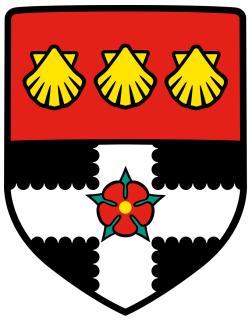
UNIVERSITY OF READING

Department of Mathematics and Statistics

The spatial statistical distribution for multiple rainfall intensities over Ghana



Jennifer E. Israelsson

A thesis presented for the degree of Doctor of Philosophy

2021

Abstract

Accurately measuring rainfall is important in most parts of the world due to our reliance on it for food, energy and drinking water. The seasonality and interannual variability at a daily scale needs to be understood for our current climate in order to understand how it is changing with global warming. However, obtaining this information for large land areas, such as the African continent, down to the few km level instead of a regional or country levels is difficult due to the very large number of rain gauges required. Rain gauges is the preferred measurement method since it provides the most accurate amount for a specific location. Satellites on the other hand can easily collect information over a whole continent, but these cannot directly measure rainfall. It therefore needs to be calibrated against ground observations and in addition only returns an area average estimate instead of a point estimate. An optimal observation product would draw information from both of these sources when appropriate, but this requires a detailed understanding about the radius the rain gauge information can be extrapolated to and the relation between the two sources of information.

This thesis aims to achieve improvements in these two areas. The first contribution is to provide new methods for estimating the correlation distance for all rainfall intensities, information which can be used to inform about the radius a gauge measurement can be extrapolated. The second is to provide an improved distribution function for daily rain gauge measurements associated with satellite estimates at a 4km scale. The combination of these two can improve the merging of information from rain gauges and satellite estimates by drawing information from the most accurate source at each location. The application of the new methodologies are demonstrated by applying these to a new, dense daily rain gauge data set collected over Ghana.

A non-parametric methodology for estimating the correlation distance is developed, which can easily be adapted for a given study region. Based on comparing the observed with the expected co-occurrence probabilities, it by design takes into account the rainfall climate for the specific time period and rainfall intensity considered. The annual variation of the correlation range for four intensity classes is estimated over southern Ghana, and compared with estimates from previous studies for other west African countries.

To estimate the dependence structure in extreme values, and especially for values larger than the ones observed, multivariate Extreme Value Theory provides the appropriate framework. A semi-parametric estimator for the coefficient of tail dependence is proposed and the performance is evaluated in a finite sample simulation study. The extremal dependence structure for different times of the year is evaluate by applying the estimator to the daily rain gauge data set collected over Ghana. Limitations stemming from strong seasonality and missing values are addressed.

A qualitative study for assessing the distributional fit of rain gauge measurements conditioned on a satellite rainfall estimate is performed, with the 4km satellite estimates given by the TAMSAT data set. A skewed distribution with heavier than normal distributed tails is found to generally be suitable.

Declaration

I declare that the work in this thesis has been done by myself and the use of all material from other sources has been properly and fully acknowledged.

Chapter 3 corresponds to published work: Israelsson, J., Black, E., Neves, C., Torgbor, F.F., Greatrex, H., Tanu, M., Lamptey, P.N.L., 'The spatial correlation structure of rainfall at the local scale over southern Ghana', *Journal of Hydrology: Regional Studies*, vol. 31, 2020, https://doi.org/10.1016/j.ejrh.2020.100720

Jennifer E. Israelsson

Acknowledgements

Firstly, a huge thanks to my supervisors Professor Emily Black and Dr Claudia Neves for their support, encouragement and the occasional push when needed. I am so grateful that you both wanted to be a part of this project, which would not have been possible without your combined wealth of knowledge! I have enjoyed our meetings and really value all the knowledge that you have shared along our four-year long journey. Doing a PhD in normal circumstances is a challenge and under pandemic conditions even worse, but together we made it through. I would also like to thank Dr David Walshaw for his collaboration on the work in Chapter 5, and all of his advice after joining for the second half of this project.

Secondly, I must thank the Ghana Meteorological Agency for sharing their rain gauge data set with me, which I know took a huge amount of time and effort to collect and digitalise. This project would not have been possible without it. I would also like to thank the MPE CDT, especially the Reading staff, for their continuous support both academically but also well-being. The regular Tuesday coffee mornings, in person and online, has been a very appreciated break from work. I would also like to acknowledge the additional funds which allowed my trip to Ghana, it was definitely one of the highlights on my PhD journey!

I would like to thank my cohort, and in particular Ieva, Elena and Mariana, for all the laughs and hugs, whether it be for celebrating our achievements or for giving support after tough feedback. The countless number of office coffee breaks (the first coffee machine even gave up) has been a true joy and is one of the things that I am really going to miss.

Continuing with coffee, I must thank Dave for the many hours at the gym, walks and coffee shops, listening to me complaining. I am honestly not sure what I would have done the past 18 months without your support.

My final and biggest thank you goes to my fantastic fiancé Bruce who has been standing next to me for this entire emotional roller-coaster, giving me endless love and support!

Contents

List of Tables

Mathematical notation

$Z(\cdot)$	Spatial process
F	Distribution function
$F^{(n)}$	Right-continuous empirical distribution function
$\bar{F}$	$1 - F$, tail distribution function
X_i	Unranked sample point
$X_{n,k}$	k^{th} ascending sample point in a sample of size n
$R(X_i)$	Rank of X_i among $(X_1, ..., X_n)$
$\mathbf{1}_A(h), \mathbf{1}_p$	indicator function, 1 if $h \in A$ or p true and 0 else
i.i.d	independent and identically distributed
$\mathcal{N}$	normal distribution
Φ^{-1}	inverse of the standard normal distribution function
$f^{\leftarrow}$	general inverse function of f
RV_α	Regularly varying with index α
$\mathcal{L}(\cdot)$	Slowly varying function
$\lfloor x \rfloor$	Integer part of x
$\xrightarrow{P}$	convergence in probability
$\xrightarrow{d}$	convergence in distribution
$a \wedge b$	$\min(a, b)$
$a \vee b$	$\max(a, b)$
O_p	If $X_n = O_p(a_n)$ then $\mathbb{P}\left(\left\lvert \frac{X_n}{a_n} \right\rvert > M\right) < \epsilon$, $\forall n > N$
o_p	If $X_n = o_p(a_n)$ then $\lim_{n\to\infty} \left(\mathbb{P}\left\lvert \frac{X_n}{a_n} \right\rvert \geq \epsilon\right) = 0$, $\forall \epsilon > 0$

Abbreviations

WAM	West African Monsoon
ITCZ	Intertropical convergence zone
TAMSAT	Tropical Applications of Meteorology using SAtllite data and ground-based observations
CCD	Cold Cloud Duration
MCS	Mesoscale Convective System
GMet	Ghana Meteorological Agency
CV	Coefficient of Variation
TIR	Thermal Infrared
PMW	Passive Microwave
TRMM	Tropical Rainfall Measuring Mission
CHIRPS	Climate Hazards group Infrared Precipitation with Stations
GPCC	Global Precipitation Climatology Centre
IDW	Inverse Distance Weighting
RFE	(Satellite) Rainfall Estimate
CDF	Cumulative Distribution Function
PDF	Probability Density Function
CI	Confidence Interval
EVT	Extreme Value Theory
CTD	Coefficient of Tail Dependence
MSE	Mean Square Error

Chapter 1

Introduction

1.1 Motivation

Consistently measuring rainfall is of great importance for most places in the world since we as a species are dependent on water. Most countries in the world depends on rain fed agriculture with limited access to irrigation, making the food production sensitive to changes in timing of the growing season and rainfall amounts. This since the crops and growing practices are usually chosen based on historic rainfall patterns. Nowadays, rainfall measurements are also needed to verify and calibrate global climate models, which is our main tool for deriving climate change projections. If we do not know the current rainfall patterns in terms of amounts, seasonality and variability it is impossible to understand how this is likely to change under global warming.

The most accurate way of measuring rainfall is through ground based rain gauge measurements, since these can capture exactly how much rain that fell in that particular location. However, in order to consistently monitor rainfall in this way, a dense and uniformly spaced rain gauge network is required to capture rainfall variability everywhere. Unless the rain gauges are automatic and can transmit their measurements, the amount in each rain gauge must be manually recorded each day, preferably in even 24 hour intervals. These two factors combined makes this a very expensive and labour intensive method for recording rainfall. It further comes with the issue that records cannot be created afterwards if someone forgot to record the rainfall one day. This requires a long-term commitment and persistence to collect long enough records to be of use, usually 30 years of length for climate studies.

Another method for collecting rainfall records is through satellite estimates. There are a number of different ways of doing this (Section 4.1) but they are all based on approximating cloud measurements recorded by satellites and converting this into rainfall amounts. Once the

satellite is launched, it can monitor and record the variable of interest over entire continents with very limited human interaction, and the information can easily be shared widely around the world. This solves both the coverage and labour issues with rain gauges, but these estimates instead have their own drawbacks. The first issue is that the rainfall is not measured directly but instead approximated, meaning that the rainfall record will be an estimated amount rather than an actual one. The estimates are further given as an area average since estimated on a grid in contrast to rain gauges that record point estimates, meaning that satellite estimates will miss any variabilities within a grid cell. The biggest limitation is however that these rainfall estimates must be calibrated against rain gauge measurements, which requires both rain gauge records and an understanding of how these two are related.

Given the two information sources different strengths and limitations, a combination of the two would provide the most and best information. With a dense rain gauge network, the unmonitored areas between the gauges can be estimated by interpolating the measurements (Section 2.1). This however requires information about the correlation structure, meaning over which distances two locations can be assumed correlated and how this tails of with distance. This is something that is not well understood so far for daily rainfall in west Africa due to the historic lack of dense enough rain gauge datasets. Since this area experiences both small drizzly events and mesoscale convective storms (Section 1.2.2) it is further likely, but so far not widely researched, that the correlation distance is not equal for all rainfall events.

In order to use satellite estimates to fill in the blank areas where rain gauges are missing or too far away to be interpolated, one first needs to understand what distance is 'too far away' and how this is related to the rainfall amount observed at the rain gauge. We must further understand how the ground based rain gauge values are related to the satellite cloud observations in order to convert the latter to a well calibrated rainfall estimate. With rain gauges recording point wise and continuously and satellites area averages and at an interval, there is naturally not a one-to-one relation between the two but rather a distribution of rain gauge values observed for each satellite cloud observation. First when a solid understanding of these two things, the correlation structure and the distributional relation, is achieved can the two information sources be combined to obtain a merged product which maximises the available rainfall records. It is to advance this area the following thesis aims are proposed.

1.2 Thesis aims

The aim of this thesis is to develop improved methodologies for estimating the key components in spatial statistics, specifically the correlation structure and conditional distributions, which are required for relating rain gauge data with satellite imagery. The main scientific questions addressed to meet this aim are:

1) *How can we accurately estimate the correlation structure in daily rainfall?*

2) *Does the correlation distance vary with storm intensity?*

3) *What is a suitable distribution to model the full range of observed gauge observations related to a given satellite image?*

The following sections provide the background and motivation for each of the questions raised above and how the work in this thesis addresses them.

1.2.1 How can we accurately estimate the correlation structure in daily rainfall?

The correlation range is essential in spatial statistics, and more specifically geostatisitics, which is the class of statistical tools aimed at describing the spatial continuity that is present in most environmental phenomena. The First Law of Geography by Waldo Tobler states that: *"Everything is related to everything else, but near things are more related than distant things"* (Sui (2004)). This might be true in principle, but the correlation range specifies when two locations can be assumed to be approximately independent, and therefore not related anymore.

Getting an accurate estimate at an unobserved location is of interest in any field where a spatial process is observed as a point process, such as weather stations or drill holes collecting data at specific locations, since observing the whole area is not possible. To get information at the unobserved locations, some form of interpolation must be performed. A key decision is which observations to include for the estimate, and whether these should be weighted depending on the assumed similarity to the unobserved location. Inverse Distance Weighting estimates the value at the new location as a linear combination of either a set number of closest observations or all observations within some area, and weight them based on some function of the Euclidean distance and potentially additional covariates. This does however not take into account the relation between the observed locations, possibly leading to spatially clustered observations more heavily influencing the weighted estimate compared to isolated observations.

A standard method in geostatistics that attempts to reduce this issue is *kriging*, which is the *best linear unbiased estimator*, since it is designed to minimise the error variance by assigning weights based on the relation between all observations and the new location (see Chapter 4 for full method; Isaaks and Srivastave (1989)). To assign the weights, a covariance function defined by three parameters; sill, nugget and range (see Section 2) are used. The nugget represents the measurement error and the sill the variance of the variable, but the work here will primarily be focused on the range. The range parameter determines at what distance two locations are no longer correlated, and should therefore not contribute in the weighted estimate. This range parameter is what will be called the correlation range, and needs to be estimated.

The commonly used correlation estimation method 'Pearson correlation coefficient' is very restricted, since it can only detect linear relationships between the two variables. This is unsuitable for rainfall, since the dependence changes at an exponential rate rather than a linear (Section 3.2.4). A method commonly used in geostatistics is the sample variogram (Section

3.2), which estimates the sample variance for pairs located at various distance lags. This has been used for daily rainfall (Greatrex et al. (2014)), but the large absolute spread in rainfall amounts leads to a large variance already for short distances and it is often difficult to determine a stabilising plateau, which indicates the correlation range.

Alternative methods for estimating rainfall correlation at short time scales, for both gauge measurements and satellite observations, have been proposed (Section 3.1), however a majority of these chooses an arbitrary 'null' value for determining when the correlation range is reached. In agreement with Tobler, a region share a general rainfall climate, which can be denoted the background rainfall climatology, which leads to the observations being weakly correlated despite no true dependence still present (Section 3.2.4) In other words, for a particular region or country there is a certain probability of observing light or heavy rainfall because of the atmospheric conditions during the different seasons. In Ghana, there is a much higher probability of observing heavy rainfall during June compared to April because of the position of the Intertropical convergence zone (ITCZ), which drives the west African monsoon (WAM). Since any two locations can observe heavy rainfall with a certain seasonally varying probability, the two can observe this on the same day even if they are completely independent of each other.

The methodology derived in Chapter 3 addresses this issue by both estimating the background and the observed dependence, and can therefore incorporate information about the local seasonal rainfall climatology. This provides a data determined rather than a user decided correlation range. The method is non-parametric, further reducing the uncertainty by removing the need for making distribution or relation assumptions. It additionally provides a straightforward way to estimate the correlation range for various intensity classes, since it takes into account the background probability of any subset of observations.

For the highest observed intensities, also called the extreme values, a different method is however needed since these are too few to accurately model the dependence with the first model. Often when one wants to model the most extreme values, methods in the field of Extreme Value Theory (EVT) are used since these are developed specifically to only work with the largest values in a sample. For estimating dependence in the extremes, multivariate EVT must be used which has become a very active field of research in the past decades due to the pressing need to better understand current and future multiple risks (Huser and Wadsworth (2020)). A major limitation in this framework, is that one needs to decide if the model one works with is asymptotically dependent or independent (Sibuya (1960)), where the former corresponds to there being a non-zero probability of the two variables being extreme simultaneously. A great deal of research is being done to construct models that allows for both, but they are often

restricted to either dependent or independent in practice (Section 6.1).

With the correlation range defined as the distance where two locations can be assumed independent, an asymptotically independent model is preferred since it does not impose that two locations must be dependent. Ledford and Tawn (1997) introduced a submodel, meaning an extension of a previous model with a larger set of possible outcomes, that allows one to measure the association between two variables despite them being asymptotically independent. This is called the *coefficient of tail dependence* and it provides a method to investigate changes in association (a weak type of dependence) in the case of asymptotic independence (Section 5.2). The regular extreme value index estimators can be used to estimate this coefficient, but these suffer from the classical *bias-vs-variance* trade-off, which means that the bias is small (large) for smaller (larger) sample sizes included in the estimation and vice versa for the variance. In Chapter 5 an estimator extending the mean-of-order-p univariate estimator (Gomes and Caeiro (2014)) is developed along side a reduced bias variation, taking into account the uncertainty arising from the marginal transformation (Section 5.3) in the bivariate setting.

The two models developed compliment each other by together making it possible to estimate the correlation range for any intensity, a key feature for answering thesis aim question 2. That is, given a rain gauge data set one can use the non-parametric model for estimating the correlation ranges for the regular rainfall values, which are observed every year from low to high intensity events. The parametric extreme value method can then be used on the most extreme values, usually the 3-5% largest, to understand over what distances it is likely to observe two extreme events simultaneously. These observations might only have occurred a few times, and therefore could not be estimated by the first method due to the small sample size. This information can then be utilised to estimate the risk of co-occurring extremes, and to understand how far away gauges must be in order to be pooled together and thereby create a larger sample size for univariate estimates.

1.2.2 Does the correlation distance vary with storm intensity?

There is a large discrepancy between the research community that models the regular part of the daily rainfall regime, and the extreme value community which only considers the most extreme daily observations. A common assumption for the first group is the one of an intensity invariant correlation range. That is, given a time of the season, the correlation range will be the same regardless of the storm intensity and therefore also the rainfall generating process, meaning that the correlation range used is a seasonally averaged estimate instead of changing with the event. This assumption can appear rather unrealistic when modelling daily rainfall, considering the widespread possible rainfall processes and the physical behaviour difference between small, drizzly events and large frontal systems.

In the extreme value statistics community on the other hand, it is generally accepted that environmental variables tend to become more localised as the intensity increases due to physical constraints (Huser and Wadsworth (2020)). However, very few studies have confirm this assumption; due to the aforementioned issue of appropriate methodology and data availability. These two opposing assumptions pose the question of whether the correlation distance is approximately constant up to some intensity before it starts to decrease, or if variations are present for the entire spectrum of rainfall intensities and nearly always are ignored.

To address this question, west Africa and specifically Ghana is chosen as the study region. This is a region which experiences a large number of storms in a year, and equally importantly a wide range of rainfall processes with varying physical features (McGregor and Nieuwolt (1998)). Maranan et al. (2018) defined, based on satellite observations, seven different rainfall processes over west Africa based on their depth, horizontal extent and reflective factor, which roughly measures the water content. It was concluded that even though the vast majority of the number of rainfall events are of the 'Moderate' or 'Strong' convection type, and a relatively large proportion warm rain along the coast, nearly all of the rainfall amount stems from the 'Strong convection' and Mesoscale convective type events. This indicates that there is a connection between the physical extent of the rainfall process and the intensity of it, but potentially in the other direction than what is assumed for very extreme events. Since the most frequently occurring events are smaller in extent than stronger and rarer events, estimating a general correlation range for all these events reduces the amount of information that could be extrapolated and risks underestimating the rainfall amounts at ungauged locations.

The second motivation for this particular region is the availability of a dense, long-term rain gauge data set, curated by the Ghana Meteorological Agency (GMet), of which a detailed description will be provided in Chapter 3. This consists of 590 unevenly distributed stations, with some of the records dating back to 1940, which is incredibly rare on the African continent. Washington et al. (2004) highlighted nearly 20 years ago that the gauge density over Africa is about 8 times less dense than recommended by the World Meteorology Organisation (WMO), and Nicholson et al. (2018) among others has showed that the gauge network has steadily declined since the 1980s. Figure 1.1 displays the gauge network currently reporting to the WMO Integrated Global Observing Systems (WIGOS) (**WMO2019**) in Europe and Africa, clearly highlighting the density difference between the two continents. The distances between the African gauges are often larger than one might expect to be the correlation distance, making it impossible to derive any conclusions about the smaller scale differences.

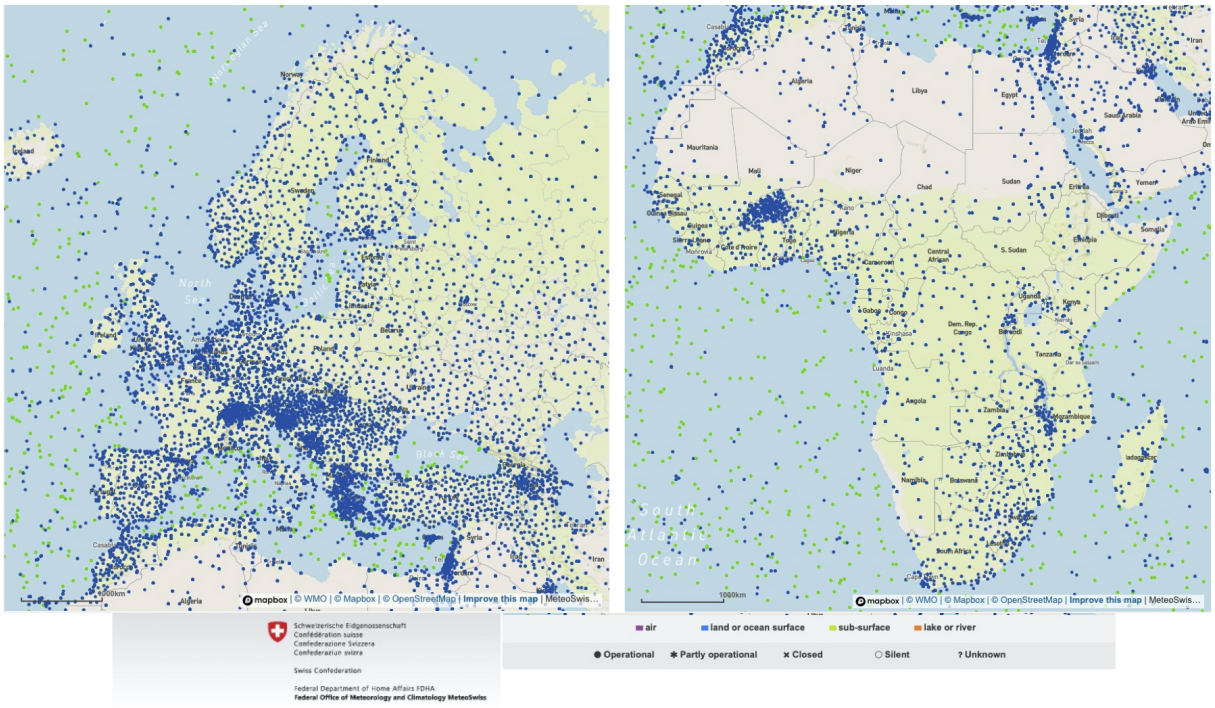


Figure 1.1: Location of all surface bases stations and platforms reporting to WIGOS (**WMO2019**), (left) Europe (latitude (71, 29), longitude (-20, 56)) and (right) Africa (latitude (37, -34), longitude (-20, 56)).

Chapter 3 and 6 tackle the two rainfall regimes, regular and extremes, by estimating the dependence as a function of distance for 5 intensity bands. This is done at the fine resolution of 10km distance steps, filling in an information gap present in many previous studies (Moron et al. (2007), Greatrex et al. (2014)). By using the same data set for both of the regimes, differences and similarities can more easily be compared, something that is often not possible. This since analysis of this sort are often performed by different research groups, either coming

from the meteorology side or statisticians from the extreme value side, and therefore not using the same data set or even the same region since what is available to one group might not be to another. The results in these chapters demonstrate the significant difference in correlation structure for different rainfall intensities and provide support for considering an intensity varying correlation range.

1.2.3 What is a suitable conditional distribution to model the full range of observed gauge observations related to a given satellite image?

Due to the scarcity of high quality rain gauges and weather stations over Africa, satellites are being used to provide consistent monitoring of the rainfall, covering the entire continent. Good coverage from geostationary satellites providing thermal infrared (TIR) imagery has been available since the early 80s, providing nearly 40 years of consistent data (Maidment et al. (2020)), therefore exceeding the 30 years of data that is usually required for climate studies. Weather observations are crucial in any region for monitoring changes and more recently validating climate models, but in Africa even more so since a large part of the population depend on rain-feed crops and energy is increasingly being generated by hydropower (IEA (2020)).

However mapping observed rain gauge measurements with satellite imagery is far from trivial, since one needs to understand the relation between the rain gauge measurements and the area average satellite estimates. There exists a multitude of satellite rainfall products over Africa, providing data at a wide range of spatial and temporal scales (Section 4.1), and a common feature for many of them is that gauge measurement are used to calibrate the chosen estimation model.

The TIR data does not provide rainfall information, but cloud cover temperature, from which the duration of clouds colder than a specific temperature (CCD) can be related to the observed rainfall amount (Section4.2.1; Maidment et al. (2020)). There is however not a one-to-one relation between these two values, but rather a distribution of gauge observations are related to a given CCD value as demonstrated in Figure 1.2.

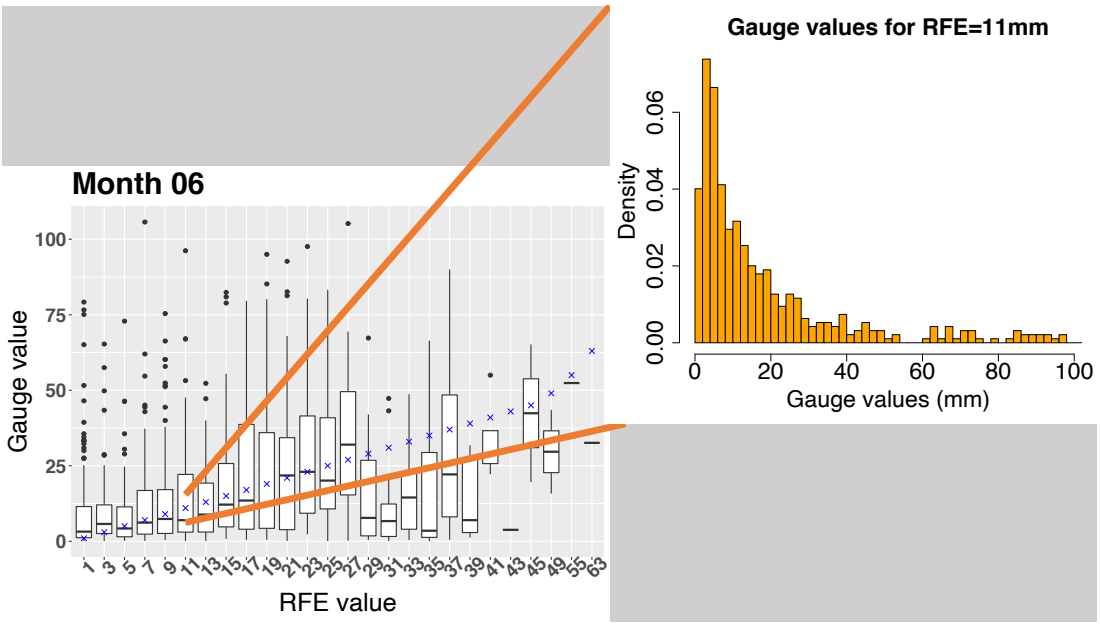


Figure 1.2: Box and Whiskers graph over all gauge measurements associated with a satellite rainfall estimate (RFE), histogram demonstrating the distribution of gauge values for a particular RFE value.

An important difference between satellite estimated rainfall and ground observations, is the difference in scale. Rain gauges are point processes that accurately captures the amount at that specific location, but leaves large gaps of information. Satellite data is collected on a fine or coarse scale (Section 4.1), but nevertheless as a smooth spatial process with one grid cell representing an average of the precipitation within that area.

The difference in observational scale and the distributional relation between the two information sources results in the satellite estimated map providing a good mean estimate. It however fails to represent the most extreme values, and there is also a relatively large uncertainty associated with the estimate. Teo and Grimes (2007) developed a geostatistical sequential simulation algorithm to generate an ensemble of rainfall estimates, given a single TIR imagery, in order to capture the uncertainty in the estimates. The basis for the algorithm is to randomly sample a rainfall amount given the CCD value at a selection of 'seed' cells, and then apply kriging to generate a realistic rainfall field with the correct spatial dependence structure (cf. Section 4.2.2).

A similar approach can be used to 'merge' ground observation data with satellite estimates, and thereby incorporate all the available information into a single estimation map (Section 4.2.2). This merged estimate will then draw information from the most reliable source at each instance, and can return the more extreme values from the satellite estimate conditional distribution when these are observed.

A key aspect for both of these applications, is that the conditional distribution accurately describes the full range of observed gauge measurements for a given satellite value. It is well known that rainfall follows a skewed distribution, but a large selection of these exists with varying tail behaviour. A too short-tailed distribution leads to an underestimation of the uncertainty and unrealistically many observations will correspond to the very highest quantiles of the distribution, and vice versa for a too heavy tailed.

This is addressed in Chapter 4 by comparing the gamma distribution used in Teo and Grimes (2007) and Greatrex et al. (2014) for describing the conditional gauge distribution, with the heavier tailed lognormal distribution. The results in the chapter shows that a well-parametrised lognormal distribution better models the full range of values, especially the tails of the observations. This provides an improvement in the understanding of the uncertainty associated with the satellite rainfall estimates.

This knowledge, coupled with the answers derived from aim 1 and 2, provides a deeper understanding of how to combine and map gauge observed measurements with satellite derived estimates and therefore addresses the overarching thesis aim. Starting with a set of gauge measurements and satellite rainfall estimates, the intensity varying correlation range will inform over what distances the gauge information can be utilised. By using an intensity dependent range instead of a mean range, the overall uncertainty will decrease since observations with a shorter ranges will not be extended outside of this and observations with longer ranges will be fully utilised. The improved conditional distribution will allow for a closer-to-the-true range of rainfall values, especially the larger values, which will result in a monitoring product which better represents the ground observed rainfall.

1.3 Thesis structure

Chapter 2 provides an introduction to the two main areas of statistics used in this thesis, geostatistics and Extreme value statistics. The first part presents the method of kriging, by detailing the semivariogram and covariogram functions and how these two are related. The second part briefly explains the concept of Extreme value statistics and how this is different from other statistical areas.

Chapter 3 provides a detailed description of the daily rainfall data set collected over Ghana from 1940-2017, that will be further used in Chapter 4 and 6, and presents the rainfall climatology for the different climate regions of the country. An algorithm for estimating the correlation range for different intensities, based on estimating the co-occurrence probability and comparing it to the background probability, is developed. This is used to estimate the correlation distance for storms in four different intensity classes for each month of the season. This information is used to answer question 1 and 2 in the thesis aims.

In Chapter 4 the focus is on answering question 3 from the thesis aim by evaluating the fit of a lognormal distribution to measured rain gauge values from the previously introduced Ghana rain gauge data set, associated with the TAMSAT satellite rainfall estimates. The evaluation of the fit is considered for four months, representing different phases of the monsoon, and a wide range of rainfall estimate values to determine the overall best performing distribution parameters.

Chapter 5 also relates to thesis aim question 2, where a new estimator for the coefficient of tail dependence, which measures the dependence strength in the case of asymptotic independence, is developed. The asymptotic normal distribution is derived and a reduced bias version is introduced. The performance compared to the Hill estimator is evaluated through an extensive finite sample simulation study.

In Chapter 6 the analysis done in Chapter 3 is extended by applying the reduced bias estimator developed in Chapter 5 to the extreme observations in the rain gauge data set. A non-stationarity and cluster analysis is performed for a set of baseline stations, from which the dependence structure is thereafter estimated. Shorter dependence distances are derived compared to Chapter 3, but several issues with the results are highlighted.

Finally, Chapter 7 presents the main conclusions from Chapter 3-6 along with some directions for further work.

Chapter 2

Statistical background

The following short chapter outlines the statistical knowledge needed to fully understand the methods introduced and applied throughout the thesis. The first section focuses on the geostatistical part, with a special focus on the method of kriging since this will be the main method used. The second part will introduce the area of extreme value statistics, starting with the univariate setting and following up with the multivariate extension. The multivariate section will provide an overview of the aims and issues, but leaves the theoretical definitions to Chapter 5.

2.1 Geostatistics

Geostatistics is as mentioned in the introduction the area of statistics related to spatio-temporal processes. The aim is usually to estimate the value at one or more unmeasured locations given the information from other locations, plus the knowledge that physical processes often are smooth in space. This means that locations close to each other should have values similar to each other, and the forced similarity should decay with distance and potentially other covariates, such as mountains or large bodies of water. As mentioned in the introduction, a key parameter is the correlation range, which determines over what distances we assume the spatially smooth condition to be true. In the following section, the impact of this parameter, and the others mentioned in Section 1.2.1 will be described. Using Figure 2.1 as our example problem, the aim is to determine the value at the orange star using either all or a subset of the measurements at the blue dots. The two circles represents two different correlation ranges which leads to a different number of sample points being included in the estimate.

A simple method for determining the value at an unmeasured location is the Inverse Distance Weighting (IDW). In this case, the value at the star is given as a weighted sum of all the

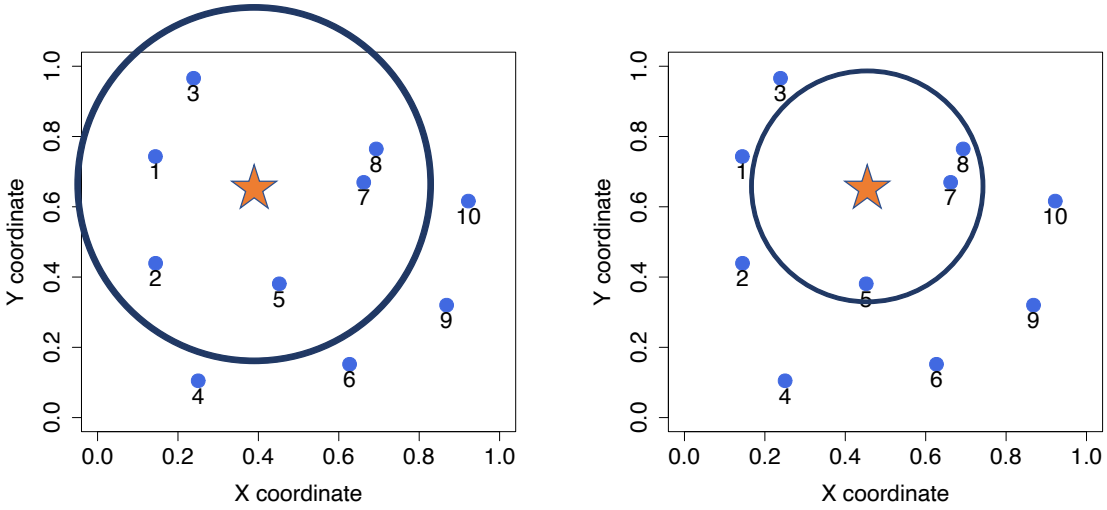


Figure 2.1: Example plot of an unmeasured location (star), 10 measured locations (blue dots) and two different correlation ranges.

values within the circles, and the weights are only dependent on the distance between the star and the point. How this weight dependence on the distance is up to the person performing the analysis, but could for example linearly decay if it is a very smooth parameter such as temperature or exponentially if it is more variable. The large drawback with this method is the lack of taking into account the dependence between the blue dots. In this example, points 7 and 8 are very close to each other so will have measured similar values. Point 5 on the other hand is at a similar distance from the star location, but in a different direction, hence will probably have a different value. In the right plot in Figure 2.1 where only these three points are included, the star value will mostly be influenced by the points 7 and 8 since they each individually carry the same weight as point 5.

2.1.1 Kriging

It is the issue of clustered observations, together with obtaining a method for deriving a measure on the uncertainty, *kriging* was developed. Kriging is an interpolation method, which given suitable assumption on the parameters, returns the *best linear unbiased estimate* (BLUE). It is modelled based on a Gaussian (normal distribution) process with some prior knowledge about the covariance structure, that is how large the circle should be and how the correlation should decay within it. The empirical idea of kriging was originally introduced in a master

thesis by Danie G. KRIGE in the 1950s, but developed and the theoretical basis formalised by the mathematician Georges Matheron in the early 1960s (Chilès and Desassis (2018)), and is now a widely used tool in the geostatistical field. There are several different types of kriging depending on the level of assumptions made and the smoothness of the field one wants to model. *Simple* kriging is the most restricted version, where we assume that the expected value $\mu = 0$ everywhere and that the covariance function is known. A slightly more flexible version is *ordinary* kriging where one assumes that the expected value is constant everywhere but unknown. One further level of complexity is to assume that the mean is following a general polynomial trend, which is referred to as *Universal* kriging.

For all versions of kriging, the key part is to determine the shape of the covariance function and the values of the three parameters; sill, nugget and range. In order to estimate the covariance function one first needs to estimate the semivariogram function, $\gamma(h)$, which estimates the variance between two locations, to then translate this to the desired covariance function, $c(h)$. Figure 2.2 demonstrates the three parameters and describes the inverse relationship between the two functions, which is given by

$$c(h) = \sigma^2 - \gamma(h) \tag{2.1}$$

where σ^2 is the sill value.

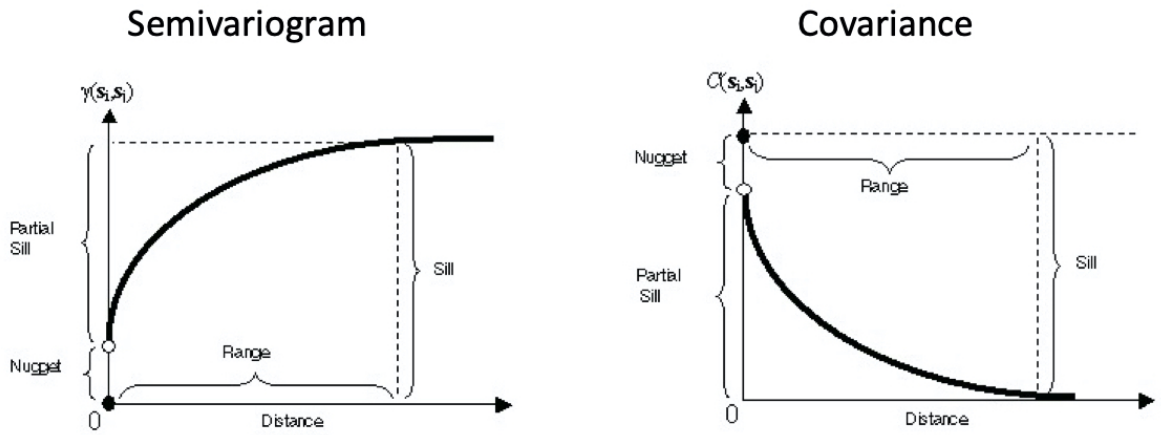


Figure 2.2: Relation between the (left) semivariogram and the (right) covariance function with the three parameters nugget, sill, range visualised. Figures produced by pro.arcgis.com

The semivariogram function can be obtained by first creating an empirical semivariogram and then fitting a semivariogram function through the points. In the isotropic setting where

we assume that the variability is the same in all directions, the empirical semivariogram is constructed by calculating the variance between all locations as a function of the distance between them. In the case of different variance structures in different directions, due to for example mountains or atmospheric patterns, only locations in the same directions can be included. Since points with similar values will result in a smaller variance, we expect the empirical semivariogram values to be close to 0 for short distances and converge to the spatial process variance for longer distances. If we assume that the mean is the same for all points, then this can be estimated by Matheron's classical estimator (Matheron and Blondel (1962)), defined as

$$\hat{\gamma}(h \pm \delta) = \frac{1}{2|N(h \pm \delta)|} \sum_{(i,j) \in N(h \pm \delta)} |z_i - z_j|^2 \tag{2.2}$$

where z_i is the observed value at location i, $N(h \pm \delta)$ is the set of pairs within the spatial lag $h \pm \delta$ and $|N(h \pm \delta)|$ is the number of pairs in that distance range. This will however only provide estimates for a limited number of distances, marked by the points in Figure 2.3, which is why we need a function to get an estimate for all distances. The line in Figure 2.3 shows the fitted semivariogram function through the empirical sample points

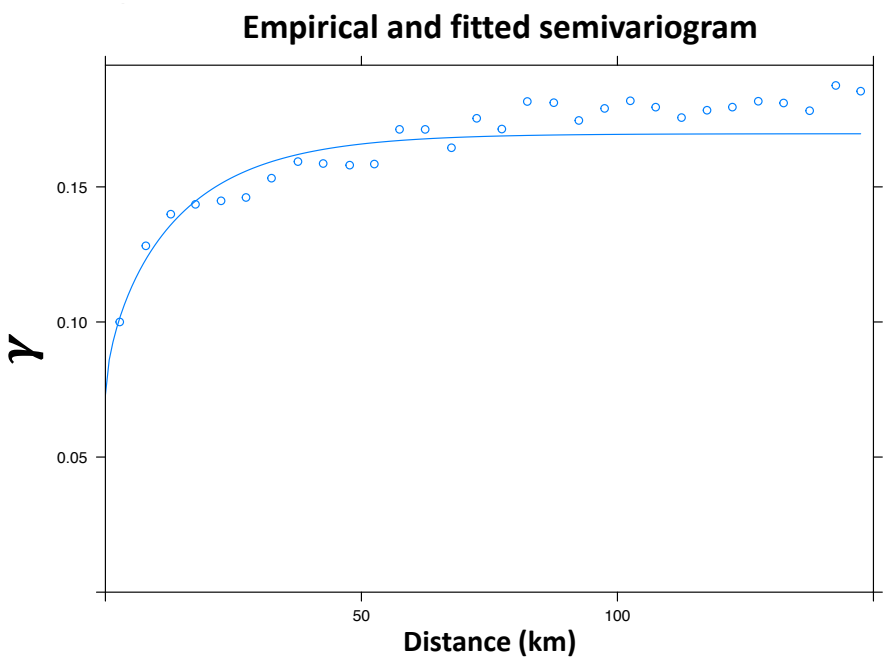


Figure 2.3: Empirical (points) and fitted (line) semivariogram.

The fitted line is given by a semivariogram function and depends on the three before mentioned parameters nugget, sill and range. The interpretation of the three parameters are as follows (Isaaks and Srivastave (1989))

- **Sill:** the maximum value that the functions attains, either definitely or asymptotically depending on the function. A larger sill value corresponds to a large area variance but does not impact the kriging estimate.
- **Range:** the distance at which the sill is reached and therefore the two locations are uncorrelated. Mainly has an effect on the number of points included in the estimation.
- **Nugget:** the jump at the near-0 distance, representing the measurement error and small scale variability. In a perfect world, measurements taken very close to each other should have the same value and therefore be perfectly correlated and zero variance. This is however nearly never the case in reality where small scale variation and measurement errors prevents this, which is reflected in the nugget effect. A large nugget results in all locations within the range having a similar weighting, since the difference in value is small.

The most commonly used semivariogram functions are given below with h denoting the distance between two locations, n, s, r denoting the nugget, sill, range respectively and the indicator function $\mathbb{1}_A(h)$ is 1 if $h \in A$ (h belongs to the range A) and 0 else. Figure 2.4 demonstrates the difference in shape for the first three functions.

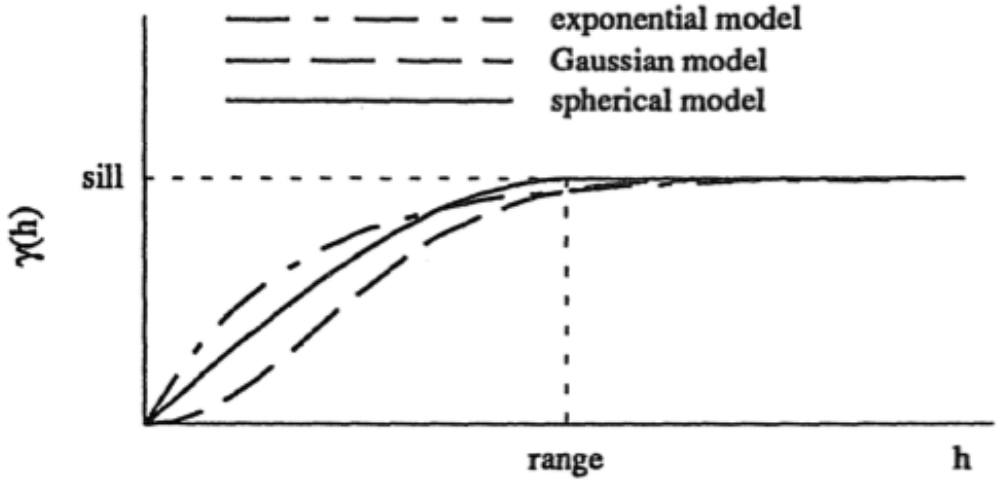


Figure 2.4: Examples of the three semivariogram functions exponential, spherical and gaussian with the same nugget, sill and range values. Figure from Isaaks and Srivastave (1989).

Exponential

$$c(h) = s^2 - (s-n)\left(1 - e^{-h/(r)}\right) + n\mathbb{1}_{(0,\infty)}(h)$$

Spherical

$$c(h) = s^2 - (s-n)\left(\left(\frac{3h}{2r} - \frac{h^3}{2r^3}\right)\mathbb{1}_{(0,r)}(h) + \mathbb{1}_{[r,\infty]}(h)\right) + n\mathbb{1}_{(0,\infty)}(h)$$

Gaussian

$$c(h) = s^2 - (s-n)\left(1 - e^{-h^2/(r^2)}\right) + n\mathbb{1}_{(0,\infty)}(h)$$

The indicator function in the last term in the functions clarifies that the nugget effect is included at any distance $\epsilon > 0$. Another model that is popular in the statistics community but less used in the meteorological setting is the following flexible, but more complicated model

Matérn

$$c(h) = s^2 \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{h}{r}\right)^{\nu} K_\nu\left(\frac{h}{r}\right)$$

where $\nu > 0$ is a shape parameter, $\Gamma(.)$ is the gamma function defined by $\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}dx$ and K_ν is the modified Bessel function of the second kind and order ν.

The covariance function can now be defined through Equation 2.1. After specifying the covariance function, one can proceed with estimating the value at the non-measured location x_0 (orange star). Denoting the observed values $Z(x_i) = z_i$ (blue dots) and the weights $w_i(x_0)$, $i = 1, .., N$, the estimate at our new location x_0 is given by

$$\hat{Z}(x_0) = \sum_{i=1}^{N} w_i(x_0) \times Z(x_i) \tag{2.3}$$

In the case of simple kriging, the weights $w_i = w_i(x_0)$ are estimated by

$$\begin{pmatrix} w_1 \\ \vdots \\ w_N \end{pmatrix} = \begin{pmatrix} c(x_1,x_1) & \cdots & c(x_1x_N) \\ \vdots & \ddots & \vdots \\ c(x_N,x_1) & \cdots & c(x_N,x_N) \end{pmatrix}^{-1} \begin{pmatrix} c(x_1,x_0) \\ \vdots \\ c(x_N,x_0) \end{pmatrix}$$

From this, the kriging error can also be estimated by

$$\text{Var}\left(\hat{Z}(x_0) - Z(x_0)\right) = c(x_0, x_0) - \begin{pmatrix} c(x_1, x_0) \\ \vdots \\ c(x_N, x_0) \end{pmatrix}' \begin{pmatrix} c(x_1, x_1) & \cdots & c(x_1 x_N) \\ \vdots & \ddots & \vdots \\ c(x_N, x_1) & \cdots & c(x_N, x_N) \end{pmatrix}^{-1} \begin{pmatrix} c(x_1, x_0) \\ \vdots \\ c(x_N, x_0) \end{pmatrix} \tag{2.4}$$

where the first term is the variance of the Gaussian process at the point x_0 and the second term the variance of the estimate.

2.2 Extreme value theory

Extreme value theory (EVT) is as the name suggests, the study of the most extreme values in a distribution. Anyone studying statistics will early on learn that one cannot extrapolate outside of the measured sample, since regular statistics are based on minimising the total error between the data points and the assumed distribution. With this aim, most of the weight to determine a good fit will naturally be given to the bulk of the data, since this is where most of our sample points lie. This unfortunately means that the fit in the tails can be rather poor, hence any extrapolation further into the tail would be based on an already poor fit. There are however a great number of situations where one would like to make inferences about values that has not been observed yet, especially when it comes to estimating risk. Insurance companies and infrastructure planners often want to know for example what intensity of rainfall or the largest pay out amount to expect once every 100 years, even though only 20 years of data is available. It is for these kind of situations extreme value has been developed, to provide a theoretically justified way of estimating return periods (e.g. how often we will see 200mm rain in a 24 hour period) or very high quantiles (e.g. the strongest wind we will measure in a 100 year period). This is achieved by only focusing on the tails and ignoring the rest of the distribution, since we already have suitable methods for estimating that. Figure 2.5 demonstrates the issue with a near perfect fit in the main part of the distribution but a rather poor fit in the tail.

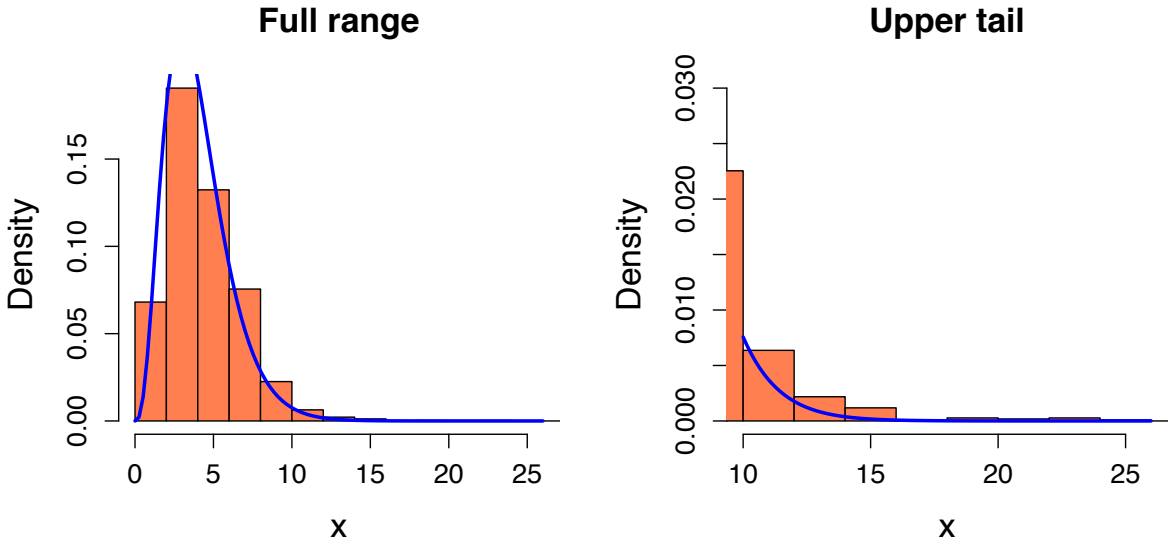


Figure 2.5: Density fit to a sample for the full sample (left) and tail (right).

2.2.1 Univariate statistics

Since we are only considering the largest values of our sample, the concept of *order statistics* is fundamental in EVT. For a sample of n independent and identically distributed (same distribution and not correlated, i.i.d) sample points $X_1, X_2, ..., X_n$, the order statistics of this sample is $X_{1,n} \leq X_{2,n} \leq ... \leq X_{n,n}$. By considering this new notation of ordered samples we can easily describe and model the largest values in the sample, where we commonly say that we choose the k largest values with k much smaller than n. In traditional statistics the aim is often to find the distribution of the mean, that is its central value and some variability around it. The basis for doing so is the central limit theorem, which states that

$$\sqrt{n}\left(\frac{(X_1 + X_2 + ... + X_n)/n - E(X)}{\sqrt{\text{var}(X)}}\right) \tag{2.5}$$

converges to a standard normal distribution as $n \to \infty$. We say that Equation 2.5 is *asymptotically* normally distributed, since it only holds true for n infinitely large. In EVT we are instead interested in finding the distribution of the maximum of the distribution, which is usually denoted $M_{X,n} := \max_{1 \leq i \leq n} X_i$. If we in Equation 2.5 replace the sum with the maximum $M_{X,n}$ and the two constants $nE(X)$, $\sqrt{n\,\text{var}(X)}$ by a sequence of numbers b_n and $a_n > 0$ respectively, we reach a similar distribution expression

$$\mathbb{P}\left(\frac{M_{X,n} - b_n}{a_n} \leq x\right) \to G(X) \tag{2.6}$$

as $n \to \infty$. The main question here is for which distributions of X there exists sequences of numbers $b_n \in \mathbb{R}$, $a_n > 0$ such that Equation 2.6 exists and $G(X)$ is not equal to a constant, also called non-degenerate. If this holds true, then we say that G belongs to the family of extreme value distributions. If $M_{X,n}$ was not standardised, it would clearly converge to the largest value of the distribution of X, denoted x^*, and $G(X)$ would be a degenerate distribution function only taking the value of x^*. We say that the distribution X is *max-stable* if it satisfies

$$a_n X + b_n \stackrel{d}{=} M_{X,n}$$

for appropriate constants $a_n > 0$, $b_n \in \mathbb{R}$. This means that the extreme value distributions are the max-stable, hence if we know that a distribution is max-stable then we also know that it is an extreme value distribution. It has been proved by Fisher and Tippett (1928) and Gnedenko (1943) that G can only be one of three possible distributions, and Haan (1970) further showed that all of these three distributions can be written as one extreme value distribution

$$G_\xi(x) = \exp\left(-(1+\xi x)^{-1/\xi}\right), \quad \text{for } 1+\xi x > 0$$

where $\xi \in \mathbb{R}$ is the *extreme value index*. This is a key quantity in EVT because it determines how heavy tailed a distribution is depending on if ξ is larger, equal or smaller than 0. Simply speaking, it gives information about if the distribution has a finite maximum value or how quickly the right tail converges towards a density value of 0, which is a measure of how extreme values this distribution can attain and how frequent. Figure 2.6 displays this, where we can see that the Weibull distribution ($\xi < 0$) has a finite right endpoint and the Fréchet/Pareto distribution ($\xi > 0$) converges much slower towards the x-axis compared to the Gumbel distribution ($\xi = 0$).

One can also flip the question and ask, 'Given that $G(x)$ is a possible limit distribution for the sequence $a^{-1}(X_{n,n} - b_n)$, what are the necessary and sufficient conditions on the distribution X for this to hold true?'. This is called the *domain of attraction problem*, and we therefore say that X is in the *domain of attraction of G*, often denoted $\mathcal{D}(G)$. More details on this will not be provided here since it is deemed out of scope, but a full mathematical description can be found in Haan and Ferreira (2006).

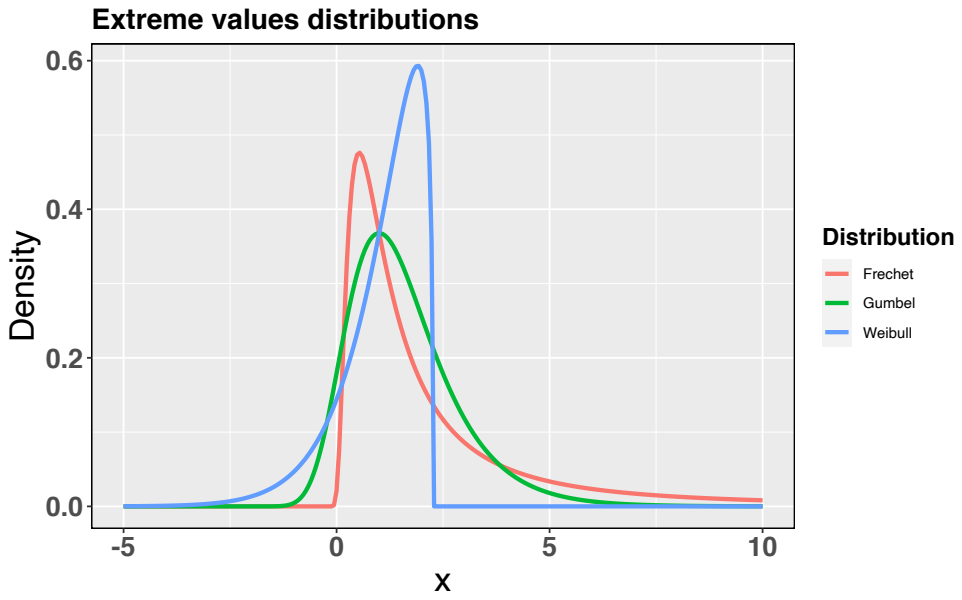


Figure 2.6: Density plots of the three possible extreme value distributions.

Since the key quantity in extreme value statistics is the extreme value index ξ, the main aim is usually to try and accurately estimate this. There are two main methods for doing this, the block maxima (BM) where one decides a block size (e.g. a year) and extracts the maximum from this, or peaks-over-threshold (POT) where we select all observations above a high thresholds. Both of these methods involve choosing a suitable value which is large enough to make sure that all the sample points are extreme but low enough to ensure that enough sample points are chosen. For the BM method, if one chooses a year as the block length but only have 5 years of data, only 5 data points will be used to estimate ξ, leading to a very uncertain estimate. If we instead choose a month as block length, non extreme values will be selected if there is a strong seasonality in the variable, such as temperatures in Europe or rainfall over west Africa. If we instead use the POT model, we need to decide a threshold with the same limitations. This is the root of the bias-vs-variance trade-off, many sample points will mean that some are not extreme and therefore bias the estimate, but very few will lead to a large variance and therefore an uncertain estimate. Figure 2.7 shows these two methods where the red dots are the selected sample points with the BM method and the blue and red in the POT method. The horizontal line marks the threshold for the POT method and the vertical lines the blocks for the BM.

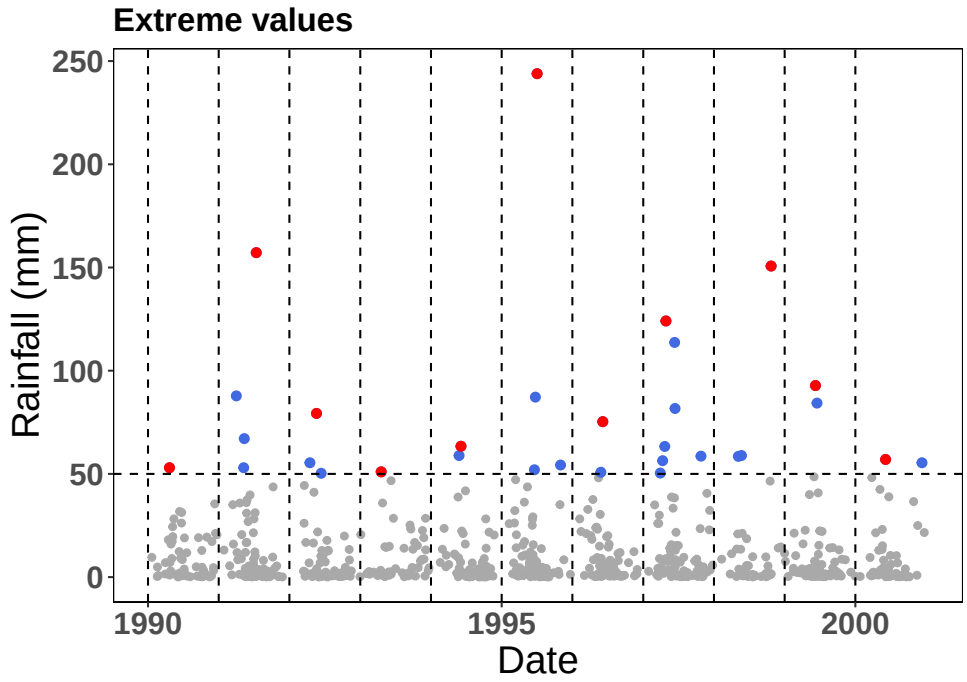


Figure 2.7: The sample points included in the extreme value analysis using the BM method (red points) and POT (blue and red points). Horizontal dashed line marks the threshold for the POT method and the vertical dashed lines the year breaks for the Block maxima.

To decide the block length or threshold, a combination of knowledge about the data and qualitative tools can be used. The block length will usually be decided based on a natural time sectioning that matches when one would expect the most extreme observations to occur. If one is modelling cold temperatures in Europe, then the best block would be the minimum temperature in October-March rather than the annual minima to capture the seasonal minima. To decide a suitable threshold in the POT model, stability plots can be used, alongside a sensibility check. One useful stability graph is to plot the estimate of ξ as a function of the k number of top order statistics included. This can be used since we expect ξ to be constant for all extreme values, since they belong to the same extreme value distribution $G(x)$, but become biased as we include non extreme values. Therefore by selecting a k, which essentially is the same as picking a threshold equal to the sample point $X_{n-k,n}$, for which the estimate of ξ is constant for smaller values of k but different for larger, we can find the balance in the bias-variance trade-off. This is visualised in Figure 2.8 where the horizontal line is the value of ξ that the estimate is stable around and the vertical dashed line indicates the optimal k value.

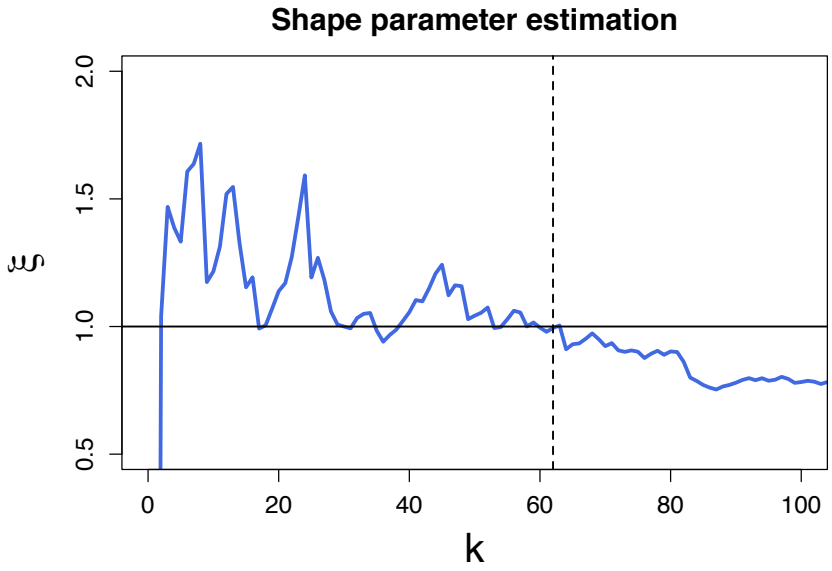


Figure 2.8: Estimate of the shape parameter ξ as a function of the number of upper order statistics k, including the top 10% values. The horizontal solid line is the correct value and the vertical dashed line the optimal k value.

To estimate ξ, a whole range of estimators are available with different advantages and drawbacks such as smaller bias, minimum variance or ability to estimate ξ for one or several of the value ranges (positive or negative values). The most classical estimator is the Hill estimator (Hill (1975)), which has a smaller variance than many others but can only estimate positive values of ξ and therefore only works for Fréchet/Pareto type distributions. Another positive thing with the Hill estimator is that it is a consistent and asymptotically normal estimator, meaning that it always converges to the true value for a large enough sample ($n \to \infty$) and it is possible to get a confidence interval on the estimate. It is however often biased due to a rather slow convergence, which means that a very large sample is needed for there to be enough extreme sample points to get an accurate estimate. Because an estimator can be unbiased in theory but not necessarily in practice, theoretical derivation of consistency and asymptotic distribution are often complemented with finite samples simulations to investigate this discrepancy.

2.2.2 Multivariate statistics

In the previous section we assumed all sample points to be i.i.d, which often is the case if they all come from the same weather station. However in many application we are interested in the interaction or dependence between either two different variables or locations. This information can for example be used to estimate the joint probability of two variables being large, e.g. rain and temperature, and thereby calculate their combined risk. It can also be used to estimate

the dependence as a function of distance and thereby understand both the area that might be experiencing extreme events at the same time, and the distance required for two locations to be independent. The second application can be very useful if only short time series exists for each weather station but there are several weather stations in a region with a similar climate, since one can then pool the independent stations together and obtain a significantly larger i.i.d sample size.

It is however not as straight forward in the multivariate setting to define and rank extreme events. Is it an extreme event only if all of the variables are extreme at the same time, or is it enough if one is extreme? And how can we rank points when only one variable is extreme?

Because the different variables might be on different scales, and to simplify the joint estimation, modelling multivariate extremes consists of two parts; the marginal distribution and the dependence structure. By first transforming the marginal distributions to be equal, we remove any potential influence from them being on different scales and can therefore estimate the true dependence between the variables. The exact choice of marginal distribution is not important from a theoretical point of view, since it is just a method for standardising the variables (Beirlant et al. (2004)). The most common choice is to standardise to unit Fréchet marginals because it leads to simple expressions to work with in the dependence estimation. Another common choice is the unit Pareto distribution, which as mentioned in the previous section also has a positive extreme value index and therefore has a similar tail behaviour to the Fréchet distribution but a different distribution function. Even though the choice of marginal distributions should not have an impact on a theoretical level, it often has an impact in practice since we need to replace the actual unknown distribution function with an empirical estimate (Section 5.3 for more details).

In contrast to the univariate case, there does not exist a finite number of parametric distribution functions that we can fit our sample points to, but instead a number of different parametric and non-parametric models for describing the dependence structure. This gives more flexibility in choosing the type of method one wants to use, but also makes it significantly more complex since we need to know which of these different method is the most suitable in our particular case. An important distinction is the one of asymptotic independence and asymptotic dependence (Sibuya (1960)), which essentially determines if the two variables are extreme at the same time. The asymptotic comes from the fact that everything in EVT is only defined in the limit $n \to \infty$, hence can never be fully but only asymptotically true.

A useful way to model the dependence structure between two variables while ignoring their marginal distributions is through copulas. A copula is a joint distribution function of two or more variables where each marginal distribution is uniform on the interval $[0, 1]$. This is based on Sklar's theorem (Sklar (1959)) which stats that if H is a 2-dimensional distribution function with marginal distribution function F, G, then there exists a copula $\mathcal{C}$ such that

$$H(x, y) = \mathcal{C}\left(F(x), G(y)\right), \quad (x, y) \in \mathbb{R}^2$$

The above stems from the fact that if X follows a distribution F then $F(X)$ is uniformly distributed on the interval $[0, 1]$. The joint behaviour is therefore completely determined by the copula function $\mathcal{C}$, of which there exists a huge variety with different behaviours (Nelsen (2006)). The main feature of a copula is how dependent, or correlated, the two variables X, Y are with each other. The dependence is controlled by the type of copula function, but also on a dependence parameter, θ, which all copula functions includes (see Section 5.5 for examples). In Figure 2.9 three different copula functions are demonstrated with different values on the dependence parameter θ. The possible values on θ for the three models are: bivariate normal $-1 \leq \theta \leq 1$, Frank $\theta > 0$ and FGM $-1 \leq \theta \leq 1$. We can see that there is a big difference in how correlated the points are for the different copulas even for θ close to their limit values.

What is more difficult to see in these graphs, is that these copula functions has two 'types' of dependence/correlation, one in the main part of the distribution and one in the tail. A prime example is the Bivariate normal distribution which can be nearly completely correlated in the main part of the distribution (Figure 2.9 top row) but is in fact asymptotically *independent* in the tail. This means that for sample points that are close to 1 (the maximum value) in one variable, is not for sure going to be close to 1 in the second variable, and therefore be located in the top right corner.

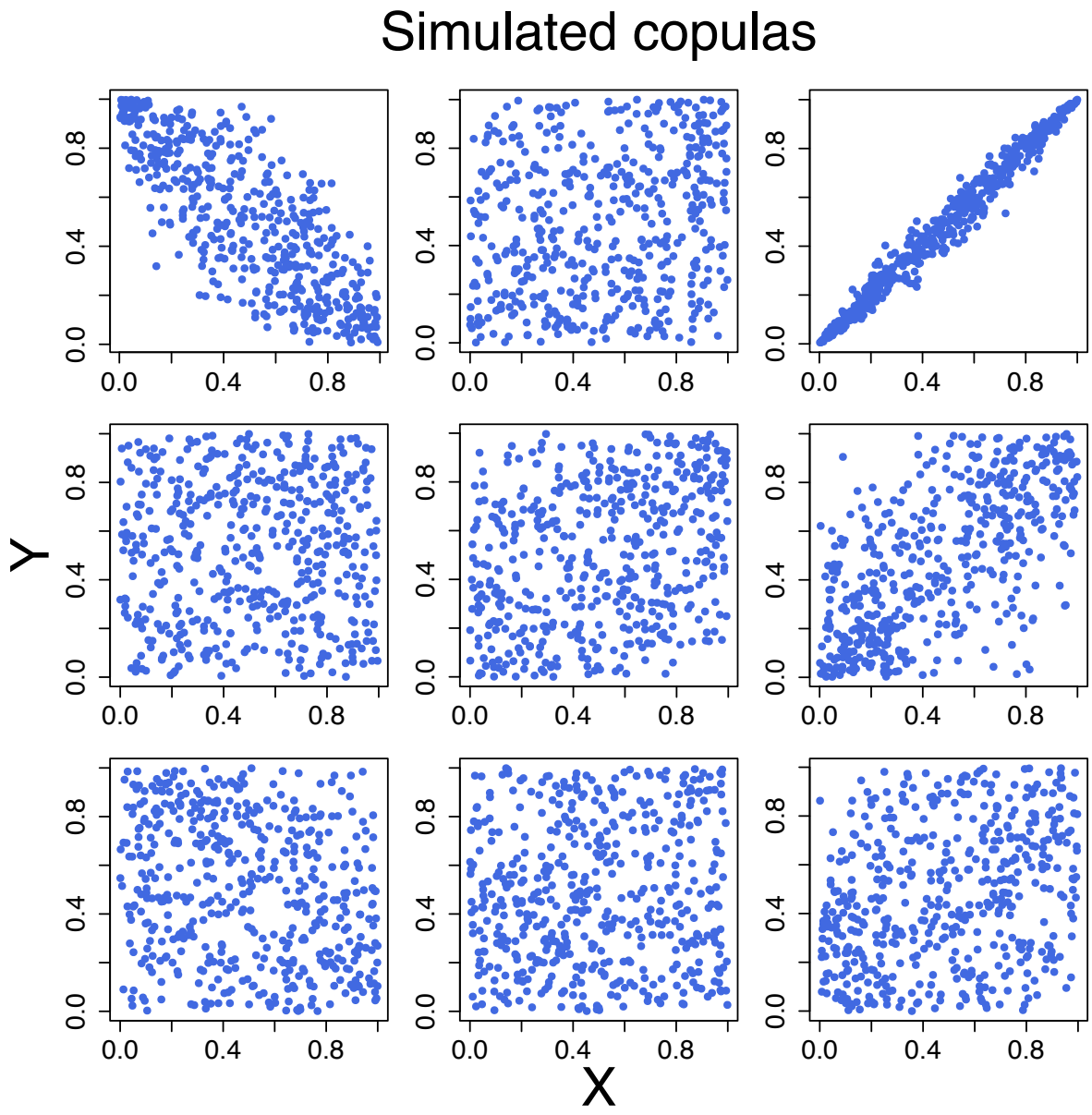


Figure 2.9: Simulated samples from three different copulas and three different values of the dependence coefficient of each. (Top) Bivariate normal, (middle) Frank and (bottom) FGM copula. (Left) θ=(-0.8, 0.2, -0.8), (middle) θ=(0.2, 0.2, 2) and (right) θ = (0.99, 0.99, 5).

Similar to univariate statistics where we needed one set of methods and distribution functions for the main part of the distribution and others for the extremes, the same is true for the two dependence areas. The dependence in the main part is often decided by various measures such as the Pearson correlation which require many sample points and considers the full sample. For the tail dependence, the common measures are instead *the extremal dependence measure*, $\chi(x)$ (Coles et al. (1999)), in case of asymptotic dependence and *coefficient of tail dependence*, $\eta(x)$ (Ledford and Tawn (1996)), for asymptotic independence. This is a clear

example of the added complexity in multivariate EVT where one needs to decide which of the two dependence regimes is considered since the estimators are different. In this thesis, only the asymptotic independence case is consider and the reason for this is outlined in Section 5.1 and 6.1 where all the theoretical definitions are also introduced.

Chapter 3

The spatial correlation structure of rainfall at the local scale over southern Ghana

In this chapter, the two first questions from the thesis aim will be addressed by developing a non-parametric estimation method for the correlation distance (see Section 1.2.1), inspired by the method of comparing observed and expected probability of rainfall co-occurrence introduced in Ricciardulli and Sardeshmukh (2002). This therefore partly answers the first question. The developed method will be applied to different combinations of rainfall intensities to study the difference in range for these, thus addressing the second question.

The chapter also includes a detailed description of the Ghana rain gauge data set that is used to estimate the correlation range here, and later used in Chapter 4 and 6. The description provides information about the distribution of missing values, the seasonal cycle and variabilities in annual, monthly and daily amounts.

This chapter has been published in Journal of Hydrology: Regional studies (Israelsson et al. (2020)) and the supplementary material can be found in Appendix A. Since it is written to be read on its own, there is some overlap with Chapter 2.

Highlights

- Daily rainfall record since 1940 from 590 stations.
- A simple and easy to interpret method for estimating correlation is constructed and implemented.
- Decorrelation range is shown to depend on the intensity of the rainfall event.
- Long-term rainfall climatology for all of Ghana.
- First study on the local correlation structure and anisotropic patterns.

3.1 Introduction

Rainfall over west Africa has received a lot of interest the past decades due to the limited possibility of irrigation for the many farmers depending on rain fed crops. In Ghana, 50% of the population depend on rain-fed crops (SRID Ministry of food and agriculture (2017)) and a large part of the country's energy come from hydropower from the lake Volta (Nyarko Kumi (2017)), which makes the hydrological cycle of great importance. Because of the sparse rain gauge network over most parts of Africa (Washington et al. (2004)), some research has been done on describing the rainfall distribution over time for a specific station and then extrapolating this knowledge to the surrounding region (e.g. Nicholson et al. (2000)). One problem with this approach is the highly variable weather over west Africa which makes it difficult to extrapolate knowledge outside a very small region, leading to large uncertainties. The majority of the west African rainfall comes from the west African monsoon which is controlled by the movement of the *Inter tropical convergence zone, (ITCZ)*, an area where the south-east and north-east trade winds meet and a belt of convective clouds is present due to this convergence and the high amount of energy from the sun. The movement of the ITCZ results in strong seasonality in rainfall over the year and the convective nature of the rainfall is one of the reasons for the high variability in both time and space on a daily scale.

There has also been a lot of research done using satellite rainfall estimates, calibrated against the sparse network of ground stations, or using reanalysis products which in general performs worse at finer scales in the tropics when compared to gauge measurements (Diro et al. (2009), Maidment et al. (2017) and references therein). To get a more realistic description of the spatial rainfall distribution, and to generate spatially accurate satellite rainfall estimates, several papers have modelled the spatial covariance over either all of Africa or a specific country or region. This has been done using satellite data (e.g. Funk et al. (2015b),

Smith et al. (2005)) and rain gauge data (Moron et al. (2007), Ricciardulli and Sardeshmukh (2002), Greatrex et al. (2014)). Both of these types of dataset have their individual issues when collected over Africa. Satellite products may not represent fine scale variability accurately (Maidment et al. (2017)), but rain gauge data on the other hand is usually very sparse which again leads to issues when modelling the small scale behaviour (Greatrex et al. (2014), Moron et al. (2007)). A general problem is the lack of long time series, making it difficult to describe the full interannual variability (Greatrex et al. (2014)). Exceptions exists, with some time series dating back to the 1880 (Nicholson et al. (2018)), but these are highly clustered in a few countries and with the highly variable nature of the African climate this does not provide much information for other countries.

Ricciardulli and Sardeshmukh (2002) used 3 years of cloud observation data transformed into a "Deep convection activity index", with a resolution of 0.35°$\times$ 0.7°covering the entire tropics to model the correlation distance. They only focused on modelling the decorrelation distance for deep convection clouds for all active months, hence not making a distinction between the different phases of the monsoon cycle. With this method, they were not able to capture rainfall events related to any processes other than deep convection (Young et al. (2014)). Both a method of estimating the distance until the correlation was less than $1/e$, and a method to estimate the distance at which the conditional probability of rainfall, given that it rains at the grid point, approaches the overall probability of rainfall was used. In south West Africa, the decorrelation distance was estimated to roughly 150km with the first method 180km with the second. Funk et al. (2015b) instead used 0.05°resolution 5-day cumulative cold cloud duration (CCD) data to estimate the decorrelation distance for each month seperatly. Their method instead involves to estimate the average correlation at 1.5°around the grid point and then calculate the decorrelation slope by assuming the correlation to be 1 at distance 0. From this slope, the distance at which the correlation should be 0 was estimated. This method results in decorrelation distances of 500-800km over south west Africa. The longer range than obtained in previous studies is likely to be due to the use of 5 day, rather than daily, values. One limitation of using CCD instead of gauge measurements is that decorrelation in CCD is not equivalent to the decorrelation range in rainfall. In contrast to Ricciardulli and Sardeshmukh (2002) and Funk et al. (2015b), Moron et al. (2007) calculated the decorrelation distance between measured rainfall at stations instead of satellite grid points. This was calculated for five different tropical regions, to assess the generality of the results, on amount and occurrence data separately by estimating the Pearson's correlation for amount data and phi correlation for occurrence. One major limitation in this paper is the small number of stations for each region (9, 11, 13, 28, 81) which results in very wide distance bins (100km) and only a few station pairs in each bin.

Another method for estimating the correlation distance is to derive a variogram. A variogram describes the variance structure between locations depending on the distance between them. In both Greatrex et al. (2014) and Teo and Grimes (2007), climatology variograms are calculated on rain gauge data to estimate the range, which is the correlation parameter, of rainfall over Ethiopia and Gambia. Both of these papers split the analysis for occurrence and positive rainfall amounts, similar to Moron et al. (2007). This is because the dependence structure for occurrence and amount are showed to not necessarily be equal, since they come from two different processes. The occurrence process only models the distances over which rainfall occurs simultaneously whereas the amount process considers how similar these rainfall values are. Furthermore, the total rainfall amount over some period is not entirely dependent on the frequency of rainfall event. Teo and Grimes (2007) face the same limitation as Moron et al. (2007) with only 20 stations, however distributed on a small area, resulting in a relatively dense network. The range for occurrence and positive rainfall amounts are 50km and 150km, hence substantially shorter than the once estimated by Funk et al. (2015b) but similar to Ricciardulli and Sardeshmukh (2002). Greatrex et al. (2014) has a much larger dataset of 276 stations but is limited to only 5 years of data and the stations are very unevenly distributed over the country with a complex topography. Many of the variograms in the paper do not have a clearly defined range and thus a clear correlation distance is difficult to determine. The difference in correlation distance can partly be explained by the use of different methods and data types but some stems from the use of different countries, since the decorrolation range varies greatly across Africa (Funk et al. (2015b)).

Two common assumptions are that the rainfall distribution will be equal for all rainfall events and the distribution is equal in all directions. But in the recent paper of Maranan et al. (2018) it was showed that even though the vast majority of the annual amount of rainfall comes from Mesoscale convective systems (MCSs), but the events classified as moderate and strong convection has the highest frequency of events. This implies that these events can not be considered to be equal, since the less frequently occurring events generate more total rainfall. Many rainfall processes are moreover anisotropic, especially at the daily scale. There has been some work done on anisotropy in Africa, but this has either just been done on very small areas (Gyasi-Agyei and Pegram (2014), Ali et al. (2003)) or using covariates to remove the spatial variability (Laux et al. (2009)).

By describing the covariance structure in daily rainfall at the small to moderate scale (10-150km), the results in this paper will help to fill in the knowledge gap currently existing between the station level rainfall distributions and the large scale behaviour ($\sim$ 400km). This analysis

is made possible by a completely new and unique dataset from the Ghana Met Agency comprising 590 stations with daily rainfall measurements. Ghana is chosen as our study region due to this unique dataset in combination with its varying rainfall behaviour both in time and space due to the ITCZ. We will be using the method of conditional probabilities from Ricciardulli and Sardeshmukh (2002) because of the easily interpretable results and the possibility to establish a rainfall reference probability. The results will also provide a better understanding of the spatial behaviour of different intensities of rainfall events over west Africa by modelling their dependence structures separately. The final contribution is an anisotropic description of rainfall, showing the impact of large scale drivers on the local scale.

The remainder of the paper is organised as follows: An introduction of the study area, the dataset and the methods used to model the co-occurrence will be presented in Section 3.2, results on the rainfall climatology and the spatial distribution will be given in Section 3.3 and the paper will end with a discussion in Section 3.4.

3.2 Data and methodology

3.2.1 Study area

Ghana is located in the Guinea coast with boarders to Burkina Faso, Côte d'Ivoire and Togo (Figure 3.1) and is approximately 650km long and 350km wide with a 560 km long coastline. It has five distinct geographical areas: low plains in the south, the Volta Basin in the centre with the artificial lake 'Lake Volta', the Akwapim-Togo ranges to the east of the Volta Basin with many heights and folded strata, the Ashanti Uplands to the west and high plains in the north (Boateng et al. (2018)). The temperature peaks around February-March and is at its lowest around August. The rainfall is mainly associated with the west African monsoon, which is controlled by the movement of the ITCZ. The country is under the influence of the tropical maritime air mass from March to October, during which the rainy season occurs. South of 8°N, there are two rainy seasons with a short dryer period in August. North of 8°N there is only one long rainy season (see Figure 3.7). From November to February/March the country is affected by the prevailing southward winds, called the Harmattan, which brings dry and dusty air from the Sahara and gives rise to the dry season. The majority of the rainfall is generated by convective clouds with a higher contributing proportion in the north. The coast experiences, aside from convective rainfall, warm rain processes and advective rainfall from the Atlantic ocean.

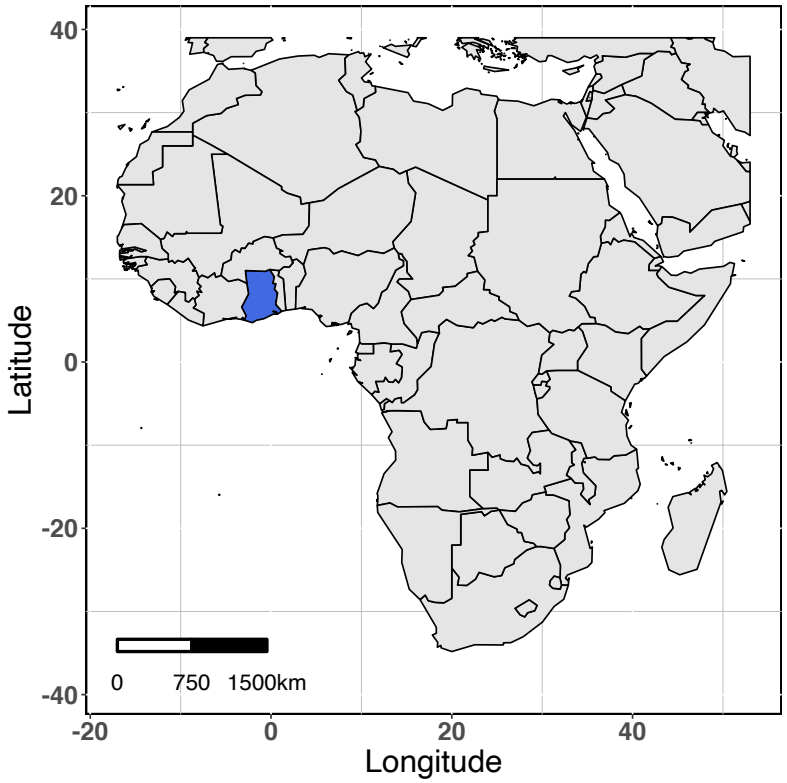


Figure 3.1: Map highlighting the location of Ghana on the African continent. Country boundaries defined by ISO 3166.

3.2.2 The dataset

The daily rainfall dataset used in this report is provided by the GMet (Ghana Meteorological Agency) and consists of daily rainfall amounts recorded at 590 stations covering all of Ghana, with a much higher density of stations in the southern half of the country. Extensive quality control was performed by the manuscript authors, along with a team of experts from the GMet. The dataset was assessed for errors in station locations, location shifts over time (over coastal data), erroneous data, the relationship with neighbouring stations and erroneous statistics and outliers. Any data flagged in this process was then checked against the original written records and other sources such as Google Earth Imagery for locations. In the case of data that was clearly erroneous, the station (or a subset) was removed. The original dataset consisted of 598 stations and 17'008'530 individual station-day data points. The Quality Control process led to a reduction of 1.55% of available data, with the controlled analysis using 590 stations and 16'744'082 individual data points. The dataset spans from 1940 until the end of 2017, with all the time series containing some missing values, but several of them only have few missing values in the period 1950-2017.

Figure 3.2 shows the number of stations for each month with less than 10% missing values, which we will refer to as valid stations. Figure 3.4 shows the number of valid stations, i.e. stations with less than 10% missing values, for each proportion of valid months in the dataset. There is a significant increase in the number of valid stations from 1950 (Figure 3.2) and then a steep decrease during the 80's, similar to the station pattern found in the datasets used in Nicholson et al. (2018). The reason for the large fluctuation between the years 2000-2010 are not known to us. Figure 3.3 shows the distribution of the median number of daily reporting stations during the year with the most available stations (1976) and in 2017. The station density has decreased coherently across the whole country, however the very sparse network in the north even during 1976 results in extremely few current stations. Due to the large increase in the number of valid stations in the 50's, our statistical analysis can be improved by only including data in the period 1950-2017, which still leaves us longer records than most other studies and includes data both before and after the Sahelian drought in the 70's and early 80's (Brooks (2004)).

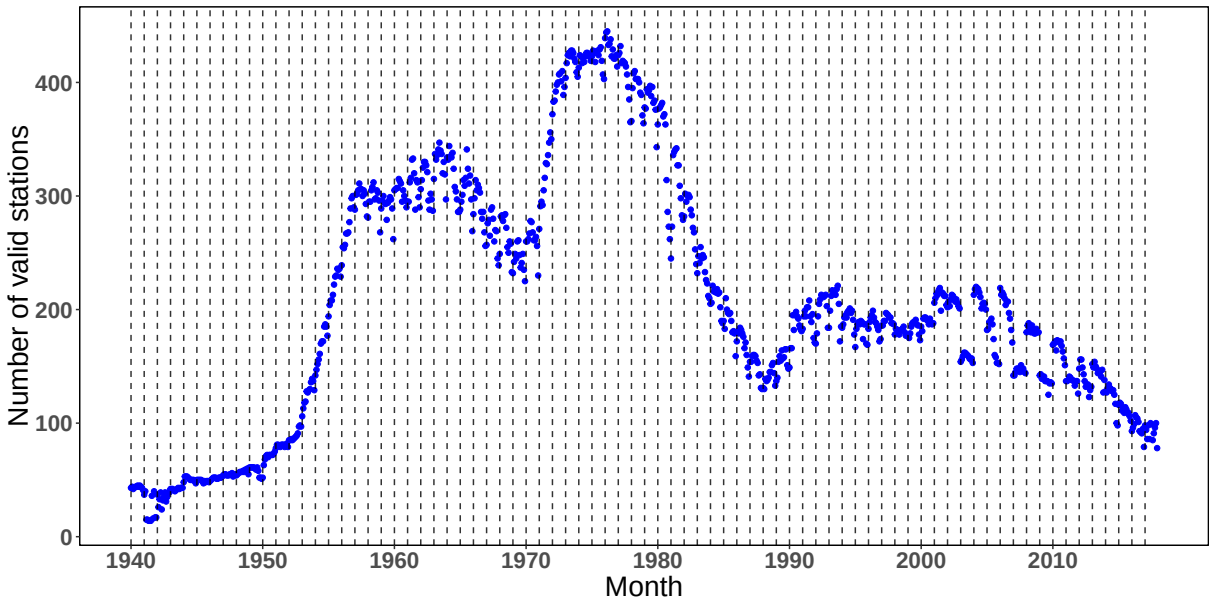
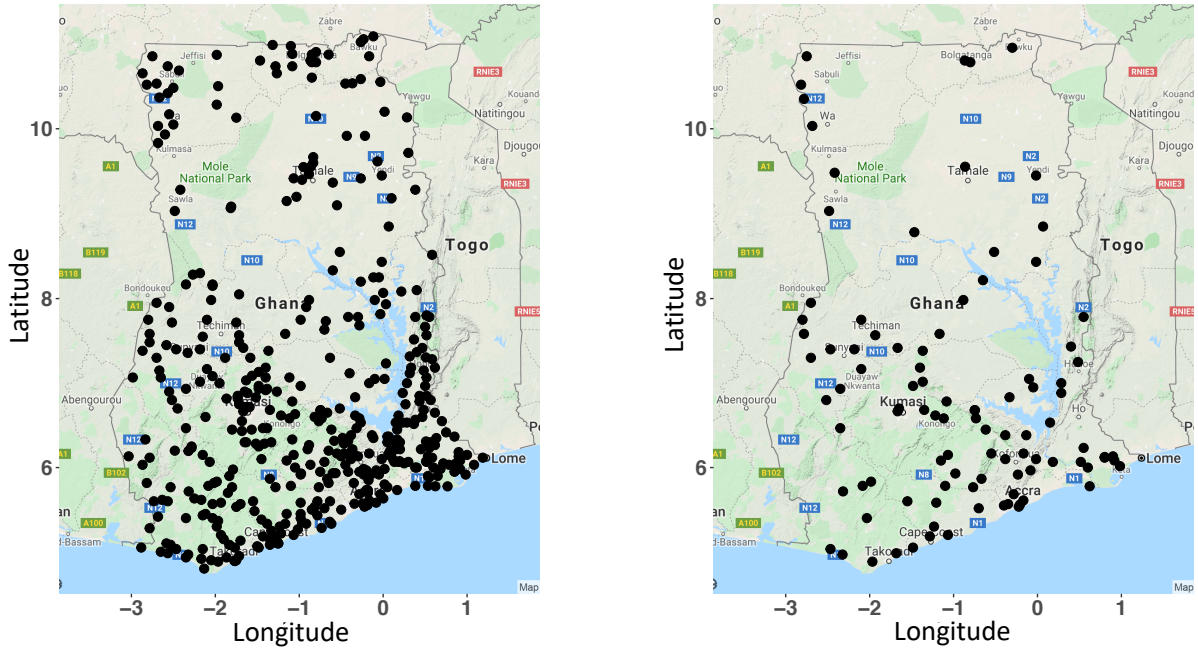


Figure 3.2: Temporal evolution of the number of stations with less than 10% missing values per month. Each vertical line marks the beginning of a year. There are 590 stations in total.



(a) (b)

Figure 3.3: Maps of Ghana showing the median number of available stations in (a) 1976 (440) and (b) 2017 (100 stations). Maps from "Google maps" using the R package *ggmap*.

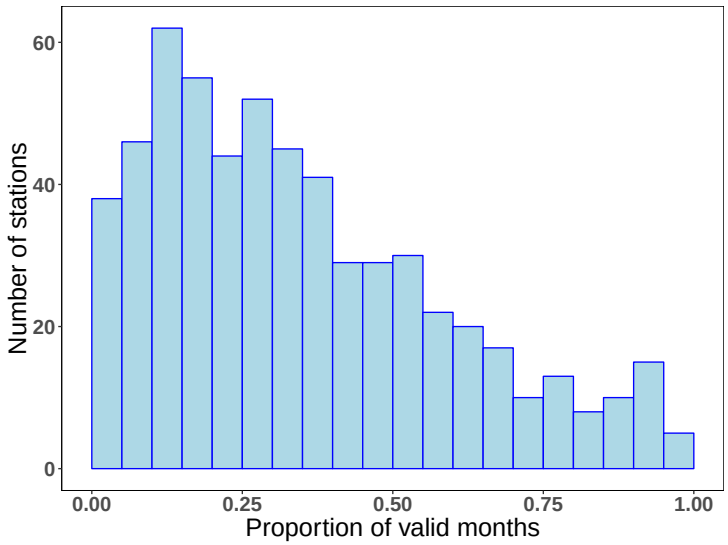


Figure 3.4: Absolute frequencies of stations against the proportion of valid months, i.e. months with less than 10% missing values. There are 590 stations and 936 months (Jan 1940-Dec 2017) in total.

3.2.3 Variability in daily to annual amounts

To provide a full picture of the rainfall climate over Ghana, made possible thanks to the dense rain gauge dataset, the variability in rainfall amounts on a daily to annual scale will also be presented in this paper. The different methods for doing this are presented below.

Because of the missing values, some adjustment must be made to the annual and monthly amount totals to compensate for the missing data points. Instead of using data fill methods, such as replacing missing values with the mean, median or most frequent measurement, the measured annual or monthly amounts are adjusted with a parameter proportional to the number of missing values in that time period. This scaling method is favoured over the fill method since we are not attempting to fill in the gaps, but rather to compensate for the expected lower amount total due to lower number of recorded days. For the monthly totals, each rainy day measurement is multiplied by ξ_m^{-1}, where ξ_m is the proportion of missing observations within that month. Similarly for the annual totals, the total amount is multiplied by ξ_y^{-1} where ξ_y is the proportion of missing observations within that year. Only individual years with less than 20% missing values are used in these graphs. This to not risk including years where the recorded days do not accurately represent the full year.

The Coefficient of Variation is defined as CV $= \frac{s_x}{\bar{x}}$, where s_x is the standard deviation of daily rainfall ≥ 1mm and $\bar{x}$ is the average rainfall over rainy days. This is calculated both for daily values and monthly aggregated values for each station with at least 23 years of data, only using the months outside of the dry season. Since the variability in the daily amounts is expected to be very different from the interannual monthly variability since longer accumulation periods usually reduces the noise, both of these will be estimated. For the estimation of the daily CV, values from all years are used, excluding missing values and amounts lower than 1mm. For the monthly CV estimation, only months with ≤ 3 missing values are used.

3.2.4 Spatial variability in the occurrence of rainfall of varying intensity

The analysis is done separately for each month to remove some of the variability due to the different phases of the monsoon, and only using every 5th day to work with independent events (see Figure A.2 in the supplementary materials for autocorrelation plots). This is done because we are interested in modelling the spatial dependence and not the dependence in time, such as lower rainfall amounts potentially follows a day with large rainfall amounts. To reduce the noise from large differences in absolute amounts between nearby stations, correlograms are estimated from occurrences within amount intervals instead of covariograms on the measured amounts.

Using the notation that "[" includes the value and ")" all values up to but not including the value, the amount intervals, hereafter denoted intensity classes, are defined as; $S_1 = [1, 10)$mm, $S_2 = [10, 30)$mm, $S_3 = [30, 50)$mm and $S_4 = [50, \infty)$mm representing low, moderate, heavy and very heavy rainfall. In order to study how the rainfall dependence changes with intensity of the convective system, we calculate within a specific distance the proportion of rain-rain occurrences, hereafter denoted co-occurrence, both within an intensity class and between an intensity class and stations in higher or lower intensity classes. That is, we will estimate the co-occurrence probability between only stations that are within the same intensity class, and in the setting with the origin station in one intensity class and surrounding stations in the same or higher/lower intensity classes.

Due to only using every 5th day, we end up with 408 independent time steps (68 years, 6 days per month except February) for each month. Only stations south of 8 °N with less than 50% missing values in the period 1950-2017 have been used which gives us 232 locations for our analysis. The reason for only using stations south of 8 °N is two-fold. Firstly, by excluding the northern region with a single rainy season, all the stations will be in the same rainfall regime (rainy or non-rainy). Secondly, the dataset is much more dense in the southern region which provides us with more robust estimations. 50% missing values is chosen as a trade-off between using stations with just a few years long record which might skew the results and discarding information. Since our method involves taking the average over a very large number of estimates of co-occurrences, we determined that stations with up to 50% of missing values will not negatively impact the results.

To model the spatial dependence structure of co-occurring rainfall events within an intensity class, the second method with conditional probabilities in Ricciardulli and Sardeshmukh (2002) was used. The full algorithms are found in Section 3.A and a summary of it is presented below. A schematic overview of the two algorithms can be found in Figure 3.16-3.18 in Section 3.B and the references in the brackets refer to these.

Algorithm 1 For each unique day, transform all amounts that are within our chosen intensity class to a 1 (green dot) and all other amounts to 0 (black dot). Choose one of the stations assigned a 1 to be the origin station (pink dot) and calculate the distance from this station to all other stations. Within each 10km distance bin (blue circles), calculate the proportion of stations assigned a 1. Calculate the proportion for all stations assigned a 1 in step 1 and repeat for each unique day (i.e. move the pink dot to each and every green dot for each unique day).

By calculating the average in each distance bin, we can get a climatological average on the probability of observing rainfall of the same intensity as the origin station for a given distance.

To model the dependence structure of co-occurring rainfall events between an intensity class and either lower or higher intensity classes, the above method is used with a few changes. Transform all amounts that are within our chosen intensity class to a 1 (green dots), all stations with a measured amount in all lower or higher intensity classes with a 2 (orange dots) and all other stations to 0 (black dots). Then calculate the proportion of 1's and 2's instead of just 1's. The rest of the algorithm is identical. Just as for measurements within an intensity class, taking the average in each distance bin, we can get a climatological average on the probability of observing rainfall of a lower or higher as the origin station for a given distance.

To compare our co-occurrence probabilities with the climatology background state, a 2-step sampling method is used. The climatology background is the by pure chance probability of observing co-occurrence due to the current stage of the monsoon. Since there in June are much more rainy days compared to April, the chance of observing co-occurrence of a certain intensity is higher regardless if there is any dependence between the two stations. The aim with algorithm 2 is therefore to 'break' the observed dependence structure without altering the rainfall distribution for each stations. By estimating this background probability we can both understand how this varies over the season for the different intensity classes and obtain a data-informed 'null hypothesis' value. A summary of the algorithm for calculating the background state within intensities is given below. The references in brackets still refers to the schematic Figure 3.18 in Section 3.B.

Algorithm 2

1. For each unique day, calculate the proportion of rainy stations (blue dots).
2. Just as in algorithm 1, select one station with a measured amount in the correct intensity band (pink dot).
3. Randomly assign rain (blue squares) or no rain (black dots) to all other station so the proportion equals the measured proportion in step 1.
4. Randomly assign a measured rainfall amount ($\geq$ 1mm) from that station from that unique month (eg. May 1980).
5. Assign a 1 to all stations in the chosen intensity class (green dots).
6. Calculate the proportion of 1's in each 10km distance bin (blue circles).

7. Repeat this for equally many times as there are stations with observations in the correct intensity band.

The last step is to ensure that there are equally many co-occurrence estimates in the observed and the climatology estimate. For the calculation between intensity classes apply the same modification to step 5 as described for Algorithm 1 (1's and 2's).

3.2.5 Anisotropy in spatial rainfall variability

The final property we will consider to fully describe the rainfall structure over Ghana, is the potential anisotropic structure in rainfall. Anisotropy means that the behaviour of the process is different depending on the direction. Here we are specifically interested in studying if the covariance of rainfall is different depending on direction. On a monthly timescales it is clear that large scale drivers such as the ITCZ will be visible, with a higher covariance in the E-W direction compared to the N-S. This has however not been explored on a sub-weekly level and on smaller scales (<100km) because of data availability.

An initial analysis of this small scale, short time scale spatial variability in covariance will be performed through the use of covariogram maps. This will be estimated on 2-, 3- and 5-day rainy aggregated amounts from all intensities, because we here are interested in how similar or different the rainfall amounts are in different directions. The reason for aggregating the amounts to a few days is to reduce some of the noise while still estimating it on a short time scale. The concept of covariogram maps will be explained by first introducing *semivariograms* and then how these are related to covariograms (see Section 2 for even more details).

In our setting, let $Z(s_i)$ represent the k-day aggregated rainfall amount at station s_i. Assuming that the mean and variance of $Z(s)$ are finite, the semivariogram is defined as the half mean squared difference

$$\gamma(s_i, s_j) = \frac{1}{2} E\left[(Z(s_i) - E(s_i)) - (Z(s_j) - E(s_j))\right]^2 \qquad i, j = 1, 2, ..., n$$

Here we are going to assume that the underlying process $Z(s)_{s \geq 0}$ generating the k-day aggregated rainfall amount is intrinsically stationary and isotropic. This simply means that the mean is constant, i.e. $E(Z(s)) = \mu$, and the semivariogram only depends on the distance between 2 locations and not the direction. Hence

$$\gamma(h) = \frac{1}{2} E\left[Z(s+h) - Z(h)\right]^2 \tag{3.1}$$

A natural method to estimate the semivariogram is by the *method-of-moments* semivariogram which essentially stems from replacing theoretical expectations with the analogous sample averages. The corresponding sample semivariogram to Equation (3.1), notably the Matheron's classical estimator (Matheron and Blondel (1962)), is defined as

$$\hat{\gamma}(h \pm \delta) = \frac{1}{2|N(h \pm \delta)|} \sum_{(i,j) \in N(h \pm \delta)} |z_i - z_j|^2 \tag{3.2}$$

where z_i is the observed k-day aggregated rainfall amount at station s_i, $N(h \pm \delta)$ is the set of pairs within the spatial lag $h \pm \delta$ and $|N(h \pm \delta)|$ is the number of pairs in that distance range.

The general shape of semivariograms can characterise and explain the dependence structure in terms of its three main indicators: *nugget, range* and *sill*. The nugget is the variance at the close-to-0 distance representing the variability at distances of a couple of meters and measurement error. The nugget is expected to be small in rainfall data because of very high correlation at small distances. The sill is the value that the semivariogram converges to as the distance increases (dependence decreases) and the range is the distance at which two stations are independent and the sill is reached (or 95% of it if only approached asymptotically) (Cressie and Wikle (2011)). The maximum distance is set to 160km since we are interested in modelling the small scale behaviour and due to the small area over which this is estimated.

The estimated semivariance values will be used to calculate the *covariogram*, defined as

$$C(h) = \sigma^2 - \gamma(h) \tag{3.3}$$

where σ^2 is the 0 distance variance given by the sill. Because the variance is linear in the number of aggregated days, $C(h)$ is divided by σ^2 to enable us to compare the results from the different aggregation periods. By using the covariogram instead of semivariogram, we can again compare our estimated values against the theoretical convergence value 0, which is reached when there is no dependence left.

For quantities with spatial dependence varying with direction, an anisotropic model must be applied instead of the isotropic model described in Equation (3.2). Isotropic models, i.e. models only depending on distance and not direction, can be turned into anisotropic models by replacing the distance parameter h with a distance vector $\mathbf{h}$, which then will be associated with both a length and a direction. The bin $\mathbf{h} \pm \boldsymbol{\delta}$ now represents both a distance range and an angular tolerance, e.g. all stations in a 45°segment.

Given that we have a dense station network, we can explore how $C(h)$ varies in different directions by estimating a covariogram map. A covariogram map is a lattice where each square represents a distance and an angle and is symmetric because of the square term in Equation (3.2). To construct a covariogram map, select one of the stations and place a square lattice over the region with your chosen station in the centre. Calculate the difference in k-day aggregated amount between your selected station and all other stations. Calculate the average within each square of the lattice. Repeat this for all stations and all k-day periods. After taking the average from all the individual maps, the resulting map describes the mean behaviour in all directions as we move away from a station, hence displaying directions with a stronger correlation.

A minimum threshold of 1000 pairs for each square is used to make the estimation more robust.

3.3 Results

3.3.1 Climatology of rainfall in Ghana

To describe the climatology of the rainfall in each agro-ecological zone shown in Figure 3.5 and defined by GMet (Owusu and Waylen (2009)), the four stations marked with red dots are used because they have the least number of missing values within each zone. Annual total amount time series, Box and whisker plots over the monthly total amounts, maps of key rainfall estimates and maps of Coefficient of Variation (CV) for each month not in the dry season are used to demonstrate this. A day is defined as rainy if the measured amount is ≥ 1mm as used by *Expert Team on Climate Change Detection and Indices, ETCCDI* and in several other papers (e.g. Moron et al. (2007) and Sillmann et al. (2013)). 30mm is chosen as the threshold for heavy rainfall because this equals the 10-20% heaviest amounts on rainy days for most of the country.

Combining the information in Figure 3.6 and 3.7, one can clearly see the reason for the partition, with the dry Coast region, the Forest region with a high proportion of rainy days, the Transition region with much fewer rainy days but still a bimodal rainy season and the North with only one rainy season. In some papers, the South-Western coastal region is classified as a separate region, which Figure 3.6c confirms with the much higher proportion of days with heavy rainfall in that area. One can notice a diagonal band of higher proportion of rainy days from the SW coast up to Lake Volta, with a significantly drier region along the coast. This was noted already in Acheampong (1982) and was explained by the complex atmospheric interaction in that region, stemming from several components.

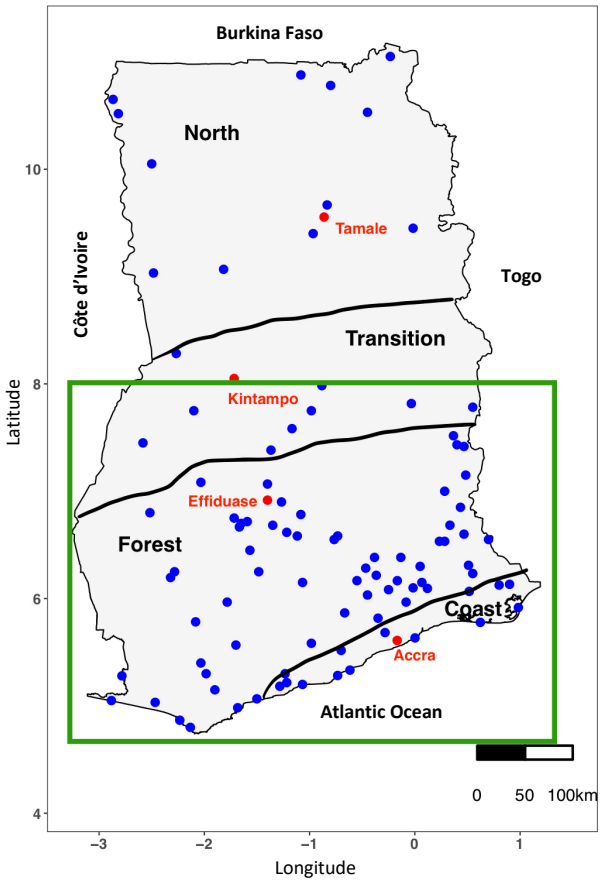
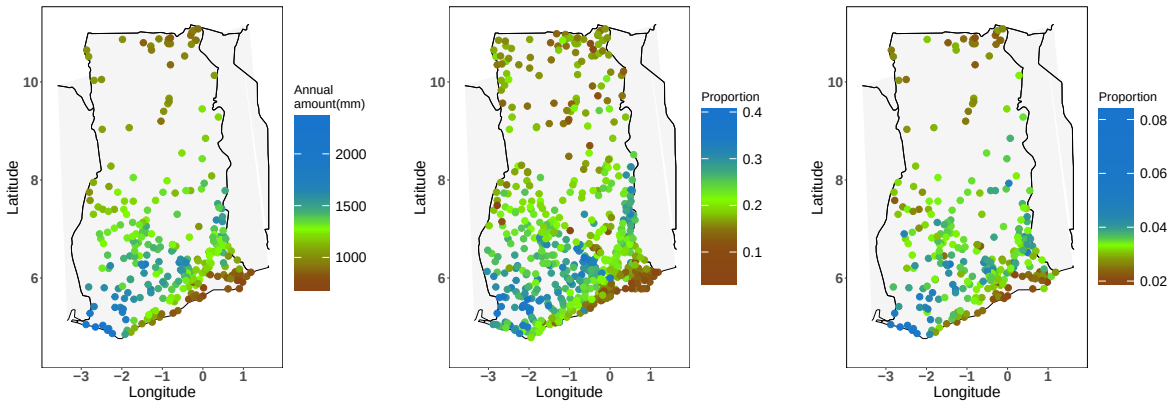


Figure 3.5: Map of Ghana showing the four agro-ecological zones defined by GMet (Owusu and Waylen (2009)) and the location of the 100 stations with the least number of missing values. The red stations are the stations used in the following climatology analysis and the region enclosed in the green box is used for the rest of the analysis.



(a) (b) (c)

Figure 3.6: Maps of Ghana. (a) average annual total rainfall amount (mm), (b) the distribution of proportion of rainy days (≥1mm) and (c) the proportion of heavy rainfall days (≥ 30mm). (a) and (c) only uses stations with at least 23 years of data. Note the different scales.

The first is that moist air from the Atlantic ocean are often coming from SW, hence reaching and releasing the rain on the south west coast instead of further East and travels parallel to the eastern coast line. A second component is the difference in latitude between the western and eastern part of the coast, resulting in convective systems generated by the ITCZ earlier and later in the season affecting the west part of the coast. We can also see the decreasing trend in rainfall as we move northward in Figure 3.6a, matching the movement of the ITCZ and a decreasing presence of rainfall coming in from the Atlantic ocean.

In Figure 3.7 we can clearly see the different rainfall modes, with an unimodal rainy season in the north zone and a bimodal in the rest of the country. This is because the ITCZ passes through the bimodal part of the country twice in a year, firstly as it moves northward in the early summer and secondly as it travels southward in September/October. The northern region is where it changes direction and therefore only passes over once. The difference between the bimodal regions is clearly visible, with both the major and the minor rainy season being of equal intensity in the Forest region whereas the Transition region has a slightly more intense minor season and the Coast has a much more intense major rainy season compared to the minor. The reason for the very intense major season compared to the minor in the Coast region is the movement of the ITCZ which brings a lot of moist air from the ocean as it propagates northward, contributing to many days with heavy rainfall.

A common feature for all regions is the large interannual variation in monthly and annual total amount. For all months during the rainy season, the difference between the whiskers are around 300mm and the mean rainfall is between 100-250mm. June in Accra, which is at the peak of the rainy season, has the largest range which is 450mm with a mean of 175mm. There is also a common pattern of a slow increase in mean rainfall up until the peak of the rainy seasons and then a quick decrease as the ITCZ retracts.

Studying the total annual rainfall in Figure 3.8 one can again see a very large interannual variation for all locations with the biggest spread in the Forest and Transition regions, both of which have two intense rainy seasons. Despite Tamale only experiencing one rainy season, the mean annual rainfall is higher than Accra, due to the longer rainy season. Accra has the lowest mean annual rainfall of the four stations, of around 750mm/year. Both Kintampo and Effiduase have mean annual rainfall of around 1400mm/year, however Effiduase has a varying pattern so the amount fluctuates over the studied time period. Accra has an interannual range of about 1000mm, Tamale 800mm and Kintampo and Effiduase of nearly 1500mm.

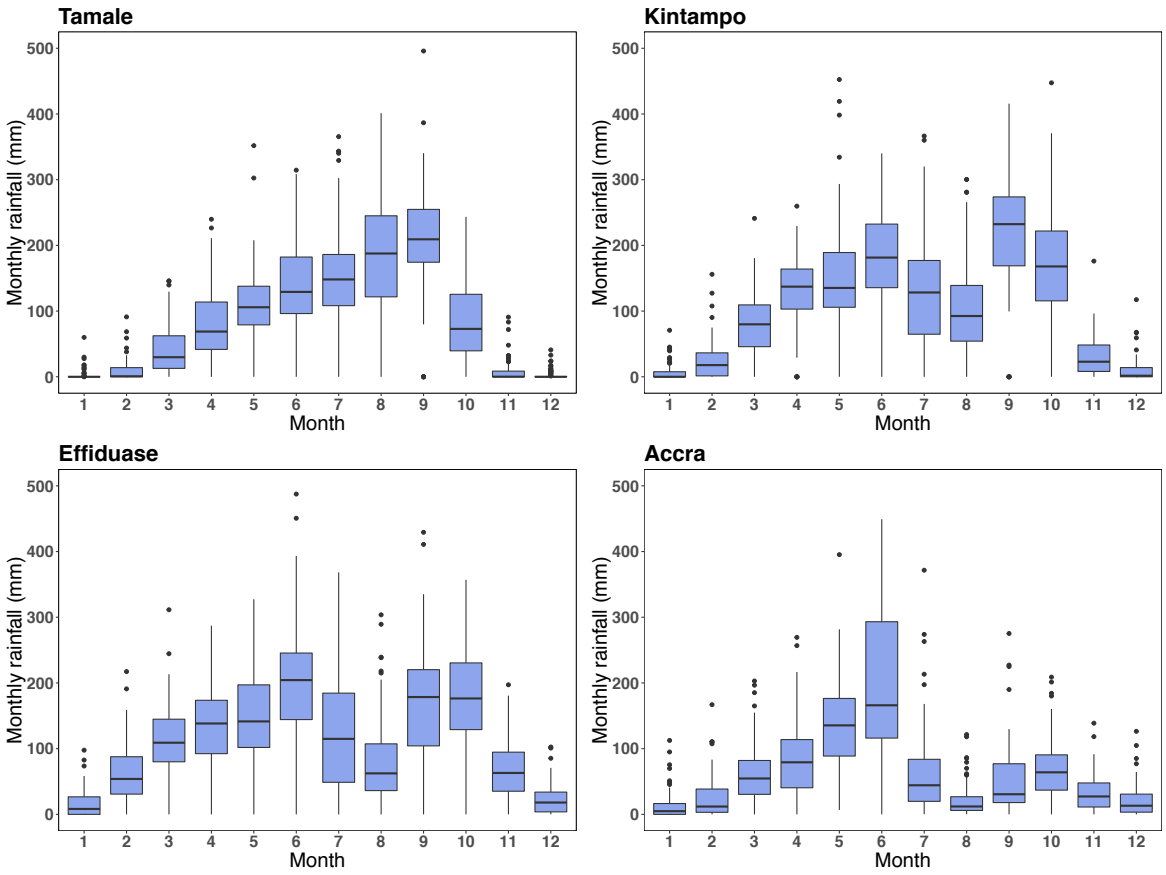


Figure 3.7: Box and whiskers plots showing the interannual variability of the total rainfall for each month. Location of all the stations are displayed in Figure 3.5 (red dots). Months with any missing values has been removed and the most extreme outliers are excluded in the graph to improve readability.

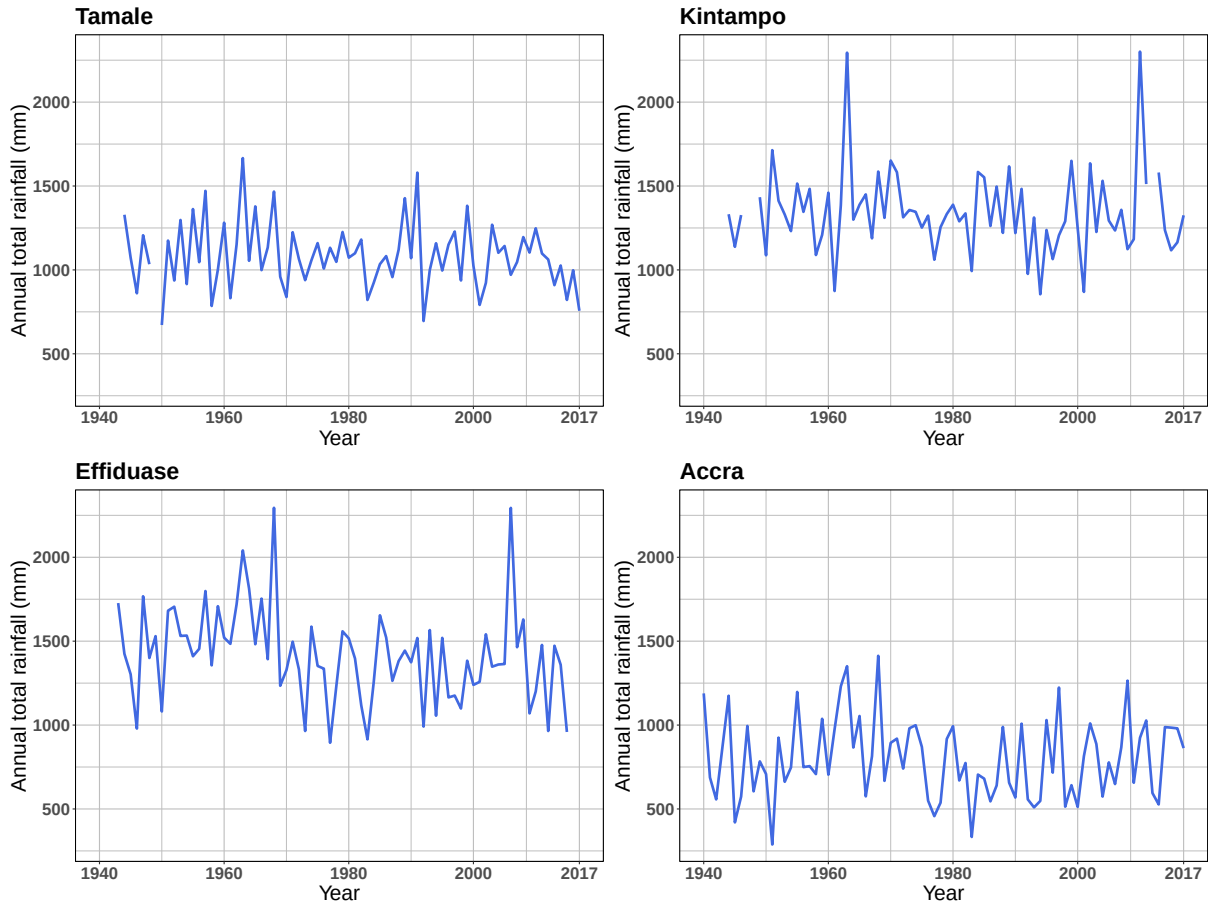


Figure 3.8: Time series over the full annual total amount for one station in each agro-ecological zone; Tamale - North, Kintampo - Transition, Effiduase - Forest and Accra - Coast. Gaps in the time series are years with more than 20% missing values.

In contrast to Owusu and Waylen (2009) but similar to Lacombe et al. (2012) and Torgbor et al. (2018), it does not appear like the Sahelian drought in the 70's and 80's had a big impact on the average annual amount in any of the regions but there is a strong decrease in the variability at Tamale and Kintampo during this period.

There is a distinct pattern of lower CV over monthly aggregated values (Figure 3.10) than daily values (Figure 3.9). This means that there is a much larger variability in the daily rainfall values compared to the interannual variability around the monthly mean. This is expected due to the convective nature of the rainfall resulting in short intense storms and that the noise reduces as the accumulation increases. We can also see a more coherent pattern over space in the monthly compared to the daily values. The monthly CV values are however still large with a mean value of 1 or higher for nearly all months, meaning that the variation around the long term mean is of the same magnitude or larger than the mean. This is similar to the results in Arvind et al. (2017) which looked at monthly values in a region of India but a lot higher than Ayanlade et al. (2018) which studied a region in Nigeria. The most northern part of Ghana is under the influence of the Harmattan in March and November as seen in Figure 3.7, hence the monthly average rainfall is very low which inflates the CV value. For the daily values, we can see a larger spread during the monsoon phase (May-October), with the exception of June which has one of the smallest spreads. This is because the higher mean in June lowers the CV estimate, even if the absolute variation is the same as May and July. There are no obvious spatial patterns outside the main monsoon season, but from June-September there is a clear pattern of highest values along the coast which decreases as we move north.

The monthly values exhibit a different pattern. Outside the main monsoon in the north (October-April), there is a bimodal pattern with high CV values in the north and along the coast and lower values inland. May and June is very non coherent, but still with high values along the coast line. During the peak of the north rainy season (July-September), there is a clear gradient with high values in the south and low values in the north. There is also a West-East gradient in the southern part for all months except July-October, with lower values in the West where we have higher annual rainfall and more rainy days.

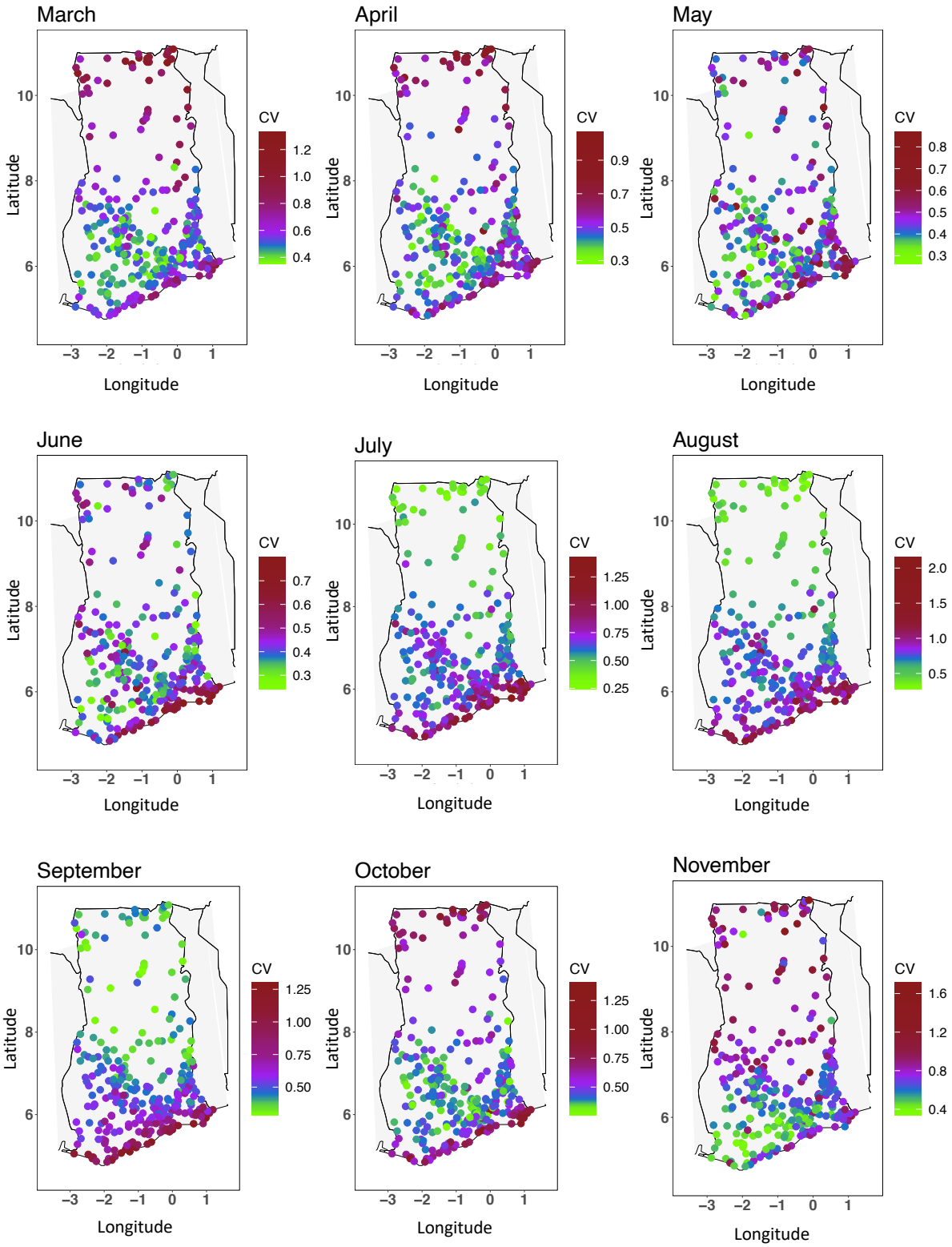


Figure 3.9: Maps of the distribution of Coefficient of Variation for daily values per month. Note the different scales on the scale bars.

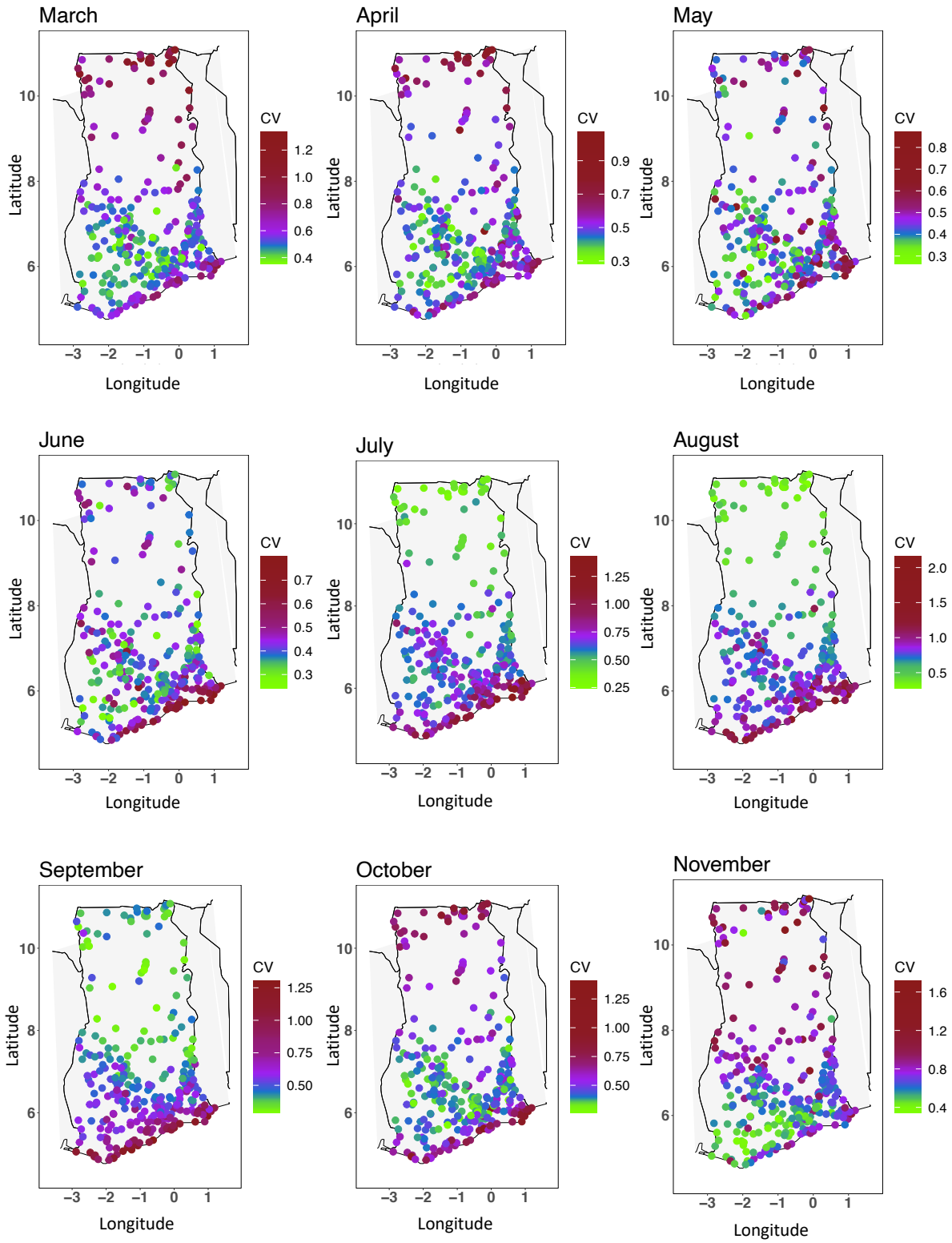


Figure 3.10: Maps of the distribution of Coefficient of Variation for monthly aggregated values per month. Note the different scales on the scale bars.

3.3.2 Spatial distribution of rainfall events

In the papers mentioned in the introduction, all rainfall events were assumed to have the same spatial variability. This must not necessarily be true since we might expect a moderately intense rainfall event to have a larger extent than a low intensity rainfall or an intense shower storm. In order to study this, the method of conditional probability curves will be applied to the intensity classes described in Section 3.2.4. By comparing our measured results with the climatology background (see Section 3.2.4), we can separate out the increased probability of rainfall at a station because of rainfall at nearby stations and the overall probability of rainfall due to the time of the year. This will both give information about the very local behaviour and the spatial extent for the different intensities. In the first section we will model the variability for increasingly larger intensity classes and the classes below, placing the highest intensity class in the origin, and in the second section we will model the variability within each intensity class.

Extent of rainfall events for different intensities

By modelling both the conditional probability of observing rainfall amounts that are lower (Figure 3.12) or higher (Figure A.1 in the Supplementary material) than the origin station, we can get information about both the extent of a large rainfall event and the probabilities of observing higher rainfall amounts close to low rainfall. By subtracting the reference rainfall probabilities from our measured co-occurrence probabilities, we can study how the anomalies changes over the season.

The results shown in Figure 3.12 suggest that low rainfall events are localised whilst heavy rainfall events have a larger spatial structure. In Figure 3.12 (a) and (c) we can see that the 5km line very closely follows the baseline, showing the small scale behaviour depending on the season. This would indicate that low rainfall events are very local, resulting in the co-occurrence probability depending on the overall probability of rain. The peak in August in (a) is most likely due to the significant difference in the rain intensity distribution in Accra. For all months except July and August, the histogram over rainfall amounts (not included) are very similar for the three southerly stations used in Section 3.3.1. However in July and August, there is much higher probability of low intensity rainfall in Accra, and since there is a high station density along the coast the results from that region has a higher weighting compared to further north. This strong connection with rainfall distribution and co-occurrence probability confirms the idea of local low intensity events.

If we instead consider Figure 3.12 (e) and (g), the 5km line is nearly constant over the year, hence the small scale behaviour of heavy events are independent of the overall rainfall

probability. The seasonality in the climatology background is due to the much smaller chance of observing rainfall > 50mm outside the rainy seasons, hence the 'by chance' probability of observing this intensity of rainfall at two locations is much lower in the dry season. The near constant 5km probability in the left plot therefore suggests that heavier rain events have a large spatial structure that dominates the area wide behaviour. From the anomalies (right column Figure 3.12), we can see that the co-occurrence probabilities follows the baseline from about 50km for moderately intense rainfall and from around 100km for heavy and very heavy rainfall. This further confirms that heavier rainfalls have a larger spatial structure.

The seasonal behaviour in the baseline comes from the changes in probability of rainfall, presented in Figure 3.11, which impacts the large scale co-occurrence probability because of a higher proportion of dry stations. From Table 3.1, we can see the number of the number of days where at least one station observed the intensity and the total number of observations for each intensity. The ratio of these two gives us the average number of stations observing a certain intensity, e.g. 30-50mm, given that at least one station observes that same intensity. For very heavy rainfall, this ratio varies between 1:2 in the dry season up to 1:6 in June. The average is 1:4 during the rainy season except August, when this drops to 1:3. This indicates that during the dry season it is more common with just one heavy storm occurring, whereas in the rainy season there are on average 4 stations affected by very heavy storms on the same day. This can explain the peak in June in the baseline probability and the relatively constant behaviour during the rest of the rainy season except August.

Similarly, the peak in October for moderate intense rainfall (Figure 3.12 (c)) can be explained by an increase in this ratio from September to October, meaning that there is a much higher probability of moderate intense rainfall in other locations in October, given that it rains, compared to September.

From the right column of Figure 3.12 we can see that the decorrelation distance increases as we increase the rainfall intensity, strengthening our claim. No formal statistical test has been applied to test the difference between the anomalies and 0 but by eye, we can see that for low intensity rainfall already at 50km the anomaly is only around 0.05. For moderately intense rainfall, the same value is not reached until 100km away. Heavy rainfall exhibits a very similar range as moderate rainfall except a slightly larger peak in August. Very heavy rainfall has not fully converged, and therefore reached its decorrelation range, even at a distance of 150km, demonstrating a large-scale impact on the rainfall probability.

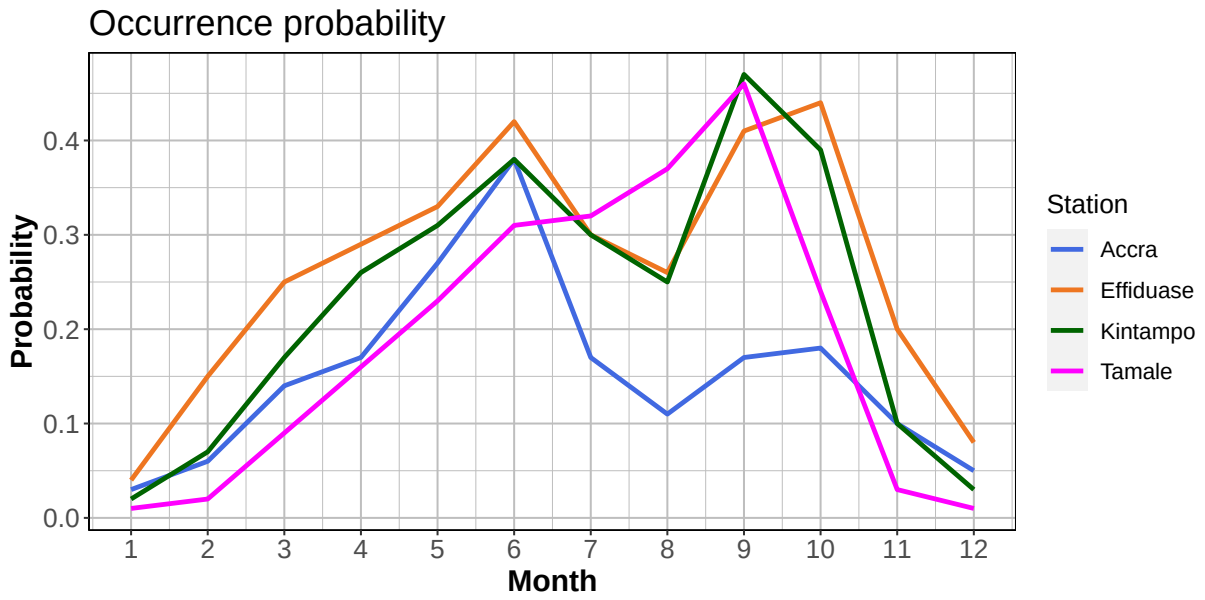


Figure 3.11: Probability of rain $\geq$ 1mm on any day in each month for the four stations Accra, Effiduase, Kintampo, Tamale.

Number of days and occurrences of each intensity

Intensity	1	2	3	4	5	6	7	8	9	10	11	12
Low	260	347	396	408	408	407	403	401	407	408	403	347
	853	6708	13111	15309	18839	23572	16715	13528	19531	22057	12212	5725
Moderate	195	296	380	397	404	404	375	345	399	408	395	279
	1255	3195	7025	8435	10178	12579	7154	4288	8678	10913	5365	2615
Heavy	98	188	313	341	344	387	280	224	341	357	287	164
	317	837	2241	2664	3102	4561	2361	1118	2562	2814	1090	680
Very heavy	54	86	201	229	252	309	208	127	229	237	132	77
	116	219	698	832	960	1798	1005	340	861	808	248	178

Table 3.1: Top row is the number of time steps with at least one station in the given intensity and the bottom row is the total number of occurrences in the given intensity. The maximum number of time steps is 408 and the maximum number of occurrences is 408*232.

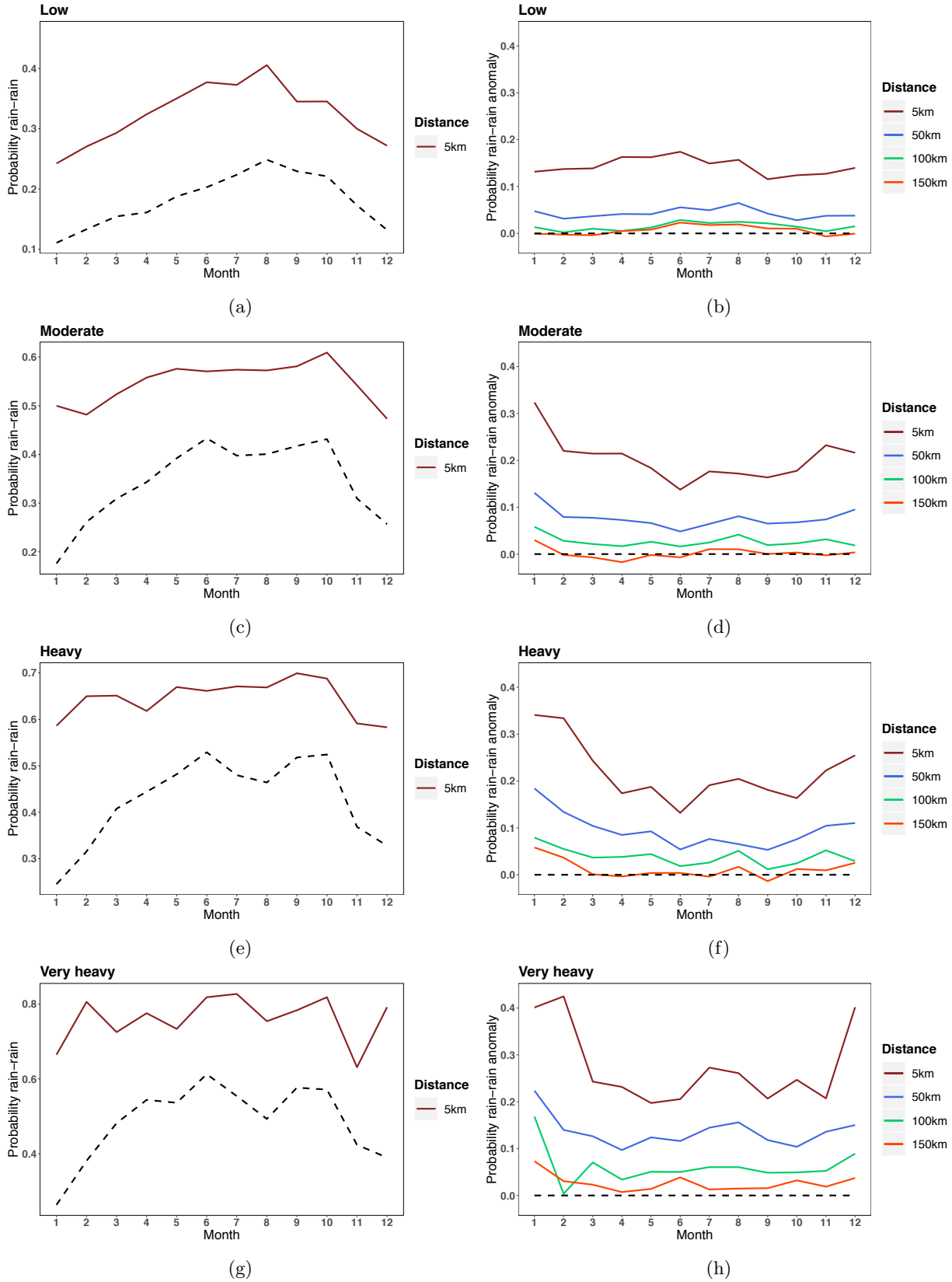


Figure 3.12: Seasonal evolution of the conditional occurrence probability for stations south of 8°N. (Left) raw co-occurrence probabilities, (right) the anomalies from the baseline. The solid lines are distances away from the origin and the dashed line (left) the climatology background at 50km (right) 0. The intensity bands are listed in Section 3.2.4. The rain-rain occurrence is 1 if the distant station is same or lower intensity. Note the different scales on the y-axis in the left column.

Spatial variability of rainfall occurrence

To better understand the spatial variability of rainfall occurrence for various intensities, and thereby the differences in spatial extent of areas with the same rainfall intensity, conditional probabilities for the separate intensity classes were calculated. The main point of the previous section was to understand the reach of a storm defined by its peak intensity, but where the other stations can record lower amounts. The aim here is instead to understand over what distances the different intensities are sustained, in other words over what distances the observations fall in the same intensity class. The same method as the previous section is used to again enable us to compare the measured probabilities with the climatology.

Figure 3.13 confirms the pattern in Figure 3.12, with the co-occurrence probability for heavy and very heavy rainfall not varying much with the season, even at long distances, whereas low and moderate rainfall has a strong seasonal pattern, from small to large scales. This implies that the probability of observing low or moderate rainfall at two nearby locations at the same time is mostly dependent on the seasonal probability of observing this intensity, but not for heavier intensities. In other words, it does not matter if we observe heavy rainfall in February or June, the chance of observing heavy rainfall at a nearby stations remains the same. The seasonal pattern in the baseline for low and moderate rainfall again highlight the different overall probability during the monsoon season.

At all distances, there is a peak in the probability in August for low (brown) intensity but a trough for all other intensities. This is most likely explained by the higher frequency of low intensity rainfall events compared to other intensities in the short dry season, which increases the probability of co-occurring low intensity events and decreases the other intensities. The climatology co-occurring probability of low and moderate (blue) intense rainfall is however close to identical during the build up phase, March-May. For moderately intense rainfall, there is an increasing overall probability until June, decreasing during July and August and then increasing again. For heavy (green) and very heavy (orange) rainfall, the climatology probability is relatively constant over time. Low intensity rainfall has nearly converged to the climatology at 100km whereas moderately intense rainfall has converged at 150km. Hence areas where all stations record moderately intense rainfall has a larger extent than areas with low intense rainfall. Neither heavy nor very heavy rainfall has converged, indicating that rainfall events that persistently releases more than 30mm of rain has a spatial dependence even at 150km away.

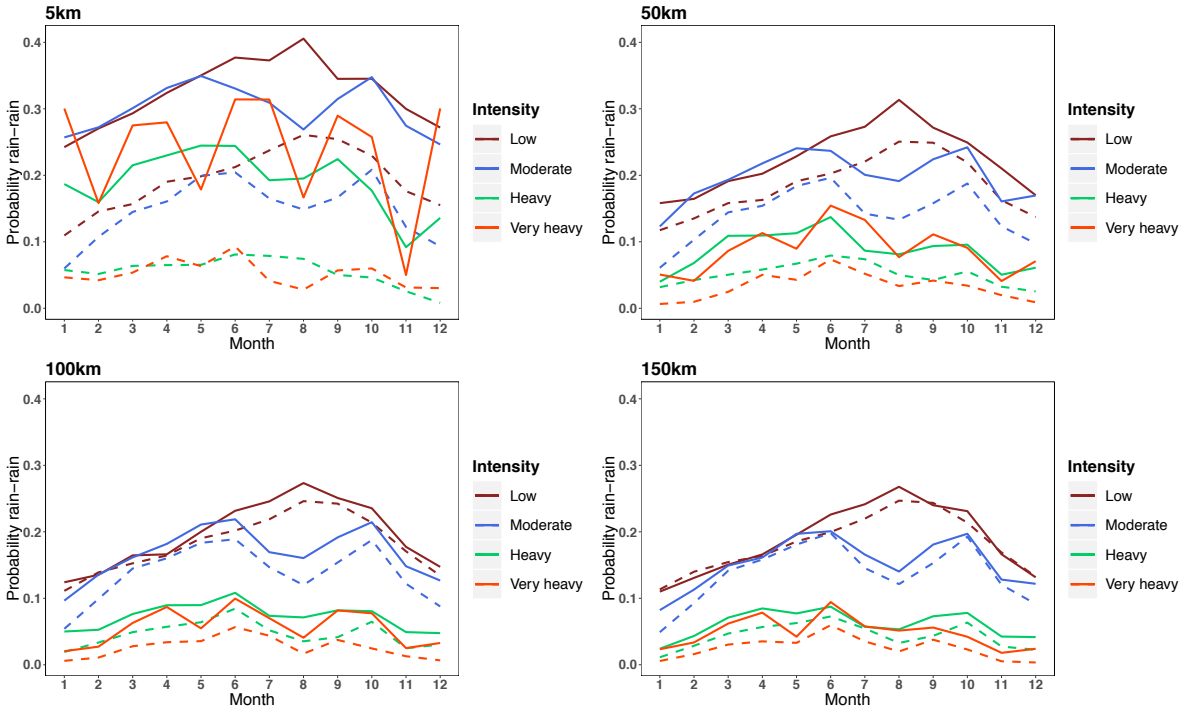
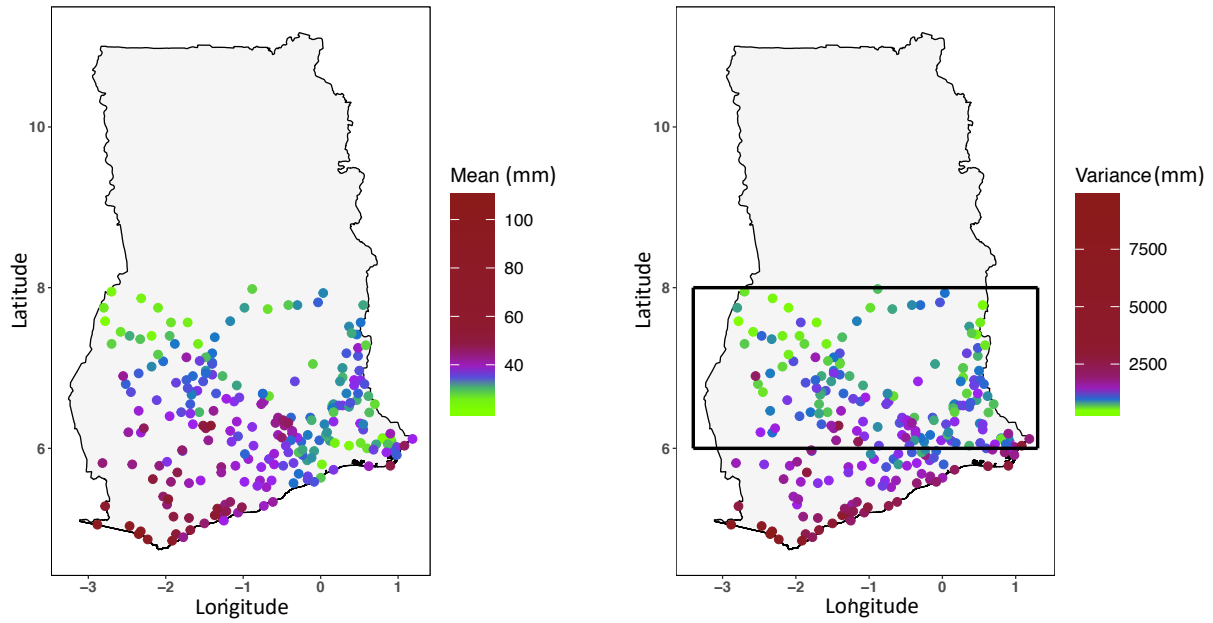


Figure 3.13: Seasonal evolution of the conditional probability at different distances for stations south of 8 °N, using Algorithm 1 in Section 3.2.4. The solid lines are the probabilities from the original dataset and the dashed lines the probabilities from the random sampling method. The rain-rain occurrence is 1 if both the distant station and the origin station are in the given intensity class.



(a) (b)

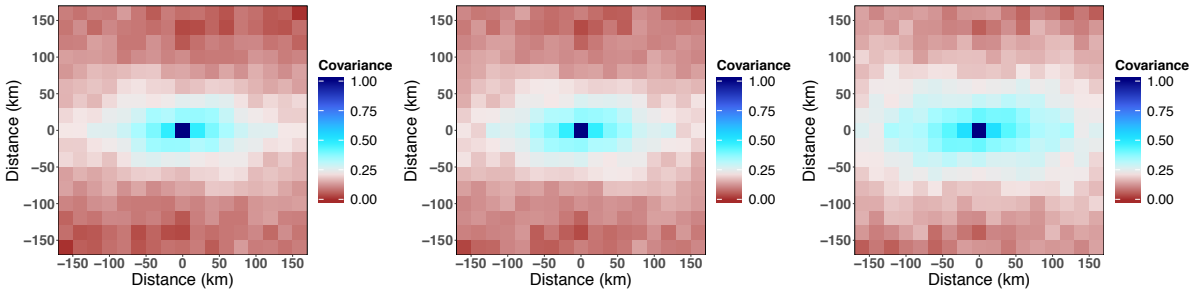
Figure 3.14: Distribution maps in June over Ghana. The maps show (a) mean of 5-day aggregated values (mm), (b) variance of 5-day aggregated values (mm) and the region used to estimate the covariogram maps.

Spatial shape of rainfall events

To study if the large scale drivers such as the East African jet also can be seen on the small scale, we estimated covariogram maps. This will be done using 2-, 3- and 5-day aggregated June data, which as previously described is to reduce the level of noise while still looking at short time scales. Since the aim here is to look at the potential anisotropic pattern generated by all rainfall, we are now including all days without splitting it into intensity classes. We know from Figure 3.6(a) that there is a strong rainfall gradient in the NW-SE direction in annual amount, but we want to see if there might be a different spatial variability pattern when looking at accumulation over only a few days. Because of the significantly larger mean and variance in the south west corner (Figure 3.14), the following analysis will be done on the indicated region in Figure 3.14(b), to work with the assumption of equal mean and variance over the entire region. This difference in mean and variance did not affect our previous results since we worked with intensity occurrence instead of amounts. Covariogram maps were also estimated over the coast region, but are excluded due to their very noisy pattern.

The patterns in the covariogram maps in Figure 3.15 have a lot of similarities for all aggregation periods but some small differences as well. One can clearly see a higher correlation distance in the E-W direction compared to the N-S direction in all aggregation periods. There

is however no clear difference in the NE-SW and NW-SE direction. Hence even on a 2-day scale we can see the pattern of dominantly westward propagating convection systems. The mean propagating speed in this region is 8ms^{-1} (Maranan et al. (2018)), hence the average storm would travel roughly 700km in a day. This would imply that multiple rainfall events could pass over the 150km window in a 2-day period, and the covariance would therefore be an average of these. This averaging is what reduces the noise and instead highlights the main patter. The correlation drops off very rapidly with a correlation of around 0.5 just 20km away and as we increase the aggregation period, the correlation gets coherently increased in all directions. The area with a correlation of 0.25 is however extended much further in the E-W direction compared to N-S, demonstrating an increased anisotropic pattern for longer aggregation periods. Even at 160km, there is some correlation which is due to the climatology similar to the previous results. The higher correlation in E-W is not due to the fact that the region is wider than long, which was checked by doing the same calculation over a square region.



(a) (b) (c)

Figure 3.15: Covariogram maps over the mid region in June. Each square is the average covariance value in that distance and direction bin and each distance bin is 20 km. The graphs shows (a) 2-day, (b) 3-day and (c) 5-day accumulated daily data.

3.4 Discussion and conclusion

Estimating and predicting rainfall over west Africa will probably remain a difficult task for some years ahead due to the sparse and degrading rain gauge network. In this paper, we have provided some insights on the spatial behaviour of daily rainfall within Ghana. In contrast to previous studies, we have not assumed that all rainfall events have the same spatial structure, but instead studied the rainfall events split into four different intensity classes to understand differences in the co-occurrence structure at small scales over the season. We have showed that the conditional probability of observing rainfall of the same intensity varies seasonally for low and moderately intense rainfall, but not for heavy and very heavy. This might partly be

explained by changes in the proportion of rainfall events in each intensity class over the season, however the same pattern is observed when treating all lower intensities as occurrences. This shows that heavy rainfall events have a stronger influence at the local scale and therefore is not affected by the overall probability of rainfall, whereas low and moderate rainfall are much more localised. The anomalies structure is also very similar for all intensities except low, demonstrating a different structure in drizzle events compared to other rainfall events.

Our results show a decorrelation distance of around 100-150km for all intensities, except low which is around 50km, similar to the one obtained by Ricciardulli and Sardeshmukh (2002) (150km) and positive amount in Teo and Grimes (2007) (150km), but about three times further than their occurrence range (50km). This difference probably results from several factors, such as the scale difference of gridded data and station data, how dense the dataset is and the method used. But a non-negligible part most likely also comes from the use of different regions. Because of the complex atmospheric systems over central Africa, the rainfall structure varies greatly, as shown in Funk et al. (2015b). This makes it difficult to directly compare large scale estimates, such as correlation ranges, with previous studies. Small scale estimates on the other hand such as hourly rainfall amounts or similar, which mostly depend on the rainfall system and not the current rainfall state, could be more comparable assuming the convective systems are similar across tropical Africa. It would however be difficult to compare with European studies since the rainfall there mostly comes from moist air masses from the Atlantic and Mediterranean being advected over land.

For the spatial shape of rainfall events, even at the small scale it is possible to see the influence of large scale drivers such as the African easterly jet. As we increase the accumulation period, the covariance range is increased in the E-W direction, which is the direction of stronger covariance. The pattern of stronger correlation at longer accumulation periods was also noted by Bacchi and Kottegoda (1995) which can be explained by the decreased dependence of the individual rainfall events, which are local scale events, and more on the large scale drivers which usually affect a hole region.

The results in this paper demonstrate the issues with describing all rainfall events with the same correlation structure, but that we can assume isotropy for short accumulation periods. We hope that this method will be applied in other regions since it is easy to adapt by changing the intensity classes to suitable country levels and the results from different studies can be directly compared. It would be especially interesting to see how this compares to other tropical regions where we might expect to see the same type of rainfall systems but with a different occurrence distribution compared to our study region.

3.A Algorithms for calculating co-occurrence probabilities

Algorithm 1 - Co-occurrence within intensity class (Figure 3.16)

1. For each unique day, that is for one of our 408 time steps, transform all amounts that are within our chosen intensity class to a 1 (green dot) and all other amounts to 0 (black dot). E.g. if we are interested in modelling the moderate intense rainfall, then all locations with a measured rainfall in the range 10-30mm will be assigned a 1 and all other locations a 0.

2. Choose one of the stations that are assigned a 1 to be the origin station (pink dot) and calculate the distance from this station to all other stations.

3. Within each 10km distance bin (blue circle), that is for all stations with a distance of 0-10km, 10-20km,..,140-150km of the origin station, calculate the proportion of stations assigned a 1 in step 1.

4. Repeat step 2-3 for all stations assigned a 1 (green dots) in step 1.

5. Repeat step 1-4 for each unique day.

To model the dependence structure of co-occurring rainfall events between an intensity class and either lower or higher intensity classes, the above method is used with a few changes (Figure 3.17). In step 1, transform all amounts that are within our chosen intensity class to a 1, all stations with a measured amount in all lower or higher intensity classes with a 2 (orange dot) and all other stations to 0. In step 3, calculate the proportion of 1's and 2's instead of just 1's. The rest of the algorithm is identical. Just as for measurements within an intensity class, taking the average in each distance bin, we can get a climatological average on the probability of observing rainfall of the same intensity or lower (higher) as the origin station for a given distance.

Algorithm 2 - Background state within intensity class (Figure 3.18)

1. For each unique day, that is for one of our 408 time steps, calculate the proportion of rainy stations (≥ 1mm) (blue dots).

2. Transform all amounts that are within our chosen intensity class to a 1 (green dots) and all other amounts to 0 (black dots). E.g. if we are interested in modelling the moderate intense rainfall, then all locations with a measured rainfall in the range 10-30mm will be assigned a 1 and all other locations a 0.

3. Choose one of the stations that are assigned a 1 to be the origin station (pink dot) and calculate the distance from this station to all other stations.

4. For all other stations, randomly assign it to be rainy (blue squares) or dry (black dots) so the proportion of rainy stations equals the proportion in step 1.

5. For each station assigned rain (blue square), randomly assign a measured rainfall amount from that station in that unique month. E.g. for the unique time step 6^{th} of May 1965 and station 120, draw any measured amount ≥ 1mm from the measurements taken at station 120 in May 1965.

6. For all stations with an amount in the chosen intensity after step 5, assign 1 (green dot) else 0 (black dot).

7. Within each 10km distance bin (blue circles), that is for all stations with a distance of 0-10km, 10-20km,..,140-150km from the station selected in step 3, calculate the proportion of stations assigned 1 in step 6.

8. Repeat step 3-7 for all stations assigned 1 in step 1.

9. Repeat step 1-8 for each unique day.

For the calculation between intensity classes the above method is applied with the modification:

- Change step 6 to the method used in the between intensities cases.
- Change step 7 to calculating the proportion of 1's and 2's instead of just 1's.

3.B Schematic overview of the algorithms

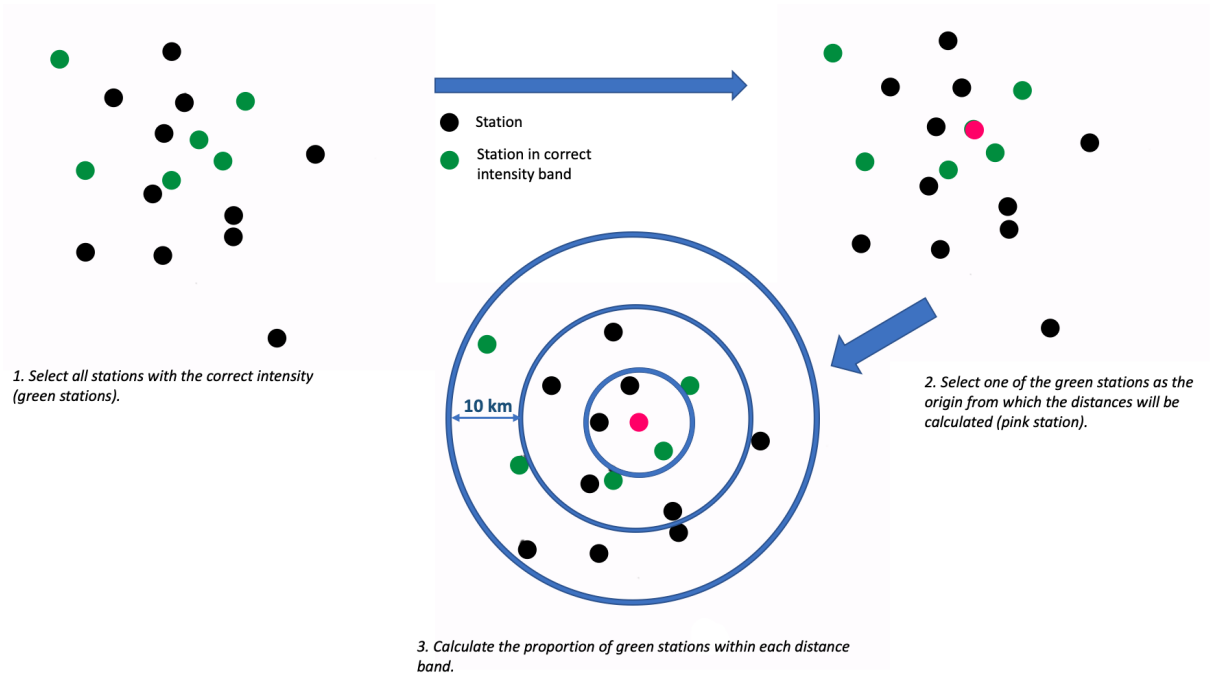


Figure 3.16: Schematic figure on the method to calculate the spatial co-occurrence dependence within an intensity band. The green stations are within the chosen intensity band and assigned a 1 and the black stations are of other amounts and assigned a 0. The pink dot is the station chosen as the origin station. Step 2 and 3 are repeated for each green station in step 1 and step 1-3 are repeated for all 408 unique days.

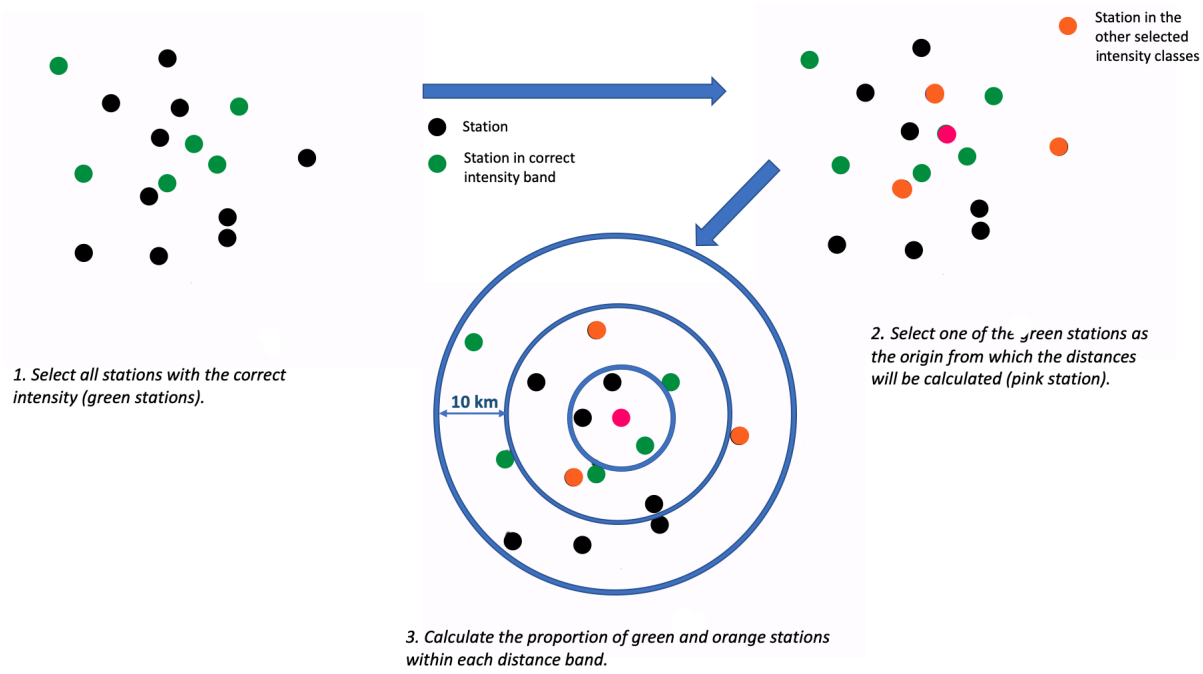


Figure 3.17: Schematic figure on the method to calculate the spatial co-occurrence dependence between lower (higher) intensity bands. The green stations are within the chosen intensity band and assigned a 1, the orange stations are the lower (higher) intensity bands assigned a 2 and the black stations are of other amounts and assigned a 0. The pink dot is the station chosen as the origin station. Step 2 and 3 are repeated for each green station in step 1 and step 1-3 are repeated for all 408 unique days.

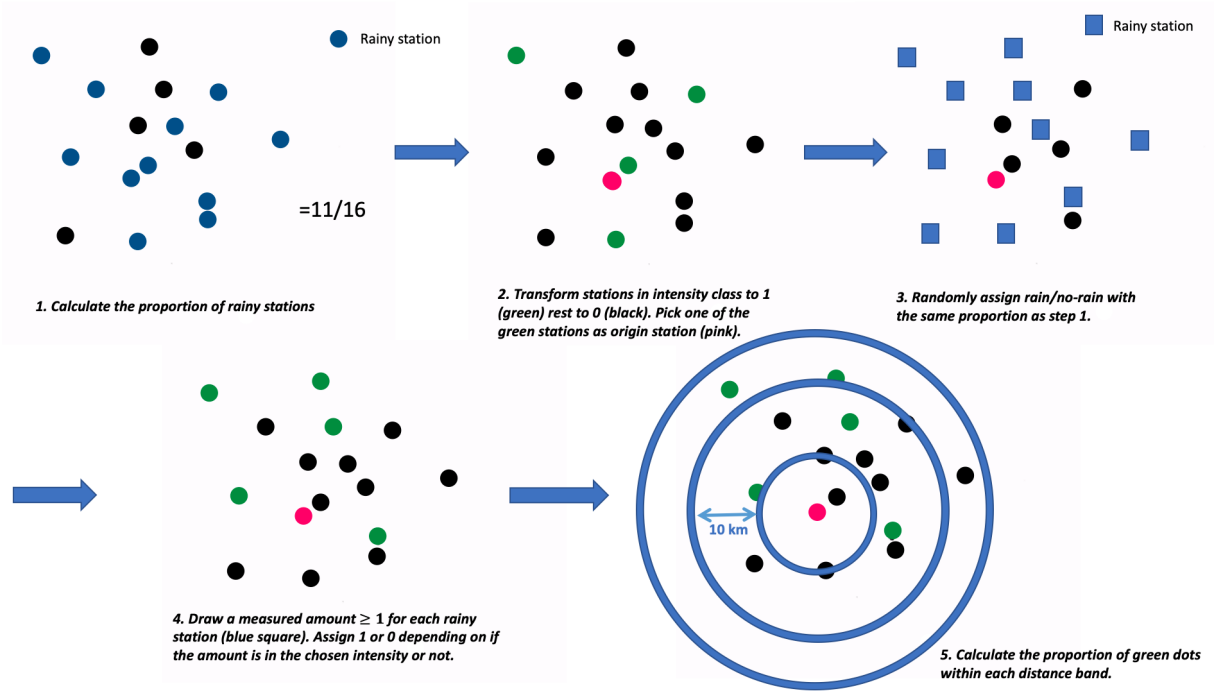


Figure 3.18: Schematic figure on the method to calculate the spatial co-occurrence dependence within an intensity band. The green stations are within the chosen intensity band and assigned a 1 and the black stations are of other amounts and assigned a 0. The pink dot is the station chosen as the origin station. Step 2 and 3 are repeated for each green station in step 1 and step 1-3 are repeated for all 408 unique days.

Chapter 4

An analysis of the conditional amount distribution for gauge observations associated with satellite measurements

In order to match ground observations to satellite rainfall estimates, the relation between the two needs to be understood. In this chapter, question 3 in the thesis aims is addressed by evaluating the fit of two different skewed distributions to daily rain gauge data that are associated with a TAMSAT satellite rainfall estimate. This is achieved by applying a number of different qualitative tools, such as histograms and QQ-plots. The improved information is important for applications such as merging gauges with satellite rainfall estimates or for generating an ensemble of rainfall estimates for a given satellite image.

A discussion of several possible extension of this work, by including a wider set of distributions and possible quantitative methods for comparing the distribution fit to the conditional gauge values, is also provided.

4.1 Overview

To estimate rainfall amounts using satellite data, two commonly used techniques are available which naturally come with their individual limitations and drawbacks. The Thermal Infrared (TIR) method, which all products described here use to some extent, in principal assumes that all cold clouds generate rainfall but not warm clouds. However the complete lack of rainfall from warm clouds can be adjusted in the calibration process. The huge advantage with these sensors is that they are placed on geostationary satellites, i.e. satellites that stay in the same place in relation to Earth, and therefore can capture everything that occurs in that region. These types of observations have been collected since the 1970s through the European Organisation for the Exploitation of Meteorological Satellite (EUMETSAT) project, meaning long available records which allows climate studies to be made on these (Maidment et al. (2020)). Passive microwave (PMW) sensors measure the thermal emission of raindrops using low frequencies, and from the earth for higher frequencies. These sensors are however only available on polar-orbiting platforms, and therefore only pass over any location for very short periods and with uneven intervals (Joyce et al. (2004)), but the 30 min interval data it returns can better capture the intensity compared to the TIR method. Unfortunately, the understanding around separating the rainfall scattering from the background (Earth) scattering is best understood over the ocean, resulting in poorer estimates over land where we have the biggest interest.

To reduce the error from the individual information sources, a combination of the two types of satellite products and gauge data is commonly used. African Rainfall Climatology v2 (ARC2) produced by NOAA includes TIR data and gauge data from Global Telecommunications Station (GTS) network to produce a climate record product (Novella and Thiaw (2013)). Their merging technique is based on an algorithm introduced in Reynolds (1988), which uses the gauge values where these are 'dense enough', called anchor points, and otherwise estimates the amount numerically by solving the Poisson equation $\nabla^2 B = \nabla^2 C$, where B is the merged product, C the satellite estimate and boundary conditions given by the anchor points (Xie and Arkin (1996)). An extension of this is the Rainfall Estimation Algorithm (RFE 2.0) which also includes PMW estimates (Climate Prediction Center (*The NOAA Climate Prediction Center African Rainfall Estimation Algorithm Version 2.0*)). The drawback with this product is the short record (1997) of PMW measurements, which means that the product cannot be used for climate applications. The three satellite sources included (two PMW and TIR) are linearly combined with associated weighting coefficients based on the satellite products random error, which is derived by comparing the estimates with the Global Precipitation Climatology Centre (GPCC) gauge measurements. These weighted satellite estimates are then combined with station data in the same fashion as in ARC2. Both of these products, ARC2 and RFE 2.0, return

daily estimates with a 0.1°resolution and a latency of around 1 day.

Another product that utilises all three information sources is the Climate Prediction Center (CPC) CMORPH , which uses TIR imaginary from geostationary TIR satellites to create back- and forwards propagation vectors, so called 'morphing', for the PMW observed rainfall clouds to return half-hour estimates at a 8km $\times$ 8km grid (Joyce et al. (2004)). These estimates are then bias reduced using CPC daily gauge analysis over land. The bias reduction technique is based on matching the satellite and gauge estimate PDFs for each 0.25°grid box. This is performed in two steps, the first one reducing the climatological bias and the second step the inter annual bias. One drawback with the method is that it is unable to shift a non rainy estimate to a positive value and the requirement of a dense gauge network to construct all the individual PDFs (Xie et al. (2017)). Due to the bias reduction method, it has a latency of 3-4 months. The 30 min product is released in the native 8km $\times$ 8km resolution and an aggregated daily product at a 0.25°resolution grid.

A very similar product to CMORPH is the NASA Integrated Multi-satellite Retrievals GPM (IMERG) product (Huffman et al. (2015)), which is the successor of the since 2014 retired Tropical Rainfall Measurement Mission (TRMM) (Huffman et al. (2007)). It produces 3-hourly rainfall estimates on a 0.1°grid, combining a multitude of PMW sensors, TIR images and where possible, bias adjusted by matching the monthly total rainfall with the GPCC monthly gridded gauge data. The monthly totals gridded gauge data is available at resolutions between 0.25°and 2.5°and stretches back to 1981. They use an interpolation method called SPHERMAP, which builds on the empirical Inverse Distance Weighting (IDW) method derived by Shepard (1968) (Becker et al. (2013)). This includes a power parameter p that controls the influence given to close and distant points, coupled with a spherical adaptation to take into account the direction of the interpolation points and the gradients of the data field. If stations only are found within a small radius r_1, a simple arithmetic mean is calculated, else all stations are interpolated using the weighting method. The GPM IMERG data is also released with a up to three months latency.

TAMSAT and CHIRP are two high resolution products only using TIR satellite derived estimates which are calibrated against gauges, with CHIRP also existing as a product with gauge data blended with the satellite estimate, called CHIRPS. The TAMSAT data set, initially named TARCAT, has been produced since the 1980s based on the TAMSAT rainfall estimation algorithm (Grimes et al. (1999), Milford et al. (1996), Dugdale et al. (1991)), which will be described in the following section. It originally was released as dekadal (10 day) data and only covered the main rainy season months in the northern and southern/eastern

Africa, before being extended to cover the full continent in 2014 (Tarnavsky et al. (2014), Maidment et al. (2014)). The extension was achieved by introducing large, irregular calibration zones for each calender month, that were determined from knowledge about the local weather, availability of gauges and the frequency bias between all gauge-CCD (Could cloud duration, see Section1.2.3) pairs. In 2017 two new daily data versions were released, one disaggregating the 10-day calibration zone data (v2.0) into daily estimates and a completely new data set, where the calibration zones are replaced by 1.0°calibration boxes and the initial estimates are made on pentadal (5 day) instead of dekadal time scale (Maidment et al. (2017)).

The satellite-only product CHIRP builds on first estimating the 0.05°monthly precipitation climatology product CHPclim (Funk et al. (2015a); Funk et al. (2015b)). CHPclim is derived from two long-term, monthly means station data sets, and incorporates information from the commonly used physiographic predictors; elevation, latitude and longitude, along with information from five satellite products that are all resampled to a common 0.05°grid. The resulting product is a pentadal means field which represents the 1980-2009 climate normals.

CHIRP data is then estimated as variations from the CHPclim, using CCD measurements at the fixed temperature threshold 235K at a 0.25°grid, from which monthly regression parameters are estimated using TRMM data. This is then resampled to the 0.05°grid before producing the pentadal rainfall estimate by multiplying the deviation fraction with the CHPclim estimate. To incorporate station data and obtain the blended product CHIRPS, a modified IDW algorithm is utilised. CHIRP is used to estimate the decorrelation range, which is described in Section 3.1. A *bias ratio* vector, **b**, is calculated from the 5 closest stations, by dividing the station observation by the CHIRP value. For stations beyond the decorrelation range, the bias is set to 1. The station values are also weighted by a factor $\alpha = R_{CHIRP}/(R_{CHIRPS} + R_{ns})$ where R_{ns} is the expected correlation with the nearest station and R_{CHIRP} is the correlation between the true rainfall value and CHIRP data. The final CHIRPS estimate is given by CHIRPS$= \alpha$CHIRP $+ (1 - \alpha)$**b**CHIRP.

Summary table over satellite rainfall products

Name	Data input	Spatial resolution	Temporal resolution (shortest provided)	Latency	Start year
ARC2	TIR, GTS gauge	0.1°	Daily	1 day	1983
RFE 2.0	TIR, PMW, GTS gauge	0.1 °	Daily	1 day	1995
CMORPH	TIR, PMW	0.07°	30-min	3-4 months	2002
GPM IMERG	TIR, PMW, GPCC gauge	0.1°	3-hour	3 months	2014
TAMSAT	TIR, gauge	0.0375°	Daily	Up to 6 days	1983
CHIRP	TIR, TRMM	0.05°	Daily	Up to 6 days	1981
CHIRPS	TIR, TRMM, GTS gauge	0.05°	Daily	3 weeks (initial 2 days)	1981

4.2 The TAMSAT estimation method and gauge-RFE merging

4.2.1 TAMSAT estimation process

TAMSAT satellite-only rainfall estimates, which throughout will be denoted as RFE (not to be confused with the rainfall product RFE 2.0), are derived from the assumption that the amount of rainfall over an area is linearly related to CCD, especially over longer accumulation periods. The CCD is the accumulated time that a satellite pixel records clouds with a certain or lower temperature. Since the CCD depends on which temperature threshold we select, we will hereafter denote it CCD_T to highlight this dependence. The linear assumption between CCD_T and rainfall has been shown to be accurate for Africa since the majority of the rainfall (around 90% of the annual rainfall) is coming from deep convective systems (Nesbitt et al. (2000); Mathon et al. (2002)). The CCD_T values are estimated from TIR sensors through a calibrated Inverse Planck function, a relation that has been improved over the past 20 years (Maidment et al. (2014)). The Meteosat TIR data is obtained every 15 min since July 2006, and every 30 min prior to this.

The correlation between CCD and RFE is stronger for longer accumulation periods, which is why the current TAMSAT v3.1 estimates the total pentadal rainfall and then disaggregates

this to daily amounts based on the proportion of the total CCD_T observed that day. The relationship between the pentadal CCD_T and RFE is given by

$$\text{RFE} = \begin{cases} a_0 + a_1 \, \text{CCD}_T & \text{CCD}_T > 0 \\ 0 & \text{CCD}_T = 0 \end{cases}$$

The specific temperature threshold T is calibrated from a range between -30°C and -65°C for 1°grid boxes that contain rain gauges. The calibration process selects the optimal temperature based on a 'rain, no rain' contingency table matched against the rain gauges. These values are then interpolated to grid boxes without any gauges (Maidment et al. (2017)), after which the linear coefficients a_0, a_1 are fitted using the gauge values.

4.2.2 Merging gauges with satellite grid data

In the current satellite-only version of TAMSAT, gauge information is only used in the calibration process, whereas a merged product would continuously incorporate this information to improve the estimates. The merging method [1] is based on estimating the rainfall using satellite data at ungauged locations and gauge measurements where these exists. This is achieved through the method illustrated in Figure 4.1, where captial letters mark input data or data conversions and lower case letters information flow. The satellite-only TAMSAT rainfall estimates (A) provides the conditional rainfall distribution to be used (a) for the observed gauge measurements (B). The normal score for the gauge measurement is derived through a cumulative density function (CDF) transformation (C) mapped to the corresponding standard normal distribution value. The normal score value is mapped (c) onto a TAMSAT resolution grid (D), with grid cells containing a gauge and therefore a normal score marked by green. Grid boxes 'close enough' to the gauged green boxes are marked by orange and are assigned an estimate and variance through kriging and the rest are assigned 0. Once the grid is filled, the normal score values are back transformed (E) to an updated rainfall estimate, which provides an adjusted rainfall map (F).

The contribution of this thesis to the methodology is in the improvement of the gauge distribution conditioned on the satellite estimate (C and E in Figure 4.1), to better capture the full range of observed values and thereby a correct distribution of normal score values (see next section).

[1] The proposed merging method developed here is in development within the wider TAMSAT programme, with a final description of the methodology in preparation for publication (E. Black and R. Maidment Pers. Comm). The methodology in Section 3.2.1 and the process in 3.2.2 are developed by the TAMSAT group. Section 3.3 and the flow charts in Section 3.2.2 are developed in this thesis.

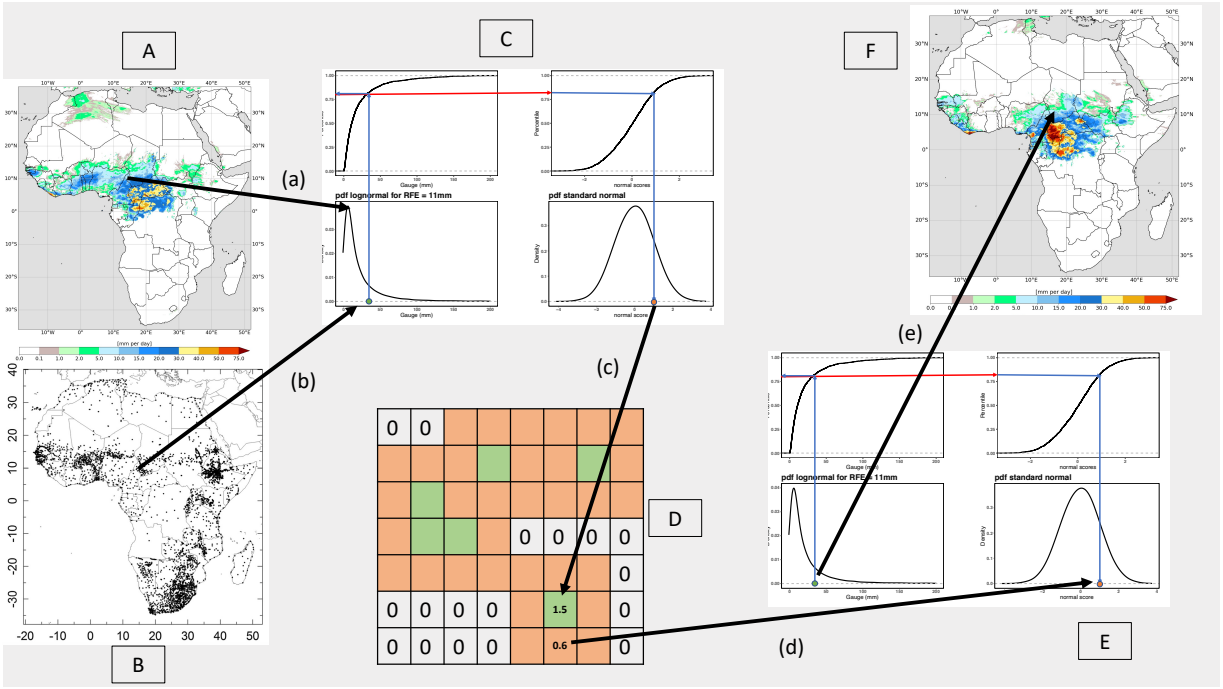


Figure 4.1: Flow chart demonstrating the steps in the merging method explained in the first paragraph in section 4.2.2. Capital letters marks input data or processes and lower case letters information flows.

The method is inspired and based on the work by Teo and Grimes (2007) and Greatrex et al. (2014), where a similar approach is used, but for other applications. The aim of these two papers were to generate a realistic ensemble of satellite rainfall estimates given one TIR image, and thereby investigate the uncertainty of the rainfall estimate which then could be compared to gauge measurements. There are however some crucial differences between their method and the one described here. Firstly, they worked with CCD data since the goal was to produce multiple maps of rainfall estimates given the same input, by for each iteration assign a grid cell wet or dry based on a Bernoulli trial (sampling 0 (dry) or 1 (rain), with the probability of choosing 1 equal to p) and then sample from a conditional CCD rainfall distribution. Secondly, they randomly selected a number of 'seeds' (green boxes) to start the kriging process, whereas here all grid boxes with gauges are naturally set as 'seeds'.

The following subsections describe the individual steps of the method before introducing where the findings in this work can make an improvement.

Normal scores

The principal idea behind the method developed within TAMSAT is that given an estimated RFE value, all rain gauge measurements that can be associated (i.e. observed in the same gridbox) with this value follow some continuous distribution, as illustrated in Figure 4.2. From this distribution, one can evaluate how anomalous an observed rain gauge measurement is compared to the expected value. This information can then be used to shift surrounding values based on the deviation from the mean by obtaining estimates through the method of kriging on the calculated deviations. The anomalies are calculated by converting the rainfall amount into a standard normal distribution z-value, here called normal score. This is done by mapping the rainfall value to its quantile value given by the Cumulative Distribution Function (CDF), which further is mapped to the corresponding z-value. This mapping method will be denoted a *CDF translation* and Figure 4.3 graphically describes this process.

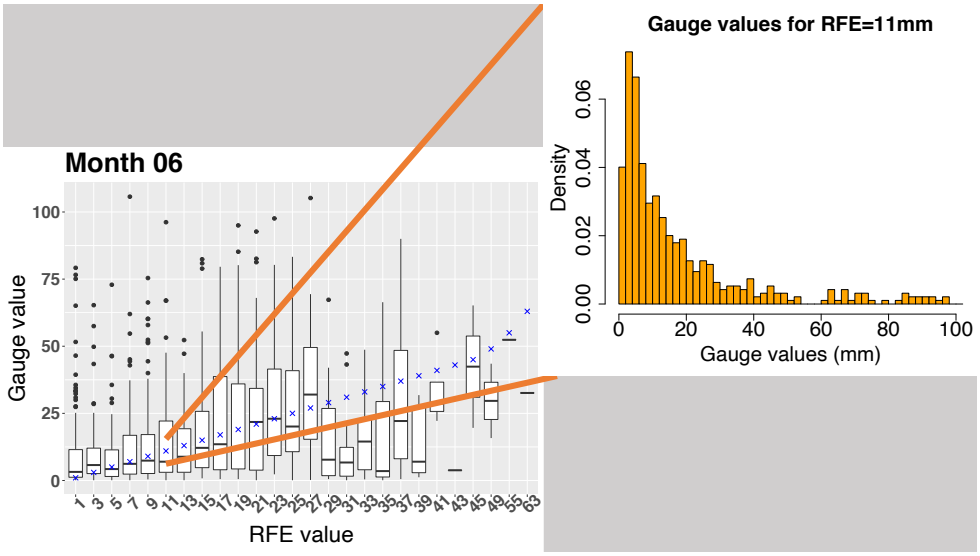


Figure 4.2: Illustration of the conditional distribution for gauge values given a RFE.

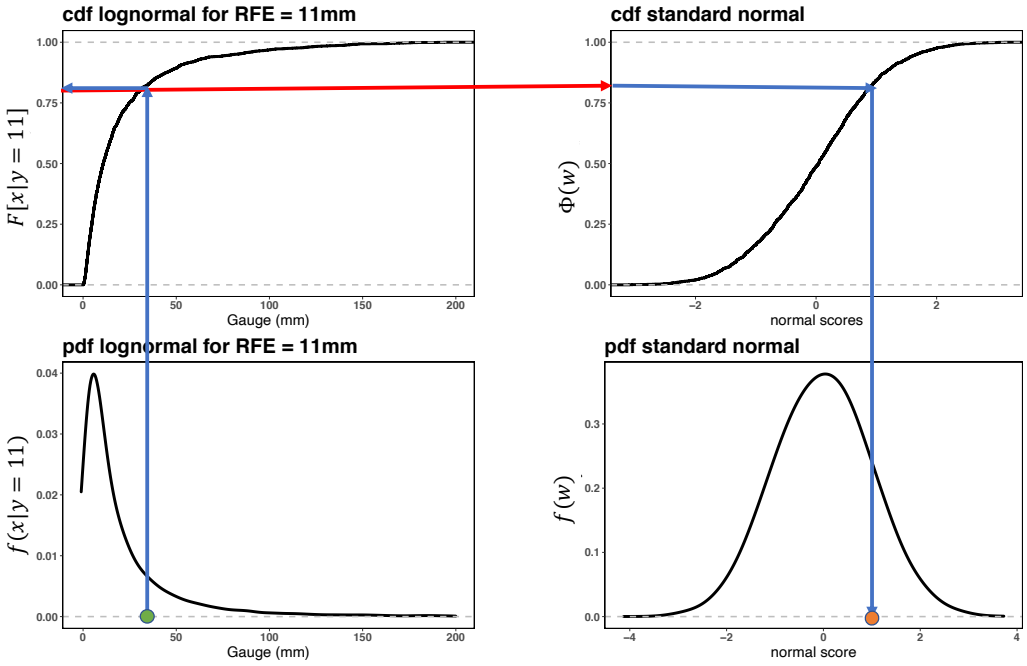


Figure 4.3: Illustration of converting a gauge value to the associated normal score through a CDF translation. The green dot marks the gauge value and the orange the corresponding normal score. The red arrow marks the CDF translation.

In more mathematical terms this means:
Let x be a rain gauge value and y the associated RFE value. We assume that the conditional distribution of x is given by

$$f(x|y) = \mathcal{D}_{RFE}(\mathbf{g}) \tag{4.1}$$

where $\mathcal{D}_{RFE}$ is some probability density function whose parameter vector $\mathbf{g}$ depends on the RFE value y. To convert x into a normal score value, ns, with F denoting the absolute continuous conditional distribution function associated with (4.1), the density value of x is mapped into its equivalent z-value through the process below

$$\begin{aligned} p &= F[x|y] \\ ns &= \Phi^{-1}(p) \end{aligned} \tag{4.2}$$

where Φ^{-1} is the inverse of the standard normal CDF. Hence if $\mathcal{D}_{RFE}$ accurately models the distribution of x given y, then ns is now by construction normally distributed with $\mu = 0$ and $\sigma^2 = 1$.

Kriging

The method of kriging is explained in Section 2.1, where several types of kriging are listed. Due to the construction of the merging method, simple kriging is used. This comes from using the transformation to a standard normal distribution as earlier done in Teo and Grimes (2007) and later Greatrex et al. (2014), which leads to the mean and variance known a priori to be 0 and 1 respectively, hence fulfilling the requirements of simple kriging.

Back transform normals score to RFE

After converting all gauge measurements into normal scores, the entire grid (given that it is within the correlation distance of some gauge) can be kriged out, returning both a kriged estimate and variance given by Equations (2.3), (2.4) respectively. Grid cells outside the range of any gauged cells are assigned 0. The process in Equation (4.2) can now be reversed on both the estimate and some confidence interval values, e.g. $\pm 2\sigma$, to obtain the adjusted rainfall estimate and its associated uncertainty. Grid cells outside the influence of gauges are by default assigned the initial RFE value (since a 0 normal score is mapped to the mean value of the distribution) and gauged cells to the measured value since their normal scores has not been changed in the kriging step. Given the sparse and unevenly distributed rain gauge network over much of Africa (see Chapter 1), this is an important feature since at each grid point the rainfall estimate is based on the most reliable source of information. This however risks generating a slightly unrealistic map since only locations with gauges, or that are nearby, can be assigned the higher adjusted values, potentially creating the skewed view that it consistently rains less in ungauged areas. But this is an issue that is nearly impossible to get around if one wants to improve the estimates where it is possible.

4.3 Conditional amount distribution

One very important part of this method is the choice of the distribution function $\mathcal{D}$ and the expressions for the associated parameter vector, $\mathbf{g}$, since this determines the normal scores for the kriging algorithm. A too light tailed distribution will result in many very large normal scores, since the heavier rain gauge values will occur much more frequently than modelled by the distribution. This will result in unrealistically large normal score values and a positive skew being applied much more frequently than expected. A too heavy tailed distribution will on the other hand concentrate all gauge values around a subset of normal scores and return too wide and therefore non-informative CI. Figure 4.4 demonstrates the impacts of the light tailed misspecification, which both can be due to wrongly specified distribution function $\mathcal{D}$ or the parameter vector $\mathbf{g}$.

A commonly used method for accurately capturing different parts of a distribution where no single model provides a good fit, is to introduce a mixed model where one distribution function models the main part of the distribution and an extreme value distribution the tails. This is however not an option here due to the back transform step (E in Figure 4.1), which requires a continuous distribution function which a mixed distribution will not have.

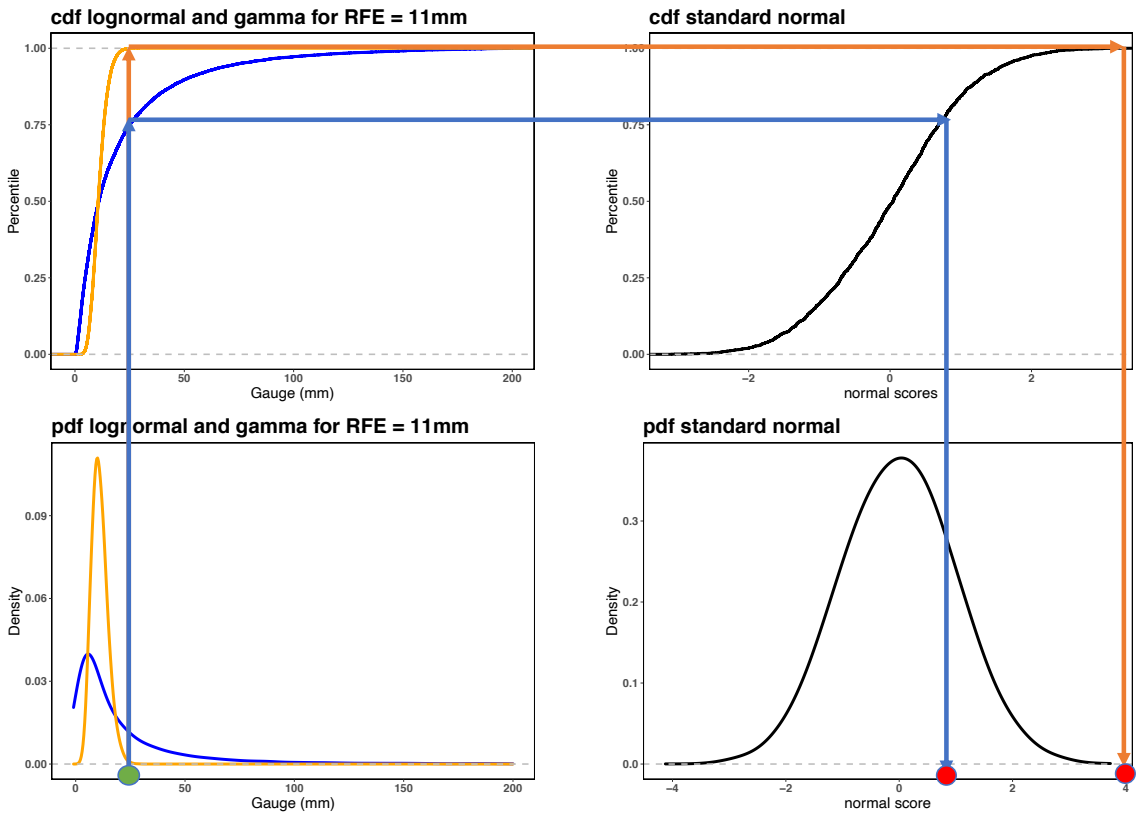


Figure 4.4: Illustration of the impact on the normal score by incorrectly defining the conditional distribution. The blue line and arrows is the lognormal and the orange the gamma distribution.

In the conceptual framework described in Teo and Grimes (2007) and Greatrex et al. (2014), $\mathcal{D}$ is a gamma distribution where the vector **g** is the shape and scale parameters, defined such that the mean and variance are given by

$$\mathbb{E}[X|y] = y \tag{4.3}$$

$$\mathsf{Var}[X|y] = \kappa\, y^{\theta} \tag{4.4}$$

That is, $\mathbf{g}$=(RFE$^2/\sigma, \sigma/$RFE), where $\sigma = \kappa *$ RFE$^{\theta}$. The justification for the mean can be seen in Figure 4.5, where the gauge values for each 2mm RFE value bin are shown in a Box and Whiskers plot. The centre RFE value, and therefore the assumed mean, is marked by blue crosses and in general align well with the observed median, especially for low and moderate RFE values where we have a large number of gauge observations. The poorer fit for the larger RFE values stems from too few gauge observations to get an accurate distribution. A slightly poorer result is shown in September, but as a first general assumption this appears to be a reasonable choice. The expression for the variance is derived in Grimes et al. (1999) and stems from the knowledge that the rainfall variance is heteroscedastic and increases with the rainfall amount.

The aim in this chapter is to demonstrate that a lognormal distribution, with suitably defined parameters, better models the conditional gauge measurements compared to the gamma distribution. In particular, it will demonstrate a significant improvement in representing the heavier rainfall amounts, resulting in a more realistic distribution of normal scores.

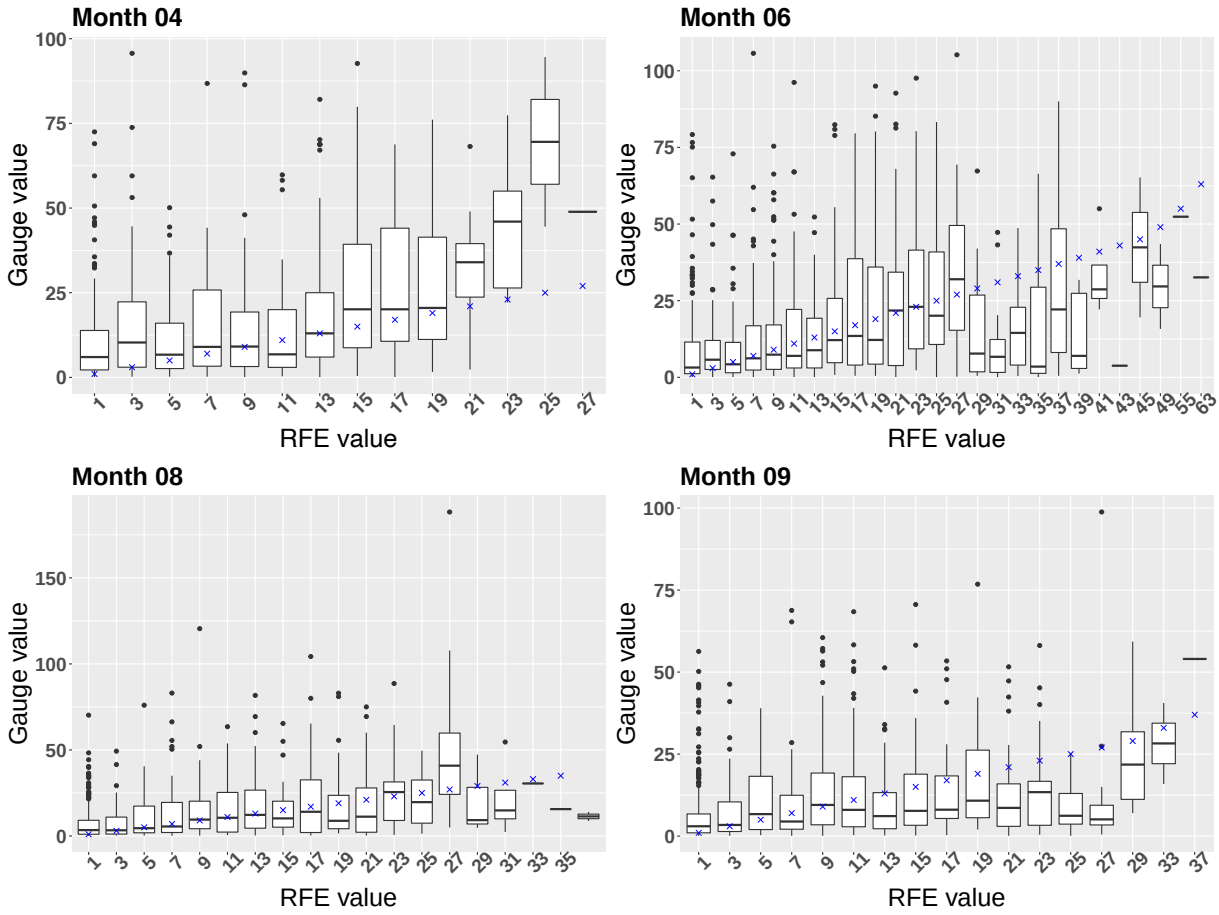


Figure 4.5: Box and Whiskers over observed gauge values given the RFE value in 2 mm bins. The blue crosses marks the RFE value.

4.3.1 Distribution functions

It is well known that rainfall follows a skewed distribution since it is capped from below at 0 but unbounded from above, and lower amounts are more common than heavier. Several different skewed distributions are commonly used to model rainfall, however in most cases the aim is to model the full rainfall distribution of all the observations from a single gauge. Here the aim is instead to model some conditional distribution, hence only a subset of the observations, which therefore might not follow the same type of distribution. Below is a short description of the gamma and the lognormal distribution that are considered here.

Gamma

The gamma distribution is defined by:

$$f(x) = \frac{x^{k-1}e^{-x/\beta}}{\Gamma(k)\beta^k}, \quad x \in (0, \infty)$$

with mean and variance given by:

$$\mathbb{E}[X] = k\beta$$
$$\mathsf{Var}[X] = k\beta^2$$

where $k > 0$ is the shape and $\beta > 0$ the scale parameter and $\Gamma(k)$ the gamma function previously defined. These two parameters control the spread and position of the peak of the distribution with the mode given by $(k-1)\beta$ for $k > 1$ and else undefined.

Lognormal

Y is said to have a lognormal distribution with mean μ and variance σ^2, if $X = \log(Y)$ follows the distribution $\mathcal{N}(\mu, \sigma^2)$. The reason for defining the distribution through the corresponding normal distribution instead of on the native scale, is the much easier methods of estimating and interpreting the parameters associated with the normal distribution.The lognormal random variable Y has probability density function f given by:

$$f(y) = \frac{1}{y\,\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log y - \mu)^2}{2\sigma^2}\right), \quad y \in (0, \infty)$$

The expected value and variance of Y are, respectively, given by:

$$\mathbb{E}[Y] = \exp\left(\mu + \frac{\sigma^2}{2}\right)$$
$$\mathsf{Var}[Y] = \left[e^{\sigma^2} - 1\right] e^{2\mu+\sigma^2}$$

The lognormal distribution has a heavier tail compared to the gamma distribution, but is however still classified as a light-tailed distribution.

4.4 Evaluate the amount distribution fit with southern Ghana as case study

To evaluate the improvement in using the heavier tailed lognormal distribution compared to the gamma, and also the values for the variance parameters κ, θ, the gauge data set over Ghana introduced in Section 3.2.2 will be used. This data set consists of 590 gauges, with a majority of them located in the southern half of Ghana, and some dating back to 1940. Only gauges from the southern region will be used because of its better data availability and only the year 2008 to keep the rest of the years as verification data. This was a relatively wet year and therefore has a large number of sample points, resulting in a more robust analysis. It

additionally has a large number of heavier rainfall events, the representation of which this work aims to improve. Due to the region being mostly dry from November-February and only small amounts of rain during March, the performance evaluation will be done over the monsoon months April-October (see Section 3.3.1 for a full description of the rainfall climatology).

For the RFE values, TAMSAT v3.1 data will be used. This is based on the same algorithm as v3.0, of which a full description is provided in Maidment et al. (2017), with additional gauges used for the CCD calibration and allowing the minimum temperature extend to -65°C.

To evaluate the change in performance when using the lognormal instead of the gamma distribution, a range of qualitative diagnostic plots will be used. The first set of plots are histograms and Quantile-Quantile plots (QQ-plots) to evaluate the distribution fit for the log-normal distribution and how it compares to the gamma distribution. Several parametrisations of the lognormal distribution with the same expression for the mean, but different combinations of κ, θ for the variance defined by Equation (4.4), are included for assessing the goodness-of-fit. The third diagnostic plot is a scatter plot of the normal scores for the gauge values as a function of the RFE value. This aims to demonstrate the improved fit with the lognormal distribution, especially for larger RFE values, with a substantial reduction of normal scores larger than 3.

From some initial analysis, it was clear that the gamma distribution was not a good fit for any combination of κ and θ. It was further clear that the choice of $\kappa = 1.19, \theta = 0.67$, derived by the method outlined in Grimes et al., 1999, with the lognormal distribution resulted in a too small variance for low RFE values but a good fit for the larger values. Three other combinations of parameters have therefore been evaluated, chosen so that Equation 4.4 returns similar values as for the choice $\kappa = 1.19, \theta = 0.67$ for large RFE values (y) but a range of values for smaller RFE (y) values. Figure 4.6 shows the standard deviation to be used, calculated from the square root of Equation 4.4, as a function of RFE (y). The initial choice of $\kappa = 1.19, \theta = 0.67$, is in blue and the three new combinations in orange, magenta and green. This clearly shows how the different choices of the parameters returns different values for the standard deviation to be used in the lognormal distribution for smaller RFE values and near equal for large values $(> 30$mm). The standard deviation value is for the normal distribution associated with the lognormal distribution, to be fitted on the log transformed amounts.

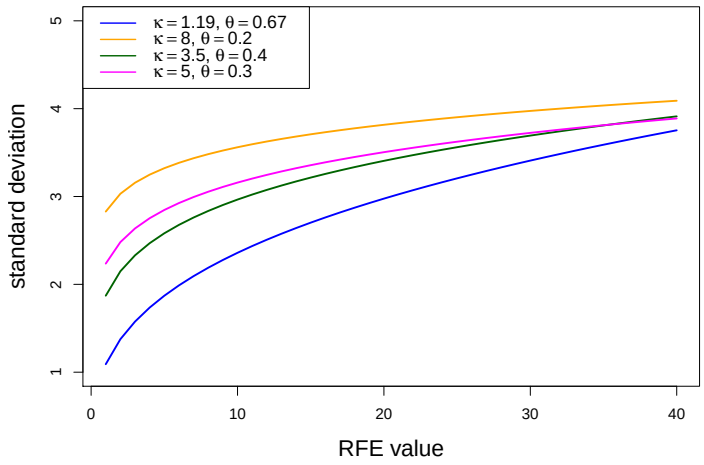


Figure 4.6: Standard deviation for the corresponding normal distribution for each non-transformed RFE value up to 40mm.

4.4.1 Histograms and QQ-plots

The first set of tools to asses the fit to the two distribution functions and the four different variance parameter combinations, are histograms and QQ-plots. Since there is a strong seasonal cycle (see Section 3.3.1), the data will be split into months to be analysed separately. If for instance one model would fit the peak monsoon months June, July, September but not the less rainy April and August, this would be obscured due to the dominance from the rainier months. With the aim being to find the most suitable conditional density function, with parameters depending on the RFE value, the data points are further split into separate groups depending on their associated RFE value, as visualised in Figure 4.2. To not have too few sample points in each group, a range of 2mm RFE values will be used, as done in Figure 4.5. For all graphs, the RFE value printed is the median in the 2mm range.

In histograms, the bin width must be decided and is a trade off between choosing a small enough width to accurately capture the different frequencies for different values, and a too small width resulting in too few samples within each bin, and thereby ending up with essentially a uniform distribution. Two set of histograms will be presented, both with the gauge measurements kept in their native scale and with a log transform. For the original measurements, the bins are 5mm wide and for the log transform log(2)mm wide. On top of each histogram, the assumed density functions are plotted. By plotting the densities on the log transformed scale, we can easily see how the lognormal distribution behaves like a normal distribution. For the gamma distribution, only the variance parameters corresponding to the blue line in Figure 4.6

are used, since it was concluded that it provided a poor fit for any combination of parameters as is therefore only included here to demonstrate the difference to the lognormal distribution. For the lognormal curves, all four sets of parameter choices are included with the colour of the density line matching the variance parameter colour. For all density functions, the mean is set to be the RFE value and therefore for the log transformed scale this is set to be the logarithm of the RFE value.

The QQ-plot maps the quantiles of a sample to the quantiles of a theoretical distribution model, and a linear relation of these points indicate that the sample distribution can be modelled by the theoretical distribution model. For a sample $x_1, .., x_n$, the unknown sample quantiles are replaced by their empirical quantiles defined at each plotting position p. There are numerous choices for the plotting positions, with the most straightforward being:

$$p \in \left\{ \frac{1}{n}, \frac{2}{n}, ..., \frac{n-1}{n}, 1 \right\}$$

To avoid overflow problems at $p = 1$, one might instead use

$$p \in \left\{ \frac{1}{n+1}, \frac{2}{n+1}, ..., \frac{n-1}{n+1}, \frac{n}{n+1} \right\}$$

In this work, the latter choice of $p_{i,n} := i/(n+1), i = 1, ..., n$ will be used, and so the sample-quantile coordinate is given by $(p_{i,n}, x_{i,n})$. The corresponding coordinate values from the theoretical distribution model can be obtained through the associated distribution function F, and are therefore given by $(p_{i,n}, F(p_{i,n}))$. Hence the plotting positions in the QQ-plot are given by $(F(p_{i,n}), x_{i,n})$ and a correctly specified distribution should result in a linear pattern.

Through these scatter points a straight line can now be fitted by more or less robust methods. The simplest method is to select two points, commonly the first and third quartile (i.e. $p = (0.25, 0.75)$) and draw a straight line through these two. A more suitable method is to fit the line by the means of linear regression on the scatter points. Through the classical least-squares algorithm, the slope and intercept is obtained by minimising the sum of squares

$$\sum_{i=1}^{n} (x_i, n - a\ F(p_{i,n}))^2$$

If the theoretical distribution is wrongly specified, the sample points will deviate from the straight line due to different tail behaviour. Further, if the theoretical distribution model is correct but the parameters differ between the sample and the theoretical model, the sample points will fall on a straight line but the values $x_{i,n}$, $F(p_{i,n})$ will differ. This is due to the

proportion of the density function for each plotting position is constant for a given distribution model, but the value associated with it is different. A simple example is the normal distribution where a change in the mean from 0 to 2, will result in an equal shift in the value of $F(0.5)$, but the shape of the distribution and therefore the quantile spacing is the same. Figure 4.7 demonstrates the QQ-plot patterns associated with correctly assumed model but difference in mean or variance between the sample points and the theoretical distribution. Here the model is assumed to be normally distributed, but this would be true for any distribution. The red line is the linear regression fitted line and the black is the $x = y$ diagonal line.

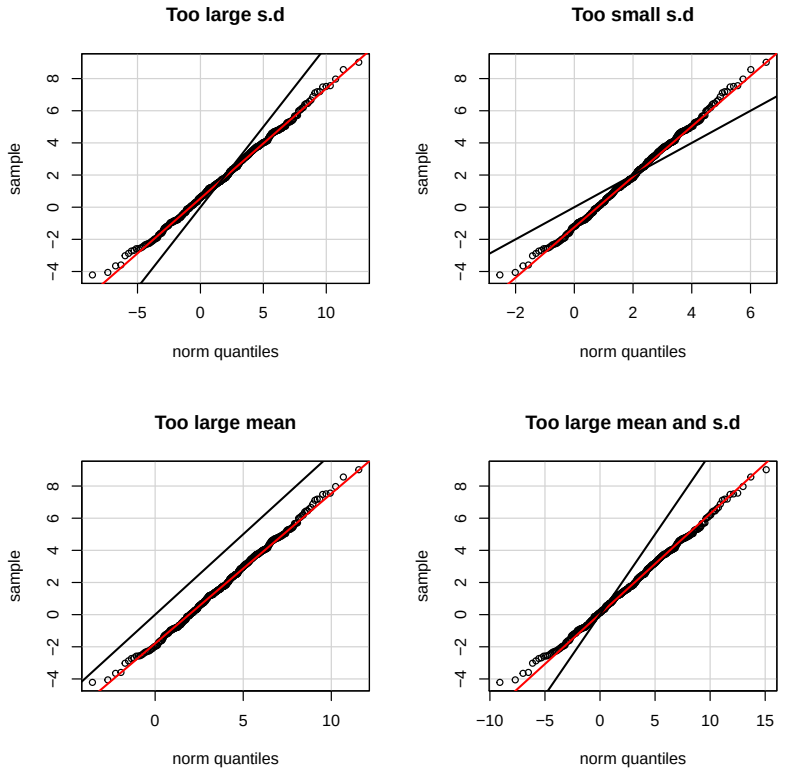


Figure 4.7: Patterns for correctly assumed model but difference in parameter values in QQ-plots. The title refers to the theoretical distribution in relation to the sample points.

Only a selection of the months, specifically April, June, August and September, will be presented here to demonstrate the fit for the different phases of the monsoon. Further, only a selection of the 2 mm RFE subsets will be presented to demonstrate the fit of both the gamma and lognormal distribution for low, moderate and large RFE values. The full range of RFE subsets with the lognormal fit are provided in Appendix B. Since we defined the lognormal distribution in terms of its associated normal distribution, we will present the results for the log transformed rain gauge values and their fit to the normal distribution. To simultaneously compare the fit for the different combinations of variance parameters κ, θ, the linear regression

fitted lines for the different choices will be presented in the same graph. That is, for a given RFE value one can get the corresponding standard deviation value for each set of κ, θ from Figure 4.6. For each of these variance values, a separate linear regression fitted line will be obtained for that set of scatter points $(x_{i,n}, F_{\kappa,\theta}(p_{i,n}))$. All of these fitted lines are then added to the same QQ-plot, with the theoretical quantiles given by the parameter pair $(\kappa = 8, \theta = 0.2)$ (orange line) and the black line is again the $x = y$ line.

Histograms

It is clear from the histograms in Figures 4.8 - 4.11 that the gamma distribution (dashed black line) is a much too light tailed distribution, with a large part of the observations found outside the tails . The lack of spread in the density curve also leads to the much higher mean peak than what is observed. This will as previously described result in a very high proportion of gauge measurement being assigned normal scores over 2, when if the distribution is correctly assumed should occur for about 2.5% of the observations. For the lognormal distribution (coloured solid lines) a significantly improved fit for all choices of the variance parameters κ, θ can be observed, especially for the moderate RFE values. In general the peak of the observations align with the density peaks for smaller RFE values, confirming that the expected value of our transformed normal scores should be 0. However for larger RFE values, the assumed mean is larger than the observed which leads to more negative normal scores than expected. For the higher RFE values (bottom rows), there tends to be a slight negative skew in the log transformed gauge values, which further increases the number of negative normal score values compared to positive. This is nevertheless still a large improvement compared to the gamma distribution, and a reasonably good fit considering that we are seeking a 'one-distribution-fits-all' for a wide range of cases.

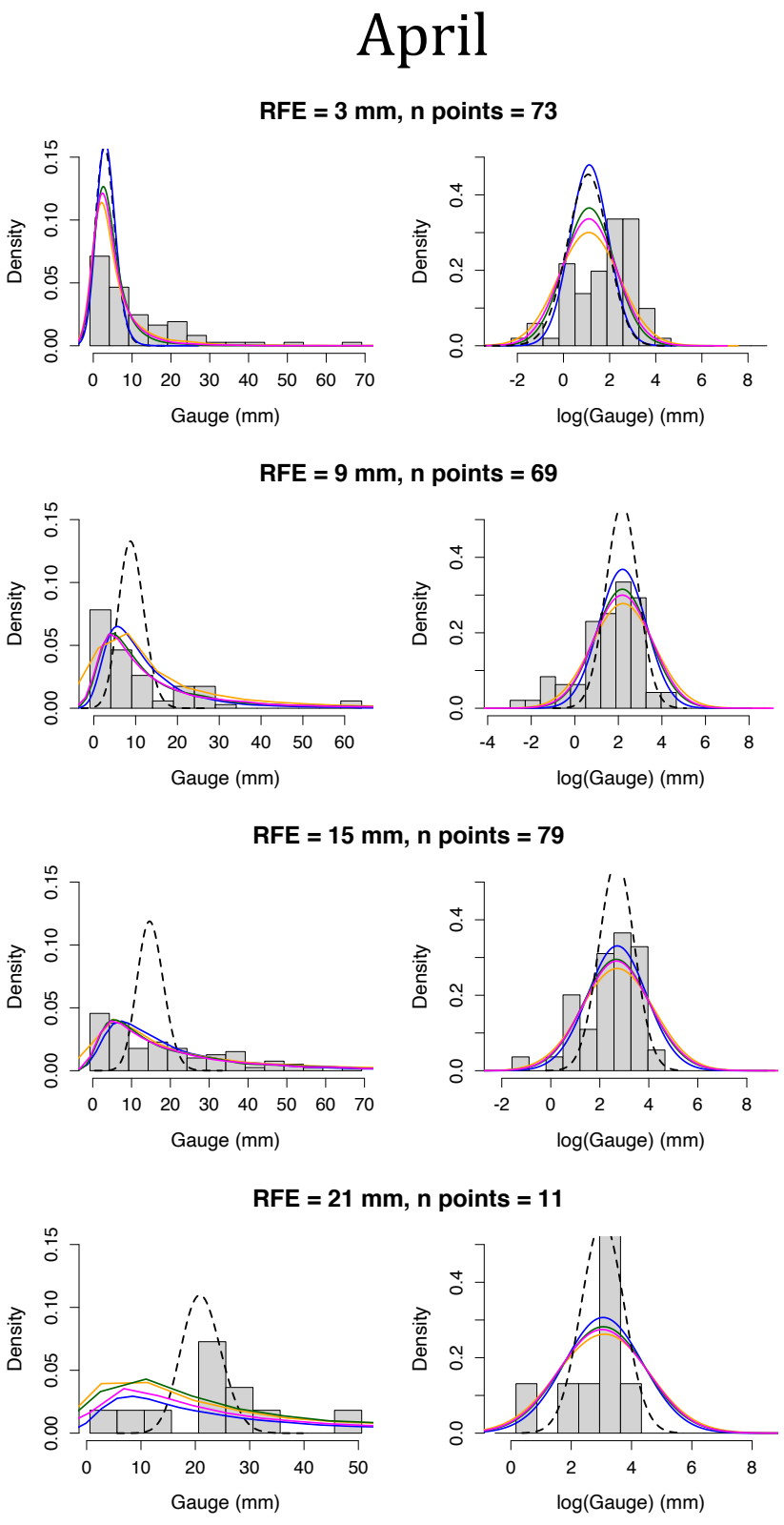


Figure 4.8: Histograms for a selection of RFE values in April, with median RFE value and number of points indicated. (Left) gauge values in 5mm wide bars, (right) logarithm of the gauge values in log(2)mm wide bars. Black dashed line is the gamma density and the coloured to the lognormal, with colour indicating the parametrisation defined in Figure 4.6.

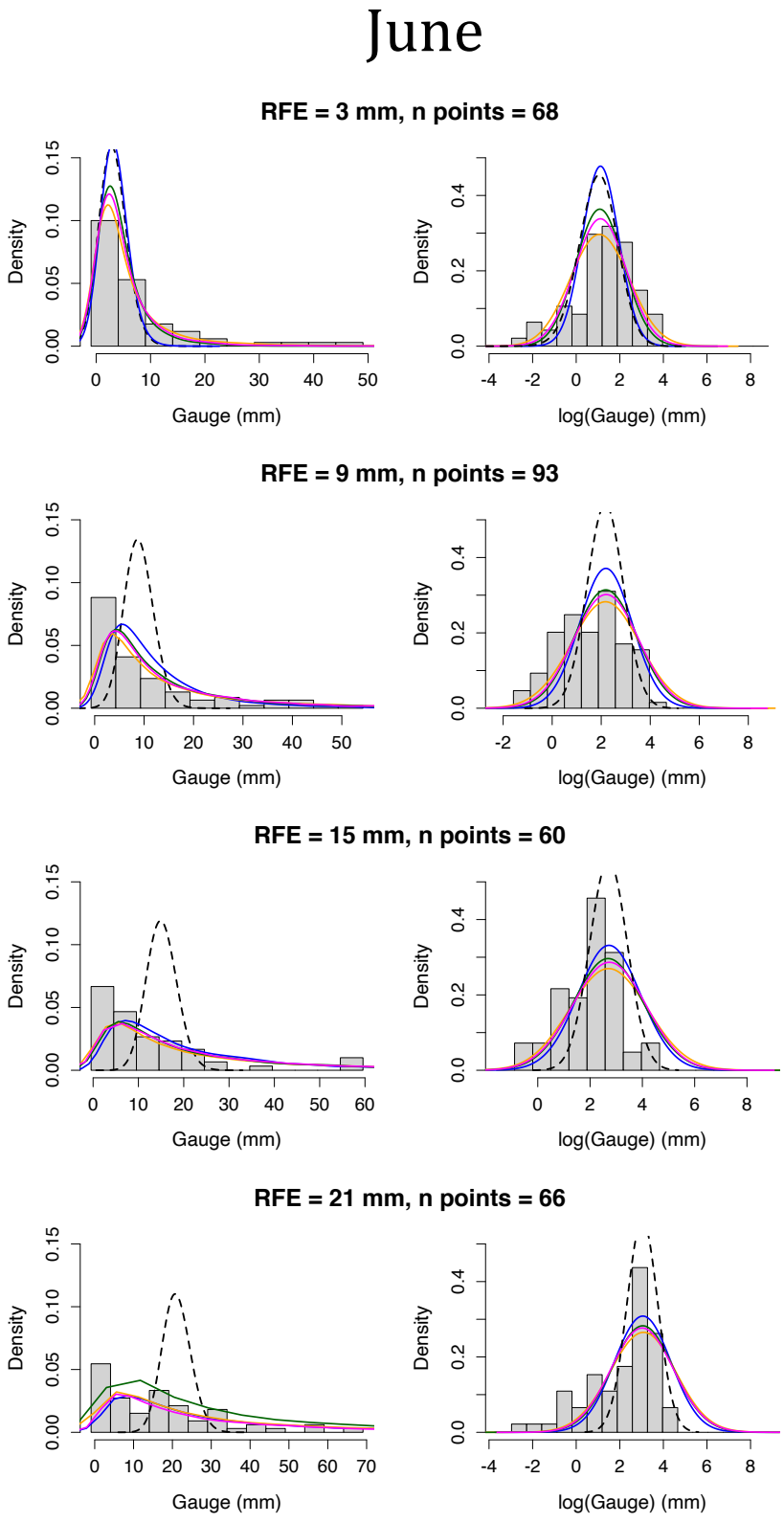


Figure 4.9: Histograms for a selection of RFE values in June, with median RFE value and number of points indicated. (Left) gauge values in 5mm wide bars, (right) logarithm of the gauge values in log(2)mm wide bars. Black dashed line is the gamma density and the coloured to the lognormal, with colour indicating the parametrisation defined in Figure 4.6.

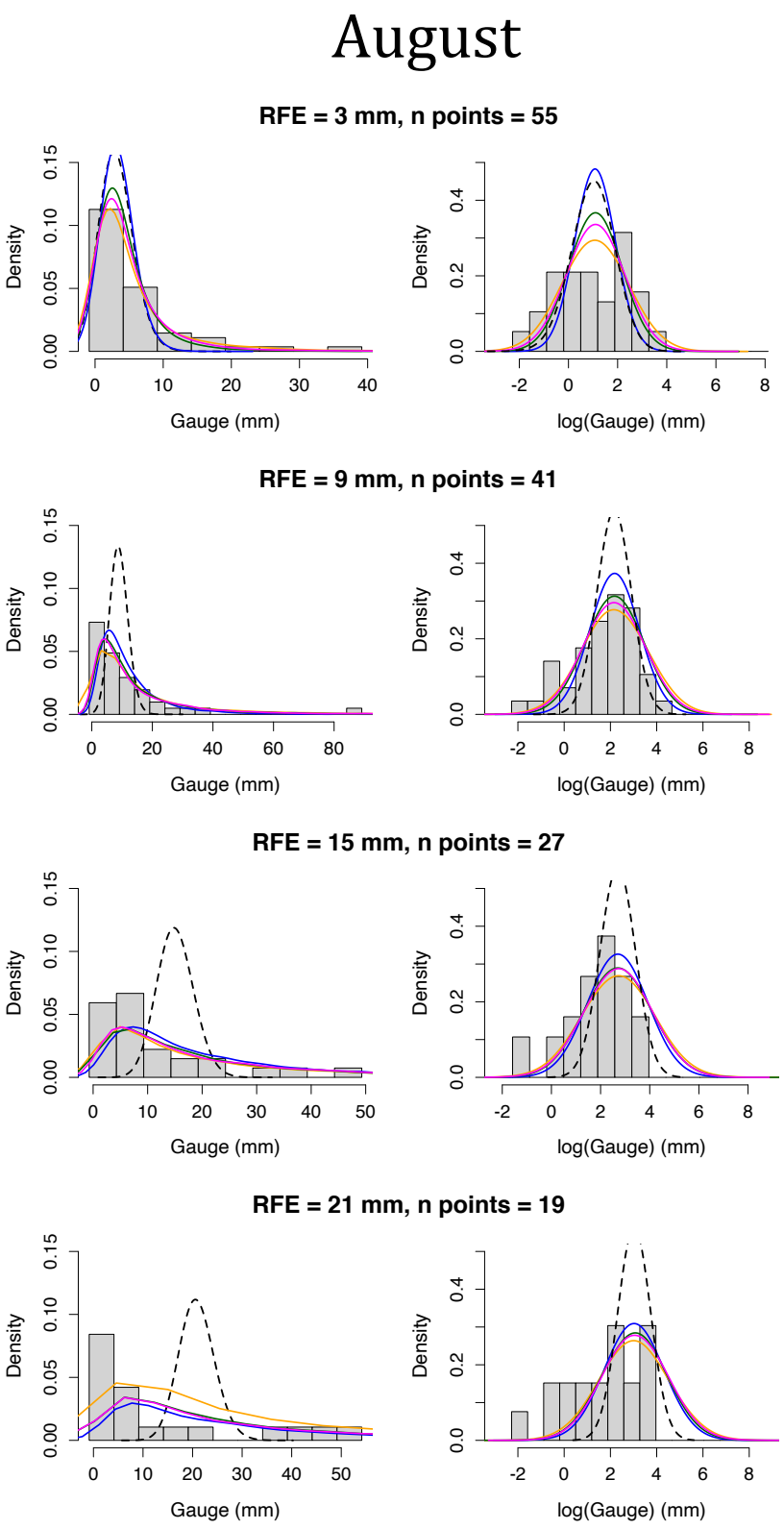


Figure 4.10: Histograms for a selection of RFE values in August, with median RFE value and number of points indicated. (Left) gauge values in 5mm wide bars, (right) logarithm of the gauge values in log(2)mm wide bars. Black dashed line is the gamma density and the coloured to the lognormal, with colour indicating the parametrisation defined in Figure 4.6.

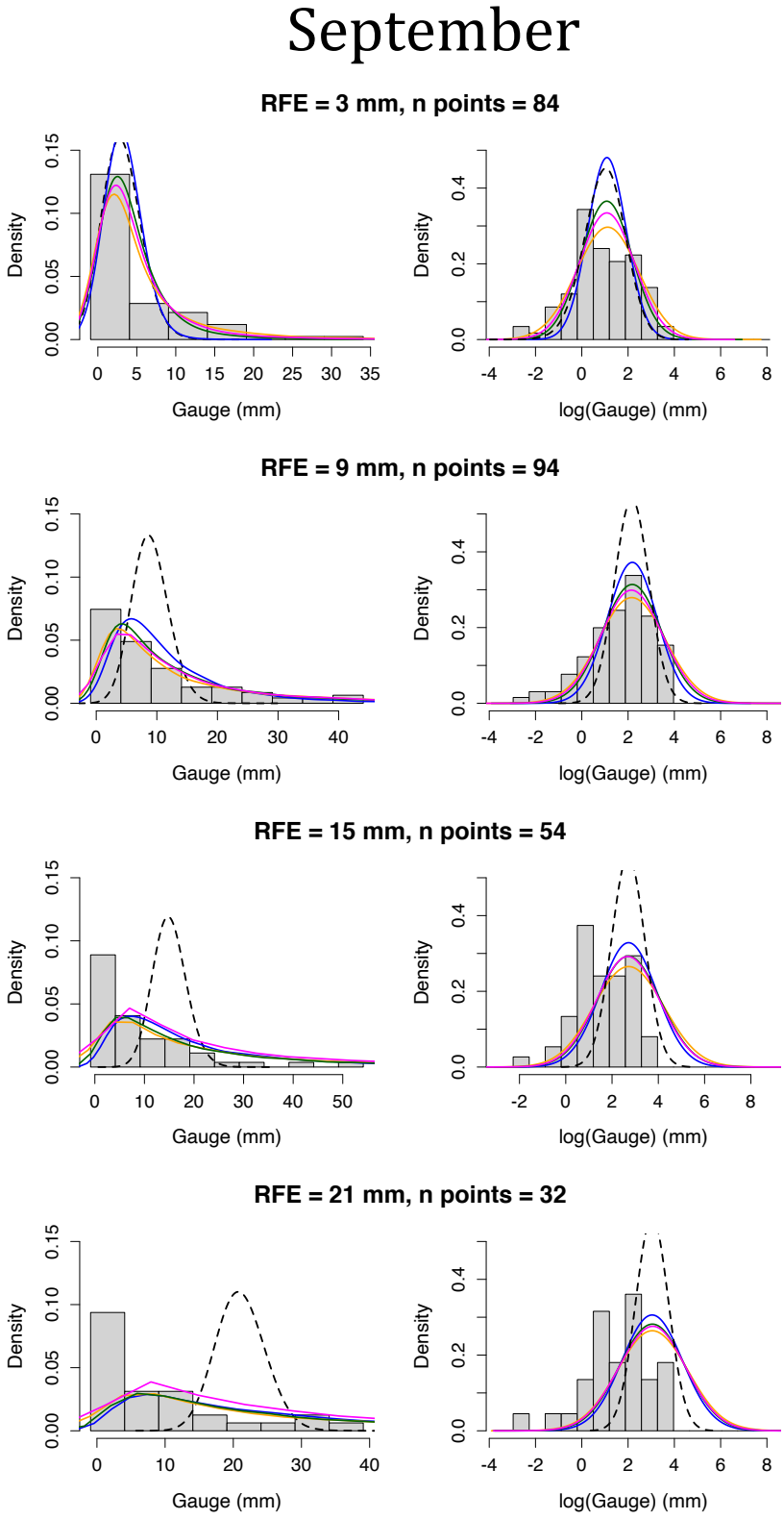


Figure 4.11: Histograms for a selection of RFE values in September, with median RFE value and number of points indicated. (Left) gauge values in 5mm wide bars, (right) logarithm of the gauge values in log(2)mm wide bars. Black dashed line is the gamma density and the coloured to the lognormal, with colour indicating the parametrisation defined in Figure 4.6.

QQ-plots

From the QQ-plots in Figure 4.12 - 4.15, with the normal quantile mapped to the log-transformed gauge values at the top and the gamma distribution mapped to the non-transformed gauge values at the bottom, we can once again see the large improvement in using the lognormal distribution. Most of the observations fall in a straight line, confirming the suitability of the assumed theoretical distribution, whereas the points fall on a skewed curve in the gamma plot, demonstrating the heavier tails in the observations compared to the theoretical distribution. In the lognormal plots, one can clearly see the improvement when parametrising the variance with a larger κ, especially for the smaller RFE values. For the smaller RFE values, the blue line fitted through the points $(x_{i,n}, F_{\kappa=1.19,\theta=0.67}(p_{i,n}))$ has a much steeper slope compared to the $x = y$ black line, which from Figure 4.7 we know corresponds to a smaller variance in the theoretical distribution compared to the sample. This is a pattern that is not present for the other combinations. For the larger RFE values, the difference between the different parameter choices are very small which is due to the close to equal values for the standard deviation values in Figure 4.6, and the mean is equal for all distributions (μ =RFE). Since the performance of the different parameter choices are equally good (or bad) for all of them, the optimal choice for κ, θ will therefore be determined based on the smaller RFE values performance. An important thing to remember is that we need one model and set of parameters for modelling all RFE intensities since we cannot fit a separate model for each, so the aim is to find the overall best performing model and not the optimal choice for each individual RFE value.

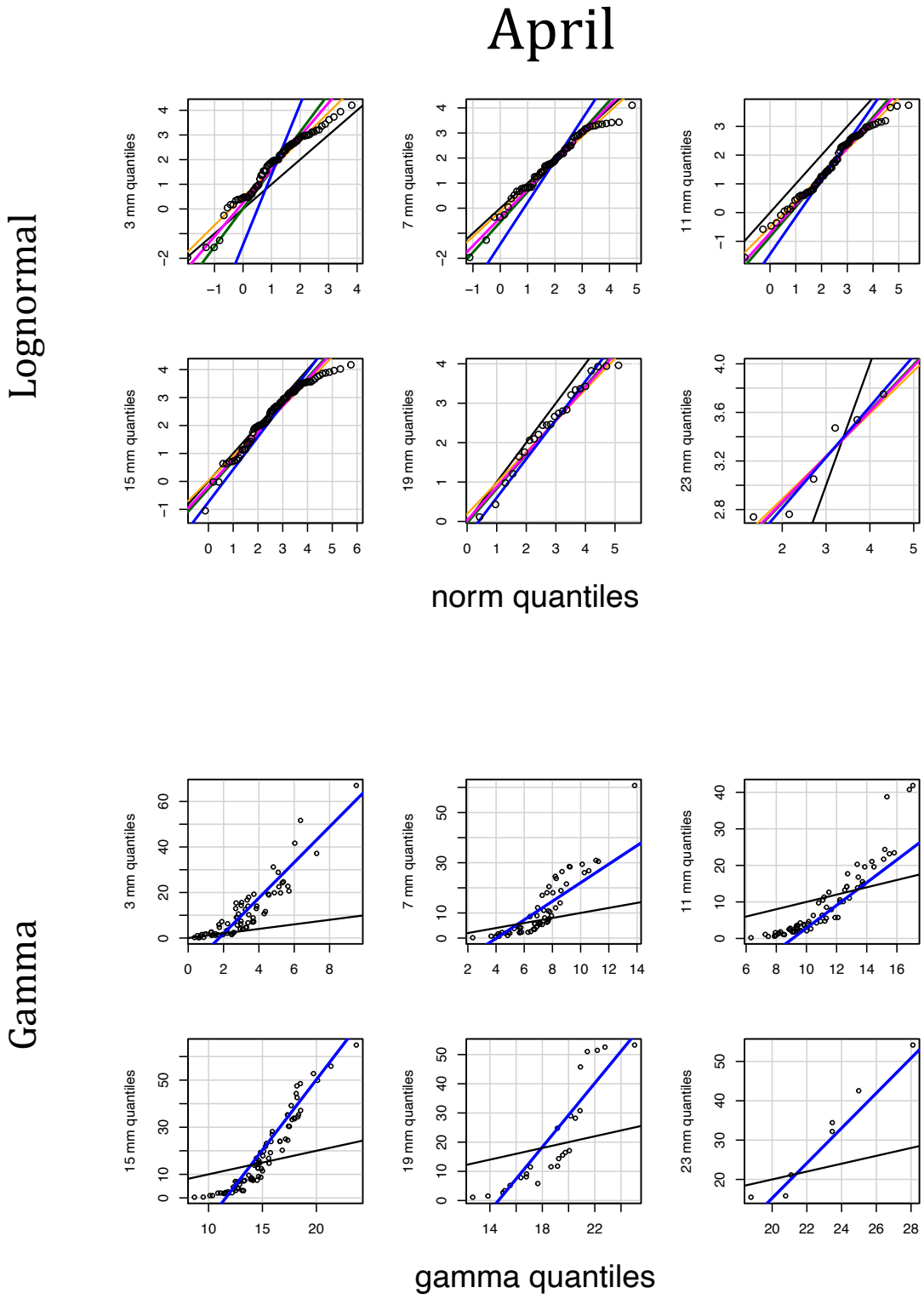


Figure 4.12: QQ-plot for a selection of RFE values in April. (Top) logtransformed gauge values and lognormal distribution associated with $\kappa = 8, \theta = 0.2$ as reference distribution. Coloured lines represents best linear fit to the lognormal distributions with the standard deviation given by the corresponding colours in Figure 4.6. (Bottom) gauge values with the gamma distribution as reference distribution. The black line marks the $x = y$ line.

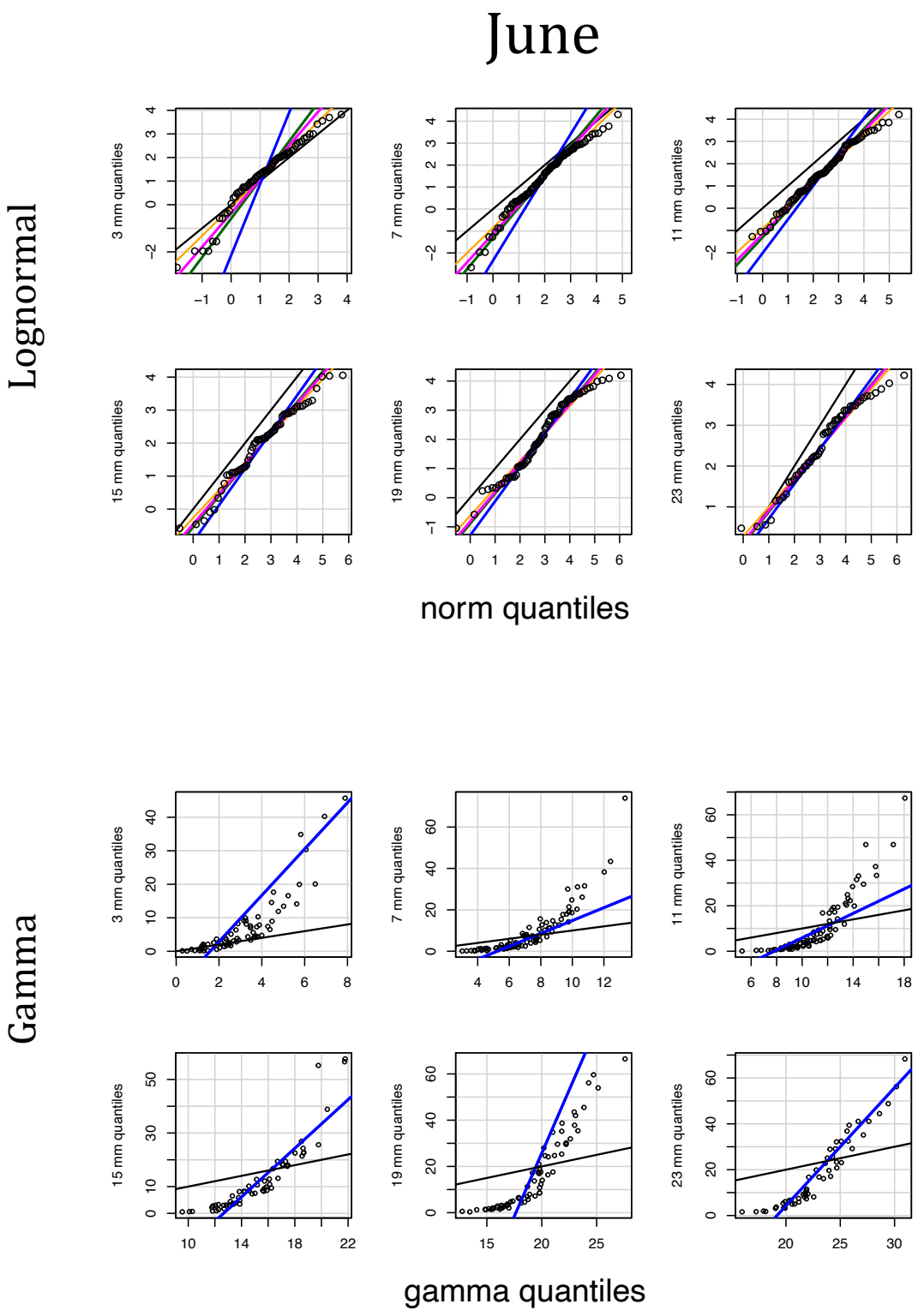


Figure 4.13: QQ-plot for a selection of RFE values in June. (Top) logtransformed gauge values and lognormal distribution associated with $\kappa = 8, \theta = 0.2$ as reference distribution. Coloured lines represents best linear fit to the lognormal distributions with the standard deviation given by the corresponding colours in Figure 4.6. (Bottom) gauge values with the gamma distribution as reference distribution. The black line marks the $x = y$ line.

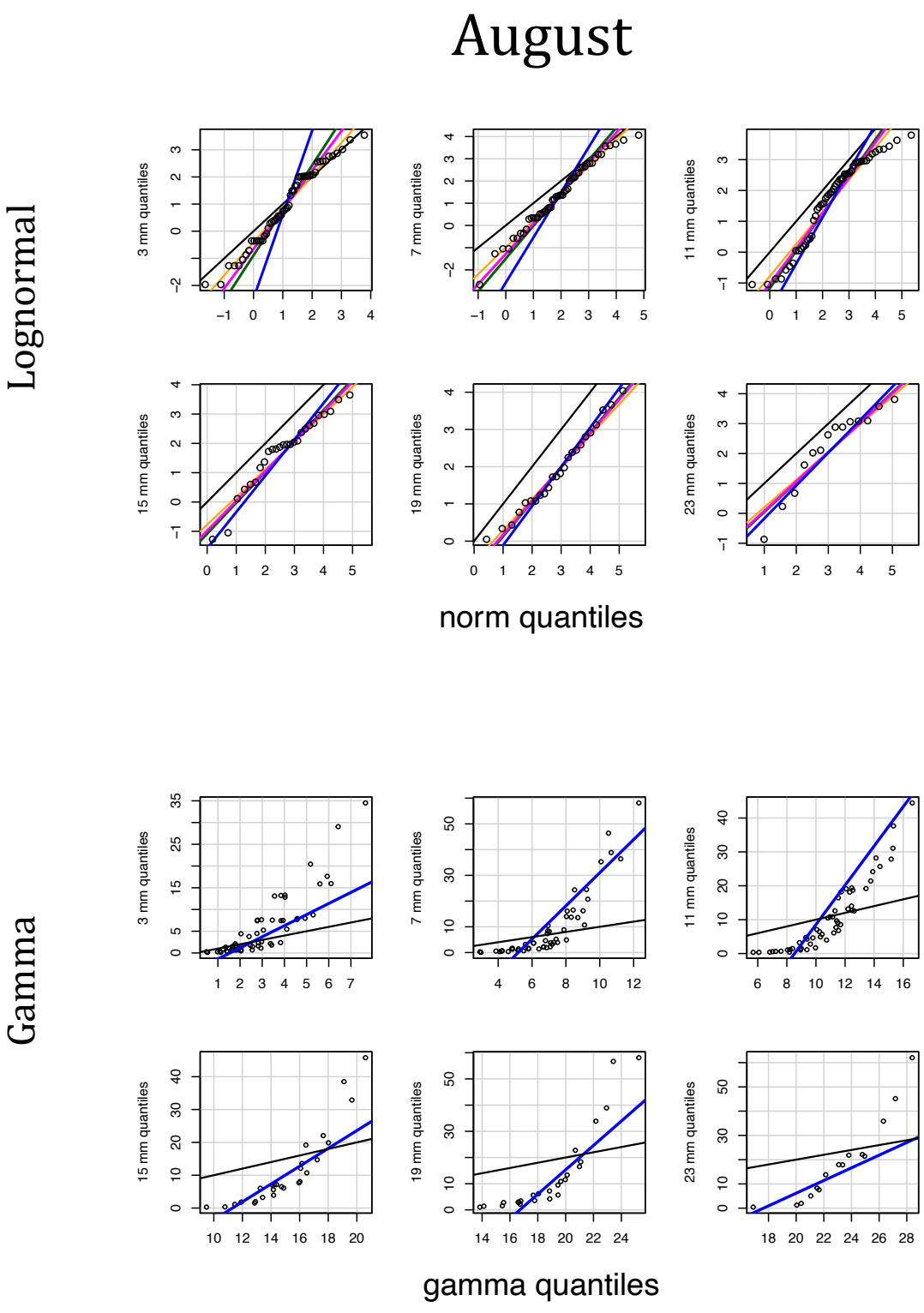


Figure 4.14: QQ-plot for a selection of RFE values in August. (Top) logtransformed gauge values and lognormal distribution associated with $\kappa = 8, \theta = 0.2$ as reference distribution. Coloured lines represents best linear fit to the lognormal distributions with the standard deviation given by the corresponding colours in Figure 4.6. (Bottom) gauge values with the gamma distribution as reference distribution. The black line marks the $x = y$ line.

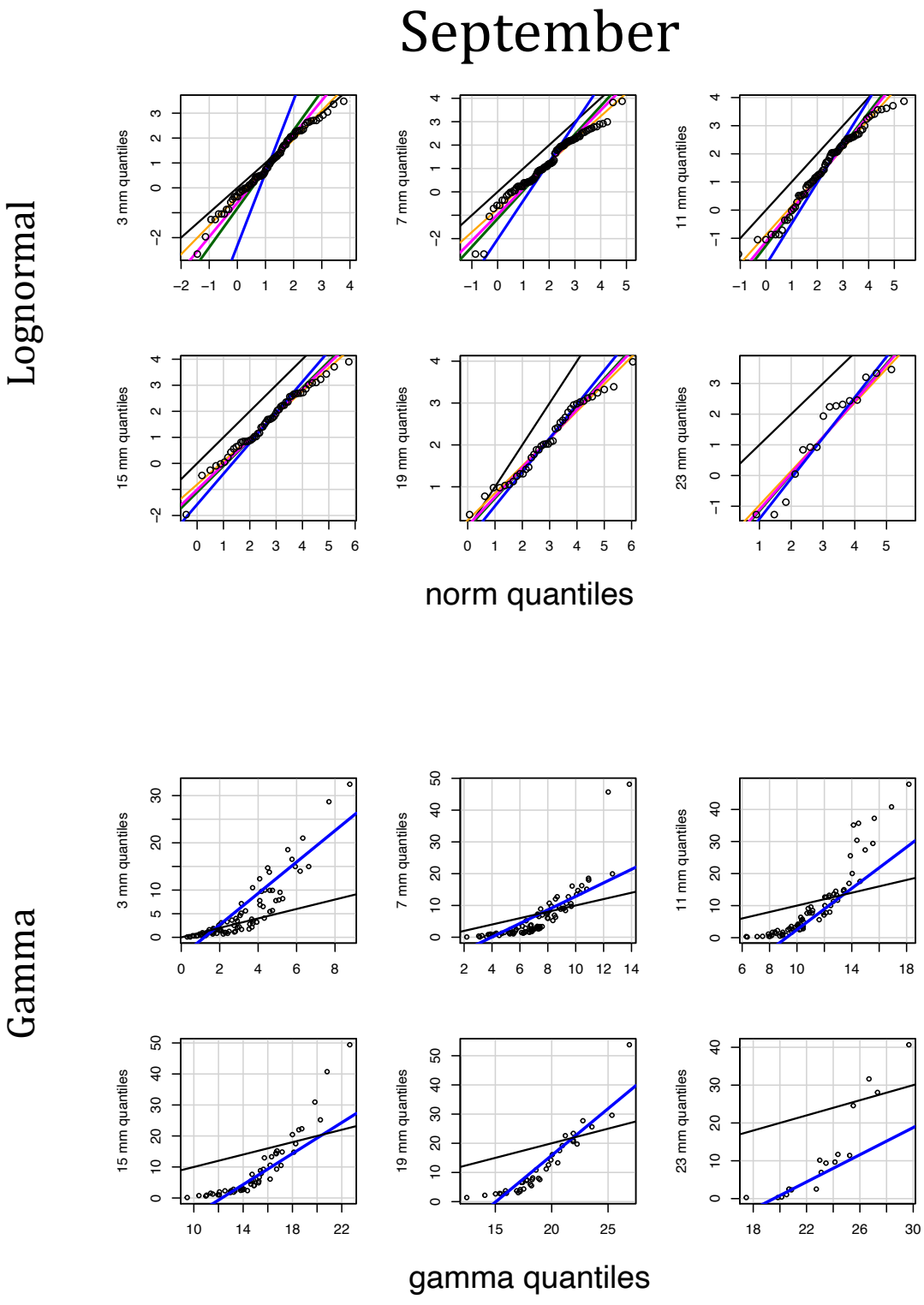


Figure 4.15: QQ-plot for a selection of RFE values in September. (Top) logtransformed gauge values and lognormal distribution associated with $\kappa = 8, \theta = 0.2$ as reference distribution. Coloured lines represents best linear fit to the lognormal distributions with the standard deviation given by the corresponding colours in Figure 4.6. (Bottom) gauge values with the gamma distribution as reference distribution. The black line marks the $x = y$ line.

4.4.2 Scatter plots

The most important aspect of getting the distribution function correct is to make sure that the resulting normal scores are sensibly and approximately normally distributed to not skew the satellite rainfall field too often (incorrect mean) or too much (incorrect variance). The following scatter plots are therefore demonstrating the normal scores for all gauge measurements as a function of the RFE values, including values larger than the ones presented in the previous section. From the above analysis, the variance setting with $\kappa = 8$ and $\theta = 0.2$ was deemed the most suitable, hence will be used here. The values when using the gamma distribution are marked with the lighter in colour triangles and the lognormal the darker circles. Orange or red markers have a normal score of more than 3, which should be highly unlikely to observe if accurately modelled, and the dashed line marks the value of ± 2 where we expect to observe 95% of the values for a normal distribution.

The first sign of the improvement in changing to a lognormal distribution in Figure 4.16 is the near elimination of normal scores larger than ± 3. A few positive large values can be observed for very low (< 5mm) RFE values, and some on the negative side for larger RFE values, which is to be expected as discussed based on the skewed histograms. The other is the good spread of points between ± 2, indicating that we have chosen a sensible function for the variance which constraints the normal scores without condensing them.

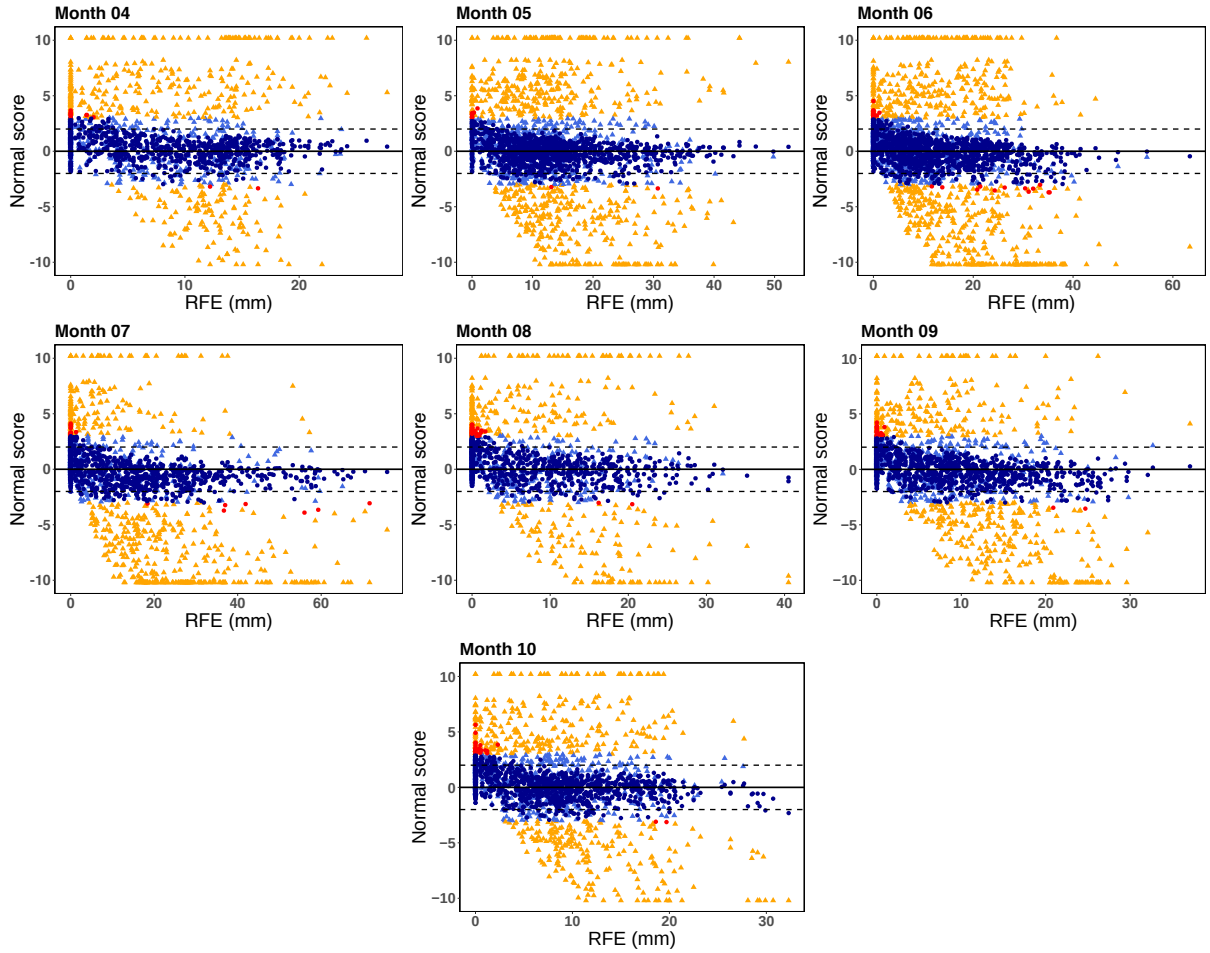
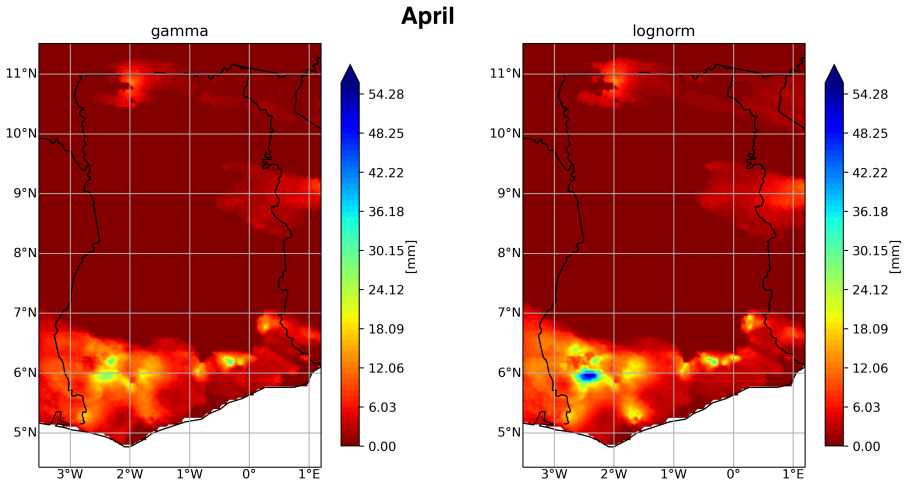


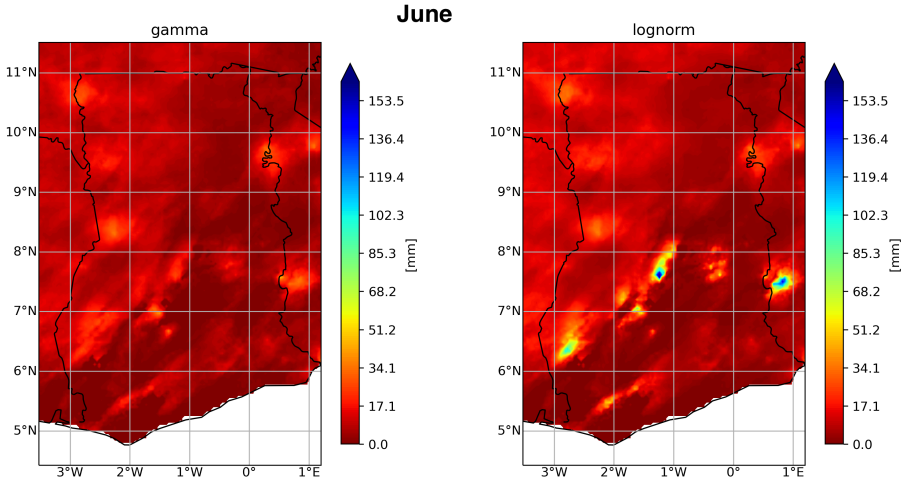
Figure 4.16: Scatter plots of lognormal (dark, circles) distribution with standard deviation corresponding to the orange line in Figure 4.6 and gamma (light, triangles) distribution for the gauge values. The red/orange dots have a normal score of larger than 3 and the dashed lines marks a normal score of ±2. All values larger than ±10 are capped to this.

4.5 Comparison in performance for the full grid

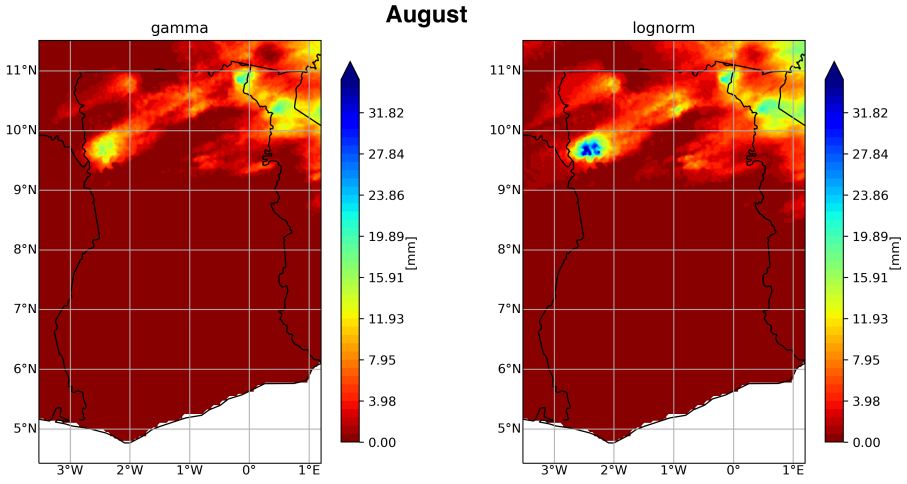
To demonstrate the impact from changing the distribution, a few examples from the merging algorithm are presented in Figure 4.17. Examples from April, June and August, are included to display the improvement for the full range of intensities. In April (top) the gauge measuring around 54mm is included with the lognormal transform but most likely capped out from the gamma due to a too large normal score, as can be noticed from the lack of grid values above 30mm. For June (middle), even the extreme values of 150mm are merged with the lognormal, which the gamma is far from achieving, with no grid values going above around 40mm. The August graph (bottom) highlights the improvement even for moderate amounts where the merged product using the lognormal distribution produces grid values of up to 32mm but gamma only 16mm, showing that the impact can be seen for all intensities.



(a)



(b)



(c)

Figure 4.17: Preliminary plots of rainfall estimates from merging TAMSAT estimates with gauge data using the (left) gamma and (right) lognormal distribution. The results are for single days in (a) April, (b) June, (c) August. Plots produced by Dr. Ross Maidment.

4.6 Discussion and further work

In this chapter, evaluation of the improvement in modelling the conditional rain gauge observations for a given RFE value with the lognormal distribution compared to the gamma distribution is presented. The motivation came from the frequently occurring very large normal scores in the TAMSAT merging algorithm with the latter distribution, indicating that the gamma distribution did not accurately model the observed gauge values. It further was not possible to use a mixed distribution with an Extreme value distribution for the tail due to the back transformation step (E in Figure 4.1), which requires a continuous distribution function. Even though the lognormal distribution is not easily interpretable in its native form, it can easily be applied after a logarithmic transform on the measurements, making it a distribution that one can easily switch to.

From the variance parametrisation analysis, done by simultaneously comparing several choices for the parameters κ, θ, we can see that even though there is some heteroscedasticity in the conditional rainfall values this is not very pronounced. The most suitable set of parameters, $\text{Var}[X|\text{RFE}] = 8 * \text{RFE}^{0.2}$ has a large and nearly constant value for the low RFE values as well as the larger.

An issue with working in the logspace is the non physical behaviour for very small rainfall values. Since $\log(1) = 0$ and negative for values smaller than that, a negative variance will be returned for very small RFE values. This can however relatively easy be solved by imposing a fixed variance for RFE values below this, and apply the RFE dependent function for values above.

4.6.1 Further work

As we in this work were interested in finding a relatively simple distribution that performed better than the gamma, many other skewed distributions were determined to be beyond the scope of this thesis. Further work would include a wider distribution comparison analysis, including skewed distributions such as Burr and Log-gamma, which are more commonly used to model economics data such as household incomes, or the log-logistic which previously has been used to model streamflows and rainfall.

An extension to the work presented here would be to complement the qualitative analysis with quantitative analysis through the use of goodness-of-fit test. This could be utilised to compare the different choices of κ and θ, but also to evaluate the most suitable model when comparing the multiple skewed distributions mentioned above. To evaluate the current lognormal distribution choice, a collection of the many tests for detecting deviations from a normal distribution would be used since they all have their individual drawbacks. A full review of the

power of 33 normality tests was undertaken in Romão et al. (2010) for symmetric, asymmetric and tainted normal distributions. From our results here, a test with a high power for asymmetric distributions would be most appropriate. In the paper they conclude that the test with the best power for asymmetric distributions are Z_A, Z_C *Zhang-Wu* tests, CS *Chen-Shapiro* and W *Shapiro-Wilk* test. Due to the close relation between the latter two, only the second would be used as it has the slightly higher power.

With a separate test needed for each RFE value distribution and month, around 100 individual tests would be performed. We would therefore have to account for the *multiple testing issue*, i.e. as the number of test increases the probability of observing false-positives (Type I error) is increased. A commonly used class of methods to adjust for this is to control the Family-wise error rate (FWER), the probability of making at least one Type I error in the family of hypothesis tests. The aim with all of these methods is to adjust the nominal significance level α, which we will denote α_a, such that FWER $\leq \alpha_a$. The most commonly used method is the "Bonferroni" test, which simply divides $\alpha/m = \alpha_a$, where m is the total number of tests. This however becomes very conservative if many tests are used. A different approach to just adjusting α, is to arrange all the hypothesis test values p_i, $i = 1, .., m$ in increasing order and let the critical value depend on the ranked position. "Holm-Bonferroni" is one of the most well-known where the null hypothesis H_0 is rejected until $p_k \geq \frac{\alpha}{m+1-k}$, k being the position among the increasingly ranked p_i. Other examples of methods in this class are "Hochberg and Holm's", "Dunn-Sidak " and "Holm-Sidak" (which combines "Holm-Bonferroni and Dunn-Sidak).

A completely different method to the one above would be to assume that the expected number of rejections can be modelled as discrete, rare events by a Poisson distribution with $\lambda = m * \alpha$, with α, m defined as above. The observed number of rejections would then be tested at the level α_{Po}, where the null hypothesis H_0 is rejected if this is larger than the critical value.

Chapter 5

Estimation and reduced bias estimation of the coefficient of tail dependence

In order to address thesis aim question number 2 fully, a separate method suitable for modelling the extreme values is needed. In this chapter, based on multivariate EVT, an improved method for estimating the association between two variables in the case of asymptotic independence will be proposed alongside a reduced bias estimator. The performance is evaluated in an extensive simulation study and compared to the well-known Hill estimator. This estimator is used in Chapter 6 to investigate the dependence in extremes as a function of distance between the stations.

5.1 Introduction

Multivariate EVT has received an increasing amount of attention in the past two decades, commonly with the aim of determining the probability of joint exccedances above a high threshold. This is often motivated by the fact that compound extremes can pose a much greater risk compared to the two individual extremes occurring separately. In the environmental science, there are numerous examples of where this occurs, both by a combination of variables or locations. Flooding can become significantly more severe if multiple locations around the same river basin experience extreme rainfall simultaneously. Extreme drought is usually a combination of very high temperature and large negative rain anomalies, and extreme coastal flooding is often a combination of high winds and wave height. Since the interest usually lies in modelling the most extreme events, and in extrapolating outside the observed values, EVT

is the most suitable framework to work in.

In bivariate extreme value statistics, the dependence is classified as either asymptotically dependent or asymptotically independent (Sibuya (1960)). The two definitions are essentially separated by if there is a non-zero probability of the two variables being extreme simultaneously, with the two being asymptotically dependent if this is true. The estimator proposed here falls into the asymptotically independent class, motivated by Sang and Gelfand (2009) recommending that such as model should be used for African rainfall. Since the aim of this improved estimator is to more accurately estimate the dependence structure in rainfall over west Africa, this was seen as the appropriate choice. The contribution in this chapter is mainly the improved estimator and the unified marginal distribution.

5.2 Modelling asymptotic independence

5.2.1 Coefficient of tail dependence definition

Like most previous work on multivariate EVT, the focus will be on the bivariate case. Extensions to higher dimensions are often possible in theory but not feasible in practice. Given that (X_i, Y_i), $i = 1, ..., n$ are independent copies of the random vector (X, Y) with joint distribution function F, we are interested in estimating the probability of

$$\mathbb{P}(X_i > u \text{ and } Y_i > v) \tag{5.1}$$

where u, v are large thresholds. We assume that that the marginals, F_X, F_Y, are known and denote the sequence of component-wise maxima by $M_{X,n} := \max_{1\leq i\leq n} X_i$ and $M_{Y,n} := \max_{1\leq i\leq n} Y_i$. We further assume that there exist normalising constants $a_n, c_n > 0$ and $b_n, d_n \in \mathbb{R}$ such that

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{M_{X,n} - b_n}{a_n} \leq x, \frac{M_{Y,n} - d_n}{c_n} \leq y\right) = G(x, y) \tag{5.2}$$

where G is a non-degenerate distribution function. When this holds, G is called a bivariate extreme value distribution. The component-wise maxima $M_{X,n}, M_{Y,n}$ are said to be asymptotically independent if $G(x, y) = G(x, \infty)G(\infty, y) =: G_1(x)G_2(y)$ for all x, y. The given expression of Equation (5.2) is of little use when attempting to estimate (5.1), something that however can be improved if we rewrite it in terms of the distribution function F

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{M_{X,n} - b_n}{a_n} \leq x, \frac{M_{Y,n} - d_n}{c_n} \leq y\right) = \lim_{n\to\infty} F^n(a_n + b_n x, a_n + b_n y)$$

and by further taking the logarithms, we obtain

$$\lim_{n\to\infty} n\mathbb{P}\left(\frac{X-b_n}{a_n} > x \text{ or } \frac{Y-d_n}{c_n} > y\right) = -\log G(x,y)$$

which ultimately, after some rearranging, results in

$$\lim_{n\to\infty} n\mathbb{P}\left(\frac{X-b_n}{a_n} > x, \frac{Y-d_n}{c_n} > y\right) = \log G(x,y) - \log G_1(x) - \log G_2(y). \tag{5.3}$$

The right hand-side in the above is equal to zero if the marginals of the limiting distribution are independent and we are presented with the case of asymptotic independence in extremes. A way to work around the problem of this equalling to zero for all cases of asymptotic independence was derived by Ledford and Tawn (1996) through the introduction of a submodel $t \mapsto q(t) := \mathbb{P}(1-F_X(X) < t, 1-F_Y(Y) < t)$, which is assumed to be regularly varying at 0 with index $1/\eta$. That is $q(t) = t^{1/\eta}\mathcal{L}(t)$ where $\mathcal{L}(t)$ is a slowly varying function, i.e. $\frac{\mathcal{L}(tx)}{\mathcal{L}(x)} \to 1$ as $t \to 0$ for all fixed $x > 0$. The coefficient $\eta \in (0,1]$ is the so called coefficient of tail dependence (CTD), or sometimes referred to as the residual dependence index, where $\eta = 1$ and $\mathcal{L}(t) \to c > 0$ implies asymptotic dependence and $\eta < 1$ asymptotic independence. Heffernan (2000) showed that a majority of the bivariate distribution functions can be written in this format and provides a list with the value of η and the expression for $\mathcal{L}(t)$ for these distributions.

The main purpose of introducing the CTD was to separate out three different types of asymptotic independence, highlighting the various joint behaviour between two variables despite being independent in the limit. The three cases of asymptotic independence identified are:

(i) $\eta \in (1/2, 1)$ or $\eta = 1$ and $\mathcal{L}(t) \to 0 : X, Y$ are positively associated and will exceed a high threshold more frequently than if exactly independent;

(ii) $\eta = 1/2 : X, Y$ are close to independent, with exact independence attained for $\mathcal{L}(\cdot) = 1$;

(iii) $\eta \in (0, 1/2) : X, Y$ are negatively associated and will exceed a high threshold less frequent than if exactly independent.

A visual interpretation of these different cases can be seen in Figure 2.9 in Chapter 2, where the Bivariate normal copula in the top row has $\eta = (0.1, 0.55, 0.995)$. For the first value of η, there is a strong negative correlation between the two variables largest values, for close to 0.5 no preference is seen, and close to 1 a very strong positive correlation for the largest values is present.

If one defines the variable $Z_i = \min\{X_i, Y_i\}$, then since q is regularly varying with index $1/\eta$, we have

$$\mathbb{P}(Z_i > z) = \mathbb{P}(X_i > z, Y_i > z) = z^{-1/\eta}\mathcal{L}(z)$$

The parameter η can therefore be seen as the extreme value index of the minimum of two components, and all the classical estimators, such as the Hill (Hill (1975)), moment (Dekkers et al. (1989)) and maximum-likelihood (Smith (1987)) estimator, are suitable. These estimators are however all biased, but unbiased versions for the univariate case have been presented in for example Feuerverger and Hall (1999) and Caeiro et al. (2005), and for the bivariate case in Beirlant et al. (2011). The estimator in Beirlant et al. (2011) is unbiased but does not take into account the uncertainty arising from the marginal transformation by means of the empirical distribution. It is also based on maximum likelihood estimation, hence no explicit expression for the estimator is available. The estimator proposed here deals with both of these issues by adjusting for the marginal transformation bias, and is given in analytical form with an explicit expressions for the variance, hence a CI can easily be obtained.

5.2.2 Estimation of the coefficient of tail dependence

Take F to be a bivariate probability distribution function in the domain of attraction of an extreme value distribution with continuous marginal distribution functions respectively defined by $F_X(x) := F(x, \infty)$ and $F_Y(y) := F(\infty, y)$ and (X, Y) a random vector following the distribution F. Suppose that for $x, y > 0$,

$$\lim_{t \downarrow 0} \frac{\mathbb{P}(1 - F_X(X) < tx, 1 - F_Y(Y) < ty)}{\mathbb{P}(1 - F_X(X) < t, 1 - F_Y(Y) < t)} =: S(x, y) \tag{5.4}$$

exists positive. Then the joint tail distribution function $q(t) := \mathbb{P}(1 - F_X(X) < t, 1 - F_Y(Y) < t)$ is regularly varying at zero with index $1/\eta$, that is $q(t) = t^{1/\eta}\mathcal{L}(t)$ for some $\eta \in (0, 1]$, where $\mathcal{L}$ is as previously a slowly varying function at zero. Condition (5.4) implies that S is homogenous of order $1/\eta$, i.e. $S(ax, ay) = a^{1/\eta}S(x, y)$. From this, we can see that condition (5.4) implies

$$\lim_{t \to \infty} \frac{\mathbb{P}\left(\frac{1}{1-F_X(X)} \wedge \frac{1}{1-F_Y(Y)} > tx\right)}{\mathbb{P}\left(\frac{1}{1-F_X(X)} \wedge \frac{1}{1-F_Y(Y)} > t\right)} = S\left(\frac{1}{x}, \frac{1}{x}\right) = x^{-1/\eta}S(1, 1) = x^{-1/\eta} \tag{5.5}$$

for all $x > 0$. Defining the tail distribution function as

$$T := \frac{1}{(1 - F_X(X)) \vee (1 - F_Y(Y))} \tag{5.6}$$

condition (5.4) implies that the tail distribution function $\bar{F}_T := 1 - F_T$ is of regular variation at infinity with index $-1/\eta$.

Since the marginal distributions F_X, F_Y are unknown in practice, their empirical counterparts need to be used instead. Let $R(X_i)$ denote the rank of X_i among $(X_1, ..., X_n)$, or specifically $R(X_i) := \sum_{j=1}^{n} \mathbf{1}_{X_j \leq X_i}$ and $R(Y_i)$ defined in a similar way. Then Equation (5.6) can be rewritten as

$$T_i^{(n)} := \frac{n+1}{n+1-R(X_i)} \wedge \frac{n+1}{n+1-R(Y_i)} \tag{5.7}$$

Note that the margins are transformed to unit Pareto, which was favoured in Draisma et al. (2004) and later in Goegebeur and Guillou (2012), due to be observed smaller bias compared to unit Fréchet margins. We denote the ascending order statistics of the sequence $\{T_i^{(n)}\}_{i=1}^n$ by $T_{n,1} \leq T_{n,2} \leq ... \leq T_{n,n}$. Since the limit $S(x, x)$ in (5.4) is completely defined up to one parameter, the CTD η, but the marginals are unspecified and therefore need to be estimated, here by the empirical counterparts, our proposed estimator will take place in a semi-parametric setting.

The class of estimators proposed is based on combining the empirical tail quantile function with a functional, similar in form to the "mean-of-order-p" estimator presented in Gomes and Caeiro (2014), but in a bivariate setting. Our approach for estimating the parameter η follows the work of Draisma et al. (2004) and Goegebeur and Guillou (2012), which stems from the univariate work in Drees (1998).

For this aim, define the empirical tail quantile function of $T_i^{(n)}$ as $Q_n(s) := T_{n,n-\lfloor ms \rfloor}$, for $0 < s < n/m$, where $\lfloor x \rfloor$ denotes the integer part of x. As shown in Goegebeur and Guillou (2012), $(m/n)Q_n(1) \xrightarrow{P} l$, with l positive, as $m \to \infty$, whereas for an intermediate sequence of positive integers $k(n) = nq^{\leftarrow}(m/n) \to \infty$, $k/n \to 0$, as $n \to \infty$, we have that $(k/n)Q_n(1) \xrightarrow{P} 1$. Our proposed class of estimators for the CTD builds on the functional

$$M_{a,b}(z) := \frac{(A_a(z))^b - 1}{b}, \qquad \text{with } A_a(z) := \left\{ \int_0^1 \left(\frac{z(s)}{z(1)} \right)^a ds \right\}^{1/a}, a, b \in \mathbb{R} \tag{5.8}$$

for any measurable function $z : [0, 1] \to \mathbb{R}$. For $a = 0$ and/or $b = 0$, this is interpreted in the limiting sense as $\log z$, which corresponds to the classical Hill estimator for the tail index given by the functional

$$M_0(z) := \int_0^1 \log^+ \frac{z(s)}{z(1)} ds \tag{5.9}$$

where $\log^+(w) = \max(\log(w), 0)$, hence only allows positive values.

Provided that a/b is set at approximately -1 and that suitable conditions stemming from a second order refinement of (5.4) are met, $M_{a,b}(Q_n(s))$ is a consistent (i.e. converges in probability to the correct value), asymptotically normal estimator for the CTD η. The second order condition will be defined and the asymptotic distribution and consistency will be demonstrated in the next section.

The main specific estimator considered in this work, is defined by the coefficient pair $(a, b) = (1/p, 1/q-1)$ with conjugate constants $p \in \mathbb{R}$, $q > 0$, i.e. $1/p+1/q = 1$. Substituting this and $Q_n(s)$ into (5.8), with k denoting the number of upper order statistics included and i the counting index, returns the estimator

$$\hat{\eta}_q(k) := \frac{\left\{\left[\frac{1}{k}\sum_{i=0}^{k-1}\left(\frac{T_{n,n-i}}{T_{n,n-k}}\right)^a\right]^{1/a}\right\}^{-(1-1/q)} - 1}{-(1-1/q)}, \quad a = \frac{1}{p} \tag{5.10}$$

with now $q = 1$ recovering the Hill estimator defined in (5.9). The coefficient pair $(a, b) = (1/(1-p), q-1)$ equals the before mentioned "mean-of-order-p" estimator.

In the original paper by Ledford and Tawn (1996) they proposed that one standardise the margins to unit Fréchet, instead of unit Pareto as outlined above, and this approach has been used in numerous papers (see for example Draisma et al. (2004) and Beirlant et al. (2011)). We will therefore use this transformation here as well. Like for the unit Pareto transformation F_X, F_Y are unknown and will therefore be replaced by their empirical counterparts, $F_X^{(n)}, F_Y^{(n)}$, on the basis of their usual plotting positions $i/(n+1)$ to avoid division by 0. Specifically,

$$V_i^{(n)} := \left(-\frac{1}{\log\ F_X^{(n)}(X_i)}\right) \wedge \left(-\frac{1}{\log\ F_Y^{(n)}(Y_i)}\right) = \left\{\left(-\log\frac{R(X_i)}{n+1}\right) \vee \left(-\log\frac{R(Y_i)}{n+1}\right)\right\}^{-1} \tag{5.11}$$

where $R(X_i), R(Y_i)$ are the ranks defined as before. We again denote by $V_{n,1} \leq V_{2,n} \leq ... \leq V_{n,n}$ the ascending order statistics and an estimator corresponding to Equation (5.10) with the T_i replaced by the V_i can be defined.

A third possible expression for the marginals that we proposed here, is to shift the pseudo unit Fréchet random variables V_i by $1/2$. This location-shifted class of estimators for η, defined through the functional $M_{a,b}(V_{n,n-\lfloor ms \rfloor} + 1/2)$ with $a/b \to -1$ defined through the

same conjugate constants p, q, is given by

$$\hat{\eta}_{a,b}^{(S)}(k) := \frac{\left\{ \left[\frac{1}{k} \sum_{i=0}^{k-1} \left(\frac{1/2+V_{n,n-i}}{1/2+V_{n,n-k}} \right)^a \right]^{1/a} \right\}^b - 1}{b} \tag{5.12}$$

We will demonstrate that the estimator $\hat{\eta}_{a,b}^{(S)}$ is asymptotically equivalent to $\hat{\eta}_{a,b}$ (Equation 5.10 before substituting in specific expressions for (a, b)) when defined with unit Pareto margins, and thereby establish the asymptotic distribution of $\hat{\eta}_{a,b}^{(S)}$. In the simulations we will further demonstrate that this simple shift of the Fréchet random variables substantially decreases the finite sample bias for the Hills' estimator, as discussed in both Draisma et al. (2004) and Goegebeur and Guillou (2012).

5.3 Asymptotic results

By establishing the asymptotic distribution of our estimator, we can learn about any potential bias that might exist because of terms not converging to 0 fast enough, and the variability around the estimate. Specifically, if the estimator is asymptotic normally distributed we can use standard methods for deriving the CI around the estimated value of η.

To establish the asymptotic distribution of the class of estimators $M_{a,b}(Q_n(s))$ defined in (5.8) in the two-dimensional setting, a second order refinement of condition (5.4) is needed. Similarly to the condition defined in Goegebeur and Guillou (2012), assume that condition (5.4) holds with S being continuously differentiable,

$$\lim_{t \downarrow 0} \frac{\frac{\mathbb{P}(1-F_X(X)<tx, 1-F_Y(Y)<ty)}{q(t)} - S(x,y)}{q_1(t)} =: D(x,y) \tag{5.13}$$

exists for all $x, y \geq 0$, $x + y > 0$, with q_1 a function of ultimately constant sign and tending to 0 as $t \downarrow 0$ and a function D which is neither constant nor a multiple of S. Moreover, we assume that the convergence is uniform on $\{(x, y) \in [0, \infty)^2 \mid x^2 + y^2 = 1\}$, and that $D(x, x) = x^{1/\eta} \frac{x^{\tau/\eta}-1}{\tau\eta}$. With the joint tail distribution function $q(t)$ being regularly varying at zero with index $1/\eta$, it can be shown that condition (5.13) implies that $|q_1|$ is regularly varying at zero with index $\tau/\eta \geq 0$. Finally, we also assume that $l := \lim_{t \downarrow 0} q(t)/t$ exists, something that is always satisfied if $\eta < 1$ and $\tau > 0$ (Goegebeur and Guillou (2012)). From all of the above assumptions, the asymptotic normality of the estimator $\hat{\eta}_q$ can be derived through the following Theorem and accompanied Corollary.

Theorem 1. *Let $T_{n,1} \leq T_{n,2} \leq ... \leq T_{n,n}$ be the ascending order statistics associated with the random variables $T_i^{(n)}, i = 1, .., n$ defined in (5.7), whose tail distribution is denoted by $\bar{F}_T = 1 - F_T$. Assume the following second order condition of regular variation for $\bar{F}_T$, implied by (5.13) and its given assumptions: there exist $\eta \in (0, 1]$, $\tau > 0$ and a function $q_\star(t) \to 0$, as $t \to \infty$, not changing sign eventually such that, for all $x > 0$,*

$$\lim_{t\to\infty} \frac{\frac{\mathbb{P}\left(\frac{1}{1-F_X(X)} \wedge \frac{1}{1-F_Y(Y)} > tx\right)}{\bar{F}_T(t)} - x^{-1/\eta}}{q_\star\left(1/\bar{F}_T(t)\right)} = x^{-1/\eta} \frac{x^{-\tau/\eta} - 1}{\eta\tau}$$

Assume the equivalent relation in terms of high quantiles:

$$\lim_{t\to\infty} \frac{\left(\frac{1}{1-F_T}\right)^{\leftarrow}(tx) / \left(\frac{1}{1-F_T}\right)^{\leftarrow}(t) - x^{\eta}}{q_\star(t)} = x^{\eta} \frac{x^{-\tau} - 1}{\tau} \tag{5.14}$$

for all $x > 0$.

For a sequence of integers $k = k(n) \to \infty$, $k/n \to 0$, as $n \to \infty$, there exists a Brownian bridge[1] B such that, with $a < 1/(2\eta)$, $a \neq 0$,

$$\sqrt{k}\left\{M_{a,b}(Q_n) - \frac{(1-a\eta)^{-b/a} - 1}{b}\right\} + \sqrt{k} q_\star\left(\frac{n}{k}\right) \frac{(1-a\eta)^{-b/a+1}}{(1-a\eta)(1-a\eta+\tau)} \xrightarrow{d} (1-a\eta)^{-b/a+1} \int_0^1 \eta s^{-(a\eta+1)} B(s) ds$$

The following proof is based on Lemma 1 in Goegebeur and Guillou (2012), in which Lemma 6.2 in Draisma et al. (2004) is modified to include a bias term.

Proof of Theorem 1. Under the second order condition (5.13), with $m := nq(k/n)$, k defined as above, and $m \to \infty$, there exists a function $q_\star$ of constant sign near infinity and tending to zero, by which $\sqrt{m} q_\star(1/q^{\leftarrow}(m/n)) = O(1)$, such that it is possible to define a sequence of standard Brownian motions $\{W_n(s)\}_{s\geq 0}$, such that, for all $s_0 > 0$, $\epsilon > 0$,

$$\sup_{0<t<t_0} t^{\eta+\frac{1}{2}+\epsilon} \left| \sqrt{m}\left(\frac{k}{n} Q_n(s) - s^{-\eta}\right) - \eta s^{-(\eta+1)} W_n(s) - \sqrt{m} q_\star\left(\frac{n}{k}\right) s^{-\eta} \frac{s^\tau - 1}{\tau} \right| = o_p(1) \tag{5.15}$$

We define, for $a \neq 0$,

$$Z_n(s) := \sqrt{m}\left\{\left(\frac{Q_n(s)}{Q_n(1)}\right)^a - s^{-a\eta}\right\}$$

and redefine the auxiliary function involved in the second order condition relating to $q_\star$

[1] A stochastic process with the conditional probability distribution given by a standard Wiener process with the condition that it equals to 0 for the start and end value.

such that $q_\star(t) \sim q_1(1/t)$, as $t \to \infty$. Coupling this with (5.15), we obtain

$$Z_n(s) = a\eta s^{-(a\eta+1)} \left(W_n(s) - sW_n(1)\right) + \sqrt{m} q_\star \left(1/q^{\leftarrow}\left(\frac{m}{n}\right)\right) s^{-a\eta} \frac{s^\tau - 1}{\tau} + o_p\left(\max(s^{-(\eta+\frac{1}{2}+\epsilon)}, s^{-\eta})\right)$$

where the o_p-term is uniform on a compact interval bound away from zero. After Taylor expanding around 0, we obtain the following asymptotic expansion with $\{B_n(s)\} \stackrel{d}{=} \{W_n(s) - sW_n(1)\}$, $s \in (0, 1]$, a sequence of Brownian bridges:

$$\int_0^1 Z_n(s)ds = a\eta \int_0^1 s^{-(a\eta+1)} B_n(s)ds + \sqrt{m} q_\star \left(\frac{n}{m}\right) \int_0^1 a s^{-a\eta} \frac{s^\tau - 1}{\tau} ds + o_p(1)$$

as $n \to \infty$. With both a, b assumed fixed, the result in the theorem follows straightforward via Cramèr's delta-method for $a < 1/(2\tau)$:

$$\sqrt{m} \left\{ M_{a,b}(Q_n) - \frac{(1 - a\eta)^{-b/a} - 1}{b} \right\} = \frac{1}{a}(1 - a\eta)^{-b/a+1} \int_0^1 Z_n(s)ds$$

□

From Theorem 1, the following corollary naturally follows

Corollary 2. *Under the conditions of Theorem 1, if* $\sqrt{k} q_\star(n/k) \to \lambda \in \mathbb{R}$ *as* $n \to \infty$, $a < 1/(2\eta)$ *and* $b/a = -1$, *then*

$$\sqrt{k} \left(\hat{\eta}_a(k) - \eta\right) \xrightarrow{d} \mathcal{N}(\lambda b_a, \sigma_a^2)$$

with

$$\hat{\eta}_a = \hat{\eta}_a(k) := M_{a,-a}(Q_n)$$

where

$$b_a = b_a(\eta, \tau) = \frac{(1 - a\eta)}{1 - a\eta + \tau}$$
$$\sigma_a^2 = \sigma_a^2(\eta) = \eta^2 \frac{(1 - a\eta)^2}{(1 - 2a\eta)}$$

Proof of Corollary 2. Theorem 1 ascertains that the random component $M_{a,b}(Q_n)$ with its deterministic bias subtracted, converges to an integral of a Brownian bridge, which essentially is a Gaussian process. Given that the increments of a Gaussian process are independent, normal random variables, this integral can be written as a sum of normals and is therefore a normal random variable itself. If $\sqrt{m} q_\star(n/m) \to \lambda \in \mathbb{R}$ as $n \to \infty$, the bias term follows from simple integration. To derive the variance of the limiting normal random variable, it is sufficient to consider the process $Z(s) := \eta s^{-(a\eta+1)} B(s)$,

$0 \leq s \leq 1$, since all other terms equal to zero. The variance of the process $Z(s)$ is simply $Var\left(\int_0^1 Z(s)ds\right) = \mathbb{E}\left[\int_0^1 \int_0^1 Z(s)Z(t)dsdt\right] = \eta^2(1-a\eta)^2/(1-2a\eta)$ □

As mentioned previously, letting $a \to 0$ the Hill estimator is recovered. As one can easily see from Corollary 2, the dominant component of the bias b_a gets smaller for larger values of $\tau > 0$. This can be explained by the higher rate of convergence of the actual underlying bivariate distribution F to its specific max-stable limit for larger values of τ. It is also apparent that the asymptotic variance does not depend on τ and is therefore unaffected by this. This is instead mainly controlled by the parameter $a < 1/(2\eta)$, a condition required to keep $\sigma_a^2 > 0$ since $\eta \in (0, 1]$. Through some straightforward algebraic manipulation one can show that $\sigma_a^2(\eta) \geq \eta^2$, for all a, η with equality easily seen attained for $a = 0$, i.e. the Hill estimator, which however does not have the smallest bias. Figure 5.1 visualises how the asymptotic variance increases with increasing η and $|a|$, with η being the dominating parameter.

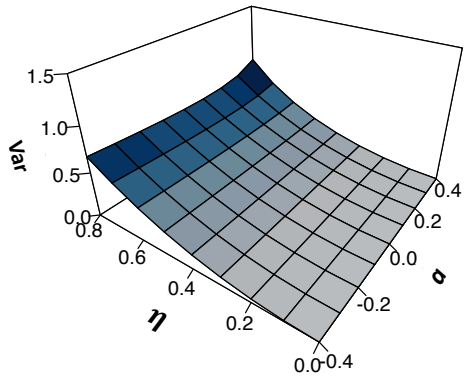


Figure 5.1: Asymptotic variance of $\hat{\eta}_a$ as a function of $|a| \leq 0.4$ and $\eta \in (0, 0.8]$

In order to establish the asymptotic distribution of the location-shifted class of estimators, $\hat{\eta}_{a,b}^{(S)}$, we will first show that it is asymptotically equivalent to $\hat{\eta}_a$ with unit Pareto margins. This is determined through the following Theorem, with the proof deferred to Appendix 5.A.

Theorem 3. *Assume the same conditions as in Theorem 1. Let $k = k(n)$ be an intermediate sequence such that $\sqrt{k}q_\star(n/k) \to 0$ as $n \to \infty$. Then, for every $a, b \in \mathbb{R}$, the estimator $\hat{\eta}_{a,b}^{(S)}(k)$ defined in (5.12), stemming from the transformation to unit Fréchet marginals with a shift by one half, is asymptotically equivalent to its standard counterpart $\hat{\eta}_{a,b}(k)$ in the class (5.10), i.e. as $n \to \infty$*

$$\sqrt{k}\ \left|\hat{\eta}_{a,b}^{(S)}(k) - \hat{\eta}_{a,b}(k)\right| \xrightarrow{P} 0$$

Hence, the two estimators have the same asymptotic distribution and only differ by a potential second order deterministic bias, which can be made explicit in terms of $\eta \in (0, 1]$,

$\rho > 0$. Thanks to the condition $\sqrt{k}q_\star(n/k) \to o(1)$ in the theorem which imposes a further mild restriction on the growth of k, this difference is dominated by the bias term $b_a(\eta, \tau)$ defined in Corollary 2. In the next section, we will present an estimator where this larger bias has been removed.

5.4 Reduced bias estimator

The estimator $\hat{\eta}_a$ presented in the previous section is clearly not unbiased except for some special cases. We will therefore in this section derive a reduced bias version of the estimator, based on subtracting the leading bias without significantly increasing the variance. Before introducing this estimator, we first need to establish a similar relation to (5.14), but with the quantile function associated with transformation to unit Fréchet instead of Pareto. The proof of this relation will also guide the particular form of the estimator. With the usual notation of $U := (1/(1-F))^{\leftarrow}$ and F the distribution function, the unit Fréchet quantile function is given by,

$$V(t) := \left(-\frac{1}{\log F}\right)^{\leftarrow}(t) = U\left(\frac{1}{1-e^{-1/t}}\right)$$

Before we are able to prove Theorem 5 we need a lemma, similar to the statement in Theorem 3, but now involving the extreme quantile functions V and U associated with the unit Fréchet and Pareto transform respectively. The proofs for both the lemma and the theorem are deferred to Appendix 5.B.

Lemma 4. *Define $V^*(t) := V(t) + 1/2$ and supposed that (5.14), the high quantile relation in Theorem 1, holds for some $\eta \in (0, 1]$ and $\tau > 0$. Then, $\lim_{t\to\infty} U(t)/V^*(t) = 1$.*

Theorem 5. *Let U be a quantile function of regular variation at infinity with index $\eta \in (0, 1]$, i.e. $\lim_{t\to\infty} U(tx)/U(t) = x^\eta$, for $x > 0$, which we denote by $U \in RV_\eta$. Assume that the second order condition of regular variation (5.14) is satisfied for U with the second order parameter $\tau > 0$. Then V^* is also of regular variation with the same index η and is such that, for $x > 0$,*

$$\lim_{t\to\infty} \frac{\frac{V^*(tx)}{V^*(t)} - x^\eta}{\tilde{q}(t)} = \begin{cases} x^\eta \frac{x^{-\tau}-1}{\tau}, & \tau < \eta \\ x^\eta \frac{x^{-\eta}-1}{\eta}, & \tau \geq \eta \end{cases} \tag{5.16}$$

where

$$\tilde{q}(t) = \begin{cases} q_\star(t) & \tau < \eta, \\ q_\star(t) + \dfrac{\eta}{2}\dfrac{1}{V^*(t)}, & \tau = \eta, \\ \dfrac{\eta}{2}\dfrac{1}{V^*(t)}, & \tau > \eta \end{cases}$$

Moreover, $|\tilde{q}(t)| \in RV_{-\tilde{\tau}}$ *with second order parameter governing the speed of convergence given by*

$$\tilde{\tau} = \begin{cases} \tau, & 0 < \tau < \eta, \\ \eta, & 0 < \eta \leq \tau \end{cases}$$

From the above theorem we can find the relevant Hall-Welsh-type model, defined in Hall and Welsh (1985), aligning with the estimator proposed here. Specifically, we are interested in the distribution functions whose extreme quantile function of the type V^* determines the following expansion as $t \to \infty$, $V^*(t) = C^\eta t^\eta (1 + \eta D_1 C^{-\tau} t^{-\tau} + \eta D_2 C^{-\eta} t^{-\eta})$, for $C > 0, D_1 \in \mathbb{R}$ and $D_2 \neq 0$. This leads to $\tilde{q}/\tilde{\tau} = \eta \tilde{\beta} t^{-\tilde{\tau}}$ with $\tilde{\beta} = D_j C^{-\tilde{\tau}}$, $j = 1$ if $\tilde{\tau} = \tau$ and $j = 2$ if $\tilde{\tau} = \eta$.

As mentioned in the beginning of the section, the reduced bias estimator stems from subtracting the leading bias term in the asymptotic distribution for the estimator $\hat{\eta}^{(S)}_{a,b}$, i.e. based on the shifted by $1/2$ unit Fréchet marginals. Through the asymptotic equivalence condition established in Theorem 3 together with the asymptotic distribution derived in Corollary 2 and Theorem 5, the following reduced bias estimator is proposed:

$$\tilde{\eta}_a(k) := \hat{\eta}^{(S)}_a(k) \left\{ 1 - \left(\hat{\beta} \left(\frac{n}{k} \right)^{-\hat{\tau}} + \frac{1}{1 + 2V_{n,n-k^*}} \right) \frac{1 - a\, \hat{\eta}^{(S)}_a(k)}{1 - a\, \hat{\eta}^{(S)}_a(k) + \hat{\tau}} \right\}, \tag{5.17}$$

with $k^* \leq \sqrt{k}$, $k \to \infty$, $k/n \to 0$, as $n \to \infty$ and $\hat{\beta}, \hat{\tau}$ denote consistent estimators for $\tilde{\beta}$ and $\tilde{\tau} > 0$. This next steps will demonstrate that this is indeed asymptotically normal distributed with zero mean for sufficiently large n. Express $V_{n,n-k^*}$ in terms of V^* and assume that $\sqrt{k}\tilde{q}(n/k) = O(1)$, then the sequence

$$\begin{aligned} \sqrt{k}\left(\tilde{\eta}_a(k) - \eta\right) = \sqrt{k} \left\{ \hat{\eta}^{(S)}_a(k) \left(1 - \hat{\beta} \left(\frac{n}{k} \right)^{-\hat{\tau}} \frac{1 - a\, \hat{\eta}^{(S)}_a(k)}{1 - a\, \hat{\eta}^{(S)}_a(k) + \hat{\tau}} \right) - \eta \right\} \\ + \sqrt{k}\, \frac{\hat{\eta}^{(S)}_a(k)}{2} \frac{1}{V^*_{n,n-k^*}} \frac{1 - a\, \hat{\eta}^{(S)}_a(k)}{1 - a\, \hat{\eta}^{(S)}_a(k) + \hat{\tau}} \end{aligned} \tag{5.18}$$

with the method developed in Caeiro et al. (2005) certifying that the remaining bias in the first $\sqrt{k}$-term converges to zero sufficiently fast as $n \to \infty$. The asymptotic expansion is therefore

$$\sqrt{k} \left\{ \hat{\eta}^{(S)}_a(k) \left(1 - \hat{\beta} \left(\frac{n}{k} \right)^{-\hat{\tau}} \frac{1 - a\, \hat{\eta}^{(S)}_a(k)}{1 - a\, \hat{\eta}^{(S)}_a(k) + \hat{\tau}} \right) - \eta \right\} = Z_a + o_p\left(\sqrt{k}\tilde{q}(n/k) \right) \tag{5.19}$$

where Z_a is a zero mean normal variable with variance $\sigma^2_a > 0$ defined in Corollary 2. The second term in (5.18) becomes negligible for sufficiently large n, as shown below. Recall Lemma 4, which proved that $\lim_{t\to\infty} U(t)/V^*(t) = 1$, where $U(t)$ is associated with the unit Pareto transform, and the following corollary from Haan and Ferreira (2006):

Lemma 6. *Let $Y_{1,n} \leq Y_{2,n} \leq ... \leq Y_{n,n}$ be the order statistics associated with a sample of n i.i.d standard Pareto random variables with common CDF $F(y) = 1 - 1/y$, $y \geq 1$. Then for an intermediate sequence $k = k(n) \to \infty, k/n \to 0$ as $n \to \infty$,*

$$\sqrt{k}\left(\frac{k}{n}Y_{n,n-k} - 1\right)$$

is asymptotically standard normal.

By the Taylor expansion $1/(1+y) = 1 - y + y^2/(y+1)$ and the condition $k^* \leq \sqrt{k}$, the second term in (5.18) is bounded by

$$\begin{aligned} 0 < \frac{\sqrt{k}\, k^*/n}{\left(\frac{k^*}{n}V^*_{n,n-k^*} - 1\right) + 1} &= \sqrt{k}\,\frac{k^*}{n}\left\{1 - \left(\frac{k^*}{n}V^*_{n,n-k^*} - 1\right) + \frac{\left(\frac{k^*}{n}V^*_{n,n-k^*} - 1\right)^2}{\frac{k^*}{n}V^*_{n,n-k^*}}\right\} \\ &\leq \frac{k}{n}\left\{1 + O_p\left(\frac{1}{\sqrt{k}}\right) + o_p\left(\frac{1}{\sqrt{k}}\right)\right\} \end{aligned}$$

hence becomes dominated by the first term for large enough n.

In the next section the finite sample performance of the estimators proposed will be demonstrated and compared to the Hill estimator. Defining the parameters by: k is the number of upper order statistics, $k^* \leq \sqrt{k}$, n is the sample size, a the tuning parameter and $\hat{\tau}, \hat{\beta}$ second-order parameters, and the three empirical marginal distributions given by

- Pareto
$$T_i^{(n)} := \frac{n+1}{n+1-R(X_i)} \wedge \frac{n+1}{n+1-R(Y_i)}$$
- Fréchet
$$V_i^{(n)} := \left\{\left(-\log\frac{R(X_i)}{n+1}\right) \vee \left(-\log\frac{R(Y_i)}{n+1}\right)\right\}^{-1}$$
- Shifted Fréchet
$$V_i^*(n) = V_i^{(n)} + 1/2$$

The estimators to be compared are

- **Hill** defined by
$$\hat{\eta}_H := \frac{1}{k}\sum_{i=0}^{k-1}\log\left(\frac{W_{n,n-i}}{W_{n,n-k}}\right)$$
and asymptotic distribution given by
$$\mathcal{N}\left(\eta + \lambda\frac{1}{1+\tau},\ \eta^2\right)$$

with $W_{n,n-i}$ being either Pareto, Fréchet or shifted Fréchet empirical marginals.

- **Extended mean-of-order-p** defined as

$$\hat{\eta}_q(k) := \frac{\left\{ \left[\frac{1}{k} \sum_{i=0}^{k-1} \left(\frac{W_{n,n-i}}{W_{n,n-k}} \right)^{\frac{1}{p}} \right]^p \right\}^{-(1-1/q)} - 1}{-(1-1/q)}$$

and asymptotic distribution given by

$$\mathcal{N}\left(\eta + \lambda \, \frac{1-a\eta}{1-a\eta+\tau}, \; \eta^2 \frac{(1-a\eta)^2}{(1-2a\eta)} \right)$$

with $W_{n,n-i}$ as above being either Pareto, Fréchet or shifted Fréchet empirical marginals.

- **Reduced bias extended mean-of-order-p** defined as

$$\tilde{\eta}_a(k) := \hat{\eta}_a^{(S)}(k) \left\{ 1 - \left(\hat{\beta} \left(\frac{n}{k} \right)^{-\hat{\tau}} + \frac{1}{1+2V_{n,n-k^*}} \right) \frac{1-a\,\hat{\eta}_a^{(S)}(k)}{1-a\,\hat{\eta}_a^{(S)}(k)+\hat{\tau}} \right\},$$

and asymptotic distribution given by

$$\mathcal{N}\left(\eta, \; \eta^2 \frac{(1-a\eta)^2}{(1-2a\eta)} \right)$$

with $\hat{\eta}_a^{(S)}(k)$ defined by the above estimator with shifted Fréchet empirical marginals.

5.5 Finite sample simulations

The aim of this section is to demonstrate the finite sample performance of the estimators proposed in Section 5.2.2 and 5.4. For the estimators presented in Section 5.2.2 the focus is to compare the results for the three possible marginal distributions and the dependence on the parameter q. The reduced bias estimator will only include the shifted Fréchet marginals V^* and the aim is instead to evaluate the impact of the bias reduction.

To simulate bivariate samples with given marginals and known values of η, τ, a selection of copula models will be used. Several copulas are chosen due to their different behaviours in both the main part of the distribution and the tails (see Figure 2.9) and we want to make sure that our estimator can accurately estimate the tail dependence for all these cases. Section 2.2.2 details the basics of copulas, hence only a very brief definition will be provided here.

The method of copulas stems from the famous Sklar's theorem (Sklar (1959)), which states that if F is a two-dimensional distribution function with continuous marginal distribution functions F_X, F_Y, then there exists a unique copula $\mathcal{C} : [0,1] \to [0,1]$ such that $F(x,y) = \mathcal{C}\left(F_X(x), F_Y(y)\right)$. Coupling this with the definition

$$\mathbb{P}(X > x, Y > y) = 1 - F_X(x) - F_Y(y) + F(x,y)$$

we arrive at the following relation

$$\mathbb{P}(1 - F_X(X) < tx, 1 - F_Y(Y) < ty) = tx + ty - 1 + \mathcal{C}_\theta(1 - tx, 1 - ty)$$

with $\mathcal{C}_\theta(x,y)$ a particular copula function of the joint distribution function F, from which we can simulate data. A thorough introduction to copulas and their relation to various dependence measures is found in Nelsen (2006). From this definition the corresponding values of η and second-order parameter τ, which controls the rate of convergence to the true bivariate distribution F, can be derived. Heffernan (2000) provides a comprehensive list for values of η for a large number of copula models. To robustly evaluate the performance, four parent bivariate distributions with different combinations of η, τ will be used, namely:

(i) **Farlie-Gumbel-Morgenstern** distribution which is defined by the copula function

$$\mathcal{C}_\theta(u,v) = uv\{1 + \theta(1-u)(1-v)\}, \quad (u,v) \in [0,1]^2, \theta \in [-1,1]$$

For $\theta \in (-1,1]$,

$$\frac{\mathbb{P}(1 - F_1(X) < tx, 1 - F_2(Y) < ty)}{\mathbb{P}(1 - F_1(X) < x, 1 - F_2(Y) < y)} = xy\left[1 - \frac{\theta t}{1+\theta}(x + y - 2) + O(t^2)\right]$$

which from (5.13) corresponds to $\eta = 0.5$, $\tau = 1$.

In the simulation we will use $\theta = -0.25$;

(ii) **Frank** distribution with copula function

$$\mathcal{C}_\theta(u,v) = -\frac{1}{\theta}\log\left(1 - \frac{\left(1 - e^{-\theta u}\right)\left(1 - e^{-\theta v}\right)}{1 - e^{-\theta}}\right), \quad (u,v) \in [0,1]^2, \theta > 1$$

Expanding the above leads to

$$\frac{\mathbb{P}(1 - F_1(X) < tx, 1 - F_2(Y) < ty)}{\mathbb{P}(1 - F_1(X) < x, 1 - F_2(Y) < y)} = xy\left[1 - \frac{\theta t}{2}(x + y - 2) + O(t^2)\right]$$

which satisfies (5.13) with $\eta = \tau = 1/2$.

In the simulation we use $\theta = 0.5$;

(iii) **Ali-Mikhail-Haq** distribution with copula defined by

$$\mathcal{C}_\theta(u,v) = \frac{uv}{1-\theta(1-u)(1-v)}, \quad (u,v) \in [0,1]^2, \theta \in [-1,1]$$

For $\theta = -1$ we get

$$\frac{\mathbb{P}(1-F_1(X) < tx, 1-F_2(Y) < ty)}{\mathbb{P}(1-F_1(X) < x, 1-F_2(Y) < y)} = xy\left[\frac{x+y}{2} - \frac{t^2}{2}(x+y)(xy-1) + O(t^3)\right]$$

satisfying (5.13) with $\eta = 1/3$ and $\tau = 2\eta = 2/3$.

(iv) **Bivariate Normal** distribution with copula function

$$\mathcal{C}_\theta(u,v) = \int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\Phi^{-1}(v)} \frac{1}{2\pi\sqrt{1-\theta^2}} \exp\left(-\frac{s^2 - 2\theta st + t^2}{2(1-\theta^2)}\right) \mathrm{d}s\mathrm{d}t, \quad (u,v) \in [0,1]^2$$

$\theta \in [-1,1]$, falls outside the scope of our study since $\tau = 0$, but is included in the simulation study to evaluate the robustness of our estimator. $\eta = \frac{1+\theta}{2}$, $D(x,y) = (xy)^{1/(1+\theta)}$ and we refer to Draisma et al. (2004) and Ledford and Tawn (1997) for detailed calculations.

We will set $\theta = 0.6$ which corresponds to $\eta = 0.8$.

From each of the above distributions, $N = 1000$ independent samples are drawn with $n = 500$ pseudo-random i.i.d sample points from (X,Y) each. The estimators are computed for $k = 5, ..., 300$, where k denotes the number of upper order statistics of the T_i, V_i for the estimation of η. To determine the range of values for the distortion parameter q, which defines our primary estimator $a = 1 - 1/q$, and recall the necessary condition $a < 1/(2\eta)$. In the case of $\eta = 1/2$, which defines the case of the bivariate pair (X,Y) being close to exactly independent, this implies that $q > 0$ (since $\eta \in (0,1]$). For $\eta > 1/2$, corresponding to a positive association between (X,Y) we get an upper bound on q, whereas for $\eta < 1/2$ which represents negative association, we get a negative lower bound, hence not adding any further restrictions to q. Since none of the copulas introduced above, satisfying $\tau > 0$, have a corresponding $\eta > 1/2$ we will take $q = 0.1(0.1)1.9$.

To estimate τ, β in the reduced bias estimator given by Equation (5.17), the algorithm outlined in Gomes et al. (2016) will be used, as described below.

For a tuning parameter $\rho \in \mathbb{R}$

$$\hat{\tau}_\rho := \min\left(0, \frac{3(T_n^{(\rho)}(k)-1)}{T_n^{(\rho)}(k)-3}\right)$$

dependent on the statistics

$$T_n^{(\rho)}(k) := \begin{cases} \frac{\left(M_n^{(1)}(k)\right)^\rho - \left(M_n^{(2)}(k)/2\right)^{\rho/2}}{\left(M_n^{(2)}(k)/2\right)^{\rho/2} - \left(M_n^{(3)}(k)/6\right)^{\rho/3}}, & \text{if } \rho \neq 0 \\ \frac{\ln\left(M_n^{(1)}(k)\right) - \frac{1}{2}\ln\left(M_n^{(2)}(k)/2\right)}{\frac{1}{2}\ln\left(M_n^{(2)}(k)/2\right) - \frac{1}{3}\ln\left(M_n^{(3)}(k)/6\right)}, & \text{if } \rho = 0 \end{cases}$$

where

$$M_n^{(j)}(k) := \frac{1}{k}\sum_{i=1}^{k}\left[\ln X_{n-i+1,n} - \ln X_{n-k,n}\right]^j, \quad j = 1,2,3$$

Thanks to the very stable sample path of τ as a function of the sample size k_1, the number of upper order statistics included in the estimation process of τ is set to be $k_1 = n^{*0.999}$, where n^* is the number of non-zero sample points (which here equals the sample length since the probability of observing $\{X = 0\}$ or $\{Y = 0\}$ is zero in our simulations). To choose ρ, we follow the stability algorithm proposed in Gomes and Pestana (2007) which selects the value with the most stable path for large k_1.

Denoting the estimate $\hat{\tau}$, β is estimated through

$$\hat{\beta}_{\hat{\tau}} := \left(\frac{k}{n}\right)^{\hat{\tau}} \frac{d_{\hat{\tau}}(k)D_0(k) - D_{\hat{\tau}}(k)}{d_{\hat{\tau}}(k)D_{\hat{\tau}}(k) - D_{2\hat{\tau}(k)}}$$

where for any $\alpha \leq 0$,

$$d_\alpha(k) := \frac{1}{k}\sum_{i=1}^{k}\left(\frac{i}{k}\right)^{-\alpha}$$

$$D_\alpha(k) := \frac{1}{k}\sum_{i=1}^{k}\left(\frac{i}{k}\right)^{-\alpha} W_i$$

where W_i, $i = 1, \ldots k$ are scaled log-spacings defined by

$$W_i := i\left(\ln\frac{X_{n-i+1,n}}{X_{n-i,n}}\right)$$

5.5.1 Marginal distribution impact

As discussed in Chapter 2, the choice of marginal distribution should theoretically not have an impact. This is however only true if we have an infinite sample, i.e. if $n \to \infty$, which naturally can never be true in practice. We further do not have the true distribution function but only the empirical distribution. These two factors means that the marginal distribution can have an impact because of the potential bias associated with them, something that can only be discovered through finite sample simulations, i.e. take n much smaller than ∞. By understanding which marginal transformation results in the smallest bias and MSE, we can learn which transformation we should apply in our real world analysis in Chapter 6.

Figures 5.2 - 5.5 shows as a function of the top sample fraction k/n on the left the estimated mean, taken as the sample mean of the 1000 samples, and the mean squared error (MSE) on the right for each of the four copulas introduced above, estimated by the proposed specific estimators $\hat{\eta}_q$ (5.10) and $\hat{\eta}_q^{(S)}$ (5.12). The three marginal distributions, unit Fréchet (top), shifted by $1/2$ unit Fréchet (middle) and unit Pareto (bottom) are all included in order to compare the bias and MSE associated with each. The green horizontal line marks the optimal value, i.e. correct η for the mean and 0 for the MSE plots. The dashed lines are for $q < 1$ and solid for $q > 1$, with the orange and red marking the limits, and the blue dot-dashed highlights the Hill estimator which is recovered for $q = 1$.

As discussed in previous papers, the bias in unit Fréchet case is significantly larger compared to the unit Pareto. This difference in bias additionally gets large for increasing values of k/n, since the bias for the unit Fréchet grows larger whereas it stays nearly constant for the unit Pareto. This is however not the case for the shifted Fréchet marginals, which very much behaves like the Pareto case for small values of k/n, with an increasing influence from the additional bias term for larger values. This nicely confirms the asymptotic results derived in Theorem 3.

For the Gaussian copula, which is out of the scope for our defined estimator due to $\tau = 0$, a similar general behaviour is seen for the different marginals with an increasing pattern for the Fréchet and stable for the others. The bias is however significantly larger, demonstrating the slower convergence rate.

In Figure 5.1 we saw that the asymptotic variance increases as a function of $|a|$, here represented mainly by $q < 1$, and η. This is again confirmed in these simulations, with significantly larger variance for $q = 0.1$, which also is partly due to the associated higher bias. This is

however only visible for the smallest value of q, showing that the value of q in general has a relatively small impact on the variance and mainly the bias. The impact from η can mainly be seen in the difference between the Frank ($\eta = 1/2$) and Ali-Mikhail-Had copula ($\eta = 1/3$) (Fig. 5.3, 5.4), where the bias is similar but the former has a significantly smaller variance for smaller values of k/n.

Focusing on the bias, $q > 1$ seems to be advisable with values close to 1 generally having a smaller bias compared to the other values of q for low values of k/n. This is however not true for the Farlie-Gumbel-Morgenstern copula with shifted Fréchet and Pareto marginals where $0.2 < q < 1$ provides the slightly smaller bias. Considering all the copulas except the Gaussian, it is advisable to only consider a small top fraction of $k/n \sim 0.05$, before the bias becomes significantly larger.

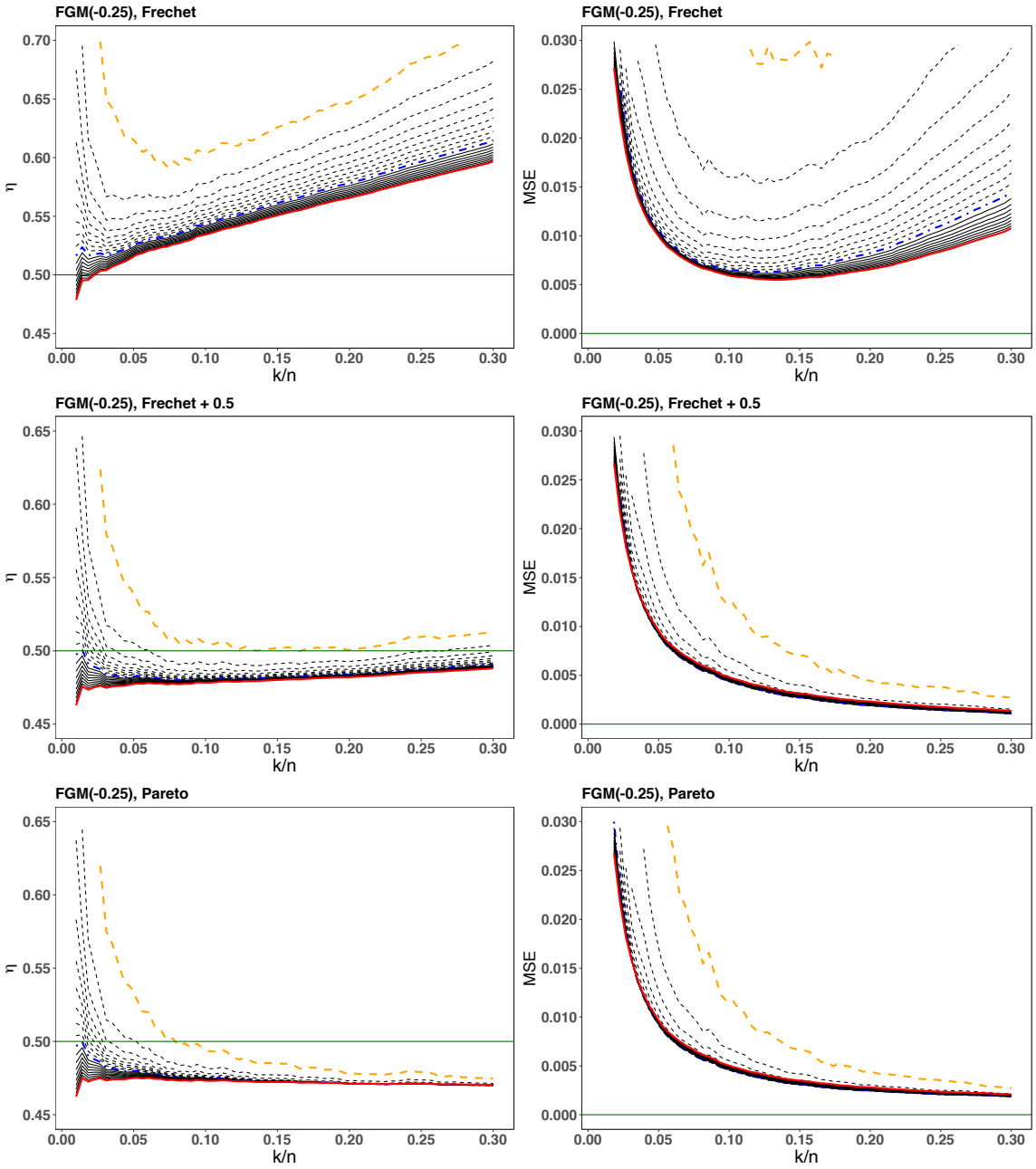


Figure 5.2: Farlie-Gumbel-Morgernstern copula with $\theta = -0.25$. The marginal distributions are (top) unit Fréchet, (middle) shifted by one half unit Fréchet and (bottom) unit Pareto. Dashed lines corresponds to $q < 1$ with orange line for the lower limit $q = 0.1$; solid lines identify $q > 1$ with red highlighting the upper bound $q = 1.9$. The blue dot-dashed line marks the Hill estimator ($q = 1$). Note the different scales on the y-axis for the mean.

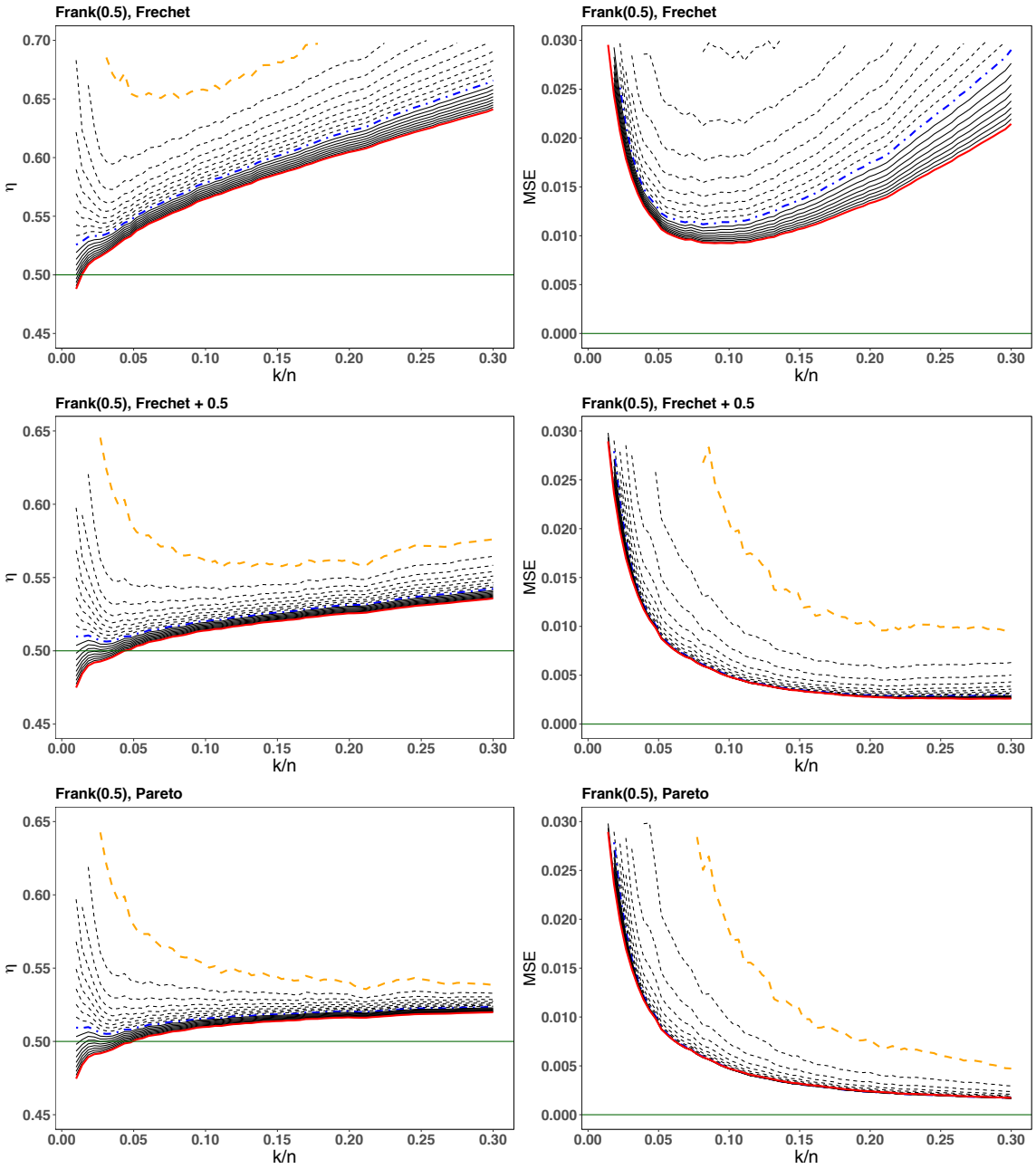


Figure 5.3: Frank copula with $\theta = 0.5$. The marginal distributions are (top) unit Fréchet, (middle) shifted by one half unit Fréchet and (bottom) unit Pareto. Dashed lines corresponds to $q < 1$ with orange line for the lower limit $q = 0.1$; solid lines identify $q > 1$ with red highlighting the upper bound $q = 1.9$. The blue dot-dashed line marks the Hill estimator ($q = 1$). Note the different scales on the y-axis for the mean.

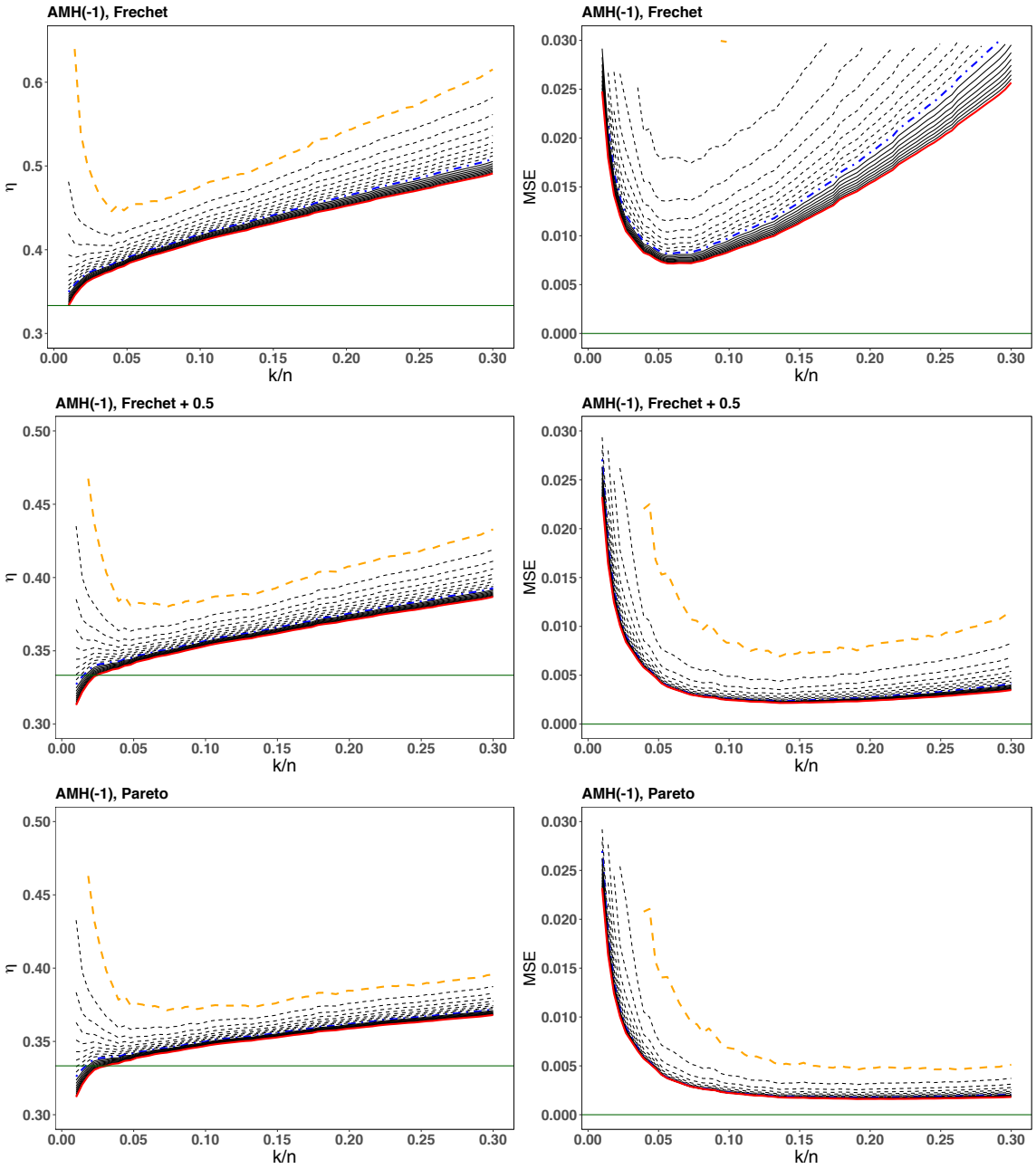


Figure 5.4: Ali-Mikhail-Had copula with $\theta = -1$. The marginal distributions are (top) unit Fréchet, (middle) shifted by one half unit Fréchet and (bottom) unit Pareto. Dashed lines corresponds to $q < 1$ with orange line for the lower limit $q = 0.1$; solid lines identify $q > 1$ with red highlighting the upper bound $q = 1.9$. The blue dot-dashed line marks the Hill estimator ($q = 1$). Note the different scales on the y-axis for the mean.

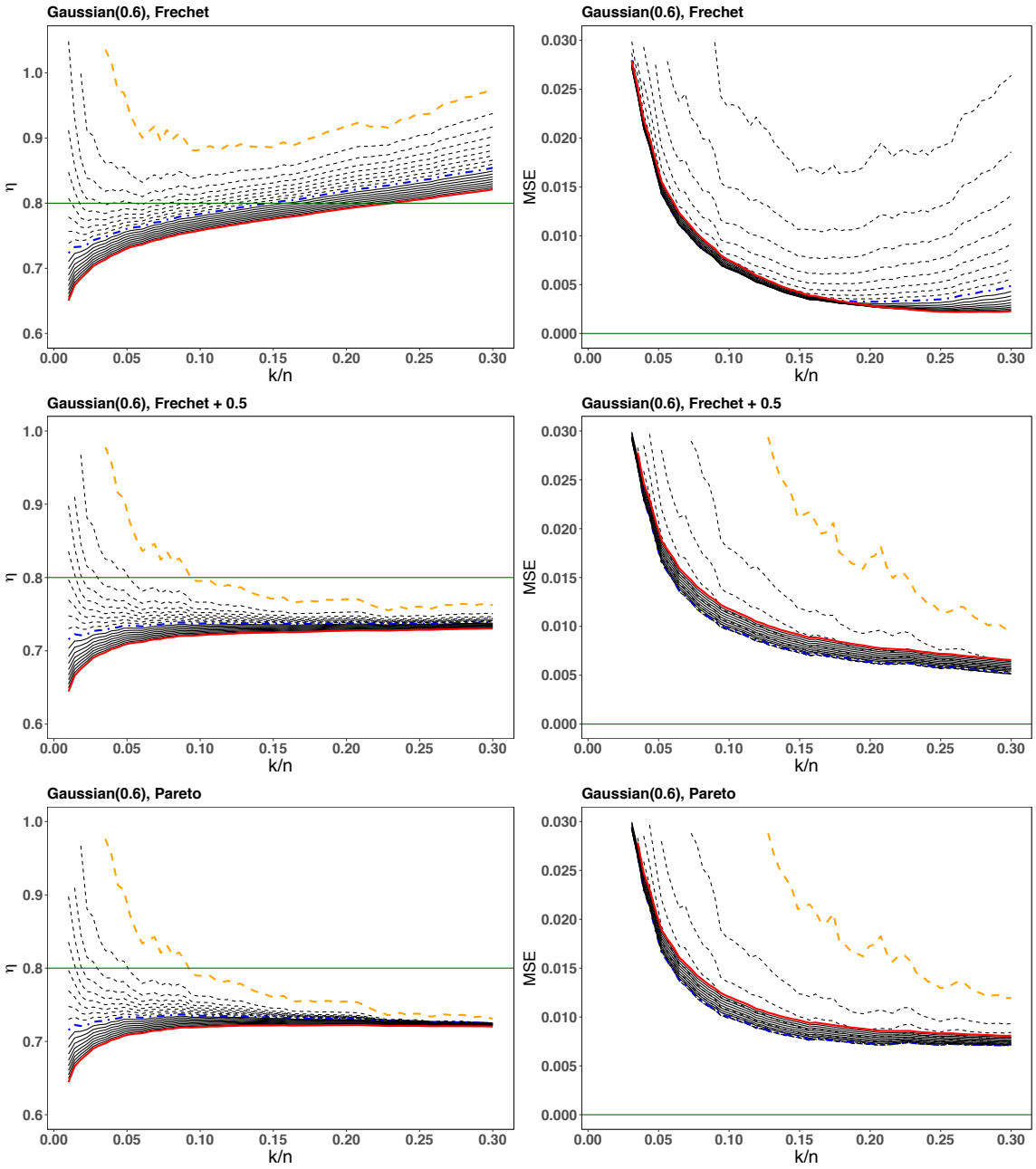


Figure 5.5: Bivariate Gaussian copula with $\theta = 0.6$. The marginal distributions are (top) unit Fréchet, (middle) shifted by one half unit Fréchet and (bottom) unit Pareto. Dashed lines corresponds to $q < 1$ with orange line for the lower limit $q = 0.1$; solid lines identify $q > 1$ with red highlighting the upper bound $q = 1.9$. The blue dot-dashed line marks the Hill estimator ($q = 1$). Note the different scales on the y-axis for the mean.

5.5.2 Reduced bias estimator

The focus is now on the reduced bias estimator $\tilde{\eta}_q^{(S)}$ (5.17), therefore only including the shifted Fréchet marginals, and only the three copulas with $\tau > 0$ are considered. The values for q are the same as above. In the definition of the reduced bias estimator we imposed that $k^* \leq \sqrt{k}$, and here two values for the fractional power of k^* are selected, specifically $0.3, 1/2$, to evaluate the dependence on this choice.

Figure 5.6 - 5.8 similarly to above displays the estimated mean to the left and the MSE to the right, with the lines representing the same range of values. From these, we can immediately see that the two copulas that had a positive bias, Frank and Ali-Mikhail-Had, has their bias nearly fully reduced to zero. The Farlie-Gumbel-Morgenstern on the other hand has a slightly increased bias, which results in the Hill estimator performing better both in terms of bias and MSE. This is due to the estimation process for β, which always returns a positive $\hat{\beta}$, leading to a subtracting term.

For all copulas, a value of q slightly below 1 seems the most appropriate since it has a stable minimal bias for all values of k/n. The specific choice of this parameter however becomes less important for increasing k/n, with very little difference between the estimates corresponding to the values $q = [0.5, 1.5]$ for $k/n \geq 0.1$. A larger fraction of the sample can also be used compared with the estimator $\hat{\eta}_q^{(S)}$, since the remaining bias is approximately constant for values $k/n \leq 0.1$. The number of top samples k^* does not seem to have a significant impact, as long as it is not greater than $\sqrt{k}$, which can be seen in the minimal difference between the plots using $k^* = k^{0.3}$ and $k^* = k^{1/2}$.

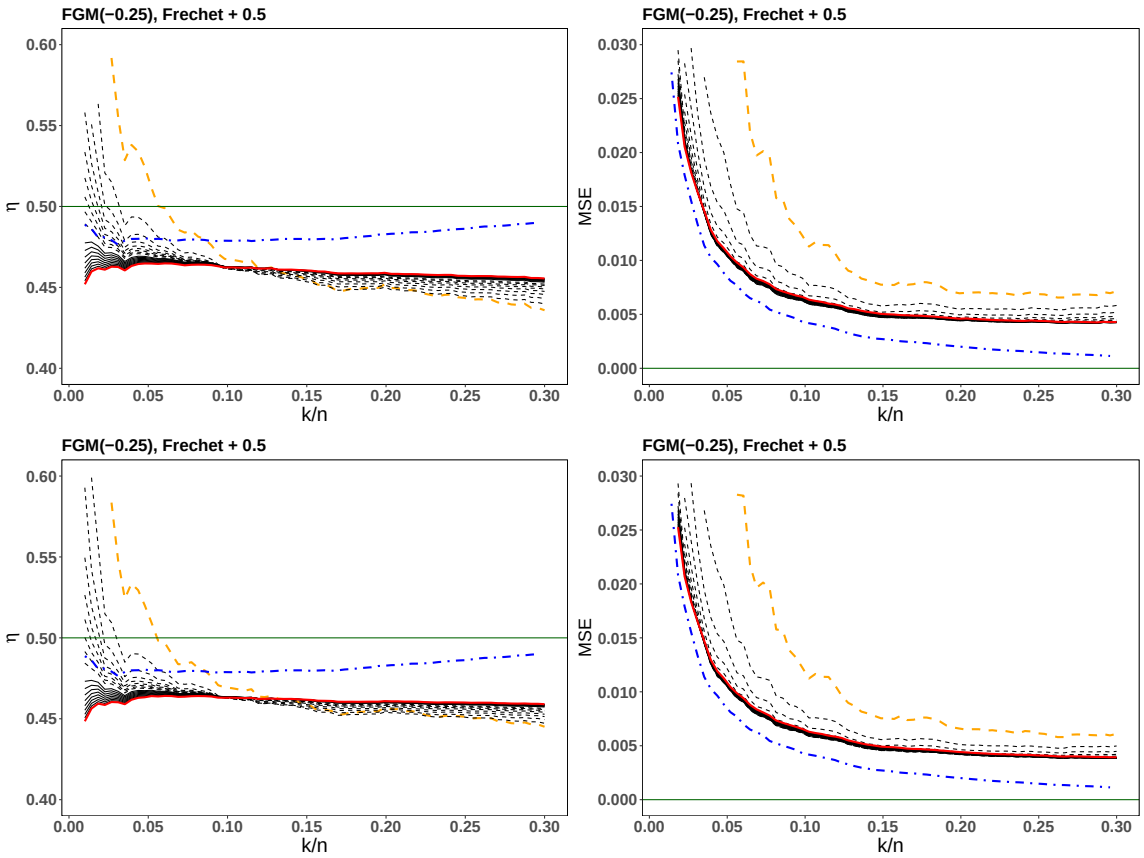


Figure 5.6: Farlie-Gumbel-Morgernstern copula with $\theta = -0.25$. Reduced bias estimator $\tilde{\eta}_q^{(S)}$ with for (top row) $k^* = \sqrt{k}$ and (bottom row) $k^* = k^{0.3}$. Dashed lines corresponds to $q < 1$ with orange line the lower limit $q = 0.1$; solid lines identify $q > 1$ with red highlighting the upper bound $q = 1.9$. The blue dot-dashed line marks the Hill estimator $(q = 1)$.

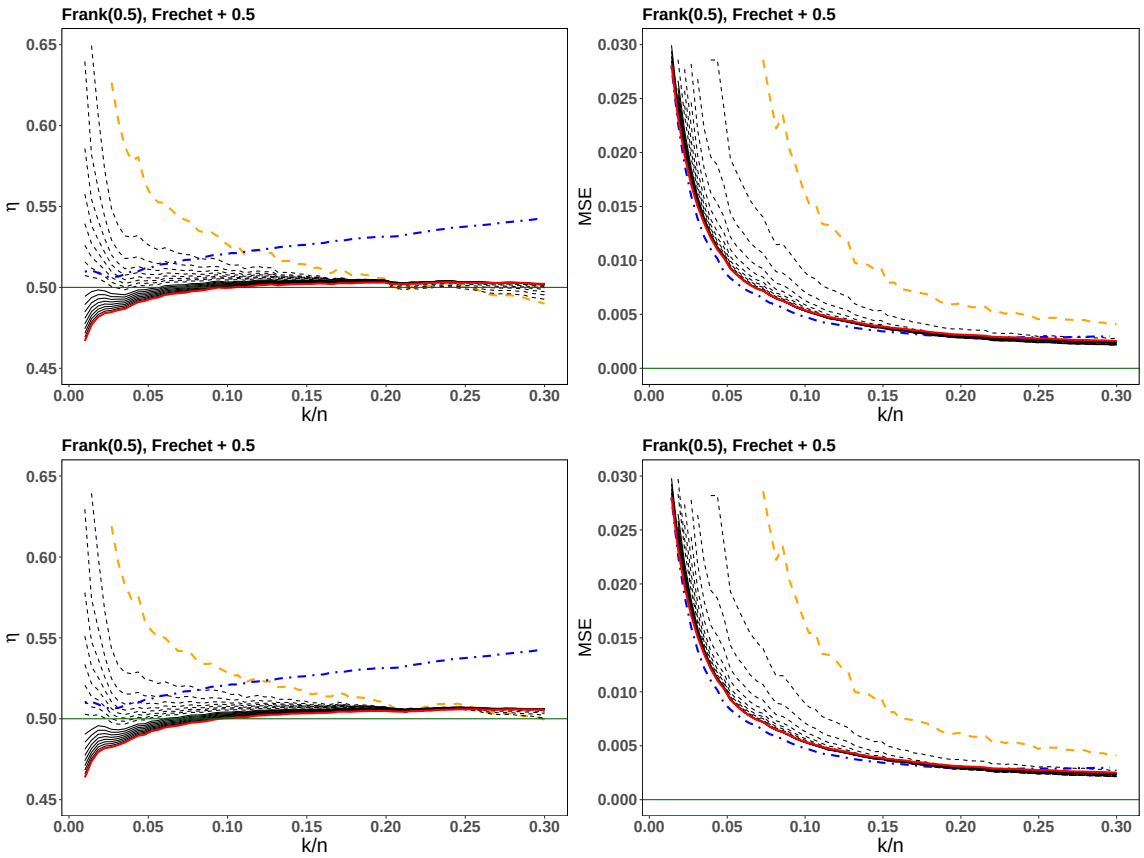


Figure 5.7: Frank copula with $\theta = -0.25$. Reduced bias estimator $\tilde{\eta}_q^{(S)}$ with for (top row) $k^* = \sqrt{k}$ and (bottom row) $k^* = k^{0.3}$. Dashed lines corresponds to $q < 1$ with orange line the lower limit $q = 0.1$; solid lines identify $q > 1$ with red highlighting the upper bound $q = 1.9$. The blue dot-dashed line marks the Hill estimator ($q = 1$).

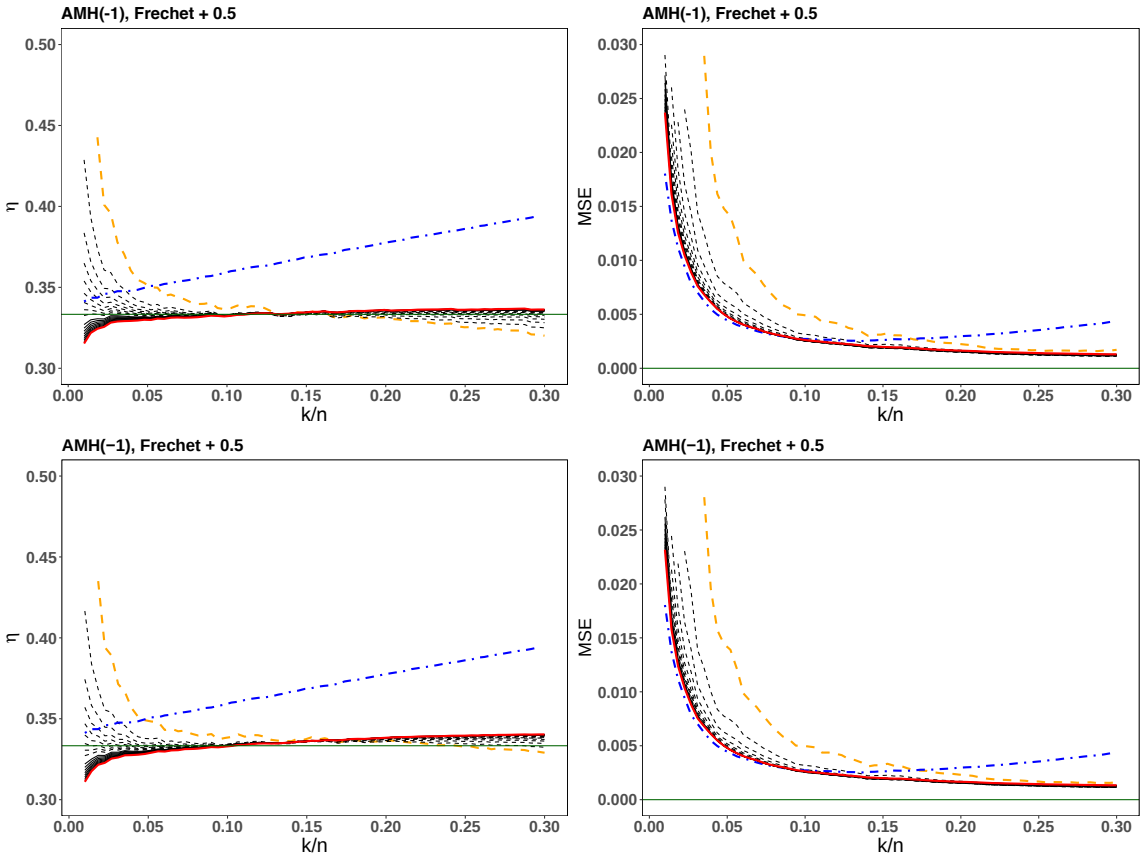


Figure 5.8: Ali-Mikhail-Had copula with $\theta = -0.25$. Reduced bias estimator $\tilde{\eta}_q^{(S)}$ with for (top row) $k^* = \sqrt{n}$ and (bottom row) $k^* = k^{0.3}$. Dashed lines corresponds to $q < 1$ with orange line the lower limit $q = 0.1$; solid lines identify $q > 1$ with red highlighting the upper bound $q = 1.9$. The blue dot-dashed line marks the Hill estimator ($q = 1$).

5.A Proof of Theorem 3

(The proofs in the Appendices 5.A and 5.B have been devised by Cláudia Neves, co-author of the submitted paper 'Estimation and reduced bias estimation of the residual dependence index with unnamed marginals'.)

Proof. Consider the random pairs (X_i, Y_i), $i = 1, ..., n$, representing i.i.d copies of (X, Y). The primary focus is on the direct empirical analogue to the copula survival function $C(1-x, 1-y)$ such that $C(x, y) = \mathbb{P}(1-F_X(X) < x, 1-F_Y(Y) < y)$, which has at its core a summation of Bernoulli random variables associated with the non-independent albeit identically distributed, $V_i^{(n)}$. For an intermediate sequence $m = nq(k/n)$ and $m \to \infty$, we have for each $x \in [0, T]$, $T > 0$ and $i = 1, .., n$ on the one hand, that

$$\begin{aligned}
&\sum_{i=1}^{n} \mathbf{1}_{\left\{X_i \geq X_{n-\lfloor kx \rfloor+1,n}, Y_i \geq Y_{n-\lfloor kx \rfloor+1,n}\right\}} \\
=&\sum_{i=1}^{n} \mathbf{1}_{\left\{1-F_X^{(n)}(X_i) \leq 1-F_X^{(n)}(X_{n-\lfloor kx \rfloor+1,n}), 1-F_Y^{(n)}(Y_i) \leq 1-F_Y^{(n)}(Y_{n-\lfloor kx \rfloor+1,n})\right\}} \\
=&\sum_{i=1}^{n} \mathbf{1}_{\left\{T_i^{(n)} \geq \frac{n}{\lfloor kx \rfloor}\right\}}
\end{aligned} \tag{5.20}$$

with $T_i^{(n)}$ defined in (5.7), which the following properly standardised version satisfies, as $n \to \infty$,

$$\sqrt{m}\left\{\frac{\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\left\{\frac{k}{n} T_i^{(n)} \geq x\right\}}}{q(n/k)} - x^{-1/\eta}\right\} \xrightarrow{d} W(x,x)$$

in $D([0,T])$, where W is a zero-mean Gaussian process (cf. Haan and Ferreira (2006) p.268). On the other hand,

$$\begin{aligned}
\sum_{i=1}^{n} \mathbf{1}_{\left\{X_i \geq X_{n-\lfloor kx \rfloor+1,n}, Y_i \geq Y_{n-\lfloor kx \rfloor+1,n}\right\}} &= \sum_{i=1}^{n} \mathbf{1}_{\left\{-\frac{1}{\log F_X^{(n)}(X_i)} \geq -\frac{1}{F_X^{(n)}(X_{n-\lfloor kx \rfloor+1,n})}, -\frac{1}{\log F_Y^{(n)}(Y_i)} \geq -\frac{1}{F_Y^{(n)}(Y_{n-\lfloor kx \rfloor+1,n})}\right\}} \\
&= \sum_{i=1}^{n} \mathbf{1}_{\left\{\left(-\frac{1}{\log F_X^{(n)}(X_i)}\right) \wedge \left(-\frac{1}{\log F_Y^{(n)}(Y_i)}\right) \geq -\frac{1}{\log\left(1-\frac{\lfloor kx \rfloor}{n}\right)}\right\}} \\
&= \sum_{i=1}^{n} \mathbf{1}_{\left\{V_i^{(n)} \geq -\frac{1}{\log\left(1-\frac{\lfloor kx \rfloor}{n}\right)}\right\}}
\end{aligned} \tag{5.21}$$

with $V_i^{(n)}$ defined in (5.11). Owing to the power-series $\sum_{n \geq 0} \frac{t^n}{n+1} = -\frac{\log(1-t)}{t}$, for $|t| < 1$, we can write the following stochastic inequalities for (5.21): there exists $\epsilon > 0$ such that

$$\mathbf{1}_{\left\{\frac{k}{n} V_i^{(n)} \geq \left(1-\frac{k}{n}\frac{x}{2}\left(1-\epsilon\left(\frac{k}{n}x\right)^{\epsilon}\right)\right)\right\}} \leq \mathbf{1}_{\left\{\frac{k}{n} V_i^{(n)} \geq -\frac{k/n}{\log\left(1-\frac{\lfloor kx \rfloor}{n}\right)}\right\}} \leq \mathbf{1}_{\left\{\frac{k}{n} V_i^{(n)} \geq \left(1-\frac{k}{n}\frac{x}{2}\left(1+\epsilon\left(\frac{k}{n}x\right)^{\epsilon}\right)\right)\right\}}$$

uniformly in x on a compact set bounded away from zero. This gives information about the error in the approximation of (5.21) to $\sum_{i=1}^{n} \mathbf{1}_{\left\{\frac{k}{n}\left(V_i^{(n)}+\frac{1}{2}\right)\frac{1}{x}\right\}}$. Specifically, for each $\delta > 0$, there exists $\epsilon' > 0$ chosen arbitrarily small, such that

$$\limsup_n \mathbb{P}\left(\max_{1 \leq i \leq n} |I_i^{(n)}(x) - I_{i,\epsilon'}^{(n)}(x)| > 1-\delta, \text{ for } 0 \leq x \leq T\right) < \delta$$

with intervening

$$I_i^{(n)}(x) = \mathbf{1}_{\left\{\frac{k}{n}V_i^{(n)} \geq -\frac{k/n}{\log\left(1-\frac{\lfloor kx \rfloor}{n}\right)}\right\}} \qquad \text{and} \qquad I_{i,\epsilon'}^{(n)}(x) = \mathbf{1}_{\left\{\frac{k}{n}V_i^{(n)} \geq \left(1-\frac{k}{n}\frac{x}{2}\left(1\pm\epsilon\left(\frac{k}{n}x\right)^{\epsilon}\right)\right)\right\}}$$

The above inequalities thus imply, for each $n \in \mathbb{N}$, with $\delta, \epsilon' > 0$ as before,

$$\mathbb{P}\left(\left|I_i^{(n)}(x) - I_{i,\epsilon'}^{(n)}(x)\right| \leq 1-\delta \text{ for all } i = 1, ..., m; \text{ for } 0 \leq x \leq T\right) > 1-\delta.$$

By noting that

$$\frac{1}{n}\left|\sum_{i=1}^{n} I_i^{(n)}(x) - I_{i,\epsilon'}^{(n)}(x)\right| \leq \frac{1}{n}\sum_{i=1}^{n}\left|I_i^{(n)}(x) - I_{i,\epsilon'}^{(n)}(x)\right| \leq \max_{1\leq i\leq n}\left|I_i^{(n)}(x) - I_{i,\epsilon'}^{(n)}(x)\right|$$

and letting $\epsilon' > 0$ be sufficiently close to zero, we arrive at

$$\mathbb{P}\left(\max_{1\leq i\leq n}\left|I_i^{(n)}(x) - I_{i,\epsilon'}^{(n)}(x)\right| = 0 \text{ for } 0 \leq x \leq T\right) > 1-\delta.$$

Now, invoking a Skorokhod construction, we have that, as $n \to \infty$, almost surely,

$$\sup_{x\in[0,T]}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\left\{V_i^{(n)} \geq -\frac{1}{\log\left(1-\frac{\lfloor kx \rfloor}{n}\right)}\right\}} - \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\left\{\frac{k}{n}\left(V_i^{(n)}+\frac{1}{2}\right)\geq\frac{1}{x}\right\}}\right| \longrightarrow 0. \tag{5.22}$$

Together with (5.20), the above entails the asymptotic, almost sure, approximation of the suitably shifted Fréchet marginals to the unit Pareto marginals. This tells us that, for each x, the normalised sum $\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\left\{\frac{k}{n}\left(V_i^{(n)}+\frac{1}{2}\right)\frac{1}{x}\right\}}$ amounts to the rescaled empirical distribution function $1-F_T^{(n)}(x) := \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\left\{\frac{k}{n}T_{n,n-i}\geq x\right\}}$ and is such that $(1-F_T^{(n)}(x))/q(k/n) = S(1/x, 1/x) + O_p\left(1/\sqrt{nq(k/n)}\right)$. Then, a functional representation of the estimator considered on the basis of the tail empirical processes involved in (5.22) follows through the identification of the order statistics $V_{n,n-i} + 1/2 = T_{n,n-i}$ in relation to the unit Pareto marginals. Such an asymptotic representation is at the origin of the Hill estimator, in particular through the functional

$$\hat{\eta}_H \equiv \frac{n}{m}\int_{T_{n,n-m}}^{\infty}(\log x - \log T_{n,n-m})\, dF_T^{(n)}(x),$$

as well as to the general class of estimators

$$\hat{\eta}_{a,b} = \frac{1}{b}\left(\left[\frac{n}{m}\int_{T_{n,n-m}}^{\infty}\left(\frac{x}{T_{n,n-m}}\right)^{a} dF_T^{(n)}(x)\right]^{b/a} - 1\right),$$

and the result in the theorem thus follows. □

Remark 1. *From the proof of Theorem 3 we find that the assertion involving* $1 - F_T^{(n)}(x) := \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\left\{\frac{k}{n}T_i^{(n)} \geq x\right\}}$ *is essential to establish the approximation* $\left|\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\left\{\frac{k}{n}V_i^{(n)} > \frac{1}{x}\left(1-\frac{k}{n}\frac{x}{2}\right)\right\}} - \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\left\{\frac{k}{n}\left(T_i^{(n)}-\frac{1}{2}\right) > \frac{1}{x}\left(1-\frac{k}{n}\frac{x}{2}\right)\right\}}\right| = o_p(1)$, *for* $k \to \infty, k/n \to 0$ *as* $n \to \infty$. *This is the change in location upon pseudo-observations for estimating the CTD* $\eta \in (0,1]$. *In essence, it follows from Einmahl (1997) and Peng (1999) that,*

$$\sqrt{m}\left(\frac{1}{m}\sum_{i=1}^{n} \mathbf{1}_{1-F_X(X_i)\leq\frac{k}{n}x, 1-F_Y(Y_i)\leq\frac{k}{n}x} - x^{-1/\eta}\right) \xrightarrow{d} W(x,x)$$

with $W(x,x) \equiv W(x)$ *a zero-mean Gaussian process as before, and this corresponds to the properly reduced and standardised stochastic process*

$$\sqrt{m}\left(\frac{\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\left\{\frac{k}{n}T_i^{(n)} \geq x\right\}}}{q\left(\frac{n}{k}\right)} - x^{-1/\eta}\right) \xrightarrow{d} W(x),$$

through the linkage $m/k := (n/k)q(n/k)$, *as* $n \to \infty$. *Upon this development, a suitable second order condition can be imposed which will determine an extra term in the approximation, akin to asymptotic bias. This condition of second order will then pin-down any potential changes in the second order parameter* τ *resulting for the shift by* $1/2$. *It also outlines the key idea to serve as basis for the shaping of Theorem 5.*

5.B Proof of Lemma and Theorem in Section 4.

Proof of Lemma. With the defined V and U it holds that $V(t) = U\left(1/(1-e^{-1/t})\right)$, whence

$$\frac{U(t)}{V^*(t)} = \frac{U^*(t)}{U^*\left(\frac{1}{1-e^{-1/t}}\right)} - \frac{1/2}{U^*\left(\frac{1}{1-e^{-1/t}}\right)}, \qquad U^* := U + \frac{1}{2} \tag{5.23}$$

with the latter term finishing as $t \to \infty$. With $\tilde{q}(t) = t^{-\eta} q_\star(t) U(t)$, $|\tilde{q}(t)| \in RV_{-\tau}$, the second order regular variation for U in (5.14) is expressed as

$$\lim_{t\to\infty} \frac{(tx)^{-\eta} U(tx) - t^\eta U(t)}{\tilde{q}(t)} = \frac{x^{-\tau} - 1}{\tau} \tag{5.24}$$

for all $x > 0$. It implies in turn that with $U^* := U + 1/2$, as $t \to \infty$

$$\frac{(tx)^{-\eta} U^*(tx) - t^\eta U^*(t)}{\tilde{q}(t)} = \frac{x^{-\tau} - 1}{\tau}\,(1 + o(1)) + \frac{x^{-\eta} - 1}{2U(t) q_\star(t)}, \qquad |U^* q_\star| \in RV_{-\tau+\eta}$$

i.e.,

$$x^{-\eta} \frac{U^*(tx)}{U^*(t)} - 1 = \left\{ \frac{x^{-\tau} - 1}{\tau} q_\star(t) \frac{U(t)}{U^*(t)} + \frac{x^{-\eta} - 1}{2} \frac{1}{U^*(t)} \right\} (1 + o(1)) \tag{5.25}$$

Additionally, we note that because $U \in RV_\eta$, $\eta > 0$, we have for any constant $c \in \mathbb{R}$,

$$\frac{U(tx) - c}{U(t) - c} = \frac{U(tx)}{U(t)} \left(1 - \frac{c}{U(t)}\right)^{-1} (1 + o(1)) = \frac{U(tx)}{U(t)} (1 + o(1)).$$

Therefore, with $c = 1/2$ in particular, we get that $U^*(t) \sim U(t)$, as $t \to \infty$, and also that by taking $y = U^*(tx)/U^*(t)$ in the equality $1/(1+y) = 1 - y + y^2/(1+y), y \neq -1$, then relation (5.25) entails

$$x^\eta \frac{U^*(t)}{U^*(tx)} - 1 = -\frac{x^{-\eta} \frac{U^*(tx)}{U^*(t)} - 1}{x^{-\eta} \frac{U^*(tx)}{U^*(t)}} = \frac{x^{-\tau} - 1}{\tau} q_\star(t)(1 + o(1)) + \frac{x^{-\eta} - 1}{2} \frac{1}{U^*(t)} (1 + o(1))$$

Finally, Taylor's expansion of $y/(1 - e^{-y})$ around zero ascertains the result:

$$\frac{U^*(t)}{U^*\left(\frac{1}{1-e^{-1/t}}\right)} = \left(\frac{1/t}{1 - e^{-1/t}}\right)^{-\eta} = 1 - \frac{\eta}{2}\frac{1}{t} + \frac{\eta(1+3\eta)}{24}\frac{1}{t^2} + o(\max(q_\star(t), 1/U^*(t))$$

as $t \to \infty$, whereby we conclude that $U(t) \sim V^*(t)$ from the stated equality (5.23) at the beginning of this proof. □

Proof of Theorem. The proof essentially hinges on translating second order regular variation into extended regular variation for an appropriate function related to the former. With the already defined quantile function V, such that $V(t) = U\left(1/(1 - e^{-1/t})\right)$, and

$\tilde{q}(t) = t^{-\eta} q_\star(t) U(t)$, $|\tilde{q}(t)| \in RV_{-\tau}$, we have that

$$\begin{aligned}\frac{(tx)^{-\eta}V(tx) - t^{\eta}V(t)}{\tilde{q}(t)} &= \frac{(tx)^{-\eta}V(tx) - t^{\eta}U(t)}{\tilde{q}(t)} - \frac{(x)^{-\eta}V(t) - t^{\eta}U(t)}{\tilde{q}(t)} \\ &= \frac{(tx)^{-\eta}U\left(1/(1-e^{-1/tx})\right) - t^{\eta}U(t)}{\tilde{q}(t)} - \frac{(tx)^{-\eta}U\left(1/(1-e^{-1/t})\right) - t^{\eta}U(t)}{\tilde{q}(t)}.\end{aligned}$$

Owing to the second order regular variation for U with index $\eta > 0$ encapsulated in (5.24), which holds locally uniformly for $x > 0$, and by noting that $x(t) = t^{-1}/(1 - e^{-1/t}) \to 1$, as $t \to \infty$, we find the representation:

$$\frac{(tx)^{-\eta}V(tx) - t^{\eta}V(t)}{\tilde{q}(t)} = \frac{\left(\frac{1/t}{1-e^{-1/(tx)}}\right)^{-\tau} - 1}{\tau}(1 + o(1)) - t^{-\eta}\frac{\left(\frac{1/t}{1-e^{-1/(t)}}\right)^{-\tau} - 1}{\tau}(1 + o(1))$$

Now it is only a matter of applying Taylor's expansion followed by judicious manipulation in order to have, for all $x > 0$, the next order representation:

$$\frac{x^{-\eta}\frac{V(tx)}{V(t)} - 1}{q_\star U(t)/V(t)} = \frac{x^{-\tau} - 1}{\tau} - \frac{1}{2t}x^{-\tau-1} + o\left(\frac{1}{t}\right), \tag{5.26}$$

as $t \to \infty$. Moving on to tackling $V^*(t) = V(t) + 1/2$, we consider the above development to approaching the extended regular variation property as follows:

$$\begin{aligned}\frac{(tx)^{-\eta}V^*(tx) - t^{\eta}V^*(t)}{\tilde{q}(t)} &= \frac{(tx)^{-\eta}V(tx) - t^{\eta}V(t)}{\tilde{q}(t)} + \frac{t^{-\eta}}{2}\frac{x^{-\eta} - 1}{\tilde{q}(t)} \\ &= \frac{x^{-\tau} - 1}{\tau} + \frac{t^{-\eta}}{\tilde{q}(t)}\frac{x^{-\eta} - 1}{2} - \frac{1}{t}\frac{x^{-\tau-1}}{2} + o(t^{-1}) + o(1).\end{aligned}$$

Given that the present setting of asymptotic independence, the range $\eta \leq 1$ is to be imposed, the third order term that trickled down from (5.26) becomes negligible (note that $|t^{\eta}\tilde{q}(t)| \in RV_{-\tau+\eta}$ and $\eta < 1 + \tau, \tau > 0$), thus resulting in the following representation for V^*:

$$\frac{x^{-\eta}\frac{V^*(tx)}{V^*(t)} - 1}{q_\star(t)U(t)/V^*(t)} = \frac{x^{-\tau} - 1}{\tau} + \frac{1}{q_\star(t)U(t)}\frac{x^{-\eta} - 1}{2} + o\left(\frac{1}{q_\star U(t)}\right)$$

Under the conditions of this theorem, Lemma 4 enables replacement of U with V^* everywhere in the expansion above and the desired result of second order regular variation for V^* arises. Specifically,

$$\frac{V^*(tx)}{V^*(t)} = q_\star(t)x^{\eta}\frac{x^{-\tau} - 1}{\tau} + \frac{\eta}{2}\frac{1}{V^*(t)}x^{\eta}\frac{x^{-\eta} - 1}{\eta} + o(q_\star(t)) + o\left(\frac{1}{V^*(t)}\right), \quad t \to \infty$$

□

Chapter 6

Seasonal and regional variability in the extremal asymptotic dependence in daily rainfall

The statistical properties of the extremes might not necessarily be the same as the bulk of the data, a feature that the bivariate Gaussian copula clearly demonstrates by potentially exhibiting strong dependence in the central part of the distribution underlying the data, but asymptotic independence in the extremes, except for the limiting case with asymptotic dependence when the correlation coefficient $\rho = 1$. This chapter also addresses question 2 in the thesis aims by considering how the dependence in the extremes potentially differ from the regular part of the data. This is to better understand how the risk of co-occurring large events changes for increasingly extreme events, hence extending the work and conclusions from Chapter 3. Since only the very largest values are considered, the work here is based on the Extreme value framework, which shifts the focus from the mean of the distribution to the tail.Since EVT is specifically developed to model the tail behaviour of the observed largest values and derive the best estimate for values larger than these, it provides support to comment on the behaviour for larger values than the ones we have observed.

6.1 Overview

Studying the spatial behaviour of extreme rainfall has so far mostly been focused on fitting a model from which one can estimate return levels. The most commonly used method to address this are *max-stable* processes (Haan and Ferreira (2006)), which are closely related to the univariate and multivariate extreme value distributions (Tawn et al. (2018)). The extremal dependence structure for the random process is summarised by the exponent measure $V_N(\cdot)$,

which in theory can be a function of N finite, but in practice is most often defined for $N = 2$ because of the very complex calculations or explosion in the number of terms involved when trying to extend it to more dimensions. This means that even if one in theory directly could estimate the dependence between 3 or more variables, in practice this can only be estimated for pairs of variables and then averaged together. There are several well established models for the max-stable processes, however a common drawback with them is that they often do not allow for asymptotic independence (e.g. Huser and Davison (2014)).

This inability to in practice attain both dependence classes, even though possible in theory, is a common issue for extremal dependence models. That is, asymptotically dependent max-stable models never reach the asymptotically independent value and vice versa for asymptotically independent models. This poses an issue for applications, such as rainfall modelling, where we expect the dependence to decrease both as the distance increases and potentially also as we move into extremes much larger than the ones we have observed. To make the distinction for the second issue, that the extremal dependence estimated holds for the observed levels of extremes and values a bit larger, but potentially not for infinitely large values, Huser and Wadsworth (2020) renamed these models *subasymptotic* instead of asymptotic. This renaming highlights the fact that the EVT framework allows us to extrapolate our findings 'a bit' outside our measured range, but we cannot be sure that this holds for 'much larger' values. Both the definition of 'a bit' and 'much larger' will depend on the application and the statistician performing the analysis.

To address the issue of changing dependence structure with distance, that is the variable exhibits some dependence for distances $d \approx 0$ but independence as $d \to \infty$, and most likely at distances far shorter than that, several methods have been proposed. Shooter et al. (2021) approach this problem by introducing a spatial conditional extremes model, building on the work by Shooter et al. (2019) and Wadsworth and Tawn (2019). The model includes two parameters α, β, both which depend on the distance between locations, and are either assumed piecewise linear or to follow a parametric function. The parametric function however does not allow for the boundary values $(\alpha, \beta) = (1, 0)$ corresponding to the asymptotically dependent case, effectively reducing the model to only allow for asymptotic independence. Wadsworth and Tawn (2012) introduced the *inverted max-stable process*, which has asymptotically independent extremes instead of asymptotically dependent extremes as is the case for the regular max-stable processes, but still do not allow for both dependence classes. This therefore still remains an active area of research that is mentioned but not further explored here.

Most of the previous studies on spatial extremal rainfall has been focused outside of the

tropics and using asymptotically dependent models, such as the UK (Atyeo and Walshaw (2012)) which used a Bayesian hierarchical model in order to model many stations at the same time. This model would be unsuitable for most of Africa because it requires relatively many stations (in there 25), with very similar rainfall structure and long time series. In Shang et al. (2011), a max stable process was fitted to the winter maximum precipitation collected in the US and Thibaud et al. (2013) also used a max-stable process, but fitted to threshold exceedances instead of annual maxima and on data collected over Switzerland. A comparison of the performance of using Latent variable, copula and max-stable models was performed in Davison et al. (2012) on annual maximum daily rainfall collected over Italy. In there they concluded that a Latent variable method performed the worst in modelling the joint distribution, hence a copula or max-stable model was preferred. Since all of these results were obtained with asymptotically dependent models and for non tropical regions, we cannot compare our results with them. However, if one wants to do a similar analysis, either the copula method or the max-stable on threshold exceedances would be the most suitable considering our sample sizes.

There has been very limited amount of work done over the continent of Africa, stemming from the re-occurring issue of data scarcity and poor representation of extremes in satellite and climate model data. A recent paper by Debusho and Diriba (2021) applied the *conditional multivariate exceedance model* by Heffernan and Tawn (2004) to daily rainfall gauge data collected over South Africa between 1991-2019. They however faced the issue of a very sparse gauge network, with some station-pairs exhibiting negative dependence due to being located in different seasonal rainfall regimes.

Sang and Gelfand (2009) also addressed the spatial extreme behaviour over South Africa. Their aim was to construct a model with given marginals and some dependence between locations, therefore naturally chose a copula model. They specifically chose to work with a Gaussian copula, hence by construction assumed asymptotic independence (Sibuya (1960)). The model is assumed to have an exponential correlation function and the fitted parameter is around 0.042, indicating a small correlation range.

Blanchet et al. (2018) is one of the very few studies in west Africa, specifically the Sahel region. They used daily data from 1950-2014, hence nearly the exact same period as the data set previously used in this thesis covers, in the central Sahel and the very dense but spatially small AMMA-CATCH rainfall network in Niger which has 5min measurements collected since 1990. Arguing that an increase in the number of storms should not imply a decrease in the extent of them, they choose the asymptotically dependent max-stable Brown-Resnick model for their

analysis, but acknowledge that their findings might not be suitable to extrapolate from. They calculated the probability of two locations co-exceeding the 99% individual quantiles, which their chosen model however overestimated compared to the pairwise estimates for distances less than 300km, a distance from which one can question if there is any true dependence left.

To determine the changes in extremal dependence for increasing distances for rainfall processes, and in particular convective tropical rainfall, the use of an asymptotically independent model seems most appropriate. This is since convective rainfall clouds are usually formed and released over a relatively short period of time and rain usually falls as intense showers, resulting in large rainfall amounts over a very local area. The aim here is to understand how the dependence changes with distance, and specifically at what distance locations can be assumed independent. This is different from the aim of most of the previously mentioned papers where the focus is instead on estimating return levels (see Section 2) which requires an asymptotically dependent model. This aim, together with the results in Sang and Gelfand (2009), and our knowledge about the rainfall climate over our study region, provides support for using the reduce bias coefficient of tail dependence estimator $\tilde{\eta}$ developed in Chapter 5 in the following analysis.

6.2 The station selection process

To estimate the spatial extent of extremal dependence and how this varies with the monsoon phase and presence of advective rainfall, the same daily rainfall data set presented in Chapter 3 will be used, comprising 590 rain gauges unevenly distributed over Ghana with daily measurements between 1940-2017. With only the largest observations being considered in the EVT statistical setting, and therefore only a very small sample size, our analysis becomes much more sensitive to missing values since removing a few of the largest values can significantly change the estimate. In the analysis in Chapter 3, a large proportion of the stations could be included based on the argument that the potential bias from missing values would be averaged out, something that cannot be relied upon here. Therefore only a small subset of high quality stations containing the smallest possible number of missing values are included in the analysis performed here.

Following the analysis in Chapter 3, only data from 1950 are used since very few stations were available before that. Similarly to Chapter 4, only April, June, August and September will be presented as representative months for the different monsoon phases. The country is split into two regions by the 8°N latitude line which separates the uni- and bi-seasonal rainfall regimes (Dunning et al. (2016)). Due to the large difference in the density of stations in the southern and northern part of the country, slightly different selection criteria will be used to define the

optimal stations. Initially, all stations with more than 50% missing values are removed since these have less than 30 years of data, which is the standard minimum for climate analysis, leaving us with 246 stations. This might still appear rather liberal, however this is set as the lowest requirement, and for nearly all station-pairs this will be improved by carefully selecting the optimal pair.

In addition, a requirement of maximum 10% missing values allowed in any month is imposed, with the station kept in the analysis but the data from that particular month is discarded if more is present. Usually, either a full month is missing or just a single day so this has a very small impact. In the southern region, between 12-15 stations has about 30 sample points discarded with this restriction and in the north it is either 1 or 2 stations with 5-26 sample points removed.

In order to estimate how the asymptotic independence estimate varies with distance, a few centre stations, S_c, are selected and for each of these stations a set of pair stations $S_{c,p}$. Each of the pair stations are located in a separate 10km distance bin, hence only one station-pair exists for each centre station and distance (see last step in Figure 3.16 for distance bin visualisation). From these station-pairs, a similar approach as in Chapter 3 can be taken by estimating the dependence in 10km distance bins, up to a distance of 200km away from the centre station.

For the southern region, there are 200 stations with less than 50% missing values and 61 of these stations have at least one pair station within each 10km distance bin. This set of 61 stations will be denoted 'distance-complete'. From this rich set of distance-complete stations, 3 stations are chosen with the least number of missing values and the identifiers of these stations are: ACC, AKU, SAL. For the northern region, there are only 43 stations and none of them are 'distance-complete'. Additionally, the station with the best distance coverage has many missing values, making it unsuitable to be the centre station for the analysis. The stations TLE is therefore selected, although only having 15 instead of 17 distances covered, it only contains 43 missing values which is determined to be more important in this case.

Since only days where both stations have measurements can be used, the optimal pair station will be the one that has the minimum number of non-overlapping missing values. Therefore, for both of the regions the pair station is selected that jointly with the centre station has the least number of missing values, thereby always extracting the maximum number of bivariate sample points. Table 6.1 presents the number of bivariate sample points for the four stations in S_c and their associated pair stations $S_{c,p}$ for each month. Figure 6.1 marks the location of the centre pairs S_c with filled circles and their associated stations $S_{c,p}$ with the same colour, but different shapes. ACC stations are blue triangles, AKU has orange crosses, SAL

green up-side-down triangles and TLE red stars. The south region stations have some common pair stations, which is to be expected since the matching station with the least number of missing values is always selected.

We note that even though there are similar number of samples for all months (accounting for the difference in number of days of the month), there are some small differences between them, highlighting that the missing values are not spread out evenly over the year. There are generally more missing values in April, but not always. Another thing to notice is the generally larger number of missing values in the northern region (TLE), however there are at least the equivalence of 30 years of data for all station-pairs except the shortest distance in September, which is just slightly below.

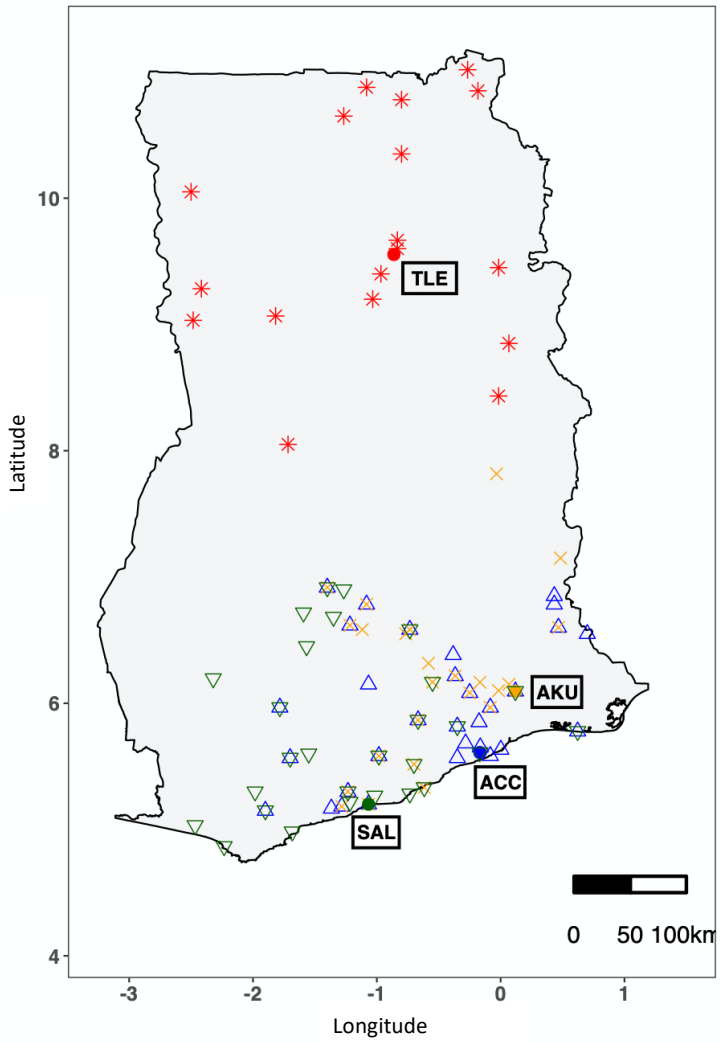


Figure 6.1: Map of Ghana with all the stations in S_c (filled circles) and $S_{c,p}$ marked. Stations associated with ACC are blue up triangles, AKU orange crosses, SAL green down triangles and TLE red stars.

Distance (km)	ACC				AKU				SAL				TLE			
	April	June	Aug	Sep	April	June	Aug	Sep	April	June	Aug	Sep	April	June	Aug	Sep
0-10	1200	1319	1240	1230	1349	1380	1426	1440	1170	1229	1240	1200	900	1020	992	840
10-20	1828	1918	2014	1920	1465	1465	1419	1436	1320	1257	1333	1170	1950	1919	1984	1859
20-30	1320	1379	1302	1320	1499	1649	1674	1619	1980	1980	2046	1950	1470	1530	1581	1560
30-40	1920	1979	2046	1950	1469	1589	1642	1525	1440	1559	1549	1440	0	0	0	0
40-50	1440	1499	1519	1500	1827	1859	1922	1828	1890	1920	1891	1830	1020	1110	1147	1110
50-60	1828	1859	1922	1828	1769	1799	1829	1770	1770	1740	1798	1680	0	0	0	0
60-70	1949	2038	2108	2040	2009	2038	2108	2040	1200	1290	1302	1260	0	0	0	0
70-80	1740	1799	1829	1770	1709	1769	1767	1620	1770	1770	1798	1680	0	0	0	0
80-90	1950	1979	2015	2040	1378	1409	1457	1440	1890	1919	1983	1828	1200	1230	1271	1200
90-100	1830	1919	1891	1860	1859	1918	1983	1858	1860	1860	1829	1740	2010	2040	2108	2010
100-110	1980	2039	2108	2010	1739	1798	1829	1710	1980	2039	218	2010	0	0	0	0
110-120	1710	1676	1704	1679	1769	1740	1798	1710	1950	1920	2015	1829	1800	1920	1922	1860
120-130	1920	1979	2046	1980	1739	1769	1828	1770	1740	1770	1767	1590	1290	1380	1519	1380
130-140	1980	2009	2108	2040	1859	1919	1891	1860	2040	2009	2108	2010	1979	1948	2077	1980
140-150	1230	1349	1240	1260	1650	1709	1581	1620	1830	1800	1984	1860	2040	2040	2107	2010
150-160	1830	1889	1953	1830	1799	1829	1891	1830	1980	1950	2015	1920	1200	1170	1209	1170
160-170	1770	1829	1891	1830	2009	2039	2108	2010	2009	2039	2108	2010	1170	1170	1146	1110
170-180	1800	1888	1890	1827	1949	1979	2046	1980	2010	2010	2077	1980	1320	1380	1457	1200
180-190	1890	1919	2015	1890	1769	1859	1921	1830	1500	1500	1550	1469	1950	1950	1953	1919
190-200	1950	1949	2015	1980	2009	2039	2107	2040	2010	1980	2015	2010	2040	2039	2108	2010

Table 6.1: Total number of bivariate sample points for each optimal station-pair in $S_{c,p}$ and month. The maximum possible number for April, June and September is 2040 and for August 2108 (time period 1950-2017).

6.3 Stationarity and clustering in time

Before investigating the tail dependence, some initial data exploration needs to be performed to establish stationarity and independence in time for the individual time series. This will complement the work done in Section 3.3.1 by focusing on the daily extreme values instead of the bulk data statistics on annual or monthly aggregates. By plotting the monthly time series for each of the stations in S_c, changes in occurrence and magnitude of extreme values can be analysed, and thereby detect non-stationarity. The processes generating the extremes must not necessarily be stationary over our time period despite the total annual amount has been largely unchanged (see Figure 3.8). Due to the very short time series, any changes in the magnitude of the extreme values will not be quantified, by for example estimating the extreme value index for different time periods. Changes in magnitude will instead solely be visually inspected and commented on from the time series. Changes in frequency are however more difficult to visually inspect from time series and can be estimated non-parametrically, making it feasible despite our small sample size. To estimate changes in frequency, the *scedasis function* developed in Einmahl et al. (2016) will be used, which builds on the tail equivalence function c introduced in Haan et al. (2015).

The skedasis function is based on estimating the kernel density for the largest observations and thereby gain a graphic representation of how frequent and even these extreme observations occur. Figure 6.2 demonstrates how this works and the limitations with this method. The sample points are marked by the black points and the blue line is the kernel density curve for the whole sample, below denoted by $G(s)$. We can view the kernel density as a 'sum' of the individual density curves associated with each individual sample point, which are plotted as orange lines. Peaks in the blue curve indicates that there are several sample points in the vicinity of that location (high frequency) and a value close to 0 that there are no sample points around that value. If the sample points are evenly distributed and the orange density curves have an appropriate width, the blue line will be horizontal, except at the ends which are outside the range of possible sample values.

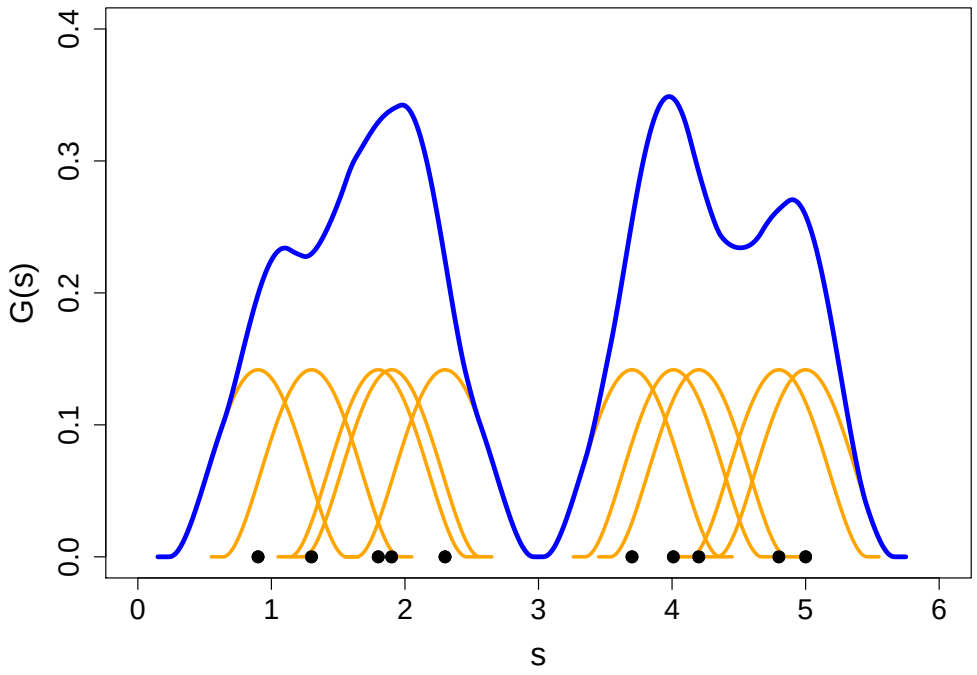


Figure 6.2: Kernel density estimate for the whole sample (blue line) and the individual sample points (orange line) with the sample points marked by the black points.

Below follows the formal definition of the skedasis functions and the mathematical definition of the kernel density function. The skedasis function is defined in the following way:
Let $i = 1, ..., n$ be the time points at which we collect the independent observations $X_1^{(n)}, ..., X_n^{(n)}$ which follow various continuous distribution functions $F_{n,1}, ..., F_{n,n}$ that share a common right end point $x^* = \sup\{x : F_{n,i}(x) < 1\} \in (-\infty, \infty]$, and there is a continuous distribution function F with the same right end point. Additionally there exists a continuous positive function c defined on $[0, 1]$ such that

$$\lim_{x \to x^*} \frac{1 - F_{n,i}(x)}{1 - F(x)} = c\left(\frac{i}{n}\right)$$

uniformly for all $n \in \mathbb{N}$ and all $1 \leq i \leq n$, and

$$\int_0^1 c(s)\mathrm{d}s = 1.$$

The case $c \equiv 1$ corresponds to the uniform density, meaning that we have 'homoscedastic extremes' or equivalently equal tail frequency. The function c is here estimated by using a kernel density estimation function G, which is a weighting function to non-parametrically estimate the density. Formally, let G be a continuous, symmetric kernel function on $[-1, 1]$ such that $\int_{-1}^1 G(s)\mathrm{d}s = 1$ and set $G(s) = 0$ for $|s| > 1$. Further let $h := h_n > 0$ be a bandwidth (width of the orange lines in Figure 6.2) such that $h \to 0$ and $kh \to \infty$ as $n \to \infty$, k as usual being

the number of top order statistics. The function c is then estimated by

$$\hat{c}(s) = \frac{1}{kh} \sum_{i=1}^{n} \mathbb{1}_{X_i^{(n)} > X_{n,n-k}} G\left(\frac{s - i/n}{h}\right) \tag{6.1}$$

and therefore only exceedances at a maximum distance of h from s are included in the estimate at the time point s.

There exists a variety of commonly used kernel functions with different shapes, giving different proportional weight to close or distant points (Figure 6.3). The uniform function assigns the same weight to all points within the bandwidth window whereas all other functions give more weight to points closer to s. We will take G to be the commonly used *biweight kernel* defined by

$$G(x) = \frac{15(1 - x^2)^2}{16}, \quad x \in [-1, 1]$$

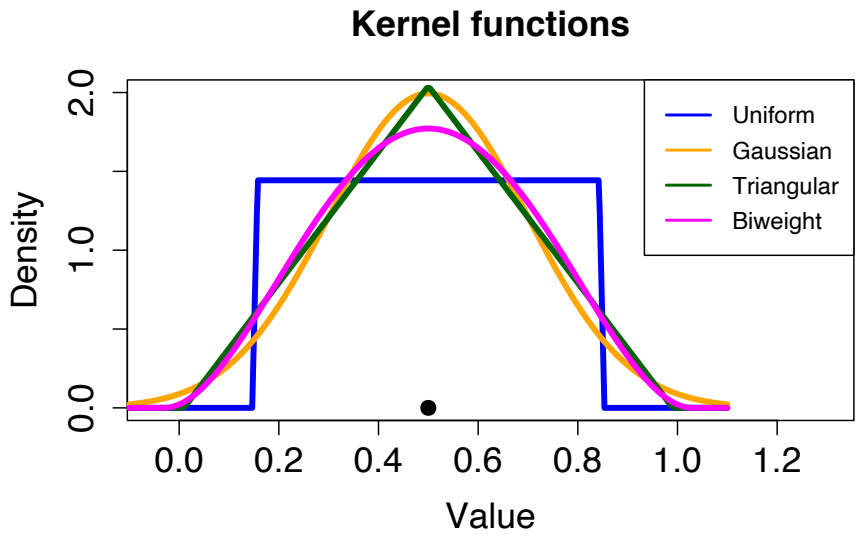


Figure 6.3: Examples of different kernel density functions with the same bandwidth.

In many environmental applications, clustering of high values are common and will result in the assumption of independent sample points being invalid. This is frequently occurring for high temperatures (Winter (2016)), as noticed with the past years heatwaves, due to the slower and smoother processes but can also be observed in rainfall despite its stochastic nature. The skedasis function described above cannot easily be used to distinguish if two extreme observation occur in two consecutive days, or if there are a few days between because of the smoothness of the function. The first case would violate the independence assumption whereas the second would not because the autocorrelation of tropical rainfall is only a couple of days (Figure A.2).

A measure of the presence of clusters in consecutive days is the *extremal index*, $\theta \in [0, 1]$, whose inverse, θ^{-1}, is roughly a measure of the mean cluster size. $\theta = 1$ indicates that there is one exceedance within each cluster and the extremes are therefore independent, whereas a smaller value corresponds to more than one exceedance within the cluster and thereby violates the independence assumption.

To avoid this issue, the most common method is to *decluster* the data by enforcing that only one extreme value exists within each cluster. To identify the cluster and to get an estimate on the presence of clusters two widely adopted methods exists: the runs method (Smith and Weissman (1994)) and the interval method (Ferro and Segers (2003)). In the runs method, we define the threshold u over which only one exceedance is allowed within a cluster, and a run length r which is the minimum number of non-exceedances required between two clusters. The extremal index can also be estimated through the *runs estimator* (Haan and Ferreira (2006)), where for $1 < l < n$ and $N(u_n) := \sum_{i=1}^{n} \mathbb{1}_{X_i > u_n}$,

$$\hat{\theta} := \frac{1}{N(u_n)} \sum_{i=1}^{n-l} \mathbb{1}_{X_i > u_n} \mathbb{1}_{X_{i+1} \leq u_n} ... \mathbb{1}_{X_{i+l} \leq u_n}$$

For this estimator both the threshold and the run length needs to be decided by the user and the choice of run length can be rather arbitrary in certain applications, as highlighted by Hsing (1991). This prompted Ferro and Segers (2003) to develop a method that estimates the extremal index without specifying the run length but instead inferring it from the data. This method builds on the limit distribution of the interexceedance times between threshold exceedances and equating theoretical moments with their empirical counterparts.

Since we have a good understanding of the physical processes behind our data and therefore know the length required between exceedances, the runs method will be used. The same approach as Roth et al. (2014) will be taken by using the 95% quantile of the non-zero values as the threshold for each location, and require there to be at least 1 day between exceedances. For univariate Peak-over-threshold methods the non-maximum exceedances within a cluster can be replaced by the threshold values, since these are ignored in the estimation process of the extreme value index (see Section 2). In our multivariate setting however all non-zero measurements are included in the coefficient of tail dependence (CTD) estimation since there might be a large value at the other location, hence the non-maxima sample point will be removed instead of replaced.

In the analysis, all time series are declustered before estimating the CTD but only results

from our base stations S_c are presented here to demonstrate the presence of clusters and the effect on the extremal index when removing them.

6.3.1 South region

Figure 6.4-6.6 shows the full monthly time series before declustering to the left and the associated skedasis function estimated by Equation (6.1) to the right, with $k = \lfloor 0.05 * n^* \rfloor$ where n^* is the number of non-zero observations and $\lfloor x \rfloor$ denotes the integer part of x. From this, the following k values for the four chosen months for each base station are obtained; ACC (21, 48, 22, 26), AKU (27, 45, 22, 36) and SAL (27, 55, 28, 31). The much and slightly larger values of k in June and September respectively are due to the larger proportion of rainy days (Figure 3.11).

Consistent for all stations is that there is not a clear trend in neither the frequency nor the magnitude of the extreme observations but there has been some variability over time, especially in the frequency. A decrease in frequency of large observations around the year 1990 can be seen for all stations in June coupled with a slight decrease in the magnitude as well. Except for this, there are no clear common patterns between the three stations, indicating that there has not been a regionally common change in the extreme rainfall distribution, but mostly local fluctuations. The close to 0 value in the skedasis function for the boundary years, but not in the centre, is due to the shrinking bandwidth window near the edges. For year close to 1950, only *later* years can be included in the estimate of 6.1 since no data for earlier years are available, hence reducing the possible number of extremes. The opposite is true for years close to 2015, where only data from earlier years are available.

ACC (Figure 6.4) has larger variability in frequency compared to the other two stations and a decrease in the most extreme values in September. SAL (Figure 6.6) on the other hand has a generally stable fluctuation around the $c = 1$ value, with the exception of the large dip in April and peak in September around 1990. AKU (Figure 6.5) has a decrease both in magnitude and frequency in September since 1990, but the frequency has recovered in the later years. Noteworthy is the lack of clear decrease in the 1970s and early 80s during the Sahelian drought (see Section 3.1), further showing that the Guinea Coast was relatively unaffected by this.

Since the non-stationarity in the extremes appears to mainly be in the frequency and not the magnitude, and the skedasis function is mostly fluctuating around 1, we will assume our time series to be stationary in time in order to keep a sufficiently large sample size. Incorporating non-stationarity into the estimation process, by for example including covariates such as

sea surface temperature or other large scale drivers could be a potential extension of this work. Slater et al. (2021) gives an extensive review of potential drivers and methods for handling non-stationarity in hydrological applications.

To address the independence assumption, Figure 6.7 displays the scatter plots of pairwise realisations of X_t, X_{t+1}, with the rainy days 95% quantile represented by the vertical and horizontal lines, for the three south stations. Since the observations for each year are independent of each other, consecutive exceedances are allowed if there is a year break between them. Hence a month with d days will result in $d-1$ points. Points in the upper right corner mark temporal clusters since two consecutive days had measurements above the high threshold, hence the smaller of these values will be excluded to obtain independent observations. To demonstrate the impact of removing these, the extremal index will be estimated with the runs method before and after this declustering.

As one would expect from the mainly convective nature of west African rainfall, most of the extreme values are isolated in time, which can be seen from the low number of points in the upper right corner. The effect of removing these few occurrences is seen in Table 6.2. Before declustering, most time series have nearly independent extremes as indicated by a value close, but not equal, to 1 and after they are as expected equal to 1. Despite the small difference and low number of sample points removed, the declustered time series will be used to not give a couple of events more weight than the rest, given the very few sample points considered for the subsequent CTD estimation.

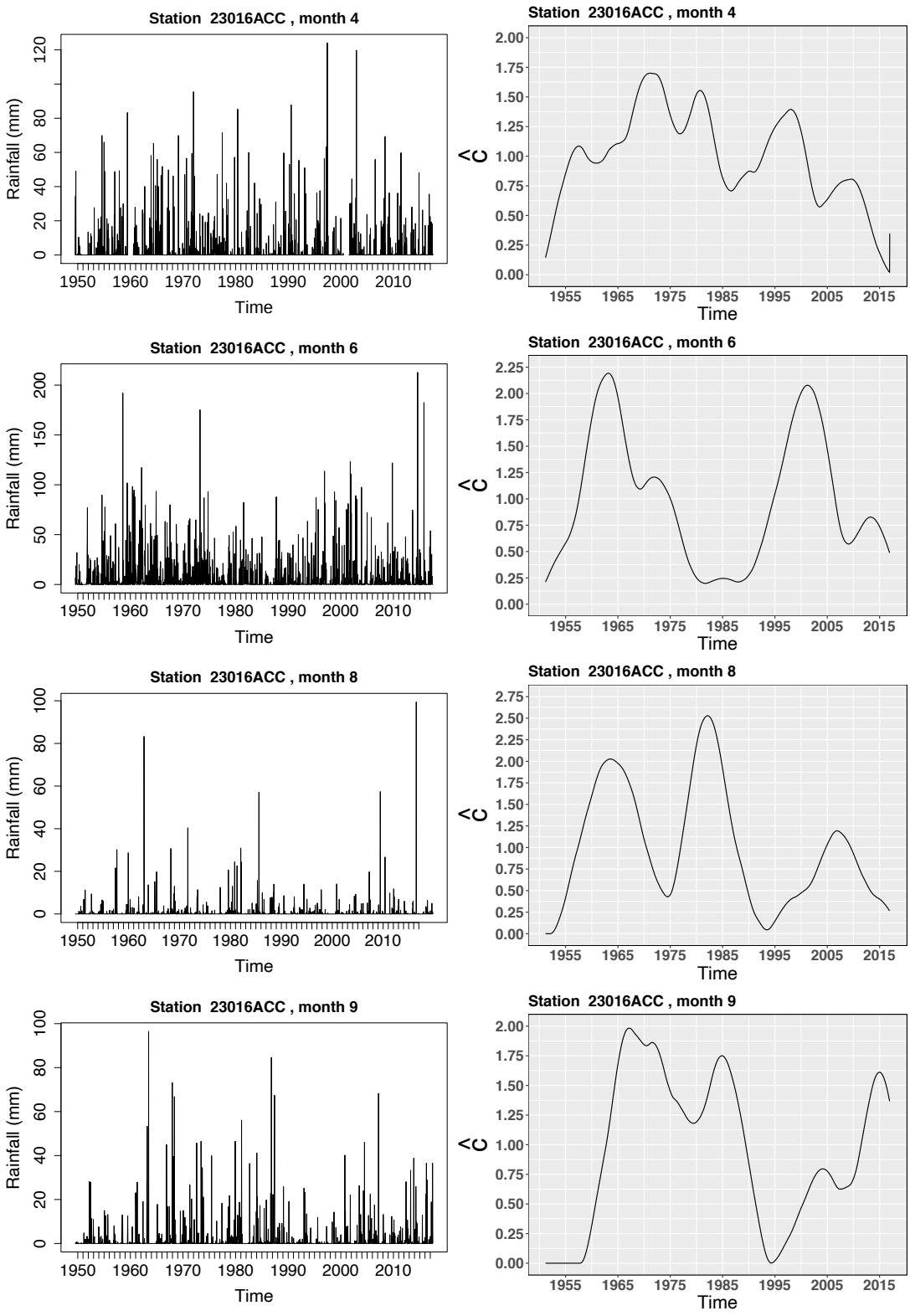


Figure 6.4: Time series of daily rainfall for the station ACC for (top to bottom) April, June, August and September. (Left) Time series with blank areas corresponds to missing values and (right) estimated skedasis function of exceedances above the 95% rainy day quantile.

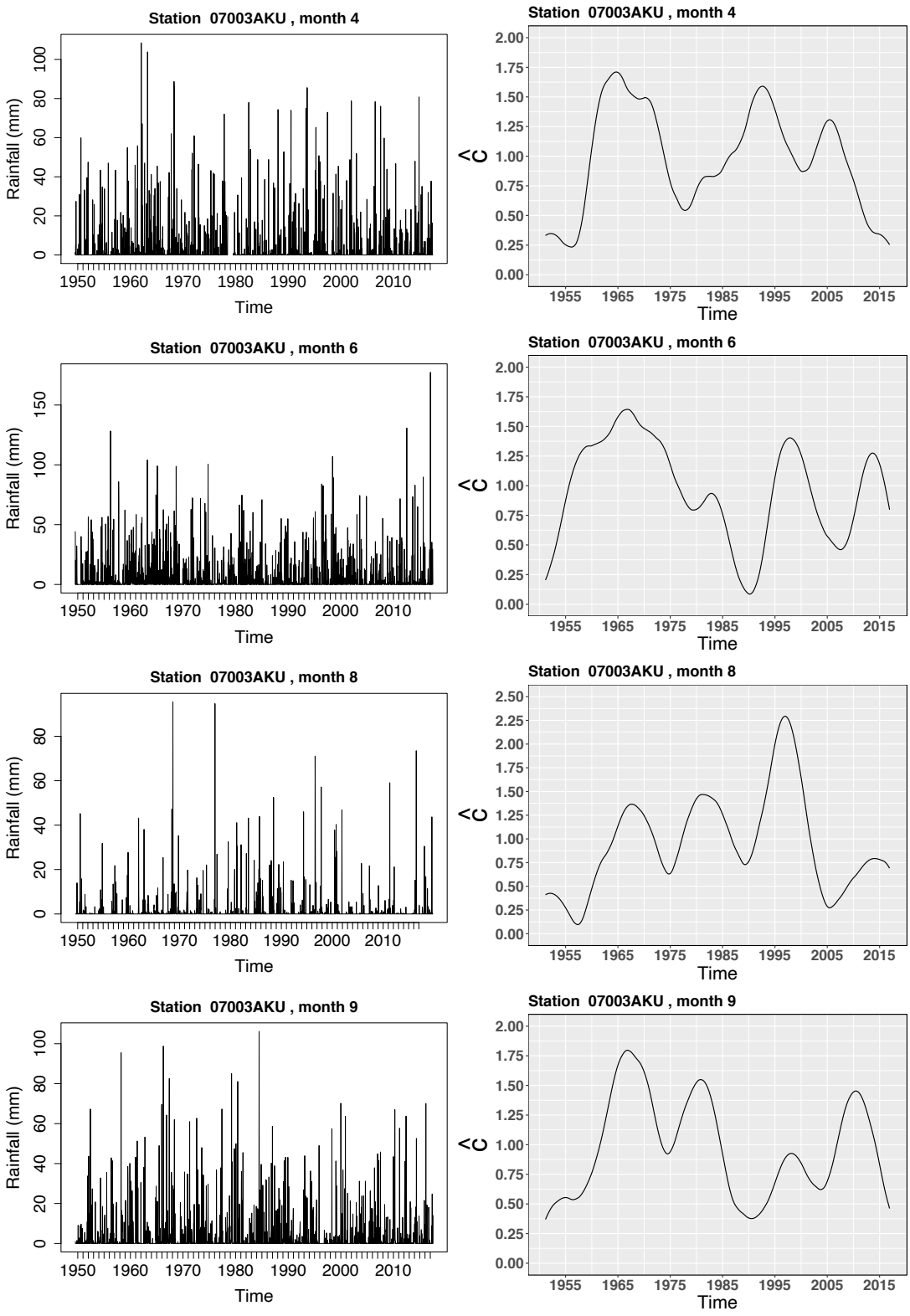


Figure 6.5: Time series of daily rainfall for the station AKU for (top to bottom) April, June, August and September. (Left) Time series with blank areas corresponds to missing values and (right) estimated skedasis function of exceedances above the 95% rainy day quantile.

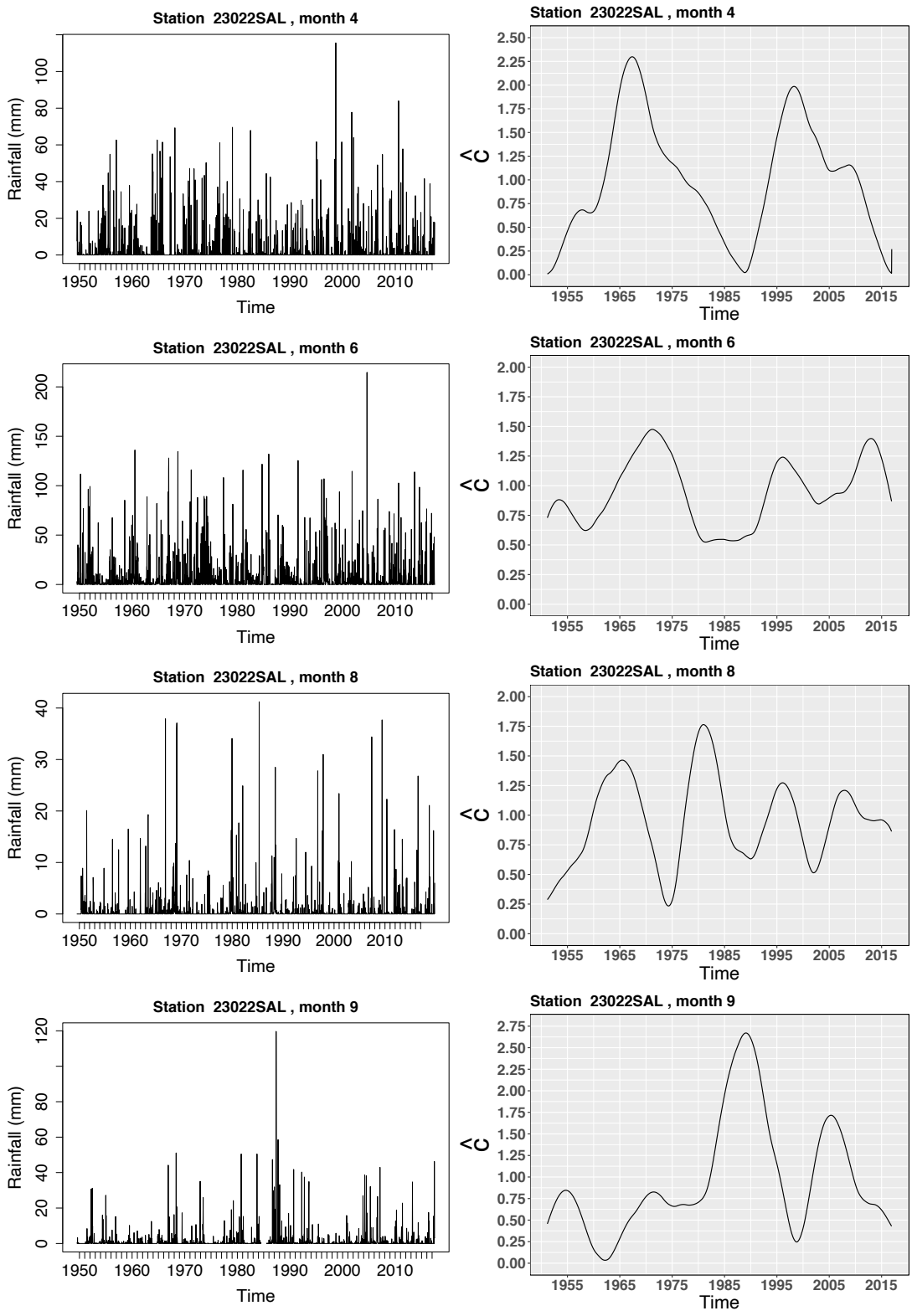


Figure 6.6: Time series of daily rainfall for the station SAL for (top to bottom) April, June, August and September.(Left) Time series with blank areas corresponds to missing values and (right) estimated skedasis function of exceedances above the 95% rainy day quantile.

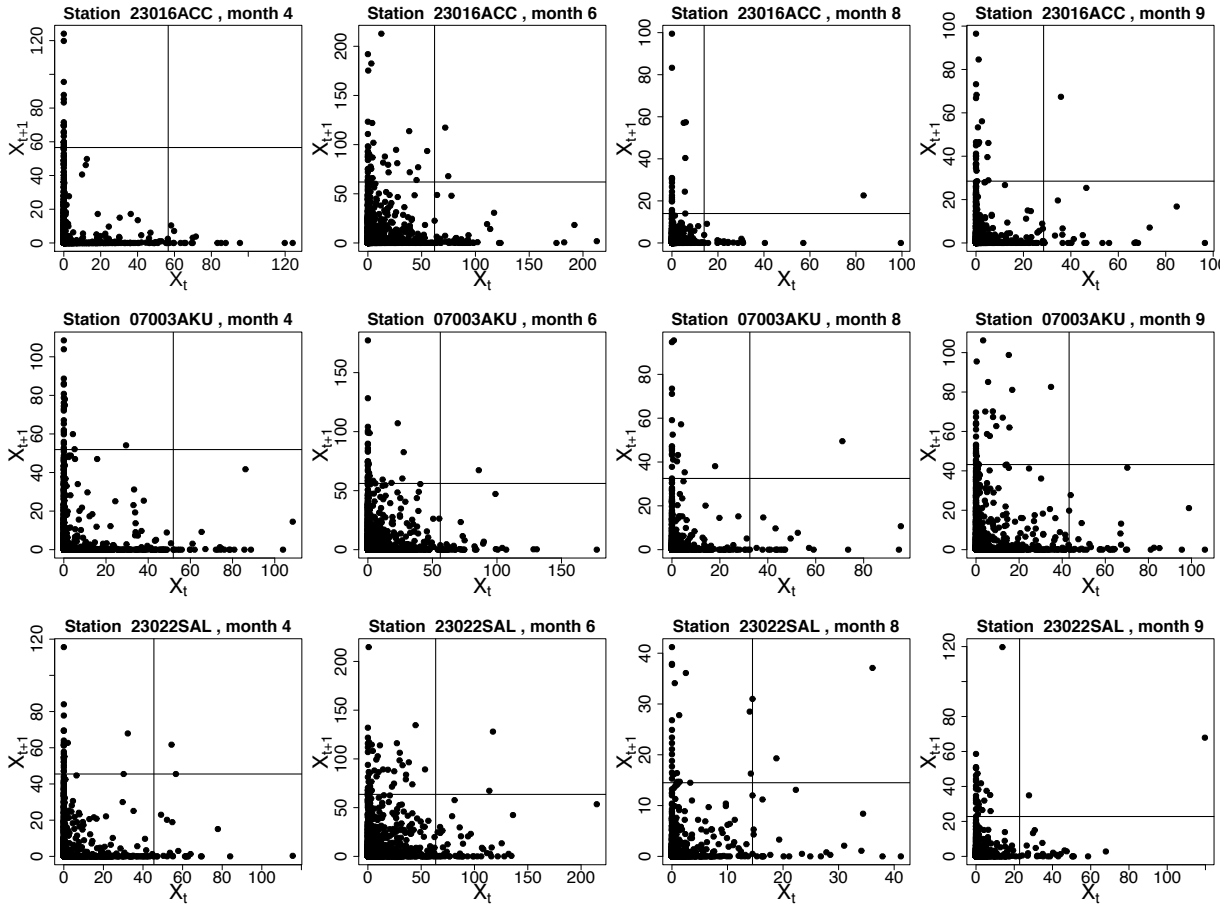


Figure 6.7: Scatter plots for X_t and X_{t+1} from the station ACC (top), AKU (middle) and SAL (bottom) for (left to right) April, June, August and September. Lines marks the rainy days 95^{th}percentile.

	ACC				AKU				SAL			
	April	June	Aug	Sep	April	June	Aug	Sep	April	June	Aug	Sep
Before	1	0.915	0.905	0.96	0.962	0.953	0.952	1	0.962	0.926	0.926	0.933
After	1	1	1	1	1	1	1	1	1	1	1	1

Table 6.2: Values of the extremal index before (top) and after (bottom) declustering the south region time series, using the runs method requiring 1 day between observations above the 95% non-zero rainfall quantile.

6.3.2 North region

Similarly to above, Figure 6.8 shows the full monthly time series before declustering to the left and the associated skedasis function to the right. Using the same method to determine k, the number of exceedances for the skedasis function for the four months are $k = (20, 38, 49, 57)$.

In contrast to the southern region, in the 1970s during the Sahelian drought a drop in the frequency of extreme observations can be seen for all months except August, with this being especially pronounced in June. Similarly to the southern region there is clear lack of extreme values around 1990, however most notably here in April instead of June. There is also a peak in the frequency of extreme observations during the 1960s in September, which also has been noted in Nicholson et al. (2018). The magnitude of the extremes has however remained fairly constant for the entire time period and all months. The overall pattern is therefore a non-regular variability around the uniform 1 line, supporting the assumption of approximate stationarity in the extremes.

Just as for the southern region, Figure 6.9 demonstrates that there are very few occurrences of consecutive days recording extreme amounts, which is confirmed by the close to 1 extremal index values in Table 6.3. Based on the same arguments as for the southern region and for consistency, the analysis in the next section will be based on the declustered time series.

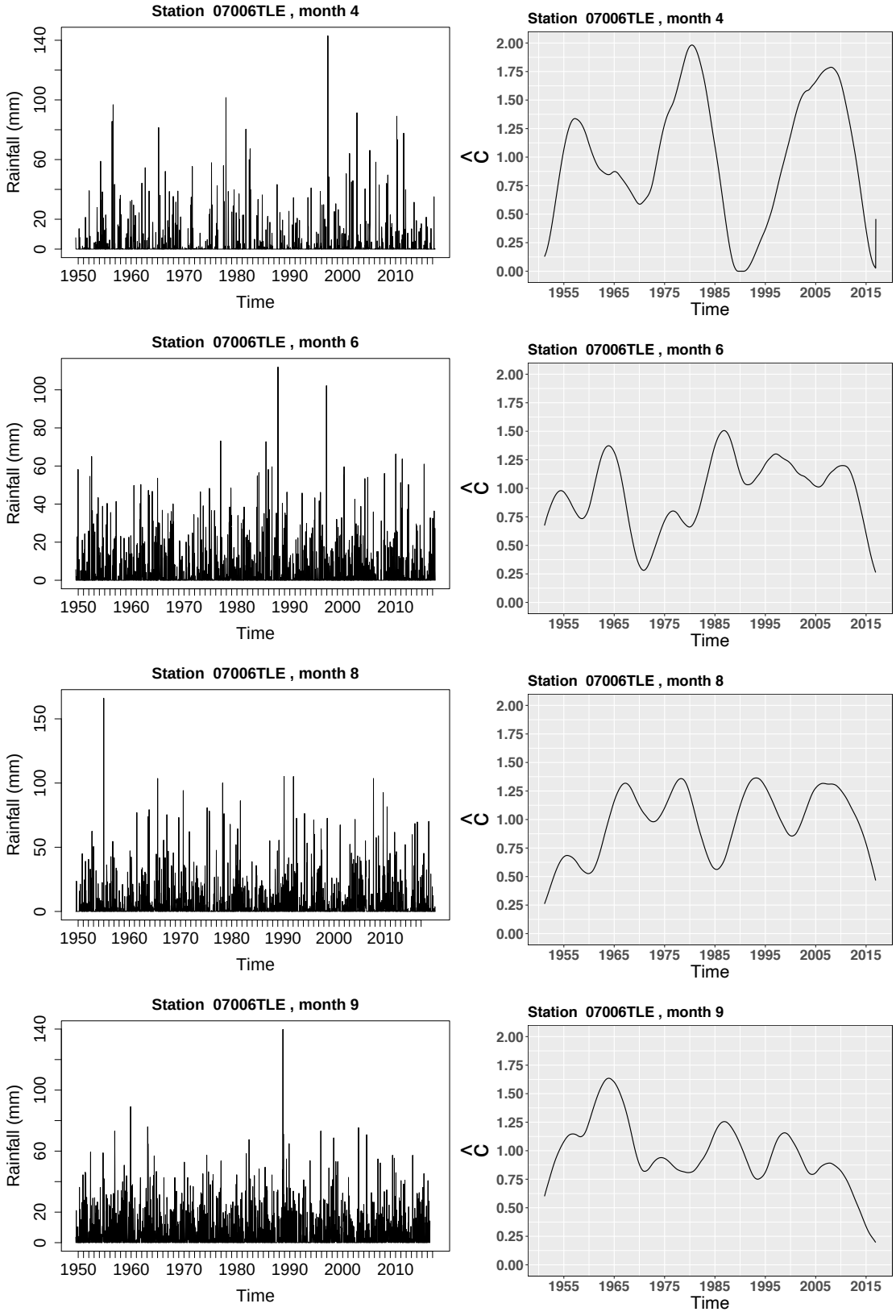


Figure 6.8: Time series of daily rainfall for the station TLE for (top to bottom) April, June, August and September.(Left) Time series with blank areas corresponds to missing values and (right) estimated skedasis function of exceedances above the 95% rainy day quantile.

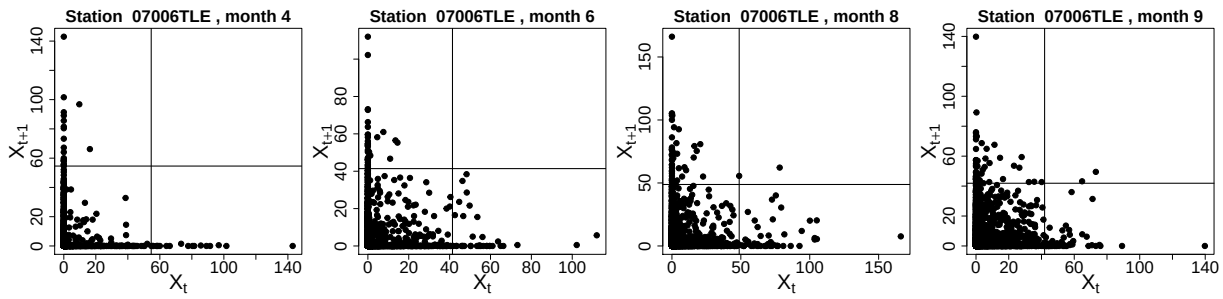


Figure 6.9: Scatter plots for X_t and X_{t+1} from the station TLE for April, June, August and September. Lines marks the rainy days 95^{th} percentile.

	TLE			
	April	June	August	September
Before	1	0.973	0.979	0.911
After	1	1	1	1

Table 6.3: Values of the extremal index before (top) and after (bottom) declustering the TLE time series, using the runs method requiring 1 day between observations above the 95% non-zero rainfall quantile.

6.4 Modelling spatial extremal dependence

To estimate the tail dependence at the different distances and months, the reduced bias estimator of the CTD developed in Chapter 5 will be used to investigate the positive or negative association between extreme events occurring simultaneously at two stations. Recalling that the CTD, η, is essentially putting a measure on the probability of X and Y being large at the same time when they are asymptotically independent, we can obtain a visual interpretation of the η value through bivariate scatter plots. By transforming the station time series data to standard uniform scale to remove differences in magnitude between the two variables, and plot the resulting bivariate sample points, we can gain an understanding of the two stations joint behaviour. To transform the data to the uniform scale, assign each sample point X_i, $i = 1, ...n$ its rank divided by the sample size, i.e. $U(X_i) := \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}_{X_j \leq X_i}$. Ties gets assigned the mean rank value, hence if there exists 5 days with the smallest possible amount of rainfall observed, each of these will be assigned $3/n$. In case there exist some positive association between the two variables, that is $0.5 < \eta < 1$ and equivalently $\mathbb{P}(X > u|Y > u) > \mathbb{P}(X > u)$, a higher density of points will be found in the top right corner compared to the exact independence case $\eta = 0.5$. In the near exact independence case, a uniform pattern should emerge for the largest values since there would be no preference for high or low values at station 2 given a

high value at station 1.

Figures 6.10 - 6.11 demonstrates how the bivariate scatter plots over rainy days are related to η and the impact from the sample size. The construction and interpretation of the two line graphs in the bottom of each plot is detailed in Section 6.4.1. For now, only the value of η in the right bottom plot, marked by arrows, are of interest.

Figure 6.10 shows examples of how the scatter pattern is related to the value of η when a sufficiently large number or samples exist.

If a smaller number of sample points exist, the estimation becomes significantly poorer and it is very difficult to graphically evaluate the pattern, as demonstrated in Figure 6.11, where the station-pair with 25km apart appears much more associated than the 5km apart pair. An important thing to remember is that the CTD only considers the tail of the distribution and provides no information about the remaining dependence structure. This can however be difficult to separate out from a visual inspection, since we are naturally drawn to the overall or dominating features of a graph. The two scatters in Figure 6.12 have the same tail dependence structure, but the 5km can easily appear more correlated since the main part of the graph has a distinct linear structure.

These three examples demonstrates that the uniformly transformed bivariate scatter plots can serve as an initial exploration tool of positive association between two variables, but a quantitative tool is needed to not draw conclusions about the tail behaviour from the main part of the distribution. They also demonstrate the sensitivity of the estimator to very small sample sizes, highlighting that the estimates needs to be evaluated in combination with their sample size.

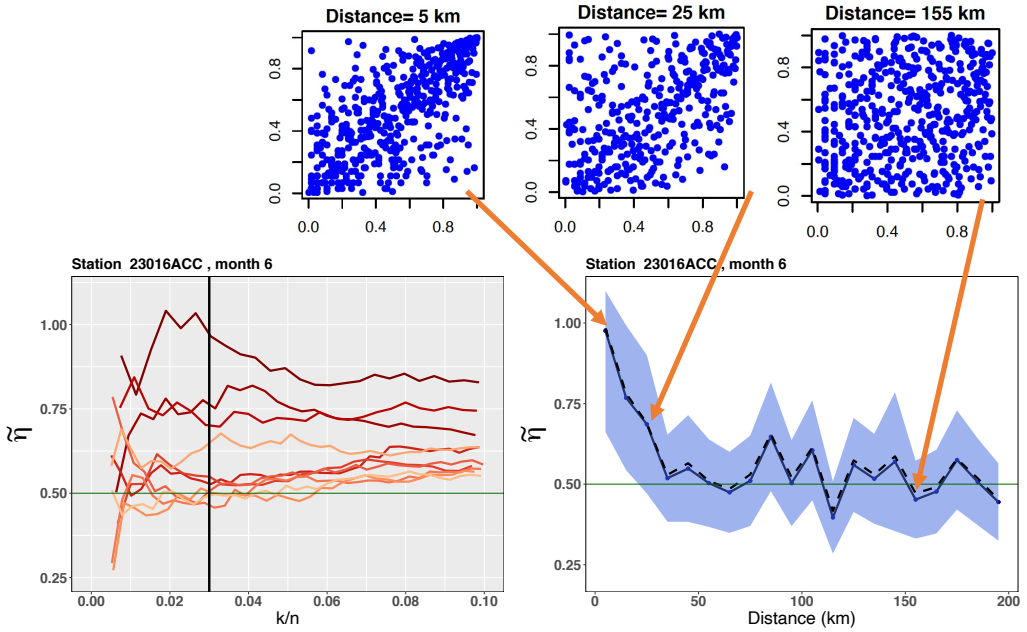


Figure 6.10: Combination plot for June at the station ACC, linking the bivariate uniform scatter pattern with coefficient of tail dependence values ranging between $\tilde{\eta} \sim 1$ and $\tilde{\eta} = 0.5$. The scatter plots are over rainy days only with sample sizes: 5km=498, 25km=356, 155km=497. Details of the two bottom graphs are given in Section 6.4.1.

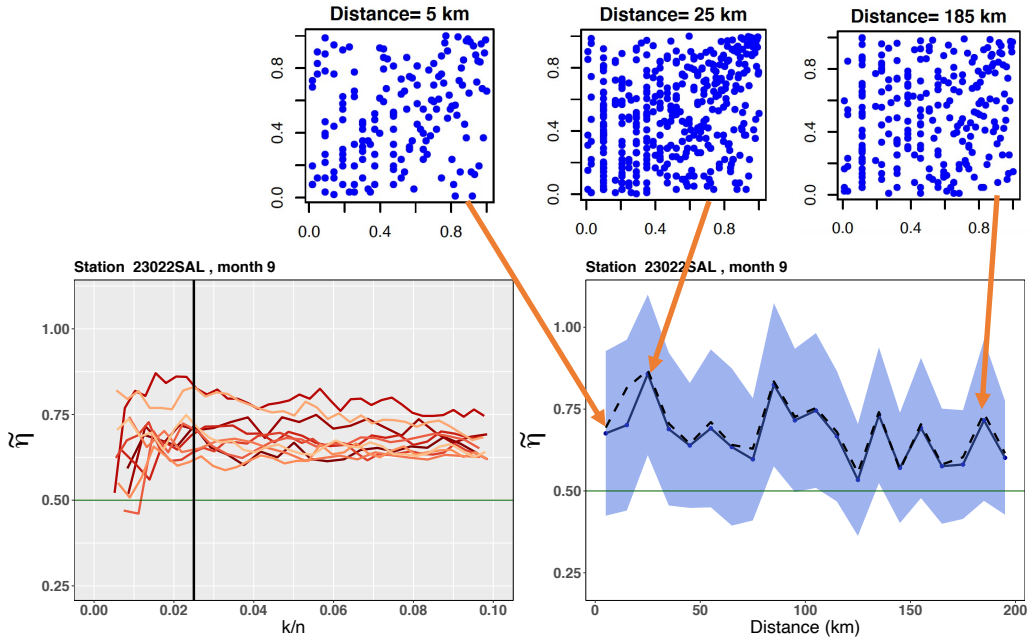


Figure 6.11: Combination plot for September at the station SAL, with examples of poor estimates of $\tilde{\eta}$ due to small bivariate sample sizes (left and right scatter) in comparison to a larger sample size (middle scatter). The scatter plots are over rainy days only with sample sizes: 5km=161, 25km=415, 185km=254. Details of the two bottom graphs are given in Section 6.4.1.

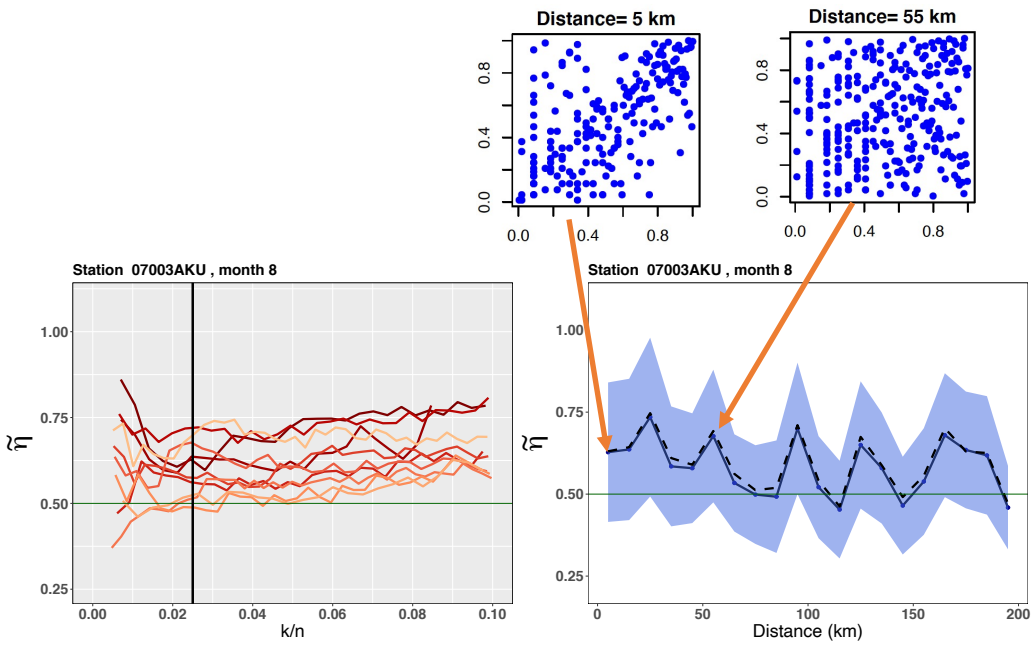


Figure 6.12: Combination plot for August at station AKU, with two examples of scatter plots associated with the same value of $\tilde{\eta}$. The scatter plots are over rainy days only with sample sizes: 5km=211, 55km=297. Details of the two bottom graphs are given in Section 6.4.1.

6.4.1 Estimation of tail dependence

Supported by the three examples shown in the previous section and the need for quantifying the positive association, the tail dependence will be estimated by applying the reduced bias estimator $\tilde{\eta}_a$ given by Equation (5.17) presented in Chapter 5, with the parametrisation $a = 1-1/q$ and shifted by $1/2$ unit Fréchet marginals. That is, given that k is the number of upper order statistics, $k^* \leq \sqrt{k}$, n the sample size, a the tuning parameter and $\hat{\tau}, \hat{\beta}$ second-order parameters, the estimator is given by

$$\tilde{\eta}_a(k) := \hat{\eta}_a^{(S)}(k) \left\{ 1 - \left(\hat{\beta} \left(\frac{n}{k} \right)^{-\hat{\tau}} + \frac{1}{1 + 2V_{n,n-k^*}} \right) \frac{1 - a\, \hat{\eta}_a^{(S)}(k)}{1 - a\, \hat{\eta}_a^{(S)}(k) + \hat{\tau}} \right\},$$

with

$$\hat{\eta}_a^{(S)}(k) := \frac{\left\{ \left[\frac{1}{k} \sum_{i=0}^{k-1} \left(\frac{1/2+V_{n,n-i}}{1/2+V_{n,n-k}} \right)^a \right]^{1/a} \right\}^{-a} - 1}{-a}$$

and

$$V_i^{(n)} := \left\{ \left(-\log \frac{R(X_i)}{n+1} \right) \vee \left(-\log \frac{R(Y_i)}{n+1} \right) \right\}^{-1}$$

The reason for working with an asymptotic independent model is supported by the results

in Chapter 3, where even for the very heavy rainfall the probability of co-occurrence was never above 0.8 and dropped down to the climatological value at a distance of ~ 150km, indicating near exact independence. Although the process behind extremal observations is often different to the regular regime, it seems suitable to work with a model that allows for exact independence rather than one that assumes the two variables to be dependent, to not artificially inflate the dependence range.

In the simulation study in Chapter 5, it was concluded that there was a significantly larger bias for q close to 0 and often a lower bias for values larger than 1, where 1 is the value at which we recover the Hill estimator. With the behaviour being more stable for $q > 1$ and the bias for small k/n generally smaller for values close to 1, the value $q = 1.1$ is chosen here.

Similarly to the co-occurrence estimation in Section 3.2.4, the aim here is to estimate the dependence in the rainfall amounts but not the occurrence since as mentioned previously these are governed by different processes. Therefore only days where both stations record non-zero values are used in the subsequent analysis. However, we still want to preserve the information about what proportion of the measured days has been used for the estimation, hence n will still be the number of measured days instead of the number of rainy days. This additionally makes it easier to compare the estimates between the different months, since we have a similar number of missing values over a given year, but a much higher proportion of dry days outside the main monsoon season (Figure 3.11).

Since there is a large difference in the number of sample points (see Table 6.1) for the station-pairs in $S_{c,p}$, especially for the north station TLE, all results will be presented as a function of $k/n =: k_n$ rather than k. To determine the optimal k_n, $\tilde{\eta}_a$ is estimated for the 10 closest station-pairs (0-100km apart) for $k = 10, 15, ..., \lfloor 0.1 * n \rfloor$ (Figures 6.13-6.16, left column), hence enforcing that at least 10 sample points are included in the estimation. By plotting the estimates as a function of k_n, we select the proportion at which the estimates lines are approximately linear to the left of this value, in Figures 6.13-6.16 indicated by the black vertical line. This is the sample principle as the one for univariate statistics for choosing the POT threshold (Figure 2.8). If they are not stable at one value, the point at which they start to stably fluctuate around a value is chosen.

After selecting the most suitable proportion, and thereby fixing k_n, the estimates are thereafter plotted as a function of the distance (Figure 6.13-6.16, right column) now extending to the full range 5-200km. This essentially corresponds to taking a vertical slice of the first plot at the fixed k_n value, which is marked by the vertical black line. The 95% CI is estimated

using the variance expression in Corollary 2 and added to the fixed k_n distance plot. As a reference these plots also include the corresponding Hill estimates (dashed line) but without the corresponding CI to ease the readability of the graph.

6.4.2 Southern region

In general, there is a lot of variation between the three stations despite being located in the same region and sharing multiple pair-stations as can be seen in Figure 6.1. This is partly due to the variability in sample size between the station pairs (see Table 6.1), but also due to the common problem in Extreme value statistics of highly variable estimates for small changes in k.

In the left hand plots in Figures 6.13 - 6.15, it can be seen how the estimate varies as a function of k/n, hence of the same format as the plots presented in the simulation study in Chapter 5. In all the left hand plots we can clearly see the convergence of all distance-dependence lines to a value of around 0.6-0.7 as we move away from the extremes and into the normal part of the sample. The black line indicating the common optimal proportion k_n at which the lines are approximately constant to the left, and before they have converged to the non-extreme estimate, is around 3% for nearly all months. This corresponds to between 35-120 sample points, depending mainly on the month but also on the station-pair. The right hand plots instead show the dependence as a function of distance, for each distance using the k corresponding to the k_n proportion marked with the vertical black line. Due to the small sample size, the 95% CI in the right hand estimates plots are relatively wide, something that we mention but cannot do anything to address at this stage. A second thing to notice is the near exact agreement between our reduced bias estimator and the Hill estimator, with the dashed Hill line only being slightly larger than the reduced bias estimator. In the simulation study in Section 5.5.2, the Hill estimator was always significantly larger than the reduced bias, indicating that the less pronounced pattern seen here could be due to difference in performance when applying the estimators to real world data compared to perfect model data. By recalling that the variance is given by η^2 for the Hill estimator and $\eta^2(1-a\eta)^2/(1-2a\eta)$ with $a = 1 - 1/1.1$ for the specific version of the reduced bias estimator, the variance of the two differ by less than 0.01 for all values of η.

For ACC (Figure 6.13), there appears to be some dependence at short but not moderately long distances in April and September, with an estimate of around 0.75 for short distances that drops down to the 0.5 exact independence line at around 50km. In June at the peak of the monsoon and when the ITCZ draws in moist air from the Atlantic ocean, there is near exact asymptotic dependence for the station-pair with 5km apart with an estimate close to 1. This is also one of the few estimates where the CI is far away from the 0.5 line, indicating that there

definitely is some positive association between the two stations. August exhibits a relatively constant pattern for the first 120km, fluctuating around the 0.7 line, which the estimator seems to converge to as we move away from the extreme values. August being the little dry season means that we have much fewer rainy days and therefore significantly less sample points, which we could see in Figure 6.12 has an impact on the estimate. If we further compare this with the co-occurrence probability in Figure A.1 and use the 95% threshold marked in Figure 6.7, we can see that August would belong to the 'moderate' intensity class ($10 - 30$mm) which in general exhibits a much longer decorrelation range compared to the higher amounts.

AKU (Figure 6.14) displays generally lower values compared to ACC and SAL, with April being the exception. Even though the estimate never reaches the independence line in April, there is a significant decrease between 5km and 60km, after which the estimate stabilises and the CI envelope encloses the exact independence line. August appears to have a decreasing trend for he first 55km, after which it fluctuates around the 0.5 line, however the estimate is close to 0.5 for all distances indicating a very weak positive association. AKU, compared to the other two south region stations is located inland with many of its pair-stations located even further inland. These stations might therefore experience less rainfall coming from moist maritime air being advected and released over land, and instead more small scale convective rainfall, which could explain some of the weaker patterns seen here.

The other coast station SAL (Figure 6.15) is more similar to ACC in patterns, with the clear difference that the estimate does not cross the exact independence line for both August and September. A comparably strong positive association for short distances in June can be seen with an estimate of around 0.85 and CI away from the exact independence line. The tendency of values around 0.6-0.7 for longer distances in August compared to the other months can be seen here as well, with a slowly decaying pattern all the way until 180km. September is very inconclusive, with a patten more similar to August at ACC, exhibiting a near constant value of 0.6-0.65 for all distances. This could be due to the same reason as for August, the largest values are not very large in the absolute sense. Looking back at Figure 6.7, we can see that the threshold line is around 20mm for SAL, whereas ACC and AKU has threshold values of 30mm and 45mm respectively, indicating generally more moderate amounts at SAL which we already know has a slower decaying correlation.

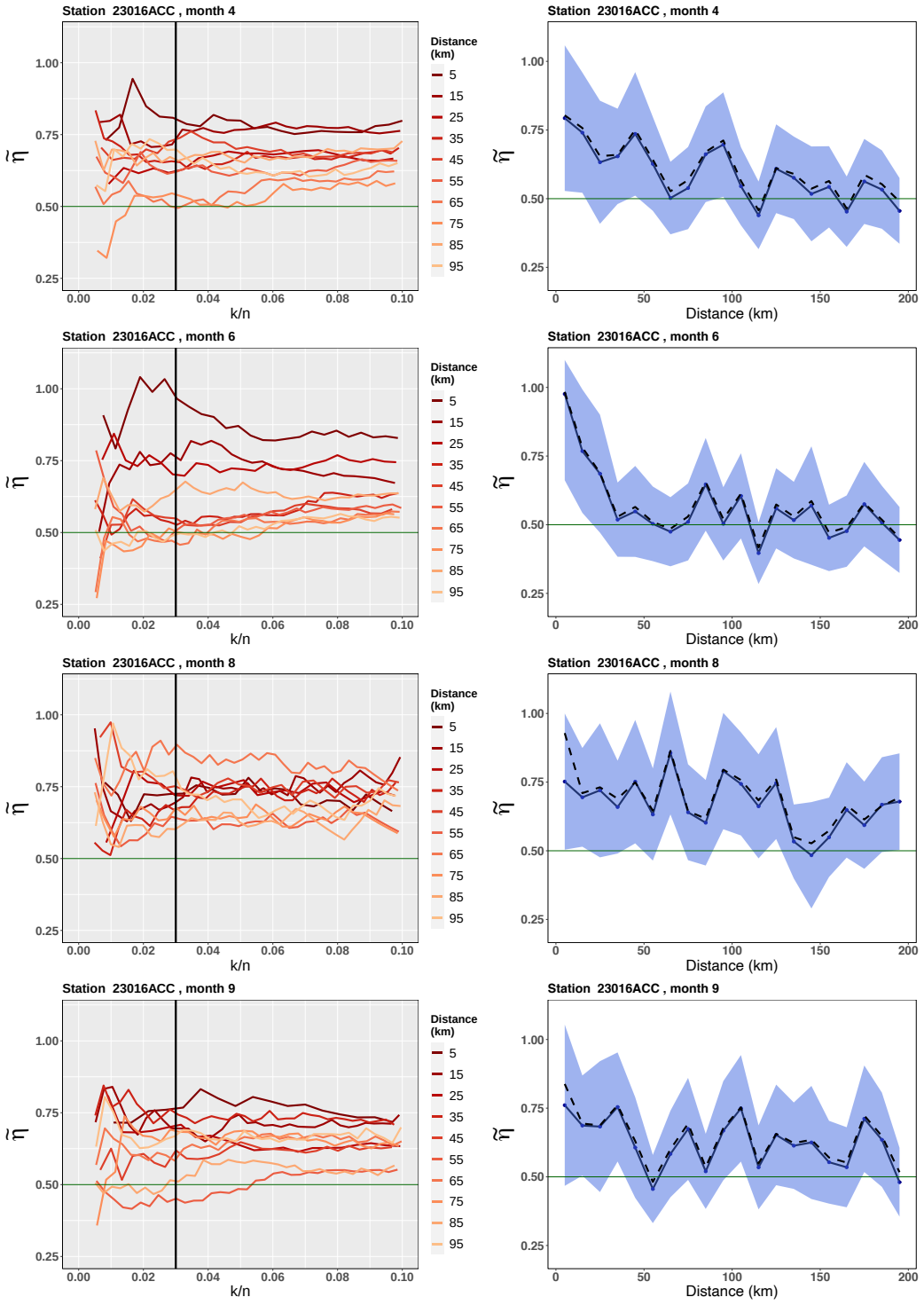


Figure 6.13: Estimation of the coefficient of tail dependence through estimator $\tilde{\eta}_a$ as a function of k/n (left) and distance (right) for the station ACC. (left) The coloured lines in the left hand plots corresponds to distances up to 100km, with darker colour representing shorter distances and the black vertical line marks the k/n value used for the right hand plots. (right) Solid line is $\tilde{\eta}_a$ and dashed Hill, the envelope on the right hand plots encloses the 95% CI.

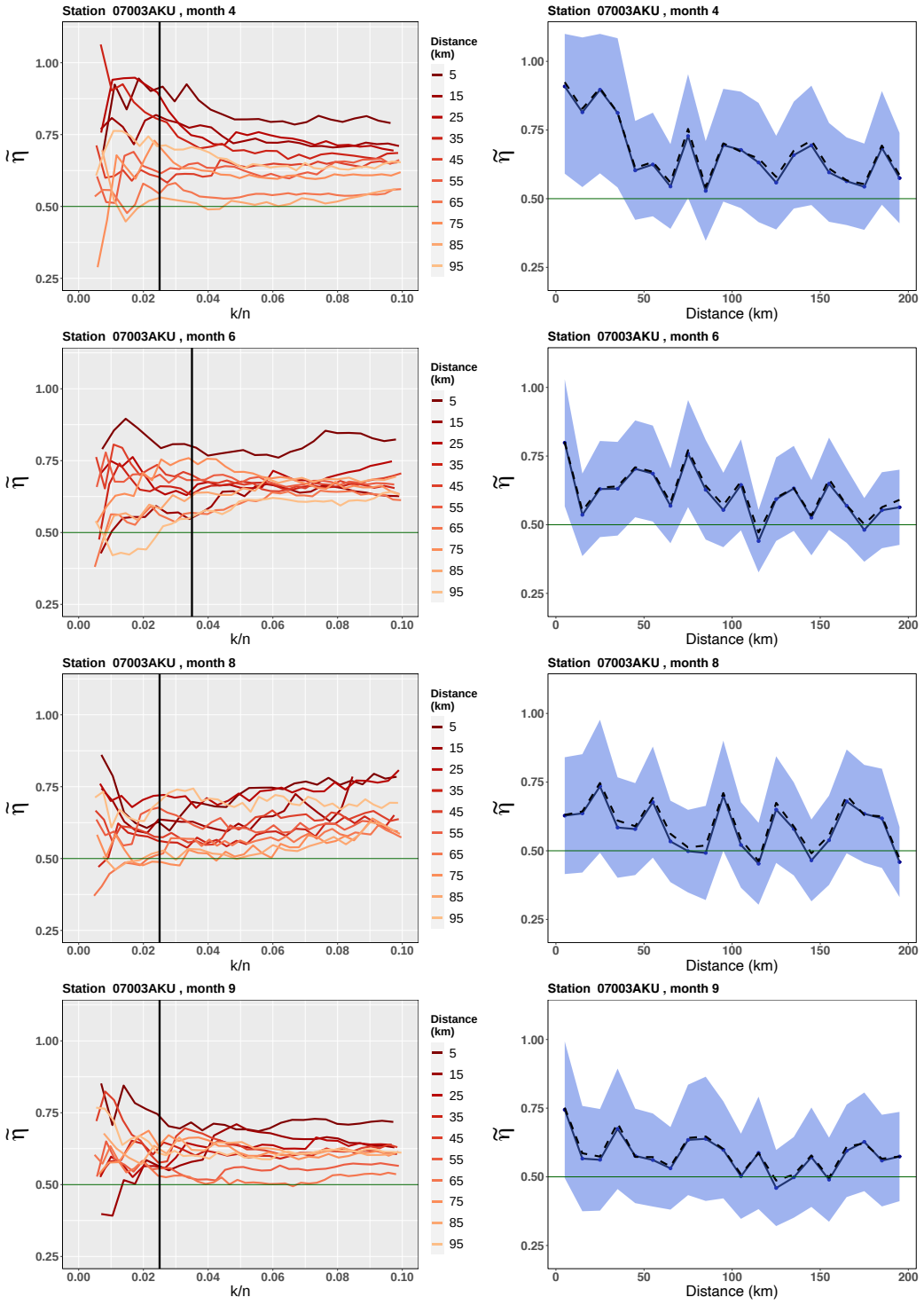


Figure 6.14: Estimation of the coefficient of tail dependenc through estimator $\tilde{\eta}_a$ as a function of k/n (left) and distance (right) for the station AKU. (left) The coloured lines in the left hand plots corresponds to distances up to 100km, with darker colour representing shorter distances and the black vertical line marks the k/n value used for the right hand plots. (right) Solid line is $\tilde{\eta}_a$ and dashed Hill, the envelope on the right hand plots encloses the 95% CI.

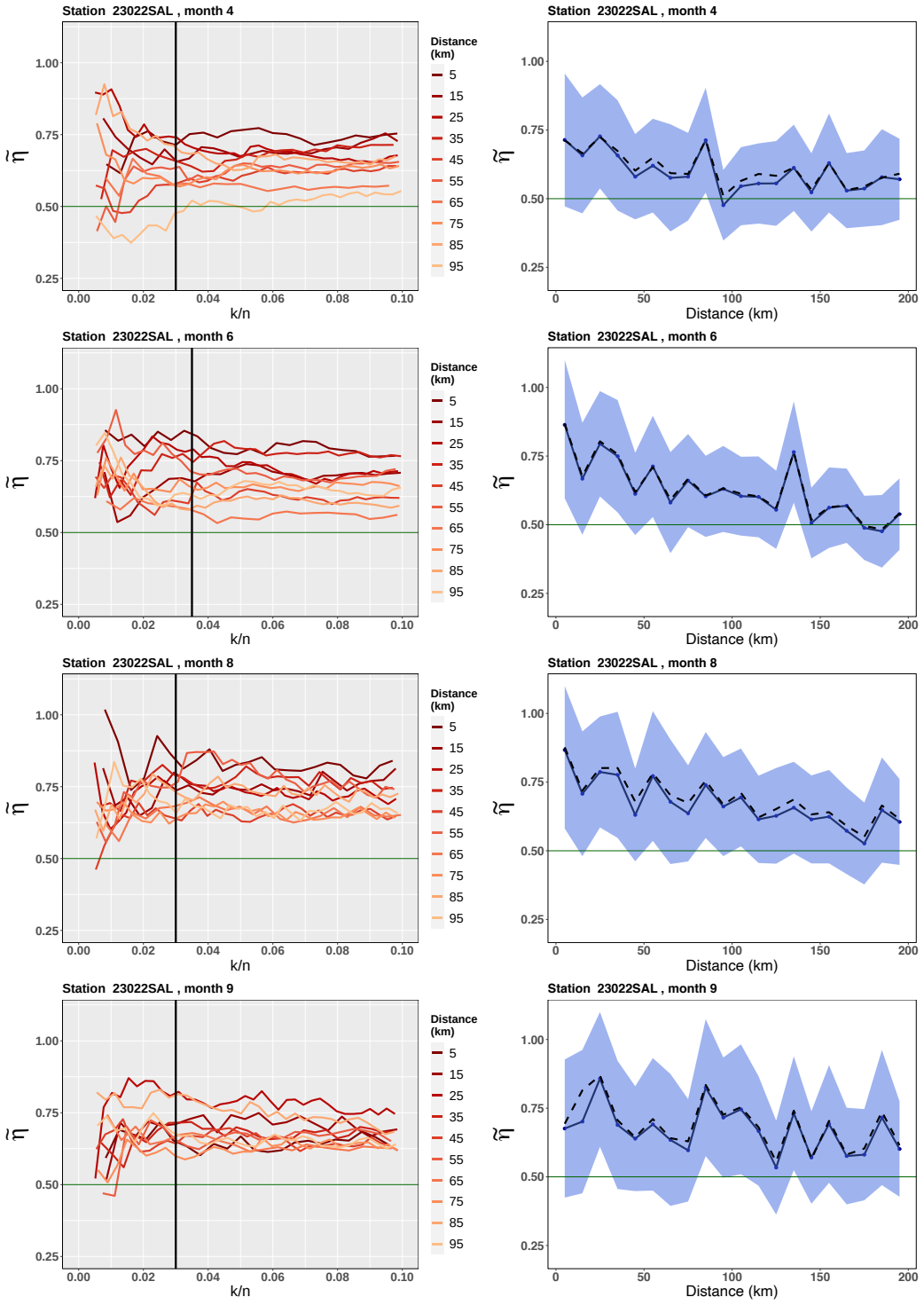


Figure 6.15: Estimation of the coefficient of tail dependence through estimator $\tilde{\eta}_a$ as a function of k/n (left) and distance (right) for the station SAL. (left) The coloured lines in the left hand plots corresponds to distances up to 100km, with darker colour representing shorter distances and the black vertical line marks the k/n value used for the right hand plots. (right) Solid line is $\tilde{\eta}_a$ and dashed Hill, the envelope on the right hand plots encloses the 95% CI.

6.4.3 North region

Since several of the crucial distances are missing, it is even more difficult to draw conclusions about the dependence structure in this region compared to the south, but some similarities and differences can be observed. In Figure 6.16 it can be seen that there is a very strong positive association at the shortest distances in both April and June, but this quickly drops of to near exact independence at 45km. Despite August being one of the rainier months in this region, there is a much weaker association even at short distances, the reason for which is still unclear. September, which is the peak of the rainy season has a relatively high, constant value for the 3 first distance bands and does not seem to reach the independence line until 115km away, which is significantly further away than for the other months. However, due to the large number of missing distances we cannot conclude if the underlying pattern follows linearly between the estimates, or rapidly drops down to a value around 0.5, and the higher value at 90km is a temporarily higher value as seen in June. This is therefore only seen as a first indication of seasonal differences and not exact estimates of decorrelation distances.

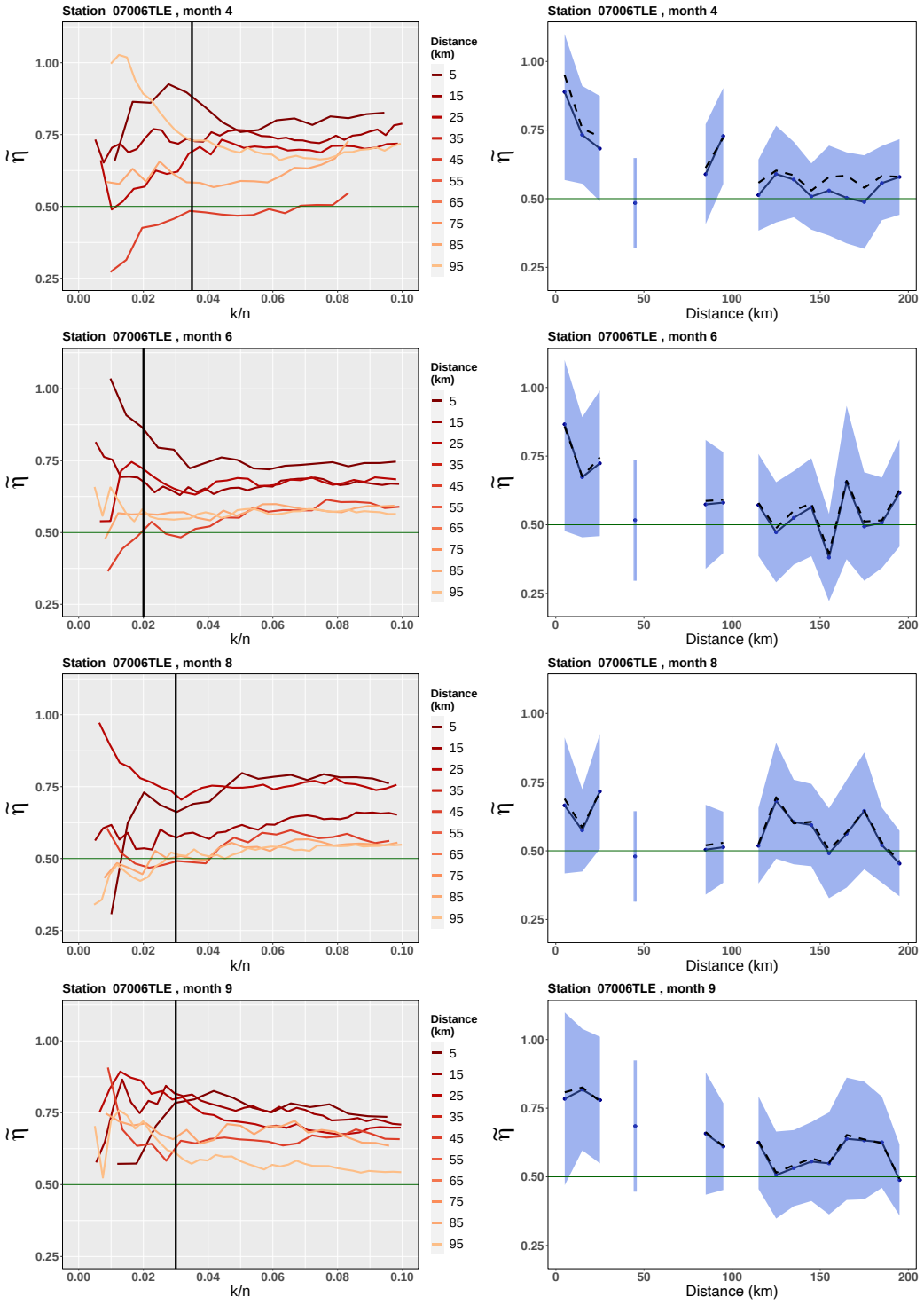


Figure 6.16: Estimation of the coefficient of tail dependence through estimator $\tilde{\eta}_a$ as a function of k/n (left) and distance (right) for the station TLE. (left) The coloured lines in the left hand plots corresponds to distances up to 100km, with darker colour representing shorter distances and the black vertical line marks the k/n value used for the right hand plots. (right) Solid line is $\tilde{\eta}_a$ and dashed Hill, the envelope on the right hand plots encloses the 95% CI.

6.5 Discussion and further work

In this chapter, we have attempted to obtain one of the first estimates of the dependence structure in extreme rainfall over west Africa through the application of multivariate EVT. This can be seen as an extension of the analysis performed in Chapter 3, drawing on asymptotic theory and only considering the largest observations instead of the full sample. Although convective systems over west Africa can spread several hundred kilometres, so called MCSs, the presence of local storms and the natural assumption that the dependence between any two locations should diminish as the distance increases, an extremal dependence model which models the asymptotic independence rather than asymptotic dependence was deemed appropriate. This has also been recommended in some recent papers when considering extreme rainfall over South Africa (Sang and Gelfand (2009)). The reduced bias estimator of the CTD developed in Chapter 5 was therefore applied to our daily rainfall data set.

One general conclusion from this chapter is the need for mixed models which can seamlessly transition between modelling extremal asymptotic dependence and asymptotic independence for increasing distances. Comparing the results here with the distances of around 100-200km obtained in Blanchet et al. (2018), one can conclude that max-stable dependence models probably inflate the dependence range. The asymptotically independent model on the other hand does not provide information about the potential asymptotic dependence existing between the two variables, leading to both type of models currently available less than ideal for this type of analysis.

A problem here, which is shared with the limited number of previous papers on extreme rainfall over Africa, is the relatively short time series which causes issues since very few rainy sample points are available for each month, especially outside the most rainy months. With the strong seasonality in rainfall due to the WAM, pooling data from several months could potentially add more variability by violating the assumption of identically distributed data. It further became clear that as with most extreme value analysis, one needs to be careful what k/n, or essentially k, one chooses since the estimator seems to converge to a general value of 0.6-0.7 when the central part of the data is included. These shortcomings added together results in very wide CI, leading to any conclusions drawn here rather weak and mainly a first indicator of how this method performs on tropical rainfall data.

One of the main conclusions drawn is that the distance at which any positive association instead of exact independence is a reasonable assumption between two stations, is much shorter than the distances seen in Chapter 3. This seems to mainly be dependent on the presence of

extreme values in the absolute sense, with months such as August in the south exhibiting a much longer dependence range but also has smaller maximum amounts and less spread between the largest amounts. The north region appears to have the opposite pattern with the longest dependence distance for the peak month September. It is however a much less robust result since we are missing several distance bands in the first 100km, which means that any conclusions at those distances assumes a linear interpolation between the available distances. We can further conclude that there is a difference between the coast stations compared to the inland in the south region, which could be highlighting the difference in nature of convective and advective rainfall.

To better understand how the estimator performs on measured data, compared to perfect model data, an additional simulation study similar to the one presented in Chapter 5 on tainted model data could be performed. This could potentially shine some light on why the reduced bias estimator $\tilde{\eta}_a$ and the Hill estimator returns the same estimate here but not in the previous simulation study for any copula model. By mixing in a certain proportion of a different distribution or introducing missing values in a similar proportion as observed in our time series, we could gain a better understanding of how large influence these deviations have on the estimate. This simulation study should also be done with sample sizes ranging from the smallest ones available here, $\sim$130, up to the ones used in Chapter 5 (n=500) to better understand at what stage this starts to have a significant impact.

Chapter 7

Conclusions

In this chapter the main conclusions drawn from Chapter 3-6 are presented and how these relate to the aim questions defined in Chapter 1. For each of the aim questions, the scientific advancements and the limitations for answering them are discussed. Several directions for further work are outlined and how these can improve on some of the limitations faced here. The chapter ends with a final conclusion on how the results presented addresses the overall thesis aim of improving the estimation of the key components for relating rain gauge observations with satellite estimates.

7.1 Summary of main outcomes

7.1.1 New methods for estimating correlation distances

Building on the methodology used in Ricciardulli and Sardeshmukh (2002), a detailed algorithm for calculating the correlation range for daily rainfall has been developed in Chapter 3. A significant advantage of using this method compared to previously proposed methods, is the random sampling step which determines the background probability for the events of interest. Instead of fixing it to a user determined value, it naturally allows for a data informed 'null hypothesis' value that changes with the season. By only incorporating the events that are of interest, such as a specific intensity band, it provides a method for investigating the differences between different rainfall events. A second advantage is the non-parametric form of this estimation method, therefore not requiring the user to make assumptions about the nature of the correlation relation, such as for the Pearson's correlation.

In Chapter 5, a semi-parametric estimator for estimating the bivariate extremal dependence in the case of asymptotic independence, as advised by Sang and Gelfand (2009) for tropical rainfall, has been developed and further a reduced bias version of the same estimator. The es-

timator is based on the mean-of-power-p estimator proposed by Gomes and Caeiro (2014) but with the bivariate extension following the work of Goegebeur and Guillou (2012) and Draisma et al. (2004). Due to the format of the submodel proposed by Ledford and Tawn (1996), all the classical estimators for the extreme value index, such as the Hill and moment estimators, can be utilised however all of these suffer from significant bias as the included sample size increases. The reduced bias estimator proposed here is one of the few available estimators incorporating the error that stems from the marginal distribution transformation. It is shown to significantly reduce the bias compared to the Hill estimator and due to its analytical instead of maximum likelihood form, confidence intervals can cheaply be obtained.

A common limitation for both of the models developed here is the need for a relatively large sample size, something that still is rare for most parts of Africa. The non-parametric estimator relies on the Law of Large Numbers for the averaging of the individual co-occurrence probabilities to converge to the true value. As with all extreme value methods, only a very small proportion of the largest values are included in the estimation, resulting in very few sample points in the case of a small sample size. This ultimately leads to very large CI, often resulting in uninformative estimates. The information obtained by the coefficient of tail dependence estimator can further not be used to estimate return periods or similar estimates of interest, which is a limitation in case of risk analysis.

7.1.2 Intensity dependent correlation range

Despite the wide range of rainfall processes occurring over west Africa, the correlation range is almost always assumed to be intensity invariant. In this work this assumption has proved to be unrealistic, with significant difference between the area of influence from varying intensity classes. Chapter 3 demonstrated that low intensity rainfall has a much smaller correlation range compared to high intensity rainfall. It further concludes that there is a strong seasonal cycle on the short-distance correlation for low intensity rainfall, but not for the heavier intensities where this stay constant. Since rainfall events of low intensity are much more frequently occurring than heavy (Maranan et al. (2018)), estimating one general correlation range will result in an underestimation of locations receiving rainfall in the case of heavier events. This also reduces the amount of information that could be obtained in data sparse regions, by not maximising the number of observations that can be utilised.

A change in the pattern of longer correlations for higher intensity is found in Chapter 6, where the dependence range of extreme values is found to be of similar distance as for low intensity events. This is in line with previous assumptions that physical phenomena tend to

become more localised for the highest quantile events (Huser and Wadsworth (2020)). The robustness of these findings are however very weak due to the limited data available. Even though the data set used here is both longer and more dense than most other African rain gauge data sets, the strong seasonality and the limited number of stations with few missing values results in very few bivariate points left for the individual estimations. In the north region, the gauge network is not dense enough to find suitable station-pairs for several of the crucial distances, leading to rather weak conclusions drawn since a linear interpolation between the distances available is not realistic as discussed previously.

7.1.3 Distributional properties of conditional rainfall

Evaluation of the conditional distribution of gauge observations for a given satellite rainfall estimate has been performed for several parametrisations of the lognormal distribution in Chapter 4. Using the same formulation for the mean and variance as proposed in Teo and Grimes (2007) and Greatrex et al. (2014), this distribution has been shown to provide a good fit based on a range of qualitative tools. It has previously been shown that the conditional rainfall is heteroscedastic as a function of the cold cloud duration, this is however shown to nearly completely be removed when taken as a function of satellite rainfall estimates instead. These findings provide an improved understanding of the nature of the relation and variability between ground point observations and satellite area estimates.

These findings can provide more realistic uncertainty estimates through the sequential simulation framework initially proposed in Teo and Grimes (2007), by better capturing the tail behaviour. It further improves the possibility to incorporate gauge observations with satellite estimates by providing a correct measure for how anomalous an observation is in relation to the estimate (see Section 4.2.2). A large limitation to this work is the very small set of distribution families investigated here, leaving the possibility that more suitable distributions exists. Further work to address this is detailed in 4.6.1.

7.2 Further work

Some directions for further work to extend the analysis performed in this thesis are already mentioned in the chapters, especially in Chapter 4 where a detailed description of how to include quantitative tools and suggestions on more sophisticated skewed distributions are detailed. Other directions for further work identified by this thesis are the following:

Other regions to establish generality of the results

All the results presented here are only from one case study area, and the regular intensity analysis in Chapter 3 only from the southern half. Even though the climate over west Africa is similar due to the large scale drivers controlling it, there are large variabilities across the region as demonstrated by Funk et al. (2015a), and further highlighted here by the differences even within Ghana. To establish how general these findings are, it would be useful to apply the methods developed here to other countries in west Africa as a first step, and later on to other regions.

Include non-stationarity

Throughout the thesis, stationarity has been assumed since no obvious trend has been identified that could easily be adjusted for, and reducing the sample size further by splitting the data into different time periods would have led to too small samples. West Africa has however experienced large scale, decadal variabilities in rainfall amounts (Nicholson et al. (2018)), and trends in the frequency and contributions from MCSs (Taylor et al. (2017)). Even though most of these strong trends has been identified for the Sahel region (see 3.1 for the Sahelian drought), which is located north of our study region, variability has been observed for the Guinea coast as well. Incorporating well-known large scale drivers, such as the sea surface temperature and the Saharan Heat Low (Parker and Diop-Kane (2017)), to remove some of this temporal variability could improve the identically distributed condition and thereby reduce the sample variability. This could especially be of interest for the extremes analysis, since variability was observed for nearly all months in Chapter 6.

Simulation study on coefficient of tail dependence estimator with increasingly tainted data

In the simulation study in Chapter 5, a large difference between the here proposed reduced bias estimator with suitable choices for the tuning parameter and the classic Hill estimator was found. This result was however not seen in the analysis on observed data in Chapter 6, where virtually no difference between the two could be detected, regardless of the sample size available. In order to better understand the reason for this behaviour, a simulation study that better mimics real world data could be carried out. The two main issues with real compared to simulated data is the deviation from all sample points stemming from the exact same distribution and the presence of missing values.

The impact from the first factor could be analysed by increasingly 'tainting' the simulated

data set by mixing in a set proportion of samples draw from a different distribution, which essentially would be the same things as simulating data from a mixed distribution. By either mixing in sample points from a distribution with the same theoretical value for the coefficient of tail dependence, or one with a different tail dependence, the sensitivity on these deviations can be thoroughly investigated. This could for example be done by taking 80% of the sample points from a Frank copula ($\eta = 1/2$) and 20% from a Ali-Mikhail-Haq ($\eta = 1/3$). A similar idea can be applied for introducing missing values, by randomly sampling a set proportion of the simulated sample and remove these observations.

The main aim of this simulation study would be to examine how quickly the potential additional bias would be visible and if the difference between the reduced bias estimator and the Hill diminishes or persists.

7.3 Conclusions

The amount of high quality rainfall data over Africa does not seem to improve, but rather deteriorate in many countries. It is therefore vital to extract the maximum amount of information possible from the available observations. The overarching aim of this thesis was to introduce improved methods for estimating some of the key components needed to relate satellite observations with ground rain gauge measurements. Specifically the focus was on estimating intensity dependent correlation ranges which is needed when performing kriging, and the conditional gauge distribution. A flexible, inexpensive, non-parametric model based on comparing estimated and observed probabilities, and a semi-parametric bivariate coefficient of tail dependence model have been developed. Applying these to the same rain gauge data set has demonstrated an increase in the correlation range for increasing intensities up to a high level, but a significantly shorter dependence range when considering the most extreme values. In a kriging setting, this would result in the use of different values for the range parameter in the covariance function. Particularly in the merging setting of Chapter 4, gauges measuring large values would be associated with a larger number of orange grid boxes compared to low intensities and extreme values. A lognormal distribution has been shown to provide a good fit for daily gauge measurements related to daily satellite estimates, providing a better understanding of the relation between the two.

Appendix A

Supplementary material, Chapter 2

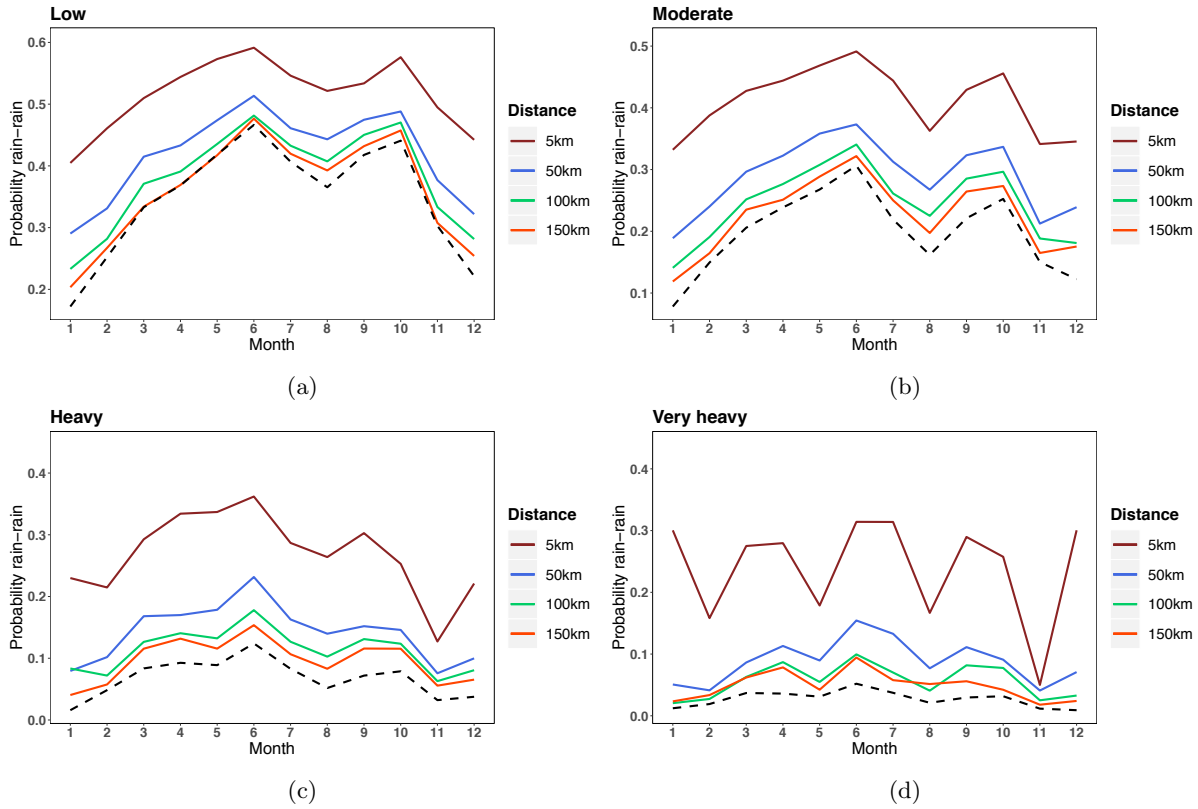


Figure A.1: Seasonal evolution of the conditional occurrence probability for stations south of 8°N. The solid lines are distances away from the origin and the dashed line is the random sampling baseline at 50km. The intensity bands are as described in Section 3.2. The rain-rain occurrence is 1 if the distant station is in the same or higher intensity class (see Algorithm 2 in Section 3.2). Note the different scales on the y-axis.

Data points for 5km line

Intensity	1	2	3	4	5	6	7	8	9	10	11	12
Low	557	1308	2212	2425	3087	4099	3452	3348	3710	3906	2436	1157
Moderate	323	866	1715	1973	2426	2857	1732	1053	2010	2828	1461	695
Heavy	58	236	570	606	747	1009	492	245	569	689	274	184
Very heavy	43	79	253	264	320	648	340	101	294	277	80	66

Table A.1: Number of data points used to estimate the 0-10km line for Figure 3.12.

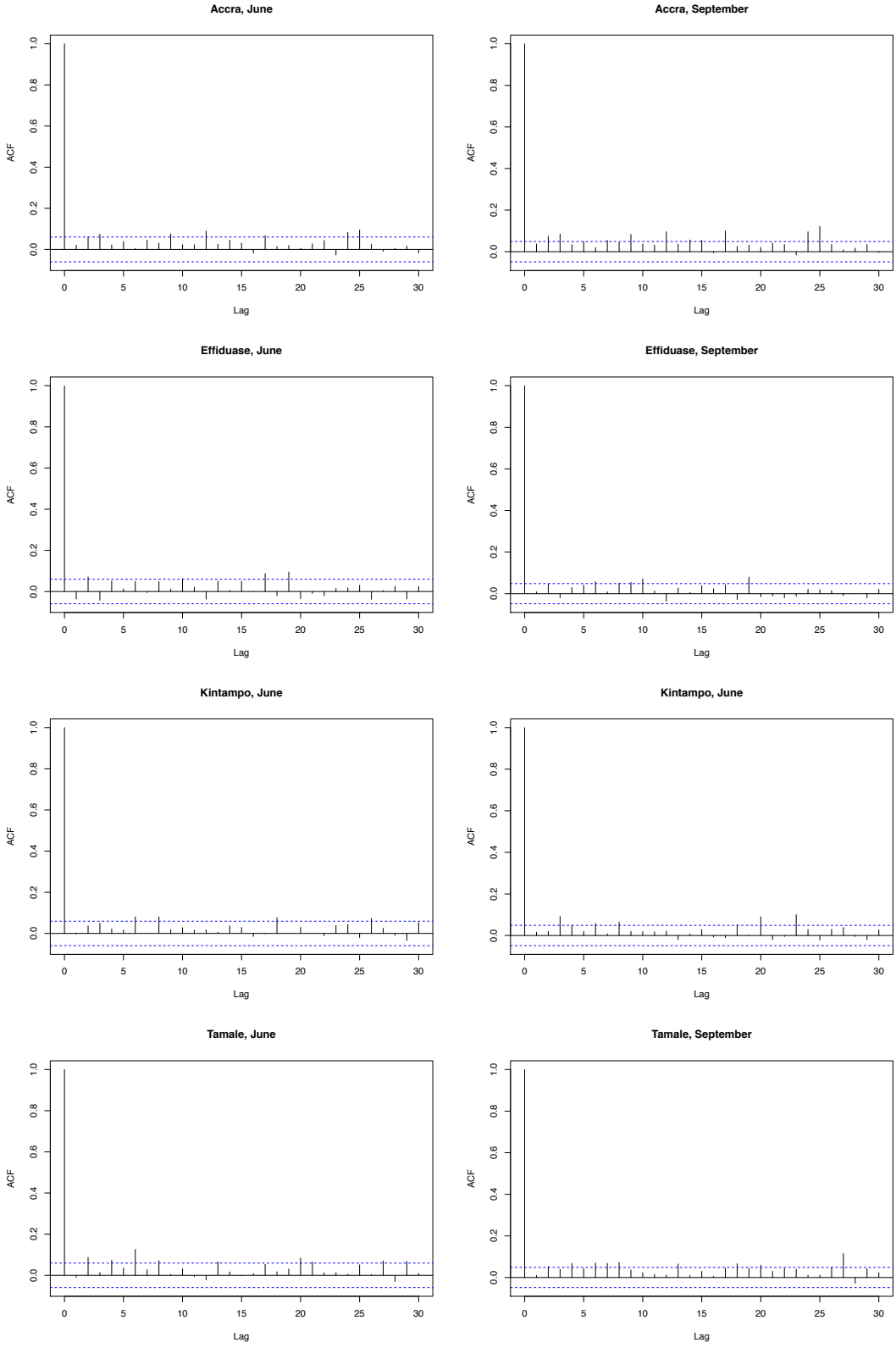


Figure A.2: Autocorrelation plot for each synoptic station used in Section 3.3.1.

Appendix B

Additional histograms and QQ-plots for Chapter 3

Histograms (Figure B.1 - B.4) and QQ-plots (Figure B.5 - B.8) for all 2mm RFE bin subsets between the RFE values 2-24mm. QQ-plots with fewer than 5 points are not included.

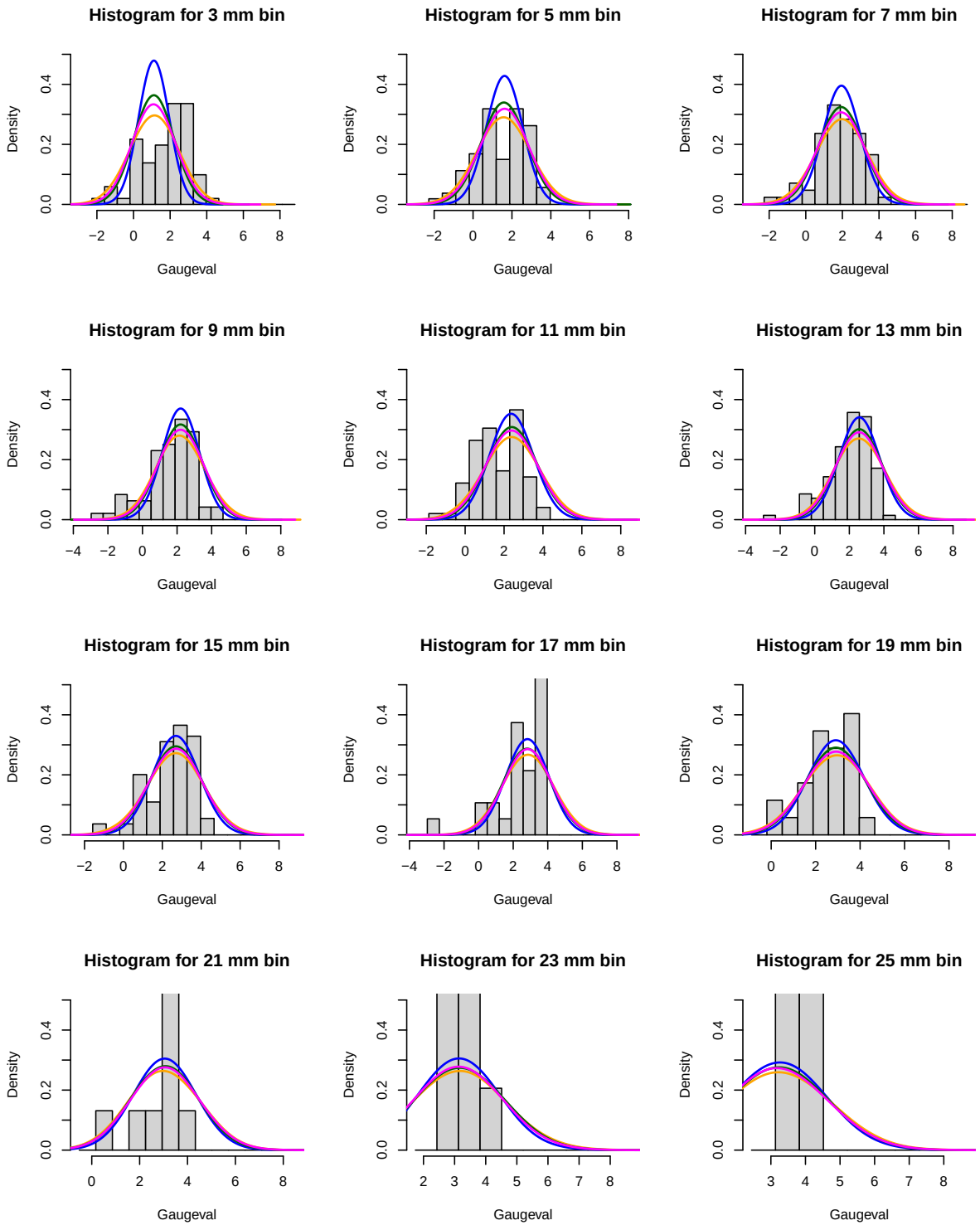


Figure B.1: Histograms for April on the logarithm of the gauge values. The mm value is the mid value for the 2mm RFE bin. Each bar is log(2)mm wide. The density curves corresponds to the normal distribution with $\mu = \log(\text{RFE})$ and standard deviation given by the coloured lines in Figure 4.6.

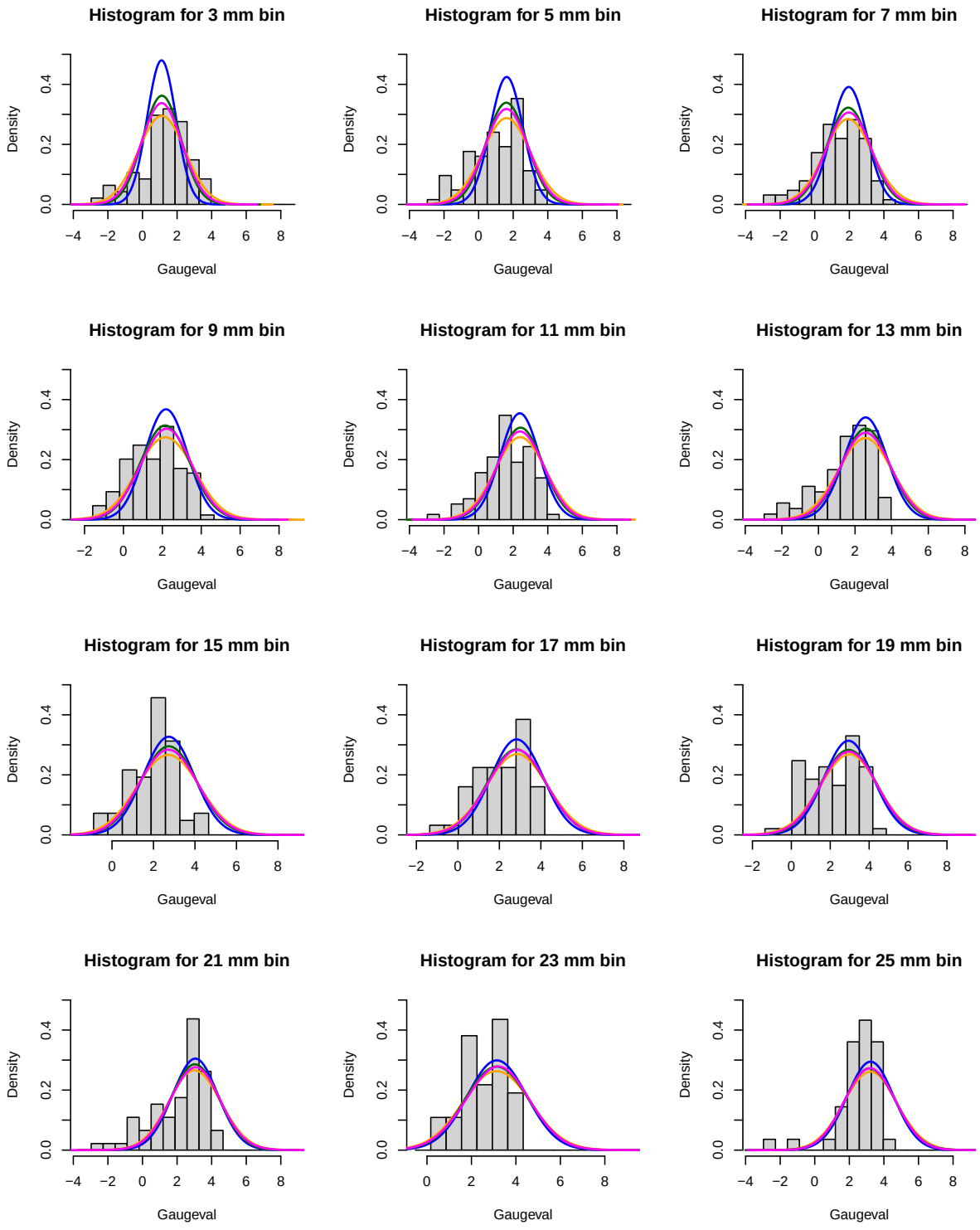


Figure B.2: Histograms for June on the logarithm of the gauge values. The mm value is the mid value for the 2mm RFE bin. Each bar is log(2)mm wide. The density curves corresponds to the normal distribution with $\mu = \log(\text{RFE})$ and standard deviation given by the coloured lines in Figure 4.6.

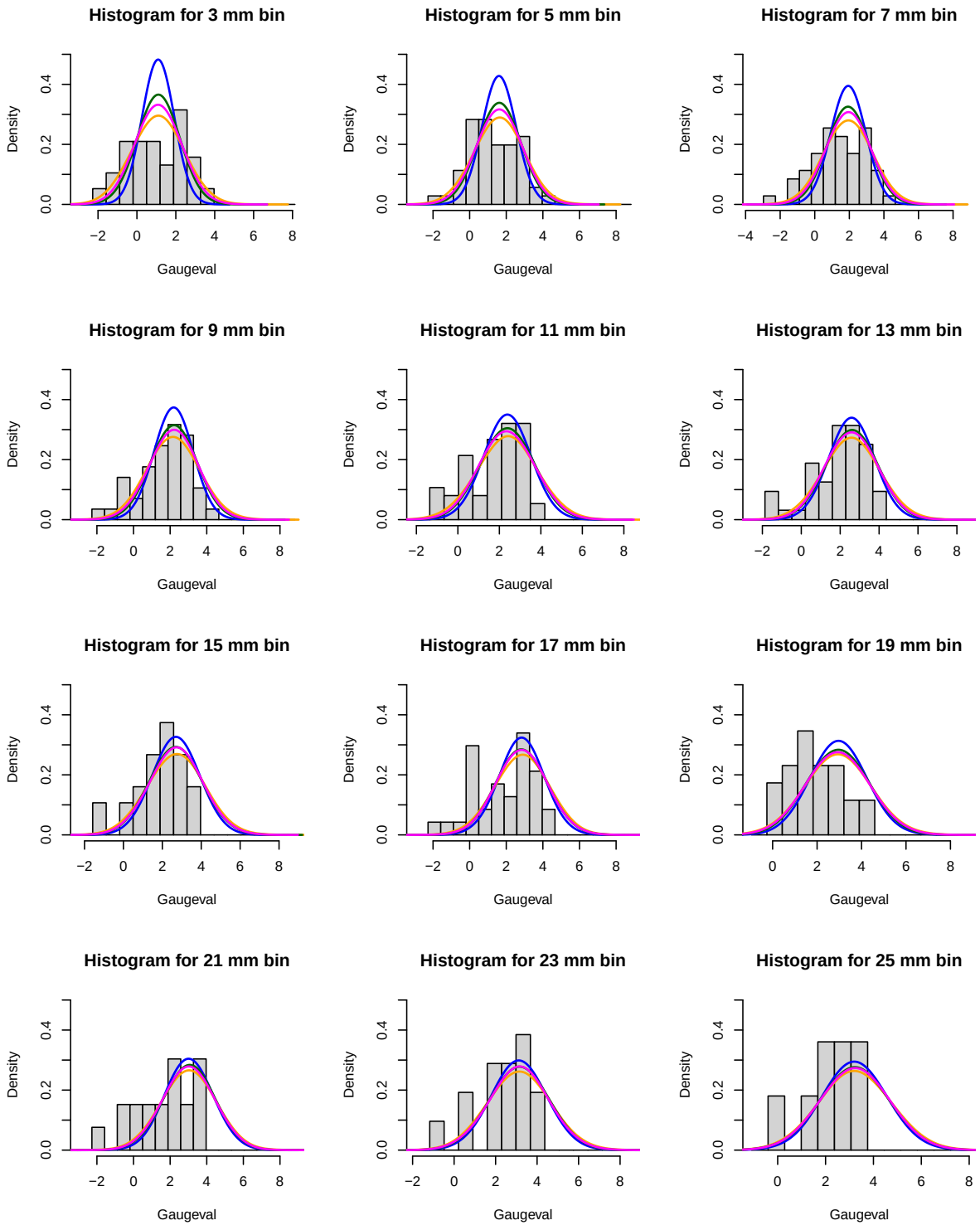


Figure B.3: Histograms for August on the logarithm of the gauge values. The mm value is the mid value for the 2mm RFE bin. Each bar is log(2)mm wide. The density curves corresponds to the normal distribution with $\mu = \log(\text{RFE})$ and standard deviation given by the coloured lines in Figure 4.6.

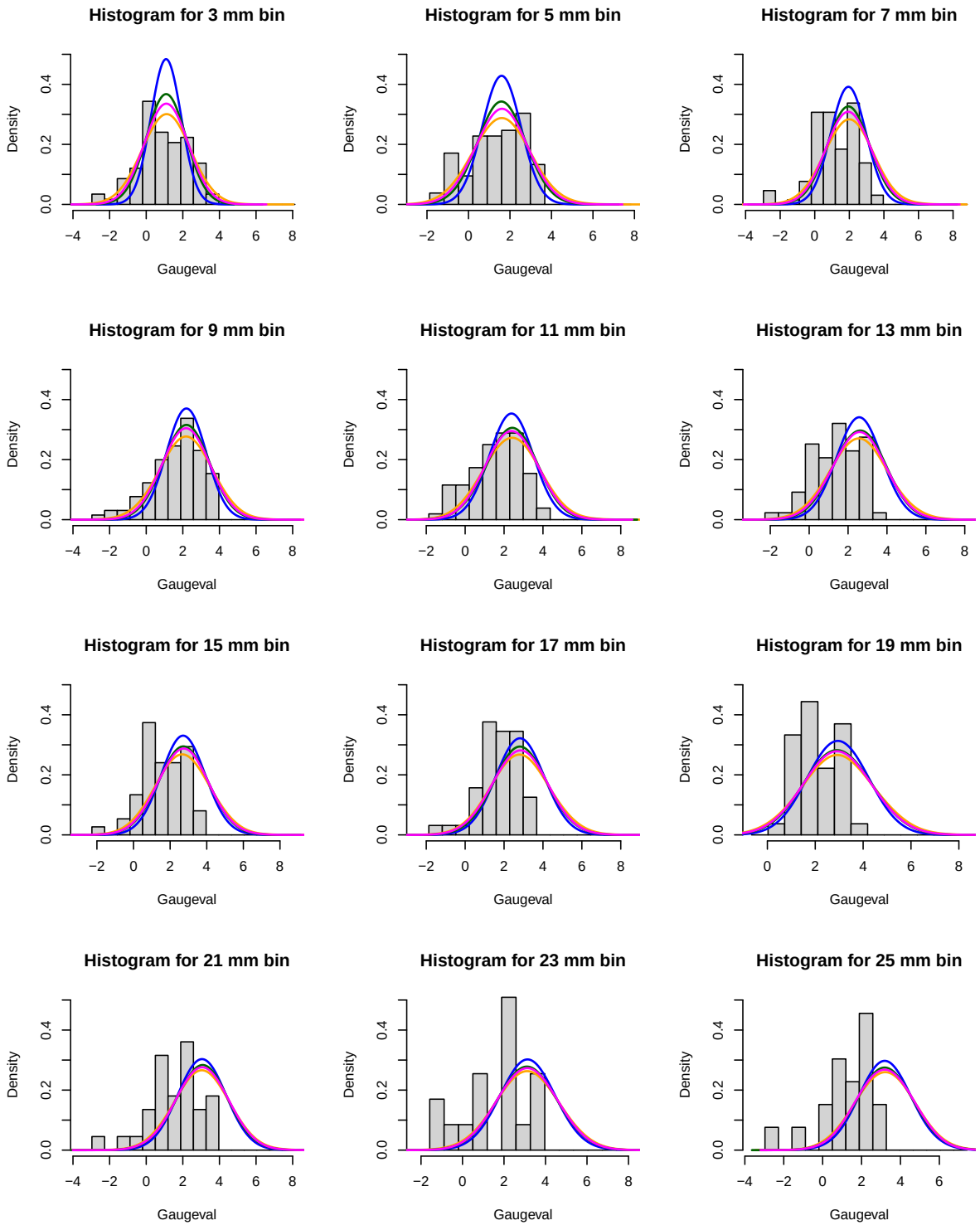


Figure B.4: Histograms for September on the logarithm of the gauge values. The mm value is the mid value for the 2mm RFE bin. Each bar is log(2)mm wide. The density curves corresponds to the normal distribution with $\mu = \log(\text{RFE})$ and standard deviation given by the coloured lines in Figure 4.6.

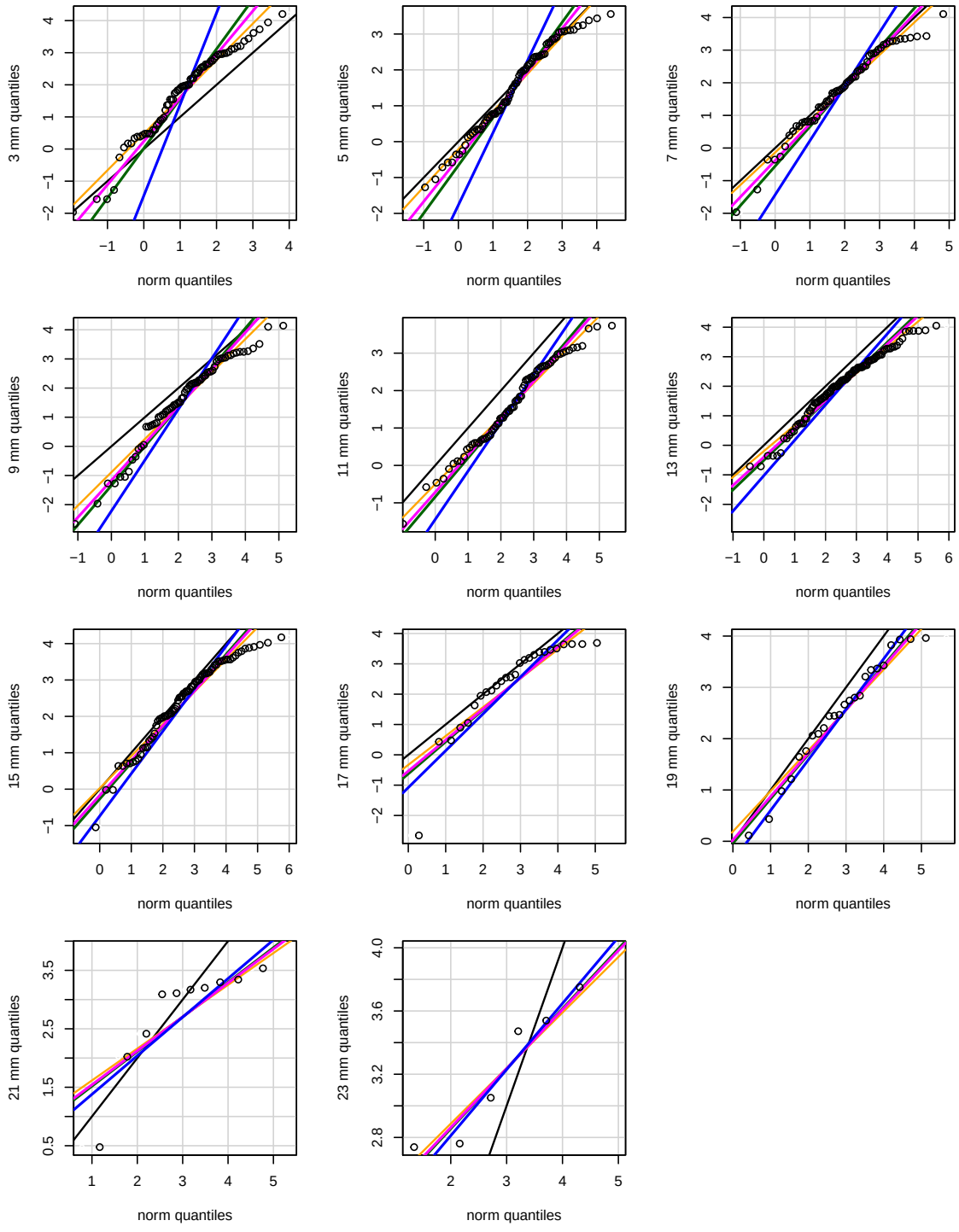


Figure B.5: QQ-plot for the logarithm of the gauge values for April with the lognormal distribution associated with $\kappa = 8, \theta = 0.2$ as reference distribution. The coloured lines are the lognormal distributions with the standard deviation given by the corresponding colours in Figure 4.6 and the black line the $x = y$ line.

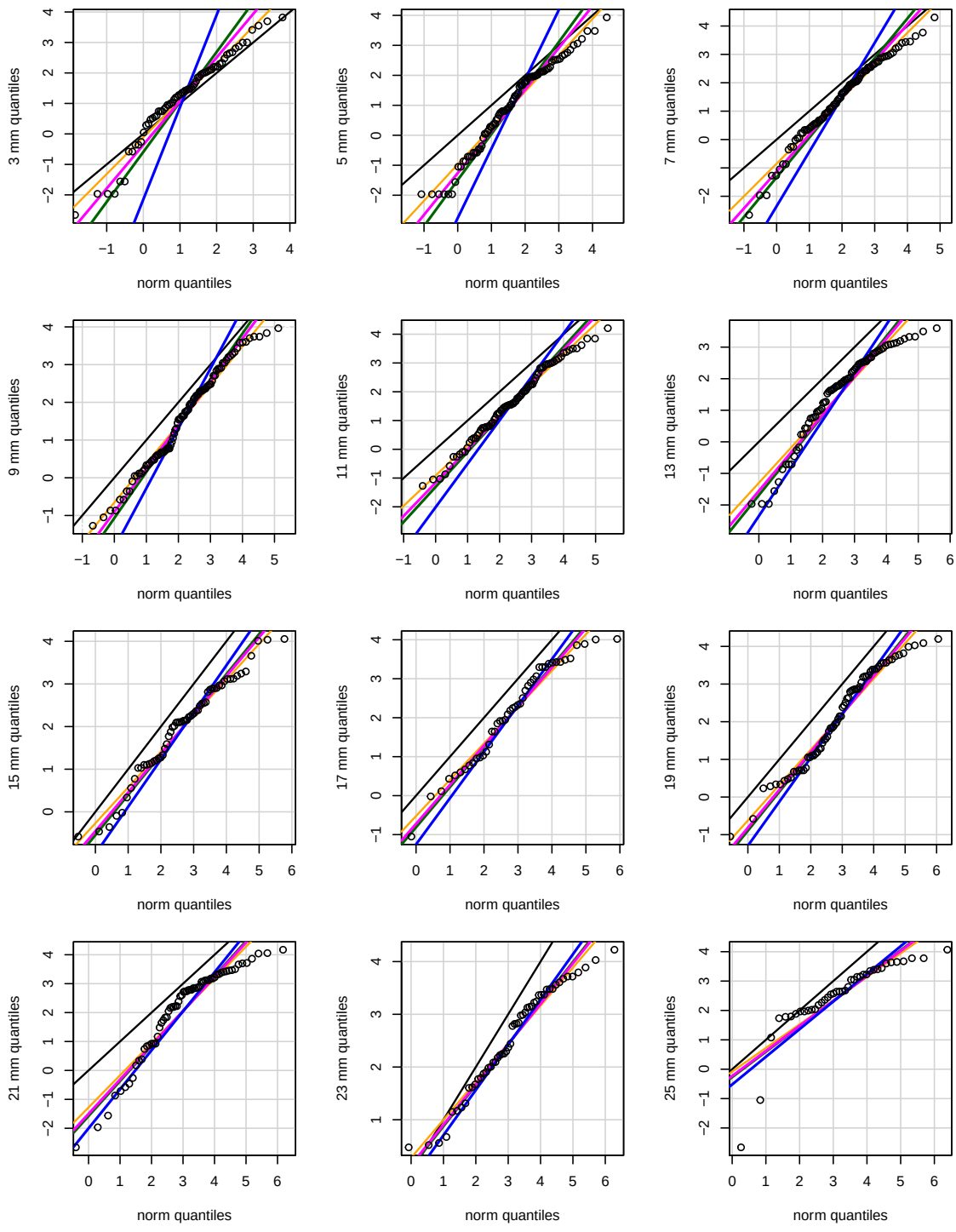


Figure B.6: QQ-plot for the logarithm of the gauge values for June with the lognormal distribution associated with $\kappa = 8, \theta = 0.2$ as reference distribution. The coloured lines are the lognormal distributions with the standard deviation given by the corresponding colours in Figure 4.6 and the black line the $x = y$ line.

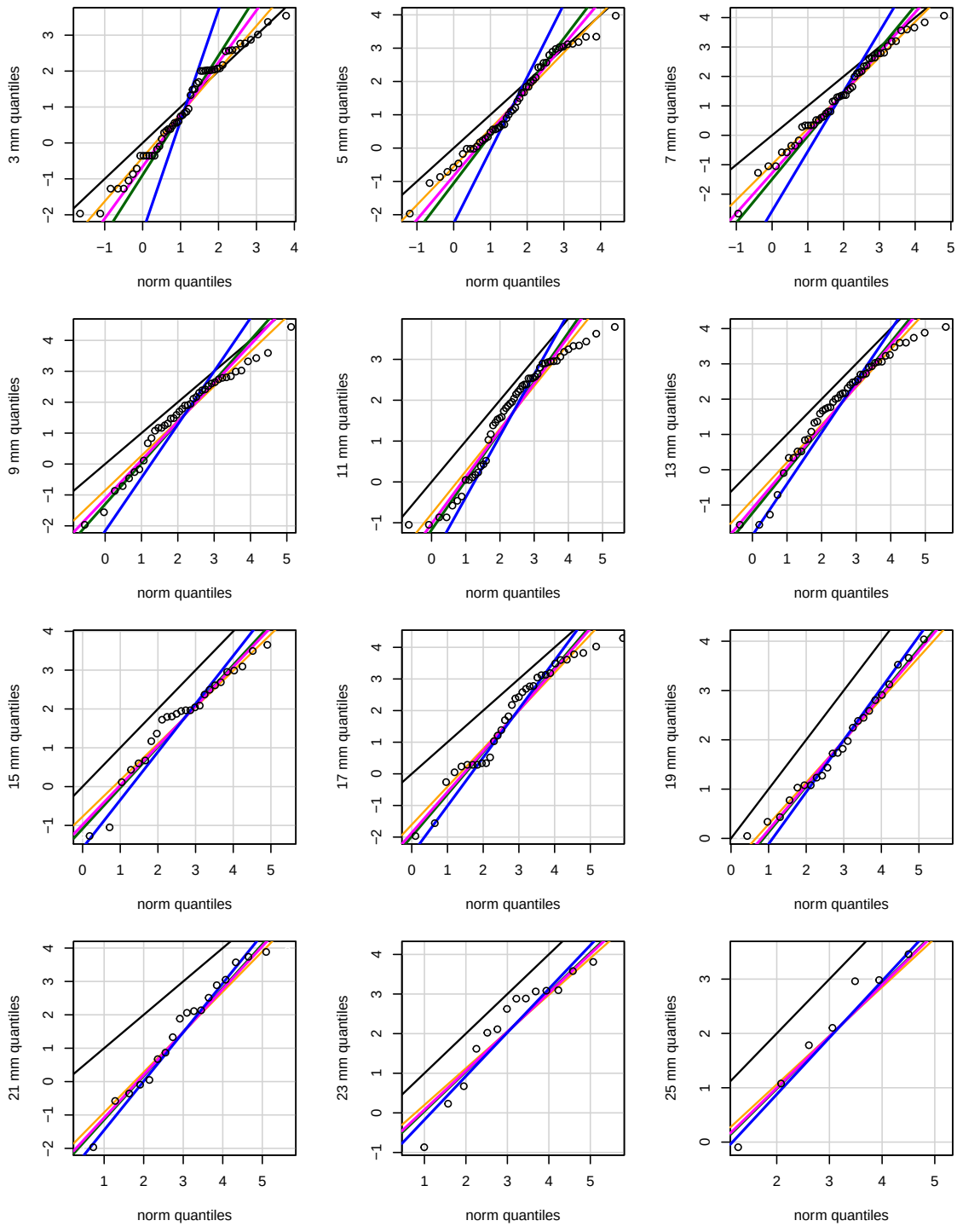


Figure B.7: QQ-plot for the logarithm of the gauge values for August with the lognormal distribution associated with $\kappa = 8, \theta = 0.2$ as reference distribution. The coloured lines are the lognormal distributions with the standard deviation given by the corresponding colours in Figure 4.6 and the black line the $x = y$ line.

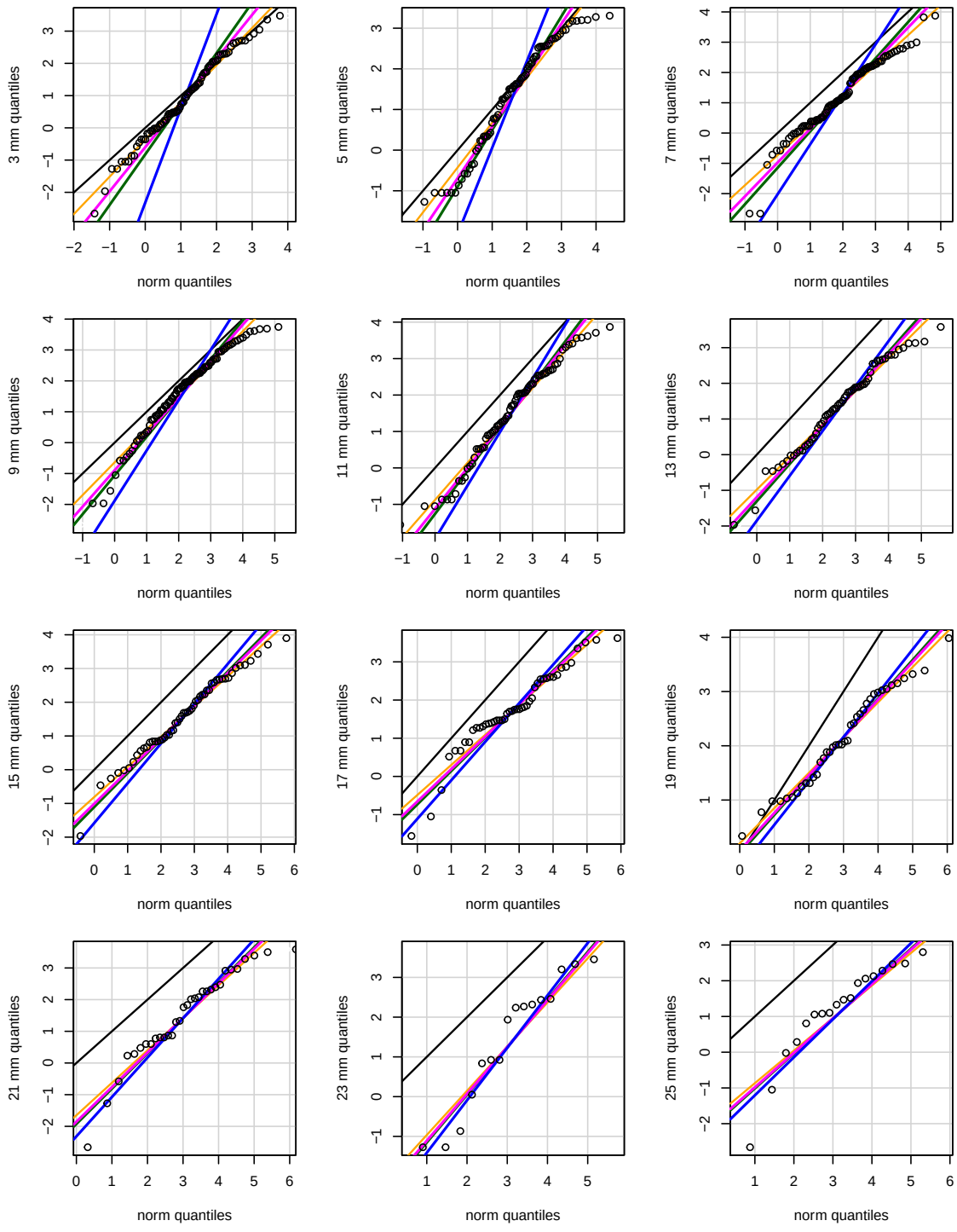


Figure B.8: QQ-plot for the logarithm of the gauge values for September with the lognormal distribution associated with $\kappa = 8, \theta = 0.2$ as reference distribution. The coloured lines are the lognormal distributions with the standard deviation given by the corresponding colours in Figure 4.6 and the black line the $x = y$ line.

Bibliography

Acheampong, P. K. (1982). "Rainfall Anomaly along the Coast of Ghana. Its Nature and Causes". In: *Geografiska Annaler. Series A, Physical Geography* 64.3/4, pp. 199–211. DOI: 10.2307/520646.

Ali, A., T. Lebel, and A. Amani (2003). "Invariance in the Spatial Structure of Sahelian Rain Fields at Climatological Scales". In: *Journal of Hydrometeorology* 4.6, pp. 996–1011. DOI: 10.1175/1525-7541(2003)004<0996:IITSSO>2.0.CO;2.

Arvind, G, P Ashok Kumar, S Girish Karthi, and C. R. Suribabu (2017). "Statistical Analysis of 30 Years Rainfall Data: A Case Study". In: *IOP Conference Series: Earth and Environmental Science* 80, p. 12067. DOI: 10.1088/1755-1315/80/1/012067.

Atyeo, J. and D. Walshaw (2012). "A region-based hierarchical model for extreme rainfall over the UK, incorporating spatial dependence and temporal trend". In: *Environmetrics* 23.6, pp. 509–521. DOI: https://doi.org/10.1002/env.2155.

Ayanlade, A., M. Radeny, J. F. Morton, and T. Muchaba (2018). "Rainfall variability and drought characteristics in two agro-climatic zones: An assessment of climate change challenges in Africa". In: *Science of The Total Environment* 630, pp. 728–737. DOI: https://doi.org/10.1016/j.scitotenv.2018.02.196.

Bacchi, B. and N. T. Kottegoda (1995). "Identification and calibration of spatial correlation patterns of rainfall". In: *Journal of Hydrology* 165.1, pp. 311–348. DOI: https://doi.org/10.1016/0022-1694(94)02590-8.

Becker, A., P. Finger, A. Meyer-Christoffer, B. Rudolf, K. Schamm, U. Schneider, and M. Ziese (2013). "A description of the global land-surface precipitation data products of the Global Precipitation Climatology Centre with sample applications including centennial (trend) analysis from 1901–present". In: *Earth Syst. Sci. Data* 5.1, pp. 71–99. DOI: 10.5194/essd-5-71-2013.

Beirlant, J, Y Goegebeur, J Segers, J Teugels, D De Waal, and C Ferro (2004). *Statistics of Extremes: Theory and Applications*. Wiley Series in Probability and Statistics. John Wiley & Sons.

Beirlant, J, G Dierckx, and A Guillou (2011). "Bias-reduced estimators for bivariate tail modelling". In: *Insurance: Mathematics and Economics* 49.1, pp. 18–26. DOI: https://doi.org/10.1016/j.insmatheco.2011.01.010.

Blanchet, J., C. Aly, T. Vischel, G. Panthou, Y. Sané, and M. Diop Kane (2018). "Trend in the Co-Occurrence of Extreme Daily Rainfall in West Africa Since 1950". In: *Journal of Geophysical Research: Atmospheres* 123.3, pp. 1536–1551. DOI: 10.1002/2017JD027219.

Boateng, E. A., J. D. Fage, O. Davies, and D. J. Maier (2018). *Ghana*.

Brooks, N. (2004). "Drought in the African Sahel: long term perspectives and future prospects". In: *Tyndall Centre for Climate Change Research, Norwich, Working Paper* 61, p. 31.

Caeiro, F., M. I. Gomes, and D. Pestana (2005). "Direct reduction of the bias of the classical hill estimator". In: *REVSTAT Statistical Journal* 3, pp. 113–136.

Chilès, J.-P. and N. Desassis (2018). "Fifty Years of Kriging". In: *Handbook of Mathematical Geosciences: Fifty Years of IAMG*. Ed. by B. S. Daya Sagar, Q. Cheng, and F. Agterberg. Cham: Springer International Publishing, pp. 589–612. DOI: 10.1007/978-3-319-78999-6_29.

Climate Prediction Center. *The NOAA Climate Prediction Center African Rainfall Estimation Algorithm Version 2.0*. Tech. rep.

Coles, S., J. Heffernan, and J. Tawn (1999). "Dependence Measures for Extreme Value Analyses". In: *Extremes* 2.4, pp. 339–365. DOI: 10.1023/A:1009963131610.

Cressie, N. and C. K. Wikle (2011). *Statistics for Spatio-Temporal data*. Hoboken, New Jersey: John Wiley & Sons, Incorporated.

Davison, A. C., S. A. Padoan, and M Ribatet (2012). "Statistical Modeling of Spatial Extremes". en. In: *Statist. Sci.* 27.2, pp. 161–186. DOI: 10.1214/11-STS376.

Debusho, L. K. and T. A. Diriba (2021). "Conditional modelling approach to multivariate extreme value distributions: application to extreme rainfall events in South Africa". In: *Environmental and Ecological Statistics*. DOI: 10.1007/s10651-021-00498-0.

Dekkers, A, J Einmahl, and L de Haan (1989). "A moment estimator for the index of an extreme-value distribution". In: *The Annals of Statistics* 17.4, pp. 1833–1855.

Diro, G. T., D. I. F. Grimes, E Black, A O'Neill, and E Pardo-Iguzquiza (2009). "Evaluation of reanalysis rainfall estimates over Ethiopia". In: *International Journal of Climatology* 29.1, pp. 67–78. DOI: 10.1002/joc.1699.

Draisma, G., H. Drees, A. Ferreira, and L. De Haan (2004). "Bivariate tail estimation: dependence in asymptotic independence". en. In: *Bernoulli* 10.2, pp. 251–280. DOI: 10.3150/bj/1082380219.

Drees, H. (1998). "A general class of estimators of the extreme value index". In: *Journal of Statistical Planning and Inference* 66.1, pp. 95–112. DOI: https://doi.org/10.1016/S0378-3758(97)00076-1.

Dugdale, G, V. D. McDougall, and J. R. Milford (1991). "Rainfall estimates in the Sahel from cold cloud statistics : Accuracy and limitations of operational systems". In: *Soil water balance in the Sudano-Sahelian Zone*. Wallingford: International Association of Hydrological Sciences.

Dunning, C. M., E. C. L. Black, and R. P. Allan (2016). "The onset and cessation of seasonal rainfall over Africa". In: *Journal of Geophysical Research: Atmospheres* 121.19, pp. 11,405–411,424. DOI: https://doi.org/10.1002/2016JD025428.

Einmahl, J. H. J. (1997). "Poisson and Gaussian approximation of weighted local empirical processes". In: *Stochastic Processes and their Applications* 70.1, pp. 31–58. DOI: https://doi.org/10.1016/S0304-4149(97)00055-0.

Einmahl, J. H. J., L. de Haan, and C. Zhou (2016). "Statistics of heteroscedastic extremes". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 78.1, pp. 31–51. DOI: https://doi.org/10.1111/rssb.12099.

Ferro, C. A. T. and J. Segers (2003). "Inference for clusters of extreme values". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65.2, pp. 545–556. DOI: https://doi.org/10.1111/1467-9868.00401.

Feuerverger, A. and P. Hall (1999). "Estimating a tail exponent by modelling departure from a Pareto distribution". en. In: *Ann. Statist.* 27.2, pp. 760–781. DOI: 10.1214/aos/1018031215.

Fisher, R. and L. Tippett (1928). "Limiting Forms of the Frequency Distribution of the Largest or Smallest Members of a Sample". In: *Proceedings of the Cambridge Philosophical Society* 24, pp. 180–190.

Funk, C., A. Verdin, J. Michaelsen, P. Peterson, D. Pedreros, and G. Husak (2015a). "A global satellite-assisted precipitation climatology". In: *Earth Syst. Sci. Data* 7.2, pp. 275–287. DOI: 10.5194/essd-7-275-2015.

Funk, C., P. Peterson, M. Landsfeld, D. Pedreros, J. Verdin, S. Shukla, G. Husak, J. Rowland, L. Harrison, A. Hoell, and J. Michaelsen (2015b). "The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes". In: *Scientific Data* 2.1, p. 150066. DOI: 10.1038/sdata.2015.66.

Gnedenko, B. (1943). "Sur la distribution limite du terme maximum of d'unesérie Aléatorie". In: *Annals of Mathematics* 44, pp. 423–453.

Goegebeur, Y. and A. Guillou (2012). "Asymptotically Unbiased Estimation of the Coefficient of Tail Dependence". In: *Scandinavian Journal of Statistics* 40.1, pp. 174–189. DOI: 10.1111/j.1467-9469.2012.00800.x.

Gomes, M. I. and D. Pestana (2007). "A Sturdy Reduced-Bias Extreme Quantile (VaR) Estimator". In: *Journal of the American Statistical Association* 102.477, pp. 280–292. DOI: 10.1198/016214506000000799.

Gomes, M. I., M. F. Brilhante, and D. Pestana (2016). "New Reduced-bias Estimators of a Positive Extreme Value Index". In: *Communications in Statistics - Simulation and Computation* 45.3, pp. 833–862. DOI: 10.1080/03610918.2013.875567.

Gomes, M. and F. Caeiro (2014). "Efficiency of partially reduced-bias mean-of-order-p versus minimum-variance reduced-bias extreme value index estimation." In: *COMPSTAT 2014*, pp. 289–298.

Greatrex, H., D. Grimes, and T. Wheeler (2014). "Advances in the Stochastic Modeling of Satellite-Derived Rainfall Estimates Using a Sparse Calibration Dataset". In: *Journal of Hydrometeorology* 15.5, pp. 1810–1831. DOI: 10.1175/JHM-D-13-0145.1.

Grimes, D. I. F., E. Pardo-Igúzquiza, and R. Bonifacio (1999). "Optimal areal rainfall estimation using raingauges and satellite data". In: *Journal of Hydrology* 222.1, pp. 93–108. DOI: https://doi.org/10.1016/S0022-1694(99)00092-X.

Gyasi-Agyei, Y. and G. Pegram (2014). "Interpolation of daily rainfall networks using simulated radar fields for realistic hydrological modelling of spatial rain field ensem-

bles". In: *Journal of Hydrology* 519, pp. 777–791. DOI: https://doi.org/10.1016/j.jhydrol.2014.08.006.

Haan, L. de (1970). "On Regular Variation and its Application to the Weak Convergence of Sample Extremes". In: *Mathematical Centre Tract 32* 32.

Haan, L. de and A. Ferreira (2006). *Extreme value theory*. 1st ed. New York, NY: Springer New York, p. 418. DOI: https://doi.org/10.1007/0-387-34471-3.

Haan, L. de, A. K. Tank, and C. Neves (2015). "On tail trend detection: modeling relative risk". In: *Extremes* 18.2, pp. 141–178. DOI: 10.1007/s10687-014-0207-8.

Hall, P. and A. H. Welsh (1985). "Adaptive Estimates of Parameters of Regular Variation". In: *The Annals of Statistics* 13.1, pp. 331–341. DOI: 10.1214/aos/1176346596.

Heffernan, J. E. (2000). "A Directory of Coefficients of Tail Dependence". In: *Extremes* 3.3, pp. 279–290. DOI: 10.1023/A:1011459127975.

Heffernan, J. E. and J. A. Tawn (2004). "A conditional approach for multivariate extreme values (with discussion)". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 66.3, pp. 497–546. DOI: https://doi.org/10.1111/j.1467-9868.2004.02050.x.

Hill, B. M. (1975). "A Simple General Approach to Inference About the Tail of a Distribution". en. In: *Ann. Statist.* 3.5, pp. 1163–1174. DOI: 10.1214/aos/1176343247.

Hsing, T. (1991). "Estimating the parameters of rare events". In: *Stochastic Processes and their Applications* 37, pp. 117–139.

Huffman, G. J., D. T. Bolvin, E. J. Nelkin, D. B. Wolff, R. F. Adler, G. Gu, Y. Hong, K. P. Bowman, and E. F. Stocker (2007). "The TRMM Multisatellite Precipitation Analysis (TMPA): Quasi-Global, Multiyear, Combined-Sensor Precipitation Estimates at Fine Scales". English. In: *Journal of Hydrometeorology* 8.1, pp. 38–55. DOI: 10.1175/JHM560.1.

Huffman, G. J., D. T. Bolvin, D. Braithwaite, K.-l. Hsu, R. Joyce, C. Kidd, E. J. Nelkin, and P. Xie (2015). *NASA Global Precipitation Measurement (GPM) Integrated Multi-satellitE Retrievals for GPM (IMERG). Algorithm Theoretical Basis Document version 4.5*. Tech. rep. NASA, p. 26.

Huser, R and A. C. Davison (2014). "Space-time modelling of extreme events". In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 76.2, pp. 439–461.

Huser, R. and J. L. Wadsworth (2020). "Advances in statistical modeling of spatial extremes". In: *WIREs Computational Statistics* n/a.n/a, e1537. DOI: https://doi.org/10.1002/wics.1537.

IEA (2020). *Climate Impacts on African Hydropower*. Tech. rep. Paris: IEA.

Isaaks, E. H. and R. M. Srivastave (1989). *An introduction to Applied Geostatistics*. New York, UNITED STATES: Oxford University Press, Inc.

Israelsson, J., E. Black, C. Neves, F. F. Torgbor, H. Greatrex, M. Tanu, and P. N. L. Lamptey (2020). "The spatial correlation structure of rainfall at the local scale over southern Ghana". In: *Journal of Hydrology: Regional Studies* 31. DOI: 10.1016/j.ejrh.2020.100720.

Joyce, R. J., J. E. Janowiak, P. A. Arkin, and P. Xie (2004). "CMORPH: A Method that Produces Global Precipitation Estimates from Passive Microwave and Infrared Data at High Spatial and Temporal Resolution". English. In: *Journal of Hydrometeorology* 5.3, pp. 487–503. DOI: 10.1175/1525-7541(2004)005<0487:CAMTPG>2.0.CO;2.

Lacombe, G, M McCartney, and G Forkuor (2012). "Drying climate in Ghana over the period 1960–2005: evidence from the resampling-based Mann-Kendall test at local and regional levels". In: *Hydrological Sciences Journal* 57.8, pp. 1594–1609. DOI: 10.1080/02626667.2012.728291.

Laux, P, S Wagner, A Wagner, J Jacobeit, A Bárdossy, and H Kunstmann (2009). "Modelling daily precipitation features in the Volta Basin of West Africa". In: *International Journal of Climatology* 29.7, pp. 937–954. DOI: 10.1002/joc.1852.

Ledford, A. W. and J. A. Tawn (1996). "Statistics for near independence in multivariate extreme values". In: *Biometrika* 83.1, pp. 169–187. DOI: 10.1093/biomet/83.1.169.

Ledford, A. W. and J. A. Tawn (1997). "Modelling Dependence Within Joint Tail Regions". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 59.2, pp. 475–499.

Maidment, R., E. Black, H. Greatrex, and M. Young (2020). "TAMSAT - Satellite Precipitation Measurement: Volume 1". In: ed. by V. Levizzani, C. Kidd, D. B. Kirschbaum, C. D. Kummerow, K. Nakamura, and F. J. Turk. Springer, Cham, pp. 393–407. DOI: 10.1007/978-3-030-24568-9_22.

Maidment, R. I., D. Grimes, R. P. Allan, E. Tarnavsky, M. Stringer, T. Hewison, R. Roebeling, and E. Black (2014). "The 30 year TAMSAT African Rainfall Climatology And

Time series (TARCAT) data set". In: *Journal of Geophysical Research: Atmospheres* 119.18, pp. 10,610–619,644. DOI: 10.1002/2014JD021927.

Maidment, R. I., D. Grimes, E. Black, E. Tarnavsky, M. Young, H. Greatrex, R. P. Allan, T. Stein, E. Nkonde, S. Senkunda, and E. M. U. Alcántara (2017). "A new, long-term daily satellite-based rainfall dataset for operational monitoring in Africa". In: *Scientific Data* 4.

Maranan, M., A. H. Fink, and P. Knippertz (2018). "Rainfall types over southern West Africa: Objective identification, climatology and synoptic environment". In: *Quarterly Journal of the Royal Meteorological Society* 144.714, pp. 1628–1648. DOI: 10.1002/qj.3345.

Matheron, G. and F. Blondel (1962). *Traité de géostatistique appliquée, Tome I: Mémoíres du Bureau de Recherches Géologiques et Minières*. 14th ed. Paris: Technip.

Mathon, V., H. Laurent, and T. Lebel (2002). "Mesoscale Convective System Rainfall in the Sahel". English. In: *Journal of Applied Meteorology* 41.11, pp. 1081–1092. DOI: 10.1175/1520-0450(2002)041<1081:MCSRIT>2.0.CO;2.

McGregor, G. and S. Nieuwolt (1998). *Tropical Climatology*. 2nd ed. Chichester: John Wiley & Sons.

Milford, J. R., V. D. McDougall, and G Dugdale (1996). "Rainfall estimation from cold cloud duration: Experience of the TAMSAT group in West Africa". In: *Colloques et Seminaires. Validation Problems of Rainfall Estimation Methods by Satellite in Intertropical Africa. Proceedings of the Niamey Workshop (1994)*. ORSTOM.

Moron, V., A. W. Robertson, M. N. Ward, and P. Camberlin (2007). "Spatial Coherence of Tropical Rainfall at the Regional Scale". In: *Journal of Climate* 20.21, pp. 5244–5263. DOI: 10.1175/2007JCLI1623.1.

Nelsen, R. B. (2006). *An Introduction to Copulas*. 2nd ed. Springer Series in Statistics. New York, NY: Springer New York, p. 272. DOI: 10.1007/0-387-28678-0.

Nesbitt, S. W., E. J. Zipser, and D. J. Cecil (2000). "A Census of Precipitation Features in the Tropics Using TRMM: Radar, Ice Scattering, and Lightning Observations". English. In: *Journal of Climate* 13.23, pp. 4087–4106. DOI: 10.1175/1520-0442(2000)013<4087:ACOPFI>2.0.CO;2.

Nicholson, S. E., B Some, and B Kone (2000). "An Analysis of Recent Rainfall Conditions in West Africa, Including the Rainy Seasons of the 1997 El Niño and the 1998 La Niña

Years". In: *Journal of Climate* 13.14, pp. 2628–2640. DOI: 10.1175/1520-0442(2000)013<2628:AAORRC>2.0.CO;2.

Nicholson, S. E., A. H. Fink, and C. Funk (2018). "Assessing recovery and change in West Africa's rainfall regime from a 161-year record". In: *International Journal of Climatology* 38.10, pp. 3770–3786. DOI: 10.1002/joc.5530.

Novella, N. S. and W. M. Thiaw (2013). "African Rainfall Climatology Version 2 for Famine Early Warning Systems". English. In: *Journal of Applied Meteorology and Climatology* 52.3, pp. 588–606. DOI: 10.1175/JAMC-D-11-0238.1.

Nyarko Kumi, E. (2017). *The Electricity Situation in Ghana: Challenges and Opportunities. CGD Policy Paper. Washington DC: Center for Global Development*. Washington, DC.

Owusu, K. and P. Waylen (2009). "Trends in spatio-temporal variability in annual rainfall in Ghana (1951-2000)". In: *Weather* 64.5, pp. 115–120. DOI: 10.1002/wea.255.

Parker, D. and M. Diop-Kane (2017). *Meteorology of Tropical West Africa: The Forecasters' Handbook*. Wiley-Blackwell.

Peng, L. (1999). "Estimation of the coefficient of tail dependence in bivariate extremes". In: *Statistics and Probability Letters* 43.4, pp. 399–409.

Reynolds, R. W. (1988). "A Real-Time Global Sea Surface Temperature Analysis". English. In: *Journal of Climate* 1.1, pp. 75–87. DOI: 10.1175/1520-0442(1988)001<0075:ARTGSS>2.0.CO;2.

Ricciardulli, L. and P. D. Sardeshmukh (2002). "Local Time- and Space Scales of Organized Tropical Deep Convection". In: *Journal of Climate* 15.19, pp. 2775–2790. DOI: 10.1175/1520-0442(2002)015<2775:LTASSO>2.0.CO;2.

Romão, X., R. Delgado, and A. Costa (2010). "An empirical power comparison of univariate goodness-of-fit tests for normality". In: *Journal of Statistical Computation and Simulation* 80.5, pp. 545–591. DOI: 10.1080/00949650902740824.

Roth, M, T. A. Buishand, G Jongbloed, A. M. G. Klein Tank, and J. H. van Zanten (2014). "Projections of precipitation extremes based on a regional, non-stationary peaks-over-threshold approach: A case study for the Netherlands and north-western Germany". In: *Weather and Climate Extremes* 4, pp. 1–10. DOI: https://doi.org/10.1016/j.wace.2014.01.001.

Sang, H. and A. E. Gelfand (2009). "Hierarchical modeling for extreme values observed over space and time". In: *Environmental and Ecological Statistics* 16.3, pp. 407–426. DOI: 10.1007/s10651-007-0078-0.

Shang, H., J. Yan, and X. Zhang (2011). "El Niño–Southern Oscillation influence on winter maximum daily precipitation in California in a spatial model". In: *Water Resources Research* 47.11. DOI: https://doi.org/10.1029/2011WR010415.

Shepard, D. (1968). "A Two-Dimensional Interpolation Function for Irregularly-Spaced Data". In: *Proceedings of the 1968 23rd ACM National Conference*. ACM '68. New York, NY, USA: Association for Computing Machinery, pp. 517–524. DOI: 10.1145/800186.810616.

Shooter, R, E Ross, J Tawn, and P Jonathan (2019). "On spatial conditional extremes for ocean storm severity". In: *Environmetrics* 30.6, e2562. DOI: https://doi.org/10.1002/env.2562.

Shooter, R., J. Tawn, E. Ross, and P. Jonathan (2021). "Basin-wide spatial conditional extremes for severe ocean storms". In: *Extremes* 24.2, pp. 241–265. DOI: 10.1007/s10687-020-00389-w.

Sibuya, M. (1960). "Bivariate extreme statistics, I". In: *Annals of the Institute of Statistical Mathematics* 11.3, pp. 195–210. DOI: 10.1007/BF01682329.

Sillmann, J, V. V. Kharin, X Zhang, F. W. Zwiers, and D Bronaugh (2013). "Climate extremes indices in the CMIP5 multimodel ensemble: Part 1. Model evaluation in the present climate". In: *Journal of Geophysical Research: Atmospheres* 118.4, pp. 1716–1733. DOI: 10.1002/jgrd.50203.

Sklar, A. (1959). "Functions de Repartition an Dimension Set Leursmarges". In: *Publications de L'In-stitut de Statistique de L'Universite de Paris* 8, pp. 229–231.

Slater, L. J., B Anderson, M Buechel, S Dadson, S Han, S Harrigan, T Kelder, K Kowal, T Lees, T Matthews, C Murphy, and R. L. Wilby (2021). "Nonstationary weather and water extremes: a review of methods for their detection, attribution, and management". In: *Hydrology and Earth System Sciences* 25.7, pp. 3897–3935. DOI: 10.5194/hess-25-3897-2021.

Smith, D., A. Gasiewski, D. Jackson, and G. Wick (2005). "Spatial scales of tropical precipitation inferred from TRMM microwave imager data". In: *IEEE Transactions*

on Geoscience and Remote Sensing 43.7, pp. 1542–1551. DOI: 10.1109/TGRS.2005.848426.

Smith, R. L. (1987). "Estimating Tails of Probability Distributions". In: *The Annals of Statistics* 15.3, pp. 1174–1207.

Smith, R. L. and I. Weissman (1994). "Estimating the Extremal Index". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 56.3, pp. 515–528. DOI: https://doi.org/10.1111/j.2517-6161.1994.tb01997.x.

SRID Ministry of food and agriculture (2017). *Agriculture in Ghana*. Tech. rep.

Sui, D. Z. (2004). "Tobler's First Law of Geography: A Big Idea for a Small World?" In: *Annals of the Association of American Geographers* 94.2, pp. 269–277. DOI: 10.1111/j.1467-8306.2004.09402003.x.

Tarnavsky, E., D. Grimes, R. Maidment, E. Black, R. P. Allan, M. Stringer, R. Chadwick, and F. Kayitakire (2014). "Extension of the TAMSAT Satellite-Based Rainfall Monitoring over Africa and from 1983 to Present". In: *Journal of Applied Meteorology and Climatology* 53.12, pp. 2805–2822. DOI: 10.1175/JAMC-D-14-0016.1.

Tawn, J., R. Shooter, R. Towe, and R. Lamb (2018). "Modelling spatial extreme events with environmental applications". In: *Spatial Statistics* 28, pp. 39–58. DOI: https://doi.org/10.1016/j.spasta.2018.04.007.

Taylor, C. M., D. Belušić, F. Guichard, D. J. Parker, T. Vischel, O. Bock, P. P. Harris, S. Janicot, C. Klein, and G. Panthou (2017). "Frequency of extreme Sahelian storms tripled since 1982 in satellite observations". In: *Nature* 544, p. 475.

Teo, C.-K. and D. I. F. Grimes (2007). "Stochastic modelling of rainfall from satellite data". In: *Journal of Hydrology* 346.1, pp. 33–50. DOI: https://doi.org/10.1016/j.jhydrol.2007.08.014.

Thibaud, E, R Mutzner, and A. C. Davison (2013). "Threshold modeling of extreme spatial rainfall". In: *Water Resources Research* 49.8, pp. 4633–4644. DOI: 10.1002/wrcr.20329.

Torgbor, F., D. A. Stern, B. K. Nkansah, and R. D. Stern (2018). "Rainfall Modelling with a Transect View in Ghana". In: *Ghana Journal of Science* 58, pp. 41–57.

Wadsworth, J. L. and J. Tawn (2019). *Higher-dimensional spatial extremes via single-site conditioning*.

Wadsworth, J. L. and J. A. Tawn (2012). "Dependence modelling for spatial extremes". In: *Biometrika* 99.2, pp. 253–272. DOI: 10.1093/biomet/asr080.

Washington, R., M. Harrison, D. Conway, E. Black, A. Challinor, D. Grimes, R. Jones, A. Morse, and M. Todd (2004). *African climate report: A report commissioned by the UK Government to review African climate science, policy and options for action*. Tech. rep. DFID/DEFRA.

Winter, H. (2016). "Extreme value modeling of heatwaves". PhD thesis. Lancaster University, p. 220.

Xie, P. and P. A. Arkin (1996). "Analyses of Global Monthly Precipitation Using Gauge Observations, Satellite Estimates, and Numerical Model Predictions". English. In: *Journal of Climate* 9.4, pp. 840–858. DOI: 10.1175/1520-0442(1996)009<0840:AOGMPU>2.0.CO;2.

Xie, P., R. Joyce, S. Wu, S.-H. Yoo, Y. Yarosh, F. Sun, and R. Lin (2017). "Reprocessed, Bias-Corrected CMORPH Global High-Resolution Precipitation Estimates from 1998". English. In: *Journal of Hydrometeorology* 18.6, pp. 1617–1641. DOI: 10.1175/JHM-D-16-0168.1.

Young, M. P., C. J. R. Williams, J. C. Chiu, R. I. Maidment, and S.-H. Chen (2014). "Investigation of Discrepancies in Satellite Rainfall Estimates over Ethiopia". In: *Journal of Hydrometeorology* 15.6, pp. 2347–2369. DOI: 10.1175/JHM-D-13-0111.1.