Universität
Augsburg
University

Statistical and Stochastic Post-Processing of Regional Climate Model Data: Copula-based Downscaling, Disaggregation and Multivariate Bias Correction

Dissertation
zur Erlangung des Doktorgrades an der
Fakultät für Angewandte Informatik
der Universität Augsburg

vorgelegt von

Manuel Lorenz

2019

Erstgutachter:	Prof. Dr. Harald Kunstmann
Zweitgutachter:	Prof. Dr. Elke Hertig
Tag der mündlichen Prüfung:	23.7.2020
Prüfer:	Prof. Dr. Harald Kunstmann, Prof. Dr. Elke Hertig, PD Dr. Christoph Beck

Abstract

In order to delineate management or climate change adaptation strategies for natural or technical water bodies, impact studies are necessary. To this end, impact models are set up for a given region which requires time series of meteorological data as driving data. Regional climate models (RCMs) are capable of simulating gridded data sets of several meteorological variables. The advantages over observed data are that the time series are complete and that meteorological information is also provided for ungauged locations. Furthermore, climate change impact studies can be conducted by driving the simulations with different forcing variables for future periods. While the performance of RCMs generally increases with a higher spatio-temporal resolution, the computational and storage demand increases non-linearly which can impede such highly resolved simulations in practice. Furthermore, systematic biases of the univariate distributions and multivariate dependence structures are a common problem of RCM simulations on all spatio-temporal scales.

Depending on the case study, meteorological data must fulfill different criteria. For instance, the spatio-temporal resolution of precipitation time series should be as fine as 1 km and 5 minutes in order to be used for urban hydrological impact models. To bridge the gap between the demands of impact modelers and available meteorological RCM data, different computationally efficient statistical and stochastic post-processing techniques have been developed to correct the bias and to increase the spatio-temporal resolution. The main meteorological variable treated in this thesis is precipitation due to its importance for hydrological impact studies. The models presented in this thesis belong to the classes of bias correction, downscaling and temporal disaggregation techniques. The focus of the developed methods lies on multivariate copulas. Copulas constitute a promising modeling approach for highly-skewed and mixed discrete-continuous variables like precipitation since the marginal distribution is treated separately from the dependence structure. This feature makes them useful for the modeling of different meteorological variables as well. While copulas have been utilized in the past to model precipitation and other meteorological variables that are relevant in hydrology, applications to RCM simulations are not very common.

The first method is a geostatistical estimation technique for distribution parameters of daily precipitation for ungauged locations, so that a bias correction with Quantile Mapping can be performed. The second method is a spatial downscaling of coarse scale RCM precipitation fields to a finer resolved domain. The model is based on the Gaussian Copula and generates ensembles of daily precipitation fields that resemble the precipitation fields of fine scale RCM simulations. The third method disaggregates hourly precipitation time series simulated by an RCM to a resolution of 5 minutes. The Gaussian Copula was utilized to condition the simulation on both spatial and temporal precipitation amounts to respect the spatio-temporal dependence structure. The fourth method is an approach to simulate a meteorological variable conditional on other variables at the same location and time step. The method was developed to improve the inter-variable dependence structure of univariately bias corrected RCM simulations in an hourly resolution.

Contents

List of Figures

List of Tables

List of Abbreviations

ACF	**A**uto **C**orrelation **F**unction
AMOC	**A**tlantic **M**eridional **O**verturning **C**irculation
BIC	**B**ayesian **I**nformation **C**riterion
BC	**B**ias **C**orrection
BLRP	**B**artlett **L**ewis **R**ectangular **P**ulse Model
BMBF	**B**undes**m**inisterium für **B**ildung und **F**orschung
BS	**B**rier **S**core
BSS	**B**rier **S**kill **S**core
CCDF	**C**onditional **C**umulative **D**istribution **F**unction
CCF	**C**ross **C**orrelation **F**unction
CDF	**C**umulative **D**istribution **F**unction
CMIP	**C**oupled **M**odel **I**ntercomparison **P**roject
CORDEX	**Co**ordinated **R**egional Climate **D**ownscaling **E**xperiment
CPDF	**C**onditional **P**robability **D**ensity **F**unction
CRU	**C**limatic **R**esearch **U**nit
DDC	**D**ry **D**ay **C**orrection
DJF	**D**ecember **J**anuary **F**ebruary
DOY	**D**ay **O**f **Y**ear
ECDF	**E**mpirical **C**umulative **D**istribution **F**unction
GCM	**G**eneral **C**irculation **M**odel
GPCP	**G**lobal **P**recipitation **C**limatology **P**roject
IDW	**I**nverse **D**istance **W**eighting
iid	**i**ndependent **i**dentically **d**istributed
IPCC	**I**ntergovernmental **P**anel on **C**limate **C**hange
JJA	**J**une **J**uly **A**ugust
KDE	**K**ernel **D**ensity **E**stimation
KS Test	**K**olmogorov **S**mirnov Test
MAE	**M**ean **A**bsolute **E**rror
MAM	**M**arch **A**pril **M**ay
ME	**M**ean **E**rror
MLM	**M**aximum **L**ikelihood **M**ethod
MLR	**M**ultiple **L**inear **R**egression
MOM	**M**ethod **O**f **M**oments
NSRP	**N**eyman **S**cott **R**ectangular **P**ulse Model
PACF	**P**artial **A**uto **C**orrelation **F**unction
PDF	**P**robability **D**ensity **F**unction
QM	**Q**auntile **M**apping
QMV	**Q**auntile **M**apping with **V**ine Copula Simulation
QQ-Plot	**Q**uantile-**Q**uantile-Plot
RCM	**R**egional **C**limate **M**odel
RCP	**R**epresentative **C**oncentration **P**athway
RMSE	**R**oot **M**ean **S**quare **E**rror
SON	**S**eptember **O**ctober **N**ovember

SYNOPSE	**S**ynthetische **N**iederschlagszeitreihen für die **o**ptimale **P**lanung und den Betrieb von **S**tadt**e**ntwässerungssystemen
WASCAL	**W**est **A**frican **S**cience Service **C**enter on Climate Change and **A**dapted **L**and Use
WRF	**W**eather **R**esearch and **F**orecasting Model

List of Symbols

a_{uni}	Lower boundary parameter of the uniform distribution
a_{vsp}	Range parameter of spherical variogram model
b_{uni}	Upper boundary parameter of the uniform distribution
c	Copula density
C	Copula
C_0	Nugget value of variogram
C_{vsp}	Sill parameter of spherical variogram model
D	Kolmogorov-Smirnov Test statistic
D^*	Threshold at a given significance level α for the Kolmogorov-Smirnov Test
$f(x)$	Univariate probability density function (PDF)
$f(x_1, ..., x_n)$	n-dimensional probability density function (PDF)
f_c	Conditional probability density function (CPDF)
$F(x)$	univariate cumulative distribution function (CDF)
$F(x_1, ..., x_n)$	n-dimensional cumulative distribution function (CDF)
F_c	Conditional cumulative distribution function (CCDF)
h	Distance / Spatial lag
$\mathbf{I}$	Identity matrix
k_{wbl}	Shape parameter of Weibull distribution
k_{gam}	Shape parameter of Gamma distribution
K_v	Modified Bessel function of second kind
k_{Θ}	Number of parameters of a parametric distribution function
$\mathcal{L}$	Likelihood
n	Sample size or number of dimensions
P	Probability
p_d	Probability of no precipitation (dry)
p_w	Probability of precipitation (wet)
R_i	Empirical rank of a value x_i in its sample
r_{mat}	Range parameter of Matérn correlogram model
s_x	Empirical standard deviation of sample of X
s_x^2	Empirical variance of sample of X
u, v	CDF value / relative rank corresponding to x, y
u_{sim}, v_{sim}	Simulated CDF value
u_{th}	Threshold CDF value
w	Uniform random number for the inversion of a CCDF
x, y	Realization of a random variable X, Y
$\overline{x}$	Mean of a random variable X
x^*	A specific value of X
x_{sim}	A realization of a random variable X simulated from a CDF or CCDF
z	Standard normal variable
α	Significance level
Δ_{DOY}	expected difference of DOY that marks of the rainy season in a future period
γ_{gam}	Lower incomplete Gamma function

$\widehat{\gamma}$	parametric variogram model
$\widehat{\gamma}_{hl}$	h-lambda variogram model
$\widehat{\gamma}_{sp}$	spherical variogram model
γ_{xy}	covariance of X and Y
$\gamma*$	experimental variogram
Γ	Gamma function
$\mathbf{\Gamma}$	Correlation matrix
λ_{exp}	Parameter of exponential distribution
λ_{ce}	Parameter of exponential correlogram model
λ_{hl}	Parameter of h-lambda variogram model
λ_K	Kriging weight
λ_{wbl}	Scale parameter of Weibull distribution
μ	First statistical moment of a random variable
μ_F	Fuzzy membership function
μ_L	Lagrange multiplier
ν_{mat}	Shape parameter of Matérn correlogram model
ϕ	Gaussian PDF
$\mathbf{\Phi}$	Gaussian CDF
ρ_{sp}	Spearman Correlation coefficient
ρ_{xy}	Pearson Correlation coefficient
$\widehat{\rho}$	Parametric correlogram model
$\widehat{\rho}_{exp}$	Exponential correlogram model
$\widehat{\rho}_{mat}$	Matérn correlogram model
σ^2	Second central statistical moment of a random variable
$\mathbf{\Sigma}$	Covariance matrix
τ	Temporal lag
τ_K	Kendall's Tau
θ_{gam}	Scale parameter of Gamma distribution
Θ	Single parameter of an unspecified distribution function
$\mathbf{\Theta}$	Parameter set of an unspecified distribution function or a copula
ϑ	dry day correction threshold

Chapter 1

Introduction

1.1 Motivation

1.1.1 Regional climate and observed data

Knowledge of a region's climatology is indispensable for the management of its water bodies, agriculture, ecosystems or technical systems like urban drainage. Reliable and long time series of meteorological variables in a sufficient spatio-temporal resolution are a prerequisite to analyze the climatology and event characteristics of the region or system at hand. Furthermore, such data is required to run impact models which simulate for example the discharge in a catchment or the potential crop yield. Long time series are required so that management decisions also take extreme events or accumulated events like dry spells into account. Since the climate is projected to change for most regions of the world (*IPCC*, 2013), decisions makers are also confronted with adapting the management strategies to the uncertain future climate. Such adaptations may be the redesign of a sewage system so that the future discharge can be safely routed. In agriculture, it may be a change of crop types or sowing and harvesting dates to accommodate the projected future climatology. Decision makers often need to make expensive long-term investments to adapt the given natural or technical system to the uncertain future conditions. Depending on the case, a failure of the system may endanger human lives or cause high damages. Therefore, a sound meteorological data base is required to derive the necessary adaptation strategies.

For many technical or natural systems, precipitation is of utmost importance. On the one hand, excessive precipitation amounts can cause a sewage system or a river to overflow into vulnerable areas. Plants in a rain-fed agriculture may suffer from stagnant moisture. On the other hand, the ecosystem in a river or hydro power generation may be adversely affected by long dry spells and related low water levels. Plants can die off if too little water is available and a sewage system starts clogging if the water level and discharge is too low to flush it.

Precipitation can be extremely variable in time and space (e.g. *Kim et al.*, 2019). Torrential rain can occur quickly while a nearby location remains dry, in particular in complex mountainous terrain. In contrast, temperature is less variable in space and time. Precipitation is also a complex variable in a statistical sense because the distribution of daily and sub-daily positive precipitation amounts is generally very skewed as some rare but impactful extreme values are much higher than the average amount. Furthermore, many values are zero and therefore, a mixed discrete-continuous distribution is required.

Under optimal conditions, long time series of highly-resolved observed meteorological variables are available at all locations of interest in a study region. Long time series without data gaps are especially important in hydrology, because extreme floods cannot always be related to extreme precipitation as the antedecent conditions of the soil also controls the magnitude of floods (*Verhoest et al.*, 2010). Meteorological data have been observed at measurement stations for decades and sometimes even centuries. However, measurement stations are often sparsely distributed across a given region. Unfortunately, the number of operating measurement stations has heavily declined since 1990 for most regions of the world (*Lorenz and Kunstmann*, 2012), which poses problems for the calibration and evaluation of remote sensing and simulation techniques. An aggravating factor is that some variables were not easily observable in the past, e.g. the automated measuring of precipitation in a temporal resolution of 5 minutes. Therefore, water resource management is typically confronted with large uncertainties regarding the meteorological input variables of local impact models.

Depending on the meteorological variable and system, spatial fields of the variable are preferable to point data, for example precipitation fields in complex mountainous regions with a quickly responding river catchment. In such a region, high rainfall amounts with a high spatio-temporal variability can occur. By relating information like the reflectivity of a radar beam (*Germann et al.*, 2006) or the attenuation of the signal of commercial micro wave links (*Haese et al.*, 2017) to gauge data, remotely-sensed products like precipitation fields can be derived for a region. The quality of these estimated products depends on how well the relation between the remotely-sensed signal and the meteorological variable can be established and on local effects like mountains that block radar beams. Since observed time series are only available for historical periods, additional uncertainties how the management should be adapted to the unknown future climatology arise. For instance, an intensification of daily and sub-daily precipitation extreme events has been reported for the last decades and this trend is expected to remain for future periods (e.g. *Barbero et al.*, 2017).

1.1.2 Climate models

Physically-based climate model simulations constitute an alternative source for climate data. A multitude of physically-based models has been developed in order to provide long gridded time series of meteorological variables for past and future conditions. General Circulation Models (GCMs) simulate the mass and energy fluxes in the atmosphere in a spatial resolution of currently up to 0.25° (*Buizza et al.*, 2017). Before this recent model development, the highest horizontal resolution was about 0.5° (*Anav et al.*, 2013). GCMs can incorporate the available observed meteorological variables as initial and boundary conditions and they generate time series for the whole globe. However, decision making in water resources management often requires a higher spatio-temporal resolution than what the GCMs can provide because many meteorological and hydrological processes are highly variable in space and time.

Regional Climate Models (RCMs) use the GCM simulations as driving boundary conditions and are set up for a confined region of interest that is nested into the numerical grid of the GCM with a higher spatio-temporal resolution (*Rummukainen*, 2009). This technique is often referred to as physical or dynamical downscaling. Further scale improvements can be attained by nesting another domain with even higher spatial and temporal resolution into the first simulation domain of the RCM (e.g. *Wagner et al.*, 2012).

RCMs offer many appealing features for regional climate analyses and the application of impact models. Most notably, time series are generated on a spatial grid without any data gaps for gauged and ungauged locations. Since the model structure of GCMs and RCMs is based on the physical processes in the atmosphere and on the ground, simulations with a good setup resemble the observed spatio-temporal dependence structures and univariate distributions. By changing boundary forcing like radiative forcing or atmospheric conditions like greenhouse gas concentrations (*van Vuuren et al.*, 2011), future conditions can be simulated with a GCM which in turn provides boundary forcing data for RCM simulations in future periods. Studies that investigated the impact of a projected increase in temperature of 0.5 K demonstrated how sensitively runoff models can react to these seemingly small differences (e.g. *Paltan et al.*, 2018). Since GCM and RCM simulations describe the physical processes they can be used to study highly complex interacting systems which in turn influence a region's climate. For instance, *Caesar et al.* (2018) modeled the Atlantic Meridional Overturning Circulation (AMOC) with GCMs and RCMs for a historical period and the simulated trend of sea surface temperature agrees with the observed trend. This study demonstrates the skill of physically-based models in simulating complex, non-linear systems. An observation-based statistical model would be hardly capable of modeling such a non-linear system under non-stationary conditions. The importance of physically-based simulations becomes especially apparent when considering tipping points at which a drastic change of a system occurs. *Lenton et al.* (2008) presented a list of potential tipping elements of the Earth, for example the Indian Monsoon system or the Greenland ice-sheet. Thus, reliable RCM simulations are a necessity to lay a foundation for regional climate change adaptation strategies under climate change conditions.

In data-scarce regions, a surrogate to missing observed time series can be attained by simulating meteorological variables with dynamical downscaling techniques. *Wagner et al.* (2009) utilized dynamically simulated and remotely-sensed information to drive a hydrological model for a region in West Africa. Also, the RCM can be run for future periods to study the local climate change. For instance, in studies by *Chiew et al.* (2010) and *Chen et al.* (2012), hydrological models were calibrated with downscaled precipitation fields for historical conditions to study the change in river discharge for future scenarios.

1.1.3 Limitations of RCM simulations

Studies by *Shrestha et al.* (2006) and *Bruni et al.* (2015) underline the importance of the spatio-temporal resolution of meteorological input data for the application of hydrological models. Especially urban hydraulic models require precipitation in sub-daily and sometimes even sub-hourly resolution because the time of concentration can be very short and any failure of an urban drainage system can lead to large damages. Since the discharge in an urban drainage system is generated by water that precipitated at different locations in the catchment, a single gauge is usually not sufficient to describe the spatio-temporal variability within a catchment.

Dynamical downscaling can be used to simulate highly resolved meteorological variables. RCM simulations with a finer spatial resolution typically reproduce the statistics of the observed variables at a finer temporal resolution better (e.g. *Sunyer et al.*, 2016). The spatial and temporal resolutions of RCMs are interlinked because the time step of the numerical simulation is chosen to ensure numerical stability at a given spatial resolution. However, the computational and storage demand increases non-linearly as a finer resolution of the time steps necessitates a refinement of the 3-dimensional grid spacing of the RCM. Also, some impact studies require even higher resolved input data than what is attainable in practice. For instance, an analysis by *Ochoa-Rodriguez et al.* (2015) in different European catchments revealed that rainfall fields should have a resolution of at least 1 km and 5 minutes for a hydraulic model to perform sufficiently well in the catchments under study. The high computational and storage demand currently impedes multidecadal RCM simulations in such a high resolution.

Another problem arises for impact modelers when systematic differences between observed and simulated distribution functions of meteorological variables exist. Such a systematic error is called bias. As meteorological variables are physically linked to one another, a bias will further propagate to other variables within the RCM and in subsequent impact models. There are several error sources of RCM simulations (*Teutschbein and Seibert*, 2013). Depending on the spatio-temporal resolution, RCMs may use simplified formulations to solve small scale processes. For example, the simulation of cloud formation or turbulence is very challenging and computationally demanding. Also, the physical processes are sometimes not fully understood and not all variables are feasibly observable. Parametrization gives a plausible estimation based on empirical dependence. Therefore, some of the output variables can be regarded as an estimation and not as a direct solution of the physical equation systems. Additionally, numerical effects, computationally limited spatio-temporal resolution, uncertainties of the initial and boundary conditions and using a single parameter to represent a certain property of a grid cell like e.g. the land use or vegetation type render it impossible for an RCM to reproduce all historical observations exactly.

1.1.4 Statistical and stochastic post-processing

In order to bridge the gap between the needs of users of impact models and the currently available RCM simulations, statistical and stochastic post-processing techniques can be employed. These techniques reduce the systematic deviations of the model's univariate distributions to the observed ones, increase the spatial and temporal resolution or transform the simulated time series in such a way that a formerly problematic statistical property agrees with the observed one. In this chapter, only a short introduction to existing post-processing techniques is given. A detailed overview is presented in each chapter of this thesis to demonstrate how the newly developed methods differ from the approaches in the literature.

The bias of RCM simulations is often removed by using different statistical or stochastic bias correction techniques. A comprehensive review of bias correction methods was presented by *Maraun* (2016). As a faster alternative to dynamical downscaling, many different statistical and stochastic downscaling approaches have been developed to increase the spatial resolution of GCMs and RCMs (e.g *Maraun et al.*, 2010; *Goly et al.*, 2014). While there are many different techniques that carry

the name downscaling (including bias correction), this work utilizes the term downscaling to refer to the spatial refinement of RCM simulations based on the statistical dependence between physical models in different spatial resolutions. A huge set of downscaling and bias correction techniques was applied to European precipitation and temperature data by *Hertig et al.* (2019). This study also examines the performance of the different techniques in great detail.
If the temporal resolution of RCM simulations is too low for further applications, they can still constitute a valuable tool to provide precipitation time series for future and present conditions. The desired temporal resolution can be attained by performing a statistical or stochastic disaggregation. Disaggregation can be used for example to estimate the sequence of hourly precipitation based on the daily precipitation simulated with an RCM if the hourly simulations are not satisfactory or available. Disaggregation models are of course not limited to RCMs but can also be used to estimate temporally finer resolved rainfall intensities of observation data.

Many statistical approaches in hydrological and atmospheric science utilize correlation or covariance to describe the dependence structure of variables. Regression techniques can be used for statistical downscaling or interpolation. A potential problem of these methods is that the parameters are influenced by the univariate distributions of the variables which complicates transferring the parameters to ungauged locations. Regression leads to a single predicted value that represents the 'best estimate' but this value may no longer follow the actual distribution of that variable and the variability can be too low. Given the non-linear behavior of hydrological systems, it is important that simulated variables follow the observed distribution function and that extreme values are present. For such purposes, simulation techniques are favorable because the predicted variables are not as smooth as with regression techniques. Furthermore, they can provide an ensemble of realizations to quantify the uncertainty of the prediction.

In recent years, copulas have been employed in many scientific disciplines as a simulation technique. Copulas allow for a separation of the marginal distribution of the variables of interest from the dependence structure which makes them a flexible tool for stochastic modeling because the model components can be determined independently. An ensemble of realizations of an unknown variable can be simulated conditionally on a set of known variables. A concise introduction to copulas can be found in *Genest and Favre* (2007). Copulas have been used in hydrology to simulate spatial fields of precipitation conditional on rain gauge observations (e.g. *Bárdossy and Li*, 2008) or to simulate continuous precipitation time series (e.g. *Vernieuwe et al.*, 2015). Within the field of RCM post-processing however, copulas are still not very common. Exceptions are the studies by *Laux et al.* (2011) and *Ben Alaya et al.* (2014).

1.2 Objectives, research questions and innovations

The objective of this work is the development of computationally-efficient statistical and stochastic post-processing methods to improve the applicability of meteorological time series simulated by an RCM. The focus of the methods presented in this thesis lies on multivariate copulas of more than two dimensions. These are the Gaussian Copula and Vine Copulas (*Aas et al.*, 2009) which decompose the multivariate copula into pairs of common bivariate copulas.

To check the suitability of the chosen approaches, evaluations of the generated time series against observations were performed. Due to the importance and complexity of precipitation, it is the central meteorological variable treated in this thesis. However, the presented methods can also be utilized for other variables since they utilize distribution functions and statistical measures of dependence which can be calculated for other variables as well. Statistical downscaling and bias correction techniques are usually calibrated with data from meteorological measurement stations. Since RCMs provide time series on a spatial grid, models that are transferable to ungauged locations were required.
The research questions in this thesis are:

1. **How can multivariate copulas be utilized to increase the spatio-temporal distribution and to improve the dependence structure of RCM simulations?**

2. **How well do the stochastic simulations agree with observed univariate and multivariate statistics?**

3. **What are the advantages of the developed models compared to other approaches and what are the limiting factors for extensions and applications to other variables?**

4. **How can the model parameters be estimated for ungauged location in a study region and what are the limitations?**

Several novel approaches have been developed in this thesis. The main innovations can be summarized as follows:

- Estimation of the distribution functions of ungauged locations for bias correction and simulation purposes with a geostatistical approach. The technique has been utilized to perform a bias correction of the complete CORDEX-Africa ensemble in a data scarce region in West Africa. A similar bias correction model has been developed independently by *Mamalakis et al.* (2017) for the Italian island Sardinia. The distinguishing feature is mainly that the model presented in this thesis utilizes anisotropic variograms. Nevertheless, geostatistical approaches to estimate unknown distribution functions for Quantile-Mapping are not very common and constitute a novel approach.

- Development of a novel model structure that employs the Gaussian Copula to simulate ensembles of spatially correlated precipitation fields conditional on a coarse scale field. The model is based on an existing copula model by *Bárdossy and Li* (2008) which was developed to simulate precipitation fields from gauge measurements. Alongside the models proposed in *Thober* (2016) and *Ben Alaya et al.* (2014), it is one of the first applications of copulas to spatially correlated downscaling of climate model simulations. The model has been presented in the publication by *Lorenz et al.* (2018).

- First application of the Gaussian Copula for the disaggregation of precipitation. The model was applied to highly-resolved RCM simulations and spatially distributed precipitation time series in a temporal resolution of 5 minutes were obtained. In order to take the spatio-temporal dependence structure into account, several spatial and temporal conditioning values were employed.

- First application of Vine Copulas to improve the inter-variable dependence structure of bias corrected RCM time series of four different meteorological variables.

1.3 Outline of the thesis

This thesis describes four statistical and stochastic techniques to refine the distribution functions and spatial and temporal resolution of meteorological time series simulated with RCMs. In each chapter that describes one of these four methods, an overview of several existing approaches is given at first. Afterwards, the study region and data is presented and the need for a new statistical or stochastic post-processing technique is discussed. Then, the mathematical principles of the respective technique are shown. Afterwards, the model is applied to a data set and analyses of the simulated precipitation time series are shown.

- Chapter 1 has formulated the motivation behind the post-processing techniques and gives an overview of the thesis.

- Chapter 2 introduces the statistical principles that are necessary for the different post-processing methods.

- Chapter 3 presents a geostatistical bias correction procedure. The chosen method utilizes Kriging to estimate the distribution of a meteorological variable at unmeasured locations and transforms the RCM simulations via Double-Quantile-Mapping. The estimated distributions were compared with the observed distribution and a cross validation was performed. The RCM simulations of daily precipitation of the CORDEX-Africa ensemble have been bias corrected for historical and future periods for a study region in West Africa where the measurement network is very irregular and sparse. The onset of the rainy season and the change of monthly and annual statistics was investigated to study how the climatology is projected to change in the future.

- Chapter 4 describes a stochastic spatial downscaling method. The developed technique uses the Gaussian Copula to generate ensembles of spatially coherent precipitation fields based on coarse scale RCM precipitation simulations for the fine scale RCM domain. The method has been applied to daily RCM precipitation simulations for Central Europe in a resolution of 42 km and 7 km. Distribution functions and dependence measures were calculated and compared with the RCM simulations and performance measures were calculated to analyze the daily values simulated with the stochastic method.

- Chapter 5 is concerned with the disaggregation of precipitation simulated by an RCM for several grid cells. Disaggregation was carried out by simulating spatio-temporally correlated time series based on observed statistics with the Gaussian Copula. Hourly RCM precipitation simulations were bias corrected for a region around Freiburg, Germany. Gauge measurements in a resolution of 5 minutes were employed to build the stochastic model and to evaluate the distribution and spatio-temporal dependence structure of the disaggregated time series.

- Chapter 6 introduces an approach to simulate a meteorological variable based on its dependence to three other variables at the same location and time step. The four-dimensional dependence structure was modeled with a Vine Copula to take the different (and partially asymmetric) dependence structures between the meteorological variables into the account. The method has been utilized to post-process bias corrected hourly RCM simulations for the Berchtesgaden National Park in Germany.

- Chapter 7 gives a short overview of the developed models and addresses the research questions stated in this chapter.

The presented methods were developed to circumvent limiting aspects of RCM simulations that hinder their direct application for a given case study. Figure 1.1 presents an overview of how these methods are related and how they are motivated. Some impact models may need a finer spatial resolution than what is attainable with dynamical downscaling which motivates the stochastic downscaling (Chapter 4) to estimate fine scale precipitation fields. In other cases, the temporal resolution may be too coarse and the temporal disaggregation technique (Chapter 5) can be employed. A common problem of RCMs is the bias of the simulated variables but this bias can be defined differently. If only a single meteorological variable is required to run an impact model, a univariate bias correction (Chapter 3) is sufficient. But for more complex impact studies, the dependence structure between the different variables may need to be corrected as well. This problem can be tackled with the multivariate bias correction method (Chapter 6).

1.4 Acknowledgements and funding information

The bias correction method presented in Chapter 3 has been developed within the BMBF research program WASCAL (West African Science Service Center on Climate Change and Adapted Land Use).

The spatial downscaling method in Chapter 4 and the temporal dissaggregation technique in Chapter 5 have been developed within the BMBF project SYNOPSE (Synthetische Niederschlagszeitreihen für die optimale Planung und den Betrieb von Stadtentwässerungssystemen).

The multivariate Vine Copula model for post-processing univariately bias corrected meteorological time series has been developed within the project BIAS II which was funded by the Bayerisches Staatsministerium für Umwelt und Verbraucherschutz (StMUV).

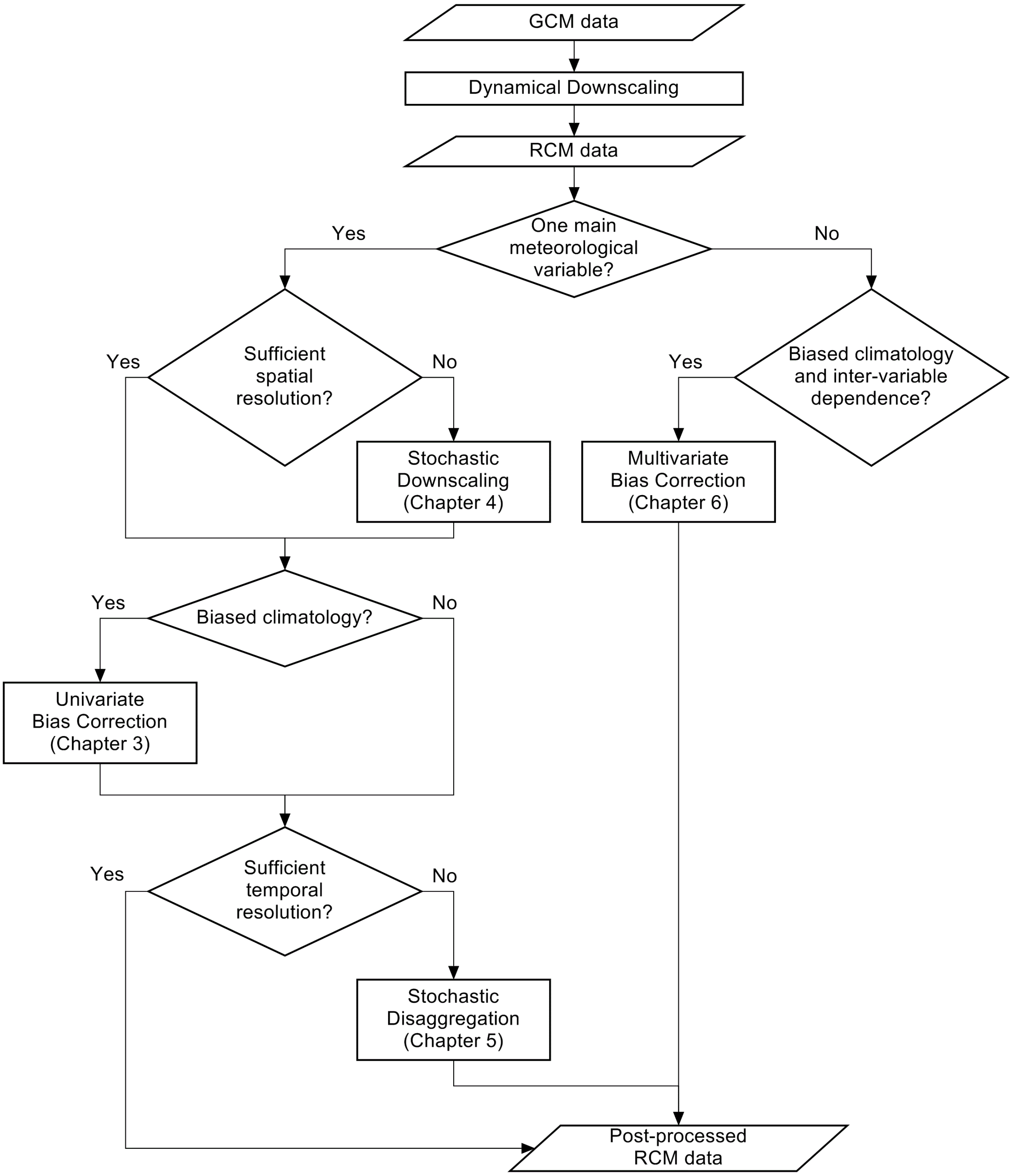
GCM data
Dynamical Downscaling
RCM data
One main meteorological variable?
Yes
No
Sufficient spatial resolution?
Yes
No
Stochastic Downscaling (Chapter 4)
Biased climatology and inter-variable dependence?
Yes
Multivariate Bias Correction (Chapter 6)
Biased climatology?
Yes
No
Univariate Bias Correction (Chapter 3)
Sufficient temporal resolution?
Yes
No
Stochastic Disaggregation (Chapter 5)
Post-processed RCM data

FIGURE 1.1: Summarizing flow chart of the four newly developed post-processing techniques and their motivation.

Chapter 2

Overview of statistical and stochastic principles

The post-processing methods presented in this thesis are based on the statistical properties of random variables. These properties are described by different measures which are required for the calibration of the stochastic models and to assess the performance of the refinement techniques. This chapter presents the statistical basics of the different methods with real world data sets to serve as a reference chapter. An extensive overview of common statistical methods that are used in atmospheric and meteorological science can be found in *Wilks* (2011). Afterwards, an introduction to copulas is given. Additional formulas that are necessary for the different developed techniques are presented in the respective chapters.

2.1 Random variables and descriptive statistics

From a statistical perspective, every measurement of a meteorological variable is a realization of a random process. To obtain information about the statistical properties of a random variable X, a sample of different realizations x_i measured at certain points in time or space i is considered. A sample of n realizations $\{x_1, ..., x_n\}$ of X is used to calculate different descriptive measures.

The mean $\overline{x}$ is the average value of a data set. If the sample size n is large enough that the sample $\{x_1, ..., x_n\}$ is representative of the random variable X, $\overline{x}$ equals the first statistical moment μ.

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{2.1}$$

The variance s_x^2 measures the spread of the random variable in relation to $\overline{x}$. Its square root s_x is the standard deviation. For large samples, s_x^2 converges towards the second central moment σ^2.

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 \tag{2.2}$$

The number of occurrences of a value x^* in a sample is called absolute frequency. The relative frequency (or probability) is calculated by dividing the absolute frequency by n. For example, the precipitation probability p_w is estimated by dividing the number of wet values $n_w = \#\{x|x > 0\}$ by n.

$$p_w = \frac{n_w}{n} \tag{2.3}$$

The number of realizations of X that are less than or equal to x^* is called absolute cumulative frequency. The corresponding non-exceedance probability $P[X \leq x^*]$ (or empirical relative cumulative frequency) is calculated by dividing by the sample size n.

$$P[X \leq x^*] = \frac{\#\{x|x \leq x^*\}}{n} \tag{2.4}$$

2.2 Distribution functions

The methods presented in the following chapters employ distribution functions which assign relative cumulative frequencies (the non-exceedance probabilities) to random variables $\{x_1, ..., x_n\}$. These functions can be empirical or parametric. To illustrate the construction of the empirical distribution and the fitting and selection of a parametric distribution function, an example data set of observed daily precipitation intensities for the city of Hamburg, Germany in the period 1993-2012 is used. The sample size is $n = 1420$ values.

2.2.1 Empirical cumulative distribution function

The histogram of the precipitation intensities (Figure 2.1) shows that most values are rather small - for example the absolute frequency of values below 0.7 $mm\ d^{-1}$ (the first bar in the histogram) amounts to 362, whereas very high values e.g. above 40 $mm\ d^{-1}$ occur rather rarely.

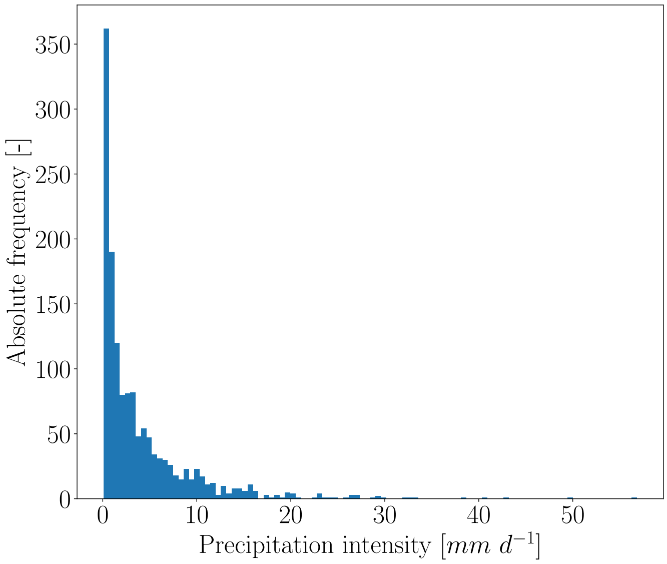

FIGURE 2.1: Histogram of daily precipitation intensities in Hamburg (1993-2012).

Ordering the precipitation intensities in ascending order and assigning the corresponding rank R_i (increasing from 1 to n) to each value x_i results in the empirical cumulative absolute frequency. The empirical cumulative distribution function (ECDF) is calculated by dividing the ranks R_i by n (Equation (2.4)). In Figure 2.2 the empirical distribution of the positive daily precipitation intensities is shown. The left y-axis corresponds to the absolute ranks R_i whereas the right y-axis are the relative ranks u_i. For instance, the rank of 11.2 $mm\ d^{-1}$ amounts to $R = 1302$. The cumulative relative frequency (the probability of non-exceedance) amounts to $u = \frac{1302}{1420} \approx 91.69\%$.

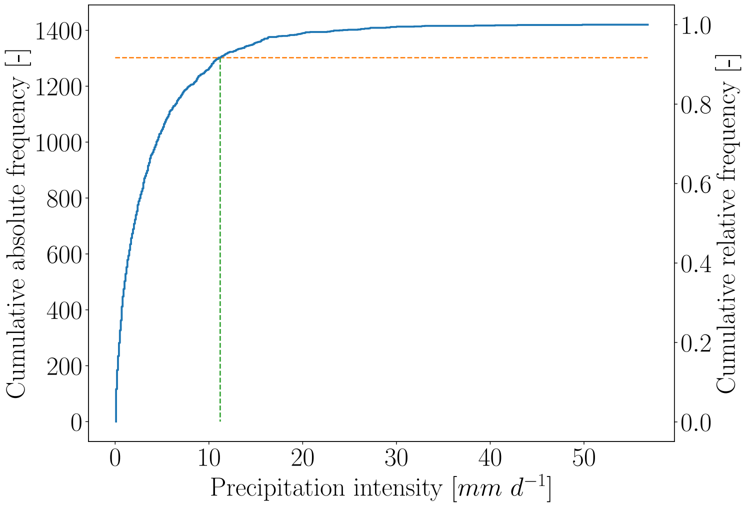

FIGURE 2.2: Empirical cumulative absolute and relative frequency of daily precipitation intensities in Hamburg (1993-2012).

2.2.2 Parametric distribution functions

Observed values are discrete because measurement devices have a resolution, e.g. 0.01 mm. For stochastic modeling, it is necessary to find an invertible function $F(x)$ defined by a parameter set $\mathbf{\Theta}$ that returns the probability of $X \leq x^*$ for arbitrary values of x^*.

$$F(x) = P[X \leq x] \tag{2.5}$$

The probability density function (PDF) $f(x)$ is the derivative of the CDF $F(x)$. The PDF is very important for the construction of parametric CDFs because fitting a distribution function to a data set is often only possible for the PDF. The CDF can then be obtained by analytical or numerical integration of the PDF.

$$f(x) = \frac{d}{dx}F(x) \tag{2.6}$$

The parametric function is fitted to a sample of the random variable with the Method of Moments (MoM) or the Maximum Likelihood Method (MLM). The MoM can be used if the distribution parameters $\mathbf{\Theta}$ can be calculated from descriptive measures of the random variable like the mean or the variance.

The MLM optimizes the function's parameters $\boldsymbol{\Theta}$ by varying their values and calculating the log-likelihood $\ln \mathcal{L}$ for each parameter set.

$$\ln \mathcal{L}(x; \boldsymbol{\Theta}) = \sum_{i=1}^{n} \ln f(x_i | \boldsymbol{\Theta}) \tag{2.7}$$

The parameter set $\boldsymbol{\Theta}$ which returns the maximum log-likelihood that the random variable X could stem from this parametric distribution function is selected.
The following section presents the most important parametric distribution functions used in this thesis.

Uniform distribution

The uniform distribution is the simplest available distribution function as it assign the same PDF value to all values on its domain $[a_{uni}, b_{uni}]$.

$$f(x) = \frac{1}{b_{uni} - a_{uni}} \tag{2.8}$$

The uniform CDF is defined as:

$$F(x) = \frac{x - a_{uni}}{b_{uni} - a_{uni}} \tag{2.9}$$

Recalling the definition of the empirical probabilities (Equation (2.4)) and setting $a_{uni} = 1/n$, $b_{uni} = 1$ and $x = R_i/n$, it can be seen that CDF values are uniformly distributed. This property will be very important later for the stochastic simulations from a distribution function and copulas.

Gaussian distribution

The Gaussian (or normal) PDF utilizes two parameters μ and σ.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \tag{2.10}$$

The MoM can be used to estimate its two parameters: the mean estimator is $\mu := \overline{x}$ and the standard deviation estimator is $\sigma := s_x$. The integration to the CDF is not straightforward and is performed numerically or by using tabulated values after the variable has been transformed to the standard normal space via $z = \frac{x-\mu}{\sigma_x}$. The standard normal PDF is denoted by ϕ and the CDF by Φ.

Exponential distribution

Another common parametric distribution function is the exponential distribution which only has one parameter λ_{exp}. In this case, the MoM can be used as well by setting $\lambda_{exp} := \frac{1}{\overline{x}}$.

$$f(x) = \lambda_{exp} e^{-\lambda_{exp} x} \tag{2.11}$$

This function can be integrated analytically to the CDF $F(x)$.

$$F(x) = 1 - e^{-\lambda_{exp} x} \tag{2.12}$$

Weibull distribution

A parametric distribution function that cannot be fitted easily via descriptive statistics is the Weibull distribution. The PDF is defined by a scale parameter λ_{wbl} and a shape parameter k_{wbl}.

$$f(x) = \frac{k_{wbl}}{\lambda_{wbl}}(\frac{x}{\lambda_{wbl}})^{k_{wbl}-1}e^{-(\frac{x}{\lambda_{wbl}})^{k_{wbl}}} \tag{2.13}$$

The corresponding CDF can be calculated analytically.

$$F(x) = 1 - e^{-(\frac{x}{\lambda_{wbl}})^{k_{wbl}}} \tag{2.14}$$

While it is possible to use the MoM for fitting the Weibull distribution (*Teimouri et al.*, 2013), it is more common to use the MLM to optimize the parameter set $\mathbf{\Theta} = \{\lambda_{wbl}, k_{wbl}\}$.

Log-normal distribution

Another parametric distribution function which is often fitted to precipitation intensities, is the log-normal distribution. It is obtained by transforming the positive precipitation amounts x to $\ln x$. Then, these transformed values are fitted with a normal distribution (Equation (2.10)).

Gamma distribution

The PDF of the Gamma distribution is defined by a shape parameter k_{gam} and a scale parameter θ_{gam}.

$$f(x) = \frac{x^{k_{gam}-1}e^{-\frac{x}{\theta_{gam}}}}{\Gamma(k_{gam})\theta_{gam}^{k_{gam}}} \tag{2.15}$$

The function Γ in the denominator is the Gamma function.

$$\Gamma(k_{gam}) = \int_0^\infty t^{k_{gam}-1}e^{-t}dt \tag{2.16}$$

The CDF is defined as:

$$F(x) = \frac{\gamma_{gam}(k_{gam}, \frac{x_{gam}}{\theta_{gam}})}{\Gamma(k_{gam})} \tag{2.17}$$

The function γ_{gam} in the numerator is the lower incomplete gamma function.

$$\gamma_{gam}(k_{gam}, \frac{x}{\theta_{gam}}) = \int_0^{\frac{x}{\theta_{gam}}} t^{k_{gam}-1}e^{-t}dt \tag{2.18}$$

Typically, the MLM is used to estimate the parameters $\mathbf{\Theta} = \{k_{gam}, \theta_{gam}\}$.

2.2.3 Choosing a parametric distribution function

Selecting a given parametric function to represent the distribution function of the random variable is based on statistical and visual tests because not every parametric function that has been fitted to a sample will be capable of reproducing the observed empirical CDF. Figure 2.3 shows the empirical CDF and the five fitted parametric CDFs.

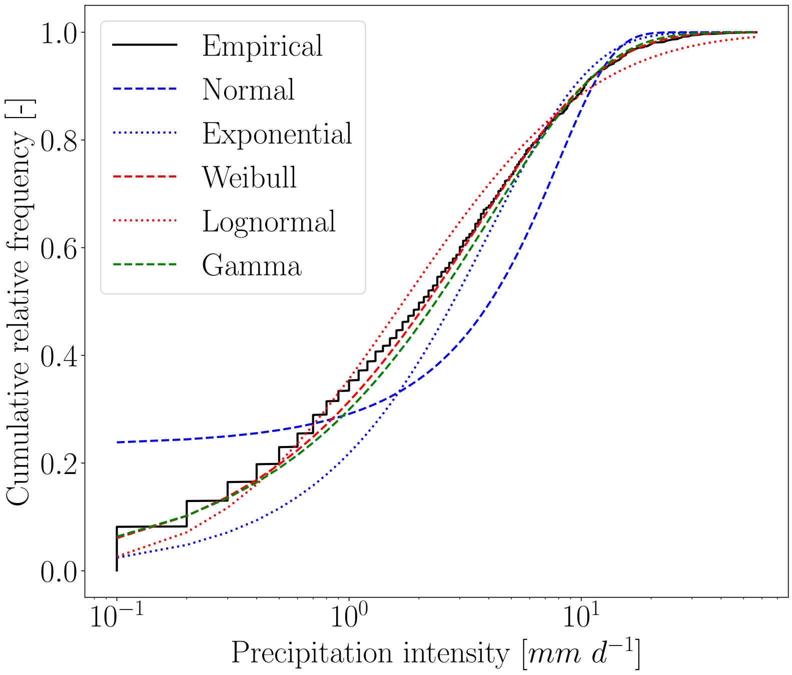

FIGURE 2.3: Parametric CDFs fitted to daily precipitation intensities in Hamburg and empirical CDF (1993-2012).

The normal and exponential distribution are rather far-off from the true distribution. For an automated and objective selection of a distribution function, statistical tests are used.

Kolmogorov-Smirnov Test

One method to check the overall performance of a parametric distribution function is the Kolmogorov-Smirnov Test (KS Test) which calculates the absolute difference $\Delta_i = |F_{emp}(x_i) - F_{par}(x_i)|$ between the parametric and the empirical CDF values for all x_i. The test statistic D is the maximum absolute difference between the two distributions.

$$D = max(\Delta_i) \tag{2.19}$$

This test statistic D is compared to a threshold D^* that depends on the significance level α of the test and on the sample size n. For a significance level of $\alpha = 5\%$, D^* becomes 0.0510 in this example. If the test statistic D is larger than the threshold D^* the hypothesis that the empirical and parametric distributions are identical is rejected. For each significance level, different test statistics are tabulated. Table 2.1 lists the D-values and whether the hypothesis of identical distributions is rejected.

Distribution	D	Hypothesis of identical distributions
Normal	0.1973	rejected
Exponential	0.1256	rejected
Weibull	0.0310	accepted
Log-normal	0.0560	rejected
Gamma	0.0458	accepted

Table 2.1: KS Test statistics of five different distribution functions.

According to the KS Test, the Weibull distribution is the best suited distribution as it has the minimum deviation D from the empirical CDF values.

Bayesian Information Criterion

The Bayesian Information Criterion (BIC, *Schwarz*, 1978) employs the likelihood $\mathcal{L}$ of a fitted distribution with a given parameter set consisting of k_Θ parameters and the sample size n.

$$BIC = -2\ln(\mathcal{L}) + k_\Theta \ln(n) \tag{2.20}$$

The higher the likelihood is, the lower the BIC gets and the better suited the respective distribution function is. The inclusion of the number of parameters penalizes distribution functions with more parameters, so that from two distribution functions with comparable likelihoods, the one with less parameters will be chosen. The reasoning behind this criterion is that models with less parameters are more robust and less prone to over-fitting.

The BIC values of the five parametric CDFs support the selection of the log-normal distribution instead of the Weibull distribution as it returned the lowest BIC (Table 2.2). The KS Test however rejected the hypothesis that the measurements stem from a log-normal distribution. Due to only having one parameter, the exponential distribution also seems like a good choice according to the BIC, even though the KS Test has rejected the hypothesis that the precipitation intensities are exponentially distributed.

Distribution	$\ln \mathcal{L}$	BIC
Normal	-4454.3	8923.1
Exponential	-3410.8	6828.8
Weibull	-3325.3	6665.2
Log-normal	-2559.7	5134.0
Gamma	-3342.8	6700.1

Table 2.2: BIC values of five different distribution functions.

This example is meant to show that the selection of a parametric CDF is not straightforward and that different criteria can lead to different choices. In the end, the choice of a parametric function should be supported by additional qualitative tests like quantile-quantile plots (QQ-Plots) which plot the sorted values of two random variables against each other.

2.2.4 Simulating from a parametric distribution function

Once a parametric distribution function has been selected, stochastic simulations can be performed. As the CDF values are uniformly distributed in $[0,1]$, it is possible to draw a set of uniform random numbers $u_{sim} \sim U(0,1)$ and invert the parametric CDF to obtain a realization x.

$$x = F^{-1}(u_{sim}) \tag{2.21}$$

In case of precipitation it is necessary to also simulate 0 mm with a dry probability of $p_d = 1 - p_w$ which is not possible with a single parametric distribution. The overall distribution is constructed as a mixed discrete-continuous distribution.

$$u_{sim}(x) = \begin{cases} p_d + p_w F(x) & \text{if } x > 0 \\ \leq p_d & \text{else} \end{cases} \tag{2.22}$$

The zero amounts obtain a censored CDF value $u_{sim} \leq p_d$. This means that u_{sim} is unknown and can take on any value between 0 and p_d. To invert this truncated distribution, the random numbers are compared with the dry probability p_d. If a random number u_{sim} is below p_d, the simulated value will become $x_{sim} = 0$. In the other case, the CDF value of the non-zero precipitation amounts is calculated as $u_w = \frac{u_{sim} - p_d}{1 - p_d}$.

$$x_{sim} = \begin{cases} F^{-1}(\frac{u_{sim} - p_d}{1 - p_d}) & \text{if } u > p_d \\ 0 & \text{if } u \leq p_d \end{cases} \tag{2.23}$$

2.2.5 Simulating from heavy-tailed distributions

For some very heavy-tailed distributions like precipitation in a temporal resolution of 5 minutes a further split might be necessary if no parametric CDF is capable of reproducing the extreme values. In such a case, the non-zero precipitation amounts of X can be split up into two parts X_{low} and X_{up} and both parts are fitted separately. This procedure requires a separation threshold x_{th} which depends on u_{th}, e.g. 0.9, to obtain individual CDFs for the values below the 90%-Quantile and the upper 10% of the measured values.

$$X_{low} = \{x | x \leq x_{th}\} \tag{2.24}$$

$$X_{up} = \{x | x > x_{th}\} - x_{th} \tag{2.25}$$

To simulate a non-zero precipitation amount x from this composite distribution, a random number $u \sim U(0,1)$ is drawn. If $u \leq u_{th}$ it will be adjusted to $u_{low} = \frac{u}{u_{th}}$, if $u > u_{th}$ it will become $u_{up} = \frac{u}{1 - u_{th}}$.

$$x = \begin{cases} F_{low}^{-1}(u_{low}) & \text{if } u \leq u_{th} \\ x_{split} + F_{up}^{-1}(u_{up}) & \text{else} \end{cases} \tag{2.26}$$

If data does not follow a single parametric distribution function, a Kernel Density Estimated distribution (KDE, *Rosenblatt*, 1956) is another alternative. This method approximates the PDF as a combination of several kernels. KDE generates a function that behaves like an empirical CDF but there are two advantages: in contrast to a discrete ECDF, the KDE-CDF is smooth and simulations from a KDE-CDF can exceed the maximum in the sample (or fall below the minimum, respectively).

2.3 Spatio-temporal dependence

Until now, only statistics of univariate random variables have been presented. However, it is also important to consider the dependence that a value has to another value in a certain distance in space or time. A review of spatio-temporal dependence modeling in hydrology can be found in *Hao and Singh* (2016).
The most common statistical measures of dependence are covariance and correlation which represent the dependence of a variable X and a variable Y with a single scalar. x_i and y_i could be precipitation measurements at two different stations separated by a distance of h or at the same station but with a temporal lag τ between the measurements.
The covariance γ_{xy} is a measure that describes the linear dependence of two variables X and Y.

$$\gamma_{xy} = \frac{\sum\limits_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{n-1} \tag{2.27}$$

Dividing the covariance by the product of the standard deviations of X and Y results in the Pearson correlation coefficient ρ_{xy}. A value of 1 indicates perfect linear dependence, -1 the opposite and 0 means that the two variables are linearly independent.

$$\rho_{xy} = \frac{\gamma_{xy}}{s_x s_y} \tag{2.28}$$

To overcome the problem of the linearity assumption in Pearson's ρ_{xy}, the rank-based statistic Spearman's ρ_{sp} can be calculated. The calculation requires a transformation of the random variable to its ranks. Then, a calculation of the correlation coefficient is done with the ranks. This transformation makes it possible to detect non-linear dependence in a data set.
Another measure of non-linear dependence is Kendall's τ_K which counts the number of concordant and discordant pairs. A pair consisting of (x_i, y_i) and (x_j, y_j) is concordant if both values are increasing or both are decreasing.

$$\tau_K = \frac{|x_i > x_j, y_i > y_j| + |x_i < x_j, y_i < y_j| - |x_i > x_j, y_i < y_j| - |x_i < x_j, y_i > y_j|}{\frac{n(n-1)}{2}} \tag{2.29}$$

A detailed example with a small data set of these two rank-based statistics can be found in (*Genest and Favre*, 2007).

An estimation of the expected correlation of censored values separated by a spatial (or temporal) distance of h can be obtained as the average correlation of all n_p CDF value pairs u and v that are separated by approximately h. A simple approach to treat the censored values is setting u to $0.5p_d$ for dry values (*Bárdossy and Pegram*, 2012). The CDF values are transformed to the standard normal space via Φ^{-1} to obtain the empirical correlation ρ^* as a function of the separation distance. Fitting a parametric correlogram model $\widehat{\rho}$ enables estimating the correlation for arbitrary distances.

$$\rho^*(h) = \frac{1}{n_p} \sum_{i=1}^{n_p} Corr\{\Phi^{-1}(u_i), \Phi^{-1}(v_i)\} \tag{2.30}$$

A more complex method utilizes the Maximum Likelihood Method (*Durban and Glasbey*, 2001). Pairs are treated differently depending on whether one or more of the values is censored, e.g. 0 mm of precipitation. The correlation coefficient ρ of pairs separated by a certain separation lag is calculated by maximizing the combined likelihood of three sets I_1, I_2 and I_3. Those three sets have individual likelihoods:

- I_1 - Both values are uncensored [> 0 mm]:

$$L_1 = \frac{1}{2\pi\sqrt{1-\rho^2}} e^{\frac{-1}{2(1-\rho^2)}\{\Phi^{-1}(u)^2+\Phi^{-1}(v)^2-2\rho\Phi^{-1}(u)\Phi^{-1}(v)\}} \tag{2.31}$$

- I_2 - u is censored [0 mm] and v is uncensored [> 0 mm]:

$$L_2 = \phi(\Phi^{-1}(p_d))\Phi(\frac{\Phi^{-1}(p_d) - \rho\Phi^{-1}(v)}{\sqrt{1-\rho^2}}) \tag{2.32}$$

- I_3 - Both values are censored [0 mm]:

$$L_3 = \int\limits_{-\infty}^{\Phi^{-1}(p_d)} \int\limits_{-\infty}^{\Phi^{-1}(p_d)} \frac{1}{2\pi\sqrt{1-\rho^2}} e^{\frac{1}{2(1-\rho^2)}(x_1^2-2\rho x_1 x_2+x_2^2)} dxdy \tag{2.33}$$

The correlation coefficient ρ of different temporal or spatial lags is then calculated from a parametric correlogram function $\widehat{\rho}$ whose parameters are optimized by maximizing the combined log-likelihood:

$$\ln \mathcal{L} = \sum \ln(L_1) + \sum \ln(L_2) + \sum \ln(L_3) \tag{2.34}$$

The optimization is performed on the parameters of the correlogram function and not on the correlation coefficient of individual lags. Otherwise, the optimization might return a non-valid correlation function as discussed in *Allard and Bourotte* (2014). *Pfaff* (2013) used this approach to estimate the parameters of a variogram function.

The parametric correlogram models that were used in this thesis to maximize the combined likelihood are the exponential and the Matérn model. These models estimate the correlation in time (auto-correlation function, ACF) or space (cross-correlation function, CCF) based on the separation lag h. In the case of an ACF, h is a temporal lag, e.g. 10 minutes. In the case of a CCF, h is the distance in space, e.g. 10 km. When both ACFs and CCFs are employed, h is used for spatial lags and τ for temporal lags to make the equations clearer.

The exponential model is the simplest correlogram model. It utilizes a single parameter λ_{ce} to describe how the correlation decreases over an increasing lag h.

$$\widehat{\rho_{exp}}(h) = e^{-\lambda_{ce}h} \tag{2.35}$$

The Matérn model (e.g. *Minasny and McBratney*, 2005) is very flexible but also rather complex. It has two parameters v_{mat} and r_{mat} and utilizes the Gamma function Γ and the modified Bessel function of the second kind K_v:

$$\widehat{\rho_{mat}}(h) = \frac{1}{\Gamma(v_{mat})2^{v_{mat}-1}}(\frac{2h\sqrt{v_{mat}}}{r_{mat}})^{v_{mat}} K_v(\frac{2h\sqrt{v_{mat}}}{r_{mat}}) \tag{2.36}$$

2.4 Multivariate distribution functions

Distribution functions can also be defined for more than one variable in which case the function is called multivariate:

$$F(x,y) = P[X \leq x, Y \leq y] \tag{2.37}$$

The multivariate CDF is not always attainable in a closed form through analytical integration of the density function $f(x,y)$ which necessitates numerical integration. One such multivariate distribution is the multivariate normal distribution with covariance matrix $\mathbf{\Sigma}$:

$$f(x_1,...,x_n) = \frac{1}{\sqrt{(2\pi)^n det(\mathbf{\Sigma})}} e^{\frac{-1}{2}(\mathbf{x}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)} \tag{2.38}$$

Alternatively, this PDF can be expressed with the correlation matrix $\mathbf{\Gamma}$ and the identity matrix $\mathbf{I}$:

$$f(x_1,...,x_n) = \frac{1}{\sqrt{(2\pi)^n det(\mathbf{\Gamma})}} e^{\frac{-1}{2}(\mathbf{x}-\mu)^T(\mathbf{\Gamma}-\mathbf{I})^{-1}(\mathbf{x}-\mu)} \tag{2.39}$$

The assumption behind the multivariate normal distribution is that the random variables $X_1,...,X_n$ are normally distributed. If this is not the case they are usually normalized with an appropriate transformation like the Box-Cox transformation.

2.5 Copulas

An alternative to using a multivariate distribution function which requires the univariate distribution of the separate random variables to stem from a certain univariate distribution is using copulas. Copulas offer two main advantages over common multivariate distribution techniques. First of all, the distribution functions of the variables are calculated separately for each variable and therefore meteorological variables that follow very different distributions can be modeled. The second advantage is that the dependence structure is modeled separately from the univariate distributions. The available copula models can describe positive or negative dependence and asymmetric dependence, for example a stronger dependence of high values. The first publication on copulas was presented by *Sklar* (1959). A detailed introduction to copulas can be found in *Nelsen* (2006).

A copula C expresses the multivariate distribution F of a set of values $(x_1,...,x_n)$ by their respective CDF values $(u_1,...,u_n)$. The univariate CDFs $F(x) = u$ are also called marginal distributions.

$$F(x_1,...,x_n) = C(u_1,...,u_n) \tag{2.40}$$

The copula density c can be calculated by partial derivation:

$$c(u_1,...,u_n) = \frac{\partial^n C(u_1,...,u_n)}{\partial u_1 ... \partial u_n} \tag{2.41}$$

For the bivariate case ($n = 2$), several parametric copula functions exist. These copulas are defined through a parameter Θ that is related to Kendall's τ_K. One of the key features of these copulas is that the conditional distribution F_c of a CDF value V given a CDF value u can be calculated analytically:

$$F_c(v|u) = Pr[V \leq v|U = u] = \frac{\partial C}{\partial u} = w \tag{2.42}$$

A realization of V can be obtained by drawing a random number $w \sim U(0, 1)$ and inverting this conditional distribution function.

$$v_{sim} = F_c^{-1}(w) \tag{2.43}$$

Finally, a transformation of the simulated CDF value v_{sim} to x_{sim} is achieved with the inverse of its univariate distribution function and a realization of the random variable X is obtained:

$$x_{sim} = F^{-1}(v_{sim}) \tag{2.44}$$

Table 2.3 lists some of the most commonly used bivariate copulas and their corresponding densities and conditional distributions. The equations were compiled from different sources, mainly *Nelsen* (2006) and *Trivedi and Zimmer* (2013). Also the relation of Kendall's τ_K to the copula parameter Θ and the range of values that Θ can take on is given. For the copula families Gumbel, Clayton, and Farlie-Gumbel-Morgenstern, a direct calculation of Θ from τ_K is possible, whereas the Frank and Ali-Mikhail-Haq copulas require a numerical procedure to calculate Θ. The calculation of the Frank Copula's parameter Θ requires a Debye function of the first kind: $D_1(x) = \frac{1}{x} \int_0^x \frac{t}{e^t - 1} dt$.

Copula family	Ali-Mikhail-Haq (AMH)
$C(u,v)$	$\frac{uv}{1-\Theta(1-u)(1-v)}$
$c(u,v)$	$\frac{1-\Theta+\frac{2\Theta uv}{1-\Theta(1-u)(1-v)}}{(1-\Theta(1-u)(1-v))^2}$
$F_c(v\mid u)$	$\frac{v-\Theta u+\Theta v^2}{(1-\Theta(1-u)(1-v))^2}$
$\tau_K = f(\Theta)$	$\frac{3\Theta-2}{\Theta} - \frac{2}{3}(1-\frac{1}{\Theta})^2 \ln(1-\Theta)$
Range of Θ	$\in [-1, 1[$

Copula family	Clayton (Cla)
$C(u,v)$	$(max(u^{-\Theta} + v^{-\Theta} - 1; 0))^{-1/\Theta}$
$c(u,v)$	$(1+\Theta)(uv)^{-1-\Theta}(u^{-\Theta} + v^{-\Theta} - 1)^{-\frac{1}{\Theta}-2}$
$F_c(v\mid u)$	$u^{-\Theta-1}(u^{-\Theta} + v^{-\Theta} - 1)^{-\frac{1}{\Theta}-1}$
$\tau_K = f(\Theta)$	$\frac{\Theta}{2+\Theta}$
Range of Θ	$\in [-1, \infty[\backslash\{0\}$

Copula family	Farlie-Gumbel-Morgenstern (FGM)
$C(u,v)$	$uv(1+\Theta(1-u)(1-v))$
$c(u,v)$	$1+\Theta(1-2u)(1-2v)$
$F_c(v\mid u)$	$v+\Theta-\Theta v^2-2\Theta uv+2\Theta uv^2$
$\tau_K = f(\Theta)$	$\frac{2}{9}\Theta$
Range of Θ	$\in [-1, 1]$

Copula family	Frank (Fra)
$C(u,v)$	$-\frac{1}{\Theta}\ln(1+\frac{(e^{-\Theta u}-1)(e^{-\Theta v}-1)}{e^{-\Theta}-1})$
$c(u,v)$	$\frac{\Theta(1-e^{-\Theta})e^{-\Theta(u+v)}}{((1-e^{-\Theta})-(1-e^{-\Theta u})(1-e^{-\Theta v}))^2}$
$F_c(v\mid u)$	$\frac{(e^{-\Theta v}-1)e^{-\Theta u}}{e^{-\Theta}-1+(e^{-\Theta u}-1)(e^{-\Theta v}-1)}$
$\tau_K = f(\Theta)$	$1+\frac{4}{\Theta}(D_1(\Theta)-1)$
Range of Θ	$\in \mathbb{R}\backslash\{0\}$
Copula family	Gumbel[-Hougaard] (Gum)
$C(u,v)$	$exp(-((-\ln(u))^{\Theta}+(-\ln(v))^{\Theta})^{1/\Theta})$
$c(u,v)$	$(\Theta-1-\ln C(u,v))exp\{\ln(C(u,v))+((\Theta-1)\ln(-\ln(u))-\ln(u))$
$F_c(v\mid u)$	$\frac{(-\ln(u))^{\Theta-1}((-\ln(u))^{\Theta}+(-\ln(v))^{\Theta})^{1/\Theta-1}}{uexp\{((-\ln(u))^{\Theta}+(-\ln(v))^{\Theta})^{1/\Theta}\}}$
$\tau_K = f(\Theta)$	$\frac{\Theta-1}{\Theta}$
Range of Θ	$\in [1, \infty[$

Table 2.3: Bivariate copulas, copula densities and conditional distribution functions.

Empirical and parametric copula densities are illustrated in Figure 2.4 for CDF value pairs of observed hourly air temperature (u) and shortwave downwelling radiation (v) in the autumn season SON in the Berchtesgaden National Park. More details on this data is given in Chapter 6.

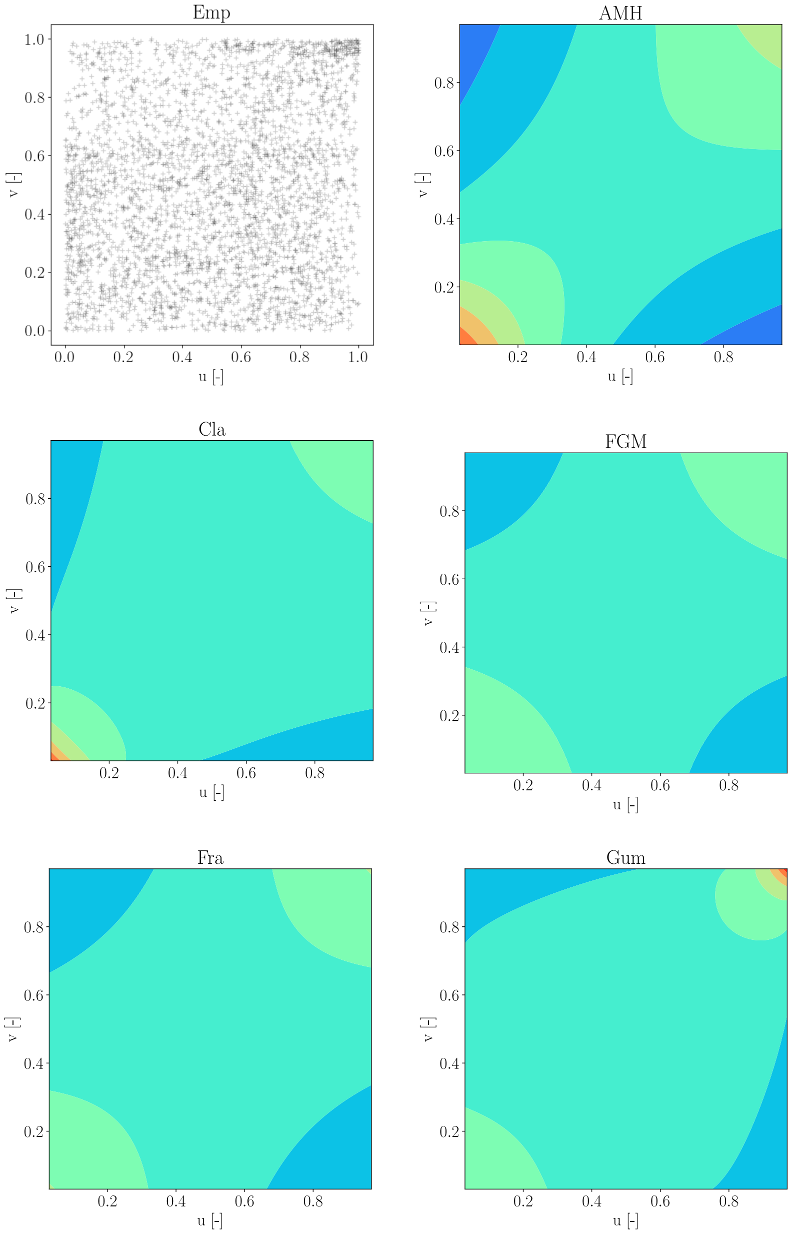

FIGURE 2.4: Empirical scatter plot of CDF values (Emp) and parametric copula densities (AMH, Cla, FGM, Fra, Gum) of hourly air temperature (u) and shortwave downwelling radiation (v) in the Berchtesgaden National Park in the season SON (2001-2010).

The empirical scatter plot of CDF values (Figure 2.4, Emp) shows asymmetric dependence: High values exhibit a stronger dependence than low values. The five parametric copula densities were calculated from τ_K. As was the case for univariate distribution functions, a selection of the best fitting copula model is necessary. To this end, the squared differences between the empirical and parametric copulas are calculated and the copula model which has the smallest differences is selected. In this example, the Gumbel copula was chosen as the best fit as it can model the observed asymmetric dependence.

The majority of copulas are bivariate. An extension to higher dimensions (n>2) is possible for the Frank and Clayton Copula (*Fischer et al.*, 2009). But those copulas are defined by a single parameter Θ which constricts their application to cases where the dependence structure between all variable pairs does not differ by a lot.
Vine copulas (*Aas et al.*, 2009) allow the decomposition of a multivariate distribution into pair copulas but the model structure is very complex. The Gaussian Copula is comparatively simple and has been used in other studies to simulate daily precipitation (*Bárdossy and Li*, 2008). This copula offers the advantage that it can deal with any number of conditioning values which makes it possible to simulate precipitation amounts conditionally on multiple spatial or temporal neighbors. The Gaussian Copula is a special form of the multivariate Gaussian distribution. Instead of the vector of CDF-values $\mathbf{u} = (u_1, ..., u_n)$, CDF-values that have been transformed to the standard normal space via the inverse of the univariate Gaussian CDF Φ are used. Another necessary transformation makes use of the univariate Gaussian PDF ϕ. The density of the Gaussian Copula is defined as:

$$c(u_1, ..., u_n) = \frac{1}{\prod_{i=1}^{n} \phi(\Phi^{-1}(u_i))} \frac{1}{(2\Pi)^{\frac{n}{2}} \sqrt{det(\mathbf{\Gamma})}} e^{-0.5(\Phi^{-1}(\mathbf{u})^T(\mathbf{\Gamma}^{-1}-\mathbf{I})\Phi^{-1}(\mathbf{u}))} \tag{2.45}$$

The term $\mathbf{\Gamma}$ is the correlation matrix of the data set and $\mathbf{I}$ is an identity matrix of the same size. In practice, $\mathbf{\Gamma}$ is estimated via a parametric function (a correlogram model) of the separation lag of points in either space or time. The two methods of calculating the correlation of censored data is given in section 2.3.
With a correlogram model, it is possible to estimate the correlation coefficients of a set of points in space or time and set up the correlation matrix. With the correlation matrix defined, conditional simulations can be performed. If a set of values $(u_2, ..., u_n)$ is given, the conditional PDF f_c (CPDF) of the unknown value u_1 is:

$$f_c(u_1|u_2, ..., u_n) = \frac{c(u_1, ..., u_n)}{c(u_2, ..., u_n)} \tag{2.46}$$

Because the denominator $c(u_2, ..., u_n)$ (the copula density of the conditioning values) is constant, it can be removed from the equation which results in the simplified conditional density function. To simulate u_1, this conditional density function is integrated numerically to the conditional CDF $F_c(u_1|u_2, ..., u_n)$ (CCDF). Afterwards, it is normed so that the codomain of F_c lies in $[0, 1]$. Then, F_c can be inverted with a random number $w \sim U(0, 1)$ to obtain a realization of u_1:

$$u_1 = F_c^{-1}(w), w \sim U(0, 1). \tag{2.47}$$

Chapter 3

Geostatistical bias correction of RCM precipitation

In this chapter, a newly developed geostatistical bias correction method for RCM precipitation and its application to a sparsely gauged catchment in West Africa is presented. In this region, many people rely on rain-fed subsistence agriculture. In order to adapt the crop cultivation to climate change, reliable precipitation time series for present and future conditions for the complete region are required. A community-based RCM ensemble has been created within the CORDEX project. While this ensemble provides a gridded data set of meteorological time series for present and future time periods, significant biases are present. Therefore, a bias correction is necessary, so that the observed climatology is reproduced by the historical simulations. Correcting the bias of RCM simulations requires observed statistics for every RCM grid cell to set up the transfer function to transform the RCM simulations to the observed distribution but this distribution is unknown for ungauged locations. The unknown local climatology was estimated by Kriging distribution parameters of observed data to ungauged sites to serve as a surrogate for the missing local information.

3.1 Overview of bias correction methods

Bias correction is applied to climate model simulations to reduce systematic differences to observed data. To this end, a transfer function is constructed which transforms a meteorological variable simulated by an RCM to a bias corrected value. The available bias correction techniques differ in how the transfer function is built. A review of different bias correction techniques, their possible short-comings and extensions and an introduction to the historical origins of bias correction of numerical weather forecasts is given by *Maraun* (2016). An application of the two statistical bias correction methods *Delta Change Approach* and Histogram Equalization to the temperature and precipitation simulations of the RCM COSMO-CLM is given in *Berg et al.* (2012).

Other studies presented copula-based bias correction schemes (*Laux et al.*, 2011; *Mao et al.*, 2015) that measure the dependence of observed and simulated precipitation at the same time step and generate an ensemble of values from the conditional distribution. Most bias correction techniques are deterministic and aim at the correction of the distribution of simulated intensities however and only a single bias corrected value is returned.

A technique to transform the simulated variables so that the spatial correlation of the observed variables is reproduced has been introduced by *Bárdossy and Pegram*

(2012). In some case studies, only a single output variable of the RCM is of interest, for example the daily temperature or precipitation. If several variables need to be bias corrected, the correction is mostly performed individually for each variable. Recently, methods that aim at a multivariate bias correction have been proposed (*Piani and Haerter*, 2012; *Cannon*, 2016; *Vrac*, 2018). These methods can strongly change the statistics of the meteorological variables since they introduce a further transformation of the RCM data. There is a debate about the applicability of bias correction in general because the variables of the uncorrected RCM are physically consistent. After the bias correction, this may no longer be the case and higher aggregated variables can exhibit a stronger bias than before the correction (*Ehret et al.*, 2012). In practice however, bias correction is still widely applied since a biased meteorological input variable is regarded as very detrimental to subsequent impact models.

3.1.1 Delta Change Approach

The easiest technique to bias correct the RCM simulations is the *Delta Change Approach*. This technique uses the average deviation of the RCM simulations from the observed values and corrects the RCM simulations via a single number Δ. This number is either added to the simulated value x_{sim} or multiplied with it to obtain a bias corrected value x_{BC}. In both cases the mean of the RCM simulations $\overline{x}_{sim}$ and the mean of the observations $\overline{x}_{obs}$ are required to calculate Δ.
The additive correction is normally used for variables like temperature. For instance, if the RCM's daily temperature is on average colder by 2 K than the observations, 2 K are added to the RCM values.

$$x_{BC} = x_{sim} - \Delta, \Delta = \overline{x}_{sim} - \overline{x}_{obs} \tag{3.1}$$

In the case of precipitation, it is common to calculate the ratio of the average daily precipitation in the RCM to the observation and divide all RCM values by this factor. So if the RCM overestimates the observed mean by 5%, each RCM value is divided by 1.05.

$$x_{BC} = \frac{x_{sim}}{\Delta}, \Delta = \frac{\overline{x}_{sim}}{\overline{x}_{obs}} \tag{3.2}$$

While the *Delta Change Approach* is easily applicable, it has the disadvantage of not being able to correct the higher moments of the variables. Another problem of this approach is its bad performance if the climatic conditions of the validation period differs from those of the calibration period (*Teutschbein and Seibert*, 2013).

3.1.2 Dry Day Correction

For precipitation, a correction of the frequency of wet values is necessary in most cases because RCM simulations typically exhibit more time steps with precipitation than is observed. *Frei et al.* (2003) analyzed the precipitation probability of RCM simulations and gridded observation data for several regions in Europe and found an overestimation of the precipitation probability for many months and RCMs.
One cause for an excessive precipitation probability is the drizzle effect. RCMs often simulate too many low intensity precipitation events when compared to observations (e.g. *Sun et al.*, 2006). The probability that a grid cell is wet is also scale-dependent. *Argüeso et al.* (2013) showed that simulations with the Weather Research and Forecasting Model (WRF, *Skamarock and Klemp* (2008)) in a spatial resolution of 10 km

have a much higher precipitation probability than the 2 km simulations for the same domain. This behavior is to be expected. If one were to construct a gridded data set from rain gauges, larger areas would also be more likely to receive precipitation than smaller areas because one gauge with a non-zero measurement would cause the grid cell to be wet. In practice, it is however often the case that there is a mismatch between the spatial resolution of the RCM and the observations: if precipitation has been measured by a gauge, X_{obs} is a variable corresponding to a single point in space and therefore a difference between $p_{w,obs}$ and $p_{w,sim}$ is to be expected. Furthermore, rain gauges may miss very light precipitation amounts that are below the detection limit. Nevertheless, gauge data is commonly used to perform bias correction - either because there is no other data available or because the bias corrected time series are desired to behave like gauge data for further applications.

An overestimated precipitation probability can be corrected by setting all values below a chosen threshold ϑ (e.g. 1.0 $mm\ d^{-1}$) to zero. This threshold should be calculated individually for each cell so that the frequency of values above the threshold is equal to the observed precipitation probability. This procedure requires calculating the observed precipitation probability $p_{w,obs}$ first. Then, only the $n_{sim}p_{w,obs}$ largest values of the RCM simulations will be considered as actual precipitation. The threshold ϑ is thus the value that fulfills:

$$n_{sim}p_{w,obs} = \#\{x_{sim}|x_{sim} \geq \vartheta\} \tag{3.3}$$

After the threshold has been found, all values below it are set to zero and the remaining values are shifted towards zero to allow for the fitting of a parametric distribution function. This approach has been used amongst others in *Volosciuk et al.* (2017) and *Lafon et al.* (2012). A correction for the rare converse case, that the wet day probability is higher in the observations than in the simulations, was developed by *Themeßl et al.* (2010) who infilled very low intensities until the observed wet day probability was matched.

3.1.3 Quantile Mapping

A more complex way to correct the bias than the *Delta Change Approach* is *Quantile Mapping* which is closely related to Histogram Equalization and Local Intensity Scaling. The distribution of the observed variable is reproduced by inverting the CDF of the observed variable F_{obs} with the CDF value of the RCM simulations $F_{sim}(x_{sim})$. Thus, the general characteristics of the RCM time series, such as when the highest values occur, remain the same but each value is mapped to its corresponding observed quantile.

$$x_{BC} = F_{obs}^{-1}\{F_{sim}(x_{sim})\} \tag{3.4}$$

F_{obs} and F_{sim} can either be empirical or parametric. One problem of choosing an empirical CDF is that the observed maximum cannot be exceeded. Also, the time series of observations should be as long as the RCM time series which can be circumvented by interpolating between the two values with a given rank or by sampling from the observation set until it is as large as the simulation set (*Piani and Haerter*, 2012). Parametric CDFs are capable of generating values larger than the observed maximum and the discrete nature of the measurements (e.g. a resolution of 0.1 mm of the measurement device) is less apparent in the bias corrected time series. Finding a function that fits the skewed precipitation intensities simulated by RCMs can be challenging, as discussed in *Gudmundsson et al.* (2012).

For climate change studies, the traditional *Quantile Mapping* cannot be used directly. Fitting a distribution function F_{sim} to the future period and inverting the observed distribution F_{obs} with the CDF values in the future, would result in a bias corrected time series with a distribution function that is identical to the one of the observations. The only difference would be how the large and small values tend to cluster in space and time in the different time periods.

3.1.4 Double Quantile Mapping for future periods

As the distribution of precipitation intensities is typically not the same for RCM simulations for future periods, bias correcting these future simulations should consider the climate change signal of the RCM. The *Double Quantile Mapping* method presented by *Bárdossy and Pegram* (2011) utilizes the historical CDF to calculate the CDF values of the future period. A parametric distribution function $F_{sim,hist}$ is fitted to the historical RCM time series and the CDF values of the future period are calculated with this CDF. This way, a change of the intensity distribution leads to a bias corrected time series whose distribution is no longer identical to the observed one.

$$x_{BC} = F^{-1}_{obs}(F_{sim,hist}(x_{sim,fut})) \tag{3.5}$$

The *Double Quantile Mapping* is illustrated in Figure 3.1 with artificial data. Since the CDFs of the historical and future RCM differ, the precipitation amount $x_{sim,fut}$ which is the 90%-Quantile in the future period attains a larger CDF value $F_{sim,hist}(x_{sim,fut}) = 0.95$.

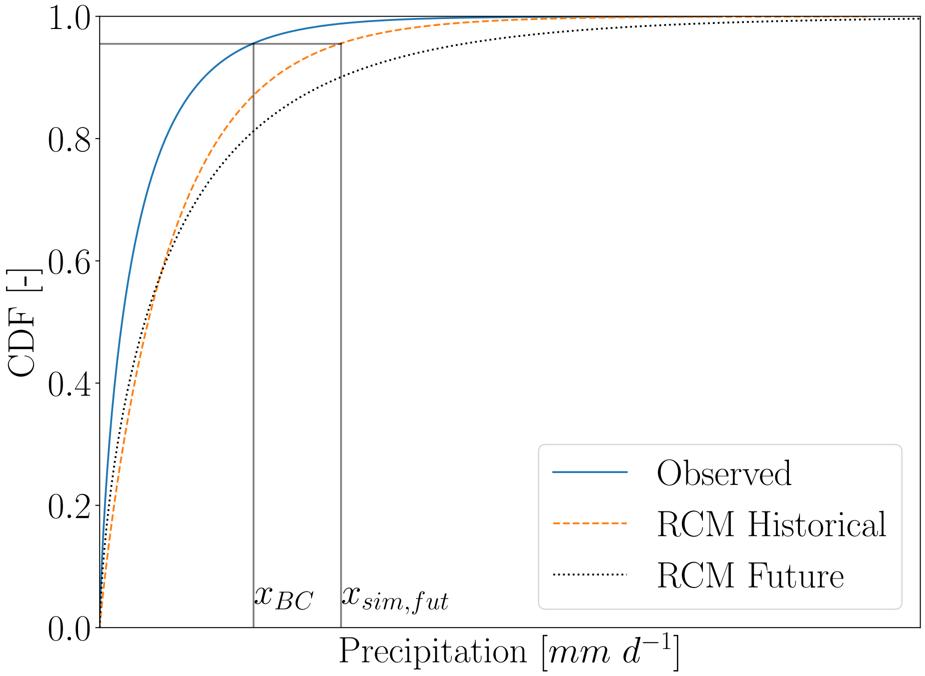

FIGURE 3.1: Transformation of non-zero precipitation amounts with *Double Quantile Mapping*.

3.2 Study region and data

In the following, a study region in West Africa is presented. The CORDEX-Africa RCM ensemble provides gridded time series of several meteorological variables for this region for historical and future periods but a bias to the observed climatology exists. The study region and observation data is introduced in subsection 3.2.1. The CORDEX-Africa ensemble and its bias is presented in subsection 3.2.2.

3.2.1 Study region and observation data

The study region consists of 173 grid cells with a spatial resolution of 0.44° in the states of Burkina Faso, Ghana, Ivory Coast, Benin, Togo, Mali and Niger. The spatial grid stems from the CORDEX-Africa RCM ensemble. From a merged data set of precipitation observations that has been collected within the BMBF research program WASCAL, 172 stations in the proximity of the study region have been extracted for the period 1950 to 2005 (Figure 3.2).

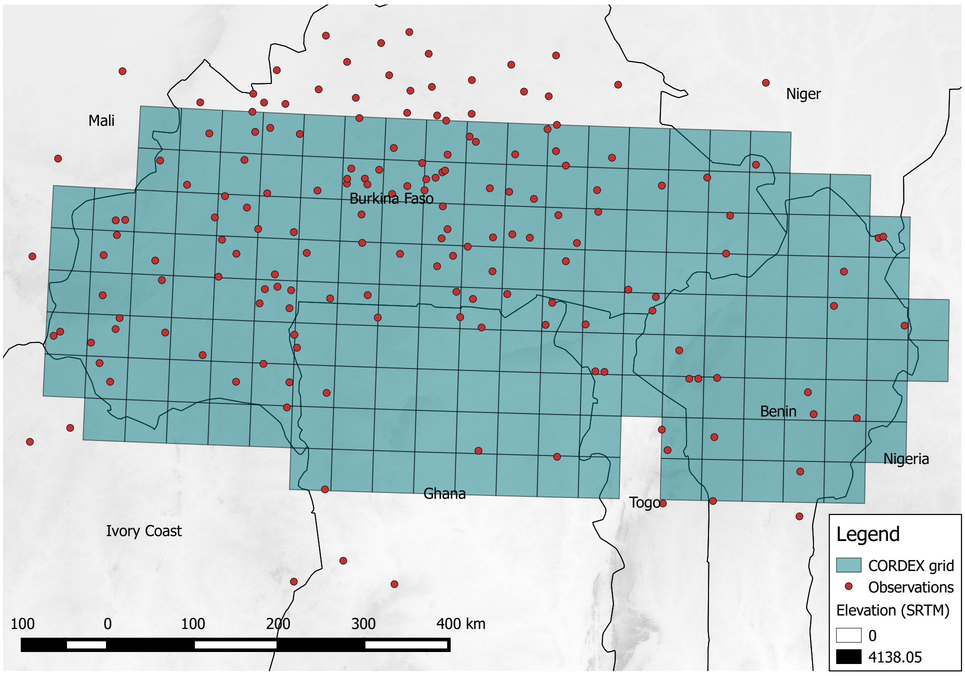

FIGURE 3.2: The study region in West Africa and the location of the observation stations.

The mean annual sum of precipitation (Figure 3.3) decreases from South (up to 1418 $mm\,a^{-1}$) to North (as low as 439 $mm\,a^{-1}$). Stations on the same degree of latitude have relatively similar annual sums.

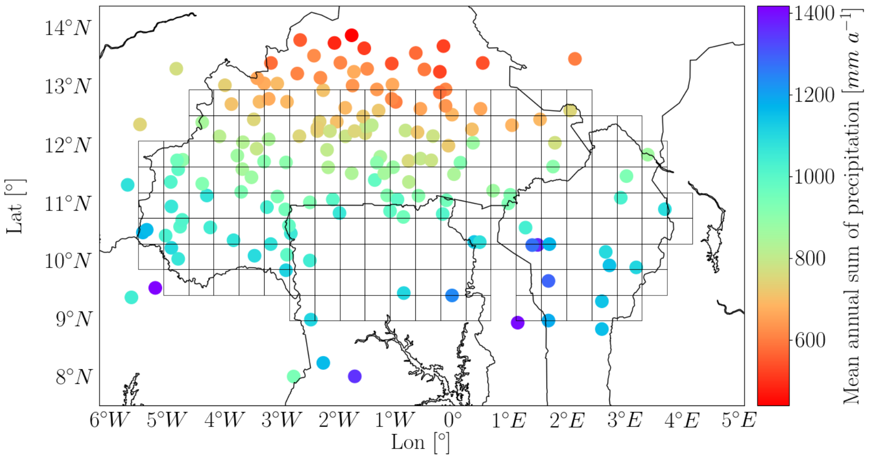
14°N
13°N
12°N
11°N
10°N
9°N
8°N
Lat [°]
6°W 5°W 4°W 3°W 2°W 1°W 0° 1°E 2°E 3°E 4°E 5°E
Lon [°]
1400
1200
1000
800
600
Mean annual sum of precipitation [mm a⁻¹]

FIGURE 3.3: Mean annual sum of precipitation of the West African observation data (1950-2005).

The rainy season is very distinct in West Africa and dominated by the West African Monsoon. From November to February, the Intertropical Convergence Zone (ITCZ) intersects the study region and dry air flows from the Sahara towards the south-west. This dry season is known as the Harmattan. In March, the ICTZ shifts towards the North and wetter air masses are transported towards the north-east. The seasonality of monthly precipitation is illustrated in the Hovmöller diagram in Figure 3.4. From February, the southern locations already receive precipitation. Over the course of the rainy season, the monthly amounts increase until August. At this time, the maxima occur at the southern border of Burkina Faso. From September, the precipitation amounts decrease quickly.

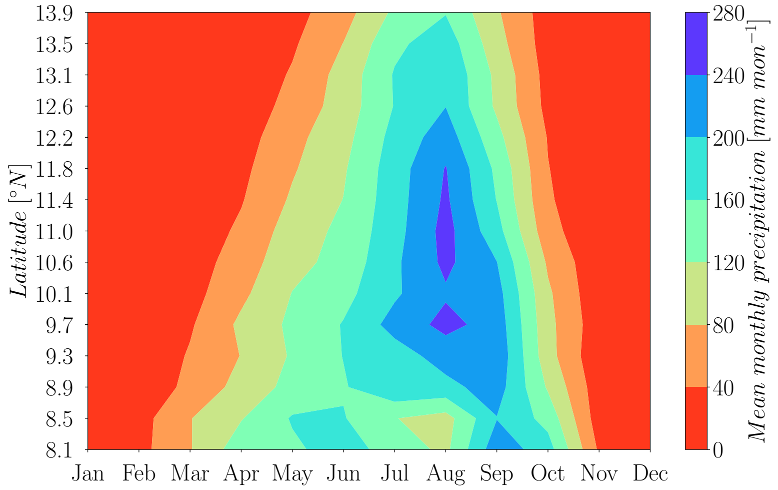

FIGURE 3.4: Hovmöller diagram of mean monthly precipitation of the West African observation data (1950-2005).

In Figure 3.5, the mean daily precipitation intensity on rainy days $\overline{x_w}$ in the season June-August is shown. As Figure 3.4 showed, the mean monthly precipitation occurs in the center of the study region in this season. The mean can vary on comparatively small scales which may be related to data gaps and decadal variability in the data set. But in general, stations on the same degree of latitude have more similar means whereas the mean varies more strongly from North to South.

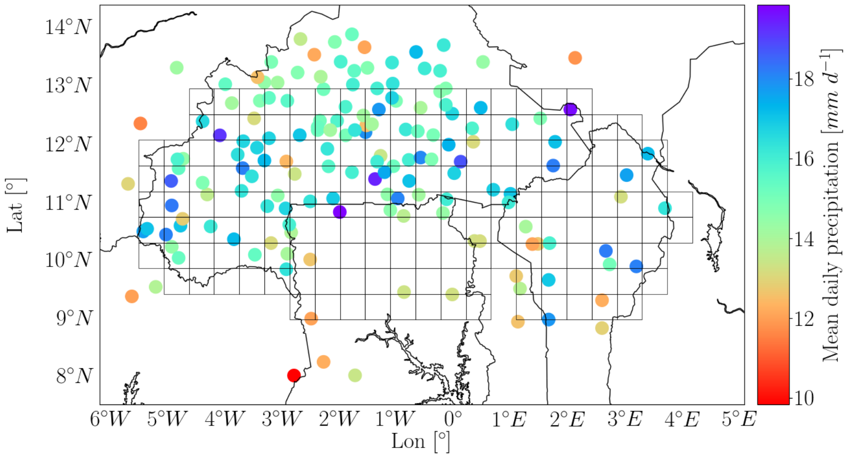
14°N
13°N
12°N
11°N
10°N
9°N
8°N
Lat [°]
6°W 5°W 4°W 3°W 2°W 1°W 0° 1°E 2°E 3°E 4°E 5°E
Lon [°]
18
16
14
12
10
Mean daily precipitation [mm d⁻¹]

FIGURE 3.5: Mean daily precipitation on rainy days $\overline{x_w}$ in the season JJA (1950-2005).

The mean daily probability of precipitation p_w (Figure 3.6) shows similar characteristics but the highest values lie still in the South. This means that the higher monthly sums in the North are related to very strong events, whereas more comparatively low daily precipitation amounts fall in the South.

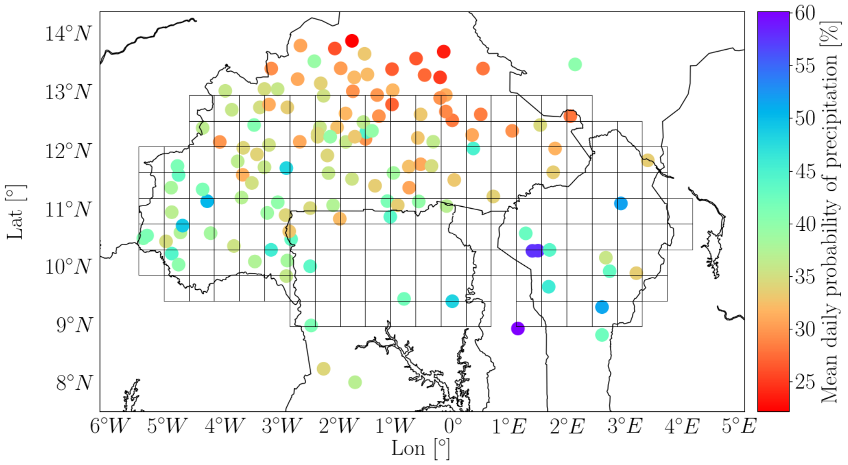
14°N
13°N
12°N
11°N
10°N
9°N
8°N
Lat [°]
6°W 5°W 4°W 3°W 2°W 1°W 0° 1°E 2°E 3°E 4°E 5°E
Lon [°]
60
55
50
45
40
35
30
25
Mean daily probability of precipitation [%]

FIGURE 3.6: Mean daily probability of precipitation p_w in the season JJA (1950-2005).

Some regions, especially Northern Ghana, have only a few measurement stations. This spatial arrangement of the measurement stations and the anisotropy of the presented statistics motivated a geostatistical approach to estimate the distribution functions F_{obs} for ungauged locations.

3.2.2 The CORDEX-Africa RCM ensemble

An ensemble of daily RCM precipitation simulations in a spatial resolution of 0.44° has been provided by the the CORDEX-Africa project (Coordinated Regional Climate Downscaling Experiment) for a historical control period (1950-2005) and future scenarios (2006-2100). The ensemble consists of 23 different model combinations. The advantage of an ensemble of RCMs is that it provides a set of scenarios which are utilized to run an impact model. The impact studies does not rely on a single simulated time series and thus the uncertainty can be quantified. A bias in simulated precipitation is especially concerning for agricultural planning in a region with a distinct rainy season like West Africa because the sowing depends on the onset of the rainy season.

For the future period, four different Representative Concentration Pathways (RCPs) have been defined as possible scenarios to accommodate the uncertainties of how anthropogenic green house gas concentrations and stratospheric adjusted radiative forcing will change (*van Vuuren et al.*, 2011). These scenarios were also adopted by the Intergovernmental Panel on Climate Change (IPCC). Within the CORDEX-Africa project, three scenarios have been used (RCP 2.6, RCP 4.5 and RCP 8.5). RCP 2.6 has been used in only five cases, while RCP 4.5 and RCP 8.5 have been used by almost all participating institutes. A list of the available simulations, time period and how leap years are treated can be found in Table A.1 and Table A.2 in Appendix A. Because some impact models require Gregorian calendar data, the simulated precipitation time series have been stretched to the Gregorian calendar by selecting the intensity of the closest day of the year. Afterwards, the intensities were rescaled such that the annual sum is consistent with the original model. There is no unique solution to this issue - one possibility is to linearly interpolate the intensities to the Gregorian calendar (*Hempel et al.*, 2013) but this approach was not used because it can severely cut the extreme values.

Bias of the CORDEX-Africa precipitation simulations reported in other studies

An inter-comparison of ten CORDEX-Africa RCMs driven by ERA-Interim reanalysis data by *Nikulin et al.* (2012) has shown that all models exhibit a significant bias in the rainy season JAS - in both positive as well as negative direction - for West Africa when compared with data from the Global Precipitation Climatology Project (GPCP) . Remarkably, ERA-Interim shows a dry bias for West Africa in this season, but the RCMs driven by ERA-Interim can lead to positive and negative biases. Some of the models simulated the onset of the rainy season too early, some have problems regarding the northward extension of the monsoon rain belt.
Mascaro et al. (2015) reported similar findings for the Niger River basin. They compared the annual precipitation of 18 GCM-RCM combinations of the CORDEX-Africa ensemble for the historical period with data from the Climatic Research Unit (CRU). In the most western sub basin close to the source of the Niger River, an underestimation of as low as -60% was found, while most models overestimated this statistic for the three following sub basins (up to $+60\%$).

The bias of RCM precipitation is not only determined by the chosen RCM but different parametrization schemes can lead to highly varying precipitation simulations with the same RCM. *Klein et al.* (2015) investigated the influence of parametrization schemes on the resulting precipitation fields with a single RCM. To this end, the WRF model was set up for West Africa in a spatial resolution of 24 km and ERA-Interim reanalysis data was used to drive the model. 27 combinations of parametrization schemes were utilized and it was found that the absolute bias in mean daily precipitation in the rainy season can be higher than 2 *mm* d^{-1} (both negative and positive). This variability illustrates the importance of the choice of parametrization schemes and explains the large uncertainties of ensemble simulations which comprise different RCMs with different driving models and not only a single RCM with different schemes.

Bias of the CORDEX-Africa precipitation simulations in the study region

The majority of the models from the CORDEX-Africa ensemble overestimate the annual sums of precipitation in the presented study region. Figure 3.7 is a scatter plot of the annual sums of the control run for the historical period (1950-2005) averaged over all GCM-RCM combinations against the annual sum of the closest measurement station.

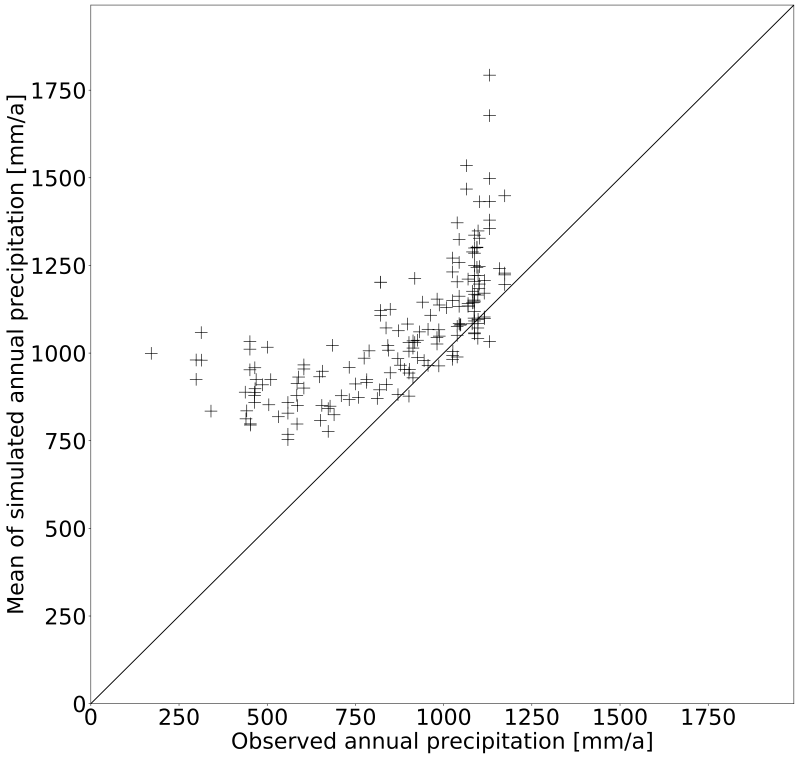

Figure 3.7: Mean of mean annual sum of simulated precipitation of 173 grid cells against annual sum of closest measurement station - uncorrected historical period (1950-2005).

Figure 3.7 illustrates that the mean annual sum of precipitation is overestimated by the ensemble for nearly every location. Single GCM-RCM combinations exhibit a much stronger bias: Annual sums of more than 2500 *mm* have been simulated by individual ensemble members. Such an overestimation poses huge problems for subsequent crop models because it may be assumed that more water is available than in reality. For the historical period this bias can be detected by comparing the simulations with observed data. For the future period, this bias is more critical because the future climatology is unknown. Assuming that the general climatology is matched by a model and that the projected climate is un-biased, wrong planting decisions could be made.

Not only then annual sums, but also the mean monthly sums of precipitation are biased in most models. Figure 3.8 shows the spread of the ensemble and its mean. The mean of the nearest observation stations is also included for reference purposes.

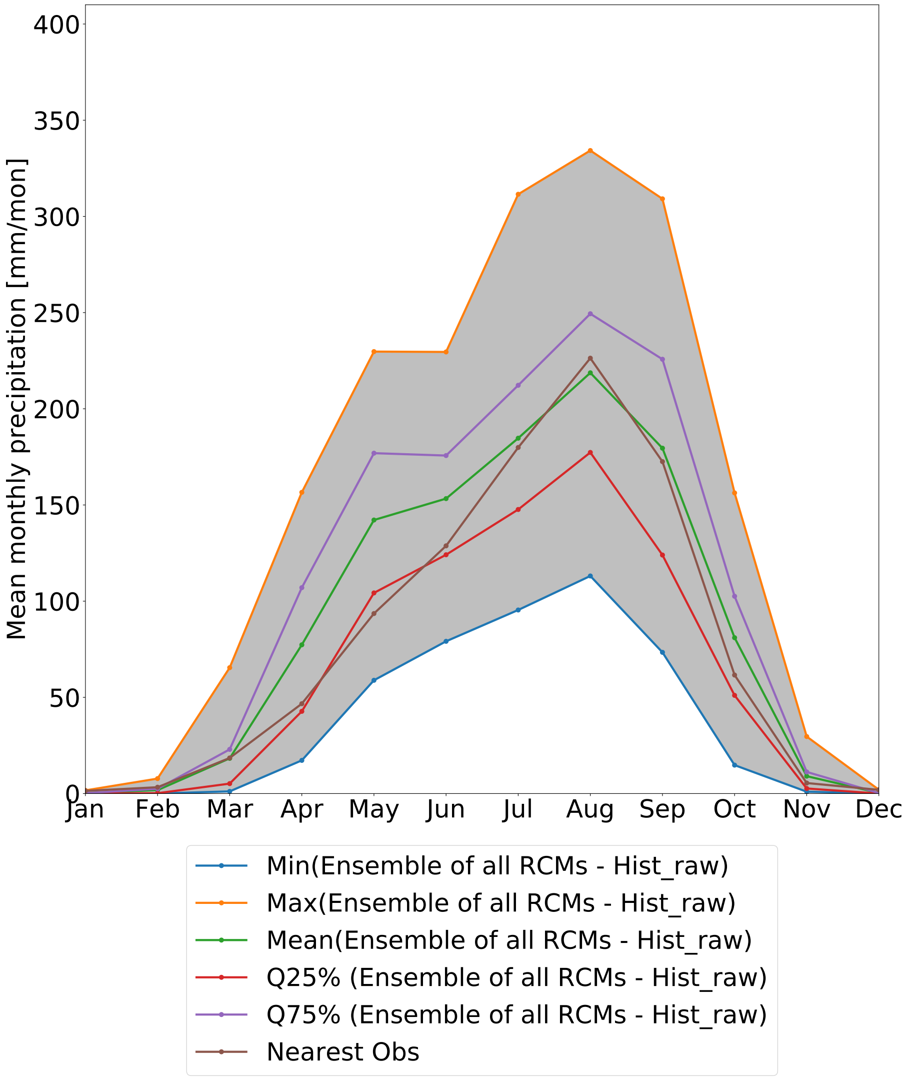

FIGURE 3.8: Mean monthly sum of precipitation averaged over 173 grid cells - uncorrected historical period (1950-2005).

As planting dates may be planned based on the onset of the rainy season, this bias calls for a correction scheme in order to avoid bad decisions like planting too early because an RCM indicates that a certain ratio of the annual precipitation has already fallen. While the mean over the 23 GCM-RCMs is closer to the observed mean monthly sums, there is still a pronounced overestimation for the months of April to June. Using the mean of an ensemble as an alternative to bias correction is also critical because doing so would not reproduce the distribution of the daily intensities and the precipitation probability.

The spread of the mean monthly sums for the uncorrected RCP 8.5 scenario looks quite similar to the historical period shown in Figure 3.8. To illustrate the climate change signal in the future period (2005-2100), the differences of the monthly sums were calculated (Figure 3.9). In general, a slight intensification of the rainfall amounts during the rainy season is projected but individual models can project highly increasing or decreasing monthly amounts. The bias correction should make use of this projected climate change signal but also reproduce the climatology of the historical period.

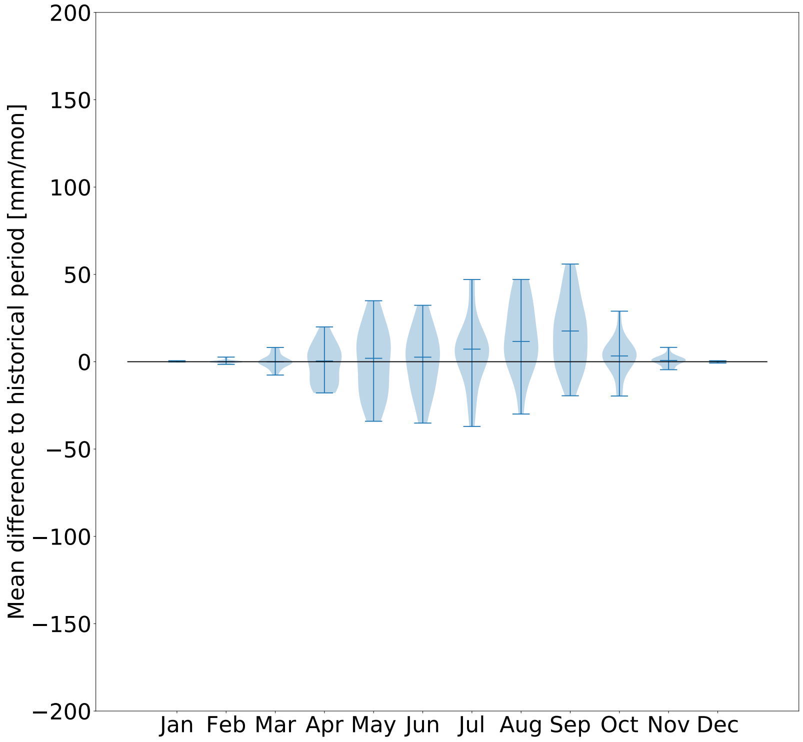

FIGURE 3.9: Difference of mean monthly sums of precipitation in the uncorrected RCP 8.5 scenario (2005-2100) to the uncorrected historical period (1950-2005).

Daily precipitation intensities also exhibit a bias but this is treated in more detail in section 3.5.4 since the bias correction was performed on daily precipitation.

While the future climate is unknown, not all model combinations should be regarded as equally suited to simulate future conditions. A model which is able to reproduce the observed climatology of the past can be assumed to be more capable of simulating reliable projections for the future. In order to evaluate the performance of the different models, an analysis of the uncorrected mean monthly sums was carried out. Before the analysis, it was intended to also investigate how well the time series of the simulations agree with the observed series. This approach was dismissed because none of the models was able to match the time series of annual sums very well. The rank correlation of the annual time series in the control period 1950-2005 ranges from −0.06 to 0.13 for all models which indicates that no model is significantly better than the others at reproducing the observed temporal structure of annual precipitation. This analysis therefore only investigates how well average monthly statistics are reproduced.

For each model and month, the average mean error ME_{avg} of the monthly precipitation amounts was calculated as a weighted average. The average mean error was calculated as the weighted average over all cells n_c with the weight w_i of each cell i depending of the distance between the cell center and its nearest station. This weighting was introduced to give a higher importance to locations with data observed close to the cell centers.

$$ME_{avg} = \sum_{i}^{n_c} w_i(\overline{x_{obs,i}} - \overline{x_{sim,i}}); \sum_{i} w_i = 1 \tag{3.6}$$

Afterwards, the average mean errors of each month and model were evaluated with a trapezoidal fuzzy rule to give a value of 1 to the model with the best performance (i.e. the lowest absolute ME) and 0 to the model with the highest absolute ME. The overall annual rating score of each model was then calculated as the weighted average of the fuzzy membership functions' value with monthly weights of $w_{mon} = \frac{\overline{x}_{obs,imon}}{\overline{x}_{obs,ann}}$ to give a higher importance to the months belonging to the rainy season. The results of this analysis can be found in Table 3.1.

Model combination	Rating [-]	Model combination	Rating [-]
CCLM4 CNRM-CM5	0.386	UQAM-CRCM5 MPI-ESM	0.663
UQAM-CRCM5 CanESM2	0.413	CCLM4 MPI-ESM	0.690
SMHI-RCA4 CSIRO-Mk3	0.458	SMHI-RCA4 HadGEM2	0.713
HIRHAM5 EC-EARTH	0.466	CCLM4 HadGEM2	0.714
HIRHAM5 NorESM1	0.473	SMHI-RCA4 EC-EARTH	0.724
KNMI-RACMO22T HadGEM2	0.529	SMHI-RCA4 ESM2M	0.751
SMHI-RCA4 IPSL-CM5A-MR	0.545	SMHI-RCA4 MIROC5	0.756
REMO2009 MPI-ESM	0.548	SMHI-RCA4 MPI-ESM	0.760
SMHI-RCA4 NorESM1	0.549	CCLM4 EC-EARTH	0.760
SMHI-RCA4 CanESM2	0.554	CanRCM4 CanESM2	0.761
KNMI-RACMO22T EC-EARTH	0.579	SMHI-RCA4 CNRM-CM5	0.856
REMO2009 EC-EARTH	0.613		

TABLE 3.1: Annual rating of all CORDEX-Africa models.

3.3 Development of a new geostatistical bias correction model for ungauged locations

RCM precipitation is usually afflicted by bias and a correction technique is required for further impact studies. This was shown to be the case for the CORDEX-Africa simulations in the study region. The bias correction techniques presented in section 3.1 build the transfer function by utilizing observed data. For instance, the observed mean $\overline{x}_{obs}$ in the *Delta Change Approach* or the CDF of the observed variable F_{obs} in *Quantile Mapping*. Due to the irregular measurement station network, it was necessary to construct the transfer function of the bias correction for ungauged sites. Utilizing the closest measurement station for each grid cell would lead to high uncertainties for grid cells that have no nearby stations. For instance, there is no observed data for several grid cells in Ghana and Benin (Figure 3.2). Assigning the closest measurement station would also lead to jumps in the statistics of the assigned rainfall gauges for neighboring cells as one cell might have a Northern station as the closest neighbor while an adjacent cell might have a Southern station as its closest neighbor. This would be especially problematic in regions with a highly-variable local climatology as in mountainous regions or in this case West Africa where the climatology depends strongly on the degree of latitude (Figure 3.4). Therefore, the method can also be used in other regions or for the estimation of CDF parameters in copula-based models.

The *Double Quantile Mapping* method presented in subsection 3.1.4 has been chosen to bias correct the CORDEX-Africa precipitation ensemble because regular *Quantile Mapping* has shown to outperform simpler methods like the *Delta Change Approach* in other studies and because the full distribution including the frequency of wet values is bias corrected. *Gudmundsson et al.* (2012) found that empirical *Quantile Mapping* resulted in the best bias corrected simulations but this approach requires complete observation time series for every location in the simulation period which are not available. The unknown point scale distribution functions F_{obs} (Equation (3.5)) for each grid cell in the study region are estimated by interpolating the parameters of the observed distribution functions to all grid cell centers in the region of interest with Ordinary Kriging. The estimated distribution F_{obs} was utilized to generate so called "simulated observations" (see Chapter 2, Subsection 2.2.4 for details on simulating from a parametric CDF). A *Double Quantile Mapping* with *Dry Day Correction* was then carried out for each RCM cell with the estimated CDFs F_{obs}. The complete process is illustrated in Figure 3.10.

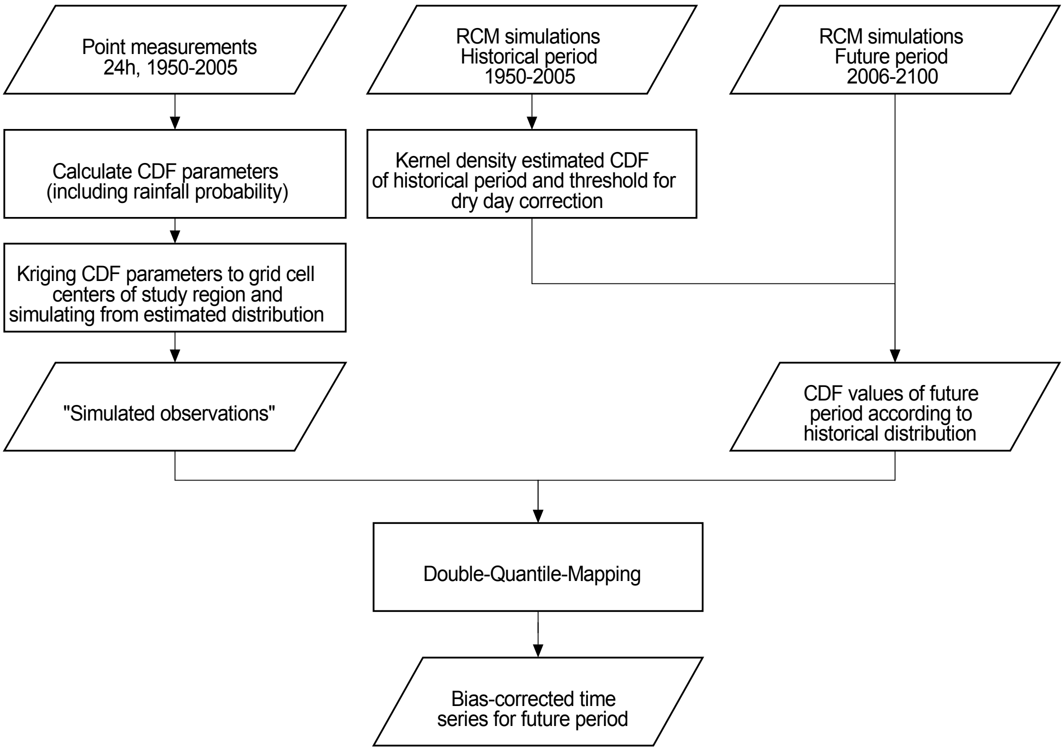

FIGURE 3.10: Flowchart of the geostatistical bias correction method.

For the sake of simplicity, the Kriging method is explained for a single CDF parameter denoted by Θ. If several parameters are required to estimate the distribution function at unmeasured locations, the following procedure has to be done for each parameter individually. *Mamalakis et al.* (2017) kriged the parameters of a Generalized Pareto distribution to ungauged locations, whereas *Mosthaf and Bárdossy* (2017) kriged precipitation quantiles to estimate unknown distributions.

It has been discussed that the estimation of the variogram and Kriging with non-normally distributed random variables can be problematic (e.g. *Cressie and Hawkins*, 1980) and to this end a transformation of non-normal variables like precipitation is often performed (e.g. *Erdin et al.*, 2012). If the parameter Θ is or can be transformed such that it is normally-distributed, the experimental variogram is calculated for different separation distances h_i. If n locations with observed data are present, an $n \times n$-distance matrix is calculated that contains the separation distance between all location combinations (k, l). The number of locations that are approximately separated by a distance of h_i is denoted by N_i and will be used to calculate the average semi-variance γ^* (the experimental variogram) of values at the given separation distance.

$$\gamma^*(h_i) = \frac{1}{2N_i} \sum_{h_{k,l} \approx h_i}^{N_i} (\Theta_k - \Theta_l)^2 \tag{3.7}$$

Afterwards a parametric function $\widehat{\gamma}$ needs to be fitted to estimate the variogram values for any separation distance. Two parametric variogram models, the h-lambda model and the spherical model, were utilized. The h-lambda model $\widehat{\gamma}_{hl}$ estimates the semi-variance with a single parameter λ_{hl}:

$$\widehat{\gamma}_{hl}(h) = h^{\lambda_{hl}} \tag{3.8}$$

The spherical model $\widehat{\gamma}_{sp}$ has two parameters, the sill C_{vsp} and the range a_{vsp}, and is of the form:

$$\widehat{\gamma}_{sp}(h) = \begin{cases} C_{vsp}(1.5\frac{h}{a_{vsp}} - 0.5\frac{h^3}{a_{vsp}^3}) & \text{if } h \leq a_{vsp} \\ C_{vsp} & \text{else} \end{cases} \tag{3.9}$$

Additionally, the nugget parameter C_0 was used to allow for the estimated semi-variance to increase rapidly for non-zero separation distances h. Depending on which model fits better, the variogram model $\widehat{\gamma}$ becomes either $C_0 + \widehat{\gamma}_{hl}$ or $C_0 + \widehat{\gamma}_{sp}$. Anisotropic variograms can be calculated as a linear combination of the directional variograms $\widehat{\gamma}_x$ and $\widehat{\gamma}_y$, so if the variogram value increases more strongly in one direction, the expected overall semivariance $\widehat{\gamma}$ between the unknown point and the observation location will be higher if the two points are mainly separated along this axis.

$$\widehat{\gamma}(h) = \widehat{\gamma}_x(h_x) + \widehat{\gamma}_y(h_y) \tag{3.10}$$

For instance, in West Africa precipitation pairs on the same degree of latitude have a lower semi-variance than pairs on the same degree of longitude because $\widehat{\gamma}_x(h) < \widehat{\gamma}_y(h)$. As Kriging minimizes the estimation uncertainty, a higher Kriging weight is given to neighboring gauges that are on the same degree of latitude.

Once the variogram model is selected, an estimate of the CDF parameters and the precipitation probability can be obtained for each unmeasured location m by Kriging the parameters of the neighboring measurement stations. Kriging estimates the unknown CDF parameter Θ^* as a linear combination of the observed parameters Θ_i which each have a weight $\lambda_{K,i}$.

$$\Theta^* = \sum_{i=1}^{n} \lambda_{K,i}\Theta_i, \sum_{i=1}^{n} \lambda_{K,i} = 1 \tag{3.11}$$

The Kriging weights λ_K are obtained by solving the Kriging equation system that minimizes the estimation uncertainty of the unknown value. The distance of the measurement locations to the unknown point is denoted as h_{im}, h_{ij} is the distance between two locations i and j with observed values and μ_L is the Lagrange multiplier.

$$\begin{bmatrix} \widehat{\gamma}(h_{11}) & \cdots & \widehat{\gamma}(h_{1n}) & 1 \\ \vdots & \ddots & \vdots & 1 \\ \widehat{\gamma}(h_{n1}) & \cdots & \widehat{\gamma}(h_{nn}) & 1 \\ 1 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_{K,1} \\ \vdots \\ \lambda_{K,n} \\ \mu_L \end{bmatrix} = \begin{bmatrix} \widehat{\gamma}(h_{1m}) \\ \vdots \\ \widehat{\gamma}(h_{nm}) \\ 1 \end{bmatrix} \tag{3.12}$$

To avoid negative weights $\lambda_{K,i}$, the adjustment procedure proposed by *Deutsch* (2005) was implemented.

With the interpolated parameter Θ, the parametric CDF F_{obs} (Equation (3.5)) is defined and a *Double Quantile Mapping* can be performed for the ungauged locations.

The statistical properties of the point measurements are transferred to the ungauged locations and a surrogate for the unknown distribution F_{obs} is provided. Alternatively to these point statistics, a Block Kriging could be performed. The choice of the Kriging method is governed by the impact studies that will be performed with the bias corrected time series. If plot scale models are used, the bias corrected time series should behave like station data. Mapping block scale simulations to local scale distributions is sometimes referred to as downscaling because this form of bias correction transfers the simulated precipitation values to the distribution of the gauge scale (e.g. *Chen et al.*, 2013).

3.4 Calibration of the geostatistical bias correction model

The bias correction model depicted in Figure 3.10 requires observed CDFs F_{obs} for the ungauged locations and CDFs $F_{sim,hist}$ fitted to the dry day corrected RCM precipitation time series in the historical period 1950-2005. The historical period stems from the CORDEX-Africa ensemble and was also chosen for the observed data.

The estimation of the CDF F_{obs} of daily precipitation is based on Kriging the parameters of the observed CDF to the ungauged locations. The required steps to calibrate the model are:

- Selection of a parametric distribution function (subsection 3.4.1)
- Testing the normality of the distribution parameters (subsection 3.4.2)
- Calculating variograms of the distribution parameters (subsection 3.4.3)

The precipitation probability p_w was also interpolated to the ungauged locations to calculate the *Dry day Correction* threshold ϑ for the historical period. The following subsections illustrate the required steps for the month of August which is the month with the highest measured precipitation amounts.

The CDF $F_{sim,hist}$ was estimated for each cell directly with KDE. This is discussed in subsection 3.4.4.

3.4.1 Parametric distribution function of observed precipitation

An analysis of the BIC values (see Chapter 2, section 2.2.3) of nine distribution functions that are implemented in the programming language MATLAB was performed in order to select an appropriate distribution function. The nine distribution functions are: Weibull (Wbl), Gamma (Gam), Exponential (Exp), Generalized Extreme Value (GEV), Generalized Pareto (Gp), Log-Normal (Logn), Logistic, Log-Logistic and Rayleigh. A boxplot of the BIC values of the respective distributions is given for the month of August in Figure 3.11. The spread of the BIC value corresponds to the gauges with enough data to fit a distribution function. The BIC was chosen as an evaluation criterion, because it accounts for the goodness of fit of a parametric distribution function via the likelihood of the fitted distribution. Furthermore, the BIC favors distribution functions with few parameters, as a low number of parameters results in a low BIC value. A parameter parsimonious distribution function is favorable for scarcely gauged regions, which is discussed in more detail later in this section. According to the BIC-values, there is no parametric distribution function that clearly outperforms all other functions as the differences are rather small. While the BIC values of some other distributions are slightly lower, the exponential distribution was chosen to model F_{obs} for the observed daily precipitation intensities.

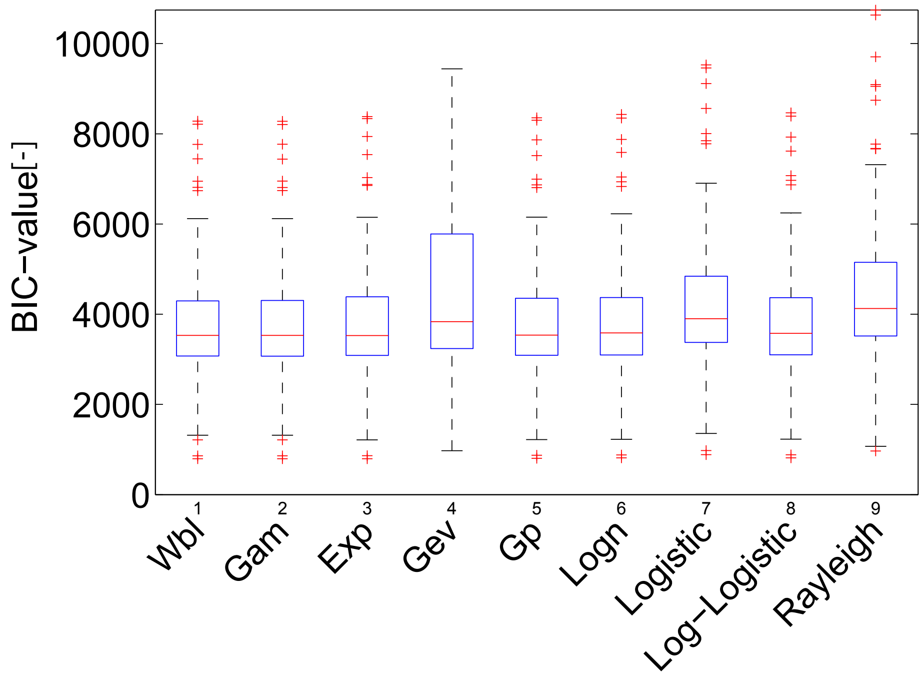

FIGURE 3.11: Boxplot of BIC-values of nine parametric distribution functions fitted to daily precipitation intensities in August (1950-2005).

The exponential distribution is defined by a single parameter λ_{exp}.

$$F(x) = 1 - e^{-\lambda_{exp}x} \tag{3.13}$$

The parameter λ_{exp} is the reciprocal value of the mean wet day amount $\overline{x_w}$. Since the mean wet day amount is easier interpretable, $\overline{x_w}$ was interpolated to the ungauged location instead of λ_{exp}.

Even though the exponential distribution did not result in the minimum BIC values, it was chosen for the following reasons:

- *Piani et al.* (2010) argue that a robust transfer function with few parameters is favorable for climate change studies. The exponential distribution's parameter $\lambda_{exp} = \frac{1}{\overline{x_w}}$ is defined by the mean of the random variable which is a simple and robust statistic that is not strongly influenced by single extreme values.

- As the rainy season is very pronounced in West Africa, a subdivision of the year into 9 seasons was made (one season for the dry season November to February and separate parameters for the other months). Fitting the seasonal CDF parameters with a moving window approach was also tested but rejected since it smoothed the monthly sums of precipitation too strongly and underestimated the strong seasonality in West Africa. As the estimation of variograms and the fitting of the distribution parameters is dependent on a sound basis of observation data, a distribution function with a single parameter can be fitted more easily.

- QQ-Plots of the simulated daily intensities against the observed ones showed a good fit for most locations.

- Stations on the same degree of latitude are generally more similar than stations on the same degree of longitude which has motivated anisotropic variograms to represent this observed feature of the West African climate system. Calculating experimental anisotropic variograms requires splitting the sample into subsets which reduces the sample size for the calculation of the directional variograms.

- Fitting a parametric CDF with more than one parameter can result in a high variability of the parameters between neighboring locations. The resulting variograms can be very flat with nearly constant values. One problem of such flat variograms is that nearly equal Kriging weights are given to all neighboring stations. Therefore, the generated maps are very smooth and local trends cannot be reproduced and a high estimation uncertainty for unknown points ensues. The other problem is that the Kriging matrix can become non-invertible if the variogram functions generate nearly identical values for a certain spatial configuration. This problem occurred for instance, for the Weibull distribution's scale parameter λ_{wbl} in March. Figure A.5 in Appendix A shows the experimental and fitted variograms of the scale parameter λ_{wbl}. The experimental variogram γ^* and the fitted spherical model $\widehat{\gamma}_{sp}$ are nearly constant.
  In contrast, the parameter of the exponential distribution is determined by the mean wet day amount which is a more stable descriptive measure. Thus, the estimation variance is more strongly influenced by the geometric configuration of the unknown point and the neighboring gauges and local trends in the observed distribution parameters can be respected. Possible extensions to lessen the problem of volatile or nearly-constant variograms could be to interpolate only one CDF parameter and make assumptions about the other CDF parameter, e.g. that it is constant for the complete region or that it correlates with the elevation or another geographical information and estimate it based on this dependence. For instance, *Marra et al.* (2019) estimated the shape parameter of the Weibull distribution based on the fitted scale parameter with linear regression.

The selection of a suitable distribution function for the observed variable could be based on different criteria. In the end, the selection of a single parametric distribution function remains somewhat subjective and can be motivated by practical problems like the sample size, the number of parameters of the distribution function or the tail behavior.

3.4.2 Normality of distribution parameters

An investigation of the distribution parameters p_w and $\overline{x_w}$ showed that they are approximately normally distributed for all seasons and no further transformation was performed. Figure 3.12 exemplifies this investigation for the month of August. The theoretical Gaussian CDFs were obtained by simulating from a Gaussian distribution with the standard deviation and mean of the observed CDF parameters within the study region.

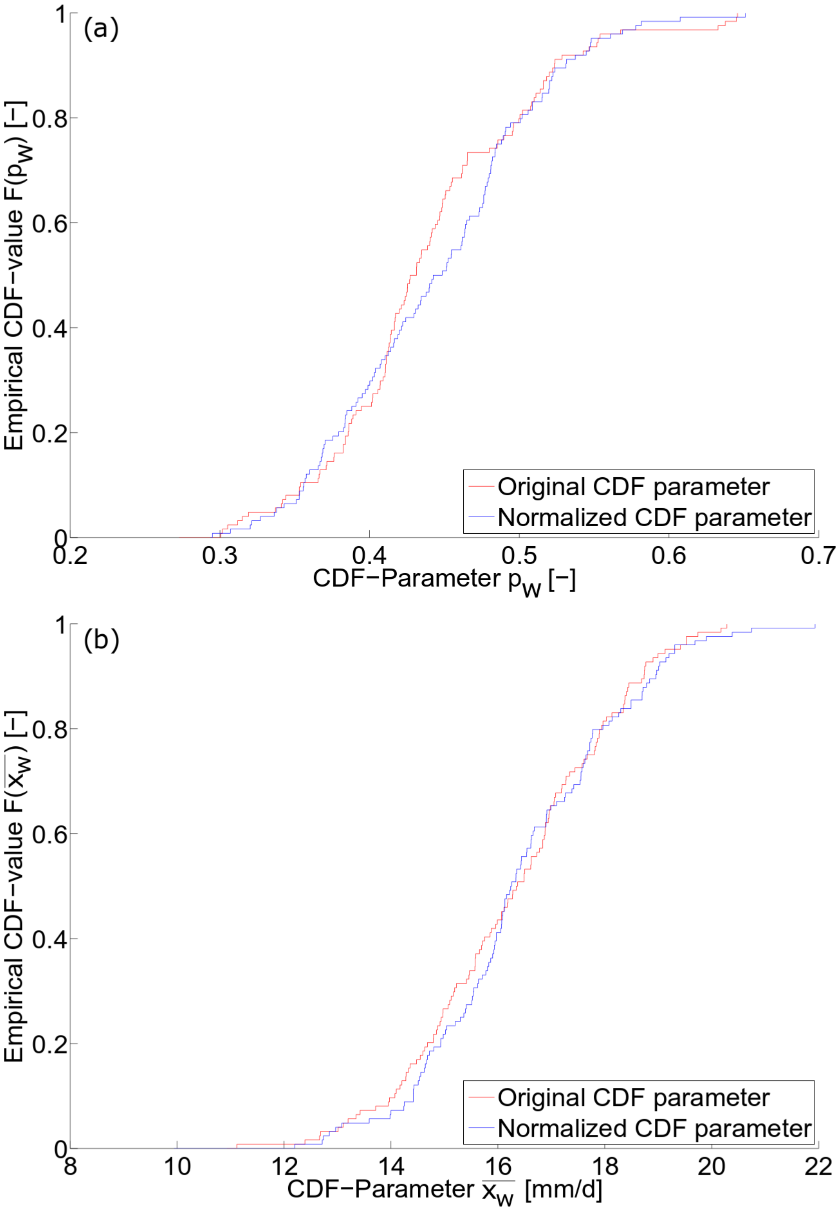

FIGURE 3.12: Empirical and Gaussian distribution of p_w (a) and $\overline{x_w}$ (b) in August (1950-2005).

3.4.3 Variograms of distribution parameters

Experimental anisotropic variograms of $\overline{x_w}$ and p_w were calculated for ten separation distances ranging from 0 to 300 km and four directions $(0°, 45°, 90°, 135°)$ to find the direction of maximum anisotropy (*Krūminiene*, 2006). It was found that lower semi-variances of $\overline{x_w}$ are to be expected when going from east to west (Figure 3.13 (a)) as when going from north to south (b) for both parameters. This anisotropy relates to the rainfall band that moves across West Africa from south to north during the rainy season. The numbers at the experimental variogram markers are the numbers of gauge pairs corresponding to the respective distances.

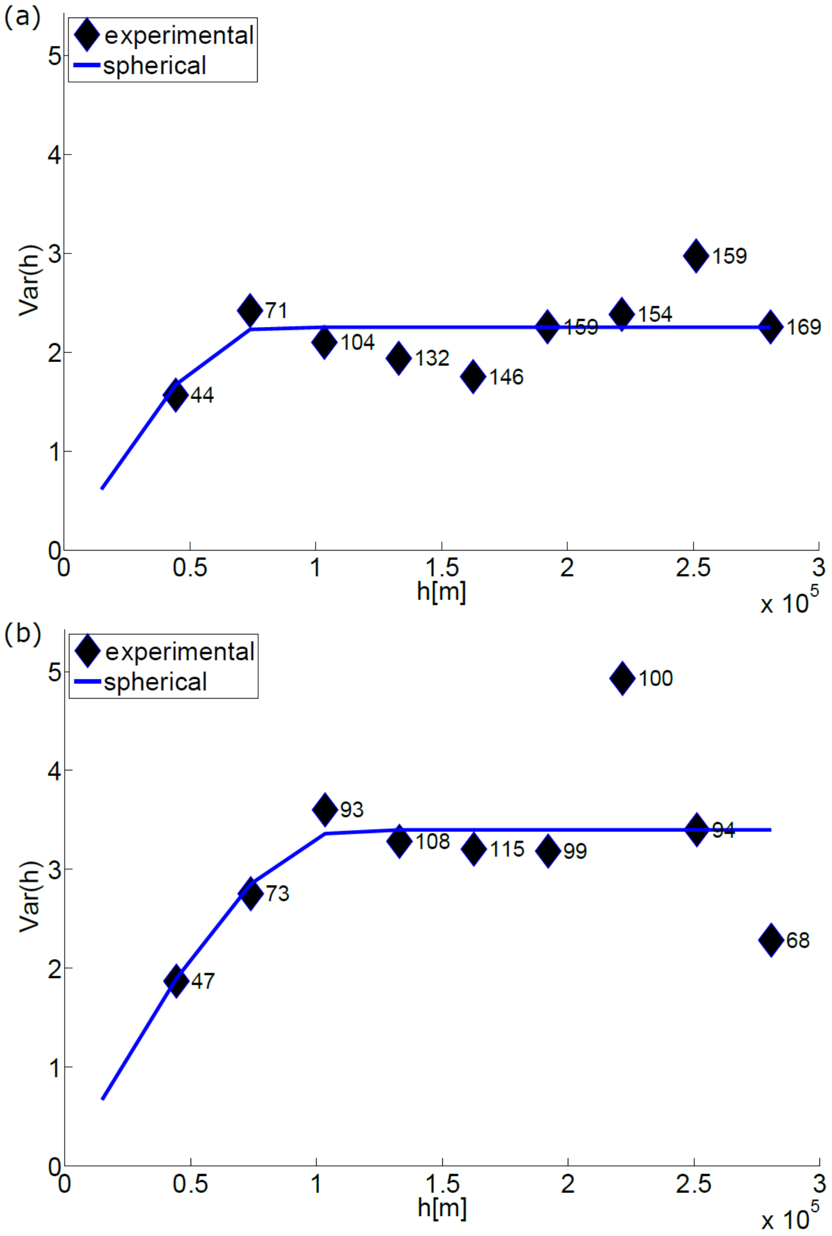

FIGURE 3.13: Experimental and fitted variograms γ_x of $\overline{x_w}$ in east-west (a) and north-south direction (b) in August (1950-2005).

3.4.4 Fitting a distribution function to the RCM simulations

RCM precipitation intensities are typically highly skewed and the probability of precipitation is generally overestimated when compared with observation data. Figure 3.14 shows the empirical distribution of one model (KNMI-RACMO22-HadGEM2) in June for one cell. The historical and future (RCP 8.5) precipitation probability p_w amounts to approximately 90% , whereas the interpolated p_w is only about 28%.

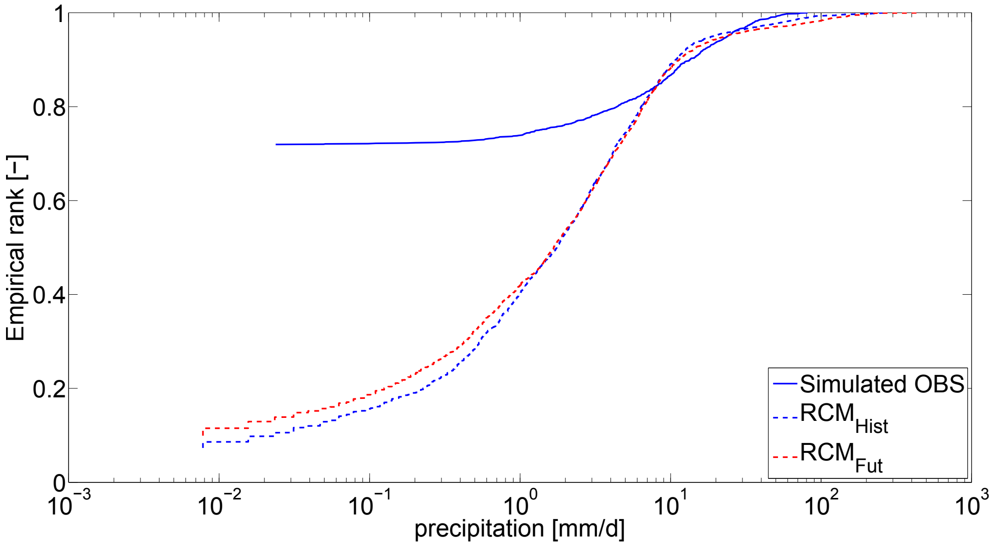

Figure 3.14: CDF of historical (1950-2005) and future (2006-2100, RCP8.5) precipitation of one model and of simulated observations in June for one cell.

While the future p_w is a bit lower than in the historical period, the values above 40 $mm\ d^{-1}$ are higher. This is shown in the Quantile-Quantile-Plot in Figure 3.15.

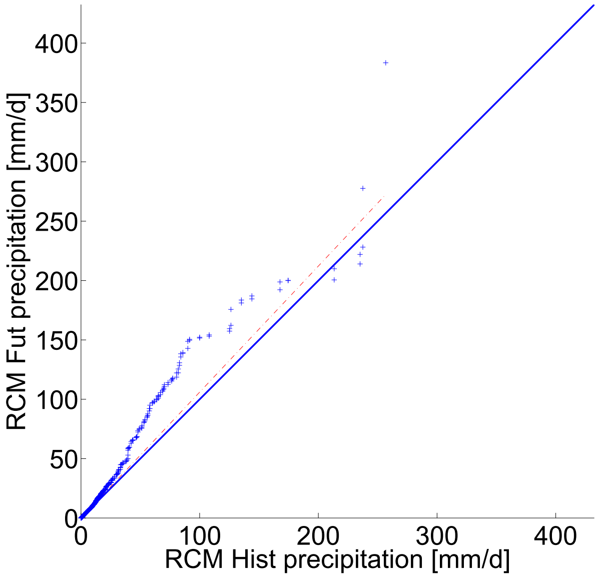

Figure 3.15: QQ-Plot of historical (1950-2005) and future (2006-2100, RCP8.5) precipitation of one model in June for one cell center with bisecting line (blue) and regression line (red).

In a first step, the RCM precipitation was dry day corrected. The *Dry Day Correction* (subsection 3.1.3) was calibrated by calculating the threshold ϑ for each season, grid cell and CORDEX-Africa model. The assumption is that the threshold ϑ remains constant for the future period.
The remaining values were shifted towards 0 and a distribution was fitted to the positive amounts. Finding a parametric distribution $F_{sim,hist}$ that fits such highly skewed data is very problematic. Instead, a KDE-CDF (Chapter 2, subsection 2.2.5) was utilized. This type of CDF was chosen due to the very high extreme values that can occur in a large ensemble of RCM simulations. The data set is large enough to allow for a robust fit of the KDE-CDF on a monthly basis with the dry season November - February pooled into one season.

3.5 Evaluation of the geostatistical bias correction model

In this section, the performance of the bias correction model is evaluated by comparing the simulated observations with observed data. Afterwards, the bias-corrected RCM precipitation is compared with observations to evaluate the suitability of the KDE-CDF to fit the historical time series of RCM precipitation.

- Maps of the interpolated distribution parameters are shown in subsection 3.5.1.
- The distribution parameters were interpolated to the gauge locations to perform a cross validation. The results of this analysis are presented in subsection 3.5.2.
- The simulated observations were also compared with the nearest observations on different aggregation levels. The difference of these analyses to the cross validation is that the simulated observations were simulated for the grid cell centers of the study region and not for the locations with observation data. Comparisons of the annual, monthly and daily precipitation are given in subsection 3.5.3.
- The fit of the KDE-CDF to the historical RCM simulations was evaluated by performing a bias correction for the historical period and comparing monthly precipitation to the nearest observed values in subsection 3.5.4.

3.5.1 Maps of interpolated distribution parameters

With the fitted variograms $\widehat{\gamma}(h)$, the Kriging Equation System was built for each ungauged location. Measurement stations were considered as supporting points of the interpolation of the precipitation probability p_w if at least 500 valid daily values had been measured in the given season. The interpolated field of p_w in August is shown in Figure 3.16 (a). For the mean wet day amount $\overline{x_w}$ (Figure 3.16 (b)) it was required that at least 100 wet values had been measured at each location.

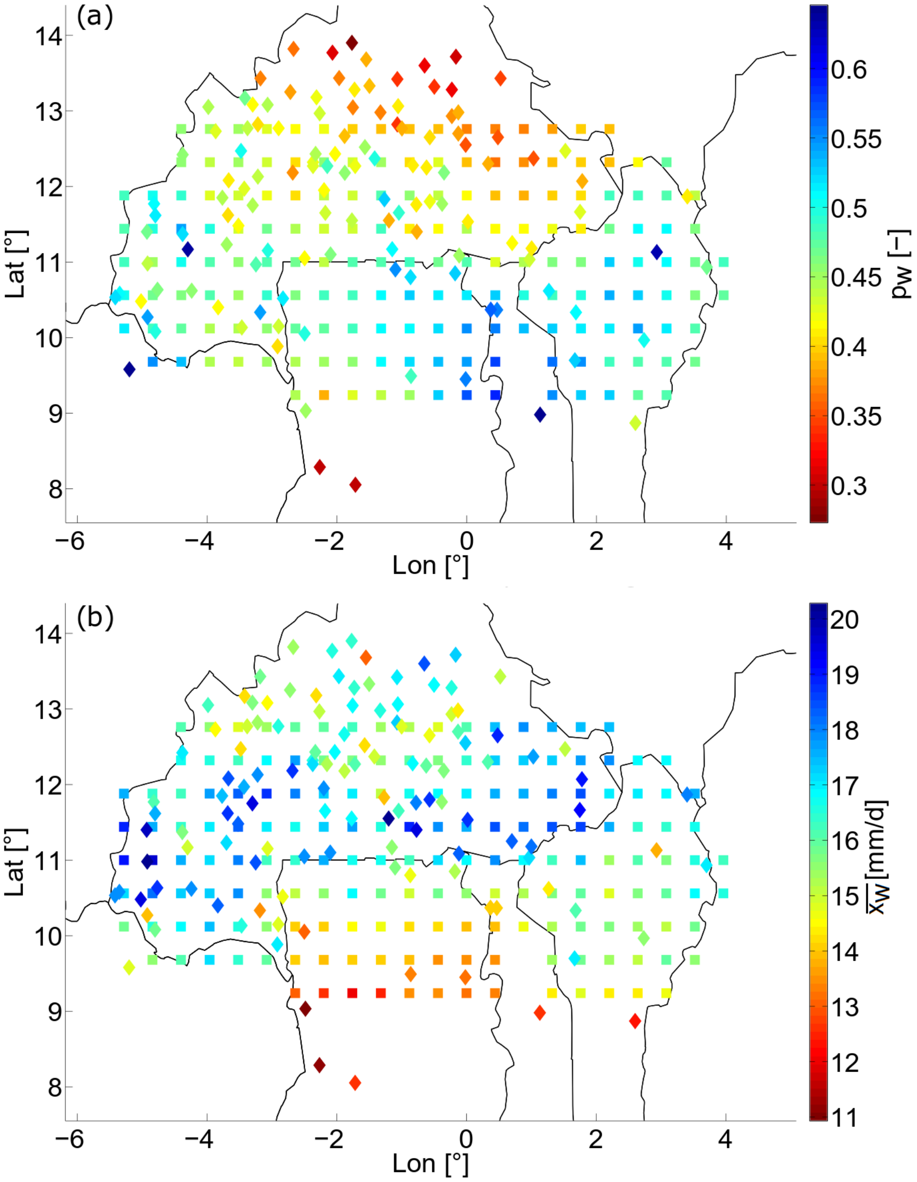

FIGURE 3.16: Kriged probability of precipitation p_w (a) and mean wet day amount $\overline{x_w}$ (b) in August (1950-2005) - Diamonds: Observed, Squares: Kriged.

3.5.2 Cross-Validation of interpolated distribution parameters

A cross-validation was performed by interpolating the probability of precipitation and the mean wet day amount $\overline{x_w}$ to every measurement location where enough data is available to fit the exponential distribution function. The interpolating was done with the Kriging procedure and a standard Inverse Distance Weighting procedure (IDW) for comparison purposes. The supporting points for the interpolation were the four closest measurement stations surrounding the unknown point. Figure 3.17 shows the performance of estimating p_w for August and the season-wise performance is given in Table 3.2.

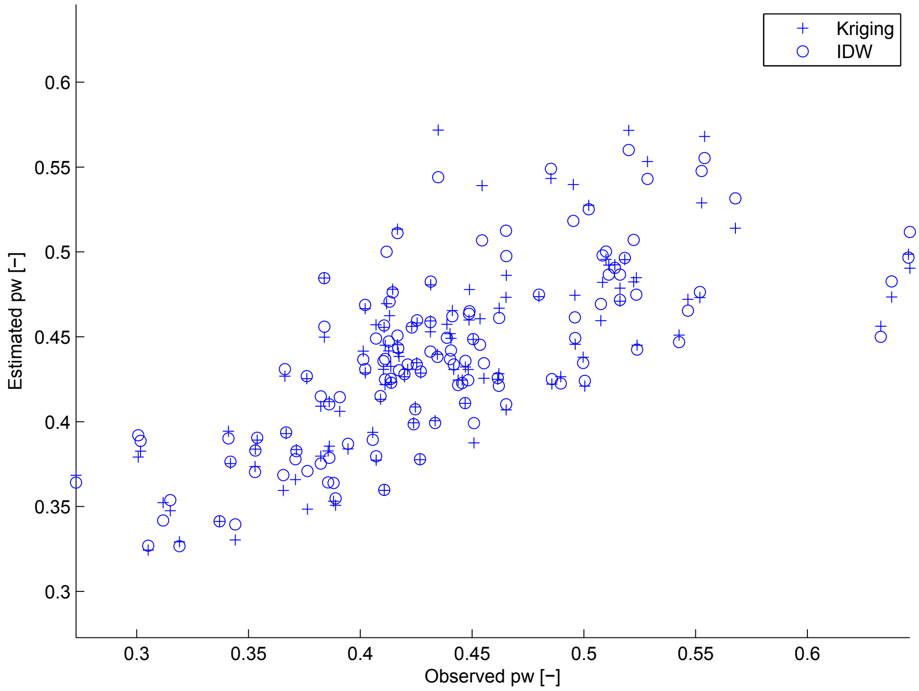

FIGURE 3.17: Scatterplot of interpolated p_w compared to observed value in August (1950-2005).

	Kriging			IDW		
Season	$r_{xy}[-]$	$MAE[-]$	$MSE[-]$	$r_{xy}[-]$	$MAE[-]$	$MSE[-]$
1 NDJF	0.82383	0.00444	**0.00007**	**0.82456**	**0.00441**	0.00007
2 Mar	0.91475	**0.00965**	**0.00024**	**0.91600**	0.01072	0.00028
3 Apr	0.94214	0.01473	0.00049	**0.94711**	**0.01380**	**0.00048**
4 May	**0.92342**	**0.01881**	**0.00064**	0.91764	0.01969	0.00070
5 Jun	0.85170	0.02492	0.00111	**0.86046**	**0.02346**	**0.00106**
6 Jul	0.71645	0.03122	0.00178	**0.73081**	**0.03047**	**0.00169**
7 Aug	0.69614	0.03837	0.00260	**0.71198**	**0.03704**	**0.00248**
8 Sep	0.89916	0.03459	0.00208	**0.90684**	**0.03200**	**0.00198**
9 Oct	0.90536	**0.02171**	**0.00115**	**0.90865**	0.02193	0.00121
Average	0.85255	0.02205	0.00113	**0.85823**	**0.02150**	**0.00111**

TABLE 3.2: Cross validation of Kriging and IDW for p_w. r_{xy}: Correlation, MAE: mean absolute error, MSE: mean squared error.

Analogously, the interpolated mean wet day amount $\overline{x_w}$ in August is illustrated in Figure 3.18 and Table 3.3 shows the seasonal performance.

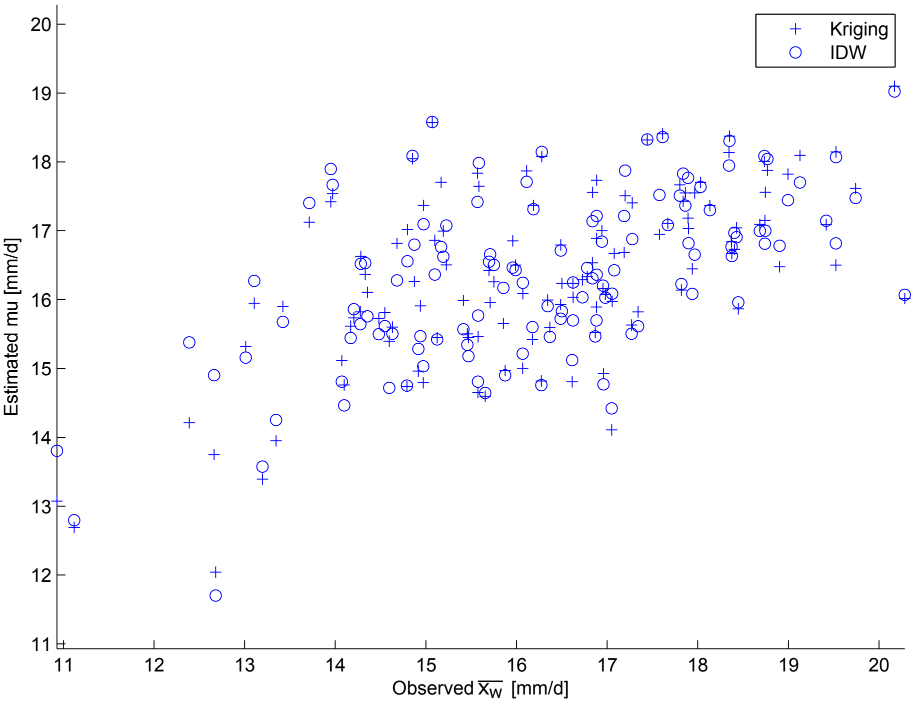

Figure 3.18: Scatterplot of interpolated $\overline{x_w}$ compared to observed value in August (1950-2005).

	Kriging			IDW		
Season	$r_{xy}[-]$	$MAE[\frac{mm}{d}]$	$MSE[\frac{mm}{d}]^2$	$r_{xy}[-]$	$MAE[\frac{mm}{d}]$	$MSE[\frac{mm}{d}]^2$
1 NDJF	0.48593	1.42200	3.26400	**0.52886**	**1.34700**	**3.03000**
2 Mar	0.51964	1.75200	**4.71900**	**0.52296**	**1.75100**	4.72300
3 Apr	**0.56343**	**1.43700**	**3.45500**	0.56281	1.44500	3.50100
4 May	**0.65652**	1.03600	**1.77600**	0.65027	**1.03400**	1.81300
5 Jun	**0.50501**	**0.96550**	**1.48800**	0.49143	0.96680	1.52900
6 Jul	**0.55925**	**1.03600**	**1.80600**	0.51425	1.08400	1.96500
7 Aug	**0.59585**	**1.21600**	**2.30800**	0.57006	1.24400	2.40800
8 Sep	0.47808	**0.93650**	**1.37900**	**0.43842**	0.97710	1.48200
9 Oct	**0.48068**	1.06200	1.66300	0.50142	**1.05900**	**1.62300**
Average	**0.53827**	**1.20700**	**2.42867**	0.53116	1.21199	2.45267

Table 3.3: Cross validation of Kriging and IDW for $\overline{x_w}$. r_{xy}: Correlation, MAE: mean absolute error, MSE: mean squared error.

Both methods reproduce the observed statistic similarly for the month of August. The probability of precipitation p_w was interpolated slightly better with IDW, whereas Kriging resulted in slightly lower errors for the mean wet day amount $\overline{x_w}$. The seasonal analysis shows that the performance of the methods in August is similar in the other months. For p_w, IDW performs slightly better on average (Table 3.2), whereas the kriged $\overline{x_w}$ values have lower mean absolute errors and higher correlation coefficients (Table 3.3). *de Amorim Borges et al.* (2006) interpolated daily precipitation in Brazil with different methods and also found that IDW may outperform more complicated interpolation methods.

To evaluate the performance of the estimated CDF, a statistical test against a CDF fitted to observation was carried out. For each location with valid CDF parameters, the measured values of the corresponding season serve as a reference set x_{ref}. With the observed CDF parameters, a set of values $x_{sim,obs}$ of the same length as x_{ref} was simulated with the truncated exponential distribution. Likewise, a set $x_{sim,krig}$ was simulated from the estimated CDF parameters that were interpolated from the neighboring stations. Both simulated sets were then tested with a KS-Test at a significance level of $\alpha = 5\%$ against the reference set. Figure 3.19 presents the ratio of accepted tests for each season.

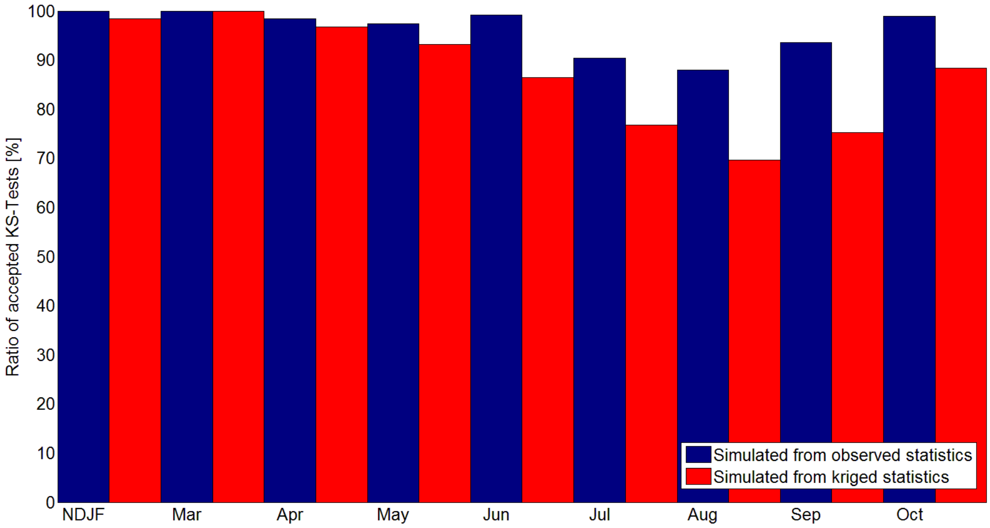

Figure 3.19: Ratio of accepted KS-Tests of simulated precipitation.

It can be seen, that the summer months do not always follow the exponential distribution and that the winter months are better represented by the exponential distribution. As Kriging tends to produce smoothed estimates, less tests were accepted for the sets simulated from the kriged parameters. The kriged parameters' ratio of accepted KS-Tests lies in the range of 69.6 to 100%. Since the exponential distribution fitted to observed precipitation was not always capable of passing the KS-Test neither, the ratios of accepted KS-Tests of kriged parameters was divided by the ratio of accepted KS-Tests with observed parameters to separate the Kriging performance from the suitability of the exponential distribution. In cases where the exponential distribution with observed parameters passed the KS-Test, between 79.1 and 100% of the estimated distributions also passed the test.

3.5.3 Comparison of simulated observations and observed precipitation

In the previous analyses, the simulated observations were generated for locations with measured data and a comparison of the distributions was performed. In the following, the simulated observations were generated for the grid cell centers of the study region and the comparison was performed with the closest measurement station. As was shown in Figure 3.2, there is no observed data for several grid cells. Therefore the closest measurement station can be separated by as much as 91 km from the grid cell centers.

Annual sums of precipitation of simulated observations

The annual sum of precipitation is a very robust statistic and can be matched well by the simulated observations (Figure 3.20) as the majority of points lie close to the bisecting line.

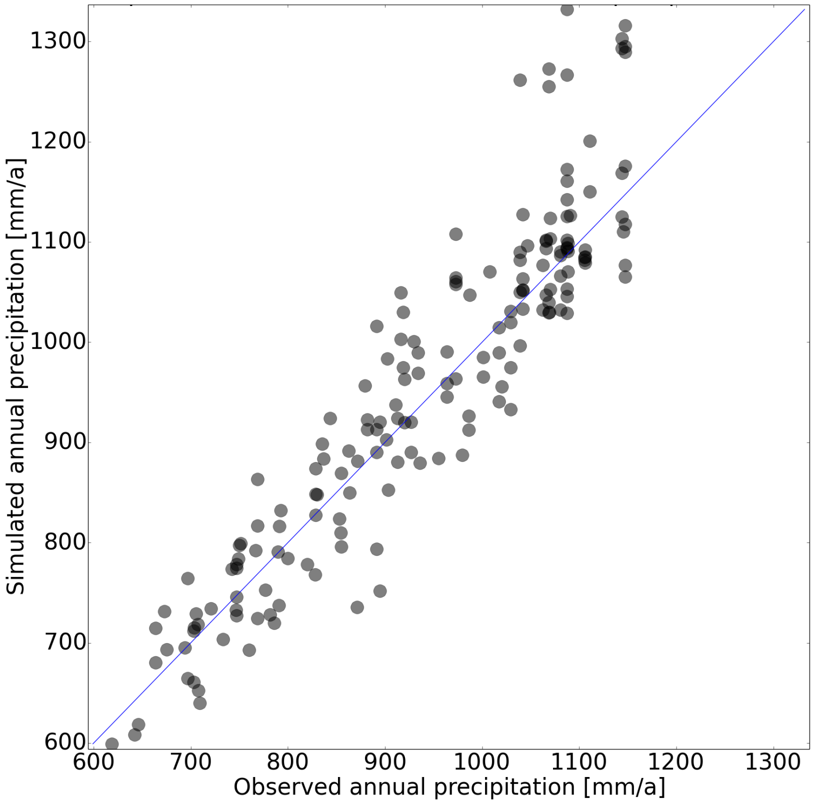

FIGURE 3.20: Mean annual sum of precipitation of simulated observations and nearest measurement station (1950-2005).

Monthly sums of precipitation of simulated observations

The average monthly sums of precipitation were also reproduced rather accurately by the simulated observations with a mean difference of 1.7 *mm mon*$^{-1}$ to the nearest observed monthly sum (Figure 3.21).

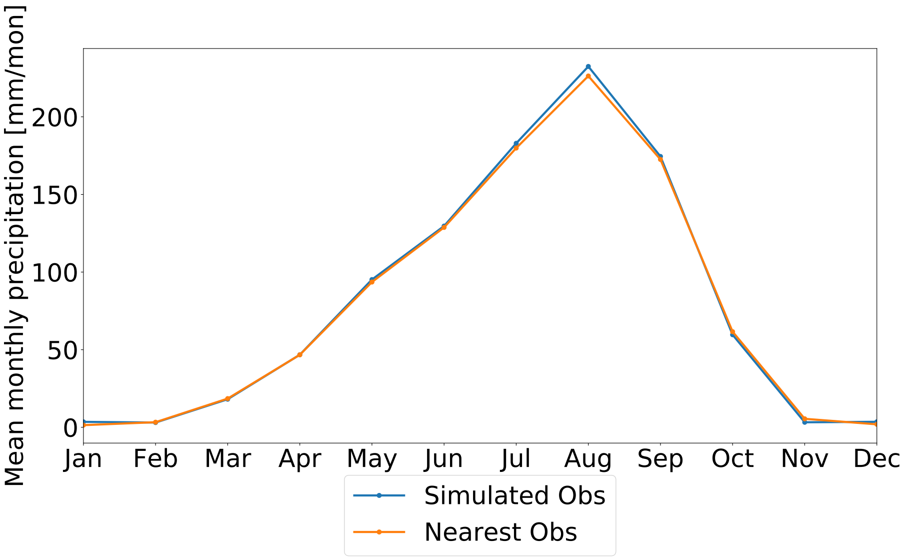

FIGURE 3.21: Mean monthly sum of precipitation of simulated observations and nearest measurement station (1950-2005).

A map of the average difference of the simulated mean monthly precipitation to the nearest observed value is given in Figure 3.22. The locations of the measurement stations that were used for this comparison are marked as crosses.

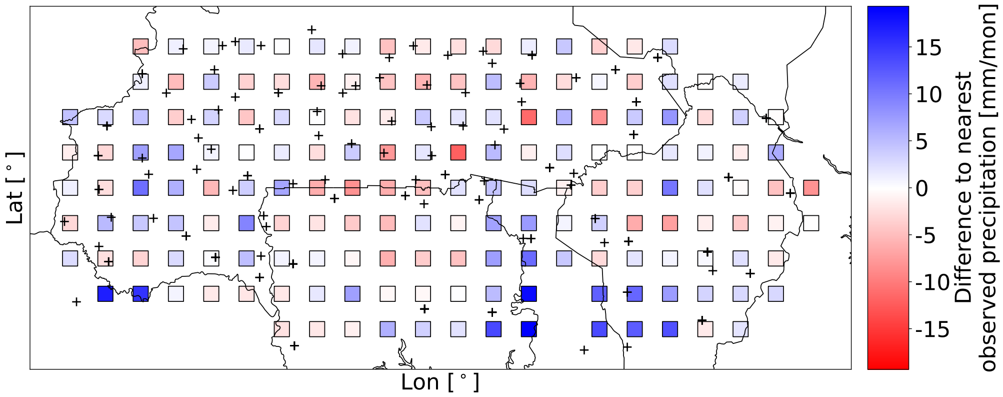

FIGURE 3.22: Difference of mean monthly sum of precipitation of simulated observations to nearest measurement station (1950-2005).

Daily precipitation of simulated observations

The Quantile-Quantile-Plots of the simulated and observed daily precipitation indicate that the data sets simulated from the exponential distribution are generally similar to the observed distribution of the nearest station. Figure 3.23(a) shows the simulated observations for Ougadougou, Burkina Faso. The QQ-Plots of some cells indicate an underestimation of the extreme values. Figure 3.23(b) is an example of one the worst fits for a cell in the north of Ghana. This region exhibits high extreme values which the exponential distribution cannot reproduce accurately.

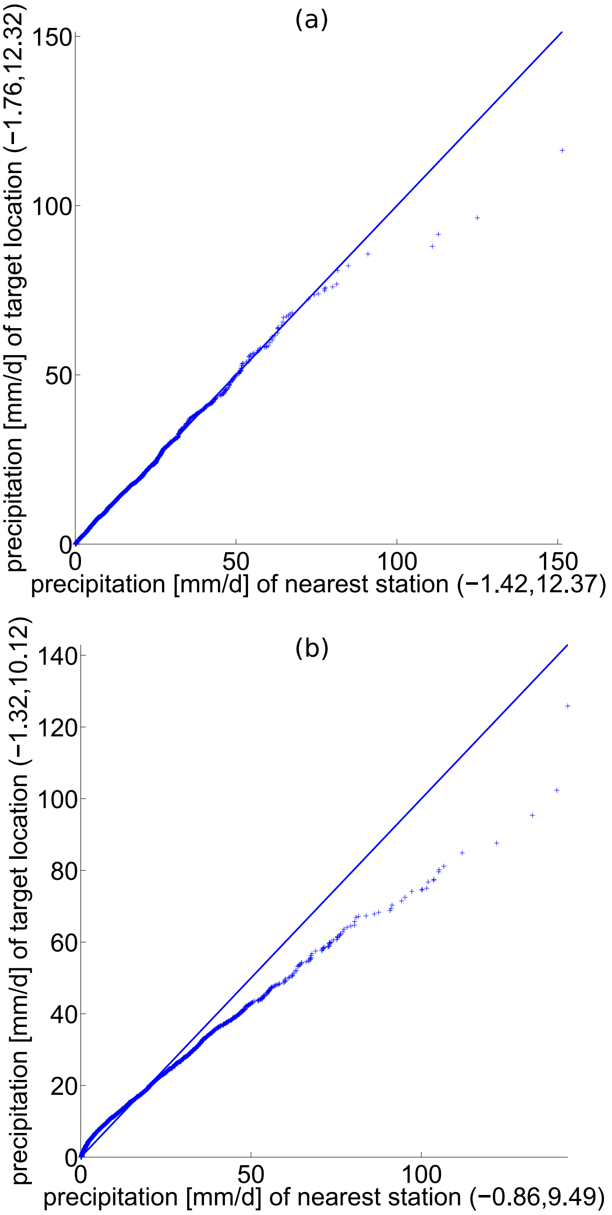

FIGURE 3.23: QQ-Plot of simulated and observed daily precipitation for Ouagadougou, Burkina Faso (a) and for a cell in the north of Ghana (b) (1950-2005).

3.5.4 Evaluation of the distribution of the regional climate models

The CDF of $F_{sim,hist}$ was fitted to the daily positive precipitation amounts $x_{RCM,hist}$ of the historical period 1950-2005 with a KDE-CDF. When the fit of a CDF is perfect, the CDF values are uniformly distributed. A regular *Quantile Mapping* (subsection 3.1.2) was performed for the historical period and the mean monthly sums of the bias corrected models of the historical period were calculated and plotted with the nearest observed monthly sums (Figure 3.24). As the diagram looks nearly identical to the comparison of the simulated and nearest observations (Figure 3.21), it can be concluded that the KDE-CDF fits the RCM precipitation well.

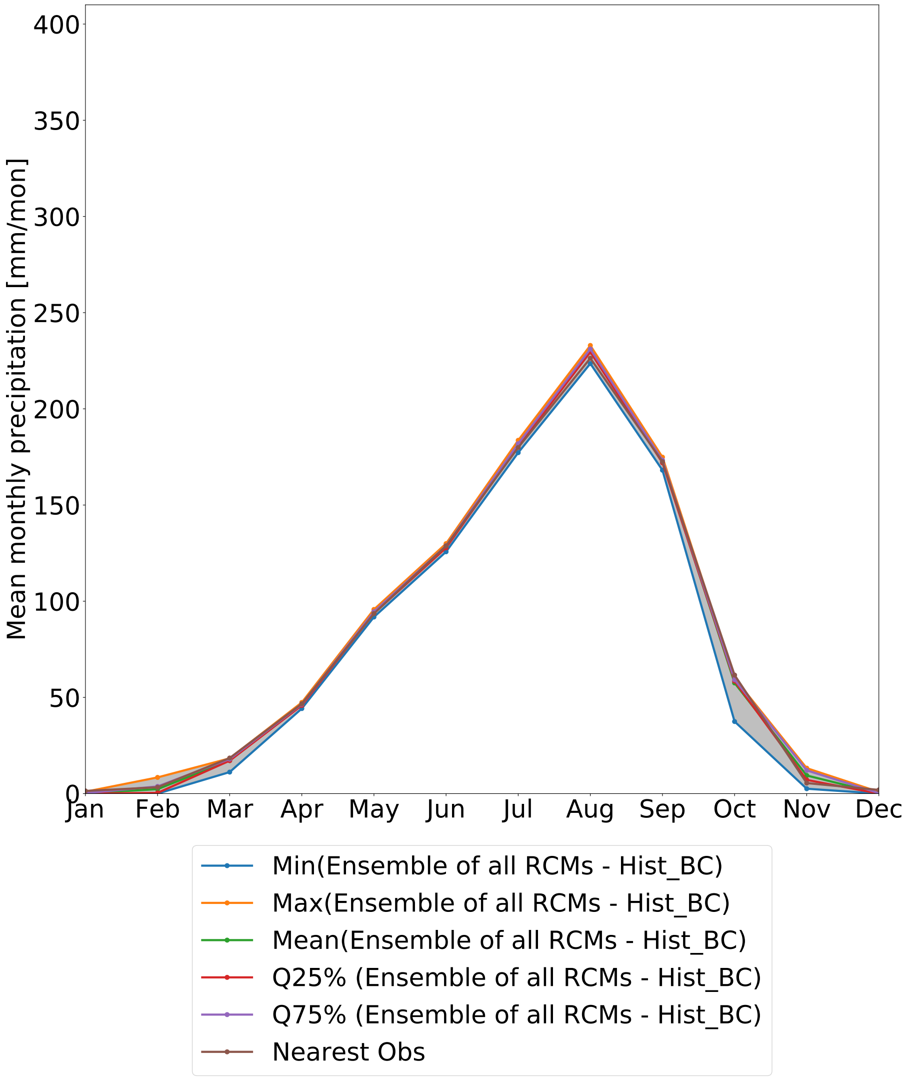

FIGURE 3.24: Mean monthly sum of precipitation of simulated observations and nearest measurement station (1950-2005).

3.6 Evaluation of the projected climatology of the bias corrected RCMs

A total of 48 models were bias corrected (Table A.2 in Appendix A). 5 models were run for the RCP 2.6 scenario, 22 for RCP 4.5 and 21 for RCP 8.5. For an analysis of the projected climatology, the future period was split up into the near future (2020-2050) and the far future (2070-2100). The historical period was chosen as 1970-2000 so that all periods are 30 years long.

The differences between the bias corrected historical and future climatology are caused by a change in the distribution. Since the *Dry Day Correction* thresholds ϑ can be very large (subsection 3.6.3), a change of its non-exceedance probability in the future period introduces a change of the bias corrected distribution that cannot be inferred from the raw simulations. Therefore the large number of uncertain low precipitation amounts can have a huge influence on the simulations' statistics. Similar findings were reported by *Polade et al.* (2014) who separated the contribution of changes in the number of wet days ($> 1\ mm\ d^{-1}$) from the changes in precipitation amounts on wet days to the annual sum of the CMIP5 (Coupled Model Intercomparison Project Phase 5) ensemble. Between 40° S and 40° N, changes in the wet day frequency contributed more than 50% to the change of annual precipitation.

3.6.1 Projected change of annual precipitation

The mean annual sum of precipitation is projected to change for all RCP scenarios but the magnitude and sign of change depend on the given RCP scenario and the geographical location. Figure 3.25 (a) is a violin plot of the difference of the annual precipitation in the near future. The differences were averaged over all grid cells and the spread of the violins relates to the different models in the ensemble. The median change amounts to $-10.6\ mm\ a^{-1}$ for the RCP 2.6 scenario. For the scenarios RCP 4.5 and RCP 8.5, the median change is positive and stronger ($28.3\ mm\ a^{-1}$ for RCP 4.5 and $79.2\ mm\ a^{-1}$ for RCP 8.5).

For the far future (2070-2100), similar differences were calculated (Figure 3.25 (b)). In the RCP 2.6 scenario, a median decrease of $-25.1\ mm\ a^{-1}$ is expected. For the other scenarios, the difference is again positive ($41.8\ mm\ a^{-1}$ for RCP 4.5 and $106.6\ mm\ a^{-1}$ for RCP 8.5). In contrast to the near future, the spread is larger and more models project very large differences of more than $400\ mm\ a^{-1}$.

This does not necessarily indicate that more water will be available for crop cultivation because the globally rising temperature might lead to a higher evapotranspiration. Also, the annual cycle is projected to change slightly, so that the larger available quantity of water precipitates later in the year. This is further discussed in subsections 3.6.2 (Projected change of monthly precipitation) and 3.6.4 (Projected change of onset of rainy season).

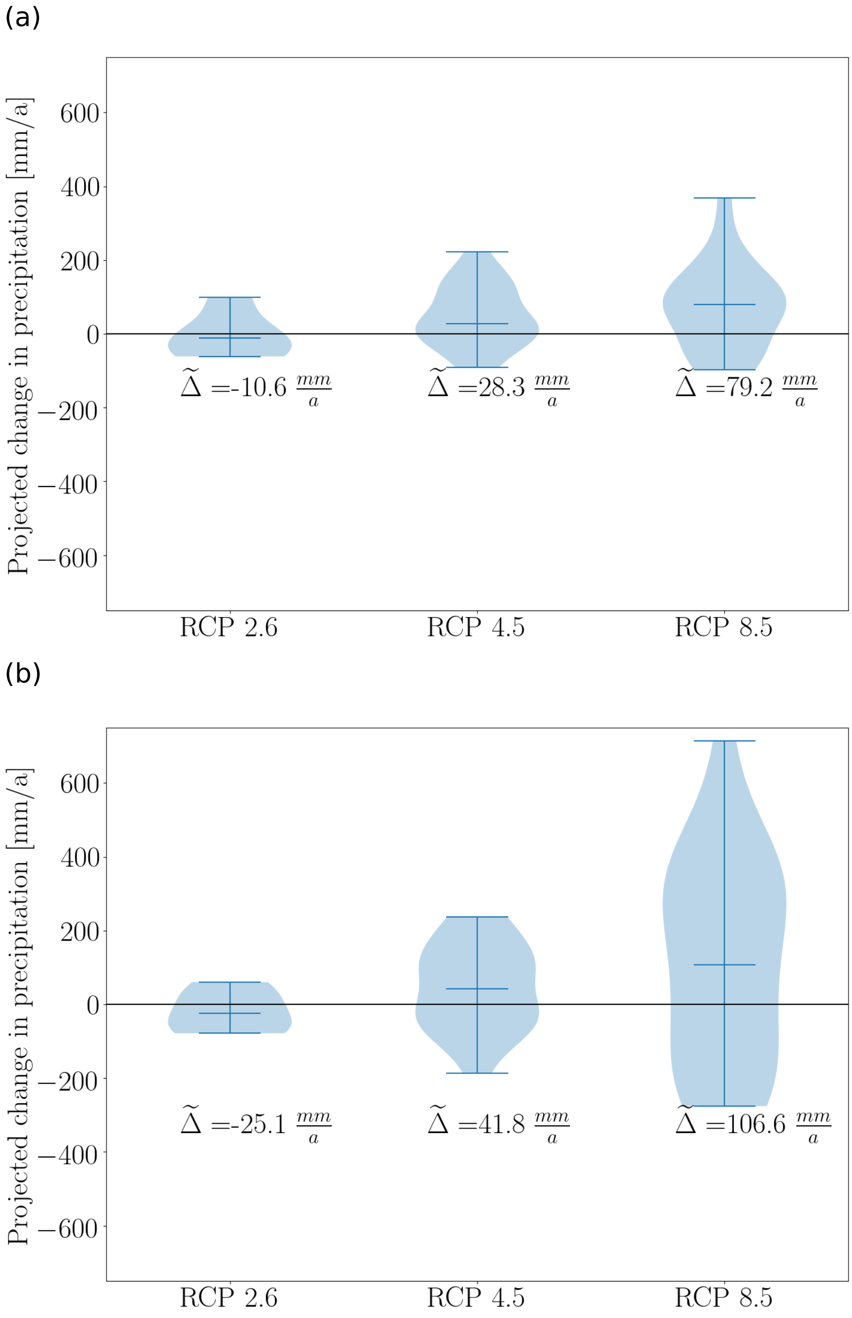

FIGURE 3.25: Violin plot of change of future annual precipitation for the three RCP scenarios in the near future 2020-2050 (a) and the far future 2070-2100 (b) in comparison to the historical period 1970-2000.

The location-specific average change of annual precipitation was calculated to identify regions where the ensemble members show a strong climate change signal. The amount of annual precipitation is generally projected to increase for most locations and RCP scenarios. In the RCP 2.6 scenario however, the annual totals decrease for the northern regions by as much as $-52\ mm\ a^{-1}$ in the near future (Figure 3.26 (a)). For the far future (b), the map looks similar and a decreased annual precipitation amount is expected for the northern regions. In comparison to the near future, the future is projected to be slightly drier.

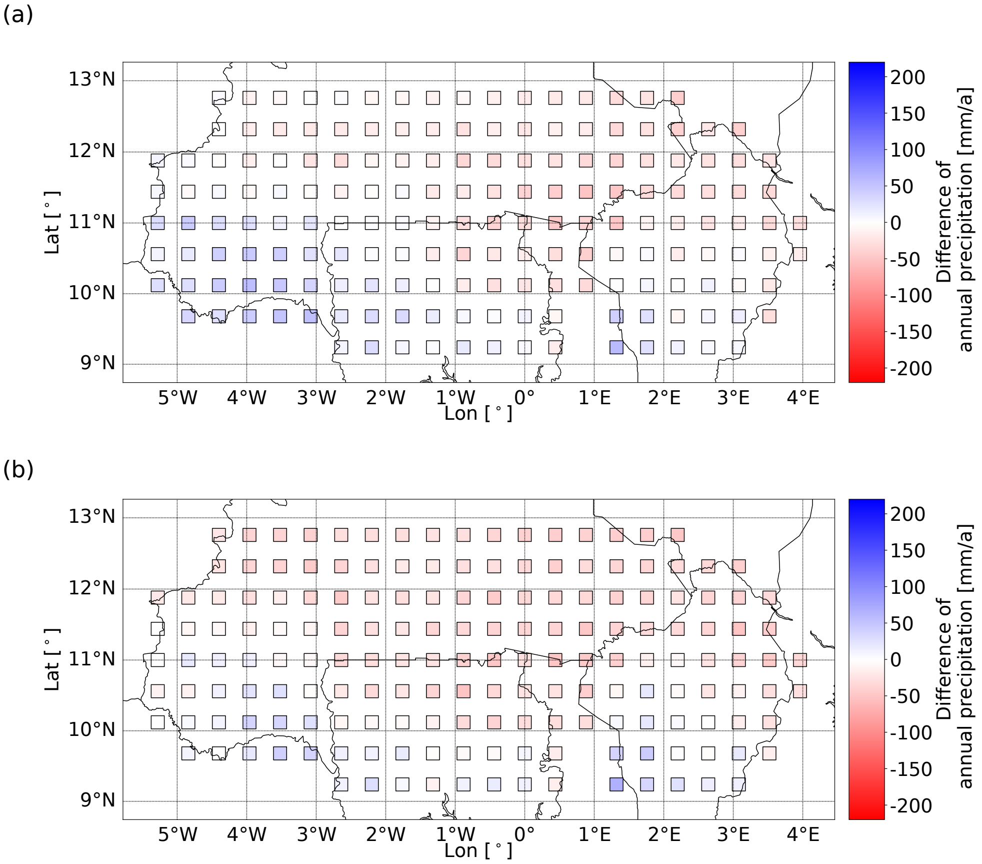

FIGURE 3.26: Absolute difference of annual precipitation in the near future 2020-2050 (a) and far future 2070-2100 (b) in comparison to the historical period (1970-2000) for RCP 2.6.

The average difference of the annual sum of precipitation in the near future (Figure 3.27 (a)) and far future (b) to the historical period is positive for the RCP 4.5 scenario. The increasing annual sums occur mainly in the southern regions as in the RCP 2.6 scenario. Differences between the far and near future are very small.

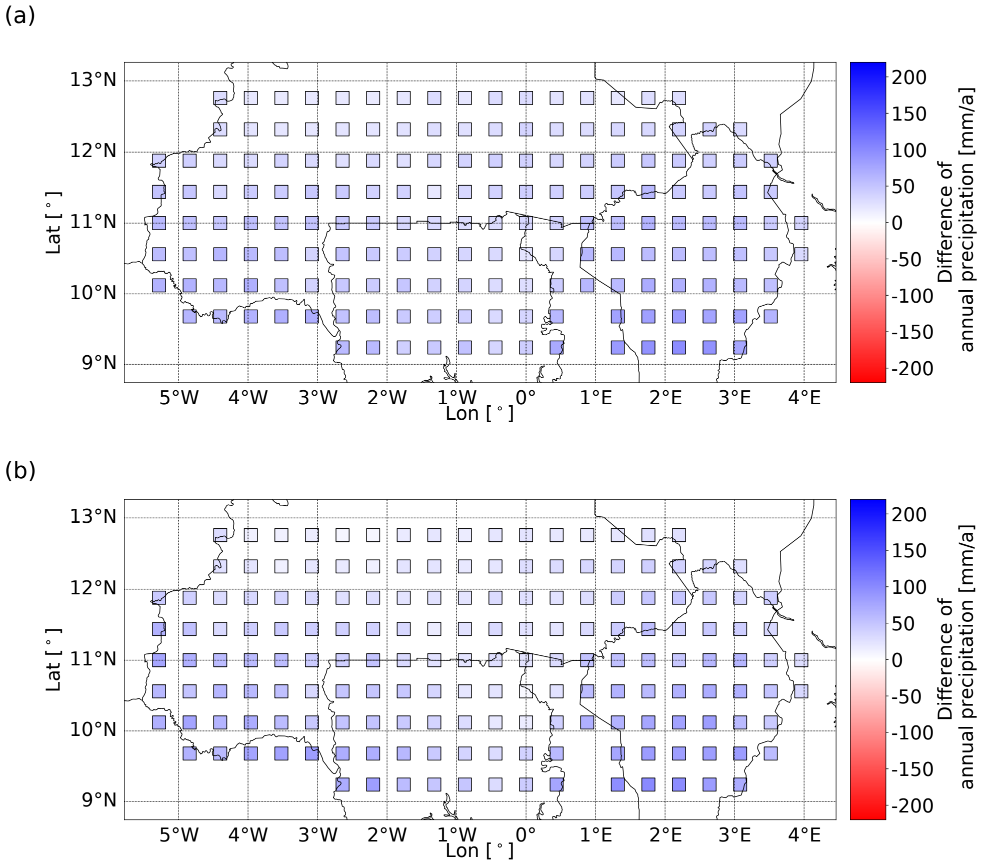

FIGURE 3.27: Absolute difference of annual precipitation in the near future 2020-2050 (a) and far future 2070-2100 (b) in comparison to the historical period (1970-2000) for RCP 4.5.

The simulations of the RCP 8.5 exhibit the highest differences in comparison to the historical period for the near future (Figure 3.28 (a)) and the far future (b). For all cells, increased annual sums of precipitation were calculated. Also, the difference of the far to the near future are larger than for the other RCP scenarios and for every cell, the annual precipitation is projected to intensify.

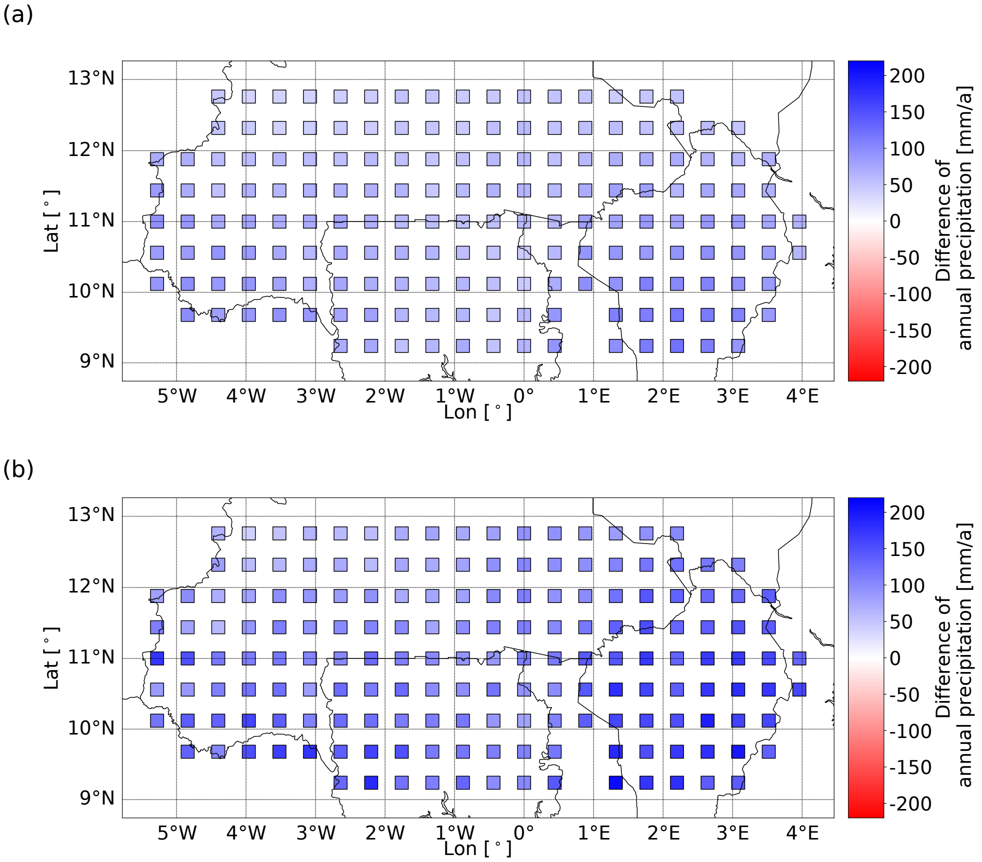

Figure 3.28: Absolute difference of annual precipitation in the near future 2020-2050 (a) and far future 2070-2100 (b) in comparison to the historical period (1970-2000) for RCP 8.5.

3.6.2 Projected change of monthly precipitation

In general, the spread of the mean monthly sums of precipitation is reduced by the bias correction method. The largest differences occur for the RCP 8.5 scenario and therefore, the spread of the projected monthly sums is only shown for this scenario (Figure 3.29). The mean monthly sums of the nearest observations stations are also included for reference purposes. The corresponding plots for the other two scenarios can be found in Appendix A (Figure A.1 and Figure A.2).

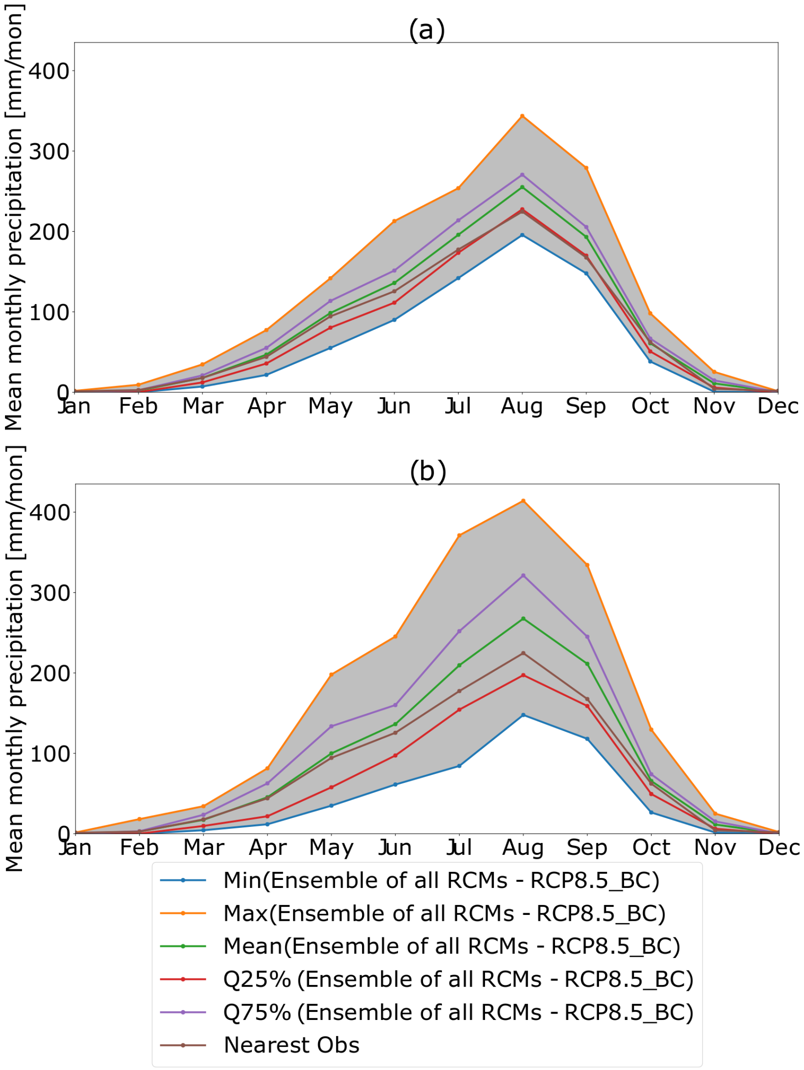

FIGURE 3.29: Mean monthly sum of precipitation averaged over 173 grid cells - bias-corrected RCP 8.5 scenario - a: near future (2020-2050), b: far future (2070-2100).

The majority of the models project increasing mean monthly sums for the RCP 8.5 scenario when compared with the historical period 1970-2000 (Figure 3.30). As the ensemble spread is very large, there exist models however which project lower mean monthly sums for the near and far future.

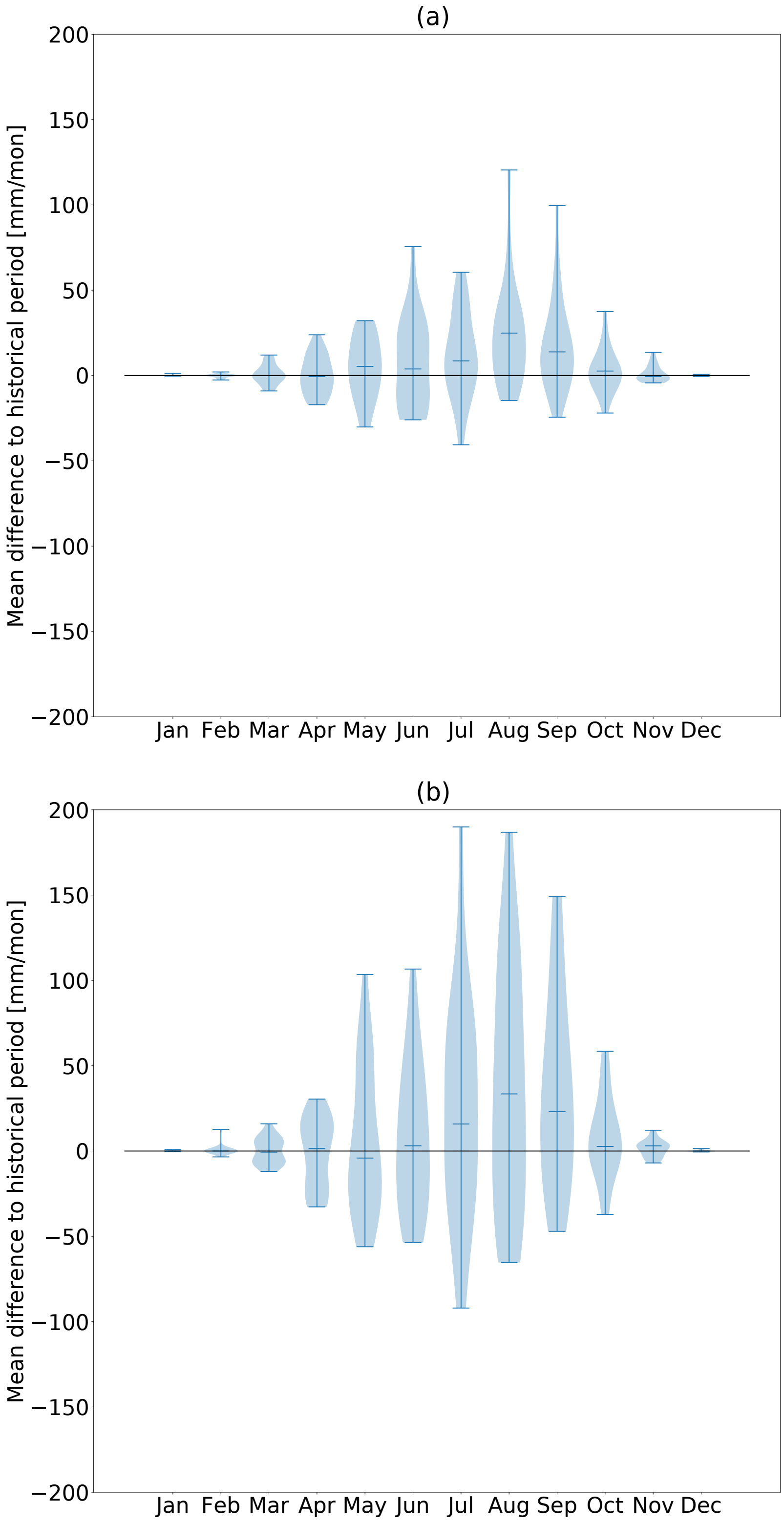

FIGURE 3.30: Difference of mean monthly sums of precipitation in the bias corrected RCP 8.5 scenario in the near future 2020-2050 (a) and the far future 2070-2100 (b) compared to historical period (1970-2000).

3.6.3 Projected change of daily precipitation

The distribution of precipitation in the future generally differs from the historical period. The *Double Quantile Mapping* calculates the CDF values of the future period according to the CDF corresponding to the historical period (Equation (3.5)) to account for the different climatic conditions projected for the future. Therefore, a shift in the distribution of the CDF values occurs and they are not uniformly distributed in the future period. A comparison of the bias corrected future and historical time series in June (Figure 3.31) shows how the large values are projected to increase for the model cell that was presented before (Figure 3.15).

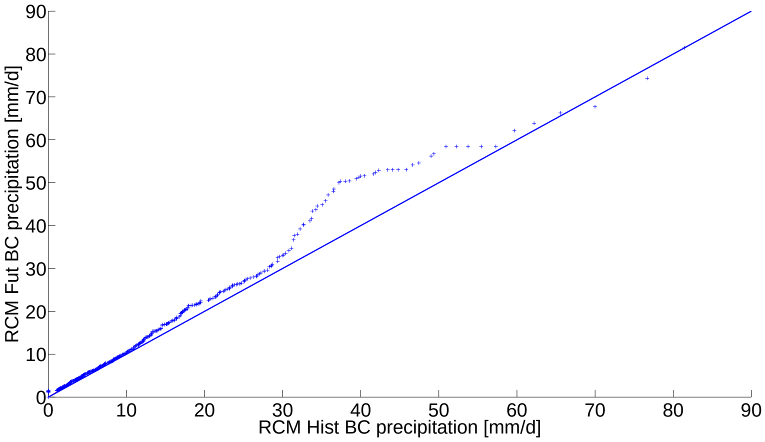

FIGURE 3.31: QQ-Plot of bias corrected historical (1950-2005) and future (2006-2100, RCP 8.5) precipitation intensities in June for one cell.

The *Dry Day Correction* method also leads to a difference of the bias corrected historical and future distributions. The assumption is that the threshold ϑ remains constant for the future period. For the presented model, location and season, the observed precipitation probability is $p_w = 28.1\%$ with corresponding $\vartheta = 4.38\ mm\ d^{-1}$. Applying this ϑ to the future period changes p_w to 29.2%. Thus, the precipitation probability is slightly higher in the bias corrected RCP 8.5 time series even though the wet day probability ($\geq 0\ mm\ d^{-1}$) in the uncorrected future period is lower than in the uncorrected historical period (Figure 3.14). This is caused by more values exceeding the threshold ϑ in the future period. It can be seen that the future CDF in Figure 3.14 intersects the historical CDF at around 1 $mm\ d^{-1}$. A change of the occurrence of these uncertain low values cannot always be utilized to estimate a trend. Therefore, ϑ was assumed to remain constant for the future period as in *Pierce et al.* (2015). Estimating a trend of the precipitation probability p_w in the raw RCM simulations to adjust this threshold for future conditions is not always feasible as there might be no discernible trend because an RCM might have no zeros at all (e.g. the Hirham-EC-EARTH model). In such a case, p_w is 100% for the historical and future period and a change in p_w cannot be estimated. Calculating ϑ for the historical period and assuming its validity for future conditions did lead to the most stable results. Because of the strong bias of most GCM-RCM combinations, very high thresholds ($\vartheta > 15\ mm\ d^{-1}$) were necessary for certain months and locations.

3.6.4 Projected change of onset of rainy season

A Fuzzy Rule-based methodology (*Zadeh*, 1965) to define the onset time of the rainy season has been developed by *Laux* (2009). To assess how the climatology is projected to change after the bias correction, the time series of daily intensities were analyzed with the *Fuzzy Method 2* developed by *Laux* (2009). This method calculates the onset of the rainy season based on two criteria: The first fuzzy membership function $\mu_{F,1}$ evaluates the amount of precipitation that has fallen in a period of five days with a triangular function that increases from 0 to 1 over the support $\{18\frac{mm}{5d}, 25\frac{mm}{5d}\}$. The second function $\mu_{F,2}$ counts the number of wet days in this 5 days period and evaluates the membership with another triangular function with a support of $\{1, 3\}$, so a single wet value will lead to $\mu_{F,2} = 0$. The combined fuzzy rule response is then calculated as the product of $\mu_{F,1}$ and $\mu_{F,2}$. The first day of the year (DOY_{on}) which fulfills $\mu_{F,1}\mu_{F,2} \geq 0.4^{2/3}$ defines the onset of the rainy season.

The DOY_{on} was calculated for all bias-corrected models for the individual grid cells. As an example, a map of DOY_{on} is given for the model with the best performance in the historical period (Figure 3.32). The general structure is well met by this bias corrected time series. In the southern parts, the rainy season starts earlier in the year (approximately in mid-April) as the monsoon rain belt moves from south to north. In the northern parts, the Fuzzy Rule estimates the rainy season to start about two months later.

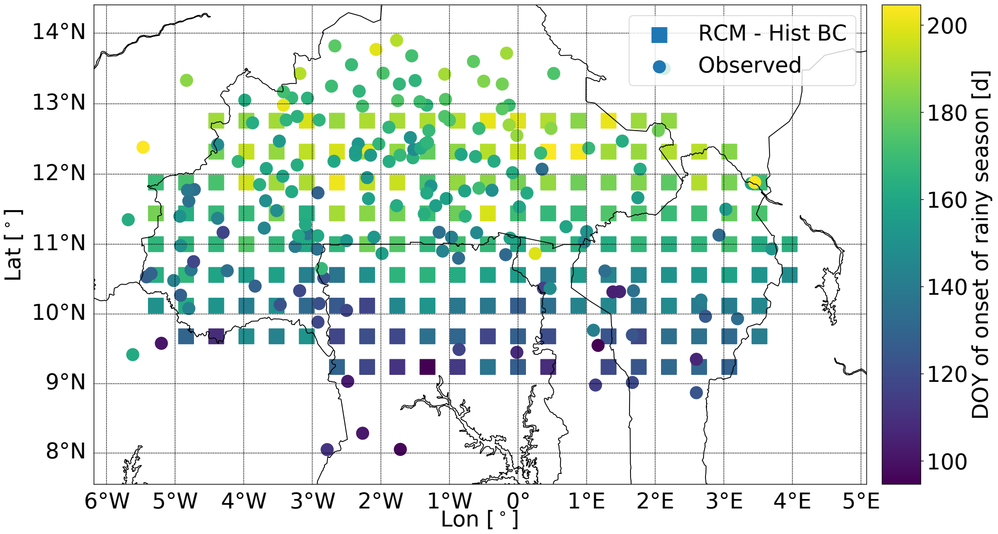

FIGURE 3.32: Map of observed (circles) and bias corrected (squares) DOY_{on} of the best model in the historical period 1970-2000.

To investigate if the rainy season is projected to start earlier or later in the future, the areal mean difference $\Delta_{DOY} = DOY_{on,fut} - DOY_{on,hist}$ was calculated for all models. The historical period covers the period 1970-2000 and the future was split up into the near future 2020-2050 and far future 2070-2100. The areal mean difference Δ_{DOY} of each model and RCP scenario as well as the average over all models of a certain RCP scenario are given in the violin plot in Figure 3.33.

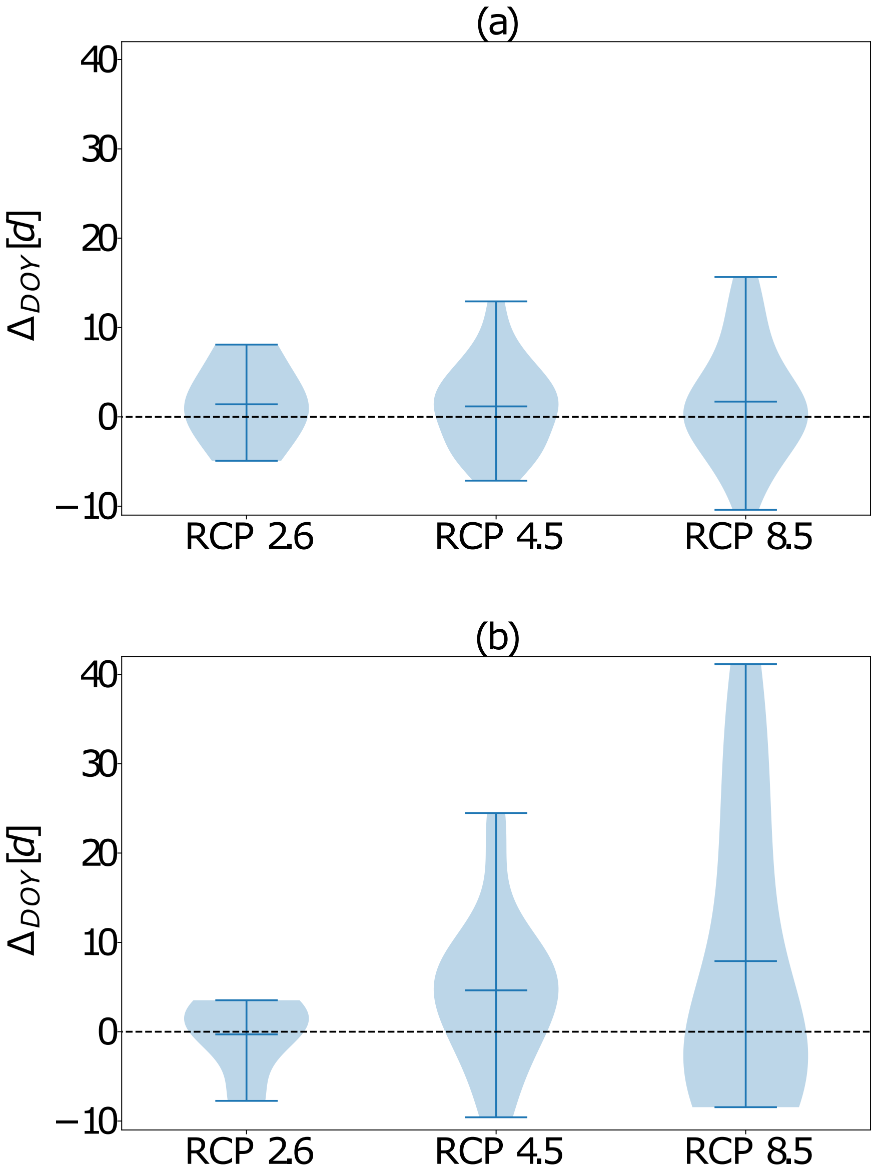

FIGURE 3.33: Violin plots of Δ_{DOY} of the bias corrected near future 2020-2050 (a) and far future 2070-2100 (b) in comparison to the historical period 1970-2000.

Depending on the RCP scenario and future period, the rainy season is projected to begin 0 to 8 days later in the future on average. While the average differences Δ_{DOY} of each RCP scenario are rather small, larger changes can be observed for individual models (between 10 days earlier and 41 days later). The different spreads of the violins is partially related to the number of available models in the different RCP scenarios (5 for RCP 2.6, 22 for RCP 4.5 and 21 for RCP 8.5). As the temperature and solar radiation is most likely increasing in the future, the fuzzy rule $\mu_{F,1}$ may need to be changed to take the higher potential evapotranspiration into consideration. Therefore, optimal planting dates are expected to occur later in the year even though the annual sum of precipitation is projected to increase (Figure 3.25). Maps of the average Δ_{DOY} can be found in Appendix A for the near (Figure A.3) and far (Figure A.4) future.

3.7 Summary and outlook

A geostatistical technique to estimate the distribution functions of daily precipitation of ungauged locations has been developed. The CDF parameters were kriged from station locations and were utilized to generate so called "simulated observations" to perform a *Double Quantile Mapping* of the precipitation time series of the CORDEX-Africa ensemble for a study region in West Africa.

The Kriging procedure is flexible since it estimates the CDF parameters of ungauged locations as a function of the distance to gauges. Different parametric CDF functions can be selected for other regions and meteorological variables. Another possibility to generate simulated observations would be to interpolate the quantiles of the observed ECDF as in *Mosthaf and Bárdossy* (2017) and then fit a parametric function to the interpolated quantiles. Alternatively, the CDF function, with which a daily precipitation amount is bias corrected, could depend on atmospheric circulation patterns and not on the month. All of these possible extensions depend on the available data and subsequent applications however. If the observed distributions are very skewed, higher-parametric CDFs might be necessary and thus more data is required to ensure that the fitting results in a stable estimate of the true distribution. On the other hand, the seasonality or the influence of circulation patterns might be dominant and that would require splitting the data set into many small sub samples which in turn impedes using a higher-parametric CDF or a quantile-based interpolation.

An analysis of the projected climate change after the bias correction revealed that West Africa will most likely experience higher rainfall amounts in the period 2006-2100. The onset of the rainy season however is not projected to change by a lot. If more water will be available for crop cultivation cannot be inferred from the projected precipitation time series. An analysis of several bias-corrected RCM simulations revealed for example a projected temperature increase between 1 and 2 K in the Black Volta basin for the near future 2019-2045 (*Kwakye*, 2016). Thus, increasing short wave radiation or temperature (and consequentially higher evapotranspiration) may counteract that higher precipitation amounts are expected in the future.

Chapter 4

Copula-based spatial downscaling of RCM precipitation[1]

This chapter describes a copula-based spatial downscaling model for RCM precipitation fields. A high spatial resolution is especially important in regions with complex orography and a high local variability of precipitation events. While the rather coarse resolution of 0.44° of the CORDEX simulations presented in Chapter 3 is sufficient to provide valuable local information in a comparatively flat region like West Africa, a higher resolution is required for a region like Central Europe. Dynamically downscaled precipitation fields can provide fine scale information for subsequent models and analyses but the computational demand is very high. Therefore, it would be helpful to generate fine scale precipitation maps that resemble RCM simulations in a faster way so that the fine scale RCM must not be run for the complete time period of interest. To this end, a copula-based downscaling technique has been developed. Nested daily RCM simulations for a region in Central Europe with the WRF model in two different spatial resolutions (7 km and 42 km) served as a training set to derive the statistics necessary to simulate fine scale precipitation values from the multivariate Gaussian Copula. The model was calibrated with RCM simulations for the year 1971 and the evaluation was performed for the period 1972-2000. The evaluation comprises the spatial correlation and statistical distributions of the simulated precipitation fields. Daily precipitation time series were analyzed by calculating Brier Skill Scores and the skill of reproducing the occurrence and amount of precipitation. An advantage of the developed model over deterministic downscaling techniques is that ensembles of predictand fields are generated 200 times faster than with the original fine scale RCM. Due to the generation of ensembles, the uncertainty that is inherent to downscaling can be estimated. The method has the potential to be used in other downscaling applications to generate ensembles of spatially correlated predictands based on other predictors. As copulas treat the dependence structure separately from the marginal distributions of the predictors and predictands, it is possible to simulate meteorological variables from any desired distribution function.

4.1 Overview of downscaling techniques

Two different approaches to increase the spatial resolution of GCM simulations are dynamical and statistical downscaling. Dynamical downscaling is performed with regional climate models for a chosen region. Statistical downscaling estimates a fine scale variable of interest (the so called predictand) based on the statistical dependence to coarse scale variables (predictors).

[1] Most contents of this chapter have been published in Hydrological Processes (*Lorenz et al.*, 2018). In this chapter, additional analyses are presented.

4.1.1 Dynamical downscaling

The advancement of processing power and more sophisticated physics schemes has enabled running regional climate models with increasingly finer spatial and temporal resolutions. The coarse scale information from a GCM is used as boundary conditions of spatially refined, physically-based simulations of the processes in the atmosphere. Further refinement can be accomplished by nesting even finer resolved RCM domains into the dynamically downscaled region.

There are many recent developments of high-resolution dynamical downscaling models which have shown to improve the simulated precipitation fields and distribution functions when compared with coarser RCMs. For instance, *Prein et al.* (2015) analyzed the added value of fine scale RCM simulations with a spatial resolution below 4 km. Such high resolution models allow the physically based simulation of convective precipitation which is usually parametrized in coarse RCMs. They compared the performance of RCMs with a spatial resolution ranging from 0.5 km to 4 km to coarser RCM simulations and found an added value in the representation of the diurnal cycle in summer and spatial patterns but not in the mean precipitation. The development of increasingly higher resolved RCMs is an area of research where new challenges have to be tackled at different scales (i.e. cloud parametrizations or new physics schemes). It can thus be expected that dynamical simulations in high spatial resolution will further improve in the future. Limiting factors are primarily the required process time and the storage demands.

4.1.2 Statistical downscaling

Statistical downscaling is faster than dynamical downscaling because the computational demand is lower. Also, because the predictands are typically observed values that are used for the calibration of the statistical downscaling model, the estimated variables tend to follow the observed distribution. The shortcoming in comparison to dynamical downscaling is that this method is not physically-based which complicates the application for future scenarios. Also, statistically downscaled fields can be too smooth in complex regions.

Many statistical downscaling techniques have been employed in the past to estimate fine scale fields of meteorological variables based on atmospheric predictors (*Wilby and Wigley*, 1997; *Maraun et al.*, 2010; *Yang et al.*, 2017). Depending on the chosen model, the downscaling is either deterministic or stochastic. A deterministic model produces a single realization of a predictand for a given set of predictors whereas a stochastic downscaling model generates multiple realizations of the predictand. A comparison of widely-used statistical downscaling techniques like multiple linear regression (MLR) or positive coefficient regression applied to meteorological predictor fields of a general circulation model was presented by *Goly et al.* (2014). Over the past years, several stochastic downscaling methods have been developed. *Gagnon and Rousseau* (2014) estimated the daily precipitation intensity of fine scale grid cells (4 km, 8 km and 12 km) based on the surrounding eight coarse scale grid cells of an RCM in a resolution of 45 km. The model incorporates anisotropy effects and wind speed, wind direction and the convective available potential energy as covariates. All fine scale cells at first obtain the value of their corresponding coarse scale cell. These values are then updated successively according to the distribution parameters that are expected based on the neighbors. *Mehrotra and Sharma* (2010) downscaled precipitation fields of a GCM with a Markov model that simulates conditional on atmospheric state variables, the past wetness state of each location and the fraction of wet values over an area around each location. *Allcroft and Glasbey* (2003) showed how spatially aggregated radar-based precipitation fields can be disaggregated back to the original resolution by transforming precipitation to a Gaussian variable via quadratic transformation. The starting point of their downscaling approach is the coarse scale value uniformly distributed across all corresponding fine scale cells. Those values are then updated with Gibbs Sampling (*Geman & Geman*, 1984). Fine scale precipitation is simulated as a Gaussian Markov Random Field that is conditioned on spatial and temporal neighbors until the fine scale values within a coarse cell agree to a prescribed threshold with the higher aggregated value. *Volosciuk et al.* (2017) developed a methodology to bias correct and downscale RCM simulations of precipitation. A gridded observation data set was used to correct the bias of the RCM fields and the downscaling to the station scale was achieved by estimating the fine scale distribution parameters with a vector generalized linear model.

In contrast to standard methods like multivariate regression, copulas model the dependence of a set of variables separately from the univariate distribution of the variables. Additionally, ensembles of an unknown variable can be simulated conditionally on predictors. A method based on bivariate copulas that generates realizations of fine scale precipitation conditional on a coarser precipitation value was presented by *van den Berg et al.* (2011). Their model was developed with radar-derived rain fields that were aggregated to a coarser spatial resolution. The simulated fine scale precipitation values are distributed randomly over the domain covered by each coarse scale cell. The methodology has also been adapted to simulate soil moisture values conditioned on satellite-derived observations in a coarser resolution (*Verhoest et al.*, 2015). Bivariate copulas have also been used to simulate ensembles of station-scale precipitation intensities based on radar fields (*Vogl et al.*, 2012) and RCM simulations (*Mao et al.*, 2015). Bias-corrected time series were attained by averaging the ensemble of realizations from the conditional distribution. *Ben Alaya et al.* (2014) developed a probabilistic regression model to estimate the parameters of the conditional distributions of minimum and maximum temperature, precipitation amount and occurrence at the station scale based on atmospheric predictors. The model was coupled with the Gaussian Copula to introduce spatial correlation of the predictands. The Random Mixing method presented by *Bárdossy and Hörning* (2015) generates spatial fields that are in accordance with linear and non-linear constraints as a combination of independent random fields that are simulated with the Gaussian copula. This method was applied by *Haese et al.* (2017) to simulate precipitation fields that are in accordance with rain gauge measurements and path-averaged rain rates of commercial microwave links.

Fine scale RCM precipitation shows an added value in comparison to coarse scale RCM precipitation which makes it an attractive predictand. Dynamically downscaled fine scale precipitation shows a high correlation to coarse scale precipitation and therefore coarse scale precipitation is a valuable predictor for fine scale precipitation. However, as it will be shown in the following, the relation is uncertain and large differences in the amount of precipitated water can be found for different sub-regions. The uncertain relation between coarse and fine scale RCM precipitation amounts is a property that cannot be replicated with techniques like interpolation, multiplicative random cascades (e.g. *Rupp et al.*, 2012) or the methods by *van den Berg et al.* (2011) and *Bárdossy and Hörning* (2015). Thus, a stochastic simulation technique is required to address this variability while respecting the spatial correlation of the estimated fine scale precipitation fields. The application of copulas for spatial downscaling is not very common, even though copulas offer a lot of flexibility for treating differently distributed variables, and therefore a copula-based model was sought. Also, downscaling with only one variable and staying within the realm of a single RCM in different resolutions is rarely done despite the steadily improving quality of fine scale RCM simulations.

4.2 Study region and RCM precipitation data

In this section, daily RCM precipitation in two different spatial resolutions for a domain that encompasses the majority of Germany, Austria and Switzerland and their surrounding countries, are presented. The daily precipitation simulations were performed by *Berg et al.* (2012) with WRF for the time period 1971 to 2000. The RCM was driven by 6-hourly ERA40-reanalysis data (*Uppala et al.*, 2005) for a coarse domain in a spatial resolution of 42 km (124 x 116 cells) and a fine scale domain in a spatial resolution of 7 km (174 x 174 cells). Figure 4.1 shows the total annual precipitation amount of the year 1971 for the coarse (a) and fine (b) RCM in the study area.

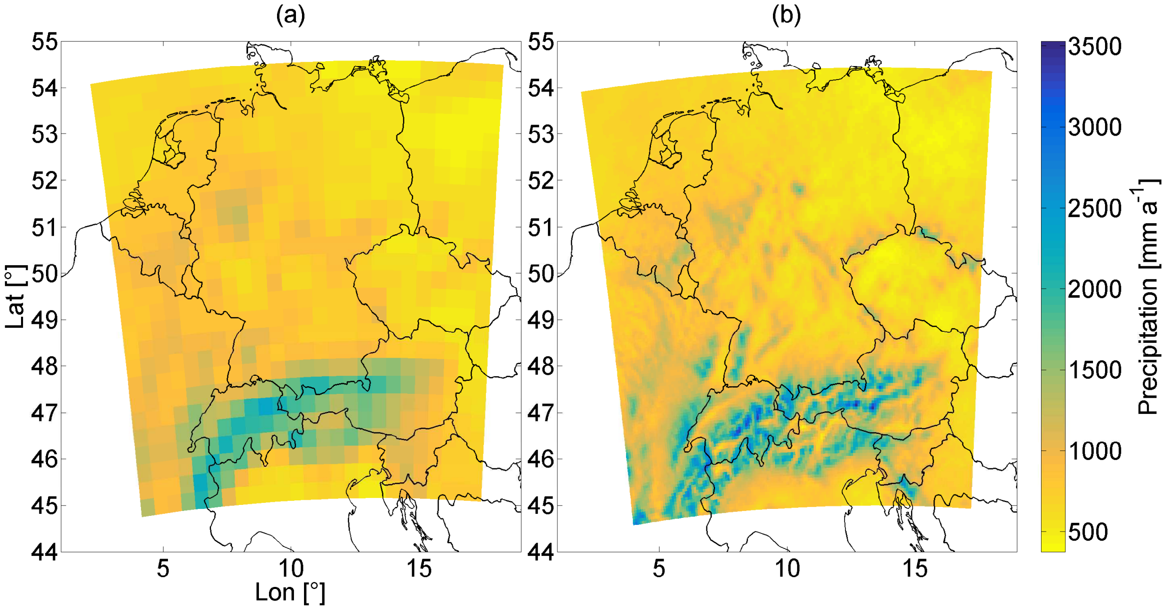

Figure 4.1: Total annual precipitation of the coarse scale (42 km, a) and the fine scale (7 km, b) RCM in the calibration year 1971.

Both model configurations led to nearly identical mean annual sums for this year (870.3 $mm\ a^{-1}$ in the 42 km simulations and 902.5 $mm\ a^{-1}$ in the 7 km simulations) but the fine scale simulations exhibit higher annual sums and more variation in many regions, e.g. in the Harz Mountains in Central Germany, the Black Forest in Germany, the Vosges and Jura Mountains in France and the Alpine region.

While the annual totals are similar, larger differences exist in the daily precipitation fields. The dynamical simulation of fine scale precipitation is not driven by the coarse scale RCM's precipitation but its state variables in the atmosphere. Precipitation is an output variable of the RCM and therefore, the coarse scale precipitation has no influence on the precipitation of the fine scale simulations. Nevertheless, it is reasonable to assume a statistical dependence between the precipitation amount in a coarse scale RCM's cell and its corresponding fine scale RCM cells' values. Figure 4.2 shows that there is a strong similarity of the daily precipitation patterns of the coarse (a) and fine (b) scale RCM for a heavy precipitation event on November 17, 1972.

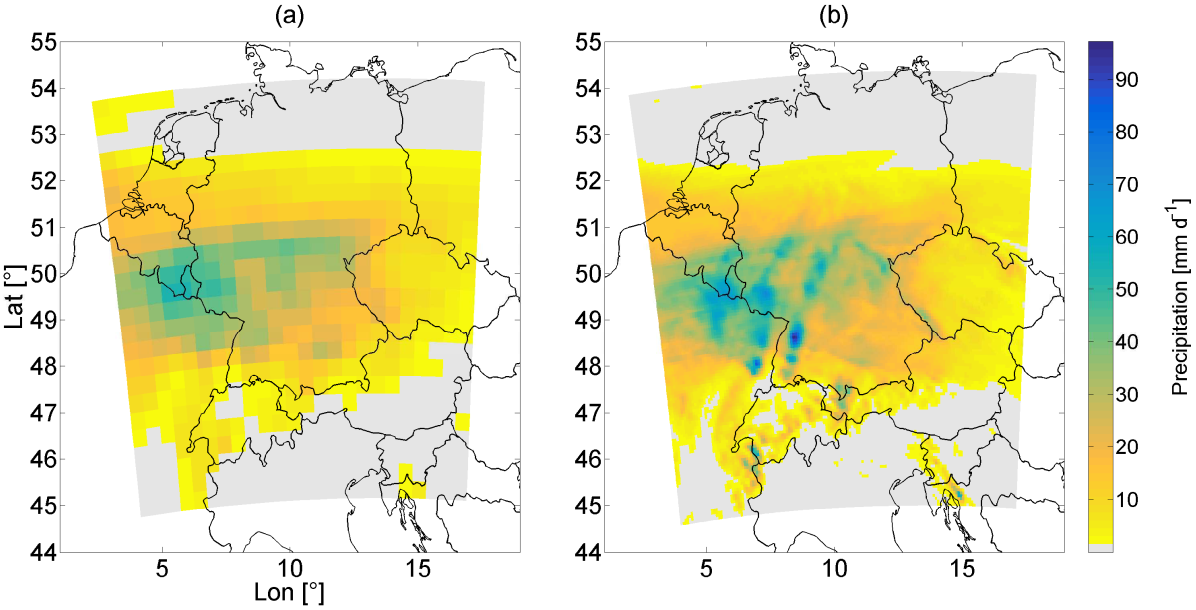

FIGURE 4.2: Example of daily precipitation patterns of the coarse scale (a) and fine scale RCM simulations (b) for a heavy precipitation event on November 17, 1972.

The main patterns of the precipitation fields are similar but large differences between the fine and coarse simulations can occur as well. One such case can be seen in Figure 4.2 in the northwest. While several coarse cells are wet, there are mostly zero precipitation amounts in the fine scale field. The opposite case can be seen in south-west Germany where the amounts in the fine scale simulations are much higher.

To demonstrate these differences of the RCM simulations, the daily precipitation amounts in the calibration year 1971 of the coarse scale simulations and the average precipitation of the corresponding 36 fine scale cells are shown in Figure 4.3. While there is a positive dependence between the coarse and fine scale precipitation amounts, there are also cases where one amount is very low and the other one is rather high. The relation between coarse and fine scale precipitation is therefore uncertain.

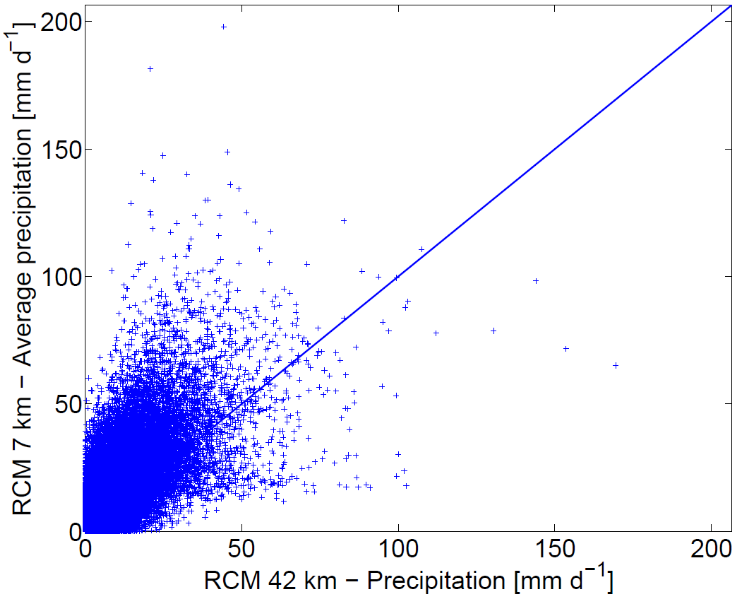

FIGURE 4.3: Scatter plot of daily precipitation of the coarse cells (x-axis) against average daily precipitation of the fine scale sub-fields (y-axis) in the calibration year 1971 for the complete study region.

4.3 Development of a stochastic copula-based downscaling model

Since a high amount in the coarse scale field can lead to high and low fine scale amounts and vice versa, the prediction of fine scale precipitation from coarse scale precipitation is uncertain. Therefore, a stochastic model that generates ensembles of fine scale precipitation fields is required to address the uncertainty of the dynamical downscaling with a stochastic method. The need for stochastic downscaling methods has also been stated by *Maraun* (2016). If the fine scale RCM can only be run for a short time period, the calibration of a downscaling is based on this limited time period which calls for a parsimonious and robust model. That fine scale fields are only available for a shorter period than coarse scale fields is a common problem in hydrology. One such case would be that the fine scale fields stem from a comparatively short measurement period (e.g. a recently set-up radar that has only been operational for one year) and that longer time series are only available for coarser observations (e.g. satellite data that has been recorded over several years). The challenge is to stochastically generate ensembles of fine scale precipitation fields which respect the spatial correlation structure and the inherent variability of the dynamically downscaled precipitation fields with a parsimonious model.

In this section, a novel copula-based downscaling model for the simulation ensembles of fine scale precipitation fields is presented. Precipitation is simulated for the fine scale domain by conditioning the simulation of each fine scale value on the surrounding coarse scale precipitation amounts with the Gaussian Copula (see Chapter 2, section 2.5). To set up the copula density c, distribution functions $F(x)$ are required to transform precipitation x to CDF values u. The Gaussian copula is defined by a correlation matrix $\mathbf{\Gamma}$. Fitting a parametric correlogram model $\widehat{\rho}$ allows setting up the correlation matrix $\mathbf{\Gamma}$ of an arbitrary set of points in space that either belong to the coarse or the fine scale domain. The correlograms were calculated for both scales. Due to a copula-based simulation of predictand fields, an ensemble of possible realizations is obtained which allows an estimation of the uncertainty of downscaling meteorological variables. Contrary to other downscaling techniques, the transfer function of predictors and predictands is not fitted for each location separately because the correlation matrices of sets of points are estimated based on the separation distance. The method was developed with RCM simulations as in *Gaitan et al.* (2014) and *Chen et al.* (2013), but it can be adapted to real observations of other meteorological variables. A summarizing flow chart of the simulation process is given in Figure 4.4.

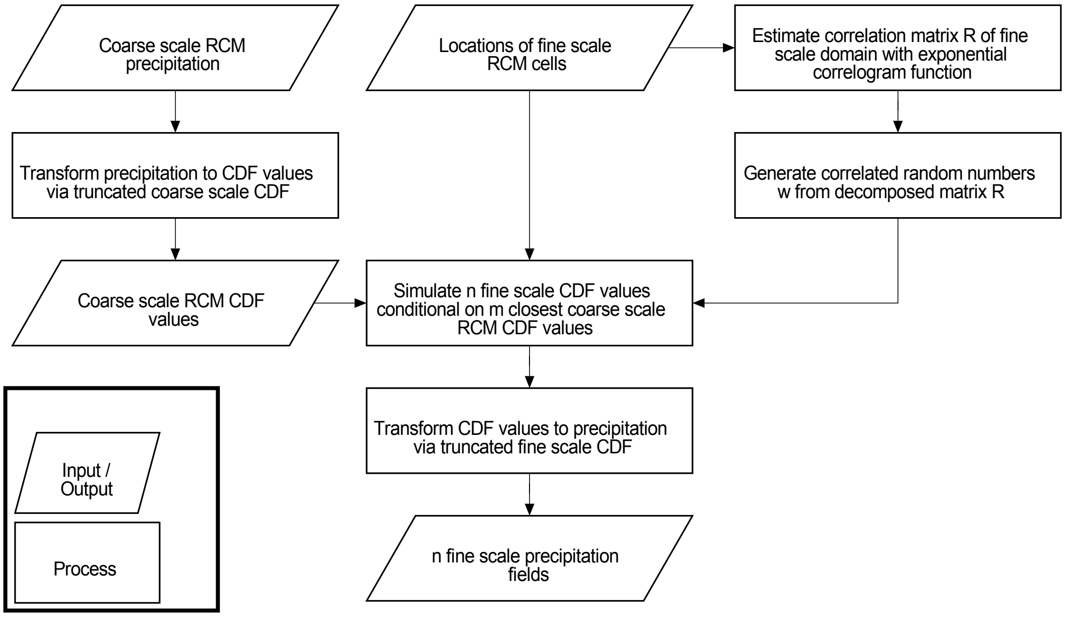
Coarse scale RCM precipitation
Locations of fine scale RCM cells
Estimate correlation matrix R of fine scale domain with exponential correlogram function
Transform precipitation to CDF values via truncated coarse scale CDF
Generate correlated random numbers w from decomposed matrix R
Coarse scale RCM CDF values
Simulate n fine scale CDF values conditional on m closest coarse scale RCM CDF values
Input / Output
Process
Transform CDF values to precipitation via truncated fine scale CDF
n fine scale precipitation fields

FIGURE 4.4: Flowchart of the simulation of fine scale precipitation fields.

The CDF value of fine scale precipitation u_1 was simulated conditionally on the precipitation amounts of $m = n - 1 = 4$ coarse scale neighbors $(u_2, ..., u_n)$. The conditional distribution from which fine scale precipitation was simulated was obtained from the Gaussian Copula density $c(u_1, ..., u_n)$ of the unknown fine scale point and the coarse conditioning values.

$$c(u_1, ..., u_n) = \frac{1}{\prod_{i=1}^{n} \phi(\Phi^{-1}(u_i))} \frac{1}{(2\Pi)^{\frac{n}{2}} \sqrt{det(\mathbf{\Gamma})}} e^{-0.5(\Phi^{-1}(\mathbf{u})^T(\mathbf{\Gamma}^{-1}-\mathbf{I})\Phi^{-1}(\mathbf{u}))} \tag{4.1}$$

The Gaussian copula was chosen because it can deal with any number of conditioning values allowing for smooth fine-scale precipitation fields that exhibit spatial correlation similar to the original RCM fields.
The unknown correlation matrix $\mathbf{\Gamma}$ was modeled with two correlogram models. A correlogram model $\widehat{\rho_{cc}}$ was fitted to the empirical correlation coefficients in different separation distances ρ^*_{cc} of the coarse RCM precipitation fields. Likewise, a correlogram model $\widehat{\rho_{fc}}$ was built to estimate the correlation between fine scale and coarse scale precipitation values in a certain separation distance.

From the copula density $c(u_1, ..., u_n)$, the conditional PDF $f_c(u_1|u_2, ..., u_n)$ was built and integrated to the conditional CDF F_c. A fine scale CDF value u_1 was simulated by inverting this conditional CDF with a uniformly distributed random number w. In a last step, u_1 is transformed to precipitation with its CDF $F(x)$.

$$u_1 = F_c^{-1}(w), w \sim U(0, 1). \tag{4.2}$$

The random numbers w were simulated from a generating correlation matrix $\mathbf{R}$ because independent random numbers were found to lead to an underestimation of the fine scale correlation of the downscaled fields. To generate spatially-correlated w fields, a Cholesky decomposition of $\mathbf{R}$ into the triangular matrix $\mathbf{L}$ was performed. $\mathbf{R}$ was calculated with the correlogram model $\widehat{\rho_{ff}}$ of fine scale precipitation to ensure that $\mathbf{R}$ is positive-definite.

$$\mathbf{R} = \mathbf{L}\mathbf{L}^T. \tag{4.3}$$

Independent uniformly distributed random numbers $u \sim U(0, 1)$ for all fine cells in the complete domain were recorrelated via the triangular matrix $\mathbf{L}$ to obtain 50 w fields for the fine scale domain. With these 50 fields, 50 precipitation fields were simulated for each day to address the uncertainty of the downscaling process.

$$w = \Phi(L\Phi^{-1}(u)) \tag{4.4}$$

Wilks (1999) simulated the rainfall amounts on wet days from a parametric CDF with spatially-correlated random numbers. In this case, w were used to invert the conditional distribution to obtain CDF values.

The conditioning process is illustrated in Figure 4.5. For the blue fine scale cell on the left, the conditional distribution $F_c(u_l|0.90, 0.85, 0.98, 0.95)$ is calculated to obtain CDF realizations u_l. For the orange cell on the right, the conditional distribution is $F_c(u_r|0.90, 0.85, 0.70, 0.75)$.

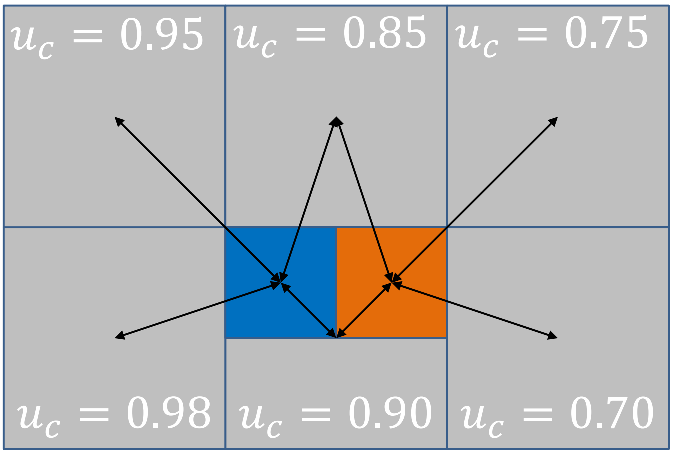

FIGURE 4.5: Artificial example of the conditioning of two fine scale value with the four closest coarse scale values.

It is expected that a realization from the conditional distribution F_c of the blue cell will be larger because its conditioning values are larger. This is illustrated in Figure 4.6. The same random number $w = 0.5$ leads to a higher CDF value u_1 for the left cell. By conditioning on four coarse scale neighbors, the general spatial pattern of the coarse scale field is taken into account, whereas conditioning on only the closest neighbor ($u_c = 0.90$) would result in identical conditional CDFs for the two fine cells.

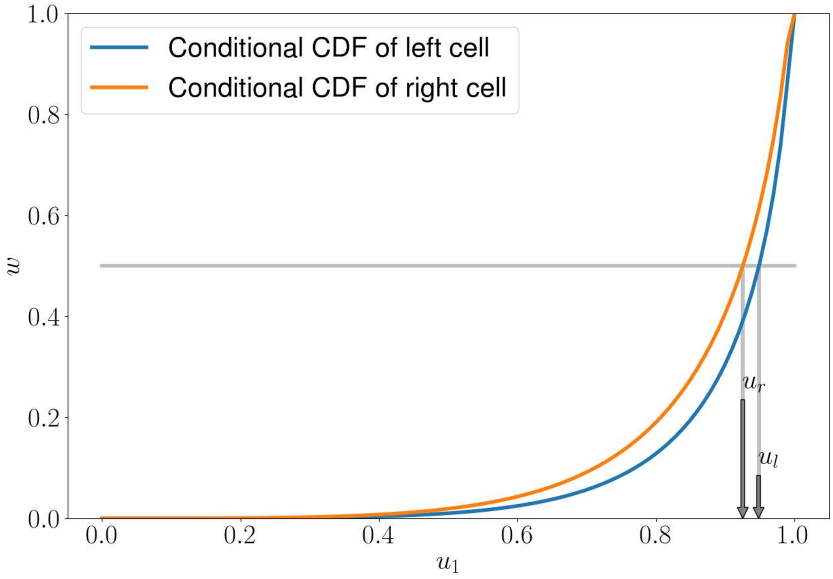

FIGURE 4.6: Conditional CDF of the two fine scale cells given their four conditioning values in the artificial example.

4.4 Calibration of the stochastic copula-based downscaling model

The calibration was carried out for the year 1971 even though fine scale RCM simulations were available for the period 1971-2000. With the derived statistics, stochastic simulations for the evaluation period 1972 to 2000 were performed. This approach was chosen because the method was developed to provide a surrogate for fine scale RCM precipitation which can only be simulated for a limited time period.

Fitting the stochastic model consists of four subsequent steps:

1. Estimating the parameters of cumulative distribution functions (CDFs) $F(x)$ for all fine and coarse scale cells,
2. Transforming precipitation amounts via the fitted CDFs to ranks $u = F^{-1}(x)$,
3. Calculating empirical correlograms ρ^* using the CDF values u of precipitation for pairs of points of different separation distances, and
4. Fitting parametric correlogram functions $\widehat{\rho}$ to the empirical correlograms ρ^* in order to obtain estimates of the correlation matrix of arbitrary sets of points in space.

4.4.1 Distribution functions

Following the approach by *Bárdossy and Pegram* (2012), the precipitation values were separated into one branch for dry values which obtain half the dry probability p_0 and one branch for positive values which were fitted with a CDF $F(x)$. This makes it possible to transform dry and wet values x to CDF values u with a single truncated distribution.

$$u(x) = \begin{cases} \frac{p_0}{2} & \text{if } x = 0 \\ p_0 + (1 - p_0)F(x) & \text{if } x > 0 \end{cases} \tag{4.5}$$

A parametric distribution function $F(x)$ was selected by calculating the BIC (see Chapter 2, 2.2.3) of nine CDF functions. Table 4.1 shows the average BIC values of the tested distribution functions.

Distribution function	RCM 7 km	RCM 42 km
Exponential	1093.5	1129.0
Gamma	979.9	997.2
Generalized Extreme Value	1914.6	3641.8
Generalized Pareto	1007.9	1046.4
Logistic	1441.8	1490.6
Log-Logistic	997.1	1031.2
Log-Normal	996.8	1034.8
Rayleigh	2042.5	2121.0
Weibull	972.3	994.3

TABLE 4.1: Average BIC-value of nine tested parametric distribution functions in the calibration period 1971.

While the Weibull distribution resulted in slightly lower BIC-values than the Gamma distribution, the Gamma distribution was selected as $F(x)$ because a scaling relation was employed to estimate the CDF parameters of the 30 year long time series based on the parameters of the calibration period. In *van den Berg et al.* (2011) the relation of the scale parameter to a change of the spatial scale was used to estimate the scale parameter of fine scale cells. The Gamma distribution is a common choice for modeling daily precipitation amounts and has been used in other studies to model daily precipitation intensities of RCM simulations (*Piani et al.*, 2010; *Tschöke et al.*, 2017; *Wetterhall et al.*, 2012). Other distribution functions can be used for other meteorological variables because copulas are flexible and can use different CDF functions as they measure the dependence of several variables separately from the univariate marginal distributions.

The Gamma distribution uses the gamma function Γ and is defined by a shape parameter k and a scale parameter θ. Its probability density function (PDF) $f(x)$ is defined as:

$$f(x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}}. \tag{4.6}$$

Analyses by *Prein et al.* (2013) of WRF simulations in spatial resolutions of 36 km, 12 km and 4 km showed that the finer resolved models capture the higher precipitation intensities better. As the distribution of precipitation amounts is dependent on the spatial resolution, separate parameter sets (k, θ) for the coarse and fine scale values are necessary. As an example, the empirical CDF of precipitation amounts for the location (6.49° E, 47.57° N) is shown in Figure 4.7. The fine scale precipitation is characterized by a slightly higher dry probability and higher extremes.

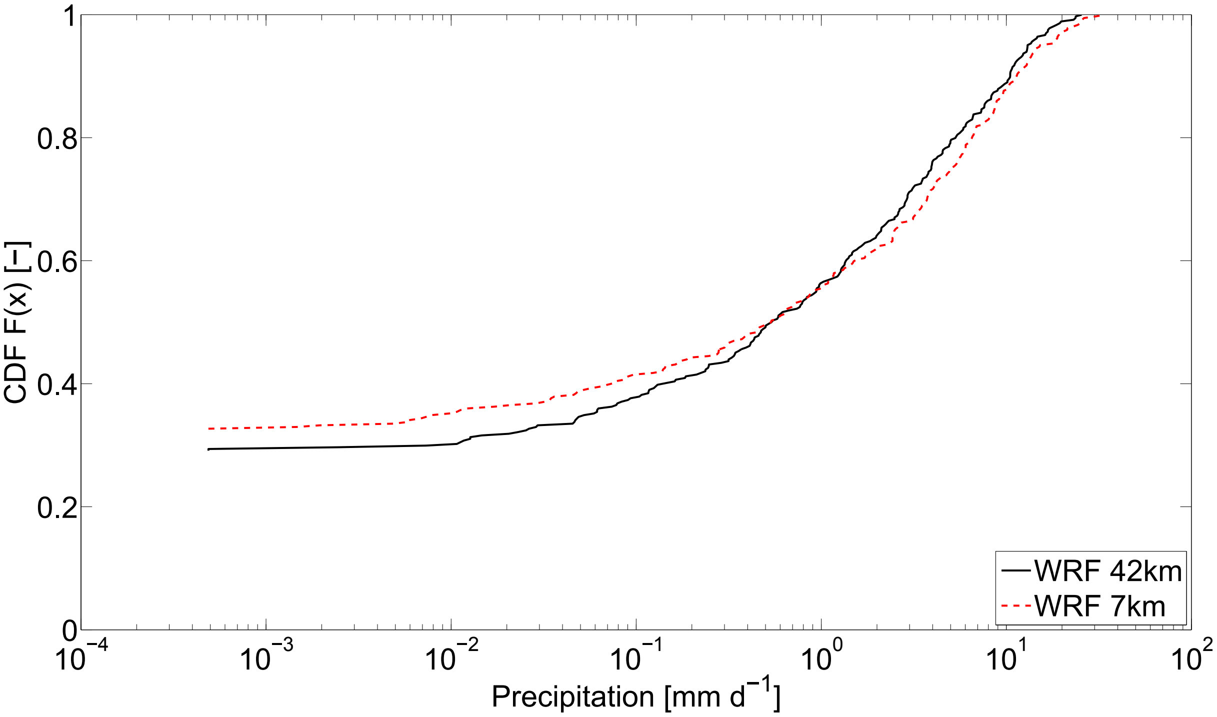

FIGURE 4.7: Empirical CDF of daily precipitation in the calibration period (1971) of one coarse scale cell and its central fine scale cell.

Under the given premises, the fine scale simulations were assumed to be only available for a very short period (1971). Instead of fitting the CDF parameters of the coarse and fine simulations to a 30 year long time series, an estimation of the fine-scale CDF parameters of the truncated Gamma distribution (p_0, k, θ) of the complete period 1971-2000 was necessary. The dry probability p_0 was calculated for each coarse and fine scale cell using the precipitation information from the calibration period 1971. In addition, the parameters k and θ were calculated for the calibration period 1971 by fitting the Gamma distribution to the positive precipitation amounts. For the complete period, stationarity of the dry probability p_0 and the shape parameter k was assumed. The scale parameter θ of the Gamma distribution was fitted to the coarse scale precipitation amounts of the complete period. On average, θ increased by 15% if 30 years were considered instead of only one year (Figure 4.8 (a)). The scaling behavior of the fine scale simulations was then assumed to be identical to the coarse scale simulations. In the present case, this assumption can be investigated: the scale parameter of the fine scale simulations increased by 18% (Figure 4.8 (b)) which is similar to the coarse scale simulations' scaling behavior.

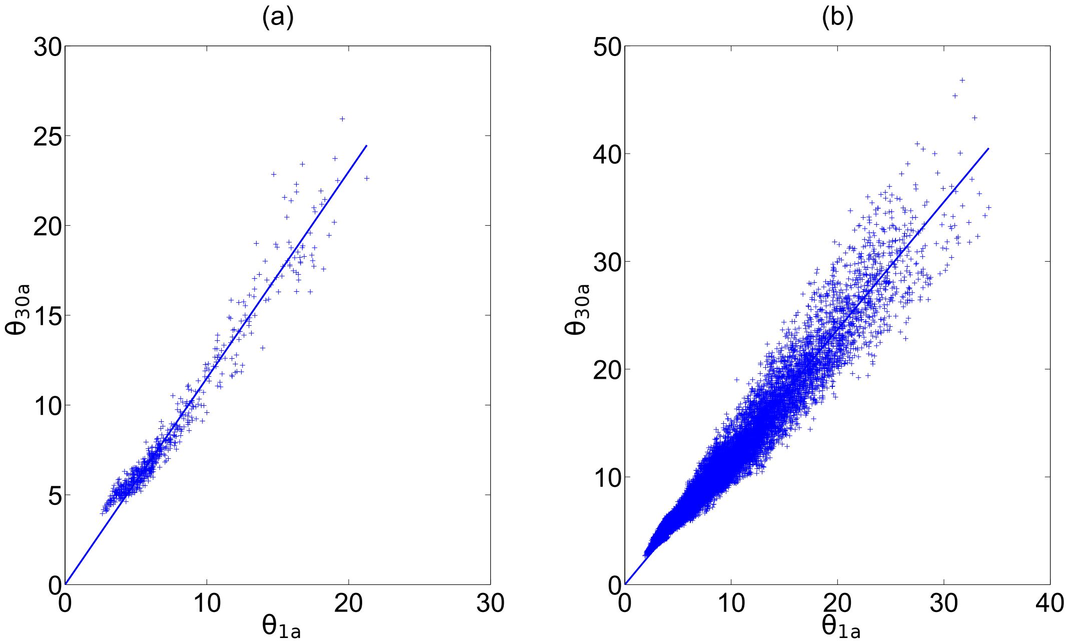

FIGURE 4.8: Scaling behavior of θ with a fixed shape parameter k from the 1-year calibration period to the complete 30-year period for the coarse scale (a) and fine scale RCM precipitation amounts (b).

The scaling assumption may not hold in some cases but in order to downscale a time series of coarse meteorological fields that is longer than the observation period of the fine scale domain, an estimation of the CDF of the longer period is necessary.

4.4.2 Spatial correlograms

The empirical correlation ρ^* of precipitation was calculated for ten separation distances h to estimate a correlogram model $\widehat{\rho}$ with Equation 2.30 as presented in Chapter 2, subsection 2.3. Precipitation time series were transformed to vectors of CDF values u and v with Equation 4.5.

An exponential correlogram model $\widehat{\rho}(h)$ with parameters r_0 and λ was fitted to the ten value pairs of the empirical correlogram by minimizing the squared differences between $\widehat{\rho}$ and ρ^*:

$$\widehat{\rho}(h) = r_0 e^{-\lambda h}. \tag{4.7}$$

The parameter r_0 is the correlation for a separation distance of 0, while λ describes how fast the correlation decays over distance. With this estimator of the correlation as a function of the separation distance, it is possible to estimate the correlation matrix of points that are separated by arbitrary distances.

The stochastic downscaling model requires correlation matrices $\mathbf{\Gamma}$ for point sets with points from the coarse and fine scale simulations. A separate consideration of pairs of values coming from different scales (coarse-coarse $\widehat{\rho}_{cc}$, fine-coarse $\widehat{\rho}_{fc}$ and fine-fine $\widehat{\rho}_{ff}$) is necessary because values from the same scale exhibit higher spatial correlation. This effect is due to the fine scale precipitation being an output variable of the RCM. Thus, the fine scale values are not directly governed by the coarse scale values even though there is correlation present.

This behavior can be seen in Figure 4.9. The correlation of fine scale values and coarse scale values with grid cell centers close to one another amounts to only about 0.83 (a), while the correlation for small separation distances converges towards 1 for fine scale values (b).

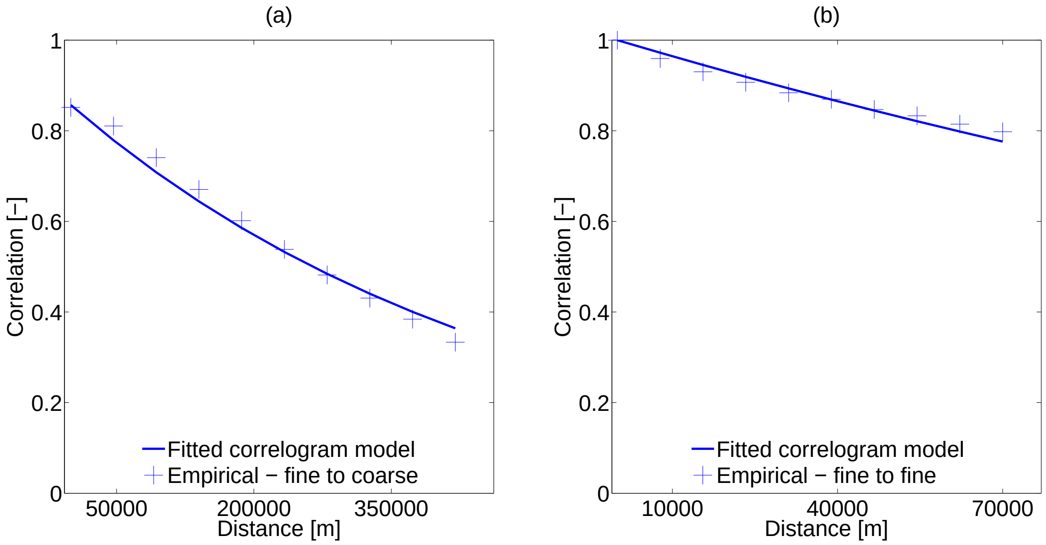

FIGURE 4.9: Empirical and fitted correlograms of fine scale RCM to coarse scale RCM (a) and fine scale RCM to fine scale RCM (b).

The CDF values of precipitation were obtained via the truncated Gamma distribution. This approach is sometimes referred to as 'inference from margins' (IFM). Alternatively, the empirical ranks ('pseudo-observations') could be employed. The suitability of the IFM-method is discussed in *Genest and Favre* (2007) and it is stated that it depends on the goodness of fit of the parametric CDF. A calculation of the correlation coefficients $\rho^*(h)$ with empirical ranks led to nearly identical values (not shown). As the proposed method utilizes a parametric CDF for the simulation, the IFM-method was employed.

Since daily precipitation exhibits temporal correlation, the influence of this auto-correlation on the spatial correlograms was analyzed. The auto-correlation of daily precipitation has decayed to zero after 10 days and the spatial correlation coefficients $\rho^*(h)$ were calculated from a reduced subset of the complete sample by only using every tenth day. The resulting correlograms closely match the ones obtained from using the complete sample and therefore the complete sample was used to calibrate the model.

4.5 Evaluation of downscaled precipitation

As the coarse cells were downscaled stochastically, every simulated precipitation field is different even though the coarse scale values remained constant. Therefore, a statistical comparison is necessary to investigate the general behavior of the simulated fields. In order to assess how well the stochastic simulation technique is able to reproduce the physical fine scale simulations, the spatial correlations and distribution functions of the stochastic simulations were investigated. In addition, the simulation technique was evaluated by calculating the ratio of correctly simulated wet and dry days, the explained variance of the amounts and Brier skill scores for ten thresholds. Additionally, different alternative model structures were employed to demonstrate the influence of model components.

4.5.1 Daily precipitation fields

A visual comparison of the generated precipitation fields for the day already shown in Figure 4.2 is given to illustrate some of the general characteristics (Figure 4.10). The areal mean precipitation of the downscaled region amounts to 9.24 $mm\ d^{-1}$ in the coarse scale simulation (a) and 10.27 $mm\ d^{-1}$ in the fine scale simulation (b). The areal mean of the 50 stochastic realizations was calculated and the ensemble member with median areal mean precipitation (9.94 $mm\ d^{-1}$) is shown in panel (c) to illustrate the general tendency of the stochastic simulations. To put the performance into perspective, the coarse scale precipitation fields were interpolated to the fine scale domain with Inverse Distance Weighting (IDW) from the four closest coarse scale cells (d). This interpolated field is too smooth and the precipitation amounts can only lie between the minimum and maximum of the coarse scale field. The copula-based technique on the other hand utilizes the fine scale CDF to simulate precipitation and larger values can be generated. In the presented case, the predictor and predictand variables are both precipitation. With copulas, meteorological variables with different distributions can be utilized easily since the dependence is modeled separately from the distribution functions.

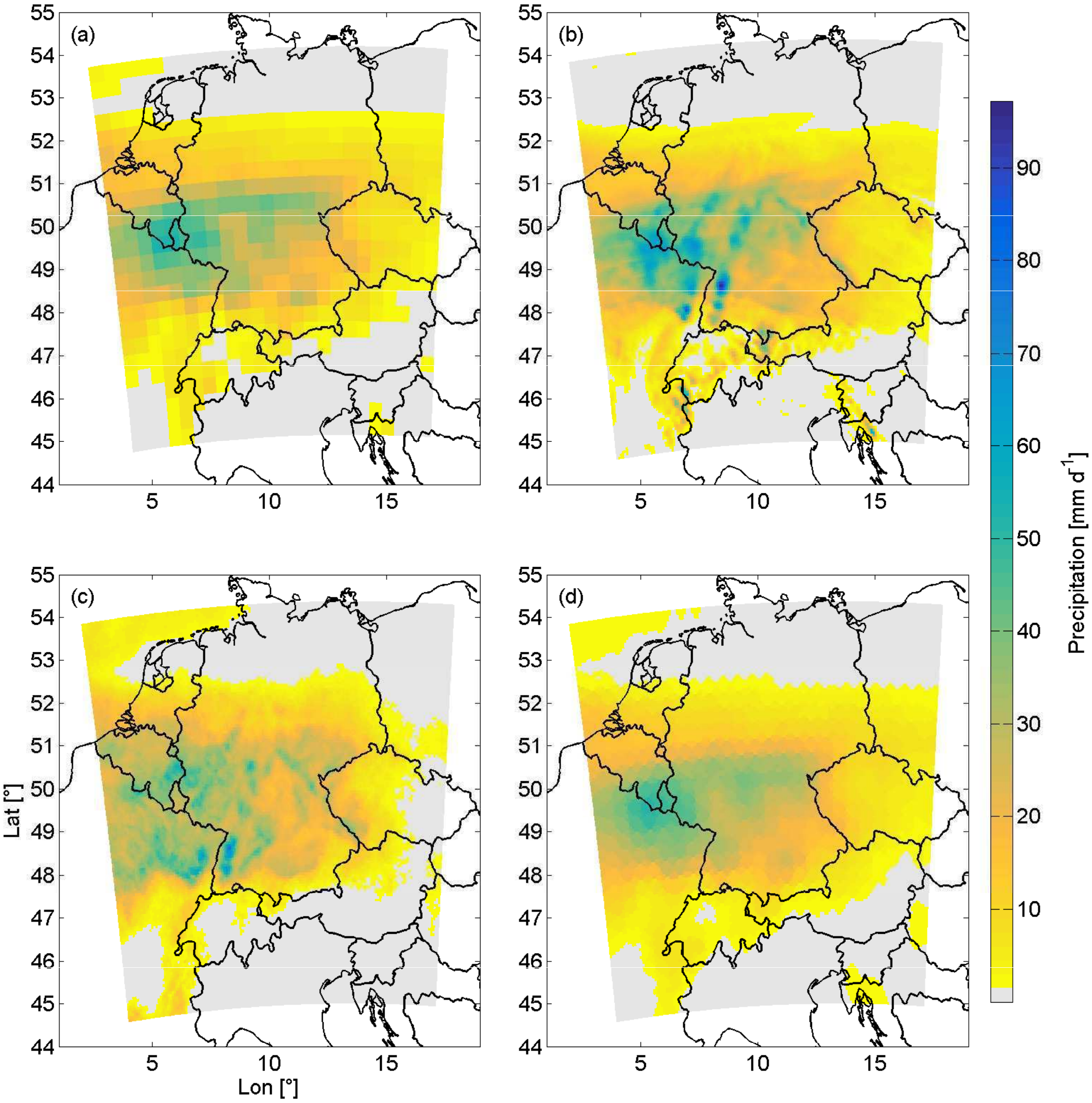

Figure 4.10: Precipitation on November 17, 1972 - Original coarse scale RCM (a), original fine scale RCM (b), stochastic simulation with median areal mean precipitation (c) and IDW interpolation of coarse scale precipitation (d).

The stochastic downscaling produces an ensemble of different realizations from a constant predictor field to address the uncertainty of the prediction shown in Figure 4.3. Figure 4.11 (a) shows the difference between the 90%-Quantile and the 10%-Quantile of each fine scale cell to demonstrate the spread of the ensemble. In Figure 4.11 (b), the mean difference of all ensemble members to the fine scale RCM field of this day is shown.

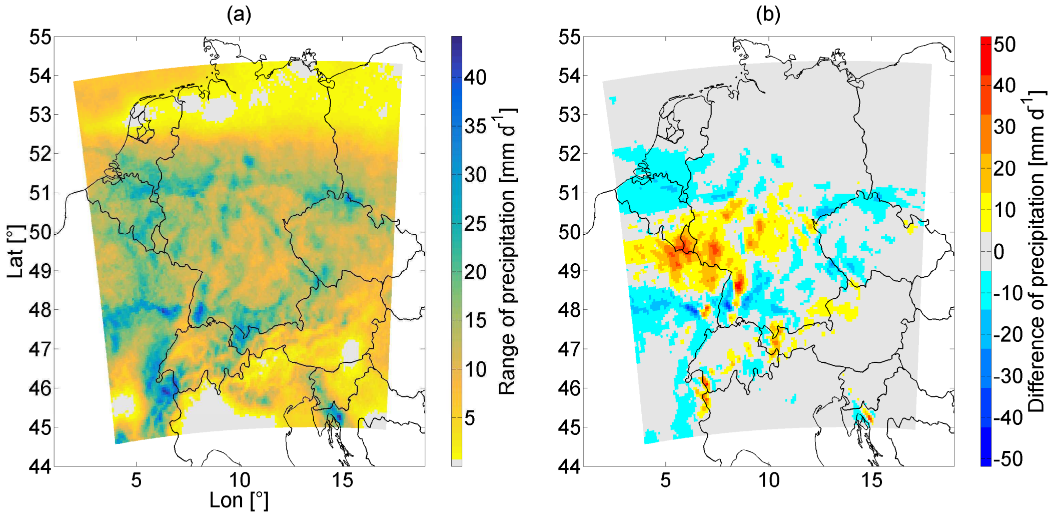

FIGURE 4.11: Stochastically simulated precipitation for November 17, 1972 - Spread of 10% and 90%-quantiles of stochastic simulations (a), mean difference of fine scale RCM precipitation field and stochastic simulations (b).

The differences are caused by the conditioning values, the estimated distribution functions and geographical effects. The coarse scale simulations serve as the only predictor. Because there is a positive precipitation amount in the northwest over the North Sea, the downscaling technique tends to generate positive amounts for this region as well. In the fine scale simulations this region is dry however. The largest differences can be seen in the region around Luxembourg where the fine scale RCM simulated high precipitation amounts.

In general, the stochastic precipitation fields show similarities to the RCM simulations. In the RCM simulations, the fine scale precipitation amounts can differ greatly from the coarse scale amounts (Figure 4.3) and this is reflected in the stochastic simulations. In the majority of other downscaling techniques, the statistical dependence of a set of points in space is calculated for this specific set (i.e., the transfer function between predictors and predictand for specific locations). In contrast, the presented method models the dependence as a function of the separation distance. A feature of the RCM simulations that cannot be captured exactly by an algorithm based on geostatistical dependence is that clouds move along a certain trajectory and precipitation can thus exhibit anisotropic dependence in space and locally isolated high amounts.

4.5.2 Spatial correlation of stochastic simulations

An evaluation of the representation of the spatial dependence was performed by calculating the correlograms of the dynamical and stochastic precipitation simulations for the first year of the evaluation period 1972 (Figure 4.12). The stochastic simulations' correlation of the fine to the coarse scale is slightly underestimated compared to the RCM simulations (a), whereas the fine scale correlation is overestimated (b).

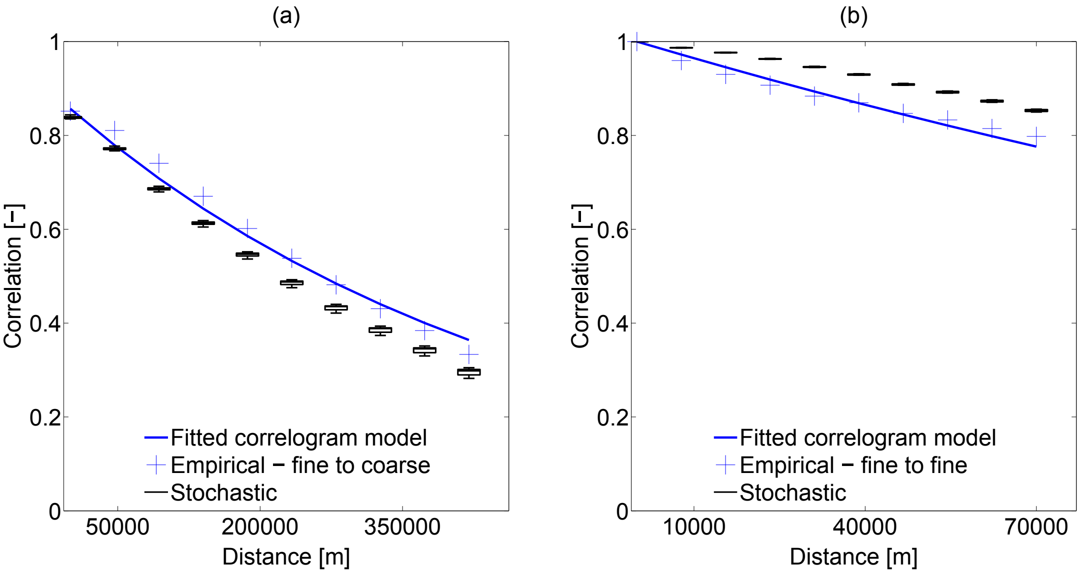

Figure 4.12: Correlograms of five stochastic ensemble members and the dynamical simulations for the fine to coarse scale values (a) and for the fine to fine scale values (b) in the year 1972.

This effect is due to a trade-off in the model structure. The maximum correlation of fine and coarse scale precipitation amounts to 0.83 (Figure 4.9 (a)) for separation distances close to zero. Thus, two neighboring fine scale cells can attain largely differing amounts if the simulation is performed independently.

During the model development, other model structures were utilized. For instance, an independent simulation of fine scale precipitation with only the closest coarse scale CDF value as conditioning value like in *van den Berg et al.* (2011) was tested. This approach resulted in noisy fine scale fields with visible jumps at the edges of coarse scale cells. The fine scale correlogram dropped to approximately 0.73 for adjacent cells with this approach (not shown). An extension to $m = 4$ conditioning values improved the fine scale correlograms of the stochastic precipitation fields slightly.

In order to strengthen the fine scale correlation of the predictands, the random numbers w were simulated with a correlation matrix that was calculated from the observed fine scale correlogram model. Further tuning of this generating matrix could potentially be achieved by using the approach of *Wilks* (1999) or *Brissette et al.* (2007), but as one correlogram is underestimated and the other one is overestimated, further improvements may not be attainable in the presented case.

4.5.3 Distribution of stochastic simulations

CDF-Parameters were estimated from fine and coarse scale simulations in the year 1971 and the CDF parameters of the evaluation period 1972-2000 were estimated by multiplying θ by 1.15. Since fine scale RCM simulations are available for the complete period, it was possible to investigate the influence of the calibration year. The standardized anomalies of the total annual precipitation, the mean amount on wet days and the rainfall probability are illustrated in Figure 4.13 for the coarse (a) and fine (b) scale WRF simulations.

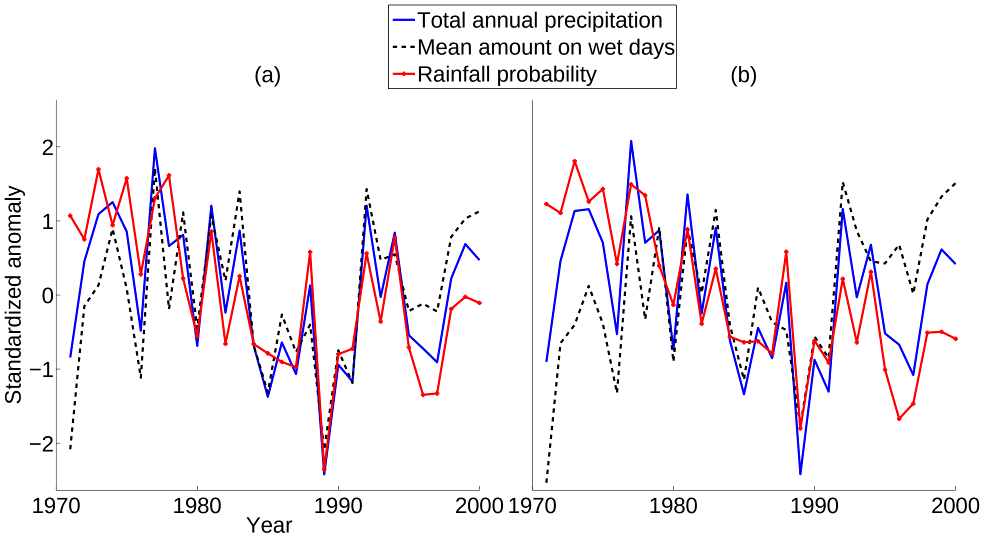

FIGURE 4.13: Standardized anomalies of the total annual precipitation, the mean amount on wet days and the rainfall probability for the coarse (a) and fine (b) scale WRF simulations

The annual total of the calibration year 1971 was only slightly below the average. The rainfall probability however was very high and the mean amount on wet days was very low. The year 1982 exhibits low anomalies for both scales. To test the robustness of the model, the downscaling model was also calibrated for the year 1982 and the remaining 29 years were downscaled.

The distribution of daily precipitation was evaluated by generating Quantile-Quantile-Plots of the downscaled precipitation fields against the fine scale RCM simulations. Because the simulated data set is very large (50 samples × 150 cells × 150 cells × 10958 days result in 1.23 10^{10} values), each sample was sorted in ascending order and 10^6 rank-equidistant values were extracted for each sample to make the comparison manageable.
The QQ-Plot of the stochastic simulations with 1971 as the calibration year (Figure 4.14 a) shows that the distribution functions of the dynamical and stochastic fine scale simulations are quite similar. However, the lower precipitation intensities were overestimated and the extreme values were underestimated due to the fitted parametric distribution but the majority of values lies relatively close to the bisecting line. To investigate the influence of the calibration year on the stochastic simulations, the calibration was performed with the less anomalous year 1982 but the QQ-Plot does not differ by a lot (Figure 4.14 b).

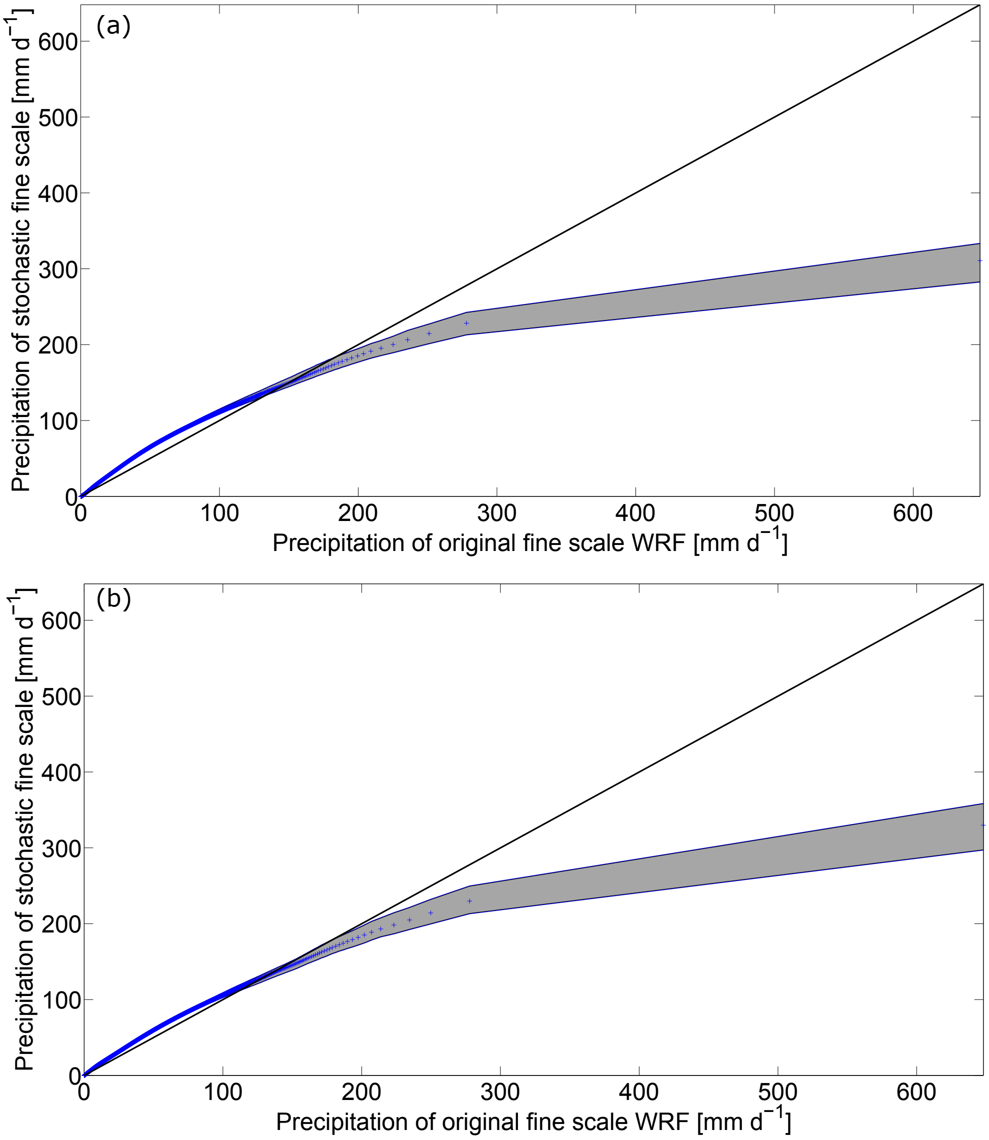

FIGURE 4.14: Quantile-Quantile-Plot of stochastic simulations with the calibration year 1971 (a) and the calibration year 1982 (b). Crosses: Average distribution of simulations, shaded area: spread of simulated distributions.

The overestimation of the low values with the calibration year 1971 may be related to the CDF parameter estimation technique because the expected value of a gamma-distributed random variable is θk. As the mean of the wet day amount of the calibration year was very low, the scaling factor became high which may have contributed to the slightly stronger bias of the downscaled fields in comparison to the simulations with the calibration year 1982. The underestimation of amounts above 200 $mm\ d^{-1}$ is caused by the Gamma distribution. *Volosciuk et al.* (2017) pointed out that very high precipitation intensities of the RCM cannot be modeled with the Gamma distribution even though it was shown to perform better than most other parametric CDFs (Table 4.1). Also, fitting a parametric distribution to heavy-tailed distributed RCM precipitation is challenging in general (e.g. *Gudmundsson et al.*, 2012). Thus, it can be concluded that the model is robust and the influence of the calibration year is not very high which indicates that the model can be used for the application to fine scale simulations that are only available for a short period.

4.5.4 Brier skill scores of daily stochastic simulations

The Brier skill score *BSS* (*Wilks*, 2011) was calculated for ten precipitation thresholds (0 $mm\ d^{-1}$, and nine quantiles of the dynamical simulations Q_{10}, Q_{20}, Q_{30}, Q_{40}, Q_{50}, Q_{60}, Q_{70}, Q_{80}, Q_{90}). The *BSS* is usually used in forecast verification to determine the skill of a probabilistic forecast. Here it has been employed to measure how well the stochastic downscaling can reproduce precipitation above or below a certain threshold. Precipitation fields were transformed to binary indicator fields I_t. If the precipitation on a day t is above a given threshold, I_t becomes 1 and 0 otherwise. This calculation was performed for each grid cell and all n_t time steps.
The Brier skill score *BSS* was calculated from the Brier scores of the stochastic simulation BS_{sto} and the Brier score of the RCM's climatology BS_{cl}. For each day and grid cell, the ratio $p_{sto,t}$ of the 50 samples which exceed a certain threshold was calculated. The Brier score of the stochastic simulations is given as:

$$BS_{sto} = \frac{\sum_{t=1}^{n_t}(p_{sto,t} - I_t)^2}{n_t} \tag{4.8}$$

Likewise, the ratio c_{cl} of days above a certain threshold in the complete time series of the dynamical simulations was calculated for each grid cell.

$$BS_{cl} = \frac{\sum_{t=1}^{n_t}(c_{cl} - I_t)^2}{n_t} \tag{4.9}$$

The *BSS* measures the skill of the stochastic simulations in comparison to the climatology of the dynamical simulations for a given threshold. The *BSS* is defined as:

$$BSS = 1 - \frac{BS_{sto}}{BS_{cl}} \tag{4.10}$$

As an example, a map of the annual *BSS* with the median quantile $Q_{50} \approx 1.55\ mm\ d^{-1}$ as threshold is shown in Figure 4.15.

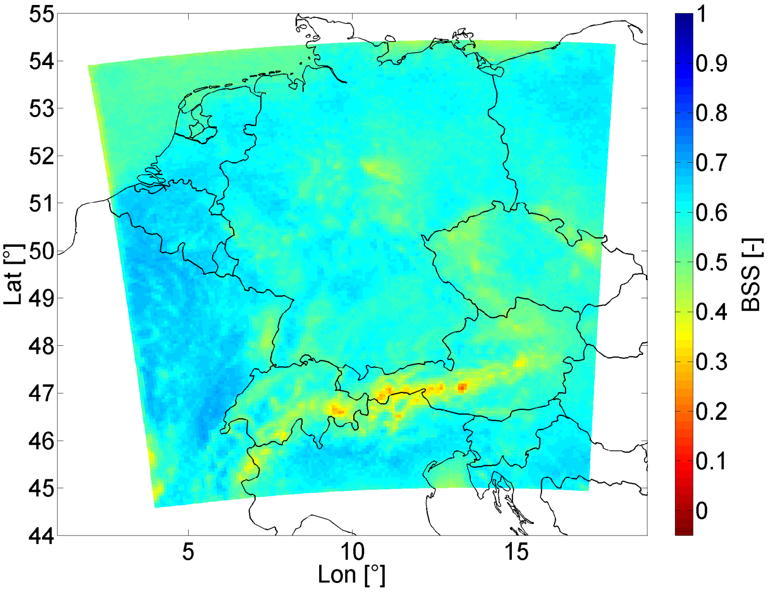

FIGURE 4.15: Brier skill score of stochastic simulations for a threshold of $Q_{50} \approx 1.55\ mm\ d^{-1}$ in the period 1972-2000.

A *BSS* of 1 indicates a perfect simulation and a *BSS* of 0 signifies that the stochastic simulations are only as good as the climatology. Values below 0 indicate no skill. The performance of the stochastic model differs depending on the chosen threshold and the geographical location. In many regions, a high performance is obtained with BSS values of more than 0.6. The best performance was attained in France and Belgium with BSS values of approximately 0.7. Over the oceans, the performance decreases ($BSS \approx 0.5$) and in the Alpine region it is the worst with a minimum BSS of 0.18. This behavior is due to the complexity of the precipitation formation in this region and the different representation of physical processes and terrain elevation in the dynamical simulations.

An overview of the performance for the ten different thresholds is given in the violin plot in Figure 4.16. In general, a positive BSS was attained and the average BSS amounts to 0.51. For a threshold of 0 $mm\ d^{-1}$ and for the Q_{90} quantile, the performance is slightly worse. The precipitation fields that were generated with the IDW interpolation show a worse performance for all thresholds.

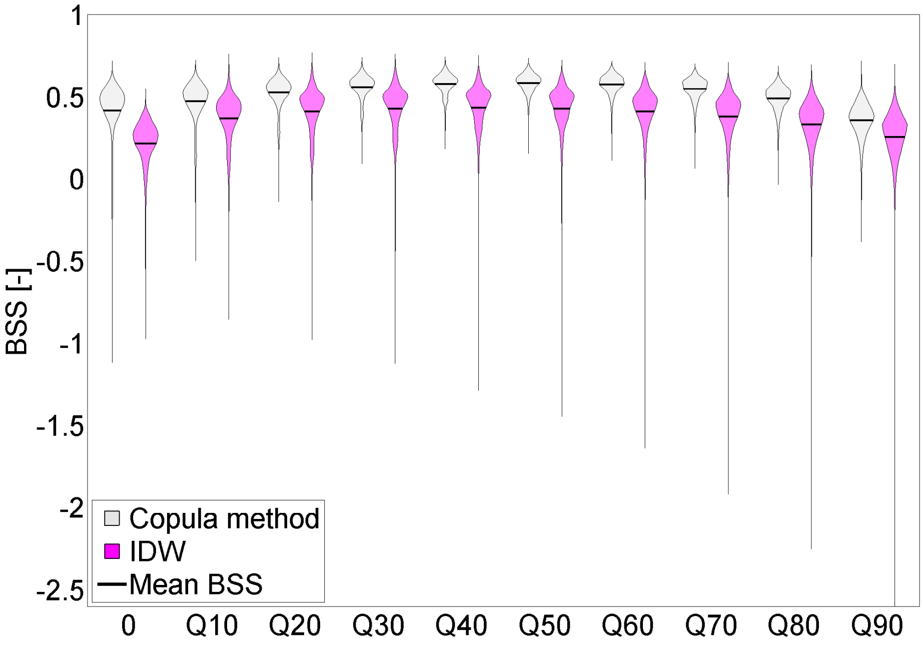

FIGURE 4.16: Violin plot of Brier skill scores of the proposed copula-based method and IDW-fields for ten ascending quantile-based precipitation thresholds in the period 1972-2000.

That the performance of the copula-based downscaling is worse for the threshold 0 $mm\ d^{-1}$ can be explained by the partially random generation of very low precipitation intensities in RCMs (drizzle effect). For instance, *Olsson et al.* (2015) demonstrated that the bias of sub-daily precipitation simulations increased with spatial resolution and that it might be caused by the drizzle effect. Depending on the definition of a wet time step, the bias even changed its sign in winter.

The performance for the higher intensities may be reduced because these intensities are related to extreme events which can often exhibit anisotropic dependence (*Niemi et al.*, 2014). The correlation matrix however is modeled with an isotropic correlogram function and thus two spatial conditioning points in the same distance from the target cell can have the same influence on the predictor even though one of the points may be on the other side of a mountain range. The BSS for the Q_{90} threshold is especially low in the Alpine Region where the precipitation amounts of the dynamical fine scale simulations exhibit large variation. Also, the Gamma distribution resulted in an underestimation of the extreme values and the BSS is typically lower for higher thresholds. Analyses of precipitation forecasts by *Liechti et al.* (2013) and *Bowler et al.* (2006) also showed that the BSS decreased for larger thresholds.

4.5.5 Comparison of daily dynamical and stochastic simulations

The performance of the stochastically simulated precipitation fields was also evaluated by calculating the mean difference of daily precipitation. For the evaluation period, the mean difference lies in the range of -1 to 1 $mm\ d^{-1}$ for 78.3% of the grid cells (Figure 4.17), even for many mountainous regions in Southern and Eastern Germany. In the Alpine Region however, the differences are larger because the downscaling method overestimated the daily precipitation amounts.

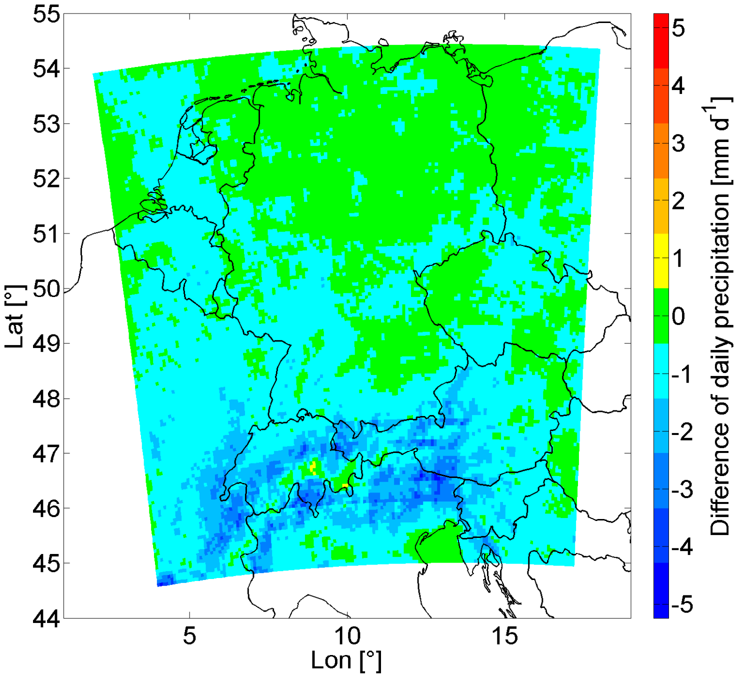

FIGURE 4.17: Mean difference of dynamical and stochastic daily precipitation (mean over 50 samples) - 1972-2000.

As additional performance measures, the ratio of correctly simulated dry and wet days and the correlation of wet day amounts were calculated to evaluate the discrete-continuous character of the generated precipitation fields. The precipitation amounts were considered as wet if they were larger than 1 *mm* d^{-1}. For wet days, the correlation of the amounts was squared to obtain the explained variance. These performance measures were chosen because they were also utilized by *Chen et al.* (2013) in a study with a similar objective but for another region. The results of this analysis are given in Figure 4.18.

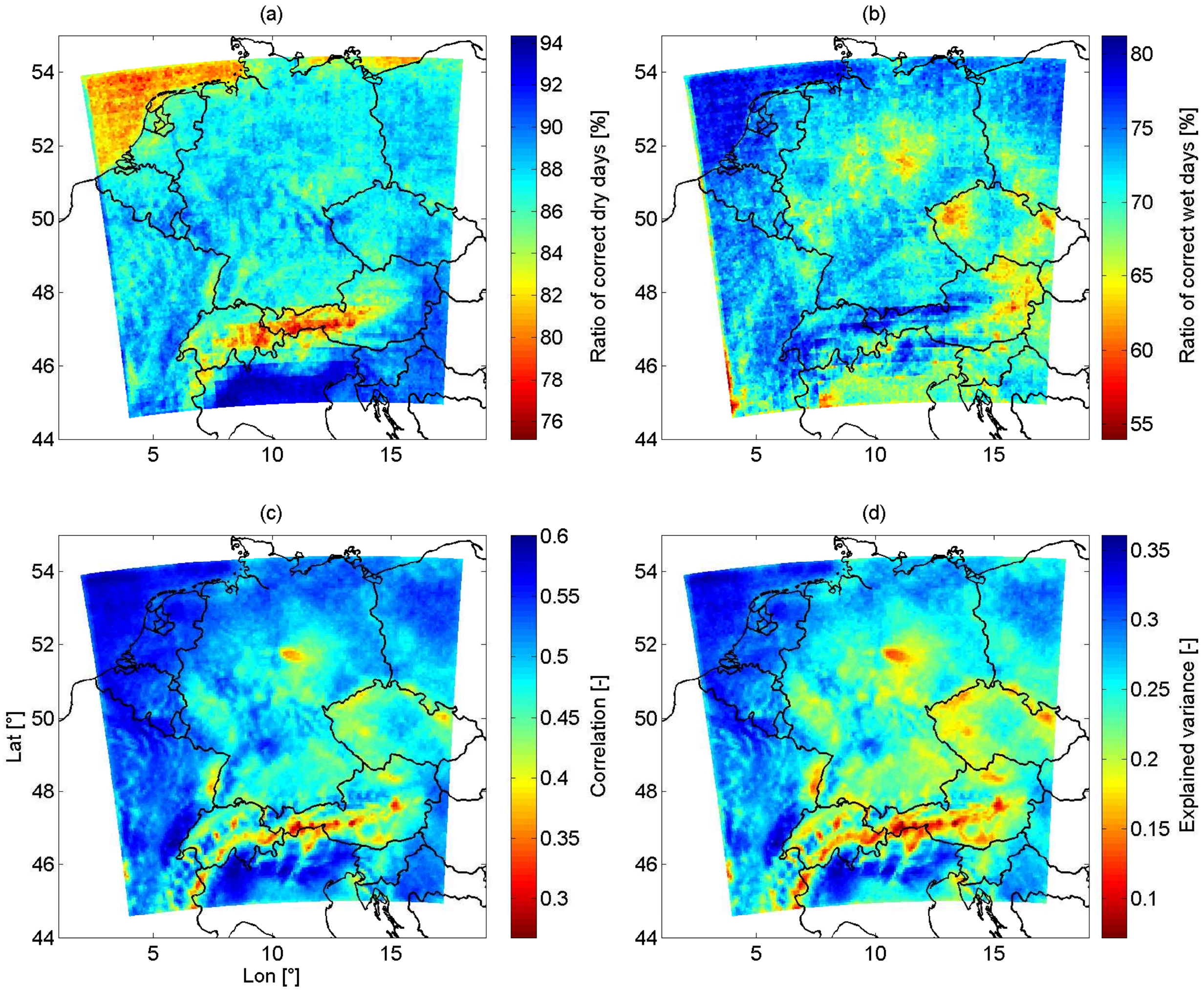

FIGURE 4.18: Performance of stochastic simulations - Reproduction of dry days (a), reproduction of wet days (b), correlation of wet day amounts (c) and explained variance of wet day amounts (d) - 1972-2000.

Dry days were correctly simulated in 87.2% of the cases, with lower skill for Alpine and maritime regions (a). Wet days were not simulated as well as dry days (b) (71.4% correctly simulated wet days). The mean explained variance (squared correlation) of wet day amounts is 25.2% (d). The worst performance can be seen in regions with complex orography, i.e. the Alps, the Ore Mountains in the Czech Republic and the Harz in Germany.

4.6 Summary and outlook

A novel copula-based method to spatially downscale RCM precipitation fields has been presented. The downscaling to the domain of the fine scale RCM simulations was based on the corresponding coarse scale RCM precipitation simulations. No auxiliary variables were used but only the precipitation amounts of the coarse scale field served as predictors. The method has shown to reproduce the spatial dependence of the original fine scale simulations well. The distribution parameters were estimated from a short calibration period to mimic the common problem of limited fine scale data availability in hydrological practice. Calibrating the model with a different year did not alter the distribution of the simulations by a lot and in both cases, extreme precipitation was underestimated. In general, RCM simulations of precipitation exhibit a bias when compared with observed data and a bias correction is necessary for further impact studies. This way, the stochastically generated precipitation fields can be regarded as a suitable surrogate for fine scale RCM simulations because the spatial correlograms and distribution functions can be reproduced for the most part even though the very high extreme values of RCM simulations were not reproducible by the Gamma distribution. Running one year of the 7km-WRF simulations required 48 CPUh on 96 nodes of a high performance cluster whereas the presented stochastic algorithm takes approximately 24 CPUh on a single standard CPU to simulate 50 estimates of this time series based on the driving RCM in a spatial resolution of 42 km. As the computation time is reduced by a factor of 200, the stochastic method can be useful in the absence of computational power or time to dynamically downscale precipitation.

A direct comparison of the daily precipitation amounts of the stochastic and dynamical simulations has been conducted. Average Brier skill scores of 0.51 have been attained. In comparison to the IDW interpolation, the skill scores are higher by 0.10 to 0.20 (42% on average) for all thresholds. Dry days were correctly simulated in 87.2% of the cases and in 71.4% for wet days. The explained variance of wet day amounts is 25.2%. These performance measures indicate that the copula-based downscaling method performs rather well. Furthermore, the presented method is not limited to model the dependence of precipitation in different spatial resolutions and can be adapted to other downscaling problems as an alternative to e.g. MLR-methods. Copula-based methods offer the advantage to be flexible as the marginal distributions are modeled separately from the spatial dependence. Therefore, predictors and predictands that follow different distributions can be used. For instance, *Yang et al.* (2017) found that near surface specific humidity from reanalysis data was the best predictor for station-scale precipitation and mean temperature at 2 m height was the best predictor for observed minimum and maximum daily temperature. In contrast to most other downscaling techniques, this method estimates the dependence of predictors and predictands based on the separation distance. Because it is not necessary to fit an estimator function of the predictand for each location individually, the downscaling can be performed for arbitrary target domains if the predictand's CDF is known or estimated. Additionally, not a single, deterministic result is obtained but an ensemble of predictand fields which allows to address the uncertainty of the downscaling process. In practice, a subset of the generated ensemble may be employed if the computational demand of running an impact model with the full ensemble is too high. To ensure that the spread of the predictor field is large enough to cover the range of uncertainty, the ensemble members may be evaluated by calculating statistics like the mean areal precipitation and by selecting only a few downscaled fields.

Further investigations may be conducted with other simulated or observed predictors and predictands which may require different distribution functions, anisotropic or location-specific correlograms, non-symmetric copulas or model statistics related to different seasons or atmospheric circulation patterns. Additionally, the model could be employed to downscale climate variables for future conditions. In that case, it would however be necessary to investigate if the transfer function that is defined by the correlograms and the univariate distribution functions can be considered stationary. *Hertig et al.* (2016) investigated temporal change points in the predictor-predictand relationship when downscaling daily precipitation from atmospheric predictors with two statistical downscaling techniques and found that 40% of all stations showed robust change points in the model parameters.

Another proposed extension of the method is related to spatial disaggregation, i.e. when the fine scale sub-fields within one coarse scale are required to add up to the value of the coarse scale value. By selecting the simulated field which deviates the least from the coarse cells' values and rescaling the sub-fields, the method could be used to disaggregate e.g. satellite-derived observations to the domain of an observed fine scale field.

Chapter 5

Copula-based temporal disaggregation of RCM precipitation

This chapter presents a novel stochastic disaggregation procedure that increases the temporal resolution of RCM precipitation time series. As in Chapter 3, a bias correction was necessary as a first step. While Chapter 4 presented a stochastic method to increase the spatial resolution of RCM precipitation, this chapter focuses on temporal disaggregation. Currently, it is not feasible to operate an RCM in a resolution of 5 minutes for long time periods but such a high temporal resolution is required for applications in urban hydrology like the design of a quickly responding sewage system. Urban hydrology is also often confronted with the problem of missing spatially distributed observations for long time periods. To meet the demands of impact modelers, a model was developed to bias correct hourly RCM precipitation and then disaggregate it to the required resolution of 5 minutes. The model was applied to a small, orographically complex region around Freiburg im Breisgau, Germany. The developed technique utilizes the Gaussian Copula to model the spatio-temporal dependence structure of precipitation.

5.1 Overview of time series and disaggregation models

Precipitation time series can be simulated with statistical or stochastic approaches. At first, an introduction to models that simulate time series independently of a higher-aggregated value are presented. These models can be used as Weather Generators to extend time series or to simulate time series at an ungauged location. Some of these models have been adapted for disaggregation purposes and those extensions are mentioned in the respective paragraphs.

Models mainly developed for the disaggregation of precipitation are treated afterwards. The existing disaggregation methods can be classified into statistical or stochastic methods. A stochastic disaggregation does not generate identical time series for the same input twice, so an ensemble of precipitation realizations can be generated. Another distinguishing feature of disaggregation techniques is how they treat the spatial dependence structure of the disaggregated time series of different points in space.

5.1.1 Time series models

Markov Chain models have been used frequently to estimate if a day is wet given the condition of the previous day (*Katz and Zheng*, 1999). A first order Markov Chain simulates the unknown state (wet or dry) conditional on the last state (the preceding time step) but higher orders are possible. Also, a further classification of the states is possible to divide the wet state into different intensity classes or the waiting time between tips of a tipping bucket respectively (*Sørup et al.*, 2012).

Point process models simulate the arrival time of individual rain pulses with a random duration and depth that are then summed to generate a time series of rainfall intensity. *Rodriguez-Iturbe et al.* (1987) presented two widely-used point process models, namely the Neyman-Scott-Rectangular-Pulse-Model (NSRP) and the Bartlett-Lewis-Rectangular-Pulse-Model (BLRP). *Hershenhorn and Woolhiser* (1987) developed a model to simulate the number of pulses conditional on the daily rainfall and then subsequently modeled the duration and amount of the individual pulses. While these models have been used to simulate very finely resolved time series and disaggregate to resolutions up to 1 minute (*Kossieris et al.*, 2016), the parameter fitting for a NSRP-model or BLRP-model is not straightforward as the required parameters are not directly observable. Nevertheless, there are still comparatively new publications like the one by *Evin and Favre* (2008) who modeled the dependence of depth and duration with copulas or *Tarpanelli et al.* (2012) who have extended the single-site NSRP model to simulate spatially-correlated time series. An extensive description of different point process models can be found in *Beck* (2013).

Bárdossy (1998) simulated precipitation values from a univariate distribution function and distributed them randomly in time over one year. Afterwards, the temporal sequence was rearranged with Simulated Annealing to optimize the ACF, the statistical moments on different aggregation levels and the ratio of the sum of different time intervals to a prescribed total to respect the annual cycle. *Brommundt* (2008) extended this model to include spatial correlation.

Another class of stochastic precipitation models are Alternating-Renewal-Models (e.g. *Haberlandt et al.*, 2008). They consist of an internal and an external structure. The external structure models rainfall events: A time period is either a dry spell period with a Weibull-distributed duration D or a wet spell period with a certain intensity I and duration W. The dependence of I and W is modeled with the Frank Copula. The internal structure is the profile model of wet spells, i.e. how the intensity changes over the course of the wet spell. This is modeled with an exponential function for the rise and decay. To introduce spatial dependence, the generated events of single sites are resampled with Simulated Annealing with three optimization criteria. An extended version (*Callau Poduje and Haberlandt*, 2017) models the dependence of wet spell duration and amount with one copula and the dependence of wet spell intensity and the maximum intensity within the wet spell with another copula. *Vernieuwe et al.* (2015) proposed a copula-based model to simulate the parameters of the external structure, namely storm duration, volume, the dry period and the fraction of dry values within a storm event with a mixture of Frank copulas. The theoretical background of this multivariate dependence model is the concept of Vine copulas (*Joe*, 1996; *Aas et al.*, 2009) which model the joint probability sequentially with bivariate copulas of pairs of CDF values. The internal structure of the model is based on Huff Curves (*Huff*, 1967) to distribute rainfall intensities over the non-zero time steps within the period of the simulated storm that sum up to the storm depth.

Wilks (1999) simulated precipitation for multiple sites simultaneously. In a first step the occurrence of a wet day was simulated with a second order Markov Chain. A day is wet if a random number is below the normalized transition probability to a wet day. Spatial correlation was introduced by drawing this random number from a multivariate Gaussian distribution whose correlation matrix is optimized in order to reproduce the correlations of the observations. The intensity of wet days is then simulated from a mixed exponential distribution with the parameter depending on the random number drawn for the occurrence model. This way, the simulated intensities exhibit more spatial coherence than would be possible with a completely random simulation from a distribution function of rainfall intensities. While the original method by *Wilks* (1999) analyzed the relation of the correlation of the random number for the occurrence model to the correlation of the observation and derived the optimal correlation coefficients empirically, *Brissette et al.* (2007) optimized the correlation matrix in an automated fashion.

Another model that generates spatially-correlated precipitation time series conditional on the previous time step was presented by *Serinaldi* (2009). He used bivariate Archimedean copulas to model the dependence of two consecutive wet values and formulated a joint probability distribution as a mixture of the copula, univariate distribution functions and the joint occurrence probabilities of two time steps that are either wet or dry. From this equation, he derived a conditional distribution and inverted it with spatially-correlated random numbers that reproduce the spatial correlation of the observations.

5.1.2 Disaggregation techniques

Disaggregating observed data to a finer temporal scale has already been started several decades ago. *Yevjevich and Lane* (1997, p. 421ff) gives an overview of the first developed models and rates the model by *Valencia and Schaake* (1973) as the "first well-accepted model". It was originally applied to the disaggregation of annual discharge into monthly values and simulates monthly discharge based on the covariance to the normalized annual discharge and a stochastic term.

Betson et al. (1980) presented a model called TVA-HYSIM to disaggregate monthly precipitation into daily, hourly and even 5-minute values. They estimated the transition probability if a day is wet depending on the previous day as in a first state Markov Chain model. For wet days, precipitation was simulated from a Weibull distribution and the simulated monthly time series was multiplied by a factor to preserve the monthly total.

In contrast to the aforementioned point process models, Random Cascade models are a pure disaggregation technique and do not model time series of rainfall intensities without a given higher aggregated value. *Onof et al.* (2005) presented such a model for the disaggregation of hourly precipitation values to a resolution of 3.75 minutes and compared their model to the STORMPAC disaggregation model developed by *Cowpertwait* (1991). A Random Cascade typically disaggregates the precipitation in a given time interval step-wise into two time intervals (in this case, 2 is the so called branching number) of the same length and assigns each interval a cascade weight that is multiplied with the coarser value. The weights are positive and sum up to 1 to ensure conservation of the precipitation amount. The model then proceeds until the desired resolution is reached. Therefore the temporal resolution is not arbitrary but a power of 2. *Olsson* (1998) also addressed this problem as he was using a final resolution of 8 minutes which is neither a common time step of measurements nor a number that can lead to common coarse scale resolutions like 60 minutes or 1440 minutes (daily resolution). There are, however, workarounds to overcome this problem. *Lisniak et al.* (2013) used a branching number of 3 for the first cascade (24h to 8h) and then the typical branching of 2 to obtain hourly precipitation intensities which is a common temporal resolution. *Müller and Haberlandt* (2015) generated spatially-correlated precipitation time series by resampling the disaggregated time series of individual sites with Simulated Annealing. *Thober et al.* (2014) used cascades to disaggregate monthly into daily precipitation. The cascade weights were drawn from a multivariate Gaussian distribution with a covariance matrix that is set up with the observed covariance of the weights. Then, the weights are multiplied with the coarse scale precipitation field to obtain a rainfall field.

Mascaro et al. (2013) applied a method developed earlier by *Deidda et al.* (1999) which is based on the theory of multifractals. The model makes use of the scaling properties of rainfall in space and time to disaggregate 6-hourly precipitation at a resolution of $104x104km^2$ to a gridded data set with a resolution of 13 km and 45 minutes. A comparison of disaggregation methods with point-process models, random cascades and the method of fragments can be found in *Pui et al.* (2012).

Knoesen and Smithers (2009) have developed a disaggregation procedure of daily to hourly rainfall for South Africa where a distinct diurnal cycle is observable. This model is based on observed statistics of the fraction (sometimes also called fragment in the literature) of each hour to the daily total. The distribution of the hour of maximum precipitation is used to simulate the time of occurrence. To allow for a clustering of the time series, the adjacent hours obtain a fraction that leads to a good agreement of the fraction of the higher-aggregated (e.g. the average maximum 6-hourly rainfall fraction). Another method to disaggregate daily precipitation to hourly values that follow the observed diurnal cycle was presented in *Waichler and Wigmosta* (2003) where the authors used the average fraction of each hour's value to the daily total. *Gyasi-Agyei* (2012) used a copula to estimate the length of a storm conditional on the total rainfall amount and disaggregated it with observed storm profiles.

Oriani (2014) disaggregated rainfall by resampling the historical observations in a random order. If the sum of a time slice is close enough to the value that needs to be disaggregated, it is selected. *Westra et al.* (2013) resampled the observed rainfall amounts in a resolution of 6 minutes based on similarity measures of atmospheric variables to disaggregate daily precipitation amounts for projected future conditions with a combination of a generalized additive model and the method of fragments.

Segond (2010) adapted the generalized linear model by *Chandler and Wheater* (2002) to simulate the occurrence of daily gamma-distributed precipitation based on a set of spatial and seasonal predictors for a single location which was coined "master station". Hourly precipitation was simulated with the BLRP-Model and the most suitable time slice was selected and multiplied by a factor such that the daily total was maintained. The temporal distribution of the fractions of each hour to the daily total was transfered from this master station to the other locations in the study area.

Koutsoyiannis et al. (2003) disaggregated daily to hourly precipitation. The model simulates spatially-correlated time series with a multivariate autoregressive model. A transformation function was developed to rescale the hourly time series such that the daily total is matched while preserving the first two statistical moments.

Bárdossy and Pegram (2016) conditioned the simulation of hourly precipitation intensities for locations where only daily measurements are available on the measured daily amount and the hourly measurements of spatial neighbors. The chosen multivariate distribution is the Gaussian Copula and covariance in both time and space is considered by formulating the total covariance as a product of spatial and temporal covariance. *Allard and Bourotte* (2014) disaggregated daily precipitation values to hourly values for two measurement stations in France. Zero and positive precipitation was modeled as a truncated Gaussian variable and the auto correlation function was fitted with MLM as described in *Durban and Glasbey* (2001). Hourly precipitation was simulated from a bivariate Gaussian distribution conditioned on the previous hourly value. A time slice of 24 simulated hourly precipitation was accepted if it

agreed to a pre-defined threshold with the daily amount.
Allcroft and Glasbey (2003) showed how spatially aggregated radar-based precipitation fields can be disaggregated back to the original resolution by transforming precipitation to a Gaussian variable via quadratic transformation. The starting point of the spatial downscaling is the coarse scale value uniformly distributed across all corresponding fine scale cells. Those values are then updated with Gibbs Sampling. Fine scale precipitation was simulated as a Gaussian Markov Random Field that was conditioned on spatial and temporal neighbors until the fine scale values within a coarse cell agreed to a prescribed threshold with the higher aggregated value.

5.2 Study region and data

The disaggregation model has been developed for a study region around the city of Freiburg im Breisgau in the German federal state Baden-Württemberg with DWD station data covering the period 1951-2013. 9 ungauged locations were selected to perform a disaggregation of hourly precipitation to time series with a resolution of 5 minutes. The hourly precipitation time series were simulated with the regional climate model WRF in a spatial resolution of 5 km for the period 1980-2009. A description of the RCM simulations can be found in the technical report by *Wagner and Kunstmann* (2016). Figure 5.1 shows the location of the nine target locations and the 16 measurement stations that were used to calibrate the disaggregation model.

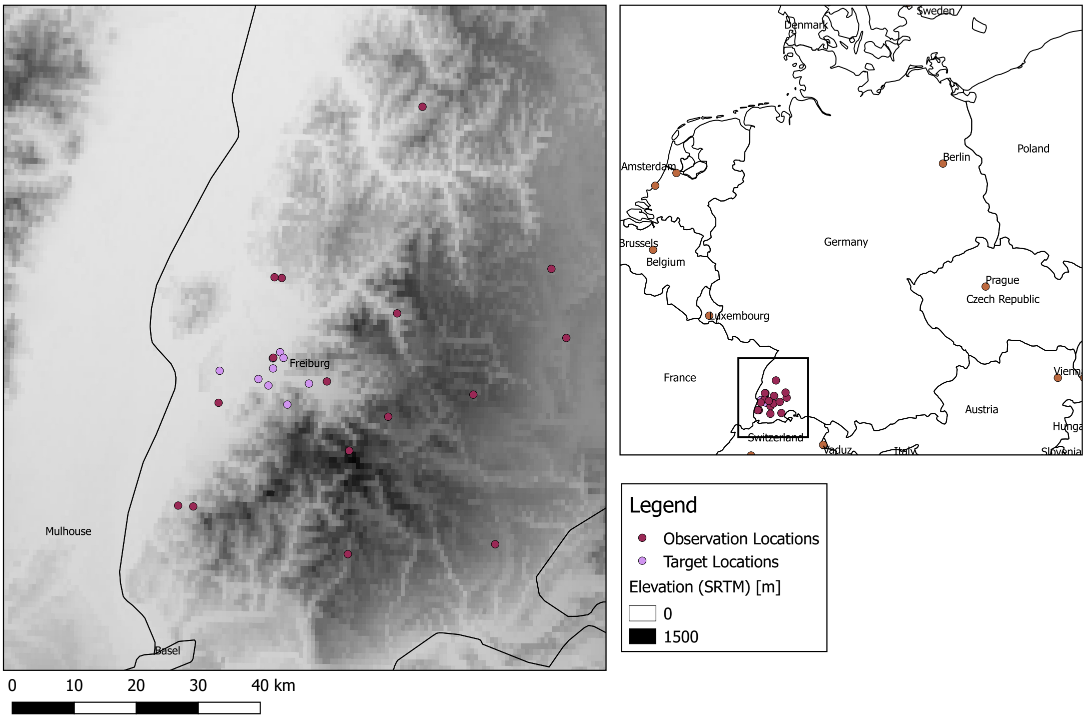

FIGURE 5.1: Location of the 16 measurement stations used for the calibration of the disaggregation model and the 9 target locations

5.3 Development of a copula-based disaggregation model

A regional climate model generates precipitation time series on a spatial grid. Since rain gauges are irregularly distributed in space, a disaggregation model should offer the possibility to estimate the model parameters for grid cells without observed data. Many of the aforementioned methods disaggregate precipitation independently for different locations. Introducing spatial correlation is one of the main problems, especially when no observed data is available for a target location.

NSRP and BLRP models are determined by non-observable parameters which makes the estimation for ungauged locations difficult. Markov Chains require a lot of data to calculate the transition probabilities and this can result in sparse transition matrices. The attainable temporal resolution of random cascade models is determined by the branching number in each cascade step. For instance, in order to disaggregate hourly values to a resolution of five minutes, the hourly value could be disaggregated to two 30 minute values, then two 15 minute values and three 5 minute values. Alternatively, a branching number of 3 in the first step would lead to 20 minute values which would then be disaggregated in two consecutive steps with a branching number of 2. This leads to a rather complex model structure where many different versions need to be evaluated. Also, the parameter estimation for ungauged locations is not straightforward. Resampling or analogue methods require long time series for all locations to sample from which impedes their applicability for the presented case. Even if such data is available, there might not exist time slices with sums similar to the higher-aggregated amount when disaggregating several locations at once.

The advantage of copula-based methods is that the marginal distribution is treated separately from the dependence structure. Therefore, the variable of interest can be fitted with distribution functions in a separate step. The distribution functions for precipitation are typically determined by two to four parameters, with the rainfall probability being one of them. Precipitation in a fine temporal resolution like 5 minutes, exhibits auto correlation in time and cross correlation in space which should be taken into account, so that subsequent impact models can be driven by precipitation time series that reproduce this dependence structure. The model presented in *Serinaldi* (2009), conditioned the simulation on only the previous time step but it would be preferable to include more conditioning values for fine temporal resolutions to more accurately model the temporal structure. The Gaussian Copula can be used to simulate conditionally on several spatio-temporal neighbors and it is determined by the correlation matrix which is easier interpretable than the parameters of some of the aforementioned models. A method that disaggregates the locations sequentially has been developed. Already disaggregated time slices were used to condition the simulation of the current location. The proposed method therefore has some similarities to the techniques presented in *Bárdossy and Pegram* (2016) and *Allcroft and Glasbey* (2003).

The disaggregation model presented in this section simulates spatio-temporally correlated time series with the Gaussian Copula. The required statistics to set up the model are:

1. Distribution parameters and dry probability to transform precipitation to CDF values and vice versa,
2. A parametric function $\widehat{\rho}_{A,\tau}$ to estimate the auto correlation of values separated by a temporal lag τ for one location and
3. A parametric function $\widehat{\rho}_{C,ij,\tau}$ to estimate the spatial cross correlation of two points (i, j) with a temporal lag τ.

The disaggregation model simulates unknown CDF values u_0 conditional on n CDF values of preceding time steps of the same location $(u_{0,\tau_1}, ..., u_{0,\tau_n})$ and conditional on the CDF values of m spatial neighbors of the same time step $(u_{1,\tau_0}, ..., u_{m,\tau_0})$. The conditional PDF of the unknown CDF value u_0 is:

$$f_c(u_0|u_{0,\tau_1}, ..., u_{0,\tau_n}, u_{1,\tau_0}, ..., u_{m,\tau_0}) = \frac{c(u_0, u_{0,\tau_1}, ..., u_{0,\tau_n}, u_{1,\tau_0}, ..., u_{m,\tau_0})}{c(u_{0,\tau_1}, ..., u_{0,\tau_n}, u_{1,\tau_0}, ..., u_{m,\tau_0})} \tag{5.1}$$

The constant denominator is dropped from the equation and the copula density of the Gaussian copula is calculated to model c. The unknown and known CDF values are denoted by $\mathbf{u} := (u_0, u_{0,\tau_1}, ..., u_{0,\tau_n}, u_{1,\tau_0}, ..., u_{m,\tau_0})$.

$$c(\mathbf{u}) = \frac{1}{\prod\limits_{i=1}^{n+m+1} \phi(\Phi^{-1}(u_i))} \frac{1}{(2\Pi)^{\frac{n+m+1}{2}} \sqrt{|\Gamma|}} e^{-0.5(\Phi^{-1}(\mathbf{u})^T(\mathbf{\Gamma}^{-1}-\mathbf{I})\Phi^{-1}(\mathbf{u}))} \tag{5.2}$$

The correlation matrix $\mathbf{\Gamma}$ is estimated with the correlogram models of the auto correlation $\widehat{\rho}_A$ and spatial cross correlation $\widehat{\rho}_C$:

$$\mathbf{\Gamma} = \begin{bmatrix} 1 & \widehat{\rho}_{A,\tau_1} & \widehat{\rho}_{A,\tau_2} & \cdots & \widehat{\rho}_{A,\tau_n} & \widehat{\rho}_{C,01,\tau_0} & \widehat{\rho}_{C,02,\tau_0} & \cdots & \widehat{\rho}_{C,0m,\tau_0} \\ \widehat{\rho}_{A,\tau_1} & 1 & \widehat{\rho}_{A,\tau_1} & \cdots & \widehat{\rho}_{A,\tau_{n-1}} & \widehat{\rho}_{C,01,\tau_1} & \widehat{\rho}_{C,02,\tau_1} & \cdots & \widehat{\rho}_{C,0m,\tau_1} \\ \widehat{\rho}_{A,\tau_2} & \widehat{\rho}_{A,\tau_1} & 1 & \cdots & \widehat{\rho}_{A,\tau_{n-2}} & \widehat{\rho}_{C,01,\tau_2} & \widehat{\rho}_{C,02,\tau_2} & \cdots & \widehat{\rho}_{C,0m,\tau_2} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \widehat{\rho}_{A,\tau_n} & \widehat{\rho}_{A,\tau_{n-1}} & \widehat{\rho}_{A,\tau_{n-2}} & \cdots & 1 & \widehat{\rho}_{C,01,\tau_n} & \widehat{\rho}_{C,02,\tau_n} & \cdots & \widehat{\rho}_{C,0m,\tau_n} \\ \widehat{\rho}_{C,01,\tau_0} & \widehat{\rho}_{C,01,\tau_1} & \widehat{\rho}_{C,01,\tau_2} & \cdots & \widehat{\rho}_{C,01,\tau_n} & 1 & \widehat{\rho}_{C,12,\tau_0} & \cdots & \widehat{\rho}_{C,1m,\tau_0} \\ \widehat{\rho}_{C,02,\tau_0} & \widehat{\rho}_{C,02,\tau_1} & \widehat{\rho}_{C,02,\tau_2} & \cdots & \widehat{\rho}_{C,02,\tau_n} & \widehat{\rho}_{C,21,\tau_0} & 1 & \cdots & \widehat{\rho}_{C,2m,\tau_0} \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \widehat{\rho}_{C,0m,\tau_0} & \widehat{\rho}_{C,0m,\tau_1} & \widehat{\rho}_{C,03,\tau_2} & \cdots & \widehat{\rho}_{C,0m,\tau_n} & \widehat{\rho}_{C,m1,\tau_0} & \widehat{\rho}_{C,m2,\tau_0} & \cdots & 1 \end{bmatrix} \tag{5.3}$$

The unknown CDF value u_0 is varied from 0 to 1 and the conditional PDF f_c is calculated for those values to obtain the complete conditional PDF. This function is then numerically integrated and normed to obtain the CCDF F_c which is then inverted with a uniformly distributed random number w to simulate a CDF value u_0.

$$u_0 = F_c^{-1}(w), w \sim U(0, 1) \tag{5.4}$$

The simulated CDF value u_0 is then transformed to precipitation via the truncated CDF of the target location. As w is random, an ensemble of fine scale precipitation time series x_{cand} is simulated. Once the length of the simulated time series x_{cand} in the fine resolution equals the resolution of the precipitation amount x_{coarse} that is being disaggregated (e.g. twelve 5-minute-values when hourly amounts are disaggregated), all candidates x_{cand} are evaluated. The rescaling factor $f_{rs} = \frac{x_{coarse}}{\sum x_{cand}}$ is multiplied with each candidate time slice x_{cand}, so that the sum of the rescaled candidate matches the coarse amount. The suitability of a candidate is evaluated by the sum of differences to the power of 4 between the rescaled candidate and the original candidate. The best-matching sample x_{sel}, which minimizes the differences to the power of 4, of all candidates x_{cand} is selected. A power of 4 has been chosen to penalize large deviations more strongly than e.g. absolute or squared differences.

$$x_{sel} = \underset{x_{cand}}{\arg\min}(f_{rs}x_{cand} - x_{cand})^4 \tag{5.5}$$

The best sample x_{sel} is then rescaled to x_{dis} to match the coarse value x_{coarse}. Then, x_{dis} is fixed and transformed to CDF values, so that it can be used to condition the simulation of further time steps or different locations. Simulated values of u_0 were set to p_0 as in *Bárdossy and Pegram* (2016) because very low values of u were found to generate too many dry events when simulating subsequent samples of x_{cand}.

$$x_{dis} = f_{rs}x_{sel} \tag{5.6}$$

The disaggregation is performed for one location at a time. For the first location, the conditioning values are only from previous time steps of this location. For the first time step and location of each month, the CDF values u_0 are drawn from a uniform distribution because no conditioning values exist. For the second location, the conditioning values stem from previous time steps (if available) and the already disaggregated values of the first location. The process of the selection of conditioning values is depicted in Figure 5.2.

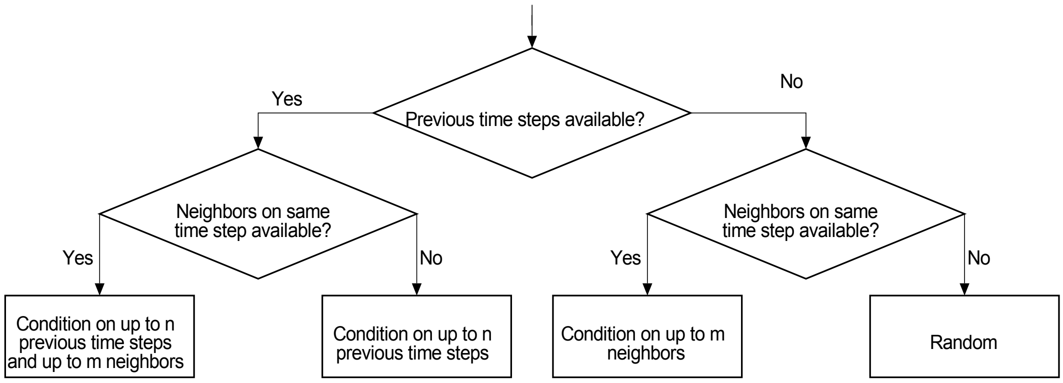
Previous time steps available?
Yes
No
Neighbors on same time step available?
Yes
No
Neighbors on same time step available?
Yes
No
Condition on up to n previous time steps and up to m neighbors
Condition on up to n previous time steps
Condition on up to m neighbors
Random

FIGURE 5.2: Flowchart of the selection of spatial and temporal conditioning values to set up the conditional distribution function.

5.4 Calibration of the copula-based disaggregation model

To disaggregate hourly precipitation simulated by an RCM to a resolution of 5 minutes, a two step procedure is necessary. First, hourly observations are utilized to perform a bias correction for the ungauged target locations. For the calibration of the disaggregation model, observed data in the target resolution of 5 minutes is required.

5.4.1 Bias corrected hourly input data

Observed precipitation data in a temporal resolution of 5 minutes from the German meteorological service DWD was employed to perform the bias correction of the hourly RCM simulations at the 9 ungauged locations. There are 129 precipitation stations in a temporal resolution of 5 minutes in Baden-Würrtemberg. The measurement period is 1951-2013 but only approximately 26 % of the measured time steps are valid.

This data was aggregated to a temporal resolution of 60 minutes and the CDF parameters were kriged to the 9 ungauged locations to perform a Quantile Mapping of the hourly precipitation intensities simulated by the WRF model. The bias correction technique is explained in more detail in Chapter 3. For this case study, not a Double-Quantile-Mapping but a regular Quantile-Mapping was utilized because the RCM simulations were performed for the historical period 1980-2009. The observed hourly precipitation amounts were fitted with a log-normal distribution and the parameters $\{\overline{\ln x}, s_{\ln x}\}$ were kriged to the nine target locations for four seasons (DJF, MAM, JJA and SON).

5.4.2 Statistics of 5 minute data

For the calibration of the disaggregation model, 16 out of the 129 observation stations with a resolution of 5 minutes were selected in the proximity of Freiburg. The observed time series cover the range 1951-2013 but there are many gaps (Figure B.1 in Appendix B). However, no station has been active for more than 30 years. In order to have enough data to fit the disaggregation model, stationarity was assumed and the complete measurement period was utilized even though it is longer than the period of RCM simulations (1980-2009). From this subset of gauge data, the CDF parameters and the correlogram models were calculated for the same four seasons as in the bias correction (DJF, MAM, JJA and SON). The parameters of the log-normal distribution were interpolated to the 9 ungauged locations with IDW. The ACF model $\widehat{\rho_{A,\tau}}$ was fitted by pooling all data of the 16 gauges into one sample. The lagged CCF $\widehat{\rho_{C,h,\tau}}$ was fitted to the gauge data pairs in a spatial distance h and a temporal lag τ.

Observed distribution functions

The 5 minute precipitation amounts were fitted with a log-normal distribution. In contrast to the hourly data, the positive precipitation amounts were additionally grouped into two classes (one part containing 99.5% of the low intensities and one part for the remaining, highest 0.5% of the data). Fitting only one log-normal distribution resulted in too low simulated extreme values because the parameter fitting was dominated by the large set of comparatively low values. By selecting a splitting CDF value of $u_{split} = 0.995$, the extremes are better represented but the trade-off is that it introduces a stepped CDF at the location of the corresponding precipitation intensity x_{split}. An example is given for the measurement station in Freiburg in the season JJA in Figure 5.3. For this station and season the splitting value is $x_{split} = 2.24\ mm\ 5min^{-1}$. Due to the importance of a correct representation of high intensities, the stepped CDF was conceded because no single parametric CDF was capable of simulating precipitation intensities that resemble the observed distribution.

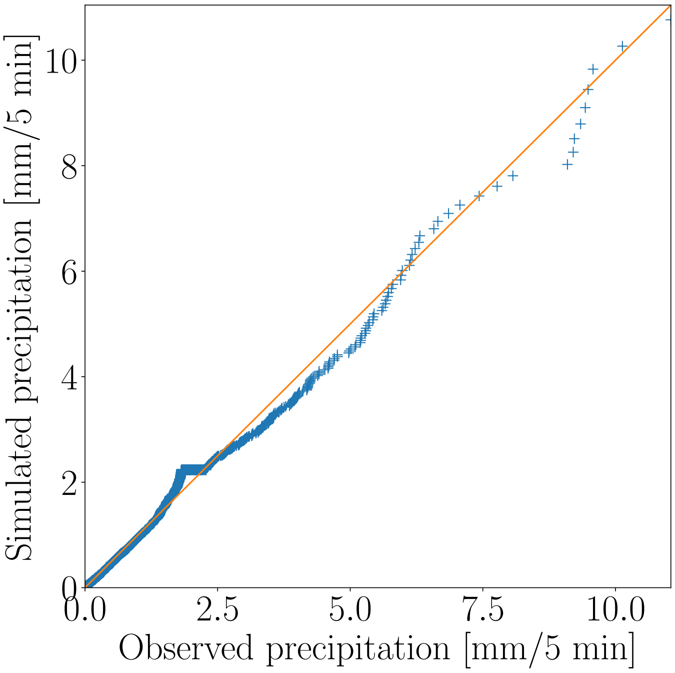

FIGURE 5.3: QQ-Plot of observed and simulated 5-minute-precipitation for Freiburg in the season JJA (1951-2013).

Observed cross correlation

The cross correlation function $\widehat{\rho}_{C,h,\tau}$ was calculated by optimizing the correlogram parameters that maximize the likelihood of the three sets introduced in Chapter 2 section 2.3. The cross correlation was calculated for different temporal lags τ to set up the spatio-temporal correlation matrix. The Matérn model was chosen as the parametric correlogram model due to its flexibility. Restricting the correlograms to just one model also facilitates the evaluation of the disaggregated series' correlation structures. As an example, the cross correlogram models of six temporal lags ($\tau_0 = 0\ min$,..., $\tau_5 = 25\ min$) in the summer season JJA is given in Figure 5.4. As expected, the spatial cross correlation decreases slightly as the temporal lag increases.

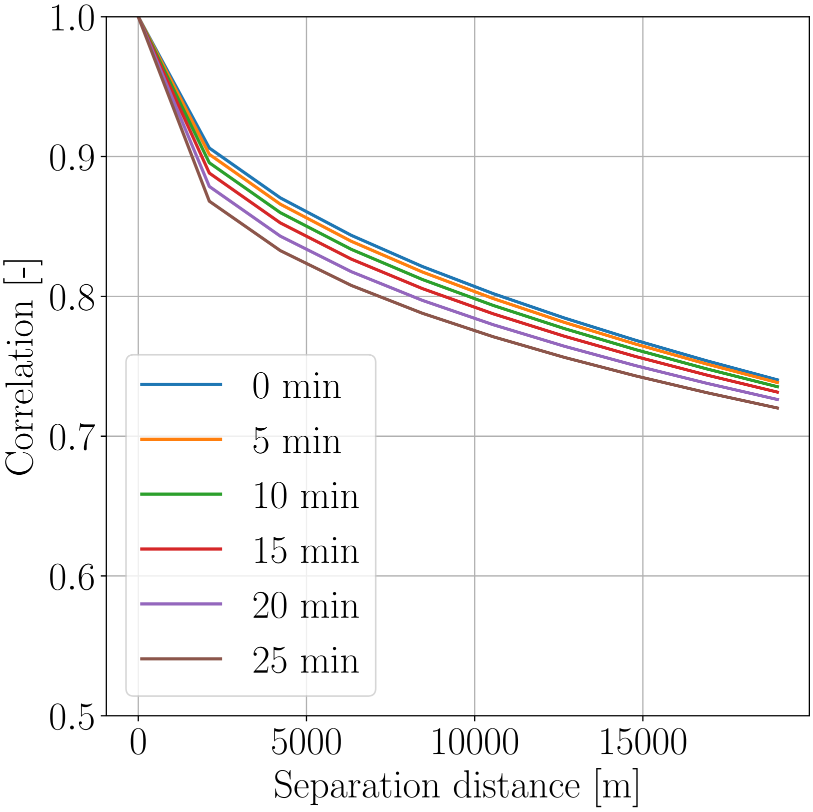

FIGURE 5.4: Fitted cross correlograms of observed 5 minute precipitation in the region of Freiburg for different temporal lags in the season JJA (1951-2013).

Observed auto correlation

Likewise, the auto correlation function $\widehat{\rho}_{A,\tau}$ was fitted by a Matérn correlogram for the four seasons DJF, MAM, JJA and SON. All 16 stations were pooled into one set to estimate how precipitation correlates with preceding time steps. Figure 5.5 shows the parametric ACFs for the four seasons. While the ACFs look very similar, the auto correlation decays a bit faster for the summer months which indicates that precipitation in summer shows less persistence than in the winter months.

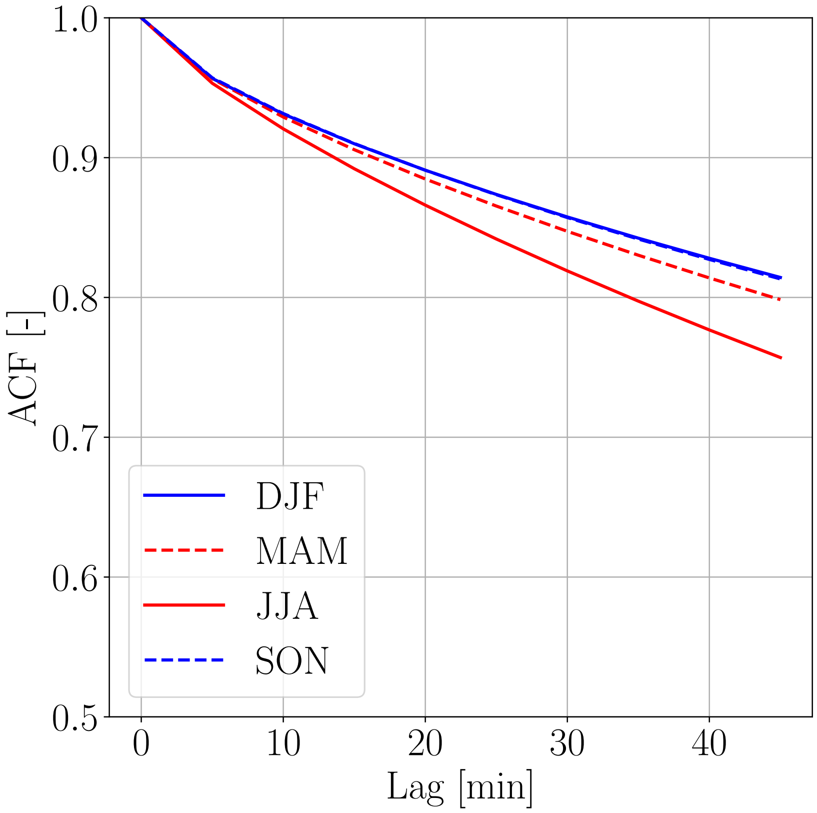

Figure 5.5: Fitted auto correlogram of observed 5 minute precipitation in the region of Freiburg in the seasons DJF, MAM, JJA, and SON (1951-2013).

5.5 Evaluation of the copula-based disaggregation model

The disaggregation model has been applied to 30 years of hourly bias corrected precipitation at 9 different locations. Different configurations were tested by varying the number of spatial (m) and temporal (n) conditioning values. At first, an example of the disaggregation of a heavy precipitation event is given to illustrate the influence of spatial and temporal conditioning values. Afterwards, the distribution functions and correlograms of the disaggregated time series are presented.

5.5.1 Example of spatio-temporal disaggregation for one event

An illustrative example of the disaggregation procedure is given for a precipitation event on July 29, 2005 (2 pm). The hourly amounts to be disaggregated lie in the range $[3.84\, mm\, h^{-1}; 16.06\, mm\, h^{-1}]$. The hourly amounts and the order of the disaggregation is depicted in Figure 5.6.

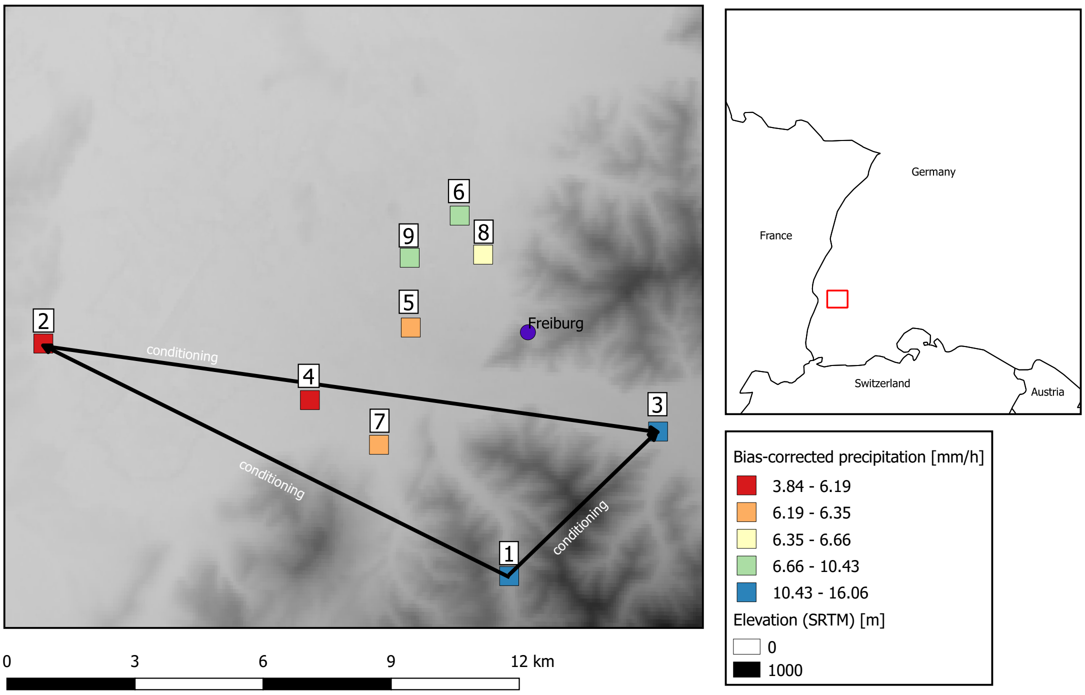

FIGURE 5.6: Order of disaggregation of bias-corrected hourly precipitation for a heavy precipitation event on July 29, 2005 (2 pm) for the first three locations.

The order of the locations is calculated on a monthly basis. The order is chosen in such a way that the monthly sum of precipitation increases with every location. In this example, the disaggregation procedure begins with the location with maximum hourly precipitation (1) but minimum monthly precipitation. Afterwards the location with minimum hourly precipitation (2) is disaggregated and the procedure continues with the location with the third lowest monthly sum of precipitation (3).

For each location candidate time series x_{cand} of varying sample sizes were calculated. For the first location, 500 realizations of x_{cand} were generated because the hourly amount is high (16.06 $mm\ h^{-1}$). 500 samples were chosen for hourly amounts above 5 $mm\ h^{-1}$, so that the sample size of x_{cand} is large enough to provide a candidate time slice that closely matches the hourly amount that needs to be disaggregated. The hourly amount of the second location is 3.84 $mm\ h^{-1}$ and only 50 candidates were simulated.

The previous hour was dry at all locations. Therefore, all conditioning values are p_0 for the first location (1) since no spatial conditioning values are available. Once x_{dis} has been selected from x_{cand} for the first location, the second location is disaggregated. For this point, the simulation of candidates x_{cand} is performed conditionally on $n = 5$ previous 5 minute time steps and the $m = 1$ values of the already disaggregated location (1). Thus, spatial conditioning values $u \geq p_0$ are available for the disaggregation of the second location. For the third location (3), $m = 2$ spatial neighbors are available as conditioning values. Figure 5.7 demonstrates the simulation and selection of candidate time slices for the disaggregation of the first three locations. The first column shows the time series of candidates x_{cand} and the previous twelve 5-minute-time steps (which are all dry). In the second column, the best fitting candidate x_{sel} and the rescaling factor f_{rs} are given. The rescaled best candidate x_{dis} which serves as a spatial conditioning value for the next locations is shown in the third column.

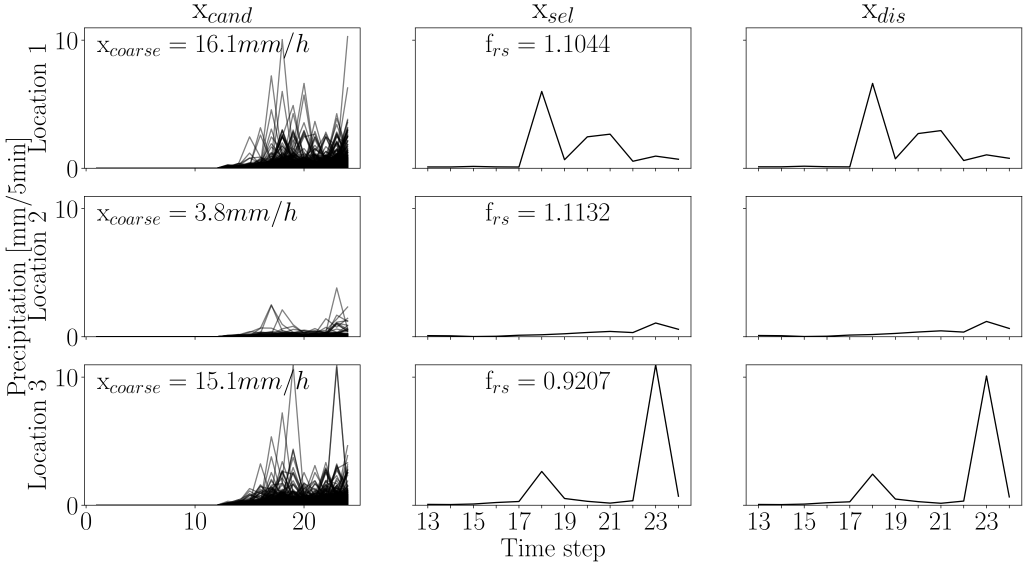

FIGURE 5.7: Selection of disaggregated time slices for the first three locations.

To further illustrate the influence of the spatial and temporal conditioning values, the mean precipitation amount of the twelve time steps of x_{cand} is given in Figure 5.8.

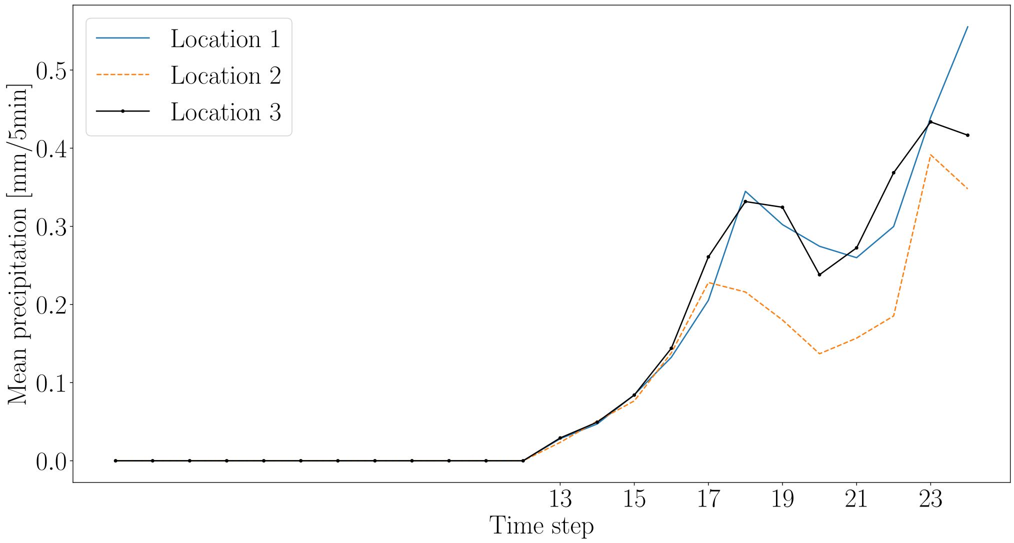

FIGURE 5.8: Mean precipitation of each time step of candidate time series for the first three locations.

For the first location, only the preceding, dry temporal conditioning values were available. The tendency of the mean simulated time series is thus that the first time steps are lower because their simulation was conditioned on $n = 5$ CDF values corresponding to $0\ mm\ 5\ min^{-1}$. The selected and rescaled time series x_{dis} of this location has its maximum at time step 18 (Figure 5.7). The mean precipitation averaged over 500 samples for the second location also has maximum precipitation at this time step because the already disaggregated time series has been used as spatial conditioning values. The maximum in this location's x_{dis} occurs at time step 17 because the corresponding candidate had the smallest difference to the power of 4 after rescaling. For the third location, the mean precipitation of each time step is mostly higher than for the second location because not only one but two non-zero spatial conditioning values were utilized.

5.5.2 Correlograms of disaggregated time series

The correlogram functions were fitted with the disaggregated time series to compare them the to correlograms that were fitted to the observed data. Because the optimization to fit the correlograms is computationally very intensive, only the first year 1980 was utilized. Tests during model development showed that longer time series do not lead to substantially different correlograms.

Simulated cross correlation

Figure 5.9 shows the spatial cross correlograms of the four seasons for a temporal lag of $\tau = 0$ *min*. With the exception of the season JJA, cross correlation is overestimated in the disaggregated series and the number of spatial conditioning values m has little influence. The behavior is very similar for the other 5 lags τ and those plots are therefore not shown here.

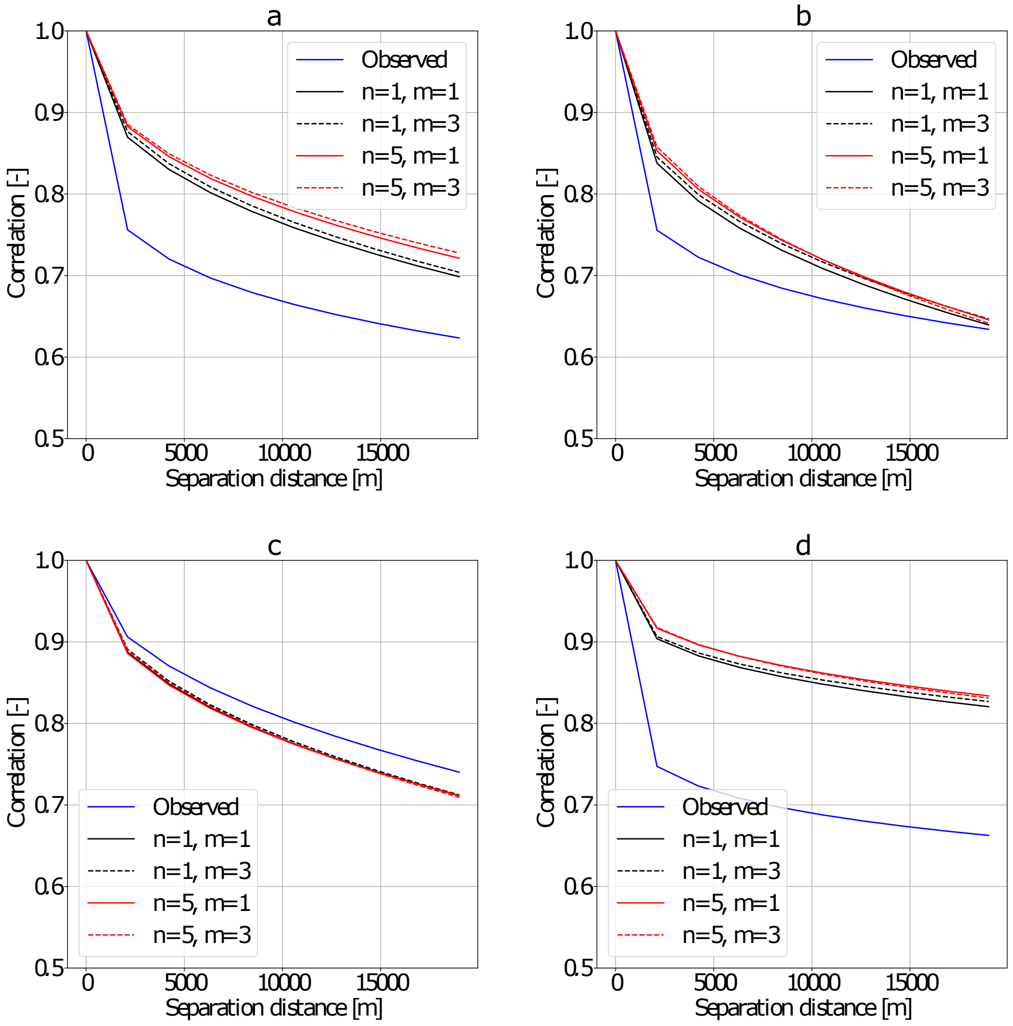

FIGURE 5.9: Fitted cross correlograms of observed (1951-2013) and disaggregated (1980-2009) 5 minute precipitation with a temporal lag of $\tau = 0$ *min* in the region of Freiburg in the seasons DJF (a), MAM (b), JJA (c) and SON (d).

Simulated auto correlation

For most seasons, the ACFs of the disaggregated time series are lower than the model fitted to the observed data. The ACFs generally improve slightly as the number of temporal conditioning values n increases (Figure 5.10). However, the results are ambiguous as the ACFs in the season JJA are nearly identical (c). Also, in the season SON, the ACF of the disaggregated time series is higher with $n = 5$, whereas $n = 1$ led to a very close agreement of the ACFs.

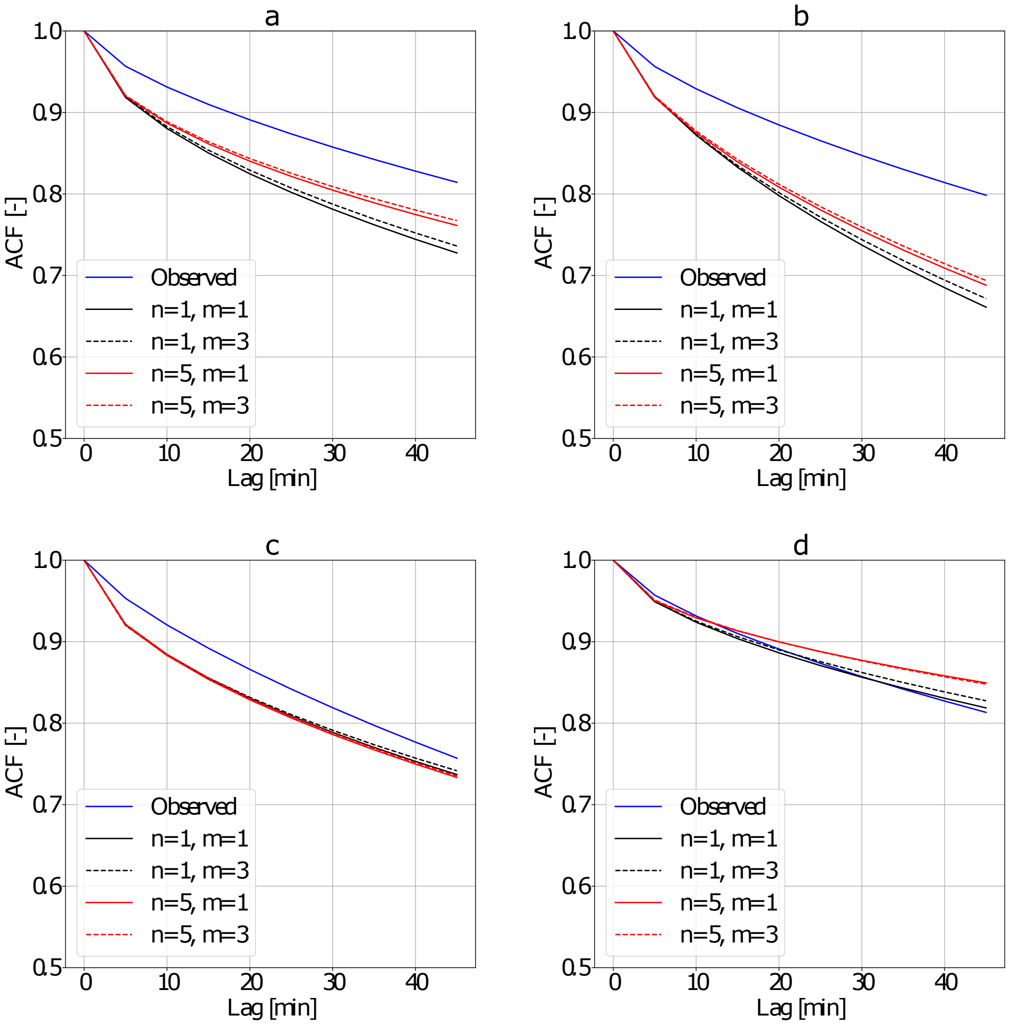

FIGURE 5.10: Fitted auto correlograms of observed (1951-2013) and disaggregated (1980-2009) 5 minute precipitation in the region of Freiburg in the seasons DJF (a), MAM (b), JJA (c) and SON (d).

Influence of bias corrected RCM on correlograms

It was found that the correlograms of the disaggregated time series remain nearly constant with different numbers of conditioning values. This behavior was assumed to be related to the high amount of dry time steps. To test this assumption, the disaggregated time series with the setup $m = 3, n = 5$ was altered by reshuffling. Since dry hours lead to 12 5-minute values that are zero, these values were left untouched. The amounts within wet hours were shuffled randomly. This data set is labeled as $m = 0, n = 0$ in the following. Fitting the lagged CCF (Figure B.2 in Appendix B) and ACF (Figure B.3 in Appendix B) with these shuffled time series revealed that the correlograms are very close to the ones shown above. This indicates that the large amounts of zeros dominate the correlogram fitting and that the influence of m and n on the correlograms is very weak. Another assumption that occurred during the evaluation was that the different spatio-temporal correlations that are introduced by the bias corrected hourly RCM time series contribute to the systematic differences between the correlograms of the disaggregated and observed time series. This was tested by calculating the ACFs and lagged CCFs of hourly observations and of the bias corrected hourly RCM simulations. The fitted CCFs can be found in Appendix B, Figure B.4. The fitted ACFs are presented in Appendix B, Figure B.5. The fitted ACFs of the bias corrected hourly RMC precipitation are relatively close to the observed ones but for the CCFs, larger differences exist and these differences in the hourly statistics propagate into the disaggregated time series as well.

5.5.3 Simulated distribution functions

Since the number of conditioning values only has a small influence on the temporal correlograms, an indirect evaluation via the distribution of aggregated precipitation was conceived. A comparison of the distribution of disaggregated precipitation to the observed distribution in Freiburg was performed for aggregation levels of 5, 10, 15 and 30 minutes (Figure 5.11). Here, the results for only one location is shown. The number of conditioning values are $m = 3$ and $n = 5$. For changing numbers of m and n, similar QQ-Plots were obtained which indicates that the influence of these parameters is rather small. The observed and simulated distributions are similar for the different aggregation levels. This indicates that the auto-correlation is respected in the simulations for the most part. However, as the aggregation level increases, the performance decreases slightly as the very high amounts are underestimated.

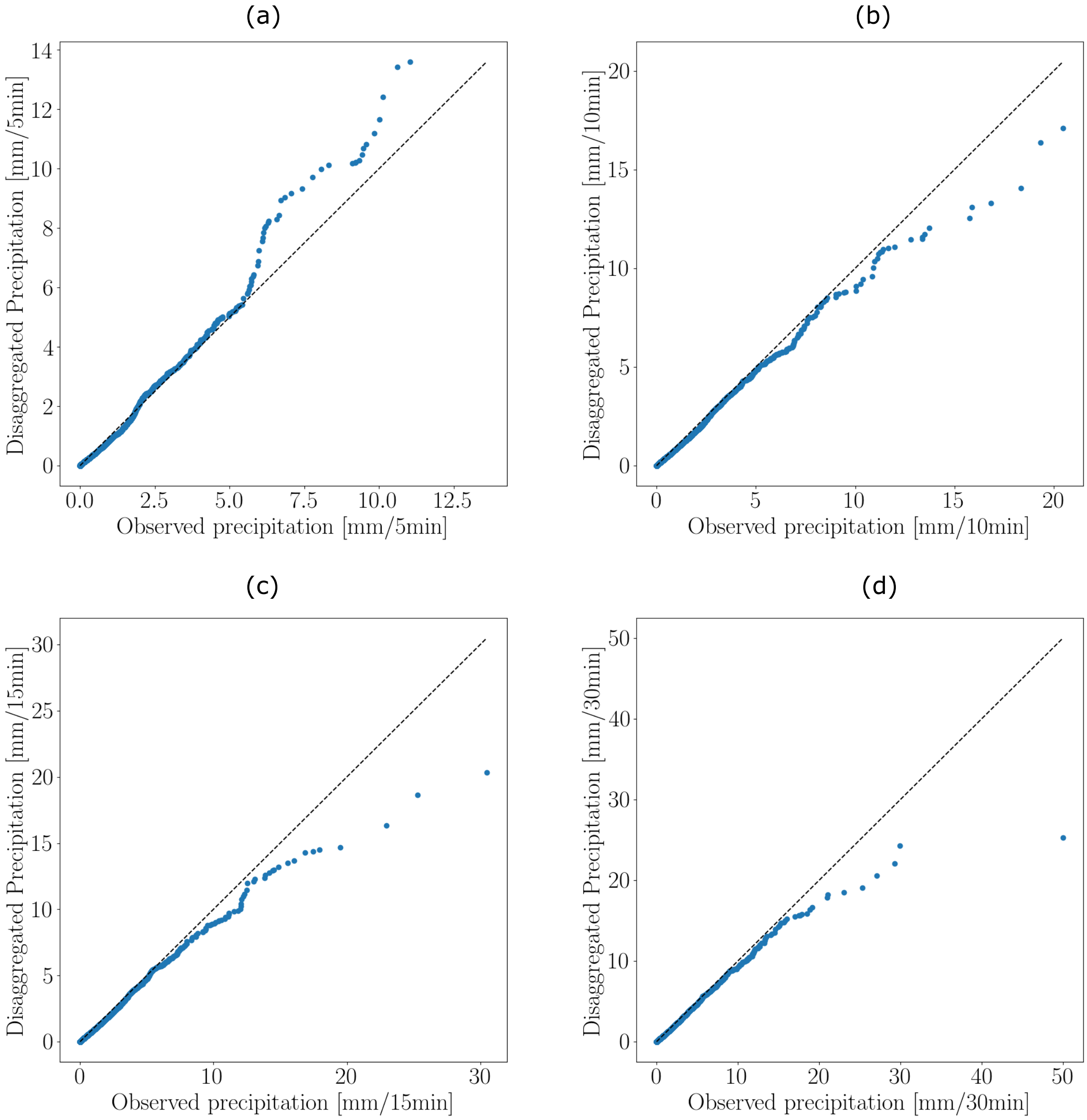

Figure 5.11: QQ-Plots of disaggregated (1980-2009) against oberseved (1951-2013) precipitation in a temporal resolution of 5 (a), 10 (b), 15 (c), 30 (d) minutes in Freiburg.

5.6 Summary and outlook

A novel copula-based technique that disaggregates hourly RCM precipitation to a resolution of 5 minutes has been developed for nine ungauged locations around the city of Freiburg im Breisgau, Germany. RCM precipitation was bias corrected with the geostatistical method presented in Chapter 3. The disaggregation model is based on the Gaussian Copula and the unknown parameters were estimated by spatial interpolation (distribution parameters) and pooling of all observations into one sample (ACF). This approach was chosen, so that a disaggregation can be performed for arbitrary, ungauged locations. Therefore, the model can be employed for other regions as well.

The distribution functions and correlograms of the disaggregated time series were shown to agree with the observed ones for the most part. However, larger differences were found in the correlograms of some seasons. This problem was partially inherited from the dependence structure of the hourly RCM simulations. A possible approach to reduce this problem would be the spatial recorrelation procedure presented by *Bárdossy and Pegram* (2012) but is unclear whether this technique would adversely affect the temporal dependence structure of the hourly RCM precipitation. If the model is not capable of matching the distribution functions of higher aggregation levels sufficiently well, the v-transformed Gaussian Copula (*Bárdossy and Li*, 2008) could be incorporated to strengthen the clustering of extreme values. In the presented case, the influence of the number of spatial (m) and temporal (n) conditioning values on the correlograms and distribution functions was very small. If the model behaves differently for another region, m and n can be set to larger values. Models that are based on the Gaussian Copula can be easily extended to higher dimensions since the correlation matrices are calculated via correlograms. However, as the dimensionality increases, the correlation matrices may become non-valid which is a common problem when dealing with large covariance and correlation matrices (e.g. *Higham et al.*, 2016). Also, while the presented disaggregation was easily manageable within approximately two days, the application to a large domain with thousands of locations, would not be particularly fast and could limit the applicability of the proposed model.

Other extensions are similar to the ones discussed in Chapter 3: the inclusion of anisotropy, circulation patterns or more seasons could improve the disaggregated time series if enough observed data is available to estimate all model components.

Chapter 6

Multivariate Vine Copula-based Bias Correction

In this chapter, a short introduction to a post-processing technique for RCM simulations of different meteorological variables is given. The method was developed to meet the demands of a hydrological model for the Berchtesgaden National Park which is a highly complex Alpine region. As the spatio-temporal resolution of the chosen RCM is sufficiently high for the impact model, a spatial downscaling or temporal disaggregation was not necessary but a new bias correction had to be developed. In the previous Chapters 3 and 5 of this thesis, only precipitation has been bias corrected for each location univariately. In the presented case, the physical processes that control snow melt and discharge are influenced by several meteorological variables that need to exhibit a realistic spatial, temporal or inter-variable dependence structure. To improve the inter-variable dependence structure of RCM simulations, a Vine Copula model has been developed to simulate a meteorological variable based on its dependence to other variables at the same location and time step.

6.1 Overview of multivariate bias correction methods

In a study by *Clark et al.* (2004), the Schaake Shuffle was introduced to reorder downscaled station scale precipitation and temperature. The limitation of this approach is that it requires historical data at all locations for the same time period as the simulated time series which are re-ordered. This is rarely the case in practice and future periods cannot be re-ordered in a straightforward manner. Furthermore, the Schaake Shuffle would lead to a reproduction of the observed time series if the transfer function of the univariate bias correction is perfect. These problems are partially addressed in *Vrac* (2018) who presented a less restrictive method to rearrange the variables. One time series was left unchanged and the remaining time series were bias corrected with an adapted Schaake Shuffle. A coarse gridded data set of precipitation and temperature was bias corrected independently at first to remove the bias of the distribution. Afterwards, the time series were rearranged with differing reference time series. The target order was obtained from a fine scale gridded data set in the same region. While the spatial correlation of the bias corrected time series was improved, the auto correlation worsened. Also, as in the study by *Clark et al.* (2004), long continuous time series of all variables are required at all locations. *Piani and Haerter* (2012) bias corrected temperature first and precipitation subsequently. The copula density of temperature and precipitation was utilized to correct the inter-variable dependence of the corrected time series.

6.2 Study region and data

The multivariate bias correction method presented in this chapter has been developed for the Berchtesgaden National Park in the southeast of Germany. In Figure 6.1, the 22 meteorological measurement stations of the region are shown. A list of the station names and coordiantes is given in Table C.1 in Appendix C.

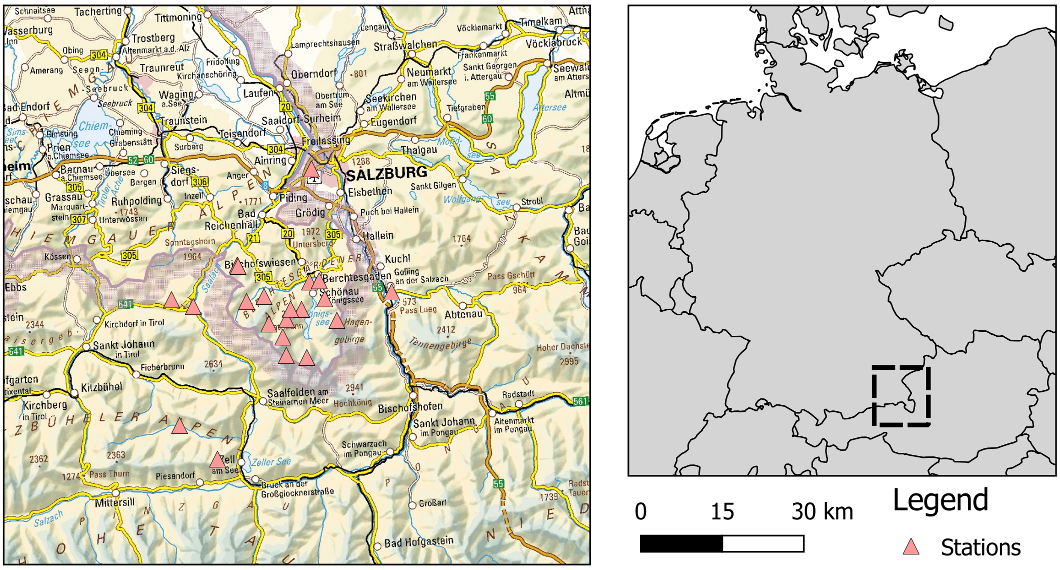

FIGURE 6.1: Locations of the 22 meteorological stations in the Berchtesgaden National Park. Background map retrieved from GeoBasis-DE / BKG (2019).

This alpine region exhibits very complex orography and accordingly the climatology of nearby locations can differ drastically for short separation distances. In addition to the high heterogeneity, a strong seasonality is present. Since the Berchtesgaden National Park is a climate sensitive region, it was investigated how the discharge will most likely change in the future. In a previous study by *Warscher et al.* (2013), the hydrological model WaSiM-ETH has been utilized to model the discharge in this complex region. It was found that an enhanced description of snow accumulation and redistribution processes improved the skill of the discharge simulations because these processes influence the amount of water that is available for discharge. For the follow-up study, the variables surface temperature (T), precipitation (P), relative humidity (H), wind speed (W) and shortwave downwelling radiation (SW) were selected as necessary input variables for WaSiM. The present bias correction method has been developed to provide meteorological time series with an improved inter-variable dependence structure for the impact model WaSiM.

The five variables have been measured at 22 sites in the Berchtesgaden National Park in the period 2001 to 2010. However, not all five variables were measured at all sites for the complete period because the data was collected from different meteorological services. Also, the permanent maintenance of a dense measurement network is very challenging in this complex Alpine region and therefore many data gaps exist. For instance, at 13 of the 22 stations not a single time step with all five variables exists. Furthermore, some variables have not been measured at all. For example, no measurements of T exist at two stations and none for P at ten stations. This data scarcity calls for an estimation of both the marginal distributions and the dependence structure for locations without enough data.

Simulations with a regional climate model provide meteorological information for all locations and time periods without observed data. For a complex and scarcely gauged region like the Berchtesgaden National Park, RCM simulations therefore allow for a better process understanding and planning of how to adapt to the projected climate impact. The RCM simulations were performed with WRF in a resolution of 5 km and 1 h. Three runs were performed: one run for the period 1980-2009 with the ERA-Interim reanalysis data which was also utilized in Chapter 5 (*Wagner and Kunstmann*, 2016) and two climate runs consisting of one control run (1980-2009) and one scenario run with the RCP 4.5 scenario (2020-2049) (*Warscher et al.*, 2019). The driving model for the control and scenario runs was the Earth-System Model MPI-ESM. For the 22 station locations, the closest grid cells were searched and the time series of T, P, H, W and SW were extracted to serve as an input for the bias correction routine.

6.3 Development of a new copula-based multivariate bias correction model

The existing multivariate bias correction methods require long observation archives of all required variables at all locations of interest which severely limits their applicability for future periods and ungauged locations. In the following, a bias correction approach which simulates one meteorological variable based on the dependence to other variables is given. The method utilizes Vine Copulas to take the highly different dependence structures into account. The difference of the proposed model to other bias correction techniques is that it requires less observed data and that one of the variables was simulated with a stochastic model and not taken from the RCM.
As illustrated in section 6.2, an estimation of the marginal distribution is required for several ungauged locations. With these estimated CDF parameters, a univariate Quantile-Mapping (QM) was performed for all five variables simulated by the RCM for all 22 stations. Because the physical processes that were to be studied with the subsequent impact model WaSiM are determined by the interaction of several variables, the dependence measure Kendall's τ_K of all variable pairs at was calculated to evaluate how well the dependence structure of the variables was reproduced by the univariately corrected RCM. This process is schematically illustrated in Figure 6.2.

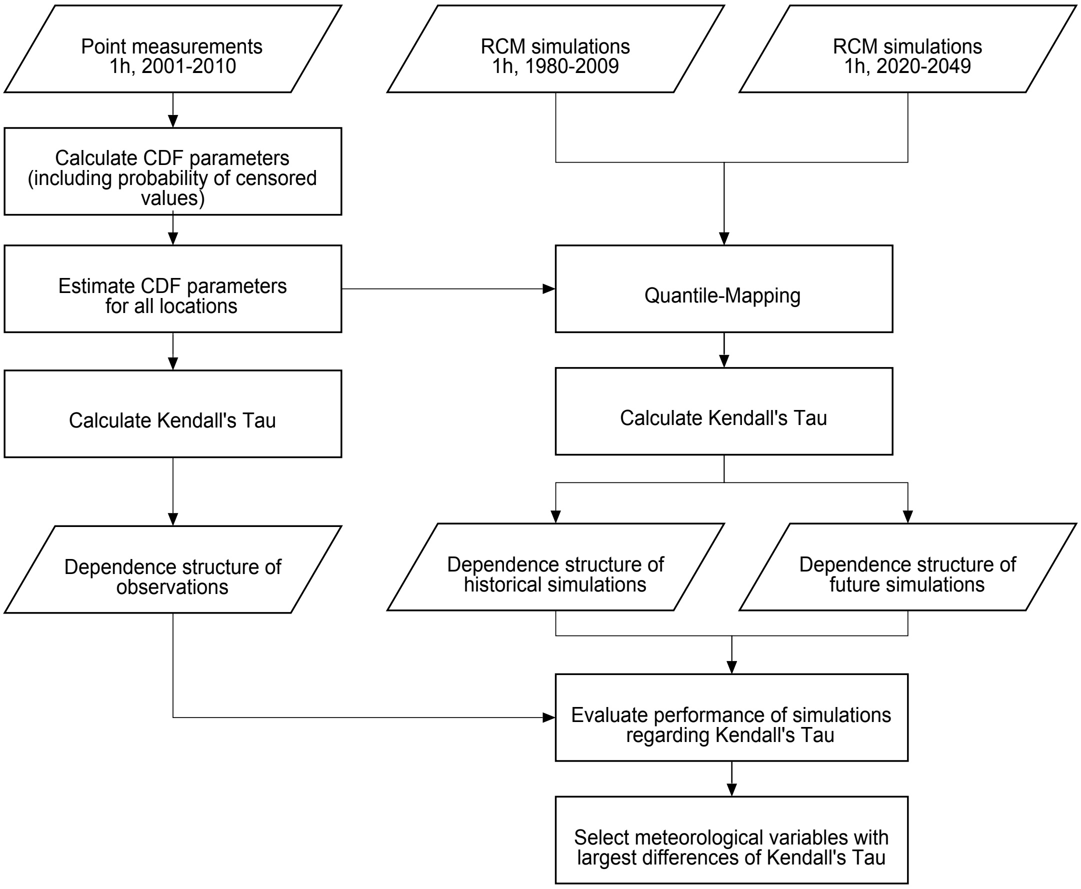
Point measurements
1h, 2001-2010
RCM simulations
1h, 1980-2009
RCM simulations
1h, 2020-2049
Calculate CDF parameters
(including probability of censored values)
Estimate CDF parameters
for all locations
Quantile-Mapping
Calculate Kendall's Tau
Calculate Kendall's Tau
Dependence structure of observations
Dependence structure of historical simulations
Dependence structure of future simulations
Evaluate performance of simulations regarding Kendall's Tau
Select meteorological variables with largest differences of Kendall's Tau

FIGURE 6.2: Flowchart of the univariate bias correction with Quantile-Mapping and selection of variable for post processing.

Once the variable, which contributes the most to the dependence bias, has been identified, it was removed from the QM time series. As a replacement, this variable was then simulated conditionally on the other univariately corrected QM values. This model is labeled as QMV (Quantile Mapping + Vine Copula). More details on the model selection and motivation can be found in sections 6.4 and 6.5.

The model is based on Vine Copulas which represent a multivariate copula as a decomposition of several bivariate copulas (*Aas et al.*, 2009). As was shown in Chapter 2, section 2.5, a conditional PDF f_c can be calculated from copula densities c:

$$f_c(u_1|u_2,...,u_n) = \frac{c(u_1,...,u_n)}{c(u_2,...,u_n)} \tag{6.1}$$

As with the Gaussian copula, the denominator was dropped from the calculation and only the four dimensional copula density was employed to calculate f_c. Numerical integration and rescaling to values between 0 and 1 then yielded the conditional CDF F_c. As in the previous chapters, F_c was inverted with a uniformly distributed number w to simulate the CDF value of the problematic variable. In contrast to the previous chapters, only one realization is provided. During model development a random number w was employed at first and while this led to time series with an improved dependence structure between the variables, the time series were very volatile and unrealistic. Therefore, w was taken from the conditional distribution F_c of the RCM.

The four-dimensional copula density $c(u_1, u_2, u_3, u_4)$ was constructed as a C-Vine for both observed and univariately bias corrected RCM time series.

$$\begin{aligned} c(u_1, u_2, u_3, u_4) = {} & c_{12}(F_1(x_1), F_2(x_2)) \cdot c_{23}(F_2(x_2), F_3(x_3)) \cdot c_{34}(F_3(x_3), F_4(x_4)) \\ & \cdot c_{13|2}(F(x_1|x_2), F(x_3|x_2)) \cdot c_{24|3}(F(x_2|x_3), F(x_4|x_3)) \\ & \cdot c_{14|23}(F(x_1|x_2, x_3), F(x_4|x_2, x_3)) \end{aligned} \tag{6.2}$$

The last term in this equation requires the values of the three-dimensional CDFs $F(x_1|x_2, x_3)$ and $F(x_4|x_2, x_3)$ which were calculated from the corresponding three dimensional Vine Copula densities $c(u_1, u_2, u_3)$ and $c(u_4, u_2, u_3)$ respectively. For instance, the first density $c(u_1, u_2, u_3)$ is given as:

$$c(u_1, u_2, u_3) = c_{12}(F_1(x_1), F_2(x_2)) \cdot c_{23}(F_2(x_2), F_3(x_3)) \cdot c_{13|2}(F(x_1|x_2), F(x_3|x_2)) \tag{6.3}$$

In the above two equations, bivariate copula densities for a given third variable are employed, e.g. $c_{13|2}$. A simplifying assumption of a constant copula that is not influenced by the value of the third variable has to be made almost always, especially if the calibration data set is small. That this necessary assumption generally performs well has been demonstrated by *Haff et al.* (2010).
The ordering of the variables of a Vine Copula is not straightforward. Also, the computational demand to calculate one multivariate copula density and the number of possible Vine structures increases drastically with the dimensionality of the copula. *Nagler et al.* (2016) presented two algorithms to select the Vine structure based on τ_K between variable pairs to reduce the number of computation steps.

Calculating the four dimensional copula density $c(u_1, u_2, u_3, u_4)$ for each time step can lead to high run times. The simulation routine has been written in Python and conditional CDFs of sets of conditioning values were saved in dictionaries to reduce computation time. The three conditioning values (u_2, u_3, u_4) were varied between 0.0001 and 0.9999 with 101 equidistant steps, leading to $101^3 = 1030301$ combinations of conditioning values for each season. During the simulation, the conditioning CDF values were rounded to two digits to access the corresponding conditional CDF from the dictionary. For instance, the conditioning values $(u_2 = 0.3251, u_3 = 0.9813, u_4 = 0.2510)$ get the key `0.33_0.98_0.25` to access the conditional CDF in the dictionary.

6.4 Calibration of the copula-based multivariate bias correction model

The calibration was carried out for the time period 2001-2010 with four seasons (DJF, MAM, JJA and SON). For each meteorological variables (T, P, H, W, SW), different parametric distribution functions were fitted. For P, W and SW it was also necessary to calculate the probability of a censored value (e.g. $P = 0\ mm/h$ or $SW = 0\ W/m^2$). The best fitting CDF was chosen by calculating the squared differences between the observed values and values that were simulated from the fitted CDFs. For ungauged locations, the parameters were estimated based on the site's elevation with linear regression. The best fitting distributions are listed in Table 6.1.

Variable	DJF	MAM	JJA	SON
T	Normal	Normal	Log-Normal	Normal
P	Weibull	Weibull	Weibull	Weibull
H	Beta	Beta	Beta	Beta
W	Weibull	Weibull	Weibull	Weibull
SW	Exponential	Normal	Exponential	Exponential

TABLE 6.1: Selected CDFs for the five observed meteorological variables (2001-2010).

With the fitted distribution functions and probabilities of censored values, all observed variables were transformed to CDF values in $[0, 1]$ (see Chapter 2, subsection 2.2.4). The RCM simulations were fitted with a KDE-CDF.

In a next step, Kendall's τ_K was calculated for the observed and univariately bias corrected QM time series of the reanalysis run. To include censored values, a small random noise was added to the CDF values as was done by *Pham et al.* (2015). It should be noted that random noise that was added to remove the ties can lead to slightly different values of τ_K with each run but these differences were found to be usually less than 0.01 in the case study presented in this thesis.
In order to find the temporal lag at which the data can be regarded as independent identically distributed (iid), the partial auto correlation function (PACF) of the normalized time series was calculated. Details on the PACF are given in *von Storch and Zwiers* (1999). It was found that the PACF after 30 hourly time steps is independent for all variables, so only CDF value pairs of two variables that are at least separated by 30 time steps were sampled. Due to the scarcity of the observation data set, all CDF value pairs of the region were pooled into one set to estimate the average observed dependence structure of the region. The absolute difference $|\Delta\tau_K|$ between the observed and simulated values of τ_K were calculated to investigate how well the QM series agree with the observed dependence structure. The values of τ_K and $|\Delta\tau_K|$ are listed in Table 6.2. Note that the numbers have been rounded to two digits but the calculations were performed with the actual values.

	T,P	T,H	T,W	T,SW	P,H	P,W	P,SW	H,W	H,SW	W,SW
DJF										
Observed	0.00	-0.26	0.10	0.10	0.09	0.04	0.01	-0.07	-0.10	0.02
Reanalysis-QM	0.02	-0.15	0.11	0.13	0.07	0.07	-0.01	-0.03	-0.04	-0.08
$\lvert\Delta\tau_K\rvert$	0.02	0.11	0.01	0.03	0.01	0.03	0.02	0.04	0.06	0.09
MAM										
Observed	-0.06	-0.38	0.04	0.25	0.12	0.04	-0.06	-0.11	-0.26	0.05
Reanalysis-QM	-0.02	-0.23	-0.07	0.24	0.07	0.02	-0.03	-0.10	-0.21	-0.04
$\lvert\Delta\tau_K\rvert$	0.04	0.15	0.12	0.01	0.05	0.01	0.03	0.01	0.05	0.09
JJA										
Observed	-0.09	-0.50	0.01	0.30	0.13	0.05	-0.04	-0.08	-0.31	0.05
Reanalysis-QM	-0.01	-0.37	-0.10	0.29	0.10	0.02	-0.01	-0.04	-0.29	0.04
$\lvert\Delta\tau_K\rvert$	0.08	0.14	0.11	0.01	0.03	0.02	0.03	0.04	0.02	0.01
SON										
Observed	-0.07	-0.33	0.00	0.19	0.10	0.05	-0.01	-0.09	-0.19	0.04
Reanalysis-QM	-0.03	-0.26	-0.02	0.21	0.11	0.03	-0.01	-0.08	-0.18	-0.03
$\lvert\Delta\tau_K\rvert$	0.05	0.07	0.02	0.02	0.00	0.02	0.00	0.01	0.01	0.07

TABLE 6.2: Kendall's τ_K of the ten meteorological variable pairs of the observations and the univariately bias-corrected QM reanalysis run.

To investigate the strength of dependence of the individual variables to the other variables at the same location and time step, the absolute values of τ_K were summed up to $\Sigma|\tau_K|$ for each observed variables. As Table 6.3 shows, the wind speed W has very low sums $\Sigma|\tau_K|$ across all seasons. Therefore, W was not utilized because it offers very little explanatory value and because the complexity of Vine Copula models increases heavily with more dimensions. In order to preserve a lean model structure, W was dropped as a potential conditioning value and for further analyses in the evaluation.

Variable	DJF	MAM	JJA	SON
T	0.466	0.734	0.913	0.597
P	0.134	0.274	0.310	0.243
H	0.521	0.873	1.021	0.711
W	0.224	0.234	0.186	0.183
SW	0.235	0.618	0.699	0.433

TABLE 6.3: Sum of absolute values of τ_k of the individual observed variables.

The absolute differences of τ_K between observations and the QM series given in Table 6.2 were summed up to $\Sigma|\Delta\tau_K|$ to identify the RCM variable which exhibits the largest deviations from the observed dependence structure. The results are given in Table 6.4. For the first two seasons DJF and MAM, H is the variable which shows the largest deviations from the observed dependence structure. In the seasons JJA and SON, T performs worst with H being the second worst variable regarding τ_K.
Thus, H was chosen as the variable to be simulated with the Vine Copula for the seasons DJF and MAM. The conditional CDF $F_c(H|T, SW, P)$ was built from the copula density $c(u_H, u_T, u_{SW}, u_P)$ of the four variables. For the seasons JJA and SON, T was simulated conditional on P, H and SW.

Variable	DJF	MAM	JJA	SON
T	0.158	0.205	0.230	0.140
P	0.059	0.120	0.140	0.051
H	0.185	0.249	0.189	0.087
SW	0.119	0.099	0.068	0.031

TABLE 6.4: $\Sigma|\Delta\tau_K|$ of the univariately bias corrected reanalysis run to the observations for individual variables.

The copula parameters were calculated from Kendall's τ_K (see Chapter 2, section 2.5). The copula families that were fitted this way are the Ali-Mikhail-Haq (AMH), Clayton (Cla), Frank (Fra), Farlie-Gumbel-Morgenstern (FGM) and the Gumbel (Gum) copulas. The best-fitting copula was selected by minimizing the squared differences between the fitted and empirical copulas. A visual example of the copula densities fitted to T, SW in the season SON was already given in Chapter 2.5, section 2.5. As was shown in Table 6.2, the sign of τ_K remains mostly identical across the four seasons but the magnitude of dependence changes. For example, the dependence of T and SW is much stronger in summer than in winter. The copula families are also mostly identical across the four seasons (Table 6.5).

	T,P	T,H	T,SW	P,H	P,SW	H,SW
DJF						
Observed	FGM	Fra	Gum	Cla	AMH	FGM
QM	Cla	FGM	AMH	Cla	AMH	FGM
MAM						
Observed	FGM	Fra	Gum	Cla	Fra	Fra
QM	FGM	Fra	Gum	Cla	FGM	FGM
JJA						
Observed	Fra	Fra	Gum	Cla	Fra	Fra
QM	Fra	Fra	Gum	Gum	Fra	Fra
SON						
Observed	FGM	Fra	Gum	Cla	Fra	FGM
QM	Fra	Fra	Gum	Gum	Fra	FGM

TABLE 6.5: Selected copulas of the four variables T, P, H, SW in the observed and QM time series.

To illustrate the model, eight different conditional CDFs $F_c(H|T, SW, P)$ fitted to the observed data are shown in Figure 6.3 for the season DJF. The three conditioning values were set to either 0.05 or 0.95 to demonstrate the influence of low and high conditioning values on the resulting conditional CDF. Blue corresponds to low temperature and red to high temperature, low precipitation is plotted as a dotted line and high precipitation as a solid line. Low values of short wave radiation are labeled with a circle marker and high values with a plus sign.

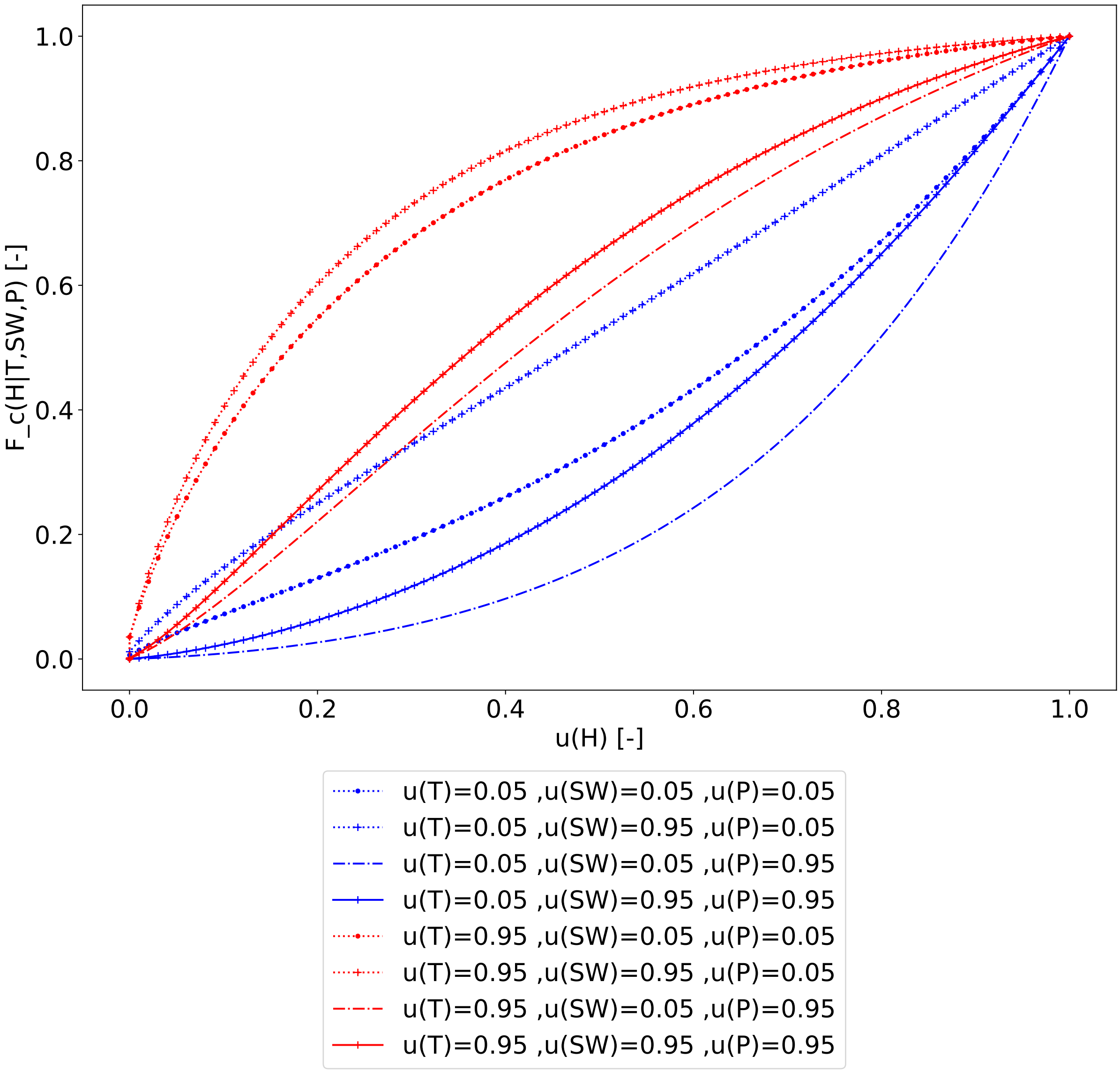

FIGURE 6.3: Conditional CDFs of relative humidity in the season DJF for different conditioning values.

It can be seen that the conditional CDF of H is influenced by all three variables to differing degrees. Low temperature (blue) leads to higher simulated values of H than high temperature (red) since the dependence parameter is $\tau_K = -0.26$. The influence of SW is also negative ($\tau_K = -0.10$), so low values of SW tend to result in high values of H. Precipitation has a positive dependence parameter ($\tau_K = +0.09$) and accordingly, the density of the conditional CDF is shifted towards higher values of H for high precipitation values.

6.5 Evaluation of the copula-based multivariate bias correction model

The copulas of variable pairs and the corresponding Vine Copulas were calculated for the observed time series. For the RCM simulations, the reanalysis and control runs were utilized to build the copulas. As was shown in Table 6.2, the QM series of the reanalysis run show deviations in τ_K to the observations. For the seasons DJF and MAM, H deviates the most from the observed dependence structure (Table 6.4). The QM time series of T, P, W and SW were not altered and H was simulated. For JJA and SON, T deviates the most and the QM series of P, W, H and SW remain constant. Thus, τ_K of these unchanged variables remains constant after the application of the QMV model. For the seasons DJF and MAM, the values of τ_K of the variable pairs (H, T), (H, P) and (H, SW) change. For JJA and SON, τ_K of (T, P), (T, H) and (T, SW) change.

The performance of the Vine Copula model was evaluated by calculating Kendall's $\Sigma|\Delta\tau_K|$ for the univariately bias corrected (QM) and for the multivariately bias corrected time series (QMV). The added value of QMV is demonstrated in Table 6.6. The percental improvement was calculated by relating $\Sigma|\Delta\tau_K|$ of QMV to the one of the univariate QM series.

	DJF	MAM	JJA	SON
Reanalysis	50.6%	40.0%	41.4%	66.0%
Control	47.6%	48.9%	22.8%	28.0%
Scenario	49.7%	50.9%	44.9%	1.9%

TABLE 6.6: Percental improvement of the absolute difference between observed and simulated Kendall's τ_K of QMV in comparison to QM.

A graphical demonstration of how QMV changes the values of τ_K is given in Figure 6.4 for the scenario run. As nearly all values of τ_K are closer to the bisecting line, it can be seen that the QMV approach leads to a more realistic dependence structure than the standard QM method.

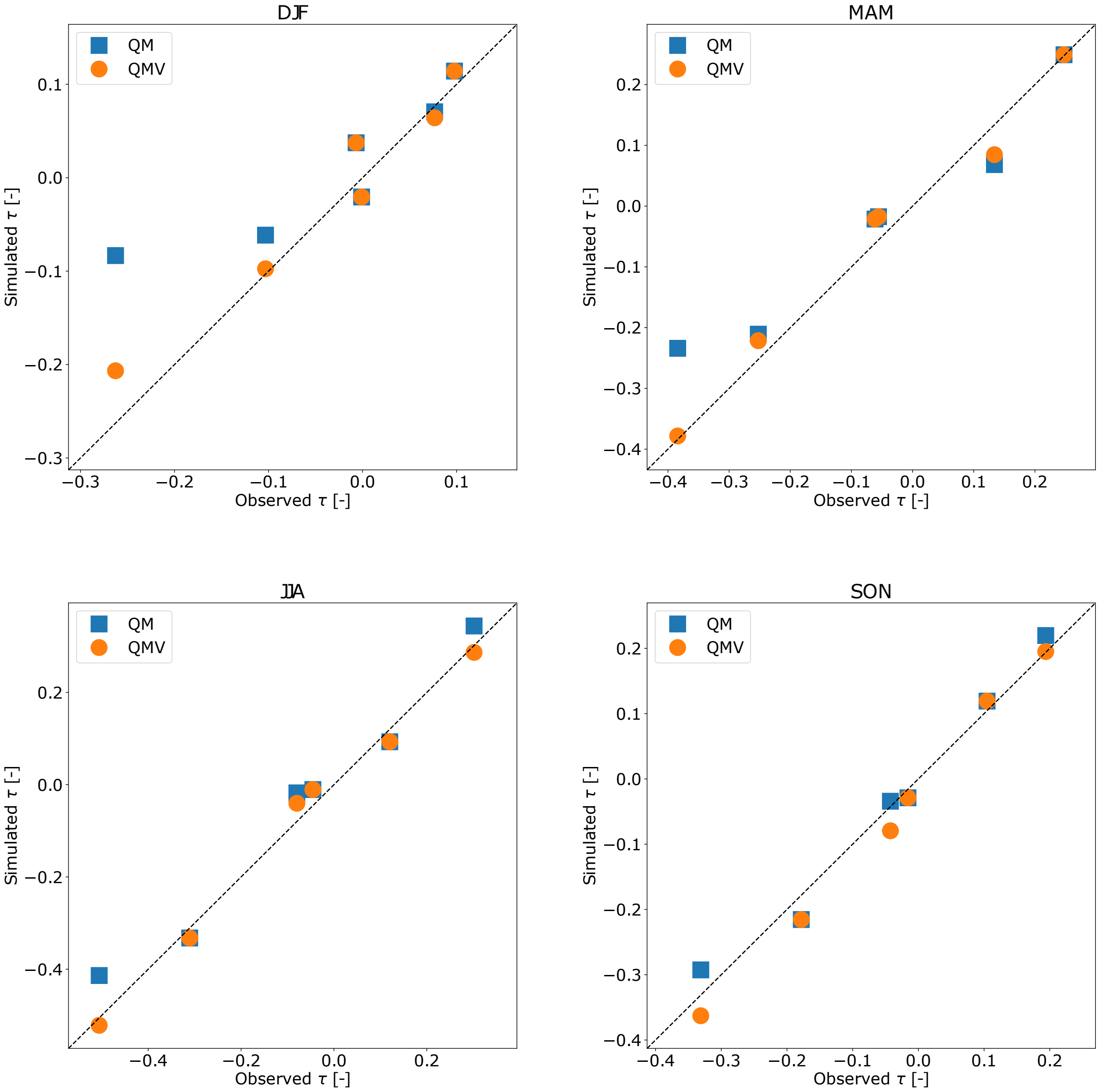

FIGURE 6.4: Scatter plot of τ_K of QM and QMV applied to the scenario run (2020-2049) against observations in the seasons DJF (a), MAM (b), JJA (c) and SON (d).

6.6 Summary and outlook

This chapter presented an approach that corrects the univariate bias and improves the multivariate inter-variable dependence structure of four meteorological variables simulated by an RCM. The method is based on Vine Copulas which decompose the multivariate copula into pair copulas. A standard univariate Quantile Mapping (QM) was performed as a baseline for the Vine Copula model (QMV). The meteorological variable with the largest dependence bias was simulated conditional on the remaining univariately bias corrected QM-variables. These conditioning values remained constant, while the problematic variable was simulated. For DJF and MAM, hourly relative humidity was simulated conditional on univariately bias corrected temperature, precipitation and short wave radiation. For JJA and SON, temperature was simulated conditional on precipitation, relative humidity and short wave radiation. It was shown that the dependence structure of the QMV time series generally matches the observed one more closely than the univariately corrected QM series. The sum of absolute differences of Kendall's τ_K was reduced by up to 66.0%. Therefore, Vine Copulas seem to be a promising method to model meteorological variables with complex dependence structure. The application of this simulation technique is of course not limited to post processing RCM simulations but it could be used for example to estimate missing values in observation data sets conditional on available variables.

The limitations of the presented method are discussed here since multivariate techniques can adversely affect dependence structures which are not part of the model structure as was shown in Chapter 3, section 3.1. In the presented case, the conditioning values stem from physically-based RCM simulations which exhibit spatio-temporal correlation. Thus, the spatio-temporal correlation of the conditioning values propagates into the simulated relative humidity to some extent. However, for an independent simulation, e.g. as a Weather Generator, it would be necessary to condition the simulation also on previous time steps and spatial neighbors. *Gräler* (2014) presented a geostatistical approach to estimate the Vine Copula parameters for the simulation of spatial fields which is already very complex. Adapting such an approach for a conditional simulation that respects the spatial, temporal and inter-variable dependence would increase the complexity tremendously. For instance, conditioning on 3 variables (e.g. T, P, SW), 3 time steps and 3 spatial neighbors would lead to a 10-dimensional Copula with unknown dependence parameters. Also, the Vine structure would be unknown and the selection of the best one would lead to huge computation times since a 10-dimensional Copula can be decomposed in 1.8110^{34} different ways (*Nagler et al.*, 2016). Also, the storage demand would increase significantly. As mentioned above, the conditional CDFs were saved in Python dictionaries. For 4 seasons and a discretization step of 0.01, two dictionaries of 3.5 *Gb* file size were obtained. For a 10-dimensional copula, the file size would be approximately 3.510^{12} *Gb*. For these reasons, the simulation was restricted to only 3 conditioning values to keep the computational and storage demand and model complexity manageable. An extended version of the model could include atmospheric circulation patterns or separate statistics and conditional distributions for different times of the day, e.g. night and day, if enough observed data is available to calibrate all model components. As illustrated before, this would however also lead to a huge increase of the computational and storage demand.

Chapter 7

Summary and conclusions

In this thesis, four newly developed post-processing techniques for time series of meteorological variables simulated by RCMs have been presented. Aside from the bias correction technique in Chapter 3, the models are based on copulas. In the following, the research questions, that were stated in Chapter 1, are addressed:

1. **How can multivariate copulas be utilized to increase the spatio-temporal distribution and to improve the dependence structure of RCM simulations?** Copulas constitute a flexible modeling approach due to the separation of the marginal distributions from the dependence structure. This two step procedure allows for an individual fitting of distribution functions to arbitrary meteorological variables in the first step. From the huge number of copula models, the one which best represents the dependence structure, is selected in the second step. However, extending regular, bivariate copulas to higher dimensions reduces the number of available copula models. Two multivariate copula models were employed in this thesis to condition the simulation of an unknown meteorological variable on more than one known value, namely the Gaussian Copula (*Bárdossy and Li* (2008), Chapter 4 and Chapter 5) and Vine Copulas (*Aas et al.* (2009), Chapter 6). All developed models take RCM simulations as input variables. Additional data with a desired property was used to construct copulas to model meteorological variables and to provide the post-processed time series.
   With the spatial downscaling model presented in Chapter 4, fine scale precipitation fields, that resemble physically downscaled precipitation fields, were simulated. The Gaussian Copula was conditioned on coarse scale RCM precipitation fields and ensembles were simulated to address the uncertainty of the physical downscaling. The motivation for this method was the high computational demand of physical downscaling and thus a faster, stochastic model was developed to provide a surrogate for the fine scale RCM precipitation fields.
   In Chapter 5, a temporal disaggregation model was presented. This model was employed to simulate precipitation time series in a temporal resolution of 5 minutes that agree with the hourly amounts of a bias corrected RCM simulation. In contrast to the spatial downscaling model, observed data was used to calibrate the model, as the aim of this technique was the simulation of precipitation time series in a finer temporal resolution than what was attainable with the RCM. The spatio-temporal conditioning values are simulated precipitation values in the target resolution of 5 minutes from spatial neighbors or at the same location.
   The other multivariate copula technique, that was employed in this thesis, is a Vine Copula which constructs a multivariate copula from bivariate pair copulas. This copula was utilized in Chapter 6 to model the dependence structure

of four meteorological variables at the same location. Univariately bias corrected RCM simulations were evaluated regarding this dependence structure and the variable which contributed most strongly to the dependence bias was removed. Afterwards, it was simulated from a four-dimensional Vine Copula so that the bias corrected time series respect the observed dependence structure more closely. In this application, Vine Copulas were chosen to take the highly differing dependence structures of variable pairs into account: for instance, one variable pair may have a symmetrical negative copula while another pair exhibits positive dependence with a strong clustering of extreme values.

2. **How well do the stochastic simulations agree with observed univariate and multivariate statistics?**
   The performance of the post-processing techniques was evaluated by analyzing both univariate (distribution functions in different aggregation levels) and multivariate (dependence structure / spatio-temporal correlation) statistics. The evaluation sections demonstrated that the developed models lead to a refinement of the RCM simulations regarding the univariate distributions, spatio-temporal resolution or inter-variable dependence structure.
   For extreme events or locations that show a deviant behavior from neighboring locations, the common problems of all statistical estimation techniques and parametric distribution functions remain. For instance, the highest daily RCM precipitation amounts were not captured by the Gamma distribution (Chapter 4). Also, the kriged CDF parameters resulted in some distributions that did not match the observed ones during cross validation (Chapter 3). Some discrepancies can also be related to errors in the RCM simulations which propagate into the results of the post-processing techniques. For instance, the 5 minute spatial correlograms in Chapter 5 are dominated by the hourly spatial dependence structure of the RCM. The disaggregated 5 minute precipitation time series have similar distributions as in the observations but as they were aggregated to longer time intervals, the high values decreased which shows that the simulated time series do not cluster as strongly as in reality. This problem may be lessened by using a non-Gaussian multivariate copula like the V-transformed copula (*Bárdossy and Li*, 2008) or Vine Copulas. However, those multivariate copulas lead to an enormous increase of the model complexity.
   As illustrated in Chapter 6, multivariate dependence can be related to spatial, temporal or inter-variable dependence. Readjusting one property with a statistical technique often comes at a loss regarding another property and this problem is not exclusive to copula models. For example, the spatial recorrelation presented by *Bárdossy and Pegram* (2012) would presumably lead to a change of the temporal structure. *Vrac* (2018) illustrated how the temporal correlation suffers from a rearrangement which aimed at an improved spatial dependence structure. The modeling strategy must therefore also be based on the importance of different statistical aspects of the final product that is required for impact studies.

3. **What are the advantages of the developed models compared to other approaches and what are the limiting factors for extensions and applications to other variables?**
   The hugest advantage of copulas over other stochastic approaches is the separation of the univariate distribution functions and the multivariate dependence structure. This way, it is possible to focus on one part at a time and to model

arbitrary meteorological variables as long as the marginal distribution can be fitted. In practice, the univariate distributions are modeled at first and then the best fitting copula is selected. The Gaussian Copula is the most common multivariate copula and it has shown to perform well regarding most evaluated statistics. Furthermore, it is comparatively simple as it is only determined by a correlation matrix. Correlation coefficients are easier interpretable than for example the parameters of a BLRP model. The selection of a suitable copula family is commonly based on choosing the one which minimizes the differences to the empirical copula. This does however not automatically imply that the selected copula is sufficiently good - especially for high resolution precipitation with many dry values. Censored variables like 0 *mm* precipitation can be included into the Gaussian Copula with the MLM-based correlogram estimation presented in Chapter 2 but for other copulas, workaround solutions are typically employed, e.g. adding random noise as in *Pham et al.* (2015). If such an approach results in a satisfactory representation of the observed distributions and dependence structure needs to be evaluated for each application individually.

For higher dimensions, less copula families are available, mainly the Gaussian Copula, its V-transformed version and the Student Copula. In theory, the dimensionality of the Gaussian Copula can be arbitrarily high but in practice, non-valid correlation matrices may occur as the number of dimensions increases. Therefore, the number of conditioning values must be limited. In spatial interpolation techniques, it is common to use only a few nearby stations to estimate an unknown value and this is also advisable when using the Gaussian Copula as was shown in *Bárdossy and Li* (2008). Another way to model multivariate dependence structures are Vine Copulas. They are more flexible in describing different forms of dependence than the Gaussian Copula but the model structure is much more complex and not all possible decompositions can be tested once a certain dimensionality has been reached (*Nagler et al.*, 2016). Furthermore, the calculation of conditional distributions of Vine Copulas of all possible conditioning values can quickly lead to very high computation times and storage demands as the dimensionality increases.

4. **How can the model parameters be estimated for ungauged location in a study region and what are the limitations?**
The adaptation of the models to different study regions is possible but some adjustments will be necessary depending on the case study: the sample size of observed data, seasonality and the importance of different aspects of the desired simulations must all be taken into account so that all model components are robust. Parameter estimation strategies were conceived for the different case studies. For instance, the individual elements of the correlation matrix of the Gaussian Copula were calculated with spatial or temporal correlogram models. This approach was chosen so that the correlation of ungauged locations can be estimated and to keep the correlation matrices invertible. As the dependence structure and marginal distributions were described by very few parameters, they can often be estimated rather well for arbitrary locations. When more parameters are necessary to model the univariate distributions, geostatistical estimation techniques may become unstable as was shown in Chapter 3. Likewise, if a region is highly heterogeneous, the estimation of local distribution functions and correlograms can become very uncertain for locations without nearby observed data. In Chapter 6, all available observed data

> was pooled into one sample to obtain an average inter-variable dependence structure for the presented data-scarce region. This was deemed a justified approach as the study region is rather small but transferring the dependence structure to a completely different region may result in poor results. Therefore, some observed data, that was measured somewhere in the vicinity of the target locations, is always necessary. The decline of observation networks reported in *Lorenz and Kunstmann* (2012) therefore poses a problem for the development and application of both RCM simulations and the presented post-processing techniques, especially in complex terrains. Also, extreme values and statistical trends in observed time series provide valuable information for modeling purposes and climate change projections.

To conclude, several novel computationally efficient statistical and stochastic methods have been developed and applied to case studies in different regions of the world. The techniques are transferable to other regions as they employ observable statistical measures like correlation. The bias correction technique in Chapter 3 aimed at the generation of a surrogate distribution to estimate the unknown local climatology. In other cases of limited data availability, data was pooled to derive for example the average auto correlation of a given region (see Chapter 5) or the inter-variable dependence structure (see Chapter 6).

Copulas were chosen as the mathematical basis of most models and their potential to successfully post-process RCM simulations was shown. The application of copula-based post-processing models to RCM simulations is however not straightforward and involves a lot of trial and error and compromises as the quality of the post-processed products is influenced by many factors: How well can the observed distribution functions be fitted and estimated with the available data, are the copulas capable of representing the actual dependence structures, are erroneous spatio-temporal dependence structures inherited from the RCM simulations, what is the required dimensionality and which statistical aspects are most important for subsequent applications?

For the presented case studies, the techniques are sufficiently fast as most computations were performed on a single computer within a few hours or days. A comparison of the observed and simulated distribution functions showed that the models are generally in close agreement to the theoretical or observed distribution functions and that the dependence structure can be approximately reproduced for different meteorological variables on different scales. Therefore, copulas can be regarded as a valuable tool to enhance the applicability of RCM simulations for very diverse impact studies.

Appendix A

Appendix to Chapter 3 - Geostatistical bias correction of RCM precipitation

Institute	Driving Model	RCM	Beginning	End
CCCMA	CCCMA-CanESM2	CanRCM4 v4	1950-01-01	2005-12-31
CLMcom	CNRM-CERFACS-CNRM-CM5	CCLM4-8-17 v1	1950-01-01	2005-12-31
CLMcom	ICHEC-EC-EARTH	CCLM4-8-17 v1	1949-12-01	2005-12-31
CLMcom	MOHC-HadGEM2-ES	CCLM4-8-17 v1	1949-12-01	2005-12-30
CLMcom	MPI-ESM	CCLM4-8-17 v1	1949-12-01	2005-12-31
DMI	ICHEC-EC-EARTH	HIRHAM5 v2	1951-01-01	2005-12-31
DMI	NCC-NorESM1-M	HIRHAM5 v1	1951-01-01	2005-12-31
KNMI	ICHEC-EC-EARTH	RACMO22T v1	1950-01-01	2005-12-31
KNMI	MOHC-HadGEM2-ES	RACMO22T v1	1950-01-01	2005-12-30
MPI-CSC	ICHEC-EC-EARTH	REMO2009 v1	1950-01-01	2005-12-31
MPI-CSC	MPI-ESM	REMO2009 v1	1950-01-01	2005-12-31
SMHI	CCCMA-CanESM2	RCA4 v1	1951-01-01	2005-12-31
SMHI	CNRM-CERFACS-CNRM-CM5	RCA4 v1	1951-01-01	2005-12-31
SMHI	CSIRO-Mk3.6.0	RCA4 v1	1951-01-01	2005-12-31
SMHI	ICHEC-EC-EARTH	RCA4 v1	1951-01-01	2005-12-31
SMHI	NOAA-GFDL-GFDL-ESM2M	RCA4 v1	1951-01-01	2005-12-31
SMHI	MOHC-HadGEM2-ES	RCA4 v1	1951-01-01	2005-12-30
SMHI	IPSL-CM5A-MR	RCA4 v1	1951-01-01	2005-12-31
SMHI	MIROC-MIROC5	RCA4 v1	1951-01-01	2005-12-31
SMHI	MPI-ESM	RCA4 v1	1951-01-01	2005-12-31
SMHI	NCC-NorESM1-M	RCA4 v1	1951-01-01	2005-12-31
UQAM	CCCMA-CanESM2	CRCM5 v1	1950-01-01	2005-12-31
UQAM	MPI-ESM	CRCM5 v1	1949-01-01	2005-12-31

TABLE A.1: CORDEX-Africa RCMs that were bias-corrected for the historical period (1950-2005).

Institute	Driving Model	RCM	Beginning	End
CCCMA	CCCma-CanESM2	CanRCM4 v4	8.5: 2006-01-01	8.5: 2100-12-31
CLMcom	CNRM-CERFACS-CNRM-CM5	CCLM4-8-17 v1	4.5: 2006-01-01	4.5: 2100-12-31
			8.5: 2006-01-01	8.5: 2100-12-31
CLMcom	ICHEC-EC-EARTH	CCLM4-8-17 v1	4.5: 2006-01-01	4.5: 2100-12-31
			8.5: 2006-01-01	8.5: 2100-12-31
CLMcom	MOHC-HadGEM2-ES	CCLM4-8-17 v1	4.5: 2006-01-01	4.5: 2099-11-30
			8.5: 2006-01-01	8.5: 2099-11-30
CLMcom	MPI-M-MPI-ESM-LR	CCLM4-8-17 v1	4.5: 2006-01-01	4.5: 2100-12-31
			8.5: 2006-01-01	8.5: 2100-12-31
DMI	ICHEC-EC-EARTH	HIRHAM5 v2	4.5: 2006-01-01	4.5: 2100-12-31
			8.5: 2006-01-01	8.5: 2100-12-31
DMI	NCC-NorESM1-M	HIRHAM5 v1	4.5: 2006-01-01	4.5: 2100-12-31
			8.5: 2006-01-01	8.5: 2100-12-31
KNMI	ICHEC-EC-EARTH	RACMO22T v1	4.5: 2006-01-01	4.5: 2100-12-31
			8.5: 2006-01-01	8.5: 2100-12-31
KNMI	MOHC-HadGEM2-ES	RACMO22T v1	4.5: 2006-01-01	4.5: 2099-11-30
			8.5: 2006-01-01	8.5: 2099-12-30
MPI-CSC	ICHEC-EC-EARTH	REMO2009 v1	2.6: 2006-01-02	2.6: 2100-12-31
			4.5: 2006-01-01	4.5: 2100-12-31
			8.5: 2006-01-01	8.5: 2100-12-31
MPI-CSC	MPI-M-MPI-ESM-LR	REMO2009 v1	2.6: 2006-01-01	2.6: 2100-12-31
			4.5: 2006-01-01	4.5: 2100-12-31
			8.5: 2006-01-01	8.5: 2100-12-31
SMHI	CCCma-CanESM2	RCA4 v1	4.5: 2006-01-01	4.5: 2100-12-31
			8.5: 2006-01-01	8.5: 2100-12-31
SMHI	CNRM-CERFACS-CNRM-CM5	RCA4 v1	4.5: 2006-01-01	4.5: 2100-12-31
			8.5: 2006-01-01	8.5: 2100-12-31
SMHI	CSIRO-QCCCE-CSIRO-Mk3-6-0	RCA4 v1	4.5: 2006-01-01	4.5: 2100-12-31
			8.5: 2006-01-01	8.5: 2100-12-31
SMHI	ICHEC-EC-EARTH	RCA4 v1	4.5: 2006-01-01	4.5: 2100-12-31
			8.5: 2006-01-01	8.5: 2100-12-31
SMHI	NOAA-GFDL-GFDL-ESM2M	RCA4 v1	4.5: 2006-01-01	4.5: 2100-12-31
			8.5: 2006-01-01	8.5: 2100-12-31
SMHI	MOHC-HadGEM2-ES	RCA4 v1	2.6: 2006-01-01	2.6: 2099-12-30
			8.5: 2006-01-01	8.5: 2099-11-30
			8.5: 2006-01-01	8.5: 2099-12-30
SMHI	IPSL-IPSL-CM5A-MR	RCA4 v1	4.5: 2006-01-01	4.5: 2100-12-31
			8.5: 2006-01-01	8.5: 2100-12-31
SMHI	MIROC-MIROC5	RCA4 v1	4.5: 2006-01-01	4.5: 2100-12-31
			8.5: 2006-01-01	8.5: 2100-12-31
SMHI	MPI-M-MPI-ESM-LR	RCA4 v1	2.6: 2006-01-01	2.6: 2100-12-31
			4.5: 2006-01-01	4.5: 2100-12-31
			8.5: 2006-01-01	8.5: 2100-12-31
SMHI	NCC-NorESM1-M	RCA4 v1	4.5: 2006-01-01	4.5: 2100-12-31
			8.5: 2006-01-01	8.5: 2100-12-31
UQAM	CCCma-CanESM2	CRCM5 v1	4.5: 2006-01-01	4.5: 2100-12-30
UQAM	MPI-M-MPI-ESM-LR	CRCM5 v1	4.5: 2006-01-01	4.5: 2100-12-30

Table A.2: CORDEX-Africa RCMs that were bias-corrected for the future period (2006-2100) - the simulation period can differ for the different RCP scenarios RCP 2.6, 4.5 and 8.5.

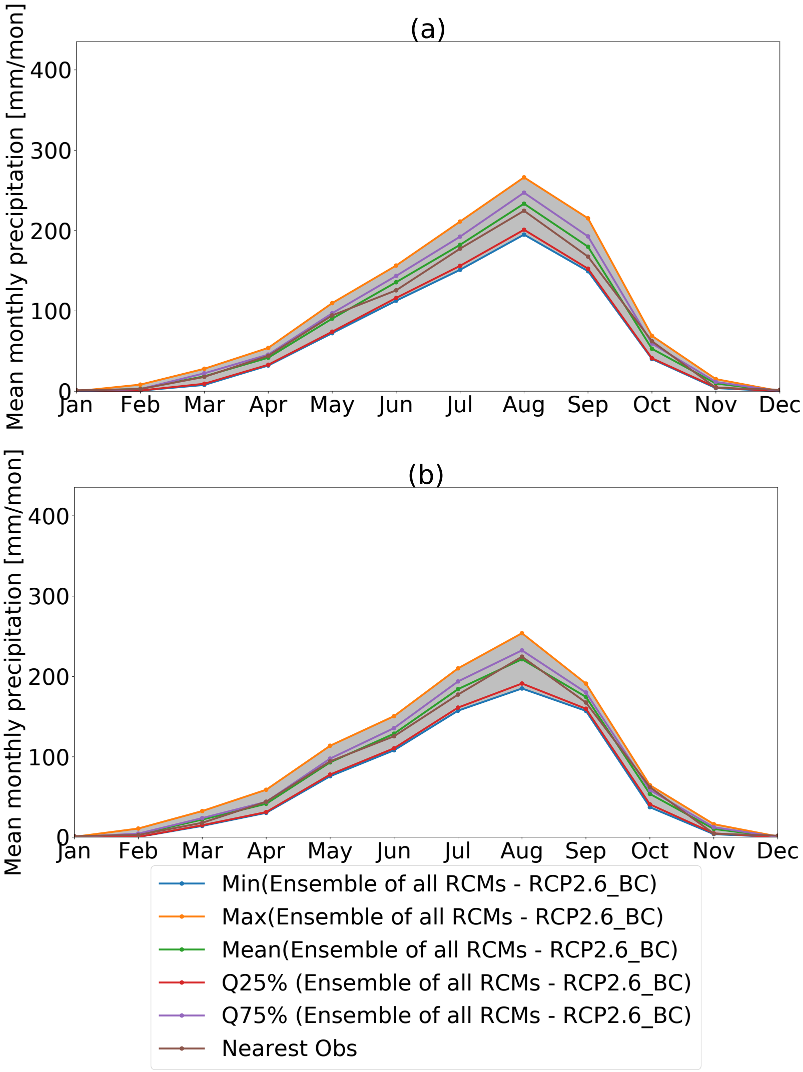

FIGURE A.1: Mean monthly sum of precipitation averaged over 173 grid cells - bias-corrected RCP 2.6 scenario - a: near future (2020-2050), b: far future (2070-2100).

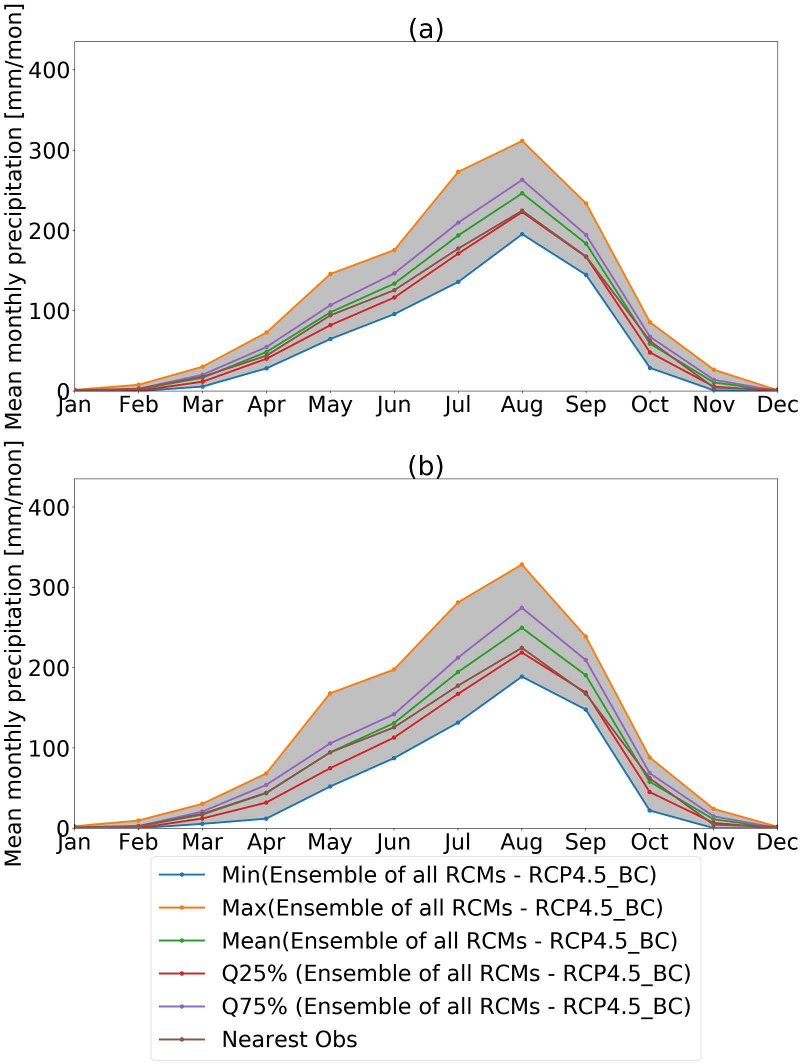

FIGURE A.2: Mean monthly sum of precipitation averaged over 173 grid cells - bias-corrected RCP 4.5 scenario - a: near future (2020-2050), b: far future (2070-2100).

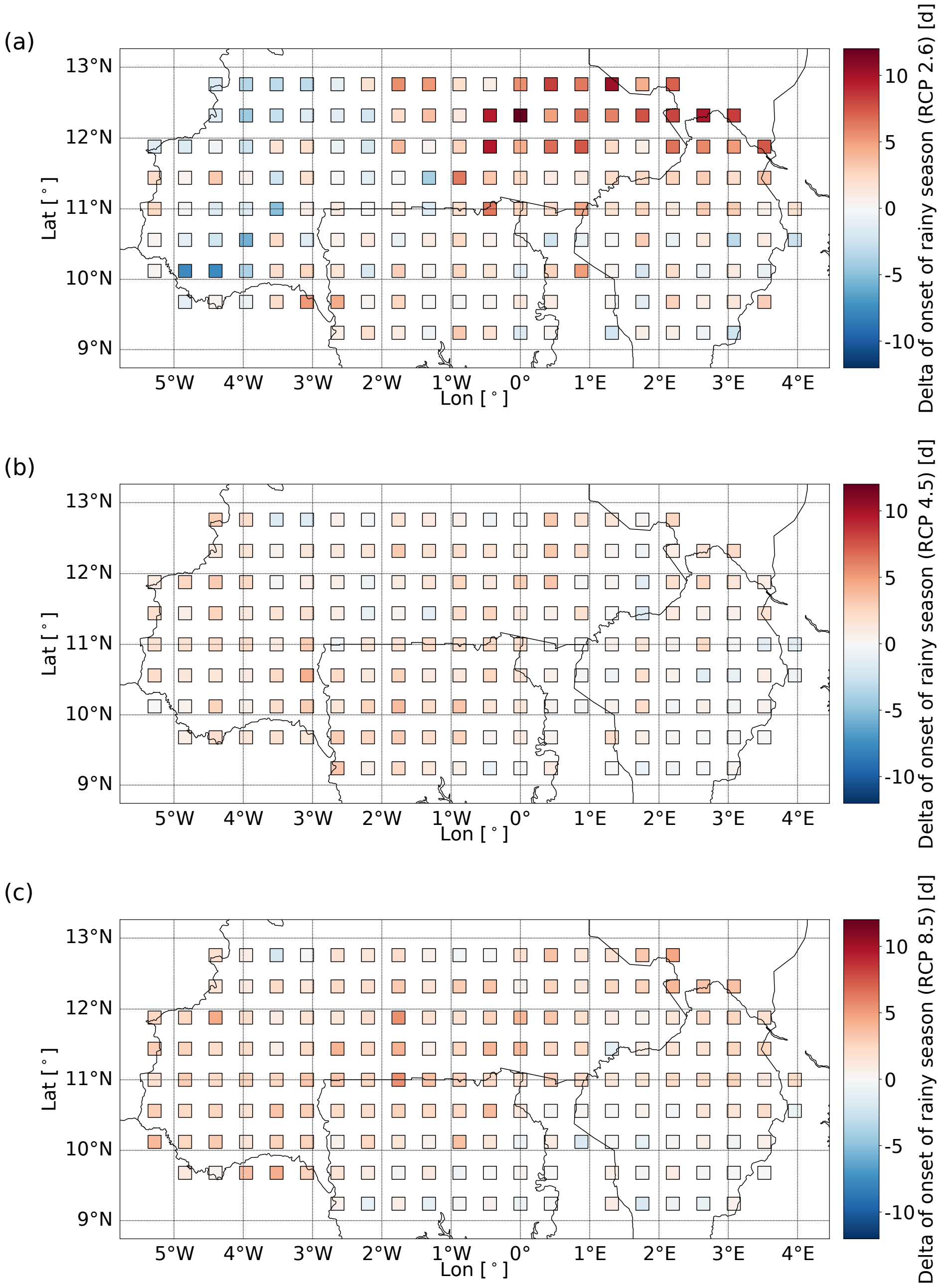
(a)
13°N
12°N
11°N
10°N
9°N
Lat [°]
5°W 4°W 3°W 2°W 1°W 0° 1°E 2°E 3°E 4°E
Lon [°]
10
5
0
-5
-10
Delta of onset of rainy season (RCP 2.6) [d]
(b)
13°N
12°N
11°N
10°N
9°N
Lat [°]
5°W 4°W 3°W 2°W 1°W 0° 1°E 2°E 3°E 4°E
Lon [°]
10
5
0
-5
-10
Delta of onset of rainy season (RCP 4.5) [d]
(c)
13°N
12°N
11°N
10°N
9°N
Lat [°]
5°W 4°W 3°W 2°W 1°W 0° 1°E 2°E 3°E 4°E
Lon [°]
10
5
0
-5
-10
Delta of onset of rainy season (RCP 8.5) [d]

FIGURE A.3: Model-averaged Δ_{DOY} in the near future (2020-2050) for RCP 2.6 (a), RCP 4.5 (b) and RCP 8.5 (c).

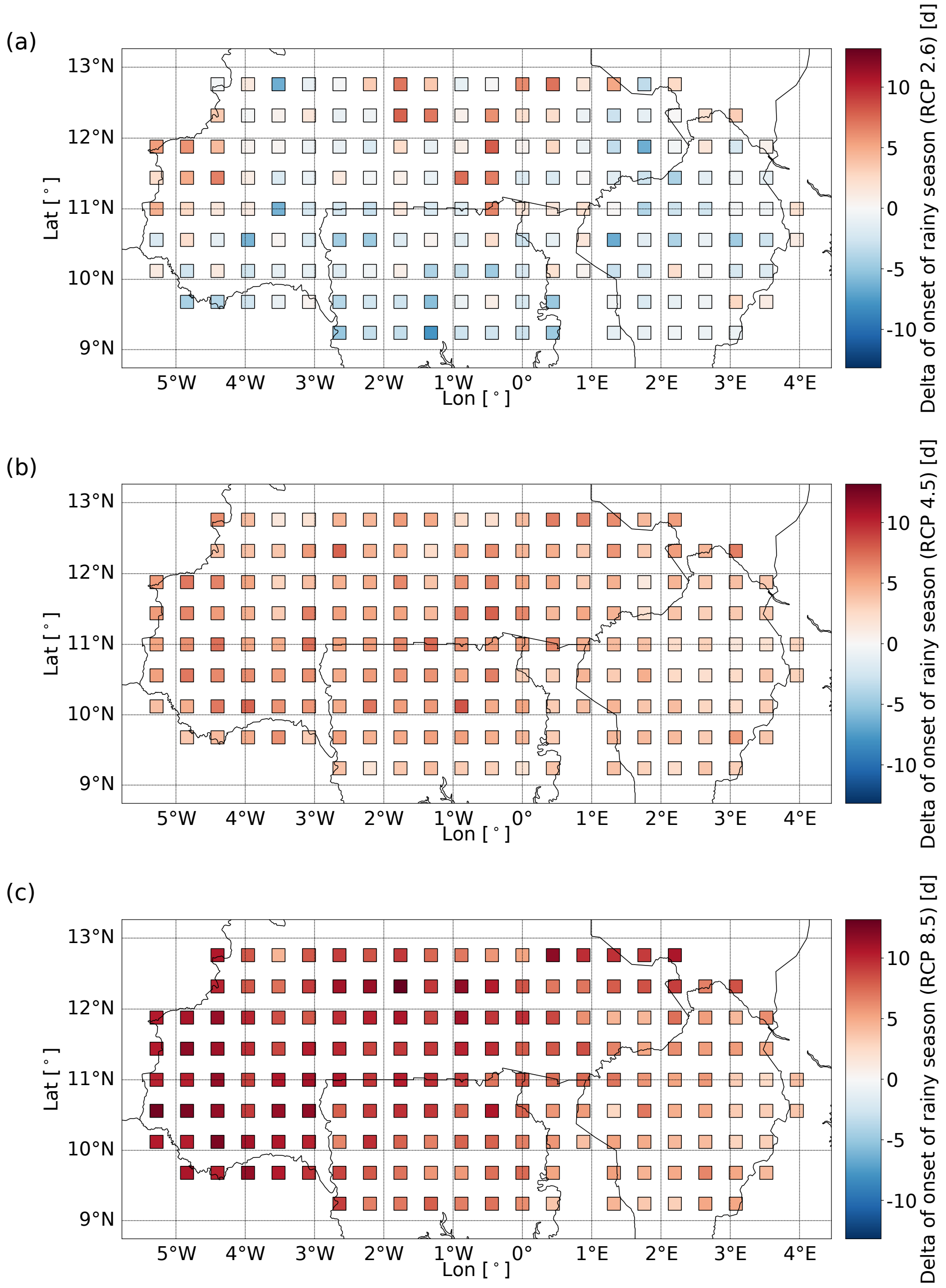
(a)
(b)
(c)
13°N
12°N
11°N
10°N
9°N
Lat [°]
5°W
4°W
3°W
2°W
1°W
0°
1°E
2°E
3°E
4°E
Lon [°]
10
5
0
-5
-10
Delta of onset of rainy season (RCP 2.6) [d]
Delta of onset of rainy season (RCP 4.5) [d]
Delta of onset of rainy season (RCP 8.5) [d]

FIGURE A.4: Model-averaged Δ_{DOY} in the far future (2070-2100) for RCP 2.6 (a), RCP 4.5 (b) and RCP 8.5 (c).

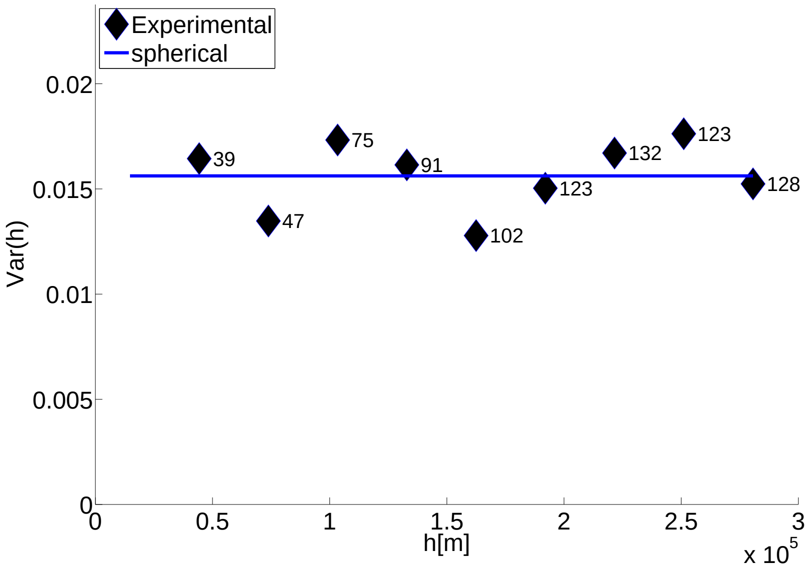

FIGURE A.5: Example of a nearly constant experimental and fitted variograms of λ_{wbl} in March (1950-2005).

Appendix B

Appendix to Chapter 5 - Copula-based temporal disaggregation of RCM precipitation

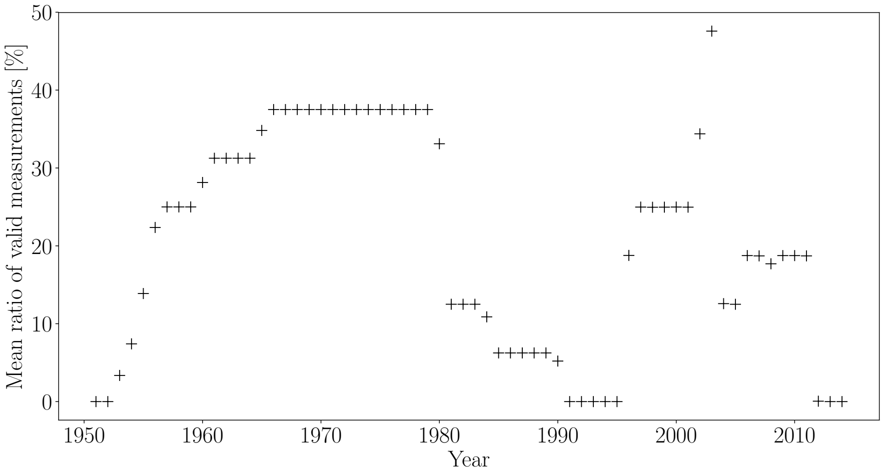

FIGURE B.1: Mean ratio of valid measurements of the 16 gauges in the proximity of Freiburg (1951-2013).

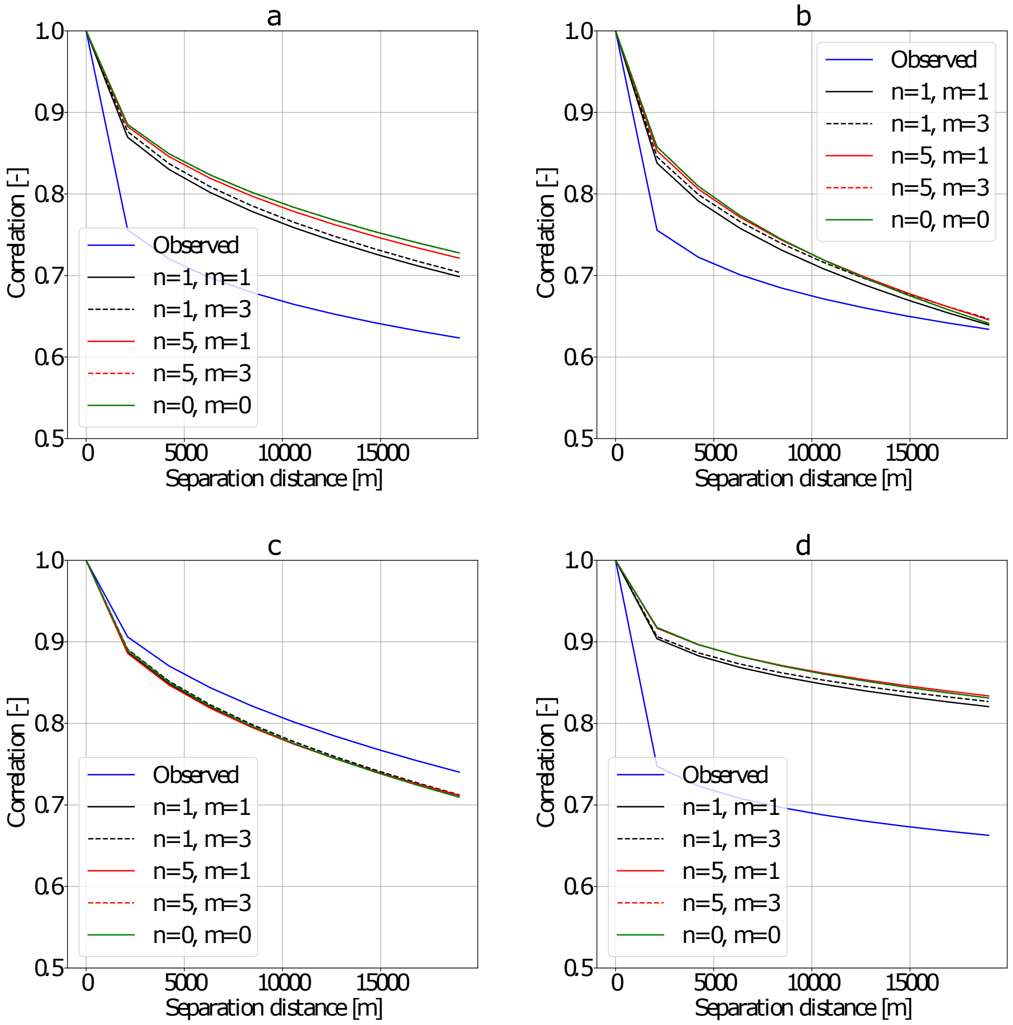

FIGURE B.2: Fitted cross correlograms of observed (1951-2013), disaggregated (1980-2009) and shuffled (1980-2009, $n = 0, m = 0$) 5 minute precipitation with a temporal lag of $\tau = 0\, min$ in the region of Freiburg in the seasons DJF (a), MAM (b), JJA (c) and SON (d).

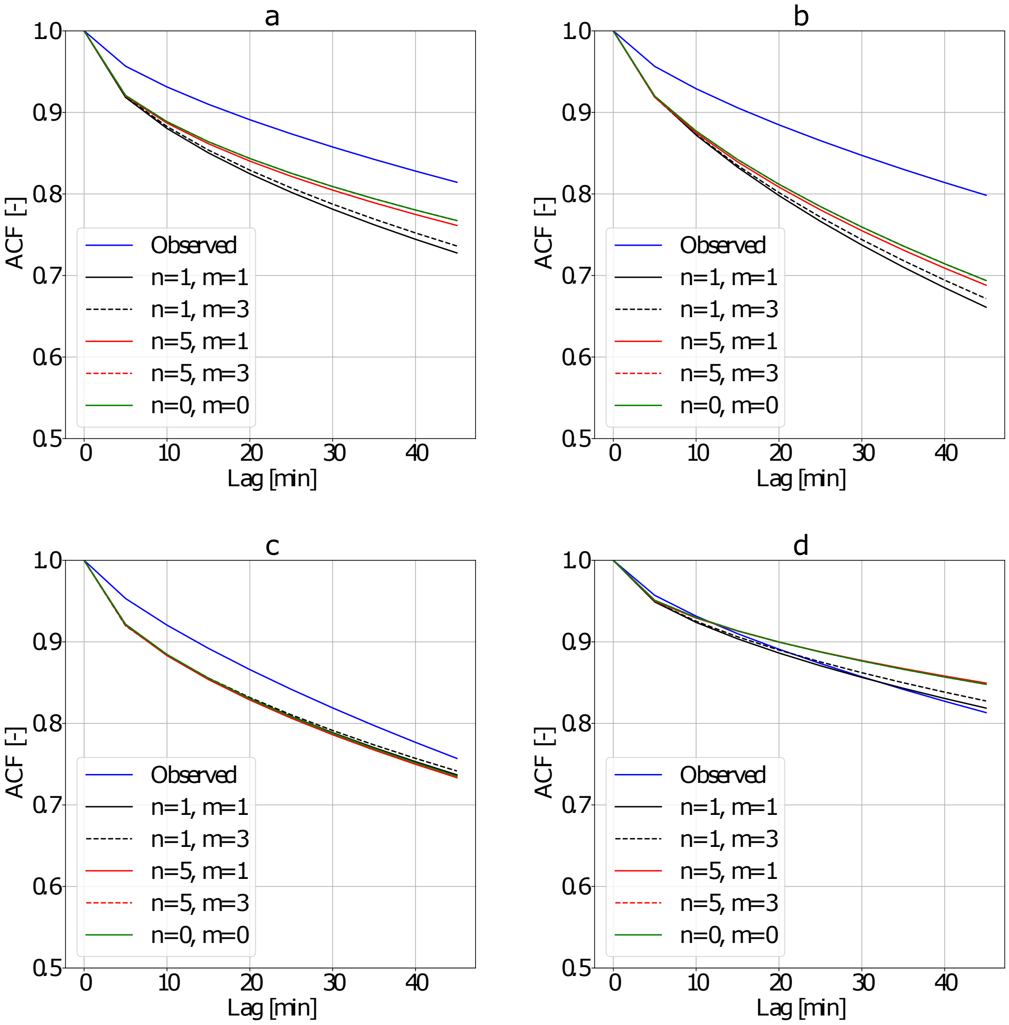

FIGURE B.3: Fitted auto correlograms of observed (1951-2013), disaggregated (1980-2009) and shuffled (1980-2009, $n = 0, m = 0$) 5 minute precipitation in the region of Freiburg in the seasons DJF (a), MAM (b), JJA (c) and SON (d).

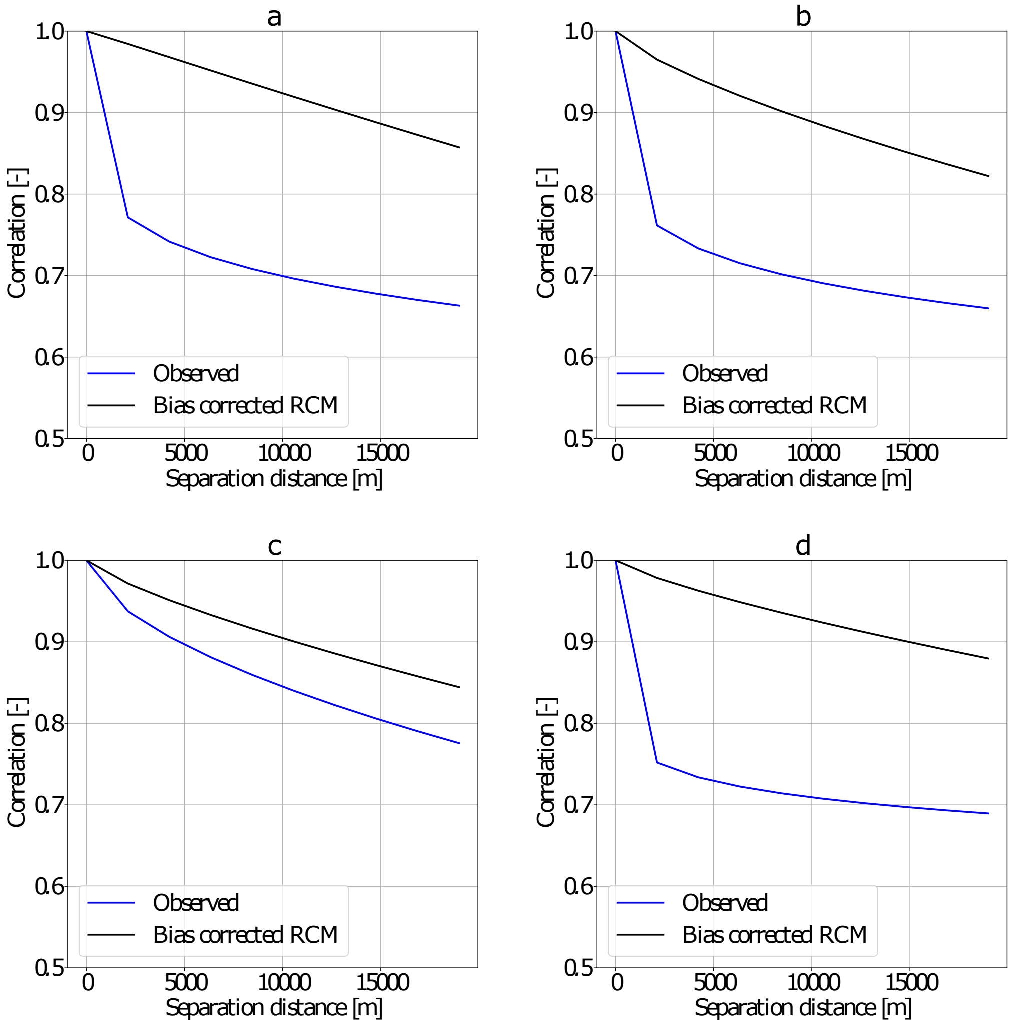

FIGURE B.4: Fitted cross correlograms of observed (1951-2013) and bias corrected (1980-2009) hourly precipitation with a temporal lag of $\tau = 0$ *min* in the region of Freiburg in the seasons DJF (a), MAM (b), JJA (c) and SON (d).

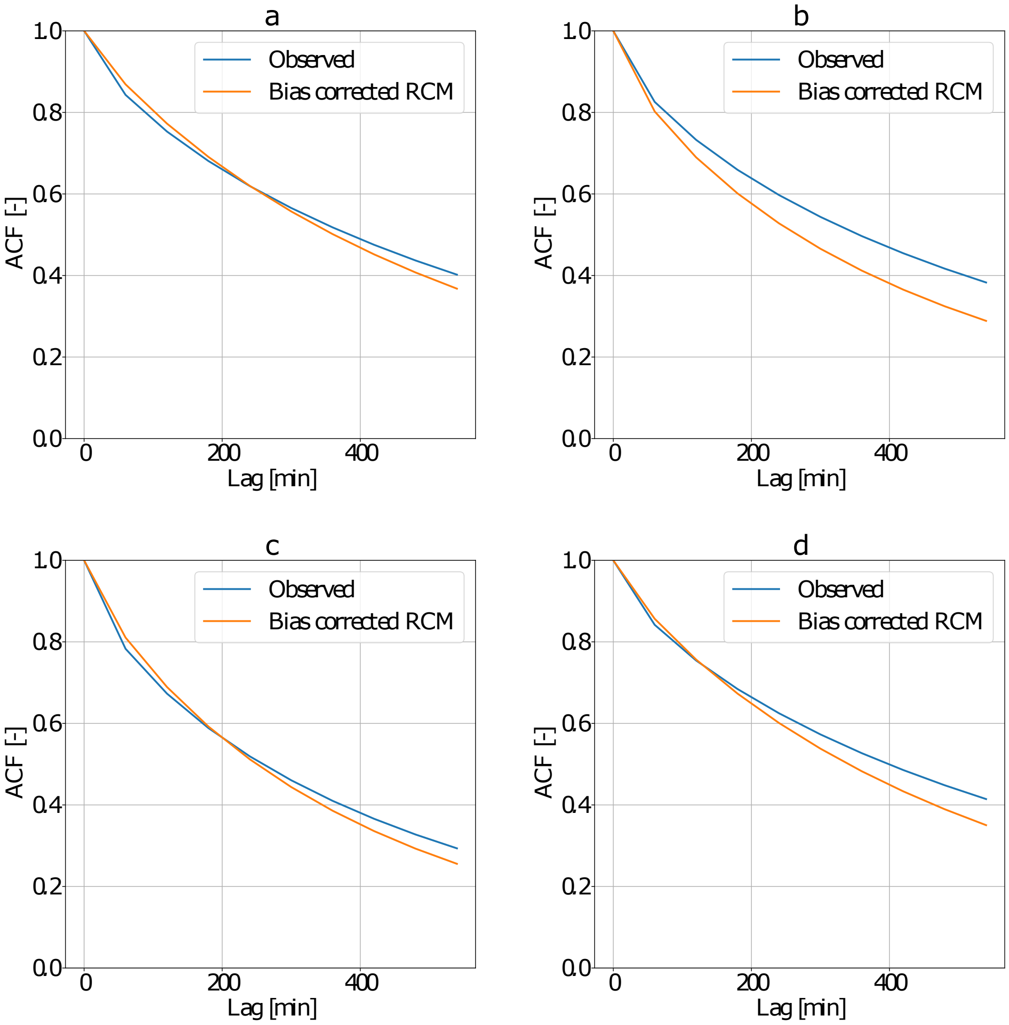

Figure B.5: Fitted auto correlograms of observed (1951-2013) and bias corrected (1980-2009) hourly precipitation in the region of Freiburg in the seasons DJF (a), MAM (b), JJA (c) and SON (d).

Appendix C

Appendix to Chapter 6 - Multivariate Vine Copula-based Bias Correction

Station Number	Name	Longitude [°]	Latitude [°]	Elevation $[m]$
1	Reiteralm 1	12.81	47.65	1755
2	Reiteralm 2	12.81	47.65	1670
3	Reiteralm 3	12.81	47.65	1615
4	Schönau	12.98	47.61	617
5	Jenner 1	13.02	47.59	1200
6	Höllgraben	13.01	47.62	653
7	Kühroint	12.96	47.57	1407
8	Funtenseetauern	12.97	47.49	2445
9	Hinterberghorn	12.92	47.55	2002
10	Trischuebel	12.91	47.53	1764
11	Schlunghorn	13.04	47.55	2155
12	Steinernes Meer	12.92	47.50	1893
13	Watzmannhaus	12.93	47.57	1919
14	Blaueis	12.87	47.59	1651
15	Hinterseeau	12.83	47.59	840
16	Brunftbergtiefe	12.88	47.55	1238
17	Lofer	12.70	47.58	625
18	Loferer Alm	12.64	47.60	1623
19	Salzburg-Flughafen	13.00	47.80	430
20	Schmittenhöhe	12.74	47.33	1973
21	Golling	13.18	47.59	491
22	Saalbach	12.65	47.39	974

TABLE C.1: Locations of the 22 meteorological measurement stations in the Berchtesgaden National Park.

Bibliography

Aas, K., Czado, C., Frigessi, A., Bakken, H. (2009). Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics*, 44, 2, 182–198, DOI: 10.1016/j.insmatheco.2007.02.001.

Allard, D., Bourotte, M. (2014). Disaggregating daily precipitations into hourly values with a transformed censored latent Gaussian process. *Stochastic Environmental Research and Risk Assessment*, 29, 2, 453–462, DOI: 10.1007/s00477-014-0913-4.

Allcroft, D. J., Glasbey, C. A. (2003). A latent Gaussian Markov random-field model for spatiotemporal rainfall disaggregation. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 52, 4, 487–498, DOI: 10.1111/1467-9876.00419.

de Amorim Borges, P., Franke, J., da Anunciação, Y. M. T., Weiss, H., Bernhofer, C. (2006). Comparison of spatial interpolation methods for the estimation of precipitation distribution in Distrito Federal, Brazil. *Theoretical and Applied Climatology*, 123, 1-2, 335–348, DOI: 10.1007/s00704-014-1359-9.

Anav, A., Friedlingstein, P., Kidston, M., Bopp, L., Ciais, P, Cox, P., Jones, C., Jung, M., Myneni, R., Zhu, Z. (2013). Evaluating the Land and Ocean Components of the Global Carbon Cycle in the CMIP5 Earth System Models. *Journal of Climate*, 26, 18, 6801–6843, DOI: 10.1175/jcli-d-12-00417.1.

Argüeso, D., Evans, J. P., Fita, L. (2013). Precipitation bias correction of very high resolution regional climate models. *Hydrology and Earth System Sciences*, 17, 11, 4379-4388, DOI: 10.5194/hess-17-4379-2013.

Barbero, R., Fowler, H. J., Lenderink, G., Blenkinsop, S. (2017). Is the intensification of precipitation extremes with global warming better detected at hourly than daily resolutions? *Geophysical Research Letters*, 44, 2, 974–983, DOI: 10.1002/2016gl071917.

Bárdossy, A. (1998). Generating precipitation time series using simulated annealing. *Water Resources Research*, 34, 7, 1737–1744, DOI: 10.1029/98wr00981.

Bárdossy, A., Hörning, S. (2015). Random Mixing: An Approach to Inverse Modeling for Groundwater Flow and Transport Problems. *Transport in Porous Media*, 114, 2, 241–259, DOI: 10.1007/s11242-015-0608-4.

Bárdossy, A., Li, J. (2008). Geostatistical interpolation using copulas. *Water Resources Research*, 44, 7, DOI: 10.1029/2007WR006115.

Bárdossy, A., Pegram, G. (2011). Downscaling precipitation using regional climate models and circulation patterns toward hydrology. *Water Resources Research*, 47, 4, DOI: 10.1029/2010wr009689.

Bárdossy, A., Pegram, G. (2012). Multiscale spatial recorrelation of RCM precipitation to produce unbiased climate change scenarios over large areas and small. *Water Resources Research*, 48, 9, DOI: 10.1029/2011wr011524.

Bárdossy, A., Pegram, G. (2016). Space-time conditional disaggregation of precipitation at high resolution via simulation. *Water Resources Research*, 52, 2, DOI: 10.1002/2015wr018037.

Beck, F. (2013). Generation of spatially correlated synthetic rainfall time series in high temporal resolution : a data driven approach. *University of Stuttgart*, DOI: 10.18419/opus-485.

Ben Alaya, M. A., Chebana, F., Ouarda, T. B. M. J. (2014). Probabilistic Gaussian Copula Regression Model for Multisite and Multivariable Downscaling. *Journal of Climate*, 27, 9, 3331–3347, DOI: 10.1175/jcli-d-13-00333.1.

van den Berg, M. J., Vandenberghe, S., De Baets, B., Verhoest, N. E. C. (2011). Copula-based downscaling of spatial rainfall: a proof of concept. *Hydrology and Earth System Sciences*, 15, 5, 1445–1457, DOI: 10.5194/hess-15-1445-2011.

Berg, P., Feldmann, H., Panitz, H.-J. (2011). Bias correction of high resolution regional climate model data. *Journal of Hydrology*, 448-449, 80–92, DOI: 10.1016/j.jhydrol.2012.04.026.

Berg, P., Wagner, S., Kunstmann, H., Schädler, G. (2012). High resolution regional climate model simulations for Germany: part I—validation, *Climate Dynamics*, 40, 1-2, 401–414, DOI: 10.1007/s00382-012-1508-8.

Betson, R., Bales, J., Pratt, H. (1980). User's guide to TVA-HYSIM. A Hydrologic Program for Quantifying Land-Use Change Effects, *U.S. Environmental Protection Agency, Washington, D.C.*

Bowler, N. E., Pierce, C. E., Seed, A. W. (2006). STEPS: A probabilistic precipitation forecasting scheme which merges an extrapolation nowcast with downscaled NWP. *Quarterly Journal of the Royal Meteorological Society*, *132*, 620, 2127–2155, DOI: 10.1256/qj.04.100.

Brissette, F. P., Khalili, M., Leconte, R. (2007). Efficient stochastic generation of multi-site synthetic precipitation data. *Journal of Hydrology*, 345, 3-4, 121–133, DOI: 10.1016/j.jhydrol.2007.06.035.

Brommundt, J. (2008). Stochastic generation of spatially related precipitation time series. *University of Stuttgart*, DOI: 10.18419/opus-278.

Bruni, G., Reinoso, R., van de Giesen, N. C., Clemens, F. H. L. R., ten Veldhuis, J. A. E. (2015). On the sensitivity of urban hydrodynamic modelling to rainfall spatial and temporal resolution. *Hydrology and Earth System Sciences*, 19, 2, 691–709, DOI: 10.5194/hess-19-691-2015.

Buizza, R., Bidlot, J.-R., Janousek, M., Keeley, S., Mogensen, K., Richardson, D. (2017). New IFS cycle brings sea-ice coupling and higher ocean resolution. *ECMWF Newsletter*, 150 - Winter 2016/17, 14–17, DOI: 10.21957/xbov3ybily.

Caesar, L., Rahmstorf, S., Robinson, A., Feulner, G., Saba, V. (2018). Observed fingerprint of a weakening Atlantic Ocean overturning circulation. *Nature*, 556, 7700, 191–196, DOI: 10.1038/s41586-018-0006-5.

Callau Poduje, A. C., Haberlandt, U. (2017). Short time step continuous rainfall modeling and simulation of extreme events. *Journal of Hydrology*, 552, 182–197, DOI: 10.1016/j.jhydrol.2017.06.036.

Cannon, A. J. (2016). Multivariate Bias Correction of Climate Model Output: Matching Marginal Distributions and Intervariable Dependence Structure. *Journal of Climate*, 29, 19, 7045–7064, DOI: 10.1175/jcli-d-15-0679.1.

Chandler, R. E., Wheater, H. S. (2002). Analysis of rainfall variability using generalized linear models: A case study from the west of Ireland. *Water Resources Research*, 38, 10, 1–11, DOI: 10.1029/2001WR000906.

Chen, H., Xu, C.-Y., Guo, S. (2012). Comparison and evaluation of multiple GCMs, statistical downscaling and hydrological models in the study of climate change impacts on runoff. *Journal of Hydrology, 434-435*, 36–45, DOI: 10.1016/j.jhydrol.2012.02.040.

Chen, J., Brissette, F. P., Leconte, R. (2013). Assessing regression-based statistical approaches for downscaling precipitation over North America. *Hydrological Processes*, 28, 9, 3482–3504, DOI: 10.1002/hyp.9889.

Chen, J., Brissette, F. P., Chaumont, D., Braun, M. (2013). Finding appropriate bias correction methods in downscaling precipitation for hydrologic impact studies over North America. *Water Resources Research*, 49, 7, 4187–4205, DOI: 10.1002/wrcr.20331.

Chiew, F. H. S., Kirono, D. G. C., Kent, D. M., Frost, A. J., Charles, S.P., Timbal, B., Nguyen, K. C., Fu, G. (2010). Comparison of runoff modelled using rainfall from different downscaling methods for historical and future climates. *Journal of Hydrology*, *387*, 1-2, 10–23, DOI: 10.1016/j.jhydrol.2010.03.025.

Clark, M., Gangopadhyay, S., Hay, L., Rajagopalan, B., Wilby, R. (2004). The Schaake Shuffle: A Method for Reconstructing Space–Time Variability in Forecasted Precipitation and Temperature Fields. *Journal of Hydrometeorology*, *5*, 1, 243–262, DOI: 10.1175/1525-7541(2004)005<0243:tssamf>2.0.co;2.

Cowpertwait, P. S. P. (1991). The stochastic generation of rainfall time series. *Newcastle University*.

Cressie, N., Hawkins, D. M. (1980). Robust estimation of the variogram: I. *Journal of the International Association for Mathematical Geology*, 12, 2, 115–125, DOI: 10.1007/BF01035243.

Deidda, R., Benzi, R., Siccardi, F. (1999). Multifractal modeling of anomalous scaling laws in rainfall. *Water Resources Research*, 35, 6, 1853–1867, DOI: 10.1029/1999wr900036.

Deutsch, C. V. (1996). Correcting for negative weights in ordinary kriging. *Computers & Geosciences*, 22, 7, 765–773, DOI: 10.1016/0098-3004(96)00005-2.

Durban, M., Glasbey, C. A. (2001). Weather modelling using a multivariate latent Gaussian model. *Agricultural and Forest Meteorology*, 109, 3, 187–201, DOI: 10.1016/s0168-1923(01)00268-4.

Ehret, U., Zehe, E., Wulfmeyer, V., Warrach-Sagi, K., Liebert, J. (2012). HESS Opinions - "'Should we apply bias correction to global and regional climate model data?"'. *Hydrology and Earth System Sciences*, 16, 9, 3391–3404, DOI: 10.5194/hess-16-3391-2012.

Erdin, R., Frei, C., Künsch (2012). Data Transformation and Uncertainty in Geostatistical Combination of Radar and Rain Gauges. *Journal of Hydrometeorology*, 13, 4, 1332–1346, DOI: 10.1175/jhm-d-11-096.1.

Evin, G., Favre, A.-C. (2008). A new rainfall model based on the Neyman-Scott process using cubic copulas. *Water Resources Research*, 44, 3, DOI: 10.1029/2007wr006054.

Fischer, M., Köck, C, Schlüter, S., Weigert, F. (2009). An empirical analysis of multivariate copula models. *Quantitative Finance*, 9, 7, 839–854, DOI: 10.1080/14697680802595650.

Frei, C., Christensen, J. H., Déqué, M., Jacob, D., Jones, R. G., Vidale, P. L.(2003). Daily precipitation statistics in regional climate models: Evaluation and intercomparison for the European Alps. *Journal of Geophysical Research: Atmospheres*, 108, D3, DOI: 10.1029/2002jd002287.

Gagnon, P., Rousseau, A. N. (2014). Stochastic spatial disaggregation of extreme precipitation to validate a regional climate model and to evaluate climate change impacts over a small watershed. *Hydrology and Earth System Sciences*, 18, 5, 1695–1704, DOI: 10.5194/hess-18-1695-2014.

Gaitan, C. F., Hsieh, W. W., Cannon, A. J. (2014). Comparison of statistically downscaled precipitation in terms of future climate indices and daily variability for southern Ontario and Quebec, Canada. *Climate Dynamics*, 43, 12, 3201–3217, DOI: 10.1007/s00382-014-2098-4.

Geman, S., Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-6*, 6, 721–741, DOI: 10.1109/tpami.1984.4767596.

Genest, C., Favre, A.-C. (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering*, 4, DOI: 10.1061/(ASCE)1084-0699(2007)12:4(347).

Germann, U., Galli, G., Boscacci, M., Bolliger, M. (2006). Radar precipitation measurement in a mountainous region. *Quarterly Journal of the Royal Meteorological Society*, 132, 618, 1669–169, DOI: 10.1256/qj.05.190.

Goly, A., Teegavarapu, R. S. V., Mondal, A. (2014). Development and Evaluation of Statistical Downscaling Models for Monthly Precipitation. *Earth Interactions*, 18, 18, DOI: 10.1175/EI-D-14-0024.1.

Gräler, B. (2014). Modelling skewed spatial random fields through the spatial vine copula. *Spatial Statistics*, 10, 87–102, DOI: 10.1016/j.spasta.2014.01.001.

Gudmundsson, L., Bremnes, J. B., Haugen, J. E., Engen-Skaugen, T. (2012). Technical Note: Downscaling RCM precipitation to the station scale using statistical transformations — a comparison of methods. *Hydrology and Earth System Sciences*, 16, 9, 3383–3390, DOI: 10.5194/hess-16-3383-2012.

Gyasi-Agyei, Y. (2012). Use of observed scaled daily storm profiles in a copula based rainfall disaggregation model. *Advances in Water Resources*, 45, 26–36, DOI: 10.1016/j.advwatres.2011.11.003.

Haberlandt, U., Ebner von Eschenbach, A.-D., Buchwald, I. (2008). A space-time hybrid hourly rainfall model for derived flood frequency analysis. *Hydrology and Earth System Sciences*, 12, 6, 1353–1367, DOI: 10.5194/hess-12-1353-2008.

Haese, B., Hörning, S., Chwala, C., Bárdossy, A., Schalge, B., Kunstmann, H. (2017). Stochastic Reconstruction and Interpolation of Precipitation Fields Using Combined Information of Commercial Microwave Links and Rain Gauges. *Water Resources Research*, 53, 12, 559–570, DOI: 10.1002/2017wr021015.

Haff, I. H., Aas, K., Frigessi, A. (2010). On the simplified pair-copula construction — Simply useful or too simplistic? *Journal of Multivariate Analysis*, 101, 5, 1296–1310, DOI: 10.1016/j.jmva.2009.12.001.

Hao, Z., Singh, V. P. (2016). Review of dependence modeling in hydrology and water resources. *Progress in Physical Geography*, 40, 4, 549–578, DOI: 10.1177/0309133316632460.

Hempel, S., Frieler, K., Warszawski, L., Schewe, J., Piontek, F. (2013). A trend-preserving bias correction — the ISI-MIP approach. *Earth System Dynamics*, 4, 2, 219–236, DOI: 10.5194/esd-4-219-2013.

Hertig, E., Merkenschlager, C., Jacobeit, J. (2016). Change points in predictors-predictand relationships within the scope of statistical downscaling. *International Journal of Climatology*, 37, 3, 1619–1633, DOI: 10.1002/joc.4801.

Hertig, E., Maraun, D., Bartholy, J., Pongracz, R., Vrac, M., Mares, I., Gutiérrez, J. M., Wibig, J., Casanueva, A., Soares, P. M. M. (2019). Comparison of statistical downscaling methods with respect to extreme events over Europe: Validation results from the perfect predictor experiment of the COST Action VALUE. *International Journal of Climatology*, 39, 9, 3846–3867, DOI: 10.1002/joc.5469.

Hershenhorn, J., Woolhiser, D. A. (1987). Disaggregation of daily rainfall. *Journal of Hydrology*, 95, 3-4, 299–322, DOI: 10.1016/0022-1694(87)90008-4.

Higham, N. J., Strabić, N., Šego, V. (2016). Restoring Definiteness via Shrinking, with an Application to Correlation Matrices with a Fixed Block. *SIAM Review*, 58, 2, 245–263, DOI: 10.1137/140996112.

Huff, F. A. (1967). Time distribution of rainfall in heavy storms. *Water Resources Research*, 3, 4, 1007–1019, DOI: 10.1029/wr003i004p01007.

IPCC (2013). Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change (Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)). *Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 1535 pp*,

Joe, H. (1996). Families of m-variate distributions with given margins and m(m-1)/2 bivariate dependence parameters. *Institute of Mathematical Statistics Lecture Notes - Monograph Series*, 120–141, DOI: 10.1214/lnms/1215452614.

Katz, R. W., Zheng, X. (1999). Mixture Model For Overdispersion of Precipitation. *Journal of Climate*, 12, 8, 2528–2537, DOI: 10.1175/1520-0442(1999)012<2528:mmfoop>2.0.co;2.

Kim, J., Lee, J., Kim, D., Kang, B. (2019). The role of rainfall spatial variability in estimating areal reduction factors. *Journal of Hydrology*, 568, 416–426, DOI: 10.1016/j.jhydrol.2018.11.014.

Klein, C., Heinzeller, D., Bliefernicht, J., Kunstmann, H. (2015). Variability of West African monsoon patterns generated by a WRF multi-physics ensemble. *Climate Dynamics*, 45, 9-10, 2733–2755, DOI: 10.1007/s00382-015-2505-5.

Knoesen, D., Smithers, J. (2009). Mixture Model For Overdispersion of The development and assessment of a daily rainfall disaggregation model for South Africa. *Hydrological Sciences Journal*, 54, 2, 217–233, DOI: 10.1623/hysj.54.2.217.

Kossieris, P., Frieler, K., Makropoulos, C., Onof, C., Koutsoyiannis, D. (2016). A rainfall disaggregation scheme for sub-hourly time scales: Coupling a Bartlett-Lewis based model with adjusting procedures. *Journal of Hydrology*, 556, 980–992, DOI: 10.1016/j.jhydrol.2016.07.015.

Koutsoyiannis, D., Onof, C., Wheater, H. S.(2003). Multivariate rainfall disaggregation at a fine timescale. *Water Resources Research*, 39, 7, DOI: 10.1029/2002wr001600.

Krūminiene, I. (2006). Analysis of anisotropic variogram models for prediction of the Curonian lagoon data. *Mathematical Modelling and Analysis*, 11, 1, 73–86, URL: http://www.tandfonline.com/doi/abs/10.1080/13926292.2006.9637303.

Kwakye, S. O. (2016). Study on the effects of climate change on the hydrology of the West African sub-region. *University of Stuttgart*, DOI: 10.18419/opus-9034.

Lafon, T., Dadson, S., Buys, G., Prudhomme, C.(2012). Bias correction of daily precipitation simulated by a regional climate model: a comparison of methods. *International Journal of Climatology*, 33, 6, 1367–1381, DOI: 10.1002/joc.3518.

Laux, P. (2009). Statistical Modeling of Precipitation for Agricultural Planning in the Volta Basin of West Africa. *University of Stuttgart.*, DOI: 10.18419/opus-303.

Laux, P., Vogl, S., Qiu, W., Knoche, H. R., Kunstmann, H. (2011). Copula-based statistical refinement of precipitation in RCM simulations over complex terrain. *Hydrology and Earth System Science*, 15, 7, 2401–2419, DOI: 10.5194/hess-15-2401-2011.

Liechti, K., Panziera, L., Germann, U., Zappa, M. (2013). The potential of radar-based ensemble forecasts for flash-flood early warning in the southern Swiss Alps. *Hydrology and Earth System Sciences*, *17*, 10, 3853–3869, DOI: 10.5194/hess-17-3853-2013.

Lenton, T. M., Held, H., Kriegler, E., Hall, J. W., Lucht, W., Rahmstorf, S., Schellnhuber, H. J. (2008). Tipping elements in the Earth's climate system. *Proceedings of the National Academy of Sciences*, *105*, 6, 1786–1793, DOI: 10.1073/pnas.0705414105.

Lisniak, D., Franke, J., Bernhofer, C. (2013). Circulation pattern based parameterization of a multiplicative random cascade for disaggregation of observed and projected daily rainfall time series. *Hydrology and Earth System Science*, 17, 7, 2487–2500, DOI: 10.5194/hess-17-2487-2013.

Lorenz, M., Bliefernicht, J., Haese, B., Kunstmann, H. (2018). Copula-based downscaling of daily precipitation fields. *Hydrological Processes*, 32, 23, 3479–3494, DOI: 10.1002/hyp.13271.

Lorenz, C., Kunstmann, H. (2012). The Hydrological Cycle in Three State-of-the-Art Reanalyses: Intercomparison and Performance Analysis. *Journal of Hydrometeorology*, 13, 5, 1397–1420, DOI: 10.1175/jhm-d-11-088.1.

Mamalakis, A., Langousis, A., Deidda, R., Marrocu, M. (2017). A parametric approach for simultaneous bias correction and high-resolution downscaling of climate model rainfall. *Water Resources Research*, 53, 3, 2149–2170, DOI: 10.1002/2016wr019578.

Mao, G., Vogl, S., Laux, P., Wagner, S., Kunstmann, H. (2015). Stochastic bias correction of dynamically downscaled precipitation fields for Germany through Copula-based integration of gridded observation data. *Hydrology and Earth System Sciences*, 19, 4, 1787–1806, DOI: 10.5194/hess-19-1787-2015.

Maraun, D., Wetterhall, F., Ireson, A. M., Chandler, R.E., Kendon, E.J., Widmann, M., Brienen, S., Rust, H. W., Sauter, T., Themeßl, M., Venema, V. K. C., Chun, K. P., Goodess, C. M., Jones, R. G., Onof, C., Vrac, M., Thiele-Eich, I. (2010). Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. *Reviews of Geophysics*, 48, 3, DOI: 10.1029/2009rg000314.

Maraun, D.(2016). Bias Correcting Climate Change Simulations - a Critical Review. *Current Climate Change Reports*, 2, 4, 211–220, DOI: 10.1007/s40641-016-0050-x.

Marra, F., Zoccatelli, D., Armon, M., Morin, E. (2019). A simplified MEV formulation to model extremes emerging from multiple nonstationary underlying processes. *Advances in Water Resources*, 127, 280–290, DOI: 10.1016/j.advwatres.2019.04.002.

Mascaro, G., Piras, M., Deidda, R., Vivoni, E. R. (2013). Distributed hydrologic modeling of a sparsely monitored basin in Sardinia, Italy, through hydrometeorological downscaling. *Hydrology and Earth System Sciences*, 17, 10, 4143–4158, DOI: 10.5194/hess-17-4143-2013.

Mascaro, G., White, D. D., Westerhoff, P., Bliss, N. (2015). Performance of the CORDEX-Africa regional climate simulations in representing the hydrological cycle of the Niger River basin. *Journal of Geophysical Research: Atmospheres*, 120, 24, 12425–12444, DOI: 10.1002/2015jd023905.

Mehrotra, R., Sharma, A. (2010). Development and Application of a Multisite Rainfall Stochastic Downscaling Framework for Climate Change Impact Assessment. *Water Resources Research*, 46, 7, DOI: 10.1029/2009wr008423.

Minasny, B., McBratney, A. B. (2005). The Matérn function as a general model for soil variograms. *Geoderma*, 128, 3-4, 192–207, DOI: 10.1016/j.geoderma.2005.04.003.

Mosthaf, T., Bárdossy, A. (2017). Regionalizing nonparametric models of precipitation amounts on different temporal scales. *Hydrology and Earth System Sciences*, 21, 5, 2463–2481, DOI: 10.5194/hess-21-2463-2017.

Müller, H., Haberlandt, U. (2015). Temporal Rainfall Disaggregation with a Cascade Model: From Single-Station Disaggregation to Spatial Rainfall. *Journal of Hydrologic Engineering*, 20, 11, DOI: 10.1061/(asce)he.1943-5584.0001195.

Nagler, T., Schellhase, C., Czado, C. (2016). Nonparametric estimation of simplified vine copula models: comparison of methods. *Dependence Modeling*, 5, 1, 99–120, DOI: 10.1515/demo-2017-0007.

Nelsen, R. B. (2006). An Introduction to Copulas. *Springer New York*, DOI: 10.1007/0-387-28678-0.

Niemi, T. J., Kokkonen, T., Seed, A. W. (2014). A simple and effective method for quantifying spatial anisotropy of time series of precipitation fields. *Water Resources Research*, 50, 7, 5906–5925, DOI: 10.1002/2013wr015190.

Nikulin, G., Jones, C., Giorgi F., Asrar, G., Büchner, M., Cerezo-Mota, R., Christensen, O. B., Déqué, M., Fernandez, J., Hänsler, A., van Meijgaard, E., Samuelsson, P., Sylla, M. B., Sushama, L. (2012). Precipitation Climatology in an Ensemble of CORDEX-Africa Regional Climate Simulations. *Journal of Climate*, 25, 18, 6057–6078, DOI: 10.1175/jcli-d-11-00375.1.

Ochoa-Rodriguez, S., Wang, L., Gires, A., Pina, R. D., Reinoso-Rondinel, R., Bruni, G., Ichiba, A., Gaitan, S., Christiano, E., van Assel, J., Kroll, S., Murlà-Tuyls, D., Tisserand, B., Schertzer, D., Tchiguirinskaia, I., Onof, C., Willems, P., ten Veldhuism, M.-C. (2015). Impact of spatial and temporal resolution of rainfall inputs on urban hydrodynamic modelling outputs: A multi-catchment investigation. *Journal of Hydrology*, 531, 389–407, DOI: 10.1016/j.jhydrol.2015.05.035.

Olsson, J.(1998). Evaluation of a scaling cascade model for temporal rainfall disaggregation. *Hydrology and Earth System Sciences*, 2, 1, 19–30, DOI: 10.5194/hess-2-19-1998.

Olsson, J., Berg, P., Kawamura, A. (2015). Impact of RCM Spatial Resolution on the Reproduction of Local, Subdaily Precipitation. *Journal of Hydrometeorology*, *16*, 2, 534–547, DOI: 10.1175/jhm-d-14-0007.1.

Onof, C., Townend, J., Kee, R. (2005). Comparison of two hourly to 5-min rainfall disaggregators. *Atmospheric Research*, 77, 1-4, 176–187, DOI: 10.1016/j.atmosres.2004.10.022.

Oriani, F., Straubhaar, J., Renard, P., Mariethoz, G. (2014). Simulation of rainfall time series from different climatic regions using the direct sampling technique. *Hydrology and Earth System Sciences*, 18, 8, 3015–3031, DOI: 10.5194/hess-18-3015-2014.

Paltan, H., Allen, M., Haustein, K., Fuldauer, L., Dadson, S. (2018). Global implications of 1.5°*C* and 2°*C* warmer worlds on extreme river flows. *Environmental Research Letters*, 13, 9, 094003, DOI: 10.1088/1748-9326/aad985.

Pfaff, T. (2013). Processing and Analysis of Weather Radar Data for Use in Hydrology. *University of Stuttgart*, DOI: 10.18419/opus-487.

Pham, M. T., Vernieuwe, H., De Baets, B., Willems, P., Verhoest, N. E. C. (2015). Stochastic simulation of precipitation-consistent daily reference evapotranspiration using vine copulas. *Stochastic Environmental Research and Risk Assessment*, 30, 8, 2197–2214, DOI: 10.1007/s00477-015-1181-7.

Piani, C., Haerter, J. O., Coppola, E. (2010). Statistical bias correction for daily precipitation in regional climate models over Europe. *Theoretical and Applied Climatology*, 99, 1, 187–192, DOI: 10.1007/s00704-009-0134-9.

Piani, C., Haerter, J. O., (2012). Two dimensional bias correction of temperature and precipitation copulas in climate models. *Geophysical Research Letters*, 39, 20, DOI: 10.1029/2012gl053839.

Pierce, D. W., Cayan, D. R., Maurer, E. P., Abatzoglou, J. T., Hegewisch, K. C. (2015). Improved Bias Correction Techniques for Hydrological Simulations of Climate Change. *Journal of Hydrometeorology*, 16, 6, 2421–2442, DOI: 10.1175/jhm-d-14-0236.1.

Polade, S. D., Pierce, D. W., Cayan, D. R., Gershunov, A., Dettinger, M. D. (2014). The key role of dry days in changing regional climate and precipitation regimes. *Scientific Reports*, 4, 1, DOI: 10.1038/srep04364.

Prein, A. F., Holland, G. J., Rasmussen, R. M., Done, J., Ikeda, K., Clark, M. P., Liu, C. H. (2013). Importance of Regional Climate Model Grid Spacing for the Simulation of Heavy Precipitation in the Colorado Headwaters. *Journal of Climate*, 26, 13, 4848–4857, DOI: 10.1175/jcli-d-12-00727.1.

Prein, A. F., Langhans, W., Fosser, G., Ferrone, A., Ban, N., Goergen, K., Keller, M., Tölle, M., Gutjahr, O., Feser, F., Brisson, E., Kollet, S., Schmidli, J., van Lipzig, N. P. M., Leung, R. (2015). A review on regional convection-permitting climate modeling: Demonstrations, prospects, and challenges. *Reviews of Geophysics*, 53, 2, 323–361, DOI: 10.1002/2014rg000475.

Pui, A., Sharma, A., Mehrotra, R., Sivakumar, B., Jeremiah, E. (2012). A comparison of alternatives for daily to sub-daily rainfall disaggregation. *Journal of Hydrology*, 470-471, 138–157, DOI: 10.1016/j.jhydrol.2012.08.041.

Renard, B., Kavetski, D., Kuczera, G., Thyer, M., Franks, S. W. (2010). Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. *Water Resources Research*, 46, 5, DOI: 10.1029/2009wr008328.

Rodriguez-Iturbe, I., Cox, D. R., Isham, V. (1987). Some Models for Rainfall Based on Stochastic Point Processes. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 410, 1839, 269–288, DOI: 10.1098/rspa.1987.0039.

Rosenblatt, M. (1956). Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics*, 27, 3, 832–837, DOI: 10.1214/aoms/1177728190.

Rummukainen, M. (2009). State-of-the-art with regional climate models. *Wiley Interdisciplinary Reviews: Climate Change*, 1, 1, 82–96, DOI: 10.1002/wcc.8.

Rupp, D. E., Licznar, P., Adamowski, W., Leśniewski, M. (2012). Multiplicative cascade models for fine spatial downscaling of rainfall: parameterization with rain gauge data. *Hydrology and Earth System Sciences*, *16*, 3, 671–684, DOI: 10.5194/hess-16-671-2012.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6, 2, 461–464, DOI: 10.1214/aos/1176344136.

Segond, M. L. (2010). Stochastic Modelling of Space-Time Rainfall and the Significance of Spatial Data for Flood Runoff Generation. *Imperial College London*, ISBN: 3843355460

Serinaldi, F. (2009). A multisite daily rainfall generator driven by bivariate copula-based mixed distributions. *Journal of Geophysical Research*, 114, D10, DOI: 10.1029/2008jd011258.

Shrestha, R., Tachikawa, Y., Takara, K. (2006). Input data resolution analysis for distributed hydrological modeling. *Journal of Hydrology*, 319, 1-4, 36–50, DOI: 10.1016/j.jhydrol.2005.04.025.

Skamarock, W. C., Klemp, J. B. (2008). A time-split nonhydrostatic atmospheric model for weather research and forecasting applications. *Journal of Computational Physics*, 227, 7, 3465–3485, DOI: 10.1016/j.jcp.2007.01.037.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8, 229–231.

Sørup, H. J. D., Madsen, H., Arnbjerg-Nielsen, K. (2012). Descriptive and predictive evaluation of high resolution Markov chain precipitation models. *Environmetrics*, 23, 7, 623–635, DOI: 10.1002/env.2173.

von Storch, H., Zwiers, F. W. (1999). Statistical Analysis in Climate Research. *Cambridge University Press*, DOI: 10.1017/cbo9780511612336.

Sun, Y., Solomon, S., Dai, A., Portmann, R. W. (2006). How Often Does It Rain? *Journal of Climate*, 19, 6, 916–934, DOI: 10.1175/jcli3672.1.

Sunyer, M. A., Luchner, J., Onof, C., Madsen, H., Arnbjerg-Nielsen, K. (2016). Assessing the importance of spatio-temporal RCM resolution when estimating sub-daily extreme precipitation under current and future climate conditions. *International Journal of Climatology*, 37, 2, 688–705, DOI: 10.1002/joc.4733.

Tarpanelli, A., Franchini, M., Brocca, L., Camici, S., Melone, F., Moramarco, T. (2012). Assessing the importance of spatio-temporal RCM resolution when estimating sub-daily extreme precipitation under current and future climate conditions. *Journal of Hydrology*, 472-473, 63–76, DOI: 10.1016/j.jhydrol.2012.09.010.

Teimouri, M., Hoseini, S. M., Nadarajah, S. (2013). Comparison of estimation methods for the Weibull distribution. *Statistics*, 47, 1, 93–109, DOI: 10.1080/02331888.2011.559657.

Teutschbein, C., Seibert, J. (2013). Is bias correction of regional climate model (RCM) simulations possible for non-stationary conditions? *Hydrology and Earth System Sciences*, 17, 12, 5061–5077, DOI: 10.5194/hess-17-5061-2013.

Themeßl, M. J., Gobiet, A., Leuprecht, A. (2010). Empirical-statistical downscaling and error correction of daily precipitation from regional climate models. *International Journal of Climatology*, 31, 10, 1530–1544, DOI: 10.1002/joc.2168.

Thober, S., Mai, J., Zink, M., Samaniego, L. (2014). Stochastic temporal disaggregation of monthly precipitation for regional gridded data sets. *Water Resources Research*, 50, 11, 8714–8735, DOI: 10.1002/2014wr015930.

Thober, S. (2016). Evaluation and disaggregation of climate model outputs for european drought prediction. *University of Jena*, URL: `https://www.db-thueringen.de/receive/dbt_mods_00029187`.

Trivedi, P. K., Zimmer, D. M. (2006). Copula Modeling: An Introduction for Practitioners *Foundations and Trends® in Econometrics*, 1, 1, 1–111, DOI: 10.1561/0800000005.

Tschöke, G. V., Kruk, N. S., de Queiroz, P. I. B., Chou, S. C., de Sousa Junior, W. C.(2017). Comparison of two bias correction methods for precipitation simulated with a regional climate model. *Theoretical and Applied Climatology*, 127, 3, 841–852, DOI: 10.1007/s00704-015-1671-z.

Uppala, S. M., KÅllberg, P. W., Simmons, A. J., Andrae, U., Da Costa Bechtold, V., Fiorino, M., Gibson, J. K., Haseler, J., Hernandez, A., Kelly, G. A., Li, X., Onogi, K., Saarinen, S., Sokka, N., Allan, R. P., Andersson, E., Arpe, K., Balmaseda, M. A., Beljaars, A. C. M., Van De Berg, L., Bidlot, J., Bormann, N., Caires, S., Chevallier, F., Dethof, A., Dragosavac, M., Fisher, M., Fuentes, M., Hagemann, S., Hólm, E., Hoskins, B. J., Isaksen, L., Janssen, P. A. E. M., Jenne, R., Mcnally, A. P., Mahfouf, J.-F., Morcrette, J.-J., Rayner, N. A., Saunders, R. W., Simon, P., Sterl, A., Trenberth, K. E., Untch, A., Vasiljevic, D., Viterbo, P., Woollen, J. (2005). The ERA-40 reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 131, 612, 2961–3012, DOI: 10.1256/qj.04.176.

Valencia, D. R., Schaake, J. C. (1973). Disaggregation processes in stochastic hydrology. *Water Resources Research*, 9, 3, 580–585, DOI: 10.1029/wr009i003p00580.

Verhoest, N. E. C., Vandenberghe, S., Cabus, P., Onof, C., Meca-Figueras, T., Jameleddine, S. (2010). Are stochastic point rainfall models able to preserve extreme flood statistics? *Hydrological Processes*, 24, 23, 3439-3445, DOI: 10.1002/hyp.7867.

Verhoest, N. E. C., van den Berg, M. J., Martens, B., Lievens, H., Wood, E. F., Pan, M., Kerr, Y. H., Al Bitar, A., Tomer, S. K., Drusch, M., Vernieuwe, H., De Baets, B., Walker, J. P., Dumeda, G., Pauwels, V. R. N. (2015). Copula-Based Downscaling of Coarse-Scale Soil Moisture Observations With Implicit Bias Correction. *IEEE Transactions on Geoscience and Remote Sensing*, 53, 3507-3521, DOI: 10.1109/TGRS.2014.2378913.

Vernieuwe, H., Vandenberghe, S., De Baets, B., Verhoest., N. E. C. (2015). A continuous rainfall model based on vine copulas. *Hydrology and Earth System Sciences*, 19, 6, 2685–2699, DOI: 10.5194/hess-19-2685-2015.

Vogl, S., Laux, P., Qiu, W., Mao., G., Kunstmann, H. (2012). Copula-based assimilation of radar and gauge information to derive bias-corrected precipitation fields. *Hydrology and Earth System Sciences*, 16, 7, 2311–2328, DOI: 10.5194/hess-16-2311-2012.

Volosciuk, C., Maraun, D., Vrac, M., Widmann, M. (2017). A combined statistical bias correction and stochastic downscaling method for precipitation. *Hydrology and Earth System Sciences*, 21, 3, 1693–1719, DOI: 10.5194/hess-21-1693-2017.

Vrac, M. (2018). Multivariate bias adjustment of high-dimensional climate simulations: the Rank Resampling for Distributions and Dependences (R^2D^2) bias correction. *Hydrology and Earth System Sciences*, 22, 6, 3175–3196, DOI: 10.5194/hess-22-3175-2018.

van Vuuren, D. P., Edmonds, J. A., Kainuma, M., Riahi, K., Weyant, J. (2011). A special issue on the RCPs. *Climatic Change*, 109, 1-2, 1–4, DOI: 10.1007/s10584-011-0157-y.

Wagner, S., Kunstmann, H., Bárdossy, A., Conrad, C., Colditz, R. R. (2009). Water balance estimation of a poorly gauged catchment in West Africa using dynamically downscaled meteorological fields and remote sensing information. *Physics and Chemistry of the Earth, Parts A/B/C, 34*, 4-5, 225–235, DOI: 10.1016/j.pce.2008.04.002.

Wagner, S., Berg, P., Schädler, G., Kunstmann, H. (2012). High resolution regional climate model simulations for Germany: Part II—projected climate changes. *Climate Dynamics*, 40, 1-2, 415–427, DOI: 10.1007/s00382-012-1510-1.

Wagner, S., Kunstmann, H. (2016). High resolution regional climate model simulations for Germany: Part II—projected climate changes. *Internal annual report. Steinbuch Centre for Computing (SCC). Karlsruhe Institute of Technology (KIT)*, 11.

Waichler, S. R., Wigmosta, M. S. (2003). Development of Hourly Meteorological Values From Daily Data and Significance to Hydrological Modeling at H. J. Andrews Experimental Forest. *Journal of Hydrometeorology*, 4, 2, 251–263, DOI: 10.1175/1525-7541(2003)4<251:dohmvf>2.0.co;2.

Warscher, M., Strasser, U., Kraller, G., Marke, T., Franz, H., Kunstmann, H. (2013). Performance of complex snow cover descriptions in a distributed hydrological model system: A case study for the high Alpine terrain of the Berchtesgaden Alps. *Water Resources Research*, 49, 5, 2619–2637, DOI: 10.1002/wrcr.20219.

Warscher, M., Wagner, S., Marke, T., Laux, P., Strasser, U., Kunstmann, H. (2019). A Very High-Resolution Regional Climate Simulation for Central Europe: Performance in High Mountain Areas and Projected Near Future Climate. *Journal of Geophysical Research: Atmospheres*, under review.

Westra, S., Evans, J. P., Mehrotra, R., Sharma, A. (2013). A conditional disaggregation algorithm for generating fine time-scale rainfall data in a warmer climate. *Journal of Hydrology*, 479, 86–99, DOI: 10.1016/j.jhydrol.2012.11.033.

Wetterhall, F., Pappenberger, F., He, Y., Freer, J., Cloke, H. L. (2012). Conditioning model output statistics of regional climate model precipitation on circulation patterns. *Nonlinear Processes in Geophysics*, 19, 6, 623–633, DOI: 10.5194/npg-19-623-2012.

Wilby, R. L., Wigley, T. M. L. (1997). Downscaling general circulation model output: a review of methods and limitations. *Agricultural and Forest Meteorology*, 96, 1-3, 85–101, DOI: 10.1177/030913339702100403.

Wilks, D. S. (1999). Simultaneous stochastic simulation of daily precipitation, temperature and solar radiation at multiple sites in complex terrain. *Progress in Physical Geography*, 21, 4, 530–548, DOI: 10.1016/s0168-1923(99)00037-4.

Wilks, D. S. (2011). Statistical Methods in the Atmospheric Sciences (3rd ed.). *Oxford; Waltham, MA: Academic Press.* , ISBN: 9780123850225.

Yang, C., Wang, N., Wang, S. (2012). A comparison of three predictor selection methods for statistical downscaling. *International Journal of Climatology*, 37, 3, 1238–1249, DOI: 10.1002/joc.4772.

Yevjevich, Y., Lane, W. L. (1997). Applied Modeling of Hydrologic Time Series. *Water Resources Pubns*, 37, 3, 1238–1249, ISBN: 978-0918334374.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, 8, 3, 338–353 DOI: 10.1016/s0019-9958(65)90241-x.

Acknowledgements

I would like to express my gratitude towards the following persons and institutions for supporting me during the writing of this thesis:

- *Prof. Dr. Harald Kunstmann* for your supervision, patience and continuous support during the last years.

- *Prof. Dr. Elke Hertig* for agreeing to serve as the second assessor of this thesis.

- *Dr. Jan Bliefernicht and Dr. Barbara Haese* who have spent many hours on proofreading and who provided lots of helpful advice regarding the structure and analysis.

- *Dr. Andreas Wagner* for the constructive support during the first few years and *Dr. Thomas Rummler* for always helping with administrative or technical issues.

- all the other colleagues in Augsburg and Garmisch for the nice atmosphere and discussions: *Christof, Cornelius, Diarra, Max, Patrick, Sina, Sven,...*

- *BMBF and StMUV* for the funding of the different projects.

- *IMK-IFU/KIT* for providing computation time.

- *my friends* inside and outside of the institute.

- *my parents* for their understanding and support.