out for all the 12 months and three regions. The results are shown in tables 5.6 and 5.7 (with region 0, 1 and 2 written as R0, R1, R2). As seen before, January ranks high when it comes to the correlation coefficient, especially in region 0 and 2 (table 5.6). The rank correlation coefficient is increased for multiple months this time: January, February, July, August and September depict relatively high values (table 5.6).

The second distinction that has been made is for a single season in one region. The values for this case have been calculated over all seasons and regions. Its results are shown in tables 5.8 and 5.9. Apart from an outstanding -0.5 for the correlation coefficient, there is no significant difference to be noted compared to

table 5.5.

Month	CC			SPR		
	R0	R1	R2	R0	R1	R2
Jan	-0.39	-0.34	-0.40	-0.49	-0.39	-0.46
Feb	-0.35	-0.31	-0.36	-0.48	-0.42	-0.45
Mar	-0.33	-0.29	-0.33	-0.45	-0.39	-0.42
Apr	-0.29	-0.25	-0.29	-0.42	-0.37	-0.39
May	-0.30	-0.33	-0.33	-0.41	-0.46	-0.43
Jun	-0.32	-0.34	-0.35	-0.46	-0.44	-0.44
Jul	-0.36	-0.35	-0.33	-0.51	-0.49	-0.48
Aug	-0.36	-0.30	-0.34	-0.53	-0.48	-0.46
Sep	-0.35	-0.33	-0.34	-0.49	-0.44	-0.44
Oct	-0.30	-0.28	-0.28	-0.41	-0.37	-0.37
Nov	-0.31	-0.27	-0.28	-0.37	-0.32	-0.34
Dec	-0.35	-0.32	-0.36	-0.47	-0.36	-0.43

Table 5.6: The calculated values (CC and SPR) for 2000-2015, separated per month and region.

Month	R^2: $d=1$			R^2: $d=2$			R^2: $d=3$		
	R0	R1	R2	R0	R1	R2	R0	R1	R2
Jan	0.15	0.11	0.16	0.16	0.11	0.16	0.16	0.12	0.16
Feb	0.12	0.10	0.13	0.12	0.11	0.13	0.13	0.12	0.14
Mar	0.11	0.09	0.11	0.11	0.09	0.11	0.11	0.11	0.12
Apr	0.08	0.06	0.08	0.08	0.07	0.08	0.10	0.08	0.09
May	0.09	0.11	0.11	0.09	0.11	0.11	0.10	0.12	0.12
Jun	0.10	0.11	0.12	0.10	0.13	0.13	0.11	0.13	0.13
Jul	0.13	0.12	0.11	0.13	0.13	0.11	0.14	0.14	0.12
Aug	0.13	0.09	0.11	0.14	0.11	0.12	0.14	0.11	0.12
Sep	0.13	0.11	0.11	0.13	0.11	0.11	0.13	0.12	0.12
Oct	0.09	0.08	0.08	0.09	0.08	0.08	0.09	0.08	0.08
Nov	0.09	0.07	0.08	0.10	0.08	0.08	0.10	0.08	0.09
Dec	0.12	0.10	0.13	0.12	0.10	0.13	0.14	0.10	0.13

Table 5.7: The calculated values (R^2) for 2000-2015, separated per month and region.

Season	CC			SPR		
	R0	R1	R2	R0	R1	R2
Winter	-0.36	-0.32	-0.37	-0.48	-0.39	-0.45
Spring	-0.29	-0.28	-0.31	-0.43	-0.41	-0.41
Summer	-0.34	-0.33	-0.34	-0.50	-0.47	-0.46
Autumn	-0.31	-0.28	-0.29	-0.43	-0.38	-0.39

Table 5.8: The calculated values (CC and SPR) for 2000-2015, separated per season and region.

Season	R^2: $d=1$			R^2: $d=2$			R^2: $d=3$		
	R0	R1	R2	R0	R1	R2	R0	R1	R2
Winter	0.13	0.10	0.14	0.13	0.11	0.14	0.14	0.11	0.14
Spring	0.09	0.08	0.09	0.09	0.08	0.10	0.10	0.10	0.11
Summer	0.12	0.11	0.11	0.12	0.12	0.12	0.13	0.12	0.12
Autumn	0.10	0.08	0.08	0.10	0.08	0.08	0.10	0.08	0.09

Table 5.9: The calculated values (R^2) for 2000-2015, separated per season and region.

Lastly, the values have been calculated for all months in the elevation levels. The results for this case are shown in tables 5.10 and 5.11. In this specific case, the highest values to account for are again January's correlation coefficient in the lower and middle height level together with the rank correlation coefficient corresponding to June, July, August and September in the highest stations.

Month	CC			SPR		
	low	mid	high	low	mid	high
Jan	-0.39	-0.39	-0.34	-0.45	-0.47	-0.42
Feb	-0.36	-0.34	-0.35	-0.46	-0.45	-0.45
Mar	-0.32	-0.32	-0.32	-0.42	-0.44	-0.44
Apr	-0.28	-0.28	-0.28	-0.39	-0.41	-0.40
May	-0.33	-0.30	-0.29	-0.43	-0.42	-0.42
Jun	-0.34	-0.33	-0.37	-0.44	-0.45	-0.49
Jul	-0.33	-0.35	-0.39	-0.48	-0.51	-0.54
Aug	-0.33	-0.36	-0.38	-0.46	-0.53	-0.54
Sep	-0.34	-0.35	-0.37	-0.44	-0.48	-0.50
Oct	-0.28	-0.29	-0.26	-0.38	-0.39	-0.38
Nov	-0.29	-0.30	-0.25	-0.35	-0.35	-0.30
Dec	-0.35	-0.34	-0.31	-0.42	-0.45	-0.43

Table 5.10: The calculated values (CC and SPR) for 2000-2015, separated per month and per elevation level.

Month	R^2: $d=1$			R^2: $d=2$			R^2: $d=3$		
	low	mid	high	low	mid	high	low	mid	high
Jan	0.15	0.15	0.12	0.15	0.16	0.12	0.16	0.16	0.12
Feb	0.13	0.12	0.12	0.13	0.12	0.13	0.14	0.13	0.14
Mar	0.11	0.10	0.10	0.11	0.10	0.11	0.12	0.11	0.11
Apr	0.08	0.08	0.08	0.08	0.08	0.09	0.09	0.10	0.11
May	0.11	0.09	0.09	0.11	0.09	0.09	0.12	0.10	0.11
Jun	0.11	0.11	0.13	0.12	0.11	0.13	0.12	0.12	0.14
Jul	0.11	0.12	0.16	0.12	0.13	0.16	0.12	0.13	0.16
Aug	0.11	0.13	0.15	0.12	0.14	0.16	0.12	0.14	0.16
Sep	0.11	0.12	0.14	0.11	0.13	0.14	0.12	0.13	0.14
Oct	0.08	0.08	0.07	0.08	0.09	0.07	0.08	0.09	0.08
Nov	0.08	0.09	0.06	0.09	0.09	0.07	0.09	0.09	0.07
Dec	0.13	0.12	0.09	0.13	0.12	0.11	0.13	0.13	0.11

Table 5.11: The calculated values (R^2) for 2000-2015, separated per month and per elevation level.

5.4. Regression analysis

Several of the previous tables contain values that stick out from the others. For example, the correlation coefficient for January in all of the previous tables is worth looking into. One might think that the calculated values are too low for any significant relation since they appear closer to 0 than to either 1 or -1. However, the sample size can be of great influence as well.

Since the values are mostly based on data sets that contain between 100,000 and 400,000 elements, their significance has to be investigated.

As illustrated in chapter 3, a correlation coefficient r can take values in the interval $[-1,1]$. Whenever r reaches 0, there is no correlation at all. Values close to 1 or -1 imply a relation (correlation) between the two examined variables. In this case we are looking at values in the interval $[-0.54,-0.25]$.

Another basic Python tool provides the p-value for the correlation coefficient and the Spearman's rank correlation coefficient. The following hypotheses are examined:

H_0: r is not significantly different from 0

H_a: r is significantly different from 0.

If the null hypothesis can be rejected based on the p-value, r is significantly different from 0. In that case we can say that there is some relation to be found among the CTT and precipitation time series, within one of the explored situations. By going through all the remarkable coefficients and checking the less interesting values as well, all p-values were found to be extremely small. Python returns a value of the type 'float64' for the p-value. In Python the smallest value a 'float64' can take on is 2.2250738585072014e-308.
For July and January in elevation level 'high', which correspond to remarkably high Spearman's rank correlations coefficients, the p-values are 3.20e-300 and 3.79e-304, respectively. All other cases take on even smaller p-values, in which case Python returns 0.0.
Neither of the significance levels 0.05, 0.01 or 0.001, that could have been set prior to the testing, would have rejected the null-hypothesis because of the low p-values. It can therefore be concluded that both the correlation coefficient and the Spearman's rank coefficient are statistically significant. That is, there is a negative correlation between the CTT and precipitation in all of the cases that have been discussed, some are higher and thus more interesting. Moreover, the negative rank correlation implies that there exists a decreasing monotonic relation between precipitation and CTT.
Nevertheless, correlation solely shows that there is a relation between the two variables. It is not shown whether the cloud top temperature is caused by the amount of precipitation, whether the precipitation is caused by the cloud top temperature or if both are in some way influenced by a third, fourth or even more other characteristics. In other words, correlation does not imply causation.

5.5. Future steps

As a result of these findings there are multiple other steps to take.

Combinations

Based on the mentioned characteristic and potential other characteristics it might be interesting to explore some other combinations. One might think of observing the seasons in the elevation levels, or one might change the way both the regions and the elevation levels have been separated.

(Adjusted) R-squared

In the previous paragraphs, the R-squared values have been calculated but not extensively investigated. The values in all these cases are remarkably low. Since we are working with an enormous data set, this does not instantly imply that the variance in the precipitation can not be explained by the CTT at all, or that the variance in the CTT can not be explained by the precipitation at all. It is interesting to look into this by testing the values to a significance test.
Another interesting value to explore is the adjusted R-squared.

Rain classification

It would be of great value if the types of rain can be distinguished from the variables in the time series. Ideally one type of rain would show bigger drops in temperature or significantly longer periods of precipitation than another type. If the type of rain can be determined from this kind of behaviour, it would improve the final comparison of Germany and Tanzania. We would also expect that there are less outliers within one type of rain which might improve the calculated values. For making an accurate distinction between the types of rain, more physical knowledge and data is necessary.

Logistic regression

Another statistical tool to explore is logistic regression. Based on a preset threshold for a 'rainy' or 'wet' day, each day in the 16 covered years will represent either a wet or a dry day, independent from the amount of rain that was measured. From there the distribution of the CTT can be explored.

Other cloud characteristics

As mentioned in chapter 2, there are many more vital precipitation related cloud characteristics. Investigating these characteristics in similar ways or including them to look into multivariate regression can bring new results and opportunities.

Time series analysis

Finally, one might want to look into the possibilities of applying several of the time series analysis tools that have been described in chapter 3. Based on the cases described in this chapter, a subset of the data might be found that is stationary, which would then be fit to be subjected to those tools.

6

Thresholds

Globally, different national weather institutes define thresholds for the intensity of a rainy day, based on the amount of millimetres measured in such a day. This is useful when the gauge measurements are available. Since the aim of this research is to make similar conclusions solely based on the cloud top temperature, similar thresholds need to be defined for these CTT time series.

6.1. General thresholds

The available data provides a lot of information that can be used to classify thresholds.

6.1.1. Gauge measurements

First of all, a threshold based on gauge data (the gauge threshold) has to be defined, for the amount of millimetres per day that make a day 'wet' or 'dry'. Various national weather and meteorological institutes suggest different thresholds. The Deutsche Wetterdienst (DWD) [23] and the American Meteorological Society (AMETSOC) [20] state the lowest values. According to those organisations a rainy day is any day on which, respectively, at least 0.1 and 0.2 mm is measured. UK's MET Office [13] records rainy days starting at 1.0 mm and both the Dutch KNMI [8] and Australia's Bureau of Meteorology [11] mention 10 mm as a threshold for a wet day.
Based on this information we will evaluate possible CTT thresholds for multiple gauge thresholds. The low thresholds given by DWD and AMETSOC recognise any day with a few drops of precipitation as a wet day. However, considering the potential use of the threshold information, higher thresholds could be more useful. The results in the previous chapter depict that relations among cloud top temperature and precipitation are not straightforward. It is necessary to to take this into account when deciding on CTT thresholds. African farmers would most probably appreciate a 'rainy' day of 1 mm similar to a dry day, while a 'rainy' day of 10 mm might be able to nourish their crops perfectly. Taking into account both the purpose and the indirect consequences that come with the distinction of wet or dry days, different gauge thresholds and their according CTT thresholds will be evaluated in the next paragraphs.

6.1.2. CTT thresholds

The present data provides a big sample to base simple calculations on. Assuming that 16 years of daily data is quite enough to form some general conclusions, the following thresholds were found.
While taking into account the different gauge thresholds, we can determine specific cloud top temperatures. Similar to the previous chapter, cuts have been made in the data to observe possible differences.

Obviously, within the data set there will be a cloud top temperature observed above which no rain occurs within the 16 years covered. Table 6.1 lists these upper thresholds for the different regions and elevation levels as selected in the previous chapter.
Apart from the higher placed locations, the different areas have very similar results. The higher stations measured no rain above a much lower temperature than all the other sets of stations. However, the amount of stations with available data higher then 1 km is only 8 as opposed to the 200 up to 500 stations in the other areas.

Area	0.1	0.2	1.0	2.0	5.0	10.0	25 .0 (mm/day)
All stations	304.9	304.9	303.0	302.8	302.8	302.8	294.0
Region 0	304.9	304.9	303.0	303.0	302.8	300.3	294.0
Region 1	301.2	301.2	301.2	301.2	299.3	299.3	286.6
Region 2	304.8	304.8	302.1	302.1	297.8	297.8	286.7
Low elevation	304.8	304.8	302.1	302.1	299.3	299.3	286.7
Middle elevation	304.9	304.9	303.0	302.8	302.8	302.8	294.0
High elevation	292.8	288.8	288.8	288.8	283.3	281.8	276.0

Table 6.1: The CTT temperatures in Kelvin **above** which no precipitation was observed during the 16 given years, per set gauge threshold in mm/day.

Area	0.1	0.2	1.0	2.0	5.0	10.0	25 .0 (mm/day)
All stations	0.0	0.0	0.0	0.0	205.0	205.0	209.1
Region 0	203.2	203.2	203.5	203.5	205.0	205.0	209.1
Region 1	203.0	203.0	207.0	207.0	207.6	207.6	212.3
Region 2	0.0	0.0	0.0	0.0	206.1	208.9	210.1
Low elevation	200.7	200.7	200.7	200.7	206.1	207.6	210.1
Middle elevation	188.2	188.2	203.5	203.5	205.0	205.0	209.1
High elevation	209.2	209.2	209.2	209.2	209.2	213.1	214.0

Table 6.2: The CTT temperatures in Kelvin **under** which no precipitation was observed during the 16 given years, per set gauge threshold in mm/day.

Table 6.2 shows, for different areas, the thresholds for CTT under which no precipitation was ever observed during the 16 years. Some values are 0.0, which in this case means that the lowest observed CTT was during a 'rainy' day based on the set gauge threshold. One might wonder why there is no 0.0 value in one of the areas characterised by height. This is because not all 941 stations are included in the division of the regions and height levels.
The table, again, shows no extreme differences. Only the middle and highest level stations behave slightly different from the other areas. Especially for lower precipitation thresholds, the middle level of stations has a significant lower temperature for when rain still occurred. For the high level stations again we see a difference as opposed to the other areas for all the precipitation thresholds, but again, this area consists of only 8 stations.

6.2. Estimated thresholds for different moments

Apart from the determined thresholds as described above, we want to define thresholds for CTT that come with certain chances for precipitation. Deciding whether a day is dry or wet based on the gauge measurements is very accurate. When these measurements are not available we want to be able to decide this based on the CTT. As the tables above show, there is a significant interval of temperatures for which the behaviour is still unclear. For all these temperatures we will investigate what happens to dry and wet days with a varying threshold for CTT.

6.2.1. Approach

Assume we have a fixed threshold R for precipitation, set by the official weather institutes. That is, for all gauge measurements $r_i < R$, on day $i = 1, 2, ..., 5844$ [1], we consider the day a dry day, and wet otherwise. The aim is to find a convenient thresholds T_{top} and T_{bottom}: an upper and lower threshold respectively. Ideally these thresholds tell us with great probability whether a day is wet or dry: any day with a CTT lower than T_{bottom} or higher than T_{top} will be considered a dry day, all days in between will be considered a wet day. For defining sensible thresholds T_{top} and T_{bottom}, two percentages will have to be minimised:

P_{falsedry} The percentage of days that are falsely considered to be dry. That is, based on the available data, the percentage of days that is considered to be dry according to the set thresholds T_{top} or T_{bottom}, but was actually a wet day based on the fixed precipitation threshold R.

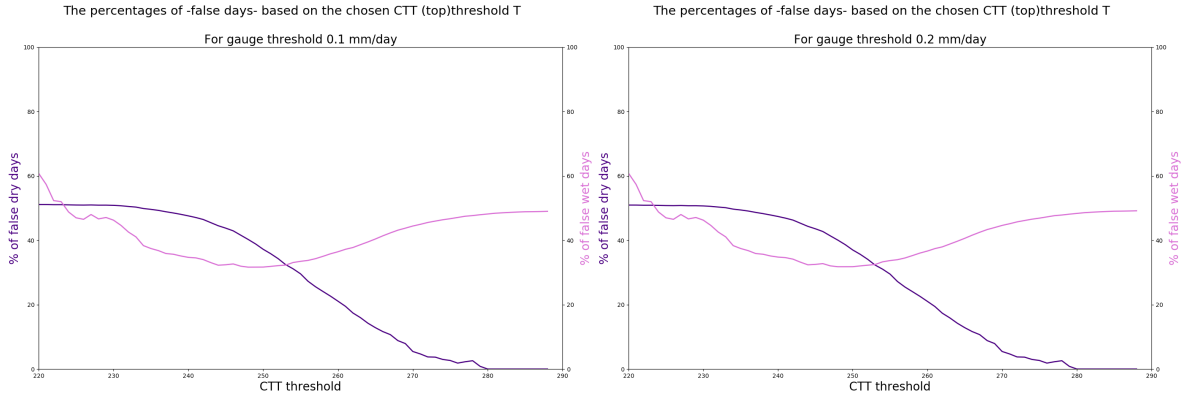
[1] The number 5844, stands for the total number of days in the years 2000-2015.

$P_{falsewet}$ The percentage of days that are falsely considered to be wet. That is, based on the available data, the percentage of days that is considered to be wet according to the set thresholds T_{top} or T_{bottom}, but was actually a dry day based on the fixed precipitation threshold R.

Both these percentages have been calculated for 4 preset precipitation thresholds R: 0.1, 0.2, 1, 5 and 10 mm/day, and various CTT thresholds for T_{top} and T_{bottom} separately.

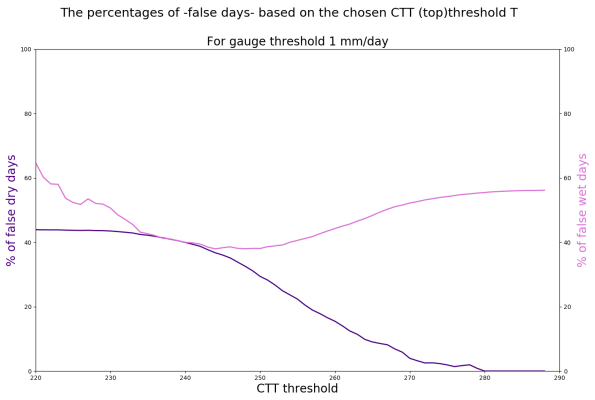
Top threshold

For determining a convenient top threshold the percentages have been set out in figures 6.1 and 6.2. The subfigures show the percentages for temperatures in the range of [220,290] Kelvin. All of the subfigures show a similar minimal percentage of false wet days ($P_{falsewet}$) around 250 degrees Kelvin and it is never lower than 30 percent. In other words, based on this data there will always be a significant percentage of days that is falsely considered a wet day for every chosen CTT threshold. On the other hand, the figures also show that $P_{falsedry}$ reaches 0 for some temperatures. Unfortunately the overall behaviour describes an increasing $P_{falsewet}$ whenever $P_{falsedry}$ decreases. As can be derived from the subfigures: a higher preset R comes with fewer false dry days simultaneously with an increased amount of false wet days.



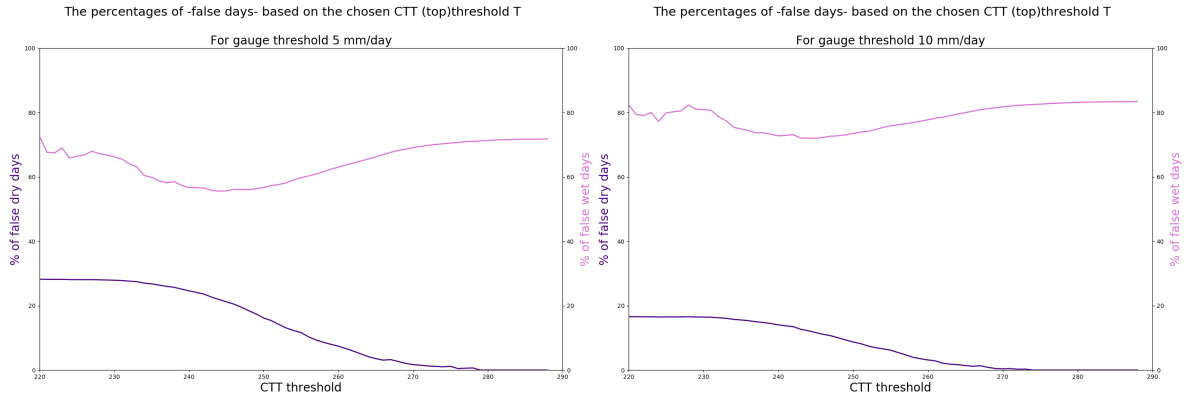
(a) $R = 0.1$ mm/day.

(b) $R = 0.2$ mm/day.



(c) $R = 1$ mm/day.

Figure 6.1: Line plot of percentages $P_{falsedry}$ and $P_{falsewet}$ for CTT top thresholds T_{top} between 220 and 290 Kelvin, with different preset precipitation thresholds R. These graphs are based on an arbitrarily chosen station 444, and show similar behaviour to the other stations.

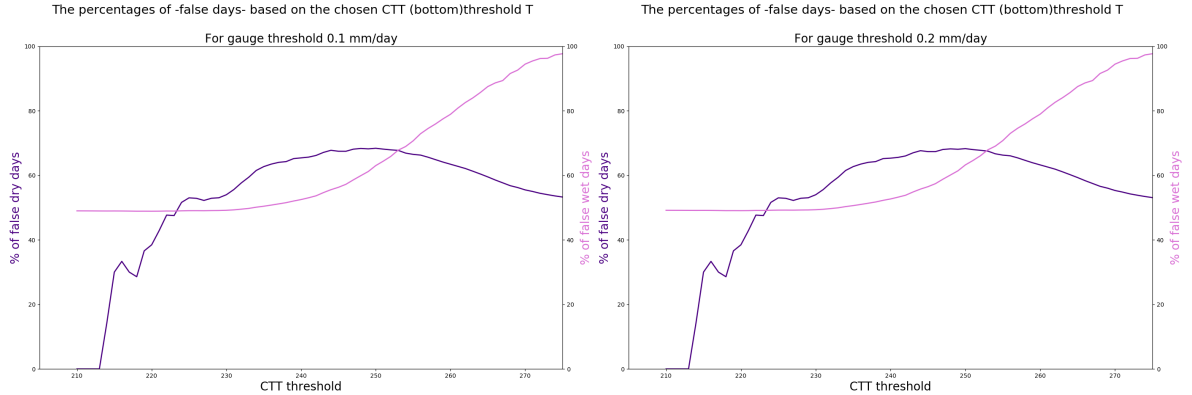


(a) $R = 5$ mm/day. (b) $R = 10$ mm/day.

Figure 6.2: Line plot of percentages P_{falsedry} and P_{falsewet} for CTT top thresholds T_{top} between 220 and 290 Kelvin, with different preset precipitation thresholds R. These graphs are based on an arbitrarily chosen station 444, and show similar behaviour to the other stations.

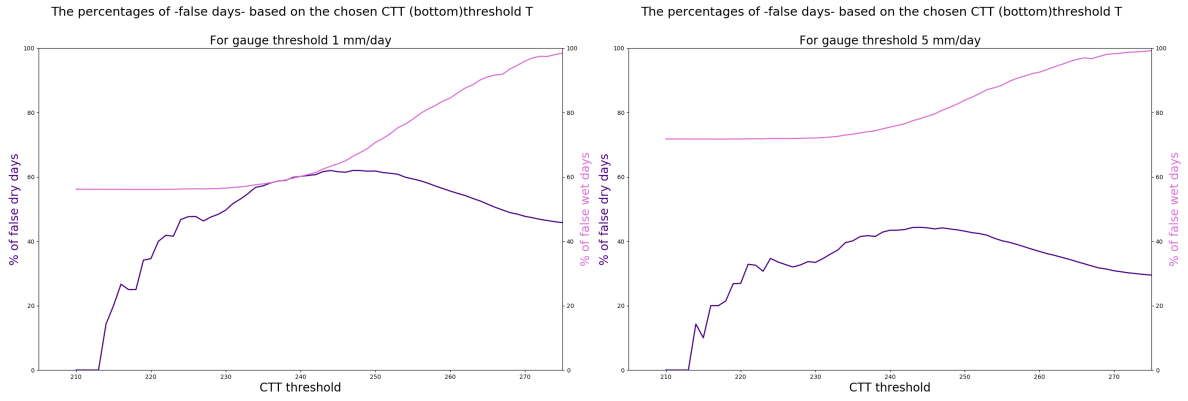
Bottom threshold

A similar process was carried out for the bottom CTT threshold T_{bottom}. Figures 6.3 and 6.4 depict the various percentages P_{falsedry} and P_{falsewet} for all possible CTT thresholds in the range of $[205, 275]$ degrees Kelvin. From the subfigures we can derive that the percentage of false wet days is at least 50 percent for all temperatures. One advantage compared to the top threshold figures is that in this case both the P_{falsedry} and P_{falsewet} are decreasing for lower temperatures. On the other hand; the increasing percentages for a higher preset R is good for nothing.



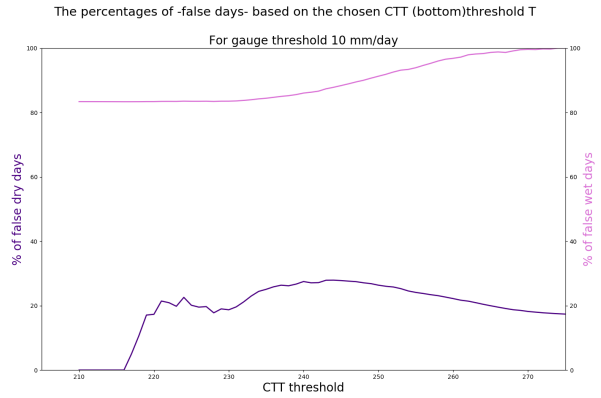
(a) $R = 0.1$ mm/day. (b) $R = 0.2$ mm/day.

Figure 6.3: Line plot of percentages P_{falsedry} and P_{falsewet} for CTT bottom thresholds T_{bottom} between 205 and 275 Kelvin, with different preset precipitation thresholds R. These graphs are based on an arbitrarily chosen station 444, and show similar behaviour to the other stations.



(a) $R = 1$ mm/day.

(b) $R = 5$ mm/day.



(c) $R = 10$ mm/day.

Figure 6.4: Line plot of percentages P_{falsedry} and P_{falsewet} for CTT bottom thresholds T_{bottom} between 205 and 275 Kelvin, with different preset precipitation thresholds R. These graphs are based on an arbitrarily chosen station 444, and show similar behaviour to the other stations.

6.3. Conclusion

As we can see from all the graphs, there is a lot open for interpretation when it comes to the context of these variables. Above all, there are a few things that need to be settled upfront.

First of all, the purpose of the investigated data should be clarified and quantified. If, for example, this data will only be used for forecasting precipitation for Sub-Saharan farmers, we need to take into account the requirements for growing crops before deciding on a model. If certain Sub-Saharan crops require a minimal amount of rain per week, it makes sense to settle the daily threshold R at a value that makes sense for reaching this amount of rain per week.

On the other hand, if for a specific crop or for a different situation, even the least amount of rain is required, we might lower the R threshold. Setting the R thus depends on the required rain intensity in this case.

Secondly, the consequences of a false wet day along with the consequences of a false dry day need to be considered. If a wet days was forecasted, the farmers might have decided to plant their crops the day before. If no wet day occurs, their whole harvest might be ruined, and thus their livelihood. This is a very big risk we want to minimise. Other decisions are made based on a forecasted dry day. Here similar consequences have to be taken into account. Estimations for these risks need to be investigated and measured against one another. Which of the consequences are better or worse?

Lastly, limits for the accuracy of these thresholds need to be settled. In some situations, we might need a very accurate relation between precipitation and CTT. Again, for planting crops one might expect a high accuracy of the amount and magnitude of the precipitation. On the other hand, for an arbitrary week throughout the growing process of these crops, it might not matter as much whether there are more days of rain one week, and less in the next.

We can thus conclude that the purpose of the data, the consequences of thresholds and the time frame that the data is used in, all influence the way that R and T_{bottom} and T_{top} should be settled and investigated. Those circumstances lead to useful conditions to base the investigation, decisions and eventual forecasts on.

For the top threshold T_{top} there are some specific things to consider. The events of a false dry or a false wet day can affect a potential user in different ways. These consequences have to be evaluated, weighed and quantified. Based on that, one of the to might be more important to minimise than the other. Moreover, a higher R influences the P_{falsewet} mostly in a negative way. If a false wet day comes with serious consequences, the rain intensity, and thus R, needs to evaluated.
For the bottom threshold, most essential will be the required intensity and accuracy of the model. Both P_{falsewet} and P_{falsedry} decrease along with the temperature. One might therefore want to consider the occurrence of these lower temperatures. That is, the most optimal bottom threshold might only rule out a small range of temperatures that occur scarcely. For such a bottom threshold to make a difference it should probably rule out more temperature.

Moreover, it will be of great value to build a model in which both the top and bottom threshold are evaluated simultaneously. For various situations and purposes only one of the two might be required for useful results. But both should at least be evaluated before ruling it out.
Lastly, in the process of creating a convenient model, we have to take into account the values from the first section in this chapter. The results suggest certain limits or boundary conditions that improve a potential model. Based on the values found in section 6.1, upper and lower limits for the CTT can be set and some temperatures can be ruled out already.

7

Conclusion and Recommendations

In this research the main goal was to investigate two vital elements within the physical precipitation system: cloud top temperature and precipitation. The main objective was to research the relation among the two variables. Besides, in this research the possibilities for CTT thresholds would be studied

7.1. Conclusion

Throughout the course of this research, many mathematical tools, methods and approaches came along. The first assumption in the process was that with advanced time series analysis and basic knowledge of the physical elements, a proper research on the available data could be done. After reviewing the essential requirements for the appliance of these mathematical tools, the assumption had to be reconsidered. More trial and error made way for more basic mathematical tools. Eventually these basic tools were the basis for the results in this research. These basic mathematical tools presented us a detailed image of the data set and its behaviour.

First of all, the time series at the basis of this research are not stationary which is why the advanced time series analysis could not be applied. This inconsistency of the data let us take the biggest turn within the research. Separating the data according to its characteristics and mutual differences was the key step towards achieving any kind of relation. By taking into account the altitude of the stations, the region in which they are situated and by selecting specific (reoccurring) time frames smaller time series as subsets of the original data could be formed. This led to new calculations. Especially the correlation coefficients and the Spearman's rank correlation coefficients were changed by the various subsets. After investigating the significance of these values we were able to conclude that cloud top temperature and precipitation are correlated negatively. However, since correlation only generates a single value, we are not able to conclude anything more on this relationship in this research. The two vital elements of the precipitation system are definitely associated. It is not possible to tell whether one influences the other or converted. There is even the option that both are influenced by a third or more elements within the system. Based on the literature covered in chapter 2, the influence of multiple elements is credible.

Secondly, the thresholds for precipitation and CTT were investigated. There are various arguments for various precipitation thresholds. Considering the required millimetres for growing crops might increase this threshold up to 5 or 10 mm/day, while accurate observations by the institutes might require lower thresholds such as 1 or 0.1 mm/day. Once a threshold for precipitation is fixed, there are multiple possibilities for deciding on a CTT threshold. Taking into account both the purpose, the user and the consequences that a certain threshold brings along is essential in the process. Based on the data set in this research there are upper and lower limits that can be set for the occurrence of rain above or under certain temperatures, but varying the threshold based on the percentages they bring, requires taking into account the said circumstances. Apart from these detailed circumstances, the graphs in chapter 6 show that the incoherence of the data complicates this process. A significant amount of days in the data will be falsely considered a wet or dry for any chosen threshold.

Based on these processes and results there are multiple possibilities for future research, as will be elaborated in the following section.

7.2. Recommendations

In the previous chapters, many suggestions have been made for future research. The suggestions and other ideas will be set out here.

Within the process of separating the the time series, other choices or combinations than the ones in chapter 5 can be made. Possibly smaller subsets generate significantly different values from what is found in chapter 5. However, since the correlation coefficients were tested to be significant, that is what should be investigated above all: What can we concluded from this correlation and how can we use it in further research or model building?

Secondly, there are some mathematical tools that can be applied to this case, that might bring more useful or very different results: logistic regression, the (adjusted) R-squared values and multivariate regression that includes other cloud characteristics are all worth considering in future research.

One thing that could be of great value is the classification of the types of rain. The different types come with their own characteristic behaviour. Specific elements like the drop in temperature before or after rainfall occurs, might make way for reconsidering all of the previously mentioned mathematical tools over again.

Lastly, for defining sensible thresholds there are several things that need to be investigated first before they can be included in the modelling of these thresholds. First of all the purpose and the potential user of the thresholds need to be taken into account. Secondly the consequences of the data and potential forecasts need to be evaluated and quantified. Quantifying and taking into account those circumstances can create a more complete and accurate model for defining the thresholds.

Reflections on the project

Looking back, different methods and approaches could have been more useful and accurate or would have made the research more complete. Unfortunately, due to time limits, this is potentially only for further research.

First of all, there are some alterations that could have been made in the early steps that are worth looking into. Removing outliers or separating the dry and wet days upfront might have created stationary time series. In that case, we would have been able to look into ARIMA models or apply some of the tools in the frequency domain after all. Once the data is fit for these methods, there is much more potential for results that can actually be used in further research and model building. However, removing the dry days might also reduce the chances for finding any relation at all. This is because the transition from wet to dry days along with the changing temperature is what will define the relation in the end. We have to be careful in making changes to the time series since this can indirectly affect the image of the weather system it represents.
Secondly, based on the low R-squared values, it might have been a good idea to look into the significance of these values right away. From there we could either have rejected the whole regression model and the correlation coefficients so that more alterations could have been made and different approaches could be applied, or the adjusted R-squared could have been introduced along with R-squared. In the latter case, the features that affect these (adjusted) R-squared values could have been investigated and, if possible, altered to improve the values. Not only the R-squared values, but also the regression on itself could have been put through a significance test. In that case, depending on the degree of the regression, the null hypothesis would pose that the unknown parameters b_i in equation 3.18 are not significantly different from 0.
Moreover, multiple questions arise by looking at the generated p-values.What makes these p-values so extremely small? Are these values reasonable? What changes in the time series would have made the p-values change? By investigating mathematical theory on p-values and the data that is present, these question might be answered.
Lastly, for defining thresholds only very basic methods were used. In the approach, different and possibly better decisions could have been made. The observation of only the percentages of false dry and wet days limits the information to base decisions for these thresholds on. Taking into account the occurrence of rainfall for different temperatures would support the thresholds that can be chosen. Also, including both the bottom and top thresholds in one model creates a more realistic image of the situation. In such a model some specific cases could have been worked out more elaborate to emphasise the basic examples that are worked out in the conclusion of chapter 6.

In retrospect, we can conclude that both of the stated objectives have been investigated thoroughly. The objectives were clear but once the data was observed, the problem to be solved became more complex. The big variety of available mathematical tools and the complexity of the data created a big challenge in finding the right way to analyse the time series.

Bibliography

[1] S. Blyth and World Conservation Monitoring Centre. *Mountain Watch: Environmental Change & Sustainable Developmental in Mountains*. UNEP-WCMC biodiversity series. UNEP-WCMC, 2002. ISBN 9781899628209.

[2] George Edward Pelham Box and Gwilym M. Jenkins. *Time Series Analysis: Forecasting and Control*. Prentice Hall PTR, 3rd edition, 1994. ISBN 0130607746.

[3] Bureau of Meteorology Commonwealth of Australia. Weather Services, Cloud Types and Precipitation. URL `www.bom.gov.au/weather-services/`.

[4] East African Community. Agriculture and Food Security. URL `https://cloudatlas.wmo.int/`. Accessed: 2019-06-06.

[5] Jochen Dibbern. Automatic Weatherstations at DWD, October 2017.

[6] University Corporation for Atmospheric Research (UCAD). Cloud Image Gallery. URL `https://scied.ucar.edu/`. Accessed: 22-6-2019.

[7] Rolf Hut, Nick van de Giesen, John Selker, and M Andreini. The trans african hydro meteorological observatory. *AGU Fall Meeting Abstracts*, 12 2010.

[8] Koninklijk Nederlands Meteorologisch Instituut. Uitleg over: Regenintensiteit. URL `https://www.knmi.nl/kennis-en-datacentrum`. Accessed: 7-2019.

[9] G.M. Jenkins and D.G. Watts. *Spectral Analysis and its applications*. Holden-Day series in time series analysis. Holden-Day, 1969.

[10] Graeme L. Stephens and Christian D. Kummerow. The remote sensing of clouds and precipitation from space: A review. *Journal of The Atmospheric Sciences - J ATMOS SCI*, 64:3742–3765, 11 2007. doi: 10.1175/2006JAS2375.1.

[11] Bureau of Meteorology. Climate statistics for Australian locations: Definitions for rainfall. URL `http://www.bom.gov.au/climate`. Accessed: 7-2019.

[12] Met Office. Learn about weather, . URL `https://www.metoffice.gov.uk/`. Accessed: 17-6-2019.

[13] Met Office. UK and regional series, . URL `https://www.metoffice.gov.uk/research/climate/maps-and-data`. Accessed: 7-2019.

[14] World Meteorological Organisation. International Cloud Atlas: Manual on the Observation of Clouds and Other Meteors, 2017. URL `https://cloudatlas.wmo.int/`.

[15] Elberly College of Science Penn State University. Statistics Online, . URL `https://newonlinecourses.science.psu.edu/statprogram/`. Accessed: 4-2019.

[16] Elberly College of Science Penn State University. Sample ACF and Properties of AR(1) Model, . URL `https://newonlinecourses.science.psu.edu/stat510/lesson/1/1.2`. Accessed: 6-2019.

[17] J.A. Rice. *Mathematical Statistics and Data Analysis*. Advanced Series. Cengage Learning, 3rd edition, 2006. ISBN 9780534399429.

[18] Nick van de Giesen Arnold Heemink Sha Lu, Marie-Claire ten Veldhuis and Martin Verlaan. On the relationship between time series of gauge precipitation and satellite-based cloud-top temperature, 2019.

[19] American Meteorological Society. Meteorology Glossary, . URL `glossary.ametsoc.org/`. Accessed: 17-6-2019.

[20] American Meteorological Society. Glossary of Meteorology, . URL http://glossary.ametsoc.org/. Accessed: 7-2019.

[21] Ruey S. Tsay. Time series and forecasting: Brief history and future research. *Journal of the American Statistical Association*, 95(450):638–643, 2000. ISSN 01621459. URL http://www.jstor.org/stable/2669408.

[22] Nick van de Giesen, Rolf Hut, and John Selker. The trans-african hydro-meteorological observatory (tahmo). *Wiley Interdisciplinary Reviews: Water*, 1(4):341–348, 2014. doi: 10.1002/wat2.1034.

[23] Deutsche Wetterdienst. Wetterlexikon. URL https://www.dwd.de/DE/service/lexikon/Functions/glossar.html. Accessed: 7-2019.

A

Appendix: Python code

```
from netCDF4 import Dataset
import numpy as np
from matplotlib.lines import Line2D
import math
import matplotlib.pyplot as plt
from matplotlib.lines import Line2D
import statistics as stat
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import PolynomialFeatures
from scipy.stats import spearmanr
from scipy.stats import pearsonr


###Precipitation (time series) from CDC weather stations, 2000-2015 daily
CDC = Dataset(r'C:\Users\Ilona\Documents\Technische Wiskunde\BEP\cdc-daily_fill_2000-
    2015.nc')
precp = CDC.variables['precipitation']
timedate = CDC.variables['time']

###Cloud Top Temperature (time series) from 2000-2015, daily
CLARACDC = Dataset(r'C:\Users\Ilona\Documents\Technische Wiskunde\BEP\CLARA2-CDG-ts-daily_2000-
    2015.nc')
ctt=CLARACDC.variables['ctt']
```

A.1. Functions

```
###Turn date into list
def timetodate(time_element): #input is an integer with first 4 digits representing the year
    timestr=str(time_element)
    #print('timestr = '+ str(timestr))
    timelst=list(timestr)
    #print('timelst = '+ str(timelst))
    date=[2000,0,0]
    #year
    a=int(timelst[2])
    if a==1:
        date[0]=2010
    b=int(timelst[3])
    if b != 0:
        date[0]=date[0]+b
```

```
    #month
    c=int(timelst[4])
    if c ==1:
        date[1]=10
    d=int(timelst[5])
    date[1]=date[1]+d
    #day
    e=int(timelst[6])
    if e != 0:
        date[2]=e*10
    f=int(timelst[7])
    date[2]=date[2]+f
    return date #date is of the form [2000,12,31]

###
def get_x_month_y_years(time_array,month,year,data_array): #in 1 year
    l=len(time_array)
    month_data=np.array([])
    for t in range(l):
        date=timetodate(time_array[t])
        if date[0]==year:
            if date[1]==month:
                month_data=np.append(month_data,data_array[t])
    return month_data

def get_x_month_all_years(time_array,month,data_array): #in all years #for 1 station
    l=len(time_array)
    month_data=np.array([])
    for t in range(l):
        td=time_array[t]
        date=timetodate(td)
        if date[1]==month:
            month_data=np.append(month_data,data_array[t])
    return month_data

###Seperate the stations in different region: 0,1,2 (-999 is no region)
###Stat(i) gives the indices where a station of region i is in the TS
def seperate(labels,precp,ctt):
    stat0=[]
    stat1=[]
    stat2=[]
    for l in range(0,len(labels)):
        if labels[l]==0:
            stat0.append(l)
        elif labels[l]==1:
            stat1.append(l)
        elif labels[l]==2:
            stat2.append(l)
    return stat0,stat1,stat2

###Provide prep and ctt time serie for region i
def TS_region(i,labels,precp,ctt):
    pregion=[]
    cregion=[]
    stations=seperate(labels,precp,ctt)[i]
    for stn in stations:
```

```
            pregion.append(precp[stn])
            cregion.append(ctt[stn])
        return pregion,cregion

    ###Seperate the stations in different heights
    def seperate_height(elevation):
        low0=np.array([]) #0-300
        mid1=np.array([]) #300-1000
        high2=np.array([]) #1000+
        for h in range(0,941):
            if elevation[h]<300:
                low0=np.append(low0,h)
            elif 300 <= elevation[h] < 1000:
                mid1=np.append(mid1,h)
            elif elevation[h] >= 1000:
                high2=np.append(high2,h)
        return low0,mid1,high2

###Ctt and precp for the different heights
def TS_elev(i,elevation,precp,ctt):
    lep=len(precp[0])
    lec=len(ctt[0])
    pelev=np.array([]).reshape(0,lep)
    celev=np.array([]).reshape(0,lec)
    stations=seperate_height(elevation)[i]
    print(len(stations))
    for stn in stations:
        pelev=np.vstack([pelev,precp[stn]])
        celev=np.vstack([celev,ctt[stn]])
    return pelev,celev

###Function for regression
def regression(x,y,degree):
    X = x[:, np.newaxis]
    Y = y[:, np.newaxis]
    XX= PolynomialFeatures(degree=degree, include_bias=False).fit_transform(X)
    model=LinearRegression()
    model.fit(XX,Y)
    ypred = model.predict(XX)
    xsort=X[X[:,0].argsort()]
    ypredsort=ypred[X[:,0].argsort()]
    values=[model.score(XX,Y), model.intercept_, model.coef_]
    return xsort, ypredsort, values

###Function for upperthreshold over all 941 stations
def general_upper_threshold(precp,ctt,gaugthr):
    T=0
    TT=1000
    for stn in range(0,len(precp)):
        for p in range(0,len(precp[stn])):
            if precp[stn][p] >= gaugthr:
                if ctt[stn][p]>T:
                    T=ctt[stn][p]
    for stn in range(0,len(precp)):
        for pp in range(0,len(precp[stn])):
            if precp[stn][pp] < gaugthr:
```

```
                if T < ctt[stn][pp] <TT:
                    TT=ctt[stn][pp]
    return TT


###Function for lowerthreshold over all 941 stations
def general_lower_threshold(precp,ctt,gaugthr):
    T=1000
    TT=0
    for stn in range(0,len(precp)):
        for p in range(0,len(precp[stn])):
            if precp[stn][p] >= gaugthr:
                if ctt[stn][p]<T:
                    T=ctt[stn][p]
    for stn in range(0,len(precp)):
        for pp in range(0,len(precp[stn])):
            if precp[stn][pp] < gaugthr:
                if TT < ctt[stn][pp] <T:
                    TT=ctt[stn][pp]
    return TT

 ###Function for upperthreshold over one long time series with multiple stations
def upper_threshold(precp,ctt,gaugthr):
    T=0
    TT=1000
    for p in range(0,len(precp)):
        if precp[p] >= gaugthr:
            if ctt[p]>T:
                T=ctt[p]
    for pp in  range(0,len(precp)):
        if precp[pp] < gaugthr:
            if T < ctt[pp] <TT:
                TT=ctt[pp]
    return TT

###Function forlowerthreshold over one long time series with multiple stations
def lower_threshold(precp,ctt,gaugthr):
    T=1000
    TT=0
    for p in  range(0,len(precp)):
        if precp[p] >= gaugthr:
            if ctt[p]<T:
                T=ctt[p]
    for pp in  range(0,len(precp)):
        if precp[pp] < gaugthr:
            if TT < ctt[pp] <T:
                TT=ctt[pp]
    return TT

###Calculate the percentage of false wet days for an upper threshold
def falsewet(ctt,precp,thresPRECP,thresCTT):
    totwetCTT=0
    dryprecp=0
    for i in range(0,5844):
        if ctt[i] < thresCTT:
            totwetCTT=totwetCTT+1
```

```
            if precp[i]<thresPRECP:
                dryprecp=dryprecp+1
    if totwetCTT==0:
        perc=0
    else:
        perc=(dryprecp/totwetCTT)*100
    return perc

###Calculate the percentage of false dry days for an upper threshold
def falsedry(ctt,precp,thresPRECP,thresCTT):
    lp=len(precp)
    totdryCTT=0
    wetprecp=0
    for i in range(lp):
        if ctt[i]>=thresCTT:
            totdryCTT=totdryCTT+1
            if precp[i]>=thresPRECP:
                wetprecp=wetprecp+1
    if totdryCTT==0:
        perc=0
    else:
        perc=(wetprecp/totdryCTT)*100
    return perc

###Calculate the percentage of false wet days for a bottom threshold
def falsewetbot(ctt,precp,thresPRECP,thresCTT):
    lp=len(precp)
    totwetCTT=0
    dryprecp=0
    for i in range(lp):
        if ctt[i] > thresCTT:
            totwetCTT=totwetCTT+1
            if precp[i]<thresPRECP:
                dryprecp=dryprecp+1
    if totwetCTT==0:
        perc=0
    else:
        perc=(dryprecp/totwetCTT)*100
    return perc

###Calculate the percentage of false wet days for an upper threshold
def falsedrybot(ctt,precp,thresPRECP,thresCTT):
    lp=len(precp)
    totdryCTT=0
    wetprecp=0
    for i in range(lp):
        if ctt[i]<=thresCTT:
            totdryCTT=totdryCTT+1
            if precp[i]>=thresPRECP:
                wetprecp=wetprecp+1
    if totdryCTT==0:
        perc=0
    else:
        perc=(wetprecp/totdryCTT)*100
    return perc
```

A.2. Chapter 4

```
###Histogram precipitation
histp=plt.figure(1)
plt.hist(precp[444], bins='auto',color='indigo')
plt.xlabel('Precipitation (mm/day)')
plt.ylabel('Frequency')
histp.show()

###Rolling mean precipitation
rm=plt.figure(2)
meann=[]
for i in range(0,5800-100,100):
    meann.append(precp[444][i:i+100].mean())
plt.plot(meann, color='indigo')
plt.ylabel('mean precipitation (mm/day)')
rm.show()

###Rolling variance precipitation
rv=plt.figure(3)
varr=[]
for i in range(0,5800-100,100):
    varr.append(precp[444][i:i+100].var())
plt.plot(varr, color='indigo')
plt.ylabel('var precipitation (mm/day)')
rv.show()

###Histogram CTT
histc=plt.figure(4)
plt.hist(ctt[444], bins='auto',color='indigo')
plt.xlabel('CTT (Kelvin))')
plt.ylabel('Frequency')
histc.show()

###Rolling mean CTT
rmm=plt.figure(5)
meann=[]
for i in range(0,5800-100,100):
    meann.append(ctt[444][i:i+100].mean())
plt.plot(meann, color='orchid')
plt.ylabel('mean CTT (Kelvin)')
rmm.show()

###Rolling variance CTT
rvv=plt.figure(6)
varr=[]
for i in range(0,5800-100,100):
    varr.append(ctt[444][i:i+100].var())
plt.plot(varr, color='orchid')
plt.ylabel('var CTT (Kelvin)')
rvv.show()

###Line plot with 2 y axis for ctt and precipitation
fig, ax1 = plt.subplots(1,1,figsize=(16,15), dpi= 100)
ax1.plot(plottime, PR, color='indigo',linewidth=1)
ax2 = ax1.twinx()
```

```
ax2.plot(plottime, CTT, color='orchid',linewidth=1)
ax1.set_xlabel('Time(days)', fontsize=15)
ax1.set_ylabel('Precipitation (mm/day)', color='indigo', fontsize=15)
ax1.set_ylim(0,140)
ax2.set_ylabel("CTT (Kelvin)", color='orchid', fontsize=15)
ax2.set_ylim(180,290)
ax2.set_title("Precipitation and CTT in time, daily 2000-2015, station 444", fontsize=17)
plt.show()
```

A.3. Chapter 5

```
###Saving files with different months in different regions
mo=1
allyrc=np.array([])
allyrp=np.array([])
for stn in range(0,277):
    yrc=get_x_month_all_years(time,mo,ctt0[stn])
    yrp=get_x_month_all_years(time,mo,precp0[stn])
    allyrc=np.concatenate((allyrc,yrc),axis=0)
    allyrp=np.concatenate((allyrp,yrp),axis=0)
np.savez('C:\\Users\Ilona\Documents\Technische Wiskunde\BEP\\marallyrreg0',allyrc=allyrc,allyrp
    =allyrp)
```

⋮

```
###Saving files with different months in different heights
mo=1
allyrc=np.array([])
allyrp=np.array([])
for stn in range(0,517):
    yrc=get_x_month_all_years(time,mo,cttlow[stn])
    yrp=get_x_month_all_years(time,mo,precplow[stn])
    allyrc=np.concatenate((allyrc,yrc),axis=0)
    allyrp=np.concatenate((allyrp,yrp),axis=0)
np.savez('C:\\Users\Ilona\Documents\Technische Wiskunde\BEP\\janallyrev0',allyrc=allyrc,allyrp
    =allyrp)
```

⋮

```
###Saving files with different seasons in different regions
allyrc=np.array([])
allyrp=np.array([])
for yr in range(2000,2016):
    for stn in range(0,277):
        yrc0=get_x_month_y_years(time,12,yr,ctt0[stn])
        yrc1=get_x_month_y_years(time,1,yr,ctt0[stn])
        yrc2=get_x_month_y_years(time,2,yr,ctt0[stn])
        yrp0=get_x_month_y_years(time,12,yr,precp0[stn])
        yrp1=get_x_month_y_years(time,1,yr,precp0[stn])
        yrp2=get_x_month_y_years(time,2,yr,precp0[stn])
        allyrc=np.concatenate((allyrc,yrc0,yrc1,yrc2),axis=0)
        allyrp=np.concatenate((allyrp,yrp0,yrp1,yrp2),axis=0)
np.savez('C:\\Users\Ilona\Documents\Technische Wiskunde\BEP\\winterallyrreg0',allyrc=allyrc,allyrp
    =allyrp)
```

⋮

```
regions = np.load('C:\\Users\Ilona\Documents\Technische Wiskunde\BEP\\regions.npz')
```

```
precp0=regions['precp0']
ctt0=regions['ctt0']
precp1=regions['precp1']
ctt1=regions['ctt1']
precp2=regions['precp2']
ctt2=regions['ctt2']

regionss = np.load('C:\\Users\Ilona\Documents\Technische␣Wiskunde\BEP\elevation.npz')
precplow=regionss['precplow0']
cttlow=regionss['cttlow0']
precpmid=regionss['precpmid1']
cttmid=regionss['cttmid1']
precphigh=regionss['precphigh2']
ctthigh=regionss['ctthigh2']

jan0el=np.load('C:\\Users\Ilona\Documents\Technische␣Wiskunde\BEP\\janallyrev0.npz')
feb0el=np.load('C:\\Users\Ilona\Documents\Technische␣Wiskunde\BEP\\feballyrev0.npz')

⋮

jan1el=np.load('C:\\Users\Ilona\Documents\Technische␣Wiskunde\BEP\\janallyrev1.npz')
feb1el=np.load('C:\\Users\Ilona\Documents\Technische␣Wiskunde\BEP\\feballyrev1.npz')

⋮

jan0elc=jan0el['allyrc']
jan0elp=jan0el['allyrp']

⋮

jan1elc=jan1el['allyrc']
jan1elp=jan1el['allyrp']

⋮

jan0 = np.load('C:\\Users\Ilona\Documents\Technische␣Wiskunde\BEP\\janallyrreg0.npz')
feb0 = np.load('C:\\Users\Ilona\Documents\Technische␣Wiskunde\BEP\\feballyrreg0.npz')

⋮

jan1 = np.load('C:\\Users\Ilona\Documents\Technische␣Wiskunde\BEP\\janallyrreg1.npz')
feb1 = np.load('C:\\Users\Ilona\Documents\Technische␣Wiskunde\BEP\\feballyrreg1.npz')

⋮

jan0c=jan0['allyrc']
jan0p=jan0['allyrp']

⋮

winter0 = np.load('C:\\Users\Ilona\Documents\Technische␣Wiskunde\BEP\\winterallyrreg0.npz')
spring0 = np.load('C:\\Users\Ilona\Documents\Technische␣Wiskunde\BEP\\springallyrreg0.npz')

⋮

winter0c = winter0['allyrc']
spring0c = spring0['allyrc']

⋮
```

```
months0c=[jan0c,feb0c,mar0c,apr0c,may0c,jun0c,jul0c,aug0c,sep0c,oct0c,nov0c,dec0c]

⋮

###Calculate CC, SPR and R_squared for all different subsets
A0_CC=[]
A0_SPR=[]
A0_R2d1=[]
A0_R2d2=[]
A0_R2d3=[]
for mo in range(0,12):
    yrc=months0c[mo]
    yrp=months0p[mo]
    A0_CC.append(np.corrcoef(yrc,yrp)[0,1])
    A0_SPR.append(spearmanr(yrc,yrp)[0])
    xsorted,ypred,values1=regression(yrc,yrp,1)
    xsorted,ypred,values2=regression(yrc,yrp,2)
    xsorted,ypred,values3=regression(yrc,yrp,3)
    A0_R2d1.append(values1[0])
    A0_R2d2.append(values2[0])
    A0_R2d3.append(values3[0])
np.savez('C:\\Users\Ilona\Documents\Technische␣Wiskunde\BEP\\A0_CC',A0_CC=A0_CC)
np.savez('C:\\Users\Ilona\Documents\Technische␣Wiskunde\BEP\\A0_SPR',A0_SPR=A0_SPR)
np.savez('C:\\Users\Ilona\Documents\Technische␣Wiskunde\BEP\\A0_R2',d1=A0_R2d1,d2=A0_R2d2,d3=A0_R2d3)
np.savez('C:\\Users\Ilona\Documents\Technische␣Wiskunde\BEP\\A0_CC',A0_CC=A0_CC)
np.savez('C:\\Users\Ilona\Documents\Technische␣Wiskunde\BEP\\A0_SPR',A0_SPR=A0_SPR)
np.savez('C:\\Users\Ilona\Documents\Technische␣Wiskunde\BEP\\A0_R2',d1=A0_R2d1,d2=A0_R2d2,d3=A0_R2d3)

⋮

###P-value for different subsets
pv=np.load('C:\\Users\Ilona\Documents\Technische␣Wiskunde\BEP\\reg0.npz')
a,b=pearsonr(pv['reg0c'],pv['reg0p'])
c,d=spearmanr(pv['reg0c'],pv['reg0p'])
print(b)
print(d)

⋮
```

A.4. Chapter 6

```
##Create precp and ctt for all 3 regions
precp0,ctt0=TS_region(0,labels,precp,ctt) #region0 (277 stations)
precp1,ctt1=TS_region(1,labels,precp,ctt) #region1 (113 stations)
precp2,ctt2=TS_region(2,labels,precp,ctt) #region2 (345 stations)

gt=[0.1,0.2,1,2,5,10,25]

for gaugthr in gt:
    ##Generate thresholds over all 941 stations
    gen_upthr1=general_upper_threshold(precp[0:310],ctt[0:310],gaugthr)
    gen_upthr2=general_upper_threshold(precp[310:620],ctt[310:620],gaugthr)
    gen_upthr3=general_upper_threshold(precp[620:914],ctt[620:914],gaugthr)
    gen_upthr=np.amax([gen_upthr1,gen_upthr2,gen_upthr3])
    gen_lothr1=general_lower_threshold(precp[0:310],ctt[0:310],gaugthr)
    gen_lothr2=general_lower_threshold(precp[310:620],ctt[310:620],gaugthr)
    gen_lothr3=general_lower_threshold(precp[620:914],ctt[620:914],gaugthr)
    gen_lothr=np.amin([gen_lothr1,gen_lothr2,gen_lothr3])
```

```
    ##Generate thresholds for regions
    thr_up_0=general_upper_threshold(precp0,ctt0,gaugthr) #region0
    thr_up_1=general_upper_threshold(precp1,ctt1,gaugthr) #region1
    thr_up_2=general_upper_threshold(precp2,ctt2,gaugthr) #region2
    thr_lo_0=general_lower_threshold(precp0,ctt0,gaugthr) #region0
    thr_lo_1=general_lower_threshold(precp1,ctt1,gaugthr) #region1
    thr_lo_2=general_lower_threshold(precp2,ctt2,gaugthr) #region2
    ###Prints
    print('The gauge threshold is: ' + str(gaugthr))
    print('The upper threshold for all stns is: '+ str(gen_upthr))
    print('The upper threshold for region 0 is: ' + str(thr_up_0) + ' degrees Kelvin')
    print('The upper threshold for region 1 is: ' + str(thr_up_1) + ' degrees Kelvin')
    print('The upper threshold for region 2 is: ' + str(thr_up_2) + ' degrees Kelvin')
    print('The lower threshold for all stns is: '+ str(gen_lothr))
    print('The lower threshold for region 0 is: ' + str(thr_lo_0) + ' degrees Kelvin')
    print('The lower threshold for region 1 is: ' + str(thr_lo_1) + ' degrees Kelvin')
    print('The lower threshold for region 2 is: ' + str(thr_lo_2) + ' degrees Kelvin')

elev0=np.load('C:\\Users\Ilona\Documents\Technische Wiskunde\BEP\elevlow.npz')
elev1=np.load('C:\\Users\Ilona\Documents\Technische Wiskunde\BEP\elevmid.npz')
elev2=np.load('C:\\Users\Ilona\Documents\Technische Wiskunde\BEP\elevhigh.npz')
elevp0=elev0['elevlowp']
elevc0=elev0['elevlowc']
elevp1=elev1['elevmidp']
elevc1=elev1['elevmidc']
elevp2=elev2['elevhighp']
elevc2=elev2['elevhighc']
for gaugthr in gt:
    ##Generate thresholds for regions
    thr_up_0=upper_threshold(elevp0,elevc0,gaugthr) #region0
    thr_up_1=upper_threshold(elevp1,elevc1,gaugthr) #region1
    thr_up_2=upper_threshold(elevp2,elevc2,gaugthr) #region2
    thr_lo_0=lower_threshold(elevp0,elevc0,gaugthr) #region0
    thr_lo_1=lower_threshold(elevp1,elevc1,gaugthr) #region1
    thr_lo_2=lower_threshold(elevp2,elevc2,gaugthr) #region2
    ###Prints
    print('The gauge threshold is: ' + str(gaugthr))
    print('The upper threshold for elev low 0 is: ' + str(thr_up_0) + ' degrees Kelvin')
    print('The upper threshold for elev mid 1 is: ' + str(thr_up_1) + ' degrees Kelvin')
    print('The upper threshold for elev high 2 is: ' + str(thr_up_2) + ' degrees Kelvin')
    print('..')
    print('The lower threshold for elev low 0 is: ' + str(thr_lo_0) + ' degrees Kelvin')
    print('The lower threshold for elev mid 1 is: ' + str(thr_lo_1) + ' degrees Kelvin')
    print('The lower threshold for elev high 2 is: ' + str(thr_lo_2) + ' degrees Kelvin')

###Scatter plot different stations
r0=plt.figure(1)
plt.scatter(x=ctt0,y=precp0,s=1,c='orchid')
plt.ylim(0, 300)
plt.xlim(180, 320)
plt.xlabel('Precipitation (mm/day)')
plt.ylabel('CTT (Kelvin)')
plt.show()
plt.title('Region 0')
r0.show()
r1=plt.figure(2)
```

```
plt.scatter(x=ctt1,y=precp1,s=1, c='darkorchid')
plt.ylim(0, 300)
plt.xlim(180, 320)
plt.ylabel('Precipitation (mm/day)')
plt.xlabel('CTT (Kelvin)')
plt.title('Region 1')
r1.show()
r2=plt.figure(3)
plt.scatter(x=ctt2,y=precp2,s=1, c='indigo')
plt.ylim(0, 300)
plt.xlim(180, 320)
plt.ylabel('Precipitation (mm/day)')
plt.xlabel('CTT (Kelvin)')
plt.title('Region 2')
r2.show()
rall=plt.figure(4)
plt.scatter(x=ctt0,y=precp0,s=1,c='orchid')
plt.scatter(x=ctt2,y=precp2,s=1,c='indigo')
plt.scatter(x=ctt1,y=precp1,s=1,c='darkorchid')
plt.ylim(0, 300)
plt.xlim(180, 320)
plt.xlabel('Precipitation (mm/day)')
plt.ylabel('CTT (Kelvin)')
plt.title('All regions')
colors = ['orchid', 'darkorchid', 'indigo']
lines = [Line2D([0], [0], color=c, linewidth=3, linestyle='', marker='o') for c in colors]
labels = ['Lower region (0)','Upperright region (1)', 'Upperleft region (2)']
plt.legend(lines,labels)
rall.show()

regionss = np.load('C:\\Users\Ilona\Documents\Technische Wiskunde\BEP\elevation.npz')
precplow=regionss['precplow0']
cttlow=regionss['cttlow0']
precpmid=regionss['precpmid1']
cttmid=regionss['cttmid1']
precphigh=regionss['precphigh2']
ctthigh=regionss['ctthigh2']


###Scatter plot different heights
r0=plt.figure(1)
plt.scatter(x=cttlow,y=precplow,s=1,c='orchid')
plt.ylim(0, 300)
plt.xlim(180, 320)
plt.xlabel('Precipitation (mm/day)')
plt.ylabel('CTT (Kelvin)')
plt.show()
plt.title('Elevation level: low')
r0.show()
r1=plt.figure(2)
plt.scatter(x=cttmid,y=precpmid,s=1, c='darkorchid')
plt.ylim(0, 300)
plt.xlim(180, 320)
plt.ylabel('Precipitation (mm/day)')
plt.xlabel('CTT (Kelvin)')
plt.title('Elevation level: mid')
```

```
r1.show()
r2=plt.figure(3)
plt.scatter(x=ctthigh,y=precphigh,s=1, c='indigo')
plt.ylim(0, 300)
plt.xlim(180, 320)
plt.ylabel('Precipitation (mm/day)')
plt.xlabel('CTT (Kelvin)')
plt.title('Elevation level: high')
r2.show()
rall=plt.figure(4)
plt.scatter(x=cttmid,y=precpmid,s=1,c='darkorchid')
plt.scatter(x=cttlow,y=precplow,s=1,c='orchid')
plt.scatter(x=ctthigh,y=precphigh,s=1,c='indigo')
plt.ylim(0, 300)
plt.xlim(180, 320)
plt.xlabel('Precipitation (mm/day)')
plt.ylabel('CTT (Kelvin)')
plt.title('All elevation levels')
colors = ['orchid', 'darkorchid', 'indigo']
lines = [Line2D([0], [0], color=c, linewidth=3, linestyle='', marker='o') for c in colors]
labels = ['Low (0-300 m)','Mid (300-1000 m)', 'High (>=1000 m)']
plt.legend(lines,labels)
rall.show()

####For the top threshold
for g in gt:
    mD=int(np.ceil(np.amax(ctt[specstn])))
    nD=int(np.floor(np.amin(ctt[specstn])))
    temp=[]
    percsD=[]
    percsW=[]
    for T in range(nD,mD+1):
        temp.append(T)
        peD=falsedry(ctt_stn,precp_stn,g,T)
        percsD.append(peD)
        peW=falsewet(ctt_stn,precp_stn,g,T)
        percsW.append(peW)

    fig, ax1 = plt.subplots(1,1,figsize=(16,15), dpi= 100)
    ax1.plot(temp, percsD, color='indigo',linewidth=2)
    ax1.set_xlim(220,290)
    ax2 = ax1.twinx()
    ax2.plot(temp,percsW, color='orchid',linewidth=2)
    ax1.set_xlabel('CTT threshold', fontsize=20)
    ax1.set_ylim(0,100)
    ax1.set_ylabel('% of false dry days', color='indigo', fontsize=20)
    ax2.set_ylabel('% of false wet days', color='orchid', fontsize=20)
    ax2.set_ylim(0,100)
    ax2.set_xlim(220,290)
    plt.suptitle('The percentages of -false days- based on the chosen CTT (top) threshold T',
        fontsize=21)
    plt.title('For gauge threshold '+str(g) + ' mm/day', fontsize=20)
    plt.show()

####For the bottom threshold
for g in gt:
```

```
mD=int(np.ceil(np.amax(ctt[specstn])))
nD=int(np.floor(np.amin(ctt[specstn])))
temp=[]
percsD=[]
percsW=[]
for T in range(nD,mD+1):
    temp.append(T)
    peD=falsedrybot(ctt_stn,precp_stn,g,T)
    percsD.append(peD)
    peW=falsewetbot(ctt_stn,precp_stn,g,T)
    percsW.append(peW)

fig, ax1 = plt.subplots(1,1,figsize=(16,15), dpi= 100)
ax1.plot(temp, percsD, color='indigo',linewidth=2)
ax1.set_xlim(205,275)
ax2 = ax1.twinx()
ax2.plot(temp,percsW, color='orchid',linewidth=2)
ax1.set_xlabel('CTT threshold', fontsize=20)
ax1.set_ylim(0,100)
ax1.set_ylabel('% of false dry days', color='indigo', fontsize=20)
ax2.set_ylabel('% of false wet days', color='orchid', fontsize=20)
ax2.set_ylim(0,100)
ax1.set_xlim(205,275)
plt.suptitle('The percentages of -false days- based on the chosen CTT (bottom) threshold T',
    fontsize=21)
plt.title('For gauge threshold '+str(g) + ' mm/day', fontsize=20)
plt.show()
```