COPULAS for INTEGRATING WEATHER and LAND INFORMATION in SPACE and TIME

Fakhereh Alidoost

COPULAS for INTEGRATING WEATHER and LAND INFORMATION in SPACE and TIME

DISSERTATION

to obtain
the degree of doctor at the University of Twente,
on the authority of the rector magnificus,
prof.dr. T.T.M. Palstra,
on account of the decision of the Doctorate Board,
to be publicly defended
on Wednesday, 24 April 2019 at 14:45 hrs

by

Fakhereh Alidoost

born on 8 June 1986

in Esfahan, Iran

This thesis has been approved by
Prof.dr.ir. A. Stein, supervisor
Prof.dr.ir. Z. Su, supervisor

ITC dissertation number 351
ITC, P.O. Box 217, 7500 AE Enschede, The Netherlands

ISBN 978-90-365-4752-9
DOI 10.3990/1.9789036547529

Data is accessible via DANS, DOI https://doi.org/10.17026/dans-xhv-s75j

Cover designed by Fakhereh Alidoost. The book cover template created by www.freepik.com and is available at http://www.vectorpicker.com. The photo in the center of the front cover created by Devany Walsh on the website http://vunature.com/. The graph of copula in the back cover is mentioned in Nelsen, 2006.

Printed by ITC Printing Department

ITC
UNIVERSITY OF TWENTE.
FACULTY OF GEO-INFORMATION SCIENCE AND EARTH OBSERVATION

Graduation committee:

Chairman/Secretary	
Prof.dr.ir. A. Veldkamp	University of Twente
Supervisors	
Prof.dr.ir. A. Stein	University of Twente
Prof.dr.ir. Z. Su	University of Twente
Members	
Prof.dr. G. van der Steenhoven	University of Twente
Prof.dr. A.D. Nelson	University of Twente
Prof.dr. A.A.M. Holtslag	Wageningen University
Prof.dr. P.J. Ward	Free University Amsterdam

This research was conducted under the auspices of the Graduate School for Socio-Economic and Natural Sciences of the Environment (SENSE).

To ambitious researchers

“It is unrealistic to walk into a room and flick a switch and lights come on. Fortunately, Edison didn’t think so.

We want to represent an idea. We want to represent possibilities. That some of you already know, that it is hard, it’s not easy, that in the process of working on your dreams, you are going to incur in a lot of disappointments, a lot of failures, a lot of pain. For those of you that have experienced some hardships, don’t give up on your dreams. The rough times are going to come, but they have not come to stay, they have come to pass.

After we face a rejection and a NO or we have a meeting and no one shows up, you’re still looking at your dreams and saying to yourself: It’s not over, until I win.” (Will Smith)

Acknowledgments

Thank God for the wisdom bestowed upon me, the strength and the talents. I’m thankful for many gifts God has given me like the great teacher who inspires me. I do not know how to express my deepest gratitude to Prof. Alfred Stein. His support throughout my research was like a heart that pumped lots of professional guidance, encouragements, and enthusiasms. With superior insight and vast experience, he suggested solutions to handle the problems that I faced. My profound appreciation is extended to Prof. Zhongbo (Bob) Su. He gave me the insights that sparked my imagination helping me to be innovative. His vision has deeply inspired me to develop ideas.

Completion of this research became possible with standing upon the shoulder of a giant: Dr. Ali Sharifi. He helped me to see further with his advice about both scientific research and life in general. Words are not enough to express my appreciation for many insightful conversations that we had.

Undertaking this Ph.D. gave me the pleasure to know amazing people. I am grateful to Gabriel Parodi and Prof. John van Genderen for their valuable friendship. I want to extend my sincere thanks to Prof. Marinus G. Bos, Dr. Benedikt Gräler, Dr. Chris Mannaerts and Dr. Yijian Zeng for their suggestions and comments through developing my Ph.D. proposal and methodology. I am gratefully indebted to Tonny Boeve, Jeroen Jansen and Dr. Ad van Dommelen for their help and supports throughout my Ph.D. journey.

One of the beautiful gifts that I am extremely grateful for having is my family. I extend my heartfelt gratitude to them for their love and empathy. Especially, my mother who is one of the strongest women I know and my identical twin sister who is an excellent researcher in remote sensing and photogrammetry. My success in achieving a Ph.D. is due to what they have done for me.

Another valuable gift that I have received is an intimate friendship with lovely and cheerful persons in Enschede, Mississauga, Tehran, and Lincoln. All of the memorable times that we shared helped me emotionally through the rough road to finish this thesis. Thank You, my dear friends.

My cordial gratitude go to my paranymphs, officemates, and colleagues for offering me a helping hand in this endeavor. My sincere thanks also go to the staff of the University of Twente and the faculty of ITC who provided me a friendly smile, a kind look, a gentle word and a hello every time we met.

Many thanks to all the people who have supported me to complete the research directly or indirectly. I acknowledge the kind cooperation of the European Centre for Medium-Range Weather Forecasts, the SAJ Consulting firm in Iran for weather data, Land Processes Distributed Active Archive Center of the U.S. Geological Survey for MODIS and Landsat 8 images, the Central Bureau for Statistics, European statistics database, the Royal Netherlands Meteorological Institute for weather data, and Cressie Communication Services for improving the English wording of chapters one, three and eight.

Summary

Environmental processes are driven by weather, land, and water variables and their interactions that change continuously in space and time. A complete process description considers both spatio-temporal dependencies and associations between those variables. Describing the dependencies is challenging because natural phenomena are often observed at a discrete set of locations and times. In this thesis I focus on reanalysis data of ECMWF[1] (ERA-I) that are being used increasingly for those process descriptions. Major dilemmas locally are that observations are sparse, and the use of reanalysis data is prone to uncertainty because of the coarse spatial resolution and systematic bias. The complete study of dependencies will also lead to an increase in the number of involved variables. To address these problems, this research demonstrates the potentials of copulas. It uses two datasets: daily mean air temperature collected from weather stations and reanalysis data in the Qazvin Plain, Iran, and daily air temperature and precipitation retrieved from weather stations and reanalysis data in the Netherlands.

First, copulas described the dependencies between measurements and reanalysis data in the absence of ancillary data in Iran. The conditional distribution of air temperature given the reanalysis data was estimated with copulas. This thesis illustrated a systematic bias in the reanalysis air temperature data as compared to weather station measurements. I predicted bias-corrected air temperatures using two new predictors based upon Conditional Probabilities (CP): CP-I offers a single conditional probability as a predictor, while CP-II is a pixel-wise version of CP-I and offers spatially varying predictors. The CPs reduced the bias with 44 – 68% as compared to commonly applied predictors. I concluded that CPs locally improved existing bias correction methods.

Second, copulas took care of the spatial dependencies between weather variables and associations between land variables. Ancillary information was obtained from remote sensing images. The classical and common method for bias correction, i.e. a univariate Quantile Mapping (QM) produced smooth maps. To locally rectify for smoothness, the conditional distribution of air temperature given reanalysis data and elevation was estimated with copulas. Three Multivariate Copula Quantile Mappings (MCQMs) were proposed to predict bias-corrected air temperature. MCQMs reduced bias with 16-63% as compared to QM. The study showed that MCQMs were well able to represent spatial and temporal variations of air temperature and its associations with elevation.

[1] ECMWF: the European Centre for Medium-range Weather Forecasts

Third, in this thesis I exploited copulas to improve the spatial resolution of air temperature data. Two new interpolators were investigated embedding remote sensing products, in particular land surface temperature, leaf area index and surface elevation: a spatial copula interpolator including covariates, and a mixed copula interpolator. The spatial copula interpolator including covariates improved the spatial predictions with 46-58% as compared to the spatial copula interpolator, the ordinary kriging predictor and the co-kriging predictor. The copula-based interpolators well represented spatial variability of air temperature and its associations with land variables at spatial resolution of 1 km. The methods are potentially useful for other sparsely and irregularly distributed weather data.

Fourth, copulas helped me to describe the multivariate dependencies of the weather extremes and yield, production, and price of potatoes in the Netherlands. In this thesis, a procedure was proposed to select the dominant driving climate indices of air temperature and precipitation in space. The conditional distributions of the non-climatic variables given the indices were estimated. The non-climatic variables were predicted with relative mean absolute errors equal to 5.4%, 3.6%, and 27.9%, respectively. I showed in this study that the proposed copula-based method optimally quantified the impact of climate extremes including their uncertainties.

The main conclusion drawn from this research is that copula-based methods can well represent the spatial variability and associations between air temperature and precipitation and other variables. They are also able to improve existing methods locally. Findings illustrate the practical advantages of copulas to describe multivariate dependencies, to define several predictors and to assess uncertainties.

Samenvatting

Processen in het milieu worden gedreven door weer-, land- en watervariabelen en hun interacties. Deze veranderen continu in ruimte en tijd. Een volledige procesbeschrijving houdt rekening met zowel spatio-temporele afhankelijkheden als associaties tussen deze variabelen. Het is een uitdaging om deze afhankelijkheden te beschrijven omdat natuurlijke fenomenen vaak worden waargenomen op discrete locaties en tijdstippen. In dit proefschrift heb ik me gericht op her-analyse weergegevens die worden verstrekt Europees Centrum voor weersvoorspellingen op middellange termijn (ECMWF). Deze gegevens worden in toenemende mate gebruikt voor procesbeschrijvingen in het milieu. Belangrijke dilemma's zijn dat waarnemingen lokaal en schaars zijn en dat het gebruik van her-analyse weergegevens gevoelig is voor onzekerheid vanwege de grote ruimtelijke resolutie en systematische vertekening. Een volledige studie van afhankelijkheden zal dan ook leiden tot een toename van het aantal betrokken variabelen. Om deze problemen aan te pakken, heb ik in dit onderzoek de mogelijkheden van copulas onderzocht. Ik heb gebruik gemaakt van twee datasets: dagelijkse gemiddelde luchttemperatuur verzameld door weerstations en her-analyse weergegevens in de Qazvin Plain, Iran, en de dagelijkse luchttemperatuur en neerslag afkomstig van weerstations en her-analyse weergegevens in Nederland.

Als eerste studie heb ik copulas gebruikt voor de Iraanse gegevens om de afhankelijkheden te beschrijven tussen metingen en her-analyse weergegevens in afwezigheid van aanvullende gegevens. De voorwaardelijke verdeling van de luchttemperatuur, gegeven de her-analyse weergegevens, heb ik geschat met copulas. Dit proefschrift liet een systematische onzuiverheid zien in her-analyse luchttemperatuurgegevens in vergelijking met metingen van weerstations. Luchttemperatuur gecorrigeerd voor onzuiverheid is voorspeld met behulp van twee nieuwe voorspellers op basis van voorwaardelijke waarschijnlijkheden (CP): CP-I biedt een enkele voorwaardelijke kans als voorspeller, terwijl CP-II een pixelgewijze versie van CP-I is en ruimtelijk variërende voorspellers biedt. De CP's verminderden de onzuiverheid met 44 - 68% in vergelijking met gangbare voorspellers. Ik kon concluderen dat CP's bestaande methoden voor de correctie van onzuiverheid lokaal hebben verbeterd.

Als tweede studie namen copulas de ruimtelijke afhankelijkheden tussen weervariabelen en associaties met landvariabelen mee. Aanvullende informatie is verkregen vanuit satellitebeelden. De klassieke, gebruikelijke methode voor correctie van onzuiverheden, namelijk een univariate Quantile Mapping (QM), produceerde continue kaarten. Om plaatselijk te corrigeren voor discontinuiteit, heb ik de voorwaardelijke verdeling van de luchttemperatuur, gegeven de her-analyse weergegevens en hoogte, geschat met copulas. Ik heb drie multivariate kwantiel karterings methoden gebaseerd op copula’s

(MCQM's) voorgesteld om luchttemperatuur gecorrigeerd voor onzuiverheid te voorspellen. MCQM's verminderden de onzuiverheid met 16-63% in vergelijking met QM’s. De studie toonde aan dat MCQM's goed in staat waren om ruimtelijke en temporele variaties van de luchttemperatuur en de associaties ervan met de hoogte weer te geven.

Als derde studie in dit proefschrift heb ik gebruik gemaakt van copulas om de ruimtelijke resolutie van luchttemperatuurgegevens te verbeteren. Twee nieuwe interpolatoren zijn onderzocht voor het inbedden van remote sensing-producten, in het bijzonder landoppervlaktetemperatuur, de bladoppervlakte-index en de hoogte van het aardoppervlak: een ruimtelijke copula-interpolator inclusief covariabelen en een gemengde copula-interpolator. De ruimtelijke copula-interpolator inclusief covariabelen verbeterde de ruimtelijke voorspellingen met 46-58% in vergelijking met de ruimtelijke copula-interpolator, de gewone Kriging-voorspeller en de cokriging voorspeller. De op copula gebaseerde interpolatoren gaven de ruimtelijke variabiliteit van de luchttemperatuur en de associaties met landvariabelen goed weer bij een ruimtelijke resolutie van 1 km. De methoden zijn mogelijk nuttig voor andere schaarse, onregelmatig verspreide weergegevens.

Als vierde studie hielpen copulas mij om de multivariate afhankelijkheden te beschrijven tussen de extreme weersomstandigheden enerzijds en opbrengst, productie en prijs van aardappelen in Nederland anderzijds. Ik stel hiervoor een procedure voor om de dominante en drijvende indicatoren van het klimaat met betrekking tot de ruimtelijke luchttemperatuur en neerslag te selecteren. De voorwaardelijke verdelingen van de niet-klimatologische variabelen gegeven de indicatoren heb ik geschat. De niet-klimatologische variabelen zijn voorspeld en gaven relatief gemiddelde absolute fouten gelijk aan respectievelijk 5,4%, 3,6% en 27,9%. De studie toonde aan dat de voorgestelde methode die gebaseerd is op copulas de impact van klimaatextremen, inclusief hun onzekerheden, optimaal kon kwantificeren.

De belangrijkste conclusie van mijn onderzoek is dat op copula gebaseerde methoden goed de ruimtelijke variabiliteit en associaties tussen luchttemperatuur en neerslag met andere variabelen kunnen weergeven. Ze kunnen ook bestaande methoden lokaal verbeteren. Mijn bevindingen illustreren de praktische voordelen van copulas om multivariate afhankelijkheden te beschrijven, om verschillende voorspellers te definiëren en om onzekerheden te beoordelen.

Summary in Farsi

محیط زیست شامل مجموعه‌ای از عوامل طبیعی مرتبط به هوا، سطح زمین و آب می‌شود. این عوامل طبیعی نه تنها به عنوان متغیرهای مکانی-زمانی در نظر گرفته می‌شوند، بلکه نوع وابستگی آن‌ها به یکدیگر نیز به طور پیوسته در طول مکان و زمان تغییر می‌کند. در نتیجه به منظور مطالعه یک فرآیند محیطی، هم تغییرات مکانی-زمانی عوامل مختلف و هم وابستگی بین آن‌ها باید در نظر گرفته شود. از آنجا که معمولا یک عامل طبیعی فقط در مکان‌ها و زمان‌های مشخصی مشاهده و اندازه گیری می‌شود، شناخت کامل وابستگی‌های عوامل طبیعی به یکدیگر مشکل می‌گردد. عامل طبیعی مورد مطالعه در این تحقیق، دمای هوا است. داده‌های هواشناسیERA-I مرکز اروپایی پیش‌بینی هوا در مقیاس متوسط[1] به طور گسترده به منظور مطالعات دمای هوا و یا به طور کلی، شناخت فرآیندهای طبیعی به کار گرفته می‌شوند. اما تهیه نقشه‌های دمای هوا در مقیاس‌های محلی با چالش های متعددی روبرو است، به عنوان مثال: تعداد ایستگاه های هواشناسی اغلب کم بوده و ایستگاه ها به صورت پراکنده توزیع شده‌اند، عدم قطعیت‌های ناشی از قدرت تفکیک مکانی پایین و خطاهای سیستماتیک در داده‌های مدل‌های جهانی هواشناسی وجود دارد، و پارامترهای بسیاری باید مورد مطالعه قرار گیرند تا بتوان شرح دقیقی از وابستگی دمای هوا با سایر عوامل طبیعی ارائه داد. در این تحقیق مزایای استفاده از توابع همبستگی[2] به منظور حل این چالش‌ها ارائه می‌گردد. در این راستا، اطلاعات مدل‌های جهانی هواشناسی و ایستگاه‌های هواشناسی در دو منطقه مطالعاتی در نظر گرفته شده‌اند: متوسط روزانه دمای هوا در دشت قزوین در ایران، و مقادیر روزانه دمای هوا و بارندگی در کشور هلند.

در مرحله اول از روش پیشنهادی، توصیف وابستگی بین داده‌های ایستگاه‌های هواشناسی و داده‌های ERA-I در ایران با استفاده از توابع همبستگی و بدون داده‌های کمکی مدنظر قرار گرفته است. بدین منظور، تابع توزیع شرطی دو-بعدی دمای هوا با استفاده از توابع همبستگی تخمین زده می‌شود. همچنین، داده‌های ERA-I با داده‌های ایستگاه‌های هواشناسی مقایسه و خطاهای سیستماتیک در این داده‌ها بررسی شدند. دو روش جدید بر اساس احتمالات شرطی[3] برای کاهش خطاهای سیستماتیک و بهبود دقت نقشه‌های دمای هوا ارائه شده است که عبارتند از: در روش اول از یک مقدار احتمال شرطی برای تمام نقاط گرید استفاده شده است، و در روش دوم یک مقدار احتمال شرطی برای هر نقطه در گرید در نظر گرفته شده است. در مقایسه با روش‌های موجود که از تابع توزیع شرطی دو-بعدی استفاده می‌نمایند، این دو روش پیشنهادی باعث کاهش خطاها در حدود 44 الی 68 درصد شده اند. در نتیجه، روش‌های ارائه شده در این تحقیق، مبتنی بر احتمالات شرطی منجر به بهبود روشهای معمول می‌شوند.

در مرحله دوم، وابستگی بین داده‌های ایستگاه‌های هواشناسی و داده‌های ERA-I در ایران به کمک توابع همبستگی و با در نظر گرفتن داده‌های کمکی سنجش از دور توصیف شده است. تناظریابی احتمالات[4] به کمک تابع توزیع یک-بعدی، رو شی معمول برای ت صحیح خطاهای سی ستماتیک ا ست. نق شه‌های تولید شده با این روش، تغییرات مکانی اطلاعات دمای هوا را به درستی نشان نمی‌دهند. به منظور در نظر گرفتن تغییرات مکانی، تابع توزیع شرطی سه-بعدی دمای هوا با استفاده از توابع همبستگی و به کمک داده‌های ایستگاه‌های هواشناسی، داده‌های ERA-I و داده‌های ارتفاع

[1] the European Centre for Medium-range Weather Forecasts
[2] copulas
[3] Conditional probabilities
[4] Quantile mapping

سطح زمین تخمین زده می‌شود. در این راستا، سه روش جدید به منظور تناظریابی احتمالات بر مبنای تابع توزیع چند-بعدی[1] ارائه شد ه است. بر اساس نتایج حاصله، روش‌های جدید چند-بعدی در مقایسه با روش تناظریابی یک-بعدی، باعث کاهش خطاها در حدود 16 الی 63 درصد شده‌اند. همچنین این تحقیق نشان داده است که نقشه‌های بدست آمده از روش‌های جدید چند-بعدی، هم تغییرات مکانی دمای هوا و هم وابستگی دما با ارتفاع سطح زمین را به خوبی نشان می‌دهند.

در مرحله سوم، از توابع همبستگی برای بهبود قدرت تفکیک مکانی نق شه‌های دمای هوا ا ستفاده شد ه ا ست. بدین منظور، دو روش درونیابی با در نظر گرفتن داده‌های کمکی سنجش از دور (از قبیل دمای سطح زمین، شاخص سطح برگ، و ارتفاع سطح زمین) ارائه می‌شود که عبارتند از: روش درونیابی شامل متغیرهای کمکی، و روش درونیابی مختلط. در مقای سه با روش‌های موجود همچون درونیابی مکانی بر ا ساس توابع همبستگی (بدون متغیرهای کمکی) و روش ordinary Kriging، روش پیشنهادی درونیابی شامل متغیرهای کمکی باعث بهبود نقشه‌های دمای هوا در حدود 46 الی 58 درصد شد هاست. همچنین روش‌های ارائه شده در این تحقیق، قادر به نشان دادن تغییرات مکانی دمای هوا و وابستگی دمای هوا با ارتفاع سطح زمین در قدرت تفکیک مکانی یک کیلومتر ه ستند. بنابراین این روش‌ها برای تهیه نقشه از داده‌های پراکنده و نامنظم هواشناسی سودمند هستند.

در مرحله چهارم، از توابع همبستگی در توصیف وابستگی‌های چند-بعدی بین بحران‌های طبیعی، میزان تولید و قیمت محصول سیب زمینی در کشور هلند ا ستفاده شده ا ست. بدین منظور در این تحقیق، روش نوینی برای انتخاب پدیده غالب[2] دما و بارش در کل کشور پی شنهاد شد. تابع توزیع شرطی چند-بعدی با ا ستفاده از توابع همبستگی و به کمک پارامترهای گیاهی و پدیده‌های غالب تخمین زده شد. طبق نتایج حاصله، مقادیر تولید در سطح، تولید و قیمت محصول با خطاهای ن سبی به ترتیب 5.4، 3.6 و 27.9 در صد بد ست آمده‌اند. این مطالعه ن شان داد که با روش پی شنهادی بر اساس توابع همبستگی می توان اثرات تغییرات آب و هوایی و عدم قطعیت‌ها را بررسی کرد.

در این تحقیق، به طور خاص نشان داده شده ا ست که روش‌های مبتنی بر توابع همبستگی با دقت خوبی قادر به نشان دادن تغییرات مکانی و وابستگی بین عوامل محیطی می باشند. همچنین، این روش های پیشنهادی باعث بهبود عملکرد روشهای معمول می شوند. یافته های حاصل از این تحقیق بیانگر کاربردهای توابع همبستگی در توصیف وابستگی های چند-بعدی، و مطالعه عدم قطعیت ها است.

[1] multivariate copula quantile mapping
[2] weather extreme

Table of Contents

List of figures

List of tables

List of abbreviations

AIC	Akaike’s Information Criteria
CBS	Central Bureau for Statistics
CDF	Cumulative Distribution Function
CE	Conditional Expectation
CM	Conditional Median
CP	Conditional Probabilities
ECMWF	European Centre for Medium-range Weather Forecasts
ERA-I	ECMWF ReAnalysis-Interim
ES	Error Score
ETCCDI	Expert Team on Climate Change Detection and Indices
Eurostat	European statistics database
ID	Identically Distributed
LAI	Leaf Area Index
LST	Land Surface Temperature
MAB	Mean Absolute Bias
MAE	Mean Absolute Error
MAPE	Mean Absolute Prediction Error
MCQM	Multivariate copula quantile mapping
MODIS	Moderate Resolution Imaging Spectroradiometer
PDF	Probability Density Function
PIW	Prediction Interval Width
QM	Quantile Mapping
RMAE	Relative Mean Absolute Error
SCS	Spatial Correlation Score
SDG	Sustainable Development Goal
SES	Spatial Error Score
TCS	Temporal Correlation Score
TES	Temporal Error Score

Chapter 1: Introduction

This chapter provides a brief overview of the research topic and the reasons for conducting the research: motivation, scientific problems, research objectives and questions, innovations and scope.

1.1 Motivation

Competition for natural resources, i.e. land and water, is increasing due to population growth, industrial development, agricultural intensification and climate change. These forces are leading to water/food scarcity, air pollution, drought and, subsequently, environmental degradation.

With respect to climate change, increasing variation in air temperature and precipitation affects agriculture (e.g. crop production), contributing to risks for food security. The crop responses to those changes are representative of many complex processes and interactions at local scales (Challinor et al. 2009a). When studying those processes, it is of interest to quantify the changes in air temperature and precipitation because those variables result into a variety of climate-related crop stresses. Indeed, they are key for assessing crop water requirements.

There are two common sources of weather data: weather stations and weather forecasting systems. The sparseness of weather stations and doubtful maintenance of their instruments create uncertainty about their data and, consequently, about the results of hydrological/agricultural studies. The European Centre for Medium-range Weather Forecasts (ECMWF), on the other hand, provides ERA-Interim (ERA-I) reanalysis weather data that are being used increasingly (Persson 2013). ERA-I generate the weather data at spatial grids that are typically of an order of 10 kilometers (see Section 3.1). Typically, an ERA-I archive can provide historical, real-time and forecast weather data. Potentially, these data could play an important role in supporting information systems, e.g. climate information services and irrigation advisory services.

With regard to hydrological/agricultural studies at regional and local scales, the report Sustainable Development Goals (SDGs) 2018 mentions that in many parts of the world, such as Asia and Africa, data at those scales are needed to produce information required for the management of natural resources. Nowadays, there is substantial potential for the use of remote sensing, in particular, satellite measurements due to improved spectral bandwidth and spatial and temporal resolutions (Mulla 2013). However, satellite data acquisition includes the quantization of continuous information, which is susceptible to uncertainty because of the influence of mixed pixels, cloud cover, and pre-processing steps for atmospheric, radiometric, and geometric correction. There are, nevertheless, growing appeals for the integration of multi-sensor, multi-resolution products and in-situ data. The research reported in this thesis was carried out in a data-scarce environment and benefits from the use of Earth observation data.

Returning to the topics of climate change and weather data, a main aspect of recent studies has been to describe their variation in both space and time, i.e. spatio-temporal variability, and dependencies between several weather parameters, i.e. covariability. For that purpose, geostatistical methods play an essential role when studying the dependencies, i.e. in modeling the underlying process. They also offer the advantage of being able to predict spatio-temporal information. In recent decades it has been suggested that copulas may be used to construct multivariate distributions (Sklar 1973). Nevertheless, the exploitation of copulas in geostatistics is still in its infancy (Bárdossy and Li, 2008; Gräler and Pebesma 2011). In this light, exploration of the potential of geostatistical methods for improving the modeling is of interest. The methods I have chosen to investigate therefore dearly belong to the domain of geostatistics and offer a wide range of potential applications in agricultural, hydrological, and climate studies.

Weather data are essential input for developing information systems, e.g., climate information services, and irrigation advisory services. Processing of weather data to generate information at regional and local scales is a challenge for the analyst. In this research, I developed new copula-based methods and compared them with several methods commonly applied for improving reanalysis weather data generated by ERA-I. For the comparison, techniques of multi-criteria evaluation and sensitivity analysis are applied. The motivation behind these comparative analyses is to explore the advantage/disadvantages of the copula-based methods. The strength and limitation of the methods are discussed through chapters 4-7 in the sections: 4.5 Discussion, 5.4 Results and discussion, 6.5 Discussion and conclusion, and 7.5 Discussion and conclusion. The findings are summarized in Section 8.1.

1.2 Problem statement

A challenging problem in many parts of the world is the use of weather data for providing information at local scales. The reason for this is that weather stations are often sparsely and irregularly distributed in many regions. Hydrological/agricultural studies may find it useful to use ERA-I reanalysis data to address the problem of the scarceness because ERA-I produce spatially well-dispersed weather data. Over- or underestimation and the coarse spatial resolution of ERA-I may, however, prohibit the use of their data for studying interactions between weather and non-climatic variables at local scales (Challinor et al. 2009a). In such cases, application of geostatistical methods for prediction purposes may provide an alternative solution. As regards predicting spatial variation of weather values, a practical side effect of the standard geostatistical methods is that they produce smooth maps.

There is a further problem that has received substantial attention in most climate change studies. Evaluation of the implications of climate change requires understanding the variation in several weather variables and non-climatic variables, i.e. covariability. A well-known technique for considering several dependencies is to estimate multivariate joint distributions. The estimation of a *d*-dimensional distribution, $d > 2$, however, is often not straightforward (Salvadori et al. 2007). Previous studies have introduced simplifications regarding the number of variables involved in the modeling of the dependencies (Miao et al. 2016).

1.3 Research questions and objectives

Weather and land variables and their interactions change continuously in space and time. Modeling spatio-temporal dependencies and associations between those variables involves a large number of variables. I investigated copula-based methods for describing the dependencies with the aims of being able to:

- Refine locally reanalysis weather data retrieved from ERA-I, to deal with data scarcity;
- Explore the potential of copulas for including ancillary remote sensing data in the modeling of dependencies;
- Produce weather maps in a data-scarce environment and to improve the spatial resolution of reanalysis weather data from ERA-I; and
- Assess the impacts of climate change on crop-related variables.

The key contributions of this research can be found in the answers it provides for the following research questions:

- How can reanalysis weather data generated by ERA-I be improved locally in a data-scarce environment by taking into consideration spatial variability and the covariability of the data?
- What are the advantages/disadvantages of applying bias correction methods as seen from the perspective of the users concerned with spatial and temporal characteristics of weather data?
- Does the integration of remote sensing data and statistical methods help improve the prediction of weather data in the spatial domain?
- How can ancillary data be embedded as additional variables in the modeling of spatial random fields using multivariate distributions?
- Can copulas describe a complex process such as the interactions between crop-related variables and weather data?
- What are the impacts of weather extremes on crop-related variables?

In line with the aims of my research, this thesis focuses on bias correction, interpolation, and joint behavior analysis in four real scenarios. The aims of my research can therefore be restated in the form of the following objectives:

Objective 1: To develop new methods to correct for bias in daily reanalysis weather data from ERA-I for an agricultural area. The methods should describe the dependencies between reanalysis weather data and weather station measurements by estimating their joint distribution.

Objective 2: To develop new methods to correct for bias in daily reanalysis weather data from ERA-I that take into consideration covariability.

Objective 3: To predict weather data that take into consideration dependencies between weather and land variables retrieved from remote sensing products.

Objective 4: To analyze the joint behavior of climate extreme indices and non-climatic variables and to determine the impacts of climate change.

1.4 Innovations and scope

This thesis focuses on a relatively new approach for describing the dependencies between weather and non-climatic variables that has emerged following the application of an advanced geostatistical technique, i.e. copulas. The novel aspects of this approach lie in the integration of data/information from several sources and definition of copula-based predictors to improve the predictions of weather and non-climatic variables. The following is a brief description of the study in context of the research objectives.

In an agricultural area in Iran in which weather stations are sparse, additional spatially distributed weather data are required for an information service (e.g. irrigarion advisory service). The gridded ERA-I reanalysis weather data is available from the European Centre for Medium-range Weather Forecasts (ECMWF) (Persson 2013). Air temperature data retrieved from ECMWF show a systematic bias concerning measurements from the weather stations. So far, copula-based methods for bias correction have mainly been applied to precipitation time-series (Laux et al. 2011; Vogl et al. 2012; Mao et al. 2015). Little attention has, however, been given to correction bias in air temperature data, in particular, in data-scarce environments. Moreover, few studies have considered the spatial variability weather data corrected for bias. Copula-based methods have been investigated with the goal of improving spatial prediction using the dependencies between air temperature data applied by ECMWF and data from weather stations.

To add more information for bias correction, I extended the one-dimensional quantile mapping to a multivariate copula quantile mapping (MCQM). To my knowledge no previous research has applied MCQM to a data-scarce environment. I, therefore, explored whether adding ancillary information can improve the spatial variability and covariability of air temperature.

Essentially, the spatial prediction of weather data needs to consider both spatial variability and dependency with other variables, i.e. covariability. Few studies have shown how to embed ancillary data in the modeling of a spatial process. Moreover, common geostatistical methods produce smooth maps. Consequently I investigated the potential of two copula-based interpolators for improving the spatial resolution of ECMWF air temperature data by using remote sensing products.

In studies of local climate change, it is of interest to quantify changes that impact crops, particularly the impact of changes on crop yield (Pirttioja et al. 2015; Challinor et al. 2013). The impact on crop production and price have rarely been studied. Copulas describe the joint behavior of climate extreme indices and non-climatic variables, e.g. yield, production, and prices of potatoes in the Netherlands. For the study I selected seven climate extreme indices related to variations in air temperature and precipitation data.

1.5 Outline

This thesis comprises eight chapters. The developed methods in chapters 4-7, each is based upon one of the above objectives. They are all based on ISI-indexed journal articles that have been already published or are being revised for publication.

Chapter 1: Introduction. The motivation, scientific problems, research questions and objectives are described. Here answers are provided for the questions of why (the motivation), what (research questions and objectives) and how (methods).

Chapter 2: Copulas. This chapter describes the main copula theorems, explains how a joint cumulative distribution is estimated by fitting copulas to data, and indicates which predictors can be defined to predict random variables.

Chapter 3: Case studies. The first three objectives of the research focus on data from Iran (my home country), while the fourth objective focuses on data from the Netherlands. The methods used are, however, generic and can be applied in different cases.

Chapter 4: The use of bivariate copulas for bias correction of air temperature data sourced from ECMWF. The study presents two methods for predicting weather data that are based upon conditional probability (CP): CP-I offers a single conditional probability as the predictor, whereas CP-II provides spatially varying predictors.

Chapter 5: Multivariate copula quantile mapping for bias correction of air temperature data generated by ERA-I. This chapter presents three multivariate copula quantile mappings (MCQMs): MCQM-I uses the dependence between air temperature and elevation, MCQM-II uses the dependence between air temperatures at a single location and its nearest neighbor; and MCQM-III combines the first two methods.

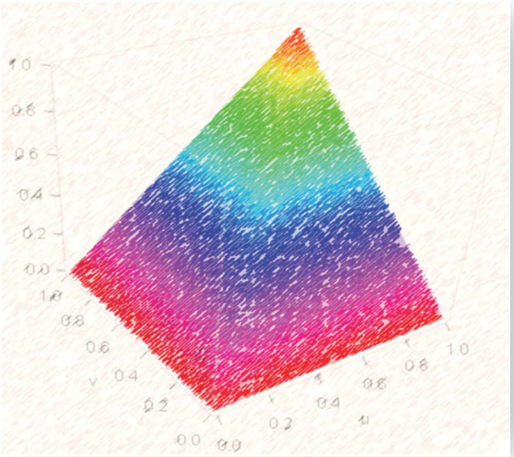
Chapter 6: Copula-based methods for interpolation of air temperature data using collocated covariates. 1) A spatial copula interpolator including covariates to consider two types of dependencies that are spatial dependences of air temperature at a single location and its nearest neighbors, and non-spatial dependencies between air temperature and its collocated covariates at that location. 2) A mixed copula interpolator extends the first method by including the non-spatial dependencies of air temperature and its collocated covariates at the nearest neighbors.

Chapter 7: Evaluating the effects of climate changes on crop production and price using multivariate distributions -a new copula application. Here a comprehensive copula-based analysis is presented for assessing the impact of climate change on the yield, production, and price of potatoes.

Chapter 8: Synthesis. I summarize the results and synthesize the research findings, pointing out significances, obstacles, prospects and limitations.

Chapter 2: Copulas

It illustrates the main copulas theorems, how a joint cumulative distribution is estimated by fitting copulas to data, and what predictors can be defined to predict random variables.



Copula /kɒpjʊlə/: the name comes from the Latin for "link" or "tie".

2.1 Main definitions

I devoted this section to giving a brief overview of copulas and basic probabilistic properties of distributions. I recommend Section 3.2 in Nelsen 2006, for a good "Geometric description" that defines copulas without a reference to distributions. In the following, the uppercase letters denote "variables," and the lowercase letters denote their "values". I, also, use a lowercase letter to indicate a density function whereas an uppercase letter for a cumulative function.

Sklar's theorem:

If F is a *n*-dimensional joint distribution function with 1-dimensional margins $F_1, \dots, F_n$, then a function C exists from the unit *n*-cube to the unit interval such that $F(x_1, \dots, x_n) = C(F_1(x_1), \dots, F_n(x_n))$ for all real *n*-tuples $(x_1, \dots, x_n)$.

The joint distribution function of two random variables X and Y is $F(X,Y)$ where the joint probability of $P[X \leq x,\ Y \leq y]$ is equal to $F(x,y)$. According to Sklar's theorem, there is a unique function $C(.,.)$ that assigns each pair of $\left(u = F_X(x), v = F_Y(y)\right)$ to $F(x,y)$, where F_X and F_Y are continuous marginal distributions, u is the probability of $P[X \leq x]$, and $v = P[Y \leq y]$ (Figure 2.1) (Sklar 1973). This function is called a copula and is a joint distribution function indicated as $C(U,V)$, where U and V are uniformly distributed random variables (Nelsen 2006). The name "copula" comes from the Latin for "tie" or "link": a copula joins (links) a joint distribution to its univariate marginals.

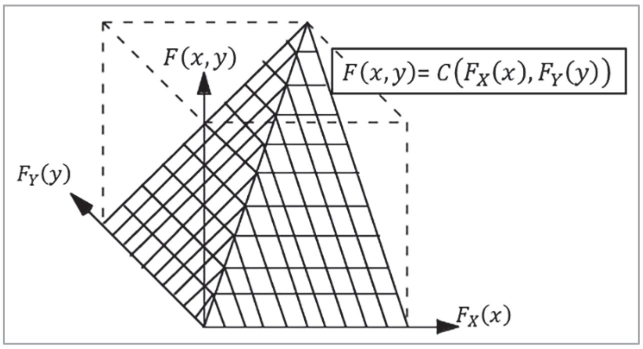


Figure 2.1 Graph of a copula (Nelsen 2006).

To understand the role of Sklar's theorem in determining the desired distribution $F(X,Y)$, I summarize the fundamental equalities between operations on distribution functions for a bivariate case as:

$$F(x,y) = C(u,v), \tag{2.1}$$

$$f(x,y) = \frac{\partial^2 F(X,Y)}{\partial X \partial Y} = \frac{\partial^2 C(U,V)}{\frac{\partial U}{f_X(x)} \frac{\partial V}{f_Y(y)}} = c(u,v) \times f_X(x) \times f_Y(y), \tag{2.2}$$

$$f_U(u) = f_V(v) = 1, F_U(u) = u,\ F_V(v) = v, \tag{2.3}$$

$$F(x|y) = \frac{1}{f_Y(y)} \times \frac{\partial F(X,Y)}{\partial Y}|_{Y=y} = \frac{\partial C(U,V)}{\partial V}|_{V=v} = C(u|v), \tag{2.4}$$

$$f(x|y) = \frac{f(x,y)}{f_Y(y)} = \ c(u,v) \times f_X(x) = c(u|v) \times f_X(x), \tag{2.5}$$

where $F(.,.)$ and $C(.,.)$ are cumulative distribution functions (CDF), $f(.,.)$ and $c(.,.)$ are probability density functions (PDF), $F(.|.)$ and $C(.|.)$ are conditional CDF, $f(.|.)$ and $c(.|.)$ are conditional PDF, U and V are uniformly distributed random variables (Kuipers and Niederreiter, 2012).

Equation (2.5) shows that the joint density probability $c(u,v)$ is equal to the conditional density probability $c(u|v)$. This equality holds only in a two-dimensional case, because $c(u|v) = \frac{c(u,v)}{f_V(v)}$ and $f_V(v) = 1$. In some literature, the conditional PDF and CDF of copulas are indicated as $c_v(u)$ and $C_v(u)$, respectively (Nelsen 2006, p. 41).

The equations can be extended to n dimensions as:

$$F(x_1, x_2, \dots, x_n) = C(u_1, u_2, \dots, u_n). \tag{2.6}$$

$$f(x_1, x_2, \dots, x_n) = \frac{\partial^n F(X_1, X_2, \dots, X_n)}{\partial X_1 \partial X_2 \dots \partial X_n} = c(u_1, u_2, \dots, u_n) \times \prod_{i=1}^{n} f_{X_i}(x_i)\,. \tag{2.7}$$

$$F(x_0|x_1, x_2, \dots, x_n) = C(u_0|u_1, u_2, \dots, u_n). \tag{2.8}$$

$$f(x_0|x_1, x_2, \dots, x_n) = f_{X_0}(x_0) \times c(u_0|u_1, u_2, \dots, u_n) = f_{X_0}(x_0) \times \frac{c(u_0, u_1, u_2, \dots, u_n)}{c(u_1, u_2, \dots, u_n)}. \tag{2.9}$$

The conditional density $c(u_1, u_2, \dots, u_n)$ in the denominator of equation (2.9) is obtained as $c(u_1, u_2, \dots, u_n) = \int_0^1 c(u_0, u_1, u_2, \dots, u_n) du_0$. This equality holds for any joint and marginal densities in probability theory, e.g., $f(x) = \int_y\ f(x,y)dy$.

I provide five aspects to point out the usefulness of copulas in real-world applications:

- The definition of copulas is without indication about the underlying process. Any joint distribution can thus be written in terms of a copula, i.e., $F(x,y) = C(u,v)$. This illustrates the growing interest in copulas and their applications in diverse studies such as finance, image analysis, geostatistics, and in particular in the environmental sciences; hydrology, disasters, agriculture, weather and climate.

- The definition can be extended to higher dimensions including several random variables/fields: spatial dependences, temporal dependences, spatio-temporal dependences, and dependences between several variables at one point in time and space. This allows one to analyze the resultant effects of several variables in modeling the underlying process.
- The family distribution of C can be different from the family of $F(X,Y)$, F_X and F_Y. Therefore, copulas describe the dependences between variables in a different configuration from marginal distributions. For example, both X and Y can follow Gaussian distributions, but C can be a non-Gaussian joint distribution.
- Some traditional statistical methods assume identically distributed (ID) variables to simplify the underlying mathematics related to multivariate joint distributions. The assumption, however, may or may not be valid in practical studies. Copulas enable to construct multivariate distributions without the assumption of ID.
- The density function $c(.,.)$ in equation (2.2) can be interpreted as a measure for the strength of the dependence between the involved variables. The function c can exhibit several types of non-linear negative or positive dependences. Hence, for mutually independent variables, $c(u,v) = 1$ and $f(x,y) = f_X(x) \times f_Y(y)$.

2.2 Estimation

For the estimation of a two-dimensional distribution using copulas, two random variables X and Y are considered with a joint distribution $F(X,Y)$ that is equal to a copula $C(U,V)$ according to Sklar's theorem. There are several copula families in the literature to determine $C(.)$ (Joe 1993; Nelsen 2003; Demarta and McNeil 2005; Manner 2007). I choose the Gaussian, Student's *t*, Clayton, Gumbel and Frank families because other families lead to computational limitations (Gräler 2014). The Gaussian and Student's *t* belong to the elliptical copulas, whereas the remainder families are Archimedean copulas (Nelsen 2006). These families describe several types of the tail dependences and have one parameter that is related to Kendall's τ correlation between variables (Table 2.1).

The parameter for each family are estimated using maximum likelihood estimation and a starting value obtained by Kendall's τ (Nelsen 2006; Brechmann and Schepsmeier 2013). Then the best family for C is the one that minimizes the Akaike's Information Criteria (AIC) (Akaike, 1974). The p value of the null hypothesis of a bivariate independence is obtained based upon the statistical test developed by Genest et al. (2007). The p values of the null hypothesis that the dependence structure is well represented by the best fitting family are obtained using 100 bootstrap runs based upon the Cramér–von

Mises statistic $S_n^{(B)}$ for the Gaussian, Clayton, Gumbel and Frank families (Genest et al., 2009), and based upon the White statistic for the Student's *t* family (Huang and Prokhorov, 2014). This number of bootstrap runs is relatively small, but a larger number would substantially increase the computational cost (Kojadinovic et al., 2011). Further note that the selection of families depends upon both the number of observations and the probabilistic nature of the dependence between variables.

I can now illustrate the estimation of a high-dimensional distribution. The five bivariate families are extendable to higher dimensional ones (Nelsen 2006). Hence, the interdependencies between these variables are restricted to one specific family of copulas. In addition, the estimation of a multivariate copula is generally a troublesome procedure (Nelsen 2006; Aas et al., 2009). In geostatistics where we have a target variable to predict, an alternative to estimate a multivariate copula, and consequently a multivariate distribution is the use of a canonical vine or C-vine structure (Aas et al., 2009). The flexibility of choosing several families in the vine structure to describe the multivariate interdependencies is one of the practical advantages of copulas. Further note that after constructing the copula, other distribution functions are retrieved from the fundamental equalities (see Section 2.1).

Table 2.1 Five families of copulas used in this study. The best fitting family is selected according to the lowest value of Akaike Information Criteria (AIC).

Index	**Name**	***C*$_\theta$*(u,v)***	**Property index**
1	Gaussian	$\emptyset_R\left(\emptyset^{-1}(u), \emptyset^{-1}(v)\right); R = \begin{bmatrix} 1 & \theta \\ \theta & 1 \end{bmatrix}$	1, 2, 6
2	Student's *t*	$t_{R,\vartheta}\left(t_\vartheta^{-1}(u), t_\vartheta^{-1}(v)\right);\ R = \begin{bmatrix} 1 & \theta \\ \theta & 1 \end{bmatrix}$; $\vartheta = degree\ of\ freedom$	1, 2, 6
3	Clayton	$\left[max\left\{\left(u^\theta + v^\theta - 1\right), 0\right\}\right]^{\frac{-1}{\theta}}$	1, 2,4,5,6
4	Gumbel	$\exp(-[(-lnu)^\theta + (-lnv)^\theta]^{\frac{1}{\theta}})$	1,2,3,6
5	Frank	$\frac{-1}{\theta}\ln(1 + \frac{(e^{-\theta u} - 1)(e^{-\theta v} - 1)}{e^{-\theta} - 1})$	1,2,6
1	Property	Permutation symmetry	
2		Symmetry about medians	
3		Extreme value copula	
4		Lower tail dependence	
5		Upper tail dependence	
6		Extendibility to multivariate copula	

Toy example, bivariate copulas:

The densities of Gaussian, Student's t, Clayton, Gumbel and Frank families are obtained for several dependence structures between two variables while the Kendall's τ is equal to 0.4 in all dependences (Figure 2.2). Kendall's τ is a non-linear measure of association between variables that can change over the range of [-1, 1] (Figure 2.3). The Clayton and Gumbel families, however, present only positive correlations (Figure 2.2).

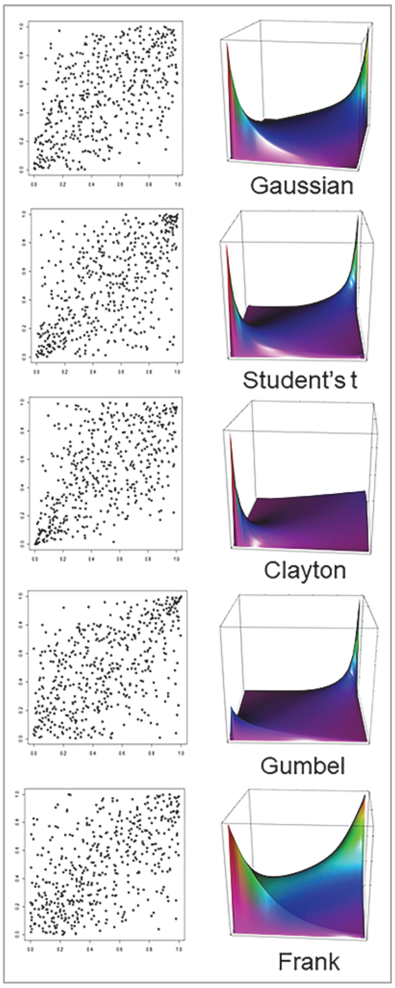


Figure 2.2 Five families of copulas. The densities of Gaussian, Student's *t*, Clayton, Gumbel and Frank families are presented for several dependence structures between two variables while the Kendall's τ is equal to 0.4 in all dependences. The horizontal axes are $\boldsymbol{u}$ and $\boldsymbol{v}$ and the third axes denote the density values. Different colors indicate different densities and are used for visualization purposes.

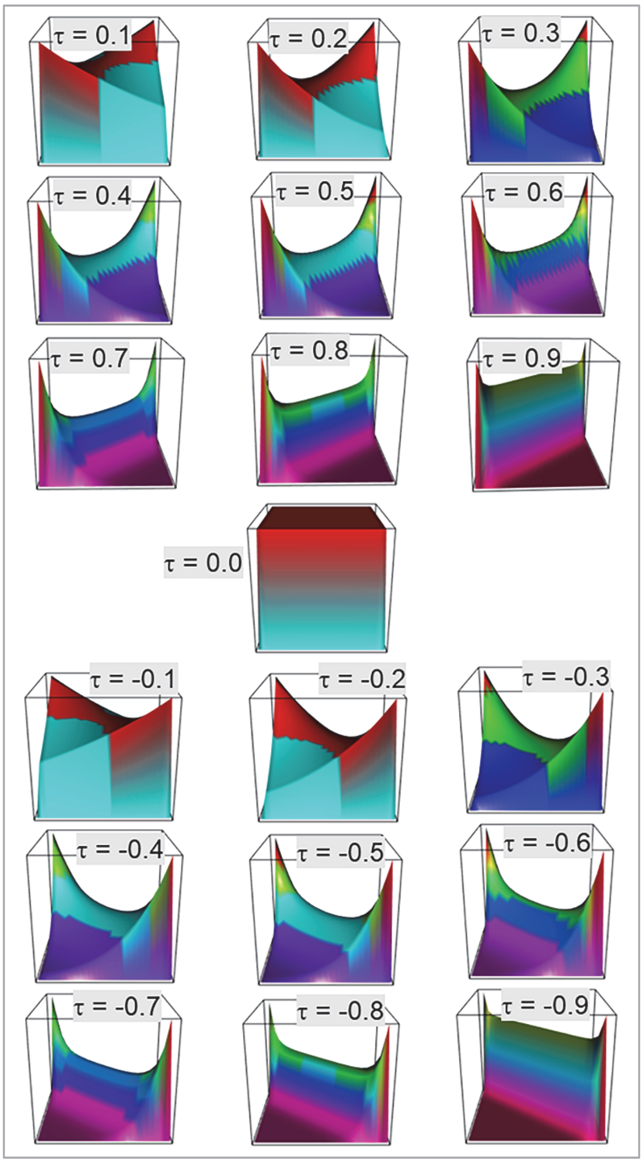


Figure 2.3 The densities of Frank copula for several values of the Kendall's τ. Different colors indicate different densities and are used for visualization purposes.

Toy example, trivariate copulas:

As an example, let's consider three random variables X, Y and Z with a copula $C(U,V,W)$, where X is the target variable. The central of the C-vine is, thus X (Figure 2.4) and the configuration of the structure is based upon bivariate copulas, Sklar's theorem and the general decomposition rule of $f(x,\, y,\, z) = f_Z(z) \times f(y|z) \times f(x|y,z)$.

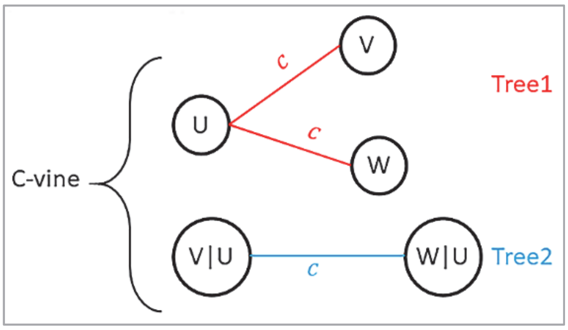


Figure 2.4 C-vine structure for three variables. A three-dimensional C-vine structure has two trees and three bivariate copulas which can belong to three different families.

In this example, the copula density $c(U, V, W)$ is first decomposed into bivariate copulas as: $c(U, W)$, $c(U, V)$ and $c\big(C(W|U), C(V|U)\big)$ (Figure 2.4). Then, each bivariate copula is estimated in a similar way to the two-dimensional case. Finally, the copula density is the product of all bivariate copula densities in the structure: $c(u, v, w) = c(u, w) \times c(u, v) \times c\big(C(W|u), C(V|u)\big)$. It follows that the dependence structure between those $n = 3$ variables is described by a combination of n different families and in total $\frac{n\times(n-1)}{2}$ parameters.

2.3 Prediction

Assume that the conditional distribution $F(X|.)$ is estimated and the random variable X is to be predicted. Any p^{th} percentile in the distribution can be used to predict X, i.e., to obtain a single value $\hat{x}$:

$$\hat{x}_p = F^{-1}(p|.), \qquad p \in [0,1], \tag{2.10}$$

$$\hat{x}_{mean} = E[X|.] = \int_x x \cdot f(x|.)dx\,, \tag{2.11}$$

$$\hat{x}_{median} = F^{-1}(0.5|.), \tag{2.12}$$

where $\widehat{\ }$ denotes $\hat{x}$ as a predicted value, $E[.]$ denotes the mathematical expectation. The choice of the p^{th} percentile in (2.10) depends upon the problem at hand. For instance, it can be obtained by a quantile mapping procedure (see Chapter 5). The conditional expectation (2.11) and the conditional median (2.12) are the optimal predictors, minimizing mean squared prediction error and mean absolute prediction error, respectively (Journel 1984; Cressie 1993). There are two common procedures using copulas to obtain $\hat{x}_{mean}$ and $\hat{x}_{median}$: the analytical evaluation (Bárdossy and Li 2008, Gräler 2014), and simulations (Salvadori et al. 2007).

For an analytical evaluation, the equations (2.11) and (2.12) are rewritten as:

$$\hat{x}_{mean} = \int_0^1 F_X^{-1}(u) \times c(u|.)du\,, \tag{2.13}$$

$$\hat{x}_{median} = F_X^{-1}\big(C^{-1}(0.5|.)\big), \tag{2.14}$$

where $u = F_X(x)$, $c(.|.)$ is the conditional PDF and $C^{-1}(.|.)$ is the inverse transformation of the conditional CDF. The new form of the conditional expectation in (2.13) is explained based upon the equalities in Section 2.1 as:

$$\hat{x}_{mean} = \int_x x \cdot f(x|.)dx \; = \int_0^1 F_X^{-1}(u) \times c(u|.) \times f_X(x)dx$$
$$= \int_0^1 F_X^{-1}(u) \times c(u|.) \times \frac{dF_X(x)}{dx}dx = \int_0^1 F_X^{-1}(u) \times c(u|.)du\,. \tag{2.14}$$

What follows is a property of the conditional expectation using a bivariate function $f(x|\mathrm{y})$. Let (x_1, y_1), (x_2, y_2), …, (x_n, y_n) be a set of paired observations for variables X and Y. If $x_1 > x_2$ and $y_1 > y_2$ or if $x_1 < x_2$ and $y_1 < y_2$, the pairs are called concordant, whereas if $x_1 > x_2$ and $y_1 < y_2$ or if $x_1 < x_2$ and $y_1 > y_2$ they are discordant. When the number of concordant pairs n_c is more than the number of discordant pairs n_d, the dependence between X and Y is positive, whereas when $n_c < n_d$, the dependence is negative (Nelsen 2006). Hence, if a bivariate copula represents a positive correlation, the conditional expectation is an increasing function of the conditioning variable, i.e., if $y_1 > y_2$, then $E[X|\,y_1] > E[X|\,y_2]$ (Dodds et al., 1990). If a bivariate copula represents a negative correlation, the conditional expectation is a decreasing function, therefore, if $y_1 < y_2$ then $E[X|\,y_1] < E[X|\,y_2]$.

Regarding the simulation method, the equation (2.10) is rewritten using copulas as:

$$\hat{x}_p = F_X^{-1}\big(C^{-1}(p|.)\big). \tag{2.15}$$

Several $\hat{x}_p$ are obtained by generating random probabilities p on [0, 1]. The mean of the obtained values provides $\hat{x}_{mean}$, whereas choosing the median of the obtained values is $\hat{x}_{median}$. For a large simulations, the results are equal to the results of the analytical evaluation (Mao et al. 2015).

Equation (2.10) is also useful to assess a γ% prediction interval width (PIW). For instance, a 95% PIW is obtained as $F^{-1}\big(C^{-1}(0.975|.)\big) - F^{-1}\big(C^{-1}(0.025|.)\big)$. The possibility of selecting several predictors is another practical advantage of copulas. Note that the three predictors have two main parts: a marginal distribution $F_X(.)$ and a multivariate copula $c(.|.)$. Hence, the overall prediction quality depends upon a good estimation of both functions.

2.4 Implementation

I provide some sample scripts for implementing the estimations and predictions in R using the packages copula (Kojadinovic and Yan, 2010), spcopula (Gräler and Pebesma, 2011), and VineCopula (Brechmann and Schepsmeier, 2013).

To estimate the five families based upon maximum likelihood and to select the best fitting family $C(U,V)$ using AIC:

```
BestFittingFamily <- BiCopSelect(U, V, familyset = c(1:5), selectioncrit =
"AIC", indeptest = T, rotations = F).
```

To construct a C-vine structure for $C(U,V,W)$ and estimate the bivariate families:

```
vineFit <- fitCopula(vineCopula(as.integer(3)), [U,V,W]).
```

```
vineStructure <- vineCopula (RVineCopSelect([U,V,W], familyset = c(1:5),
vineFit@copula@RVM$Matrix, rotations =T)).
```

To implement the three predictors, first $F_X^{-1}(.)$ is defined by the user, e.g., `InverseOfCDF`. Second, the best fitting family $C(U,V)$ is selected:

```
BestFittingFamily <- BiCopSelect(U, V, familyset = c(1:5), selectioncrit =
"AIC", indeptest = T, rotations = F).
```

```
BestFittingFamily    <-    copulaFromFamilyIndex(BestFittingFamily$family,
BestFittingFamily$par, BestFittingFamily$par2).
```

Finally, the variable X is obtained using one of the predictors:

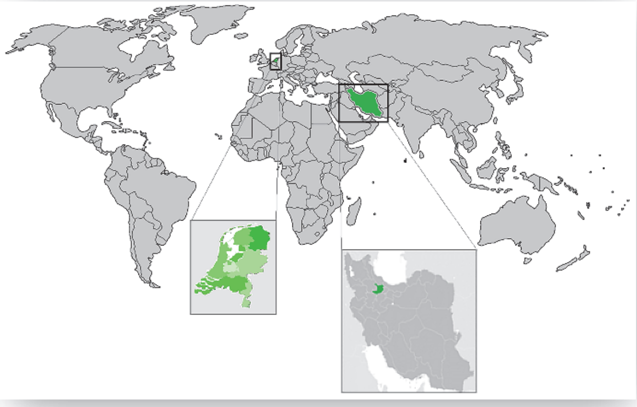
```
x_p       <- InverseOfCDF(invdduCopula(v, BestFittingFamily, p).
```

```
x_mean     <-    integrate(function(u)(InverseOfCDF(u)*dCopula(cbind(v,u),
BestFittingFamily)), 0.0, 1.0, subdivisions=1000L, stop.on.error=F).
```

```
x_median  <- InverseOfCDF(invdduCopula(v, BestFittingFamily, 0.5).
```

Chapter 3: Study area and data sets

The data used in this study consists of weather data (e.g. air temperature and precipitation), data sourced from remote sensing products and statistical databases. Bias correction and interpolation methods were applied to the data concerning Iran, whereas copula-based joint behaviors were applied to the data concerning the Netherlands.



3.1 Qazvin irrigation network

With rainfall limited in many places, Iran is a water-scarce country. This certainly applies to the Qazvin area, one of Iran's oldest and most advanced agricultural areas. Lying at an altitude of about 1,800 m above sea level, it has an arid climate, with an average annual precipitation of about 327 mm and an average daily temperature of 14°C. The Qazvin irrigation network, located on the Qazvin Plain (Figure 3.1), serves a predominantly mixed farming system: 50% of the network area is cultivated with winter crops, while some 20-25% of the area produces summer crops (Sharifi 2013). In addition to the major grain crops of wheat, barley, maize and sorghum, alfalfa, fruit, and vegetables are also grown. Urban settlements and areas of natural vegetation cover are also to be found.

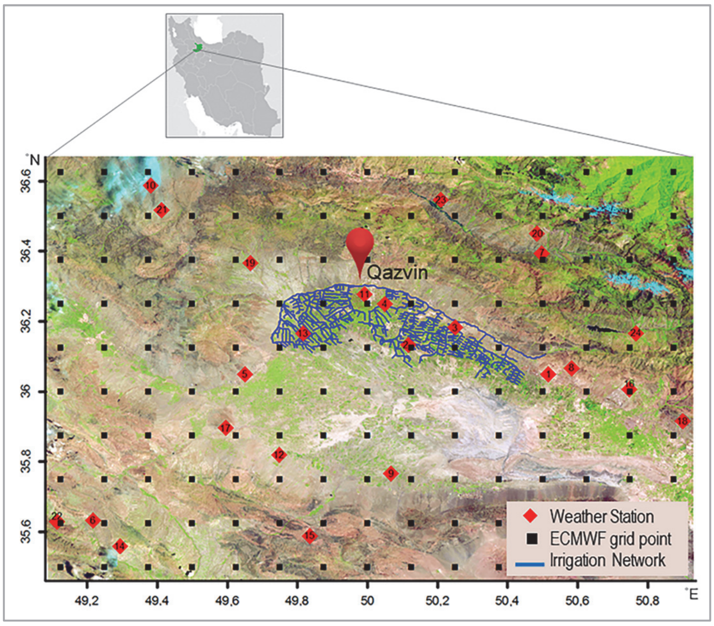


Figure 3.1 The irrigation network in Qazvin Plain, Iran. The area contains 24 weather stations and a sample subset of 10 × 15 grid cells of the ECMWF dataset. The background image has been produced from Landsat 8 RGB data.

The network is participating in a pilot study for the project "Increasing water productivity through demand management and improved operation" (Sharifi, 2013). The objective of this project was to raise water productivity by developing an information system to address problems in water management. The system provides information on crop-water demands based on crop-growth simulation models, weather data and field measurements. The study area extends between 35.44º and 36.68º latitude (N) and 49.09º and 50.92º longitude (E) so as to include as many weather stations as possible (24 stations, see Table 3.1).

Table 3.1 Air temperature is measured at 24 weather stations in the study area.

Station ID	Station name	Stations coordinates		Elevation (m)	Type	Air temperature measurements
		Latitude	Longitude			
1	Abeyk	36.05	50.52	1278	Climatology type1	6-hourly
2	Magsal	36.13	50.12	1205	Climatology type1	6-hourly
3	Nirougah	36.18	50.25	1299	Climatology type1	6-hourly
4	Qazvin	36.25	50.05	1280	Synoptic	3-hourly
5	Takestan	36.05	49.65	1326	Synoptic	3-hourly
6	Avaj	35.63	49.22	1888	Climatology type1	6-hourly
7	Baghkelaye	36.39	50.50	1256	Climatology type2	min. & max.
8	Baghkosar	36.07	50.58	1541	Climatology type2	min. & max.
9	Bouinzahra	35.77	50.07	1213	Synoptic	3-hourly
10	Bourmanak	36.59	49.38	578	Climatology type2	min. & max.
11	Camp	36.28	49.99	1311	Climatology type2	min. & max.
12	Danesfahan	35.82	49.75	1303	Climatology type2	min. & max.
13	Dolatabad	36.17	49.82	1249	Climatology type2	min. & max.
14	Estalaj	35.56	49.29	2340	Climatology type2	min. & max.
15	Hajiarab	35.59	49.84	1707	Climatology type2	min. & max.
16	Hashtgerd	36.01	50.75	1601	Synoptic	3-hourly
17	Jahanabad	35.90	49.60	1372	Climatology type2	min. & max.
18	Karaj	35.92	50.90	1657	Synoptic	3-hourly
19	Kouhin	36.37	49.67	1498	Climatology type2	min. & max.
20	Moalem	36.45	50.48	1569	Synoptic	3-hourly
21	Niarak	36.52	49.41	1184	Climatology type2	min. & max.
22	Qouzlo	35.63	49.11	2061	Climatology type2	min. & max.
23	Razmiankia	36.55	50.21	1010	Climatology type2	min. & max.
24	Taleghan	36.17	50.77	1827	Synoptic	3-hourly

Depending upon the instrument used to measure air temperature and the temporal frequency of measurement, weather stations were categorized as one of three types: synoptic and climatology type1 stations measure air temperature by thermometer; climatology type2 stations use a thermograph. The synoptic stations are supposed to be able to provide more precise measurements. The number of measured values can vary among weather stations, caused by differences in the number of missing values at each station.

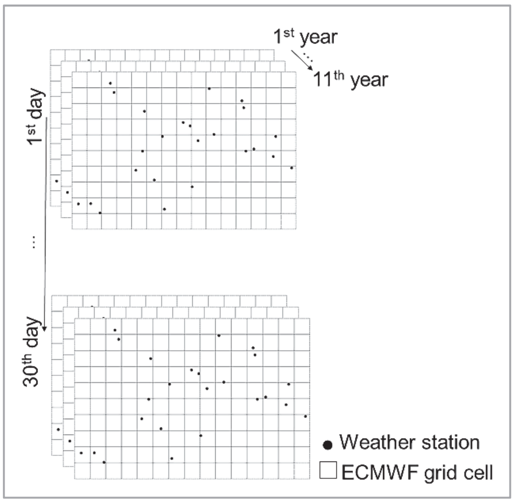


Figure 3.2 The data frame. Daily air temperatures in June are available for 24 weather stations and 150 grid cells of ECMWF over a period of 11 years.

The reanalysis air temperatures were retrieved for the 150 grid cells from the ERA-Interim data assimilation system provided by the European Centre for Medium-Range Weather Forecasts (ECMWF) (Persson, 2013). The ECMWF forecasting system consists of several components like general circulation models, an ocean wave model, a land surface model, a data assimilation and forecast ensembles (Persson 2013). Reanalysis data are a multivariate, spatially complete record of the global atmospheric circulation (Dee et al. 2011). ERA-Interim is the most widely used source of global atmospheric reanalysis (Dee et al. 2011). The data are available at 3-hourly intervals and can be retrieved for a 0.125º Lat/Long grid, corresponding to a spatial resolution of 13.5 km in the meridional direction (Persson 2013). Each station is assigned to its nearest grid cell for comparison of reanalysis values with measured values.

Daily mean air temperature was calculated by averaging the minimum and maximum temperatures at each station in June from 2004 to 2014. The measurements at the stations are assigned to the reanalysis values at the nearest grid cells. There are 150 grid cells × 11 years = 1650 reanalysis air temperatures and 24 stations × 11 years = 264 measurements at each day of June (Figure 3.2). Temperatures in June are used because of the importance of this month in the cropping calendar of the irrigation network: it is the end of the season for winter crops and the beginning for summer crops, especially maize.

My study describes the dependencies between air temperature and non-climatic variables, i.e. its covariates. For instance, in my research I investigated whether considering leaf area index (LAI), land surface temperature (LST) and surface elevation improves the results of the copula-based methods. The covariates were obtained with remote sensing retrieval techniques. Landsat 8 provides several images in panchromatic, optical and thermal bands at a spatial resolution of 30m and temporal resolution of 16 days (Zanter, 2016). Two days in June 2014 were selected as these were Landsat 8 overpass days (Figure 3.3). In the case of LST, I followed the method introduced by Jiménez-Muñoz et al. (2014) and for LAI that of Allen et al. (1998).

The NASA Land Processes Distributed Active Archive Centre (LPDAAC) provides Moderate Resolution Imaging Spectroradiometer (MODIS) products. The MOD03 product provides per-pixel digital-elevation model values in a sequence of swath-based products at 5-minute increments. This gives elevations at a spatial resolution of 1km. Also, surface elevation was obtained from the SRTM dataset at a spatial resolution of 90m (Jarvis et al., 2008). The study area is a relatively homogenous in terms land cover and topography, the main exception being mountainous terrain in the northeastern part of the study area (Figure 3.3).

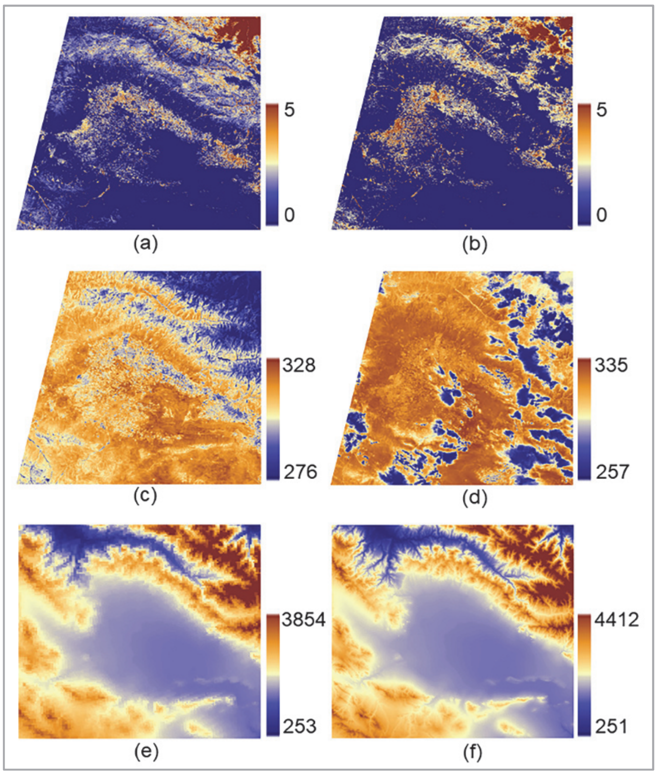


Figure 3.3 Three covariates for air temperature. a) LAI on 6 June 2014, b) LAI on 22 June 2014, c) LST in ºK on 6 June 2014, d) LST in ºK on 22 June 2014, e) MODIS surface elevation in meters, f) SRTM surface elevation in meters. LAI and LST are obtained from Landsat 8 bands at a spatial resolution of 30m. Surface elevations are obtained from the MODIS and SRTM datasets at spatial resolutions of 1km and 90m, respectively. Low values of LST on 22 June indicate a greater degree of cloud covers.

3.2 The Netherlands

Potato is a valuable crop in the Netherlands. Its growing season typically starts in April and ends in September (Figure 3.4) (Beukema and van der Zaag 1990). Potato farms occupy about one-quarter of the country's arable land area and account for approximately half the total production from arable cropping (Figure 3.5) (Beukema and van der Zaag 1990). Figure 3.6 shows the largest change in the consumer price of goods and services in the Netherlands from 2001 to 2018 (CBS 2018). The consumer price of potatoes shows the largest changes in nine years between 2001 and 2018 (CBS 2018).

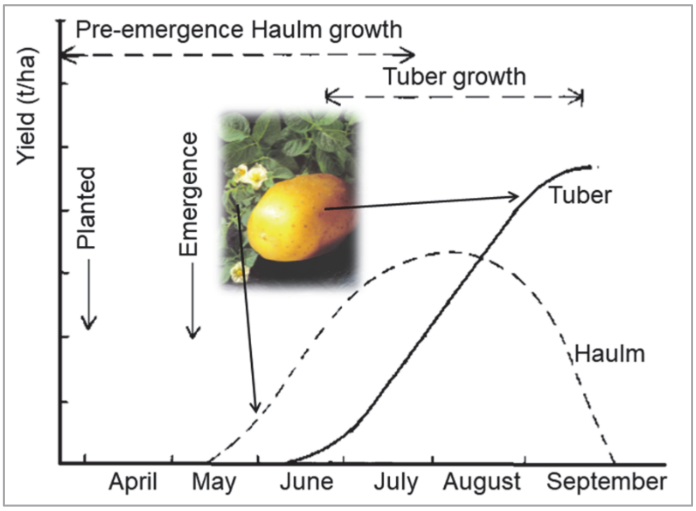


Figure 3.4 The potato growing season in the Netherlands (Beukema and van der Zaag 1990).

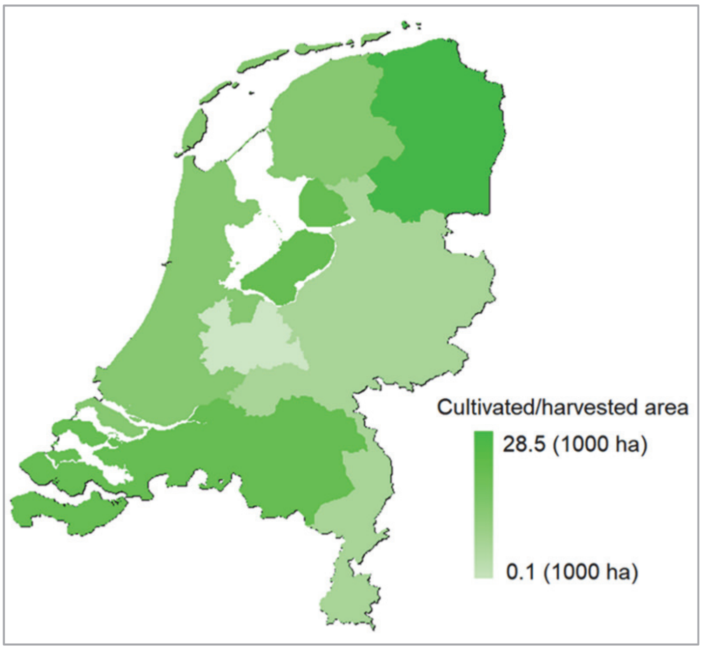


Figure 3.5 Potatoes cultivated/harvested areas in the Netherlands in 2017.

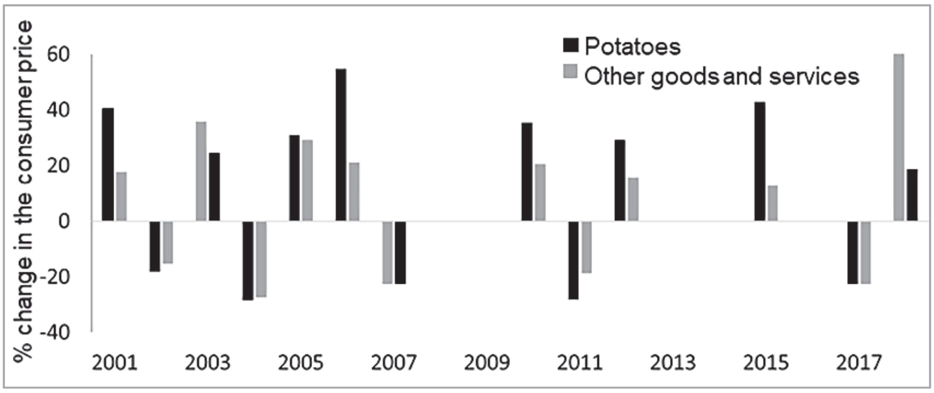


Figure 3.6 The largest change in the consumer price in the Netherlands from 2001 to 2018.

Annual absolute selling prices (€ per 100 kg of potatoes, including seed potatoes), annual harvested production (in 1000 tonnes), yield (in tonnes ha^{-1}), the harvested and cultivated area per 1000 ha as shown in Figure 3.7, were retrieved from the archive of the Central Bureau for Statistics (CBS) in the Netherlands and the statistics database of the European Union (Eurostat 2018) for the period 1980-2017.

Absolute selling prices are prices at which the producer sells to the wholesale trade and are based upon the prices of main agricultural outputs and inputs. These prices indicate direct flows of money into farmers' income and, therefore, were used for analyses of agricultural income (Eurostat 2018). Harvested production is the weight of potatoes that have been harvested and transported away from the field. Yield is the weight of potatoes produced per unit area under cultivation (Eurostat 2018).

Hourly air temperature and precipitation data from 50 automated weather stations in the Netherlands for the period 1980-2017 is available from the data centre (KDC) of the Royal Netherlands Meteorological Institute (KNMI 2018). In the potato growing season the number of measurements may differ between weather station (Figure 3.8). For my study I chose 33 stations for which both rainfall and temperature measurements were available (Figure 3.8).

Gridded reanalysis weather data at a 0.125º Lat/Long resolution is available from the ERA-interim Archive maintained by the European Centre for Medium-range Weather Forecasts (ECMWF) (Persson 2013). The ERA-Interim archive is the most widely used source of global atmospheric reanalysis data (Dee et al. 2011). For my study I selected 33 grid points from the ECMWF data nearest to the chosen KNMI stations. Daily minimum and maximum air temperatures were identified from the minimum and maximum values of the hourly data, and daily precipitation was calculated as the sum of the hourly precipitation data.

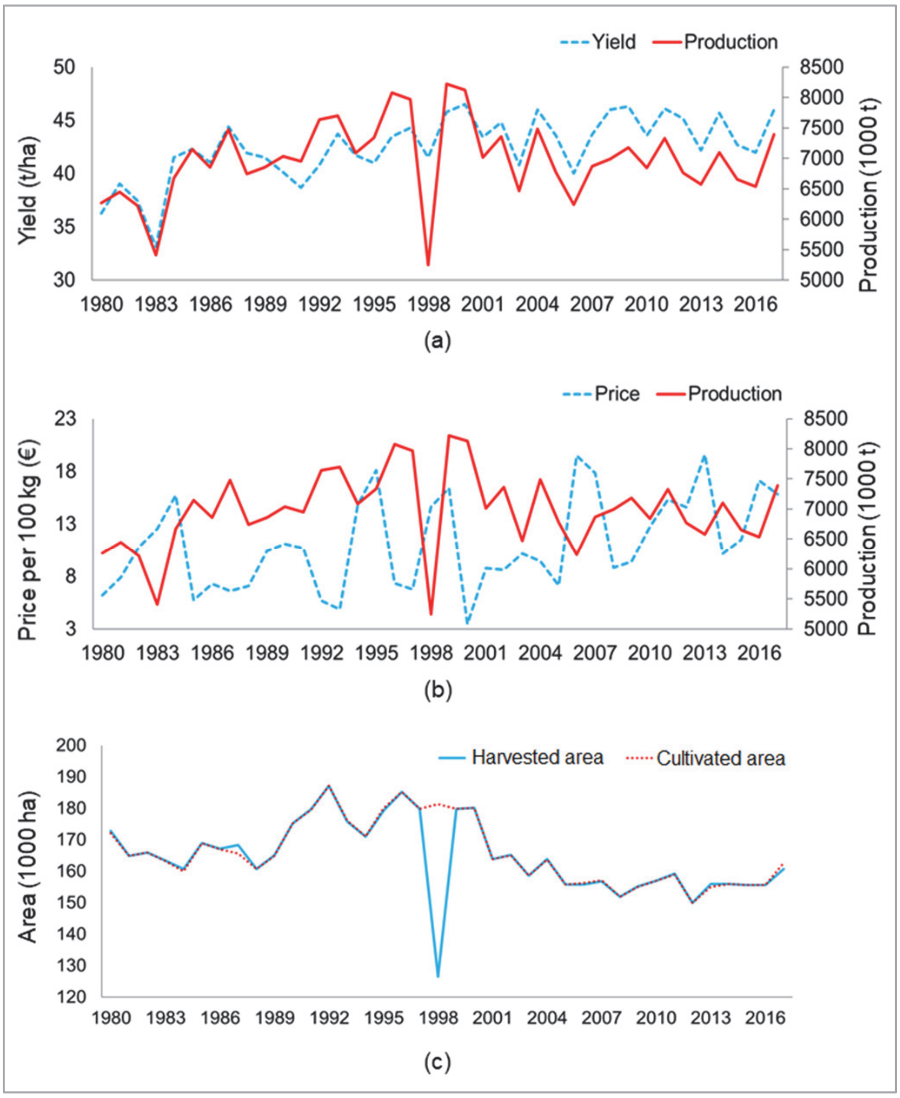


Figure 3.7 Temporal trends in the non-climatic variable: a) yield and production, b) price and production, c) cultivated and harvested areas of potatoes.

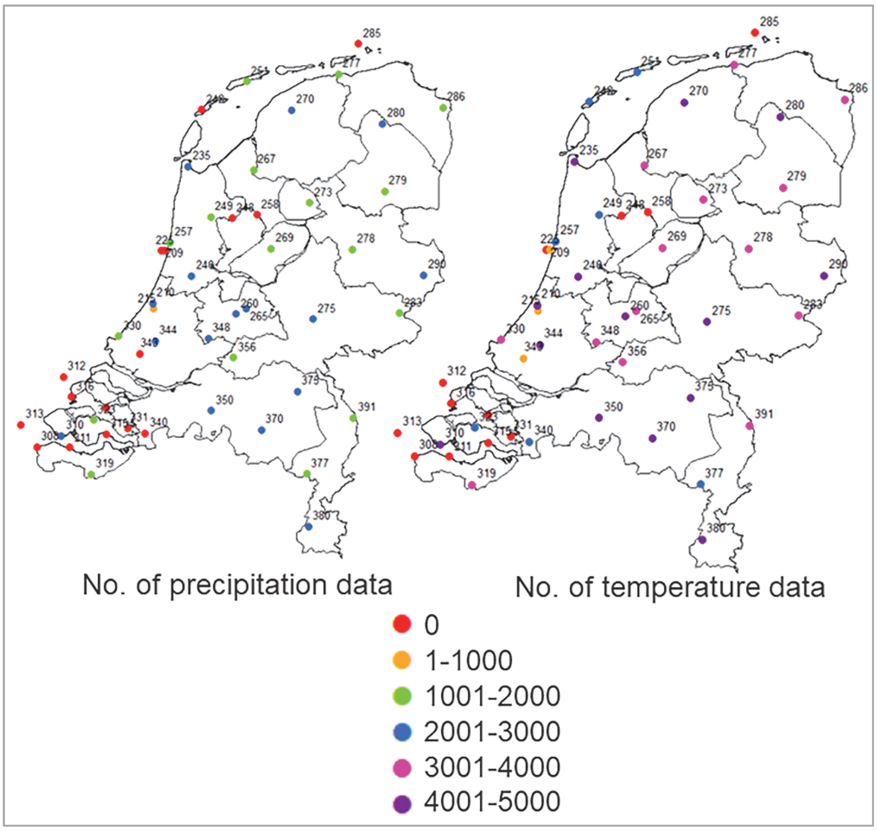
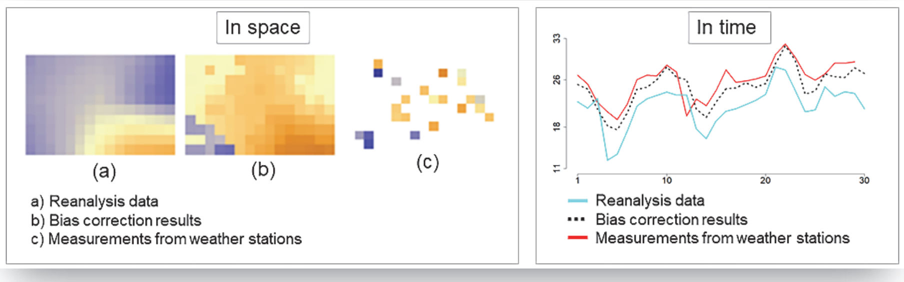


Figure 3.8 Number of daily measurements during the potato growing season at 50 automated KNMI weather stations. Colored dots indicate the range of number of measurements; the number alongside each dot is the station ID.

Chapter 4: The use of bivariate copulas for bias correction



This chapter is submitted as: Alidoost F., Stein A., Su Z. The use of bivariate copulas for bias correction of reanalysis air temperature data. *PLOS ONE*.

Abstract

Air temperature data retrieved from global atmospheric models may show a systematic bias with respect to measurements from weather stations. This is a common concern in local climate studies. The current study presents two methods based upon copulas and Conditional Probability (CP) to predict bias-corrected mean air temperature in a data-scarce environment: CP-I offers a single conditional probability as a predictor, CP-II is a pixel-wise version of CP-I and offers spatially varying predictors. The methods were applied on daily air temperature in the Qazvin Plain, Iran that were collected at 24 weather stations and 150 ECMWF ERA-interim grid cells from a single month (June) over 11 years. We compared the methods with the commonly applied conditional expectation and conditional median methods. Leave-*k*-out cross-validation and correlation scores show that the new methods reduced the bias with 44 – 68% for the full data set and with 34 – 74% on a homogeneous subarea. We conclude that the two methods are able to locally improve ECMWF air temperatures in a data-scarce area.

Keywords

Bias, copula, conditional, data scarcity, mean temperature, probability.

Author contributions

F.A. conceived and designed the analysis, collected and processed the data, developed tools, performed the analysis, wrote the manuscript, is the corresponding author.

A.S. supervised the findings of this work, verified the analytical methods, encouraged A.F. to investigate copulas, improved the English wording.

Z.S. supervised the findings of this work, encouraged A.F. to investigate bias correction of ECMWF data.

All authors contributed to the interpretation of the results, and commented on the final manuscript.

Structure of the chapter

After the introduction in section 4.1, copula-based bias correction methods are introduced in section 4.2. Our application is introduced in section 4.3, and the results are shown in section 4.4. We discuss the results and point to further directions of this work in section 4.5. This is followed by the conclusion in section 4.6, and three appendices in sections Appendix 4.1, 4.2, and 4.3.

4.1 Introduction

Assessment of the impact of climate change in agricultural areas is primarily based upon changes in weather data such as air temperature (Challinor et al. 2009). In a data-scarce area, e.g., where weather stations are sparse, additional data are required. The European Centre for Medium-range Weather Forecasts (ECMWF) provides gridded ERA-interim reanalysis weather data that are being used increasingly (Persson 2013). They are prone to uncertainty because of the coarse resolution of models (Durai and Bhradwaj 2014) and the variability of model parameters in space and time (Dee et al. 2011). When compared with the measurements from weather stations, their bias is often considerable (Hannah and Valdes 2001), in particular, if those measurements serve as benchmarks from which any measurement errors are ignored.

Recently, copula-based methods have been developed for deriving bias corrected weather data at unvisited locations (Laux et al. 2011; Vogl et al. 2012; Mao et al. 2015). A copula is a joint distribution function, describing the dependence structure between two or more variables (Sklar 1973). A good description of copula has been provided by (Nelsen 2006). The joint distribution function can be estimated using any distribution family that can be different from the marginal distribution family of the involved variables (Nelsen 2006). Mao et al. (2015) investigated bias correction methods of daily precipitation data and showed that a copula-based bias correction performs better than quantile mapping. After estimating the joint distribution, several methods can be used to obtain bias corrected values at unvisited locations. Examples are the conditional expectation (CE) (Bárdossy and Li 2008), the conditional median (CM) (Gräler 2014), and the simulation method (Salvadori et al. 2007; Nelsen 2006).

So far, Copula-based methods have been applied mainly to precipitation time-series, where bias corrected values are obtained using the simulation method (Laux et al. 2011; Vogl et al. 2012; Mao et al. 2015). Little attention, however, has been given to bias correction in air temperature data in a data-scarce area. Our main focus of bias correction is based upon the construction of the dependence structure between measurements and ECMWF reanalysis data using a joint distribution. The distribution is initially estimated using copulas and is then used to reduce bias of ECMWF air temperatures at grid cells that are often lacking a measurement from a weather station in a data-scarce area. To reduce bias in ECMWF air temperatures at those grid cells, an important aspect is the spatial variation of the data.

This study aims to introduce two copula-based predictors based upon Conditional Probabilities (CP) taking care of the spatial variation of daily air temperatures in a data-scarce area. We evaluate the performance of the predictors comparing to conventional methods like CE and CM in an agricultural area in Iran.

4.2 Bias correction methods

The structural, one-sided difference between a measured value from a weather station x, and an ECMWF reanalysis value y is defined as the bias in ECMWF reanalysis values. We assume that the data are observed from two spatio-temporal random variables X and Y. In our study, the basis of the copula-based bias correction is a distribution function that allows for modeling the dependence structure between X and Y. The purpose of bias correction is to predict $\hat{x}_0$ where $\widehat{}$ denotes a predicted value and $_0$ indicates an unvisited location. An unvisited location is an ECMWF grid point without a measurement.

We focus on a bivariate distribution $F(x, y)$; it can be extended to higher dimensions if more than two variables are available. The bivariate case is useful if ancillary information is unavailable. Regarding our main objective, we aim to introduce copula-based predictors to obtain $\hat{x}_0$. Section 4.2.1 first illustrates both the commonly applied predictors and introduces the new predictors and section 4.2.2 presents the estimation of marginals and copulas.

4.2.1 Copula-based predictors

The conditional expectation (CE), the conditional median (CM) and the simulation method are commonly applied methods to obtain $\hat{x}_0$. CE and CM are both optimal predictors, minimizing the mean squared prediction error and the mean absolute prediction error, respectively (Journel 1984; Cressie 1993). They obtain the bias-corrected value $\hat{x}_0$ as:

$$\text{CE:} \quad \hat{x}_0 = E[X|Y = y_0] = \int_x x \cdot f(x|y_0)dx\,, \tag{4.1}$$

$$\text{CM:} \quad \hat{x}_0 = F^{-1}(p|y_0), \;\; p = 0.5, \tag{4.2}$$

where $f(.|.)$ is conditional density distribution function, F^{-1} denotes the inverse transformation of the conditional distribution $F(.|.)$, and p is the conditional probability that determines the median. Both CE and CM are either an increasing or a decreasing function of the conditioning variable Y depending upon the sign of the dependence between X and Y (see Section 2.3). Therefore, the variation of bias-corrected values follows the variation of ECMWF reanalysis values rather than those of the measurements; this will be further illustrated in Section 4.4.

The third method is the simulation method. It obtains m bias-corrected values by generating m conditional probabilities p on [0, 1] as:

$$\hat{x}_{0,k} = F^{-1}(p_k|y_0), k = 1, \dots, m. \tag{4.3}$$

Note that the mean of $\{\hat{x}_{0,1}, \dots, \hat{x}_{0,m}\}$ provides a single value $\hat{x}_0$ and that both the value of m and the simulations themselves influence the results. For a large m, the results of this method are equal to the results of CE (Mao et al. 2015). In case of choosing the median of $\{\hat{x}_{0,1}, \dots, \hat{x}_{0,m}\}$, this also applies to CM.

For CE, the mean value of the distribution $F(x|y_0)$ is selected as $\hat{x}_0$, whereas for CM, this is the median value of the distribution. We may question whether mean and median values best suit bias-corrected air temperatures. In the following, two new methods are introduced to obtain a conditional probability which serves as a predictor.

CP-I and CP-II are the predictors, minimizing mean absolute bias (MAB) as:

$$MAB = \frac{1}{n}\sum_{i=1}^{i=n} |x_i - F^{-1}(p|y_i)|\,, \tag{4.4}$$

where for CP-I, $n = N$ and equals the total number of observations, whereas for CP-II, $n = M \ll N$ and equals the number of observations at the nearest M locations to x_0. The conditional probability p is iteratively estimated based upon minimizing MAB in (4.4) resulting in the optimal p^* value. The bias-corrected value $\hat{x}_0$ then equals:

$$\hat{x}_0 = F^{-1}(p|y_0), p = p^*. \tag{4.5}$$

For CP-I, the conditional probability p^* is constant for all unvisited locations, e.g. $F(x_0|y_0) = p^*$. Therefore, similar to CE and CM, CP-I is either an increasing or a decreasing function of the conditioning variable, depending upon the sign of the dependence (see Section 2.3). For CP-II, the optimal conditional probability depends upon unvisited location and is denoted now by p_0^*, e.g. $F(x_0|y_0) = p_0^*$.

Next we formulate the equations using copulas and investigate the use of copulas for the construction of distribution functions. According to Sklar's theorem, it can be shown that $F(x|y) = C(u|v)$ (see Chapter 2) and the predictors are rewritten as:

CE: $\hat{x}_0 = \int_0^1 F_X^{-1}(u) \times c(u|V = v_0)du,$

CM: $\hat{x}_0 = F_X^{-1}(C^{-1}(p|V = v_0)), p = 0.5,$

CP: $MAB = \frac{1}{n}\sum_{i=1}^{i=n}|x_i - F_X^{-1}(C^{-1}(p|V = v_i))|$, $\hat{x}_0 = F_X^{-1}(C^{-1}(p|V = v_0)), p = p^*.$

where F_X^{-1} denotes the inverse transformation of the marginal cumulative distribution function F_X, v is marginal probability i.e. $v = F_Y(y)$, $c(.|.)$ is the conditional density copula, and $C(.|.)$ is the conditional cumulative copula (see Chapter 2).

Before introducing estimation of the distribution functions, we now explain the implementation of CP-I and CP-II to identify the optimal conditional probability. Initially, a probability $p = 0.01$ is chosen and MAB is obtained from Equation (4.4). Then the probability p increases with steps of 0.01 until $p = 1$. We select the probability p^* that results into the lowest MAB. Finally, the bias-corrected value $\hat{x}_{s_0}$ is obtained from Equation (4.5). The choice for the initial probability and for a step value equal to 0.01 are based upon our experience on the variable of interest and uncertainty sources. We compare this value using a sensitivity analysis on the mean absolute prediction error to assess the effect of choosing larger or smaller increment values i.e. 0.1 or 0.001; results are reported in Section 4.4. Note that CP-I is implemented only once, whereas CP-II is implemented at each unvisited location separately and therefore has a higher computational cost.

4.2.2 Distributions estimation

In practice, finite samples on X and Y are observed in space and time without replication. Therefore, the joint distribution $F(x,y)$ is estimated using the assumption of stationarity (in space or time), i.e. marginal distributions and dependence structure between X and Y are irrespective of location or time. In the literature, reviewed in Section 4.1, the current bias correction methods have been applied to climate time-series assuming temporal stationarity. Hence, removing autocorrelation and heteroscedasticity that may exist in any climate time-series, is necessary for any estimation procedure (Laux et al. 2011). To achieve our main objective, we apply a bias correction to predict $\hat{x}_0$ at an unvisited location in space, separately at each day of time-series.

Estimation of theoretical marginal distributions may affect the estimation of the copula parameter and consequently the selection of the copula family. Therefore, we use empirical marginal distributions. By means of kernel density estimation, a continuous approximation of the marginal distribution are obtained under the assumption of stationary (Silverman 1986). We evaluate this assumption using regression analysis and the auto-correlation function (See appendix 4.1). The choice of the method to estimate empirical marginal probability is not unique and a more specific sensitivity analysis might help to show the effects of other marginal distribution functions on the results. This, however, is outside the scope of the study.

The bivariate copula *C* can be determined using several copula families. We assume spatial stationarity and evaluate the assumption using a co-correlation function (See appendix 4.1).

4.2.3 Evaluation

We apply the leave-*k*-out validation (Lafon et al. 2013). The bias-corrected values $\hat{x}_{s,t}$ at time t and location s are obtained by leaving k observations out for the same day of the year in k successive years and using the reminder of the observations. The mean absolute error $MAE_{s,t}$ is defined as:

$$MAE_{s,t} = \frac{1}{k}\sum_{i=1}^{k}\left|x_{s,t,i} - \hat{x}_{s,t,i}\right|, \tag{4.6}$$

We define three criteria based upon the mean absolute errors to compare the presented methods at N weather stations and T days:

$$MAE = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{N}\sum_{s=1}^{N} MAE_{s,t}\right), \tag{4.7}$$

$$SES = \sum_{s=1}^{N}\left(rank\left(\frac{1}{T}\sum_{t=1}^{T} MAE_{s,t}\right)\right), \tag{4.8}$$

$$TES = \sum_{t=1}^{T}\left(rank\left(\frac{1}{N}\sum_{s=1}^{N} MAE_{s,t}\right)\right). \tag{4.9}$$

where the MAE is the overall mean absolute error, SES and TES are spatial and temporal error scores (Durai and Bhradwaj 2014), $\frac{1}{T}\sum_{t=1}^{T} MAE_{s,t}$ and $\frac{1}{N}\sum_{s=1}^{N} MAE_{s,t}$ are spatial and temporal mean absolute errors, respectively. A low value of a criterion indicates a good performance.

To evaluate correlations, the bias-corrected value $\hat{x}_{s,t}$ at day t and location s is obtained using all observations. The temporal correlations r_s at location s and the spatial correlations r_t at day t are used to evaluate the performance of the presented methods in reproducing temporal and spatial variation of the measurements:

$$r_s = corr\left(\{\hat{x}_{s,t}, \dots, \hat{x}_{s,T}\}, \{x_{s,t}, \dots, x_{s,T}\}\right), \tag{4.10}$$

$$r_t = corr\left(\{\hat{x}_{1,t}, \dots, \hat{x}_{N,t}\}, \{x_{1,t}, \dots, x_{N,t}\}\right). \tag{4.11}$$

We define two criteria to evaluate the correlations as:

$$SCS = \sum_{s=1}^{N}\left(rank(r_s)\right), \tag{4.12}$$

$$TCS = \sum_{t=1}^{T}\left(rank(r_t)\right), \tag{4.13}$$

where SCS and TCS are spatial and temporal correlation scores, respectively. A high value of SCS and TCS indicates a good performance.

4.3 Application: daily mean air temperatures in Iran

The presented methods are applied to daily mean air temperatures in the Qazvin irrigation network, Iran in June from 2004 to 2014 (see Section 3.1). There are 150 grid cells×11 years=1650 reanalysis values on each day in June 2014, and there are 24 stations×11 years=264 measured values. The cross-validation is carried out for measured values on each day of June between 2004 and 2014, i.e., k=11.

The time-series of the air temperature at the climatology type2 stations, e.g., stations 11, 13 and 21 reveals that the quality of the measurements is low (Appendix 4.3, Figure 4.14). In Section 4.4, we report to which degree the results of the presented methods are affected by different qualities of the measurements at the three types of stations. Overestimation and underestimation of reanalysis data has been observed in June 2014 (Appendix 4.3, Figure 4.14). Correlations r_t between reanalysis values and measured values in space are low at most days in June 2014 (Figure 4.1). In addition, correlations r_s at the weather stations 13 and 21 are rather weak (Figure 4.1).

To extend copulas to higher dimensions by including covariates in describing the dependence structures, we investigate whether considering elevation improves the results of the bias correction method. The dependence structure between air temperature and MODIS elevation (see Section 3.1) is described using copulas as it does not follow the lapse-rate law (Figure 4.2).

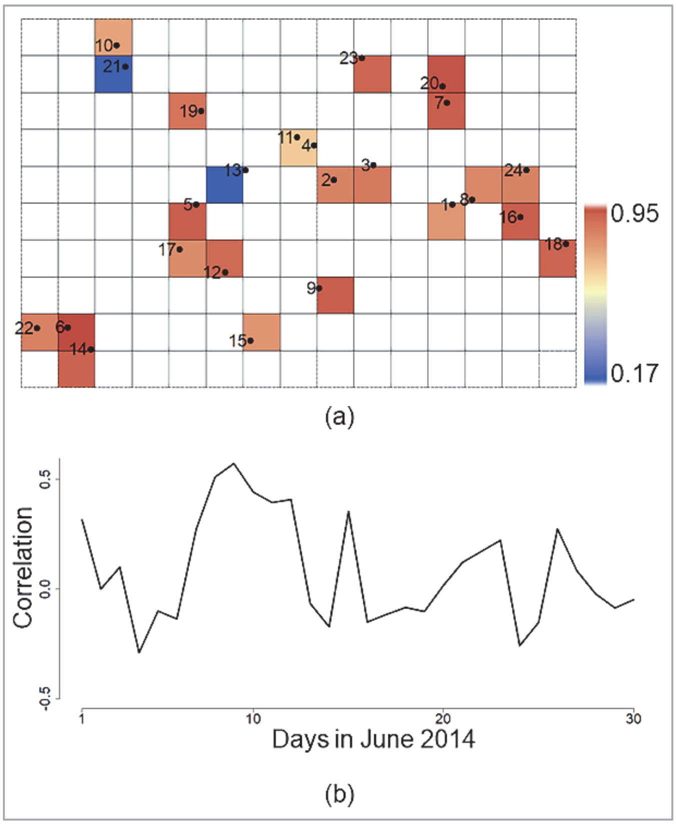


Figure 4.1 Correlations r_i and r_j that indicate temporal and spatial dependences between measurements and ECMWF ERA-interim reanalysis air temperature. a) r_i at each weather station, b) r_j at each day in June 2014.

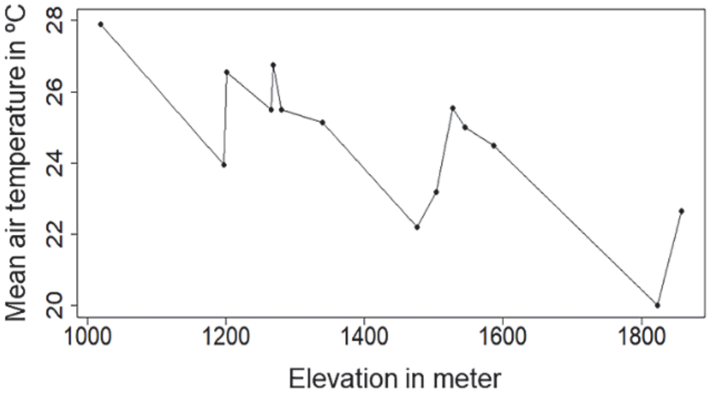


Figure 4.2 Variation of the mean air temperature on the 1st day of June 2014 compared with a variation of the elevation in the study area. The mean air temperature in °C is derived from the synoptic and climatology type 1 weather stations.

In the pooling procedure, effects of non-stationarity may exist due to climate change. For this time-series of 11 years, we ignore those effects, but for a longer time-series a correction should be applied. In our study, the dependence structures between the reanalysis values and measured values, i.e., copulas are studied in a relatively small and homogenous area and are thus likely to change spatially in a stationary way. An exception concerns the mountains in the northeastern part of the study area (Figure 4.3). To evaluate the potential effect of spatial non-stationarity, we applied the presented methods on a complete set of 24 weather stations as well as a subset of ten stations where the spatial variation of elevation is more homogenous (Figure 4.3).

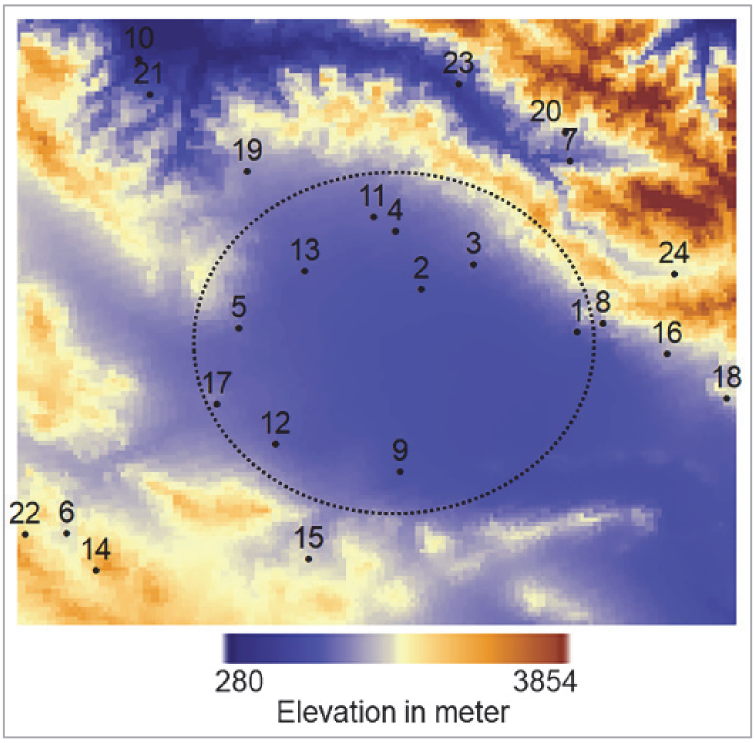


Figure 4.3 Elevations (m) are covariates for air temperature in the CP-II including covariate. It is obtained by MODIS product at a spatial resolution of 1km. Location and index of the weather stations are shown in this figure. We applied the presented methods on a complete set of 24 weather stations as well as a subset of ten stations where the spatial variation of elevation is more homogenous, i.e., the area indicated by a circle.

4.4 Results: bias-corrected values in time and space

4.4.1 Marginal distributions and copulas

Figure 4.4 shows the fit of marginal distribution functions assuming spatial stationarity. Appendix 4.1 presents the evaluation of this assumption on each day in June 2014.

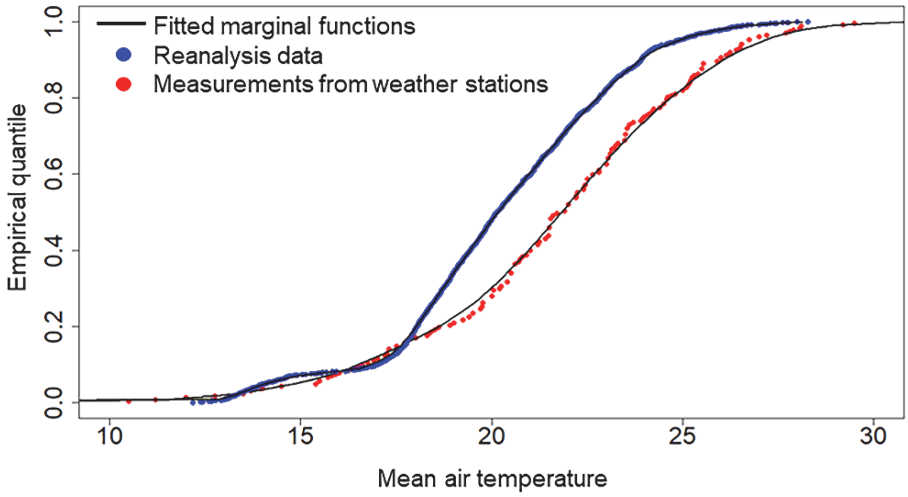


Figure 4.4 Empirical marginal probabilities on June 1st. Marginal probabilities are obtained on each day of June using eleven years series from 2004 to 2014 at 24 weather stations. A monotone cubic spline is fitted to obtain the distribution function.

The parameters of five copula families are estimated on each day of June assuming spatial stationarity. Appendix 4.1 further contains the evaluation of this assumption for copulas. Table 4.1 shows the number of data used for fitting. The *p* value of the null hypothesis of bivariate independence is zero, thus rejecting the null hypothesis (Table 4.1, third column). The best fitting family based upon the lowest AIC value turned out to be Gumbel family for 17 days in June. The *p* values of the Cramér–von Mises statistic $S_n^{(B)}$ were larger than 0.2 for all days (Table 4.1, last column), hence not rejecting the null hypothesis. We could safely assume that the best fitting family well describes the dependence structure.

Table 4.1 The *p* values and selected family on each day in June. Number of data denotes the number of available data for fitting purposes and equals the number of measurements from weather stations from 2004 to 2014 on each day in June. The *p* value-1 is obtained under the null hypothesis of bivariate independence. The copula families are: N=Gaussian, T=Student's *t*, C=Clayton, G=Gumbel, and F=Frank. The *p* values-2 are obtained by the Cramér–von Mises statistic $\boldsymbol{S_n^{(B)}}$.

Day	Number of data	*p* value-1	Selected family	*p* value-2
1	226	0.00	G	0.42
2	224	0.00	N	0.62
3	226	0.00	G	0.48
4	226	0.00	G	0.58
5	226	0.00	T	1.00
6	226	0.00	F	0.40
7	226	0.00	N	0.44
8	225	0.00	T	1.00
9	226	0.00	G	0.34
10	226	0.00	G	0.26
11	226	0.00	G	0.36

12	226	0.00	N	0.62
13	226	0.00	N	0.44
14	226	0.00	N	0.64
15	226	0.00	G	0.44
16	226	0.00	G	0.52
17	226	0.00	G	0.46
18	226	0.00	F	0.44
19	226	0.00	F	0.25
20	226	0.00	G	0.34
21	226	0.00	G	0.30
22	226	0.00	G	0.79
23	225	0.00	G	0.36
24	226	0.00	G	0.54
25	226	0.00	G	0.75
26	226	0.00	G	0.68
27	226	0.00	N	0.50
28	226	0.00	F	0.44
29	226	0.00	F	0.60
30	225	0.00	G	0.54

4.4.2 Evaluation and comparison

The optimal conditional probability obtained using CP-I, and the minimum and maximum of the optimal conditional probabilities obtained using CP-II on each day are given in Table 4.2. The conditional probability using CP-I clearly changes in time in the range of [0.30, 0.95]. For CP-II, the optimal conditional probability changes in time and space in the range of [0.02, 0.99], using M=4. Influence of the choice of the increment value in CP-I is assessed using sensitivity analysis (Figure 4.5). It revealed that the uncertainty is higher using an increment value of 0.1, whereas for 0.001 no improvements were achieved.

Table 4.2 Optimal conditional probabilities. A single optimal conditional probability is obtained using CP-I for all unvisited locations on each day whereas using CP-II, it is obtained at each unvisited location and each day. The minimum and maximum of the optimal conditional probabilities obtained by CP-II are mentioned here.

Day	Optimal conditional probability in CP-I	Minimum and maximum optimal conditional probabilities in CP-II	
1	0.79	0.13	0.90
2	0.60	0.08	0.97
3	0.30	0.04	0.92
4	0.36	0.08	0.93
5	0.50	0.02	0.90
6	0.61	0.08	0.93
7	0.71	0.12	0.96
8	0.66	0.21	0.92
9	0.64	0.25	0.90
10	0.82	0.23	0.99
11	0.87	0.28	0.98
12	0.68	0.09	0.95
13	0.58	0.06	0.84
14	0.57	0.05	0.88
15	0.65	0.10	0.86

16	0.65	0.09	0.94
17	0.76	0.07	0.84
18	0.55	0.10	0.74
19	0.73	0.07	0.88
20	0.69	0.19	0.91
21	0.50	0.13	0.95
22	0.83	0.19	0.98
23	0.91	0.23	0.99
24	0.64	0.14	0.96
25	0.65	0.09	0.94
26	0.79	0.17	0.92
27	0.74	0.13	0.98
28	0.83	0.10	0.95
29	0.92	0.21	0.98
30	0.79	0.16	0.99

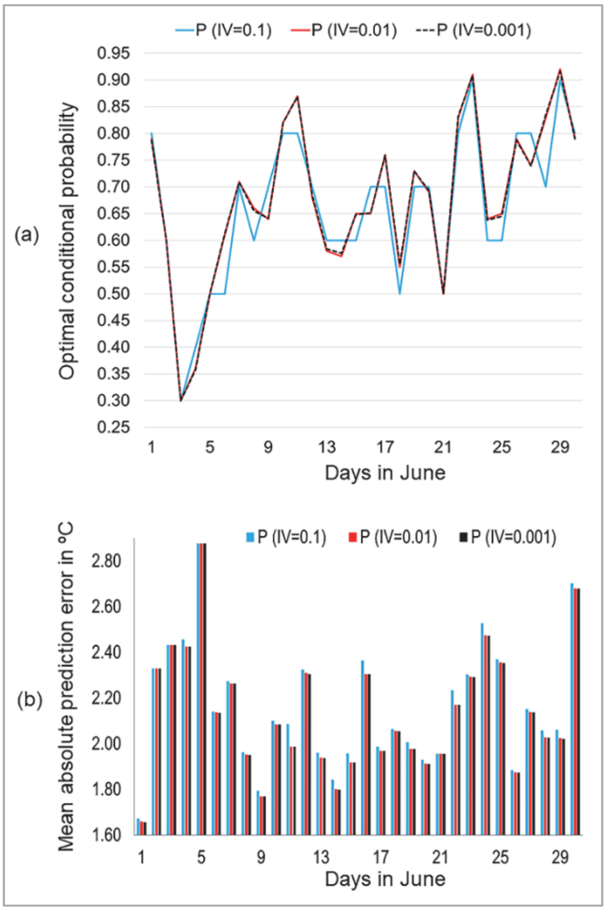


Figure 4.5 Influence of the choice of the increment value (IV) on a) the optimal conditional probability in CP-I and b) the mean absolute prediction errors. Three IVs 0.1, 0.01 and 0.001 are chosen.

Two time-series of the bias-corrected values obtained by CP-I and CP-II (Figure 4.6a and b) at the first station are compared with those of CE and CM (Figure 4.6c and d). The spatial mean absolute errors at this station for CP-II and CP-I were equal to 1.56ºC and 1.66ºC, whereas, for CM and CE, they were equal to 2.72ºC and 2.95ºC, respectively. Bias-corrected values at June 1st 2014 are shown in Figure 4.7. For CP-II and CP-I, the temporal mean absolute errors were equal to 2.17ºC and 2.23ºC at this day, whereas for CM and CE, they were equal to 2.41ºC and 2.49ºC, respectively.

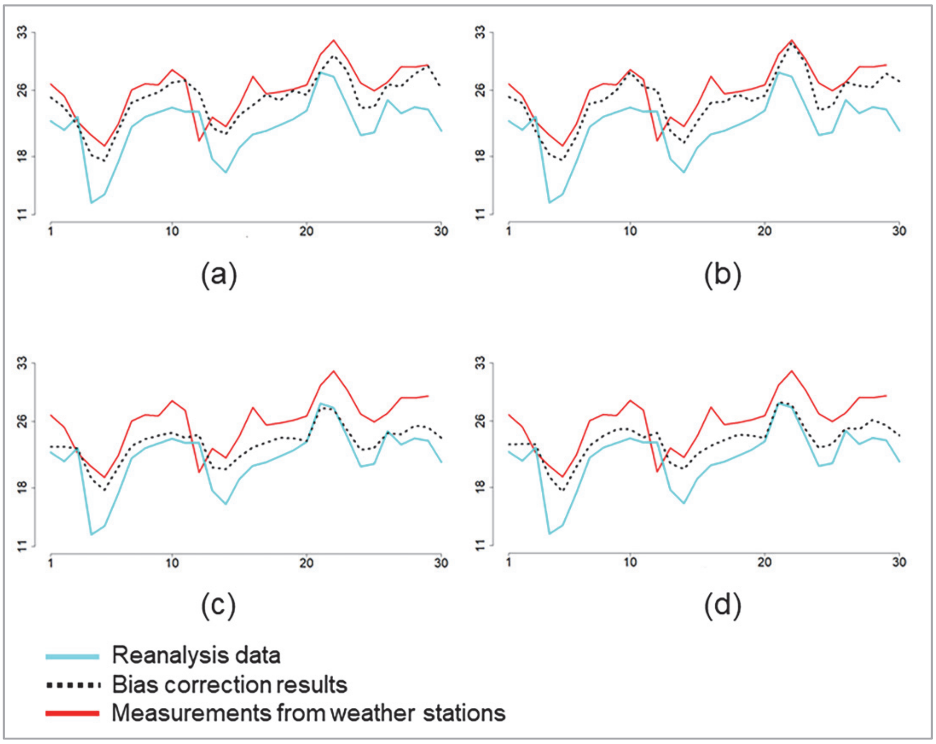


Figure 4.6 Time-series of the mean air temperatures at first station in June 2014 obtained by the measurements, the reanalysis data, and the results of a) CP-I, b) CP-II, c) CE and d) CM. The vertical axis is the daily mean air temperature in ºC. The horizontal axis is days in June 2014.

We note that CP-I fails to predict spatial variation and extremes in space (Figure 4.7c) but that CP-II is successful (Figure 4.7d) as compared to the spatial variation of the measurements at this day (Figure 4.7a). Spatial variation of the bias-corrected values obtained by CP-I (Figure 4.7c), CE (Figure 4.7e) and CM (Figure 4.7f) is similar to the spatial variation of the reanalysis air temperatures (Figure 4.7b). Spatial variation of the bias-corrected values obtained by CP-II differs from spatial variation of the reanalysis air temperatures (Figure 4.7b) because the optimal conditional probability obtained by this method changes in space. Bias and prediction errors at June 1st 2014 are shown in Figure 4.8. The mean absolute bias is 2.84ºC at this day, whereas the mean absolute prediction errors for CP-II and CP-I were equal to 1.13ºC and 1.66ºC, and for CE and CM to 2.46 ºC and 2.31ºC, respectively.

Table 4.3 Comparison of the bias correction methods for two experiments. The methods are applied to 24 weather stations in the first experiment whereas they are applied to a subset of ten stations in the second experiments. Total mean absolute error (MAE), spatial error scores (SES), temporal error scores (TES), spatial correlation scores (SCS), and temporal correlation scores (TCS), obtained by the conditional probabilities (CP-I, CP-II and CP-II including elevation), conditional expectation (CE) and conditional median (CM). The underlined values denote the best method. Only MAE is obtained for CP-II including elevation.

	Method	MAE	SES	TES	SCS	TCS
Results of the 1st experiment	CP-I	2.28	52	59	71	80
	CP-II	2.17	55	34	86	120
	CP-II including elevation	1.92	-	-	-	-
	CE	2.45	71	116	54	49
	CM	2.41	62	91	29	51
Results of the 2nd experiment	CP-I	1.44	27	70	32	80
	CP-II	1.36	19	47	37	102
	CE	1.50	28	92	20	56
	CM	1.50	26	91	11	62

MAE obtained by leave-11-out cross validation for two experiments (Table 4.3) shows that CP-II performed best, followed by CP-I, CM, and CE. The MAE is slightly above 2°C for all methods whereas the average absolute bias is 3.6°C. The horizontal distances, different height, and differences in land cover between the location of a station and the grid cell centre might affect the MAE. Investigating the CP-II including elevation, we noticed a large improvement in the results: the MAE for CP-II including elevation was equal to 1.92ºC whereas for CP-II it was equal to 2.17ºC (Table 4.3).

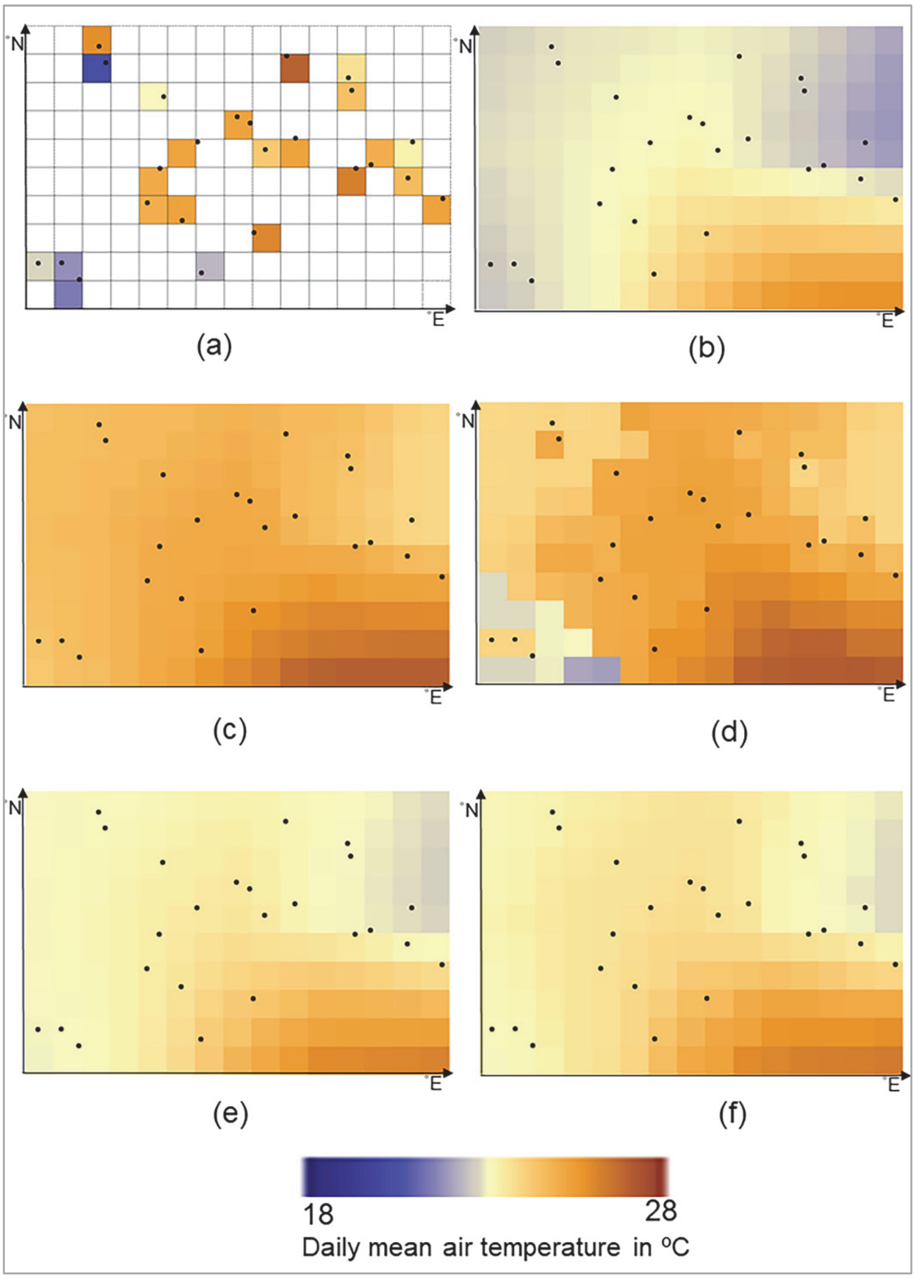


Figure 4.7 The mean air temperatures from a) weather stations, b) reanalysis data, and results of c) CP-I, d) CP-II, e) CE and f) CM, for all locations at June 1st 2014. For experimentation in this study, a sample subset of 10 × 15 grid cells of ECMWF dataset is selected at a spatial resolution of 0.125º Lat/Long. The study area extends from 35.44º to 36.68º latitudes (N) and from 49.09º to 50.92º longitudes (E).

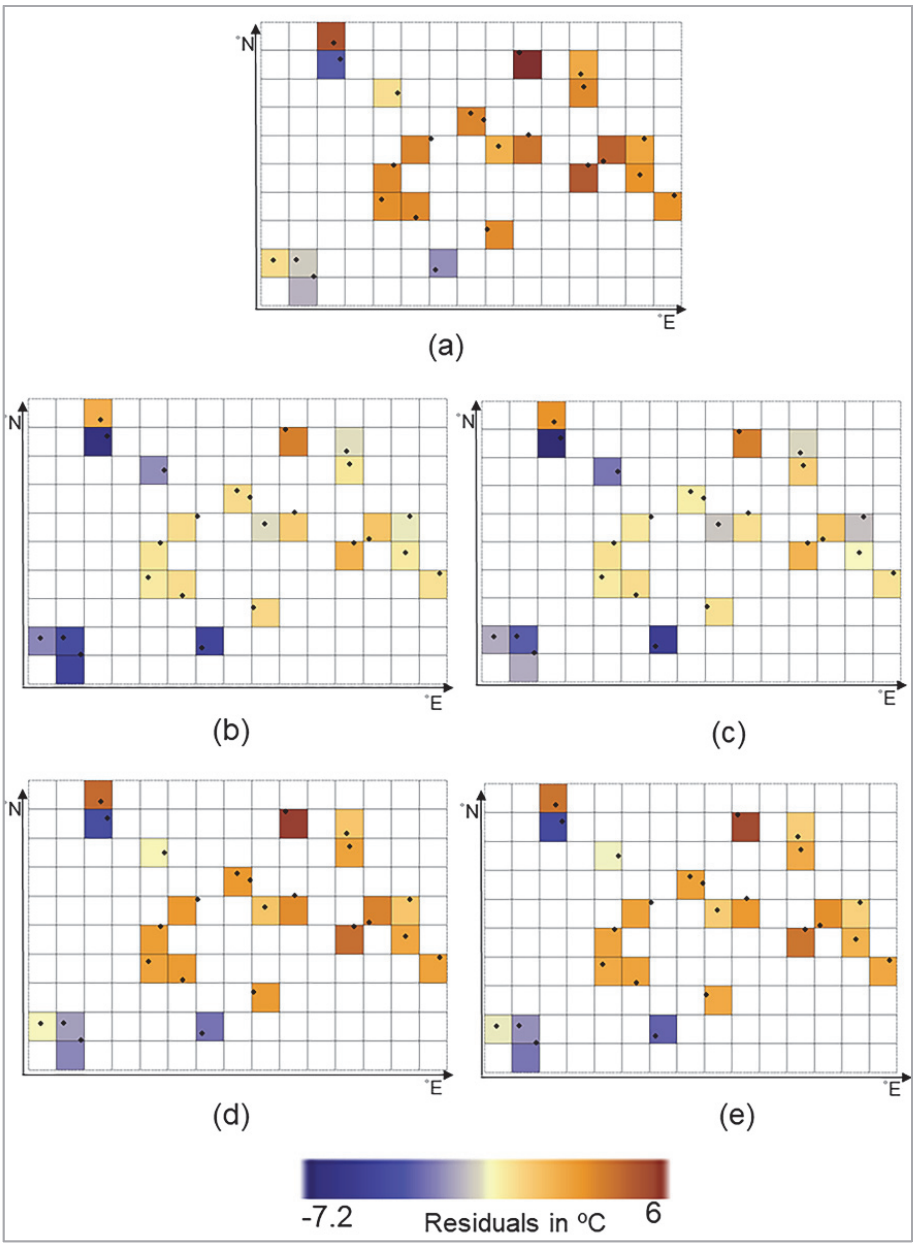


Figure 4.8 Bias (a) and prediction errors. Prediction errors are differences between the mean air temperatures from weather stations and the predictions obtained by b) CP-I, c) CP-II, d) CE and e) CM at June 1st 2014. For experimentation in this study, a sample subset of 10 × 15 grid cells of ECMWF dataset is selected at a spatial resolution of 0.125º Lat/Long. The study area extends from 35.44º to 36.68º latitudes (N) and from 49.09º to 50.92º longitudes (E).

We used SES and SCS to compare the presented methods based upon errors and correlations in time, i.e., 30 days in June (as shown in Appendix 4.2, Figure

4.14). For the comparison in space, TES and TCS were used with N=24 (as shown in Appendix 4.2, Figure 4.15). Table 4.3 shows that CP-I resulted into the lowest errors in time whereas CP-II resulted into the lowest errors in space and highest correlations in space and time. The correlations r_t show that CP-II performed better in reproducing the spatial variation of the daily air temperatures in the study area (Figure 4.9). The correlations r_t obtained by CP-I, CE and CM are similar to the correlations between the reanalysis values and the measured values (Figure 4.9). This is as expected, because the predictor is the same for all locations in space. The correlations r_s denote that CP-I performed better in reproducing the temporal variation of the daily air temperatures in June (Figure 4.10).

Table 4.4 Overall score based upon Table 4.3 for two experiments. The methods are applied on 24 weather stations in the first experiment whereas they are applied on a subset of ten stations in the second experiments. The scores are obtained for each method based upon each criterion, i.e., each column of Table 4.3 where the lowest score denotes the best method. Overall score is the sum of the scores. The underlined values denote the best method.

	Method	Score based on					Overall score
		MAE	SES	TES	SCS	TCS	
Results of the 1st experiment	CP-I	2	1	2	2	2	9
	CP-II	1	2	1	1	1	6
	CE	4	4	4	3	4	19
	CM	3	3	3	4	3	16
Results of the 2nd experiment	CP-I	2	2	2	2	2	10
	CP-II	1	1	1	1	1	5
	CE	4	4	4	3	4	19
	CM	3	3	3	4	3	16

Investigating the differences in quality of the measurements at the weather stations, we compared the spatial mean absolute prediction error (see equation (4.10)) with the spatial mean absolute bias. In this way, we assessed the performance of the bias correction methods at three types of weather stations (Figure 4.11). This investigation showed that the predictions at two synoptic stations, i.e., stations 6 and 19 are influenced by different sources of uncertainties in the measurements derived from three types of weather stations. In addition, CP-II performed better than CE and CM.

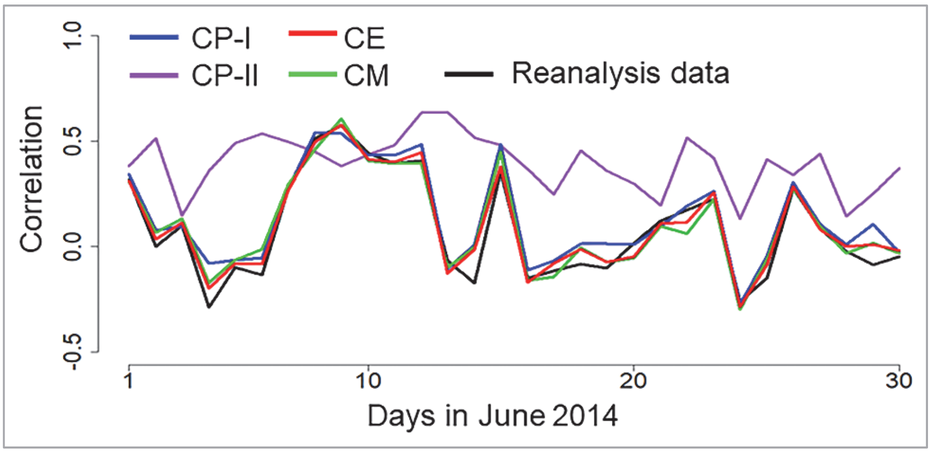


Figure 4.9 The correlation coefficients *r* in space on each day in June 2014.

Table 4.4 shows the score of each method based upon the criteria mentioned in Table 4.3. We obtained an overall score using the sum of the scores. This overall score shows that CP-II reduced the bias with 63 – 68% for the full data set and with 69 – 74% on a homogeneous subarea whereas CP-I decreased the bias with 44 – 53% for the complete data set and with 34 – 47% on a homogeneous subarea (Table 4.4 , last column).

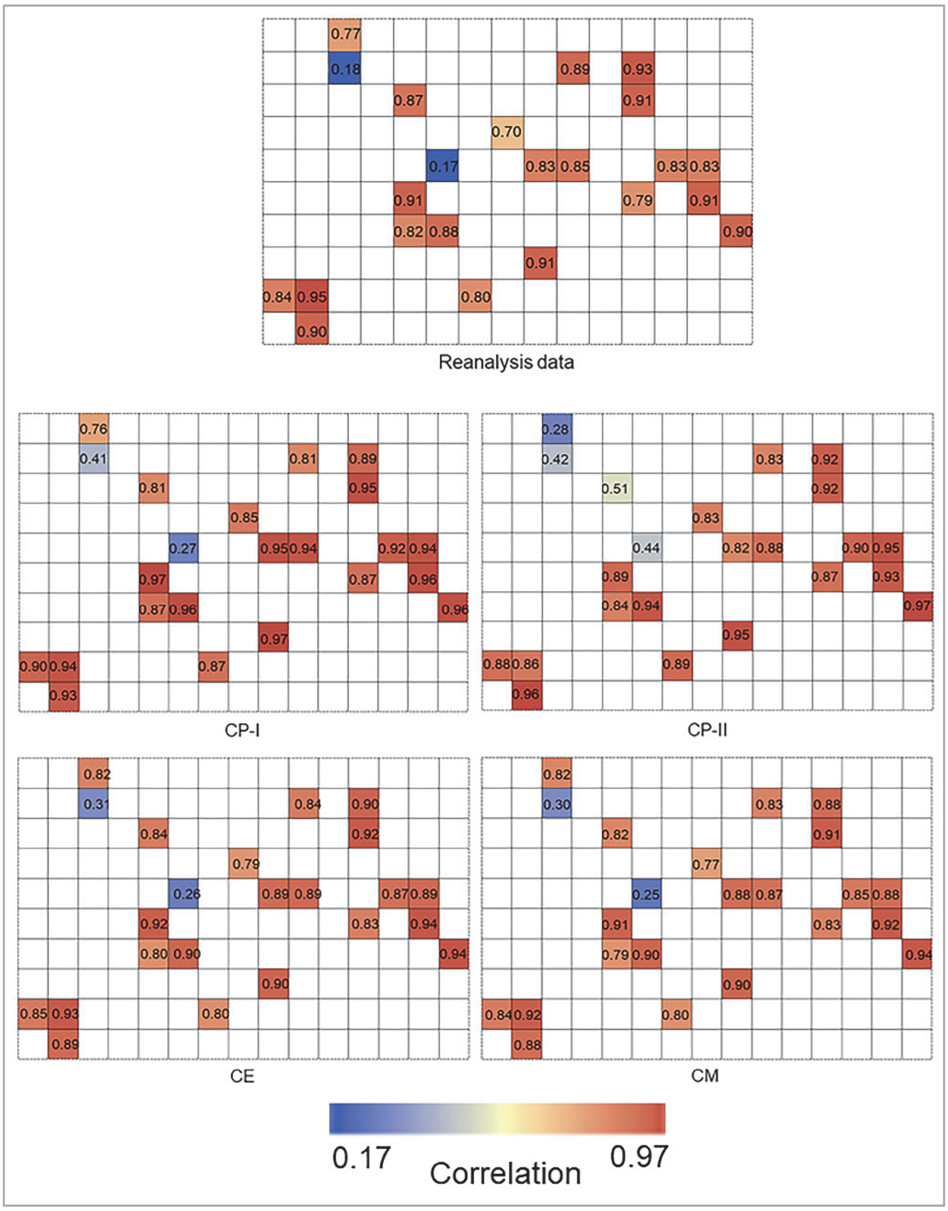


Figure 4.10 The correlation coefficients *r* in time at each weather station. The numbers on the figures denote correlations.

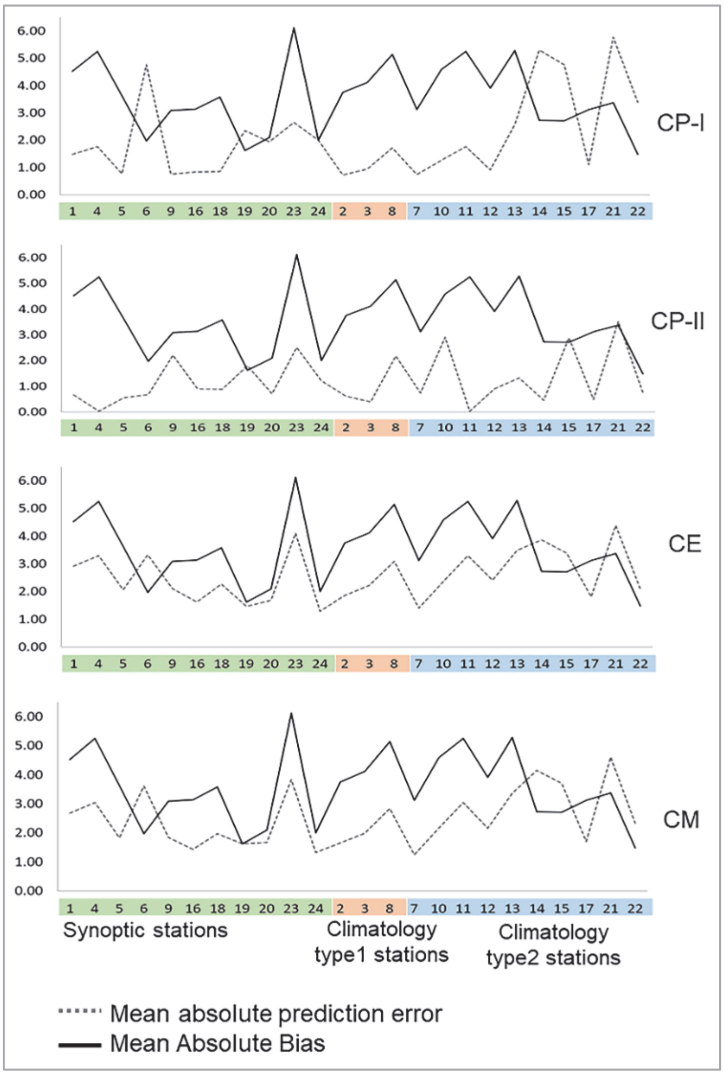


Figure 4.11 Comparing spatial mean absolute prediction error (MAPE) with spatial mean absolute bias (MAB) at three types of weather stations. The vertical axis is error/bias in °C. The synoptic stations are supposed to provide more precise measurements.

4.5 Discussion

In this paper, we presented and evaluated two new bias correction methods for air temperature that take temporal and spatial variations into account. The CE and CM methods produce smooth maps, assuming spatial stationarity when estimating the dependence structures between the measured and the reanalysis weather data. We proposed to use different conditional probabilities minimizing the bias in space to improve spatial variation of the bias-corrected values. In addition, we described the dependence structure between the measured and the reanalysis weather data using the flexibility of selecting the best fitting family among five copula families.

In our application, a bivariate copula was fitted to daily observations of the involved variables assuming spatial stationarity, and the bias correction was applied separately on each day. The results showed that our methods performed better to correct time-series of the air temperatures, i.e., the temporal variation of the daily air temperatures in June 2014. Therefore, a practical advantage of the new methods is that they are not any longer restricted to remove autocorrelation and heteroscedasticity in time-series. A novel aspect is the potential and the use of new methods for other copula-based methods such as interpolation and downscaling where the variable of interest needs to be predicted.

By means of the comparison of the methods based upon error scores and correlation scores, we demonstrated that CP-I performed best in time, whereas CP-II performed best in space. As the copulas are generally able to describe spatio-temporal dependences, the use of the spatio-temporal information in CP-II might help to improve its performance in time as well. We selected the number of neighbours based upon our experience. A more generally applicable sensitivity analysis is necessary to show the effects of the number of nearest neighbours on performance of CP-II.

We identified several routes for future research. First, we treated the measurements from weather stations as the benchmarks in the identification of bias and in the cross-validation. To address the uncertainty of the measurements and its impact on the results of the proposed methods, the proposed methods should be extended towards other datasets. In addition, further applications of the new copula-based methods in other case studies including simulation-based information should provide more insight on these methods. Second, we used the AIC to select the best fitting family. We realize though that the suitability of a copula also depends on the number of data used for fitting and the probabilistic nature of the bias. Further cross validations need to be carried out using random samples of the measurements to choose the copula family. Third, spatially varying conditional probabilities needs to be further applied in other methods, e.g., Bayes' classifier and possibly in a machine learning environment. Fourth, to extend the current study, the use of multivariate copula describing the dependence between more variables, e.g., air temperature, elevation and land cover might help to improve the performance of the presented methods. The bivariate case of the proposed methods in this paper is useful if such a covariate is unavailable. Finally, a comparison to other bias correction methods, e.g., quantile mapping might be included in further studies.

4.6 Conclusions

We proposed to use conditional probabilities to correct for bias in the gridded reanalysis weather data provided by ECMWF as compared to the measurements from weather stations taken as the benchmarks. Cross-validation results and correlation scores showed that the new methods perform better than commonly applied methods and are able to account for spatial and temporal variation of air temperatures at unvisited locations.

Appendix 4.1 Evaluating the stationarity assumption

To evaluate the second order spatial stationarity assumption in estimating marginal distribution of daily air temperature, we used two methods: linear regression and auto-correlation function. The null hypothesis H_0 and alternative hypothesis H_1 to test for second order stationarity assumption are then defined as:

$$H_0\colon\ E[Z_s] = \mu, \tag{4.12}$$

$$H_1\colon\ E[Z_s] = \beta_0 + \beta_1 . x_s + \beta_2 . y_s, \tag{4.13}$$

where Z_s is the variable of interest at location s, *E[]* denotes the expectation, x_s and y_s are the *x* and *y* coordinates of location s and the $\beta_j, j = 0, 1, 2$ are regression parameters. We obtained the parameters and their *p* values using a linear model and *F* test (Chambers et al. 1990). We found that the values of regression coefficients are not significantly different from zero and their *p* values of *F* test are above 0.05 and 0.01 at all days (Figure 4.12). The auto-correlation function, i.e., correlogram, describes dependences in space based upon the correlation per each spatial lag (Oden, 1984). The values of correlogram at five spatial lags are obtained from the measured values on each day in June between 2004 and 2014 (Figure 4.13). It is immediate that the correlations are decreasing by the separating distance. These results and the limited effects of including non-stationarity make the assumption of spatial stationarity a reasonable one.

We assess the second order spatial stationarity assumption in estimating copula using the co-correlation function. Co-correlation function, i.e., the co-correlogram, is an extension of the correlogram for two or more random fields in space. The values of the co-correlogram and the best fitting family at five spatial lags are obtained from measured and reanalysis values on each day in June between 2004 and 2014 (Table 4.5). The results show that the best fitting families and the correlations differ slightly at different spatial lags. Therefore, we conclude that the spatial stationarity is a reasonable assumption in estimating copula and point to further application of co-correlogram in a lag based bias correction methods.

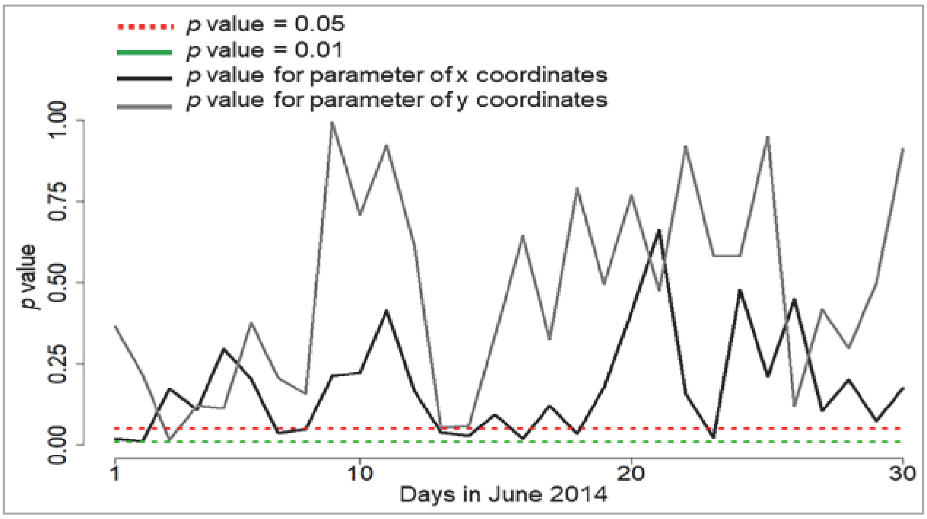


Figure 4.12 *p* values of the regression parameters in trend analysis obtained by *F* test. Based upon its results, spatial stationarity is assumed in estimating the marginal distribution.

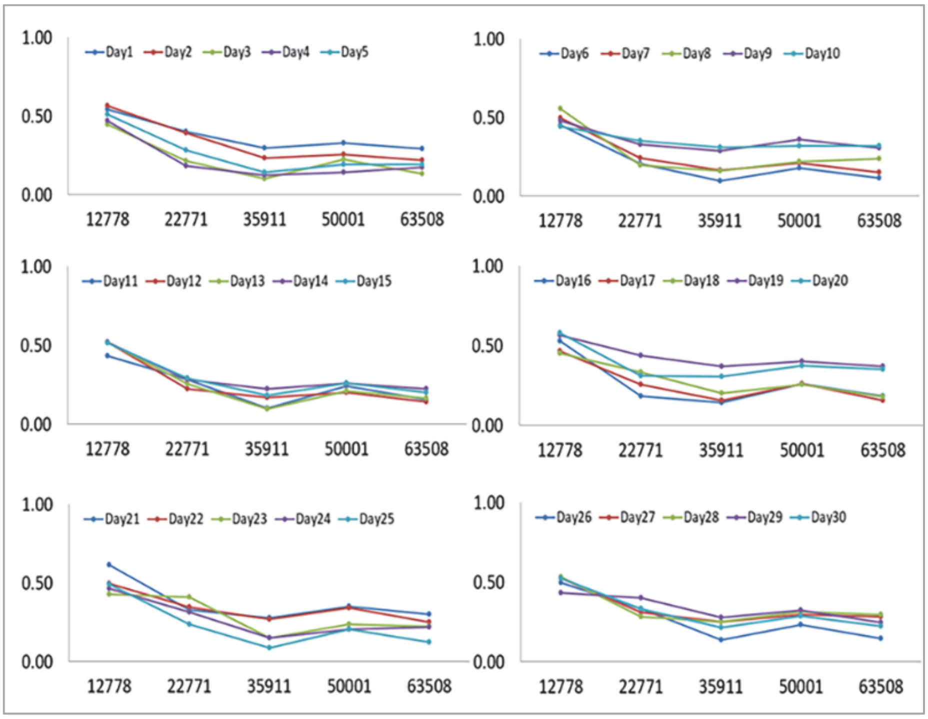
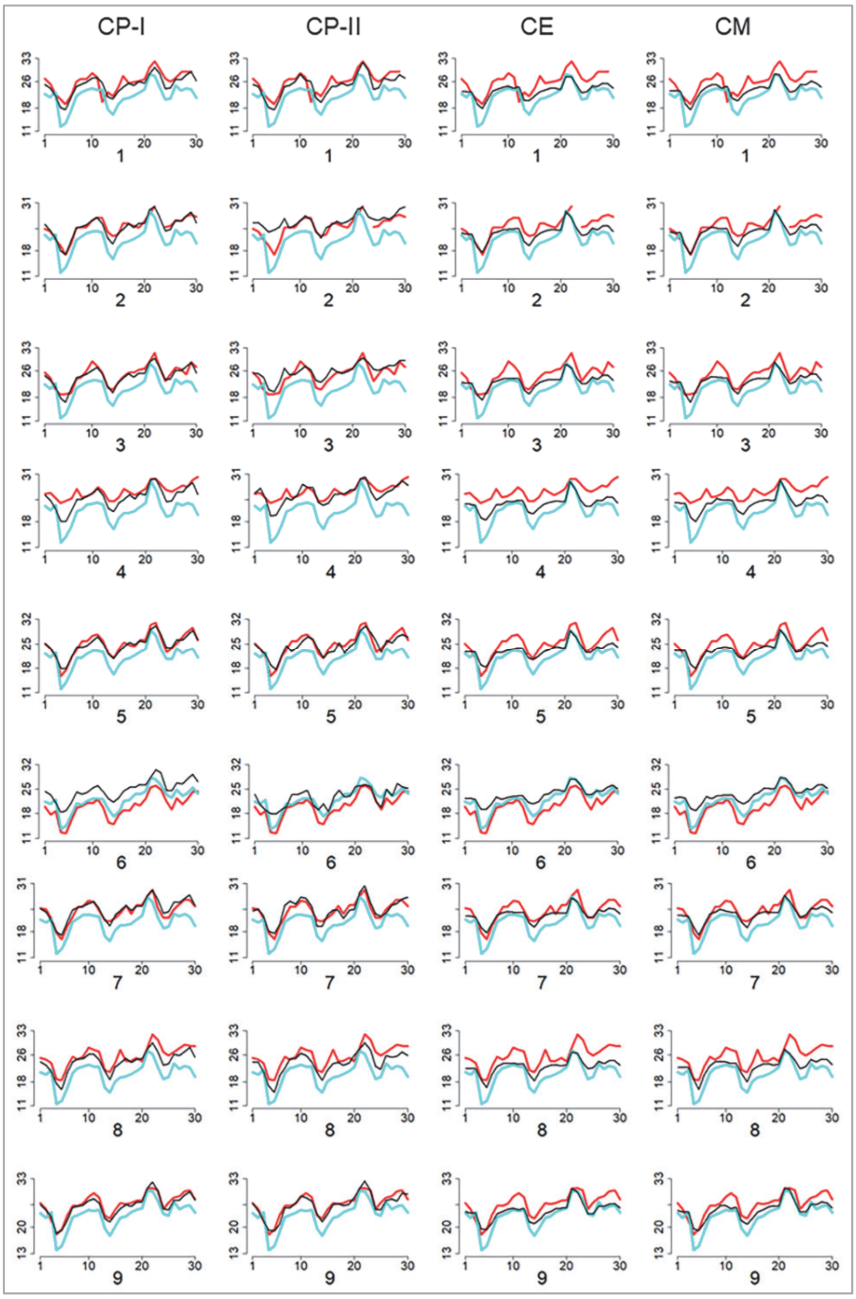


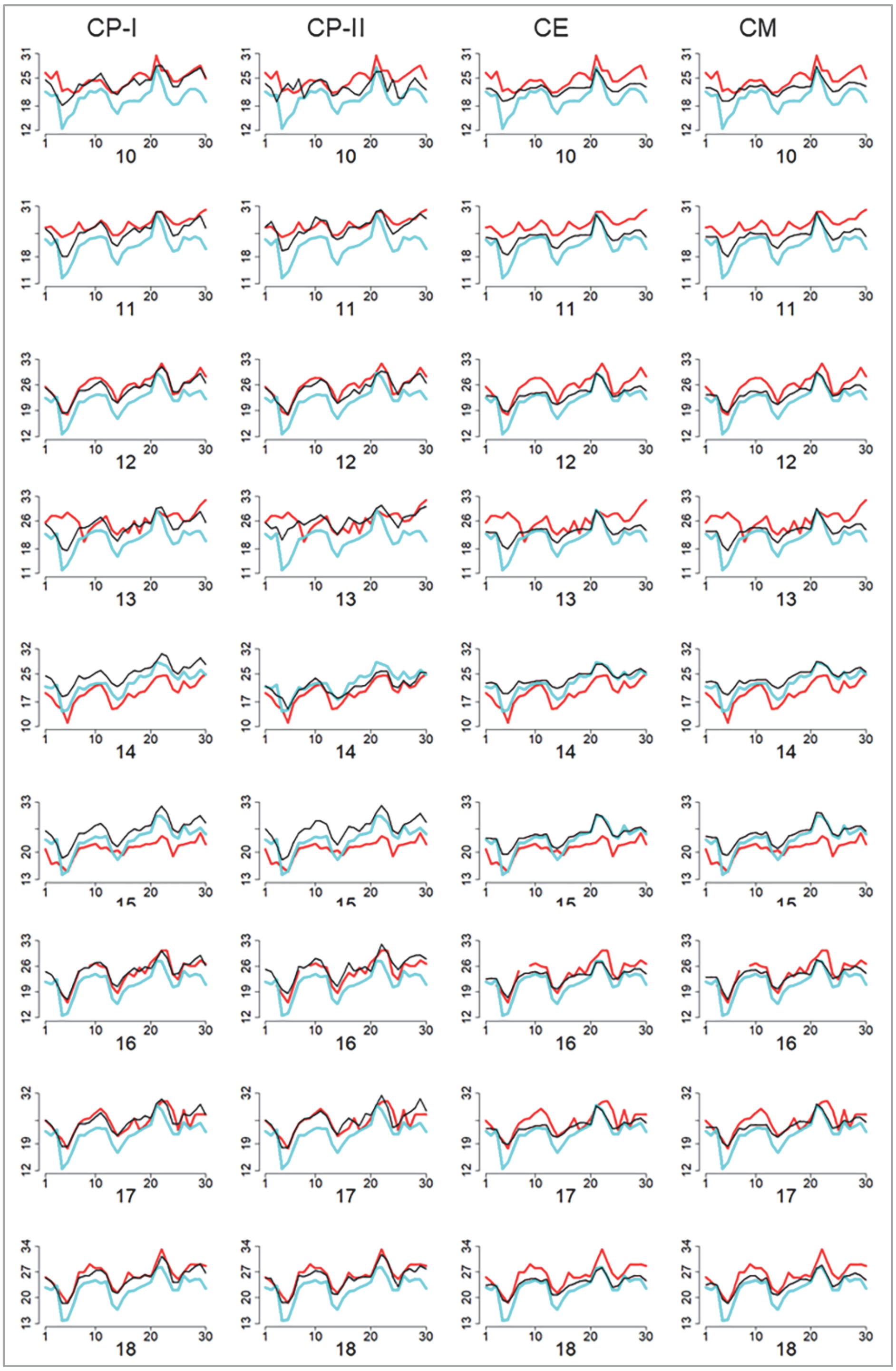
Figure 4.13 The values of correlogram at five spatial lags. The vertical axis is Kendall's τ correlations obtained using the measurements on each day in June between 2004 to 2014. The horizontal axis is spatial lags in meter.

Table 4.5 The values of co-correlogram and best fitting family at five spatial lags. Kendall's τ correlations are obtained using the measured and reanalysis values on each day in June from 24 weather stations between 2004 to 2014. The copula families are: N=Gaussian, T=Student's *t*, C=Clayton, G=Gumbel, and F=Frank.

	Best fitted family					**Kendall's τ correlation**				
Day	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
1	G	G	G	G	G	0.45	0.45	0.46	0.43	0.43
2	N	N	N	N	N	0.35	0.37	0.37	0.34	0.34
3	N	G	G	G	G	0.28	0.31	0.31	0.30	0.28
4	N	G	G	G	G	0.34	0.31	0.33	0.29	0.29
5	N	T	T	T	T	0.35	0.34	0.37	0.31	0.32
6	F	F	F	F	F	0.36	0.34	0.32	0.29	0.30
7	G	N	G	G	G	0.38	0.38	0.38	0.35	0.35
8	G	G	G	T	T	0.42	0.42	0.38	0.38	0.37
9	N	G	G	G	F	0.46	0.46	0.44	0.44	0.42
10	G	G	G	G	G	0.37	0.39	0.40	0.35	0.38
11	G	G	G	G	G	0.27	0.28	0.29	0.27	0.27
12	N	N	N	N	N	0.40	0.40	0.39	0.39	0.39
13	N	N	T	N	T	0.36	0.37	0.39	0.35	0.35
14	N	N	N	N	N	0.33	0.38	0.38	0.35	0.35
15	N	G	G	G	G	0.35	0.38	0.38	0.35	0.35
16	N	G	G	G	G	0.38	0.37	0.35	0.35	0.34
17	N	G	G	G	G	0.33	0.38	0.37	0.34	0.33
18	F	F	F	F	F	0.25	0.31	0.30	0.30	0.28
19	F	F	F	F	F	0.46	0.48	0.49	0.45	0.45
20	G	G	G	G	G	0.50	0.50	0.50	0.47	0.49
21	F	G	G	G	G	0.47	0.45	0.47	0.44	0.45
22	G	G	G	G	G	0.38	0.35	0.36	0.34	0.33
23	F	G	G	G	F	0.31	0.33	0.37	0.31	0.33
24	G	G	G	G	G	0.34	0.30	0.33	0.28	0.30
25	G	G	G	G	G	0.19	0.25	0.30	0.20	0.22
26	G	G	G	G	G	0.23	0.31	0.34	0.29	0.29
27	G	N	N	N	N	0.35	0.37	0.43	0.34	0.36
28	F	F	F	F	F	0.32	0.35	0.37	0.32	0.32
29	N	F	F	F	F	0.34	0.34	0.35	0.31	0.32
30	G	G	G	G	G	0.32	0.31	0.32	0.29	0.31

Appendix 4.2 predictions in time and space




CP-I
CP-II
CE
CM
10
11
12
13
14
15
16
17
18
1
10
20
30

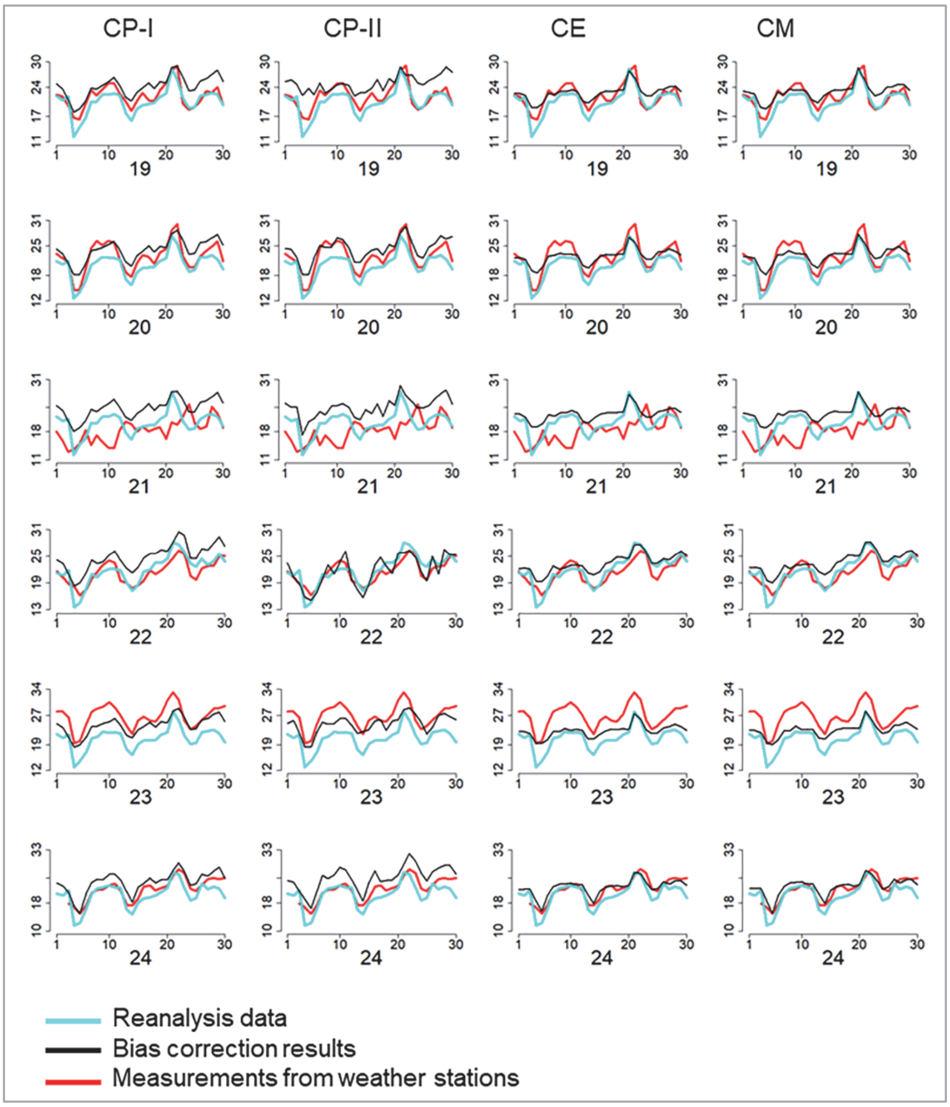
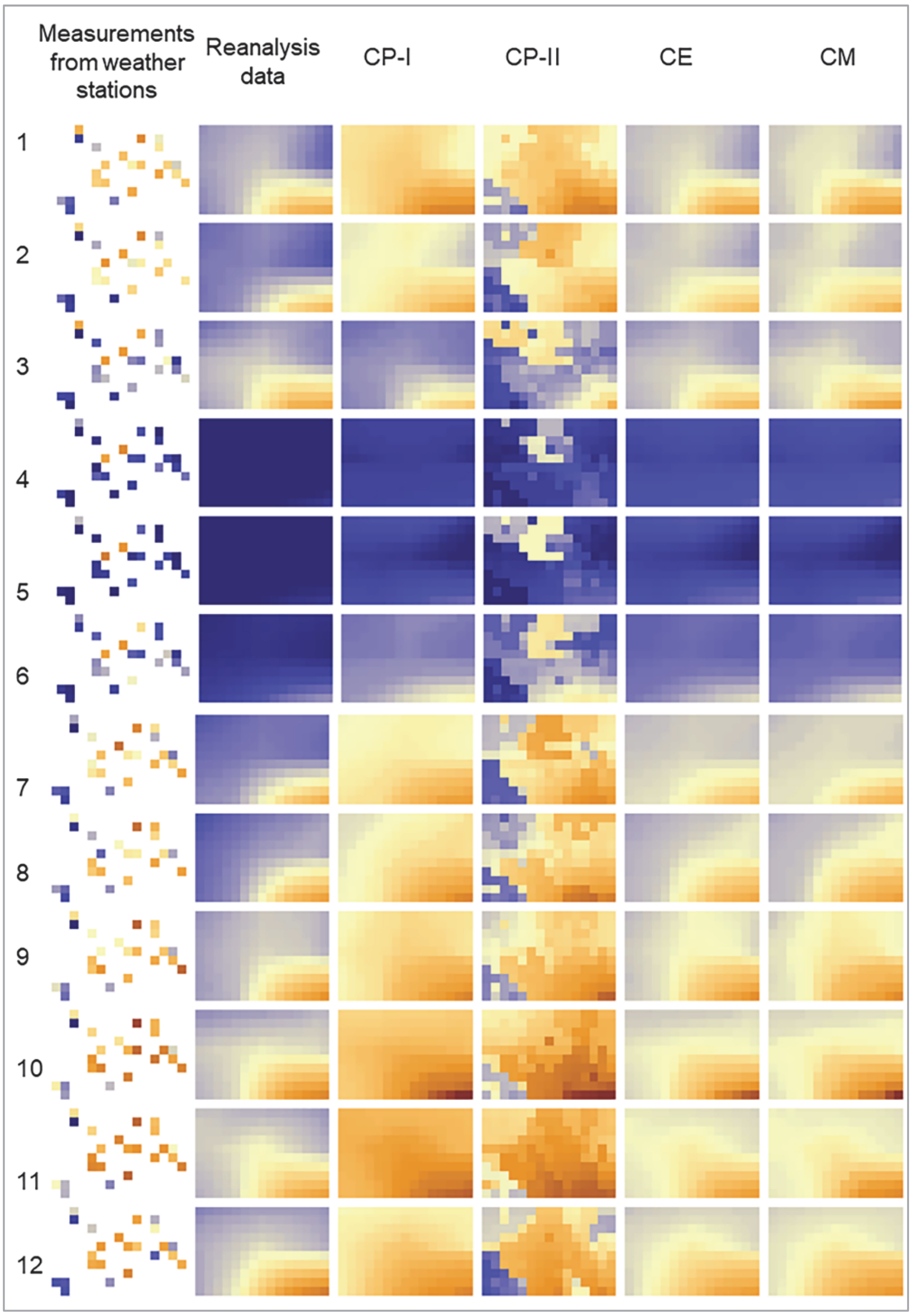
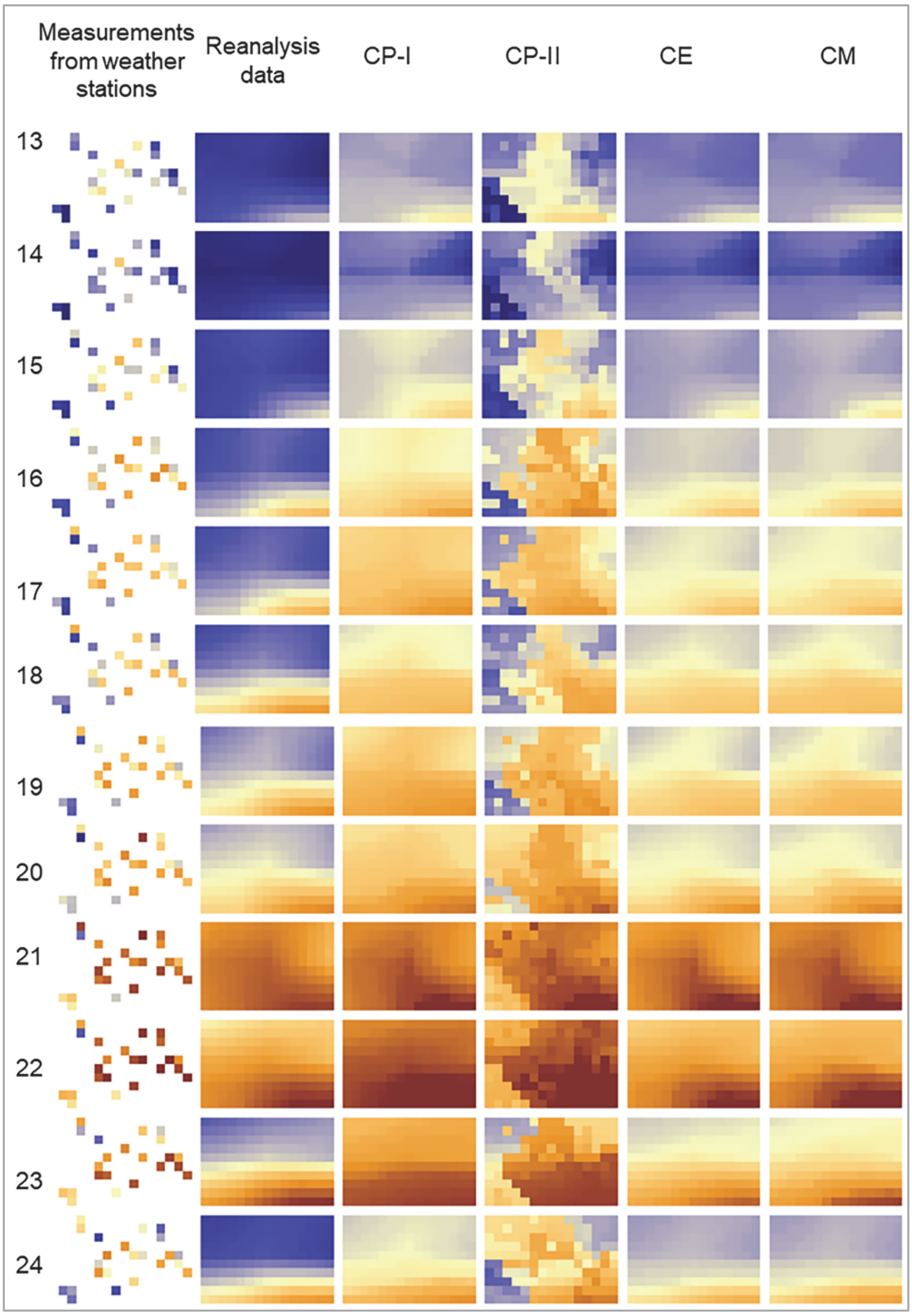


Figure 4.14 Time-series of the measurements from weather stations, reanalysis data and bias-corrected values obtained by the bias correction methods at each station in June 2014. The vertical axis is the daily mean air temperature in °C. The number on each graph denotes the weather station number.


Measurements from weather stations
Reanalysis data
CP-I
CP-II
CE
CM
1
2
3
4
5
6
7
8
9
10
11
12


Measurements from weather stations
Reanalysis data
CP-I
CP-II
CE
CM
13
14
15
16
17
18
19
20
21
22
23
24

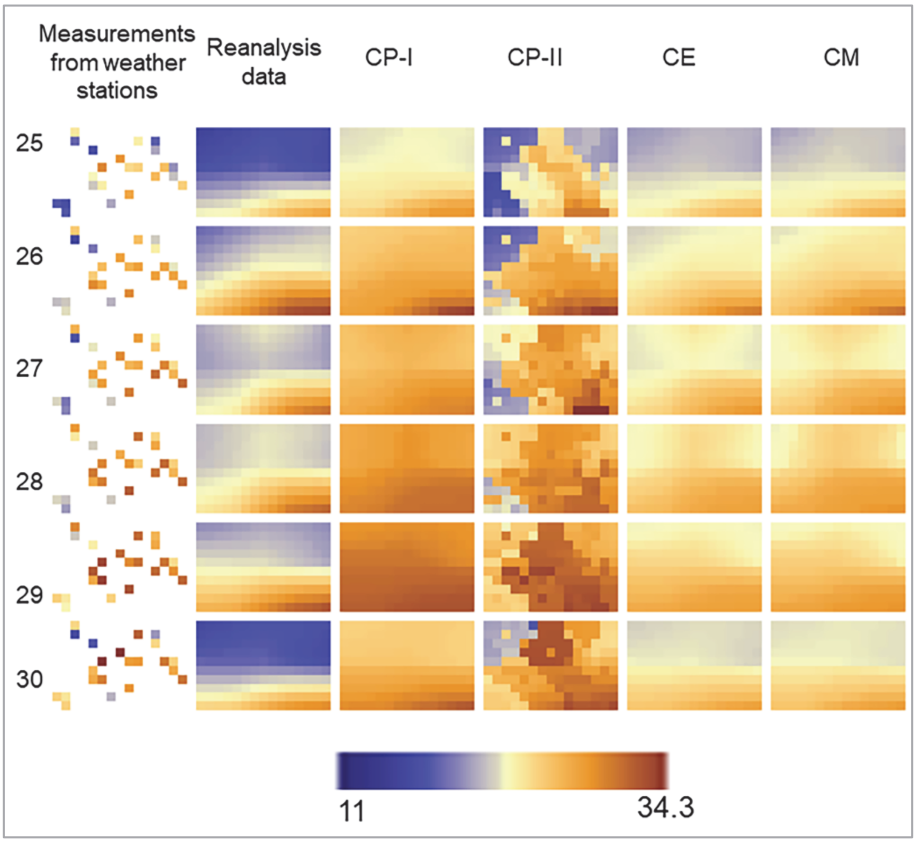
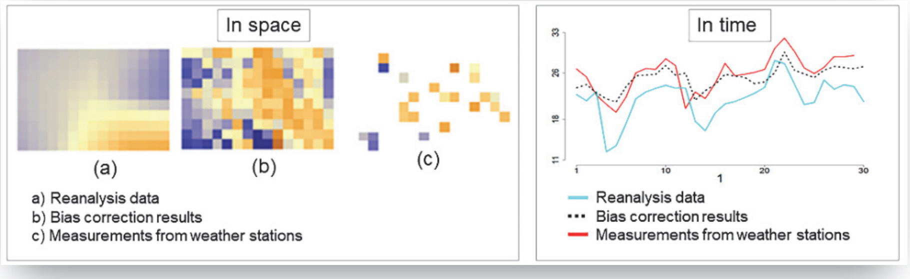


Figure 4.15 The daily mean air temperatures from weather stations, reanalysis data and bias-corrected values obtained by the bias correction methods for all locations on each day in June 2014. The number on each graph denotes the day in June 2014.

Chapter 5: Multivariate copula quantile mapping for bias correction



This chapter is published as: Alidoost F., Stein A., Su Z, Sharifi, A. 2019. Multivariate copula quantile mapping for bias correction of reanalysis air temperature data. *Journal of Spatial Science*. https://doi.org/10.1080/14498596.2019.1601138 (In Press)

Abstract

Gridded reanalysis air temperature data retrieved from the European Centre for Medium-range Weather Forecasts (ECMWF) are useful for hydrological studies in a data-scarce agricultural area. A justified use requires to correct for bias, defined as the systematic difference between reanalysis values and measurements from weather stations. We propose three multivariate copula quantile mappings (MCQMs) to predict the bias-corrected air temperature at unvisited locations. MCQMs estimate multivariate distributions using two types of covariates for air temperature. Daily air temperature was retrieved at 24 weather stations and from the ECMWF ERA-Interim archive at 150 grid cells for a single month over 11 years in the Qazvin Plain, Iran. Cross-validation and correlations showed that MCQMs reduced bias with 46% as compared with classical quantile mapping. The study concludes that MCQMs are well able to describe covariability and to represent the spatial and temporal variation of air temperature.

Keywords

bias correction, copula, conditional, mean temperature, data scarce.

Author contributions

F.A. conceived and designed the analysis, collected and processed the data, developed tools, performed the analysis, wrote the manuscript, is the corresponding author.

A.S. supervised the findings of this work, verified the analytical methods, encouraged A.F. to investigate copulas, improved the English wording.

Z.S. supervised the findings of this work, encouraged A.F. to investigate bias correction of ECMWF data.

A.Sh. collected the data, encouraged A.F. to apply the method to a data-scarce environment.

All authors contributed to the interpretation of the results, and commented on the final manuscript.

Structure of the chapter

After the introduction in section 5.1, copulas and bias correction methods are presented in section 5.2. The study area and data are introduced in section 5.3. The results are discussed in section 5.4. We conclude and point to further directions of this work in section 5.5. This is followed by two appendices in sections Appendix 5.1, 5.2.

5.1 Introduction

Hydrological studies refer to air temperature as a key variable to support water management in an irrigation network. At local scales (Sarma, 2005), sparsely and irregularly distributed data from weather stations are a challenge for hydrological studies at unvisited locations in irrigation networks. To address the problem, additional spatially distributed data may be included, e.g., gridded reanalysis weather data from the European Centre for Medium-range Weather Forecasts (ECMWF). The coarse resolution of models, the mutual dependence of weather parameters, and variability of these parameters in space and time are major sources of uncertainties when using reanalysis weather data (Dee et al. 2011; Durai and Bhradwaj 2014).

In our paper, weather station measurements are considered as benchmarks. Hence, bias is defined as the difference between the reanalysis values and the measurements from weather stations (Hannah and Valdes 2001; Persson 2013). We consider an unvisited location at the center of a grid cell characterized by a reanalysis value, but without a measurement from a weather station.

Various bias correction methods have been proposed in the literature: quantile mapping (Ines and Hansen 2006), linear-scaling factor methods (Lenderink et al. 2007) and nonlinear methods (Lafon et al. 2013). The Gamma and empirical distributions have been used for bias correction of precipitation data and the Gaussian distribution for bias correction of air temperature data (Teutschbein and Seibert 2012; Lafon et al. 2013; Kum et al. 2014).

Recently, copula-based methods have been developed for deriving bias-corrected weather data (Vogl et al. 2012; Mao et al. 2015). A copula links univariate distributions with a multivariate distribution based upon Sklar's theorem (Sklar 1973; Nelsen 2006). So far, the methods have mainly been applied to precipitation time-series retrieved from regional climate models under the assumption of temporal stationarity. Laux et al. (2011) employed bivariate copulas to describe dependences between daily precipitation time-series retrieved from a regional climate model and measurements at three locations where data are available. They fitted a bivariate copula to daily time-series at one location, ignoring the temporal variation of the copula parameters as well as any spatial dependency. In addition, fitting is required to remove autocorrelation and heteroscedasticity, which may exist in a climate time-series (Laux et al. 2011). Mao et al. (2015) investigated bias correction methods of daily precipitation data and showed that copula-based bias correction performs better than quantile mapping.

The aim of our study is to obtain bias-corrected daily air temperature at unvisited locations in a data-scarce area. To do so, we developed three multivariate copula quantile mappings. Copulas help to estimate the joint multivariate distributions of air temperature and its covariate, in our study: elevation. We investigated two types of dependences: the dependence between air temperature and elevation at a single location, the dependence between air temperatures at a single location and its nearest neighbour. The new methods are compared with classical quantile mapping.

5.2 Bias correction methods

5.2.1 Multivariate copula quantile mappings

Multivariate copula quantile mapping (MCQM) is a d-dimensional quantile mapping method that relies on two conditional copula distributions (Gräler 2014; Verhoest et al. 2015). From two random variables X and Y over the same spatial domain, n samples $\{x_1, \dots, x_n\}$ are obtained from weather station measurements and m samples $\{y_1, \dots, y_m\}$ from reanalysis weather data. Bias b_i at location i is:

$$b_i = x_i - y_i, \tag{5.1}$$

The joint distribution function $H(X, Y)$ is written in terms of a copula as $C(U, V)$, where U and V are uniformly distributed random variables (Nelsen 2006). The empirical marginal probability u_i using the rank-order-transformation equals:

$$u_i = \frac{\mathrm{rank}(x_i)}{n+1}, i = 1, \dots, n. \tag{5.2}$$

A monotone cubic spline is fitted to the pairs (x_i, u_i) to obtain a continuous approximation of the marginal distribution F_X as $u_i = F_X(x_i)$ (Fritsch and Carlson, 1980). The marginal distribution F_Y is estimated in a similar way. Use of an empirical distribution avoids estimating theoretical marginal distributions that might otherwise affect the estimation of copula parameter. Further note that the marginal distribution is assumed to be stationary (see appendix 5.1).

The purpose of quantile mapping is to predict u_i at an unvisited location i. The inverse transformation of the marginal distribution F_X^{-1} provides the bias-corrected value $\hat{x}_i$:

$$\hat{x}_i = F_X^{-1}(\hat{u}_i), \tag{5.3}$$

where the notation $\widehat{}$ denotes that $\hat{x}$ and $\hat{u}$ are predicted values. To obtain $\hat{u}_i$, we develop three MCQMs including d-dimensional joint distributions where $2 \leq d \leq 3$.

MCQM-I: let Z be a covariate for X and Y, e.g., elevation. Then two conditional distributions $C(U|W = w_i)$ and $C(V|W = w_i)$ are obtained based upon bivariate joint distributions $C(U, W)$ and $C(V, W)$ describing non-spatial dependences, where the distributions can belong to different families and $w_i = F_Z(z_i)$. The marginal probability $\hat{u}_i$ is obtained using the inverse transformation of $C(U|W = w_i)$ as:

$$\hat{u}_i = C^{-1}(C(v_i|W = w_i)|W = w_i). \tag{5.4}$$

Distributions can be extended to higher dimensions if more than one covariate is available.

MCQM-II: we consider two bivariate joint distributions $C(U, U_{-i})$ and $C(V, V_{-i})$ that describe spatial dependences between air temperatures at location i and its nearest neighbour $-i$ and two conditional distributions $C(U|U_{-i} = u_{-i})$ and $C(V|V_{-i} = v_{-i})$ are based upon the joint distributions, where $u_{-i} = F_X(x_{-i})$ and $v_{-i} = F_Y(y_{-i})$. The marginal probability $\hat{u}_i$ is then obtained as:

$$\hat{u}_i = C^{-1}(C(v_i|V_{-i} = v_{-i})|U_{-i} = u_{-i}). \tag{5.5}$$

Distributions can be extended to higher dimensions using more than one neighbour where the number of observations is sufficient to obtain a correlogram that describes dependences in space (Oden, 1984).

MCQM-III: the third method combines MCQM-I and MCQM-II. We consider two conditional distributions $C(U|U_{-i} = u_{-i}, W = w_i)$ and $C(V|V_{-i} = v_{-i}, W = w_i)$ based upon trivariate joint distributions $C(U, U_{-i}, W)$ and $C(V, V_{-i}, W)$ describing non-spatial and spatial dependences. The marginal probability $\hat{u}_i$ is then obtained as:

$$\hat{u}_i = C^{-1}(C(v_i|V_{-i} = v_{-i}, W = w_i)|U_{-i} = u_{-i}, W = w_i). \tag{5.6}$$

As for MCQM-II, distributions can be extended to higher dimensions. For MCQMs, it is assumed that the conditional probability of X conditioned on its covariate $F_X(X|.)$ is equal to the conditional probability of Y conditioned on that covariate $F_Y(Y|.)$.

5.2.2 Copula estimation in MCQMs

A bivariate copula describes the dependences between two variables. We used five copula families among several families available in the literature (see Section 2.2). In MCQM-III, we estimate the conditional distribution $C(U|U_{-i} = u_{-i}, W = w_i)$ based upon a canonical vine or C-vine structure: $c(U, W)$, $c(U, U_{-i})$ and $c(C(W|U), C(U_{-i}|U))$ (see Section 2.1). The conditional distribution $C(V|.)$ is estimated in a similar way.

5.2.3 Quantile mapping

A comprehensive study carried out by Teutschbein and Seibert (2012) showed that quantile mapping (QM) performs best among the classical bias correction methods. QM is implemented as:

$$\hat{x}_i = F_X^{-1}(v_i). \tag{5.7}$$

QM assumes that there is a perfect dependence between variables i.e. $\hat{u}_i = v_i$. It is sensitive to the number of quantile divisions when using an empirical marginal distribution. There are several names in the literature for this method, such as probability mapping, CDF matching, and quantile-quantile mapping.

5.2.4 Comparison and evaluation of the bias correction methods

We compare MCQMs with quantile mapping using leave-*K*-out cross-validation (Lafon et al. 2013). To this end, the observations in *K* successive years at day j and station i are removed from the dataset and the bias-corrected values are predicted using the reminder of the observations. The mean absolute error $MAE_{i,j}$ equals:

$$MAE_{i,j} = \frac{1}{K}\sum_{k=1}^{K}\left|x_{i,j,k} - \hat{x}_{i,j,k}\right|. \tag{5.8}$$

We determine total mean absolute error MAE, spatial and temporal error scores, i.e., SES and TES for *t* days and *n* stations as:

$$MAE = \frac{1}{T}\sum_{j=1}^{t}\left(\frac{1}{N}\sum_{i=1}^{n} MAE_{i,j}\right), \tag{5.9}$$

$$SES = \sum_{i=1}^{n}\left(rank\left(\frac{1}{t}\sum_{j=1}^{t} MAE_{i,j}\right)\right), \tag{5.10}$$

$$TES = \sum_{j=1}^{t}\left(rank\left(\frac{1}{n}\sum_{i=1}^{n} MAE_{i,j}\right)\right), \tag{5.11}$$

The lowest score indicates the best method (Durai and Bhradwaj 2014). In addition, we define correlations r_i and r_j that indicate temporal and spatial dependences between measurements and bias-corrected values, respectively as:

$$r_i = corr(\{\hat{x}_{i,1}, \hat{x}_{i,j}, \dots, \hat{x}_{i,t}\}, \{x_{i,1}, x_{i,j}, \dots, x_{i,t}\}), \tag{5.12}$$

$$r_j = corr(\{\ \hat{x}_{1,j}, \hat{x}_{i,j} \dots, \hat{x}_{n,j}\}, \{x_{1,j}, x_{i,j}, \dots, x_{n,j}\}). \tag{5.13}$$

Spatial and temporal correlation scores i.e. SCS and TCS are then obtained as:

$$SCS\ = \sum_{i=1}^{n}\left(rank(r_i)\right), \tag{5.14}$$

$$TCS = \sum_{j=1}^{t} \left(rank(r_j) \right). \tag{5.15}$$

The highest score indicates the best method.

5.3 Case study: daily mean air temperature in Iran

Our methods are applied to daily mean air temperature the Qazvin irrigation network located in the Qazvin plain, Iran in June from 2004 to 2014 (see Section 3.1). The measurements at the stations are assigned to the reanalysis values at the nearest grid cells. For instance, the measurements at stations number four and eleven are assigned to the reanalysis value at a grid cell. There are 150 grid cells × 11 years = 1650 reanalysis air temperatures and 24 stations × 11 years = 264 measurements at each day of June. Missing values in the measurements from weather stations may occur; their number differs between stations and days.

A comparison of the time-series of the measurements and reanalysis values revealed systematic overestimation and underestimation (Appendix 5.2, Figure 5.7). We noted that the time-series at stations 13 and 21 have a lower correlation with the time-series of reanalysis air temperature than the other stations (Figure 5.1b). The time-series at those stations revealed that the quality of their measurements, in particular, their accuracy is low (Appendix 5.2, Figure 5.7). In addition, spatial correlations between the measurements and reanalysis air temperature are weak at most of the days in June 2014 (Figure 5.1a).

This study focuses on obtaining the bias-corrected daily air temperature at unvisited locations at each day in June 2014. The total mean absolute bias was equal to 3.6°C for all stations and all days. We did not consider predicting the bias-corrected air temperature at an unvisited location using the mean absolute bias since there is both spatial and temporal variation. The MODIS elevations are retrieved in 22410 pixels at a spatial resolution of 1 km (see Section 3.1). We used the same elevations for eleven years assuming that elevation remains the same.

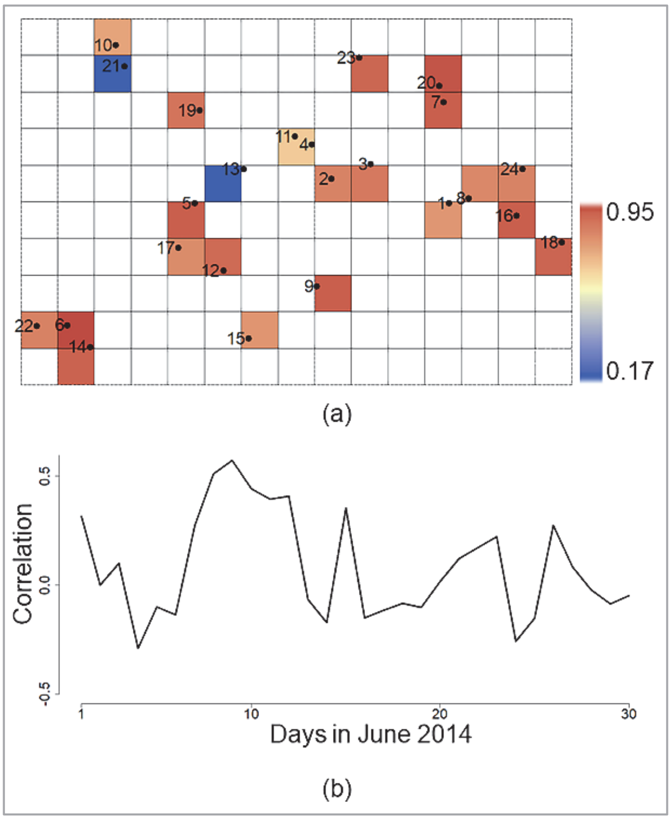


Figure 5.1 Correlations r_i and r_j that indicate temporal and spatial dependences between measurements and ECMWF ERA-interim reanalysis air temperature. a) r_i at each weather station, b) r_j at each day in June 2014.

5.4 Results and discussion

5.4.1 Marginal distributions and copulas

Marginal distributions and copulas are estimated for each day in June 2014, separately. The empirical marginal distribution on the first day is shown in Figure 5.2. The method to estimate empirical marginal distribution is not unique, and a more generally applicable sensitivity analysis might help to explain the effects of other methods on the results. For instance, we also used kernel density estimation and noticed that the final results of the bias correction methods changed only slightly (results not shown). To assess spatial stationarity, a trend surface was fitted to the measurements (Appendix 5.1). The β_1 parameter has p values in the range of [0.02, 0.80] with mean value equal to 0.19, whereas the β_2 parameter has p values in the range of [0.02, 0.99] with a mean value of 0.45. We were thus safe to assume spatial stationarity when estimating the marginal distributions.

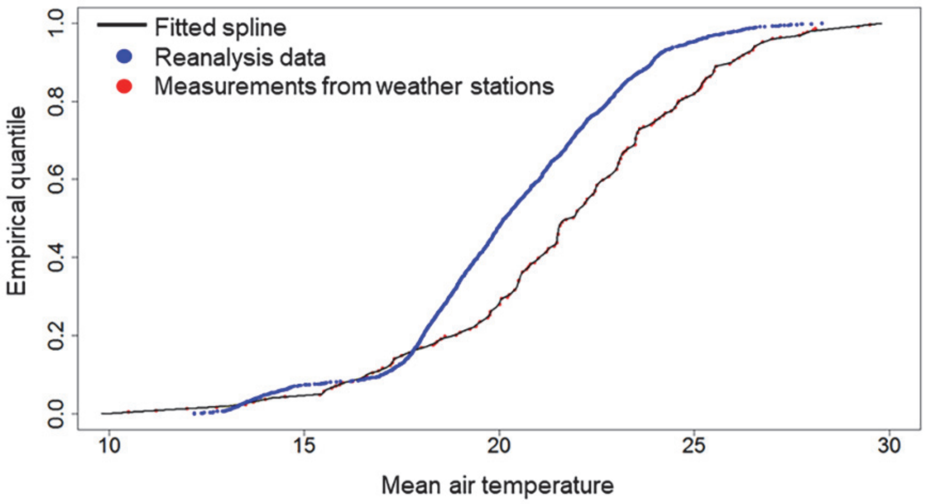


Figure 5.2 Empirical marginal probabilities on June 1st. A monotone cubic spline is fitted to obtain the marginal distribution function. Marginal distribution functions are estimated at each day of June, separately.

The parameters of five copula families and the number of data for fitting purposes are listed in Table 5.1. We considered the elevation as the covariate in MCQM-I. We found that the best fitting family was the Frank family for the joint distribution of the measurements and the elevation for all days and also, for the joint distribution of the reanalysis air temperature and the elevation for 18 days (Table 5.1). The *p* values of the Cramér–von Mises statistic $S_n^{(B)}$ were larger than 0.05 for all days showing that the best fitting family is well describing the dependences (Table 5.1).

We considered spatial dependences in MCQM-II. The Student's *t* family dominates the dependences of the measurements for 14 days and the dependences of the reanalysis air temperature for 15 days (Table 5.1). The *p* values of The Cramér–von Mises statistic $S_n^{(B)}$ and the White statistic were larger than 0.1 except for the Gumbel family at five days, showing that the best fitting family is well describing the spatial dependences (Table 5.1). The p values were close to zero and the best fitting family was the Gumbel family at days 1, 10, 17, 21 and 22. The low *p* values are related either to the limitation of the test or to the inflexibility of those five families. The p values were close to one for the Student's t.

For MCQM-III, the parameters of three bivariate copulas were estimated (Table 5.2). Best fitting families turned out to be non-Gaussian families for most of the days. The *p* values of the Cramér–von Mises statistic $S_n^{(B)}$ were larger than 0.2 for most of the days showing that the best fitting family is well describing the dependences.

Table 5.1 The *p* value and best fitting families in MCQM-I and MCQM-II. The copula families are: N=Gaussian, T=Student's t, C=Clayton, G=Gumbel and F=Frank. Number of data denotes the number of marginal probabilities of each variable used for fitting purposes and equals to the number of weather station measurements at each day in June during the years 2004 to 2014.

		MCQM-I				**MCQM-II**			
		$\boldsymbol{C(U,W)}$		$\boldsymbol{C(V,W)}$		$\boldsymbol{C(U,U_{-i})}$		$\boldsymbol{C(V,V_{-i})}$	
Day	**Number of data**	***p*-value**	**Best**	***p*-value**	**Best**	***p*-value**	**Best**	***p*-value**	**Best**
1	226	0.36	F	0.45	F	0.99	T	0.00	G
2	224	0.29	F	0.42	F	0.99	T	1.00	T
3	226	0.26	F	0.32	F	1.00	T	0.99	T
4	226	0.18	F	0.25	F	1.00	T	0.29	G
5	226	0.31	F	0.44	F	1.00	T	0.98	T
6	226	0.21	F	0.28	F	0.59	F	0.92	F
7	226	0.15	F	0.33	F	0.51	F	0.98	T
8	225	0.39	F	0.41	F	1.00	T	0.93	F
9	226	0.28	F	0.31	N	0.44	F	0.62	F
10	226	0.27	F	0.46	N	0.44	G	0.00	G
11	226	0.26	F	1.00	T	0.66	G	0.93	F
12	226	0.37	F	0.27	F	1.00	T	0.99	T
13	226	0.29	F	0.25	F	1.00	T	1.00	T
14	226	0.19	F	0.51	N	1.00	T	0.96	F
15	226	0.09	F	0.45	N	1.00	T	0.98	T
16	226	0.27	F	0.20	F	1.00	T	0.97	T
17	226	0.17	F	0.25	F	0.40	G	0.01	G
18	226	0.10	F	0.32	C	0.60	F	0.98	T
19	226	0.34	F	0.37	F	0.04	C	0.96	T
20	226	0.39	F	0.55	N	0.31	C	0.95	T
21	226	0.27	F	0.36	N	1.00	T	0.00	G
22	226	0.31	F	0.30	F	0.86	G	0.06	G
23	225	0.25	F	0.35	N	0.63	F	0.99	T
24	226	0.18	F	0.28	F	0.44	N	0.97	T
25	226	0.07	F	0.22	N	1.00	T	0.99	T
26	226	0.10	F	0.36	F	1.00	T	0.07	G
27	226	0.22	F	0.50	F	0.37	F	0.98	T
28	226	0.22	F	0.20	N	0.39	C	0.10	G
29	226	0.21	F	0.20	F	0.64	C	0.15	G
30	225	0.09	F	0.12	C	0.61	F	0.34	G

Table 5.2 The *p* value and best fitting family in MCQM-III. The copula density function $\boldsymbol{c^1} = \boldsymbol{c}(\boldsymbol{U}, \boldsymbol{U}_{-i}, \boldsymbol{W})$ consists of three bivariate copulas $\boldsymbol{c^{11}} = \boldsymbol{c}(\boldsymbol{U}, \boldsymbol{W})$, $\boldsymbol{c^{12}} = \boldsymbol{c}(\boldsymbol{U}, \boldsymbol{U}_{-i})$ and $\boldsymbol{c^{13}} = \boldsymbol{c}(\boldsymbol{C}(\boldsymbol{U}_{-i}|\boldsymbol{U}), \boldsymbol{C}(\boldsymbol{W}|\boldsymbol{U}))$. The copula density function $\boldsymbol{c^2} = \boldsymbol{c}(\boldsymbol{V}, \boldsymbol{V}_{-i}, \boldsymbol{W})$ consists of three bivariate copulas $\boldsymbol{c^{21}} = \boldsymbol{c}(\boldsymbol{V}, \boldsymbol{W})$, $\boldsymbol{c^{22}} = \boldsymbol{c}(\boldsymbol{V}, \boldsymbol{V}_{-i})$ and $\boldsymbol{c^{23}} = \boldsymbol{c}(\boldsymbol{C}(\boldsymbol{V}_{-i}|\boldsymbol{V}), \boldsymbol{C}(\boldsymbol{W}|\boldsymbol{V}))$. The copula families are: N=Gaussian, T=Student's t, C=Clayton, G=Gumbel and F=Frank. Number of data denotes number of marginal probabilities of each variable used for fitting purposes and equals to the number of weather station measurements at each day in June during years 2004 to 2014.

					$\boldsymbol{c^1}$						$\boldsymbol{c^2}$		
Day	**Number of data**	***p***	$\boldsymbol{c^{11}}$	***p***	$\boldsymbol{c^{12}}$	***p***	$\boldsymbol{c^{13}}$	***p***	$\boldsymbol{c^{21}}$	***p***	$\boldsymbol{c^{22}}$	***p***	$\boldsymbol{c^{23}}$
1	226	0.38	F	0.00	G	0.63	N	0.27	F	0.99	T	0.81	N
2	224	0.40	F	1.00	T	1.00	T	0.32	F	0.99	T	0.87	N
3	226	0.32	F	0.99	T	1.00	T	0.33	F	1.00	T	0.77	N
4	226	0.23	F	0.26	G	0.76	N	0.25	F	1.00	T	0.76	N
5	226	0.44	F	0.98	T	1.00	T	0.20	F	1.00	T	0.71	N
6	226	0.36	F	0.88	F	0.67	N	0.20	F	0.63	F	0.99	T
7	226	0.23	F	0.98	F	0.49	3	0.20	F	0.62	F	0.91	N
8	225	0.34	F	0.90	F	0.41	N	0.37	F	1.00	T	0.85	N
9	226	0.42	N	0.49	F	0.14	N	0.27	F	0.54	F	0.94	N
10	226	0.48	N	0.00	G	0.44	C	0.27	F	0.35	G	0.76	N
11	226	1.00	T	0.96	F	0.57	C	0.31	F	0.68	G	0.75	N
12	226	0.31	F	0.99	T	0.99	T	0.20	F	1.00	T	0.62	N
13	226	0.26	F	1.00	T	1.00	T	0.26	F	1.00	T	0.52	N
14	226	0.42	N	0.98	F	0.56	N	0.35	F	1.00	T	0.67	N
15	226	0.45	N	0.98	T	0.52	C	0.14	F	1.00	T	0.85	N
16	226	0.35	F	0.97	T	1.00	T	0.27	F	1.00	T	0.91	N
17	226	0.20	F	0.03	G	0.44	N	0.18	F	0.44	G	0.93	N
18	226	0.35	C	0.98	T	1.00	T	0.21	F	0.64	F	1.00	T
19	226	0.46	F	0.96	T	1.00	T	0.31	F	0.02	C	0.78	G
20	226	0.48	N	0.95	T	1.00	T	0.29	F	0.35	C	0.65	G
21	226	0.52	N	0.01	G	1.00	T	0.24	F	1.00	T	0.71	N
22	226	0.32	F	0.09	G	0.59	N	0.26	F	0.83	G	0.31	N
23	225	0.40	N	0.99	T	1.00	T	0.26	F	0.62	F	0.62	N
24	226	0.26	F	0.97	T	1.00	T	0.22	F	0.56	N	1.00	T
25	226	0.13	N	0.99	T	0.51	C	0.11	F	1.00	T	0.70	N
26	226	0.29	F	0.04	G	0.51	C	0.08	F	1.00	T	1.00	T
27	226	0.39	F	0.98	T	0.99	T	0.27	F	0.40	F	0.88	N
28	226	0.15	N	0.12	G	0.57	C	0.24	F	0.27	C	0.75	G
29	226	0.22	F	0.12	G	0.52	N	0.18	F	0.62	C	0.75	G
30	225	0.23	C	0.38	G	0.64	N	0.20	F	0.54	F	0.99	T

5.4.2 Bias-corrected values

In the following, we present the bias-corrected values at the first station for all days in June 2014 (Figure 5.3) and at 1st June 2014 for all grid cells in the study area (Figure 5.4). Detailed comparisons for all days and all grid cells are given in appendix 5.2.

Time-series of the bias-corrected values obtained by MCQM-I at the first station (Figure 5.3a) showed that MCQM-I successfully corrects for bias at most of the days as well as the days with high extremes in comparison with time-series obtained by QM (Figure 5.3d). Mean absolute bias was equal to 4.52 ºC at this station. Mean absolute error and mean absolute prediction error were equal to 1.46ºC and 1.40ºC for MCQM-I, whereas for QM they were equal to 2.84ºC and 2.82ºC, respectively. MCQM-I resulted in a heterogeneous map at June 1st 2014 (Figure 5.4c) in comparison with the map obtained by QM (Figure 5.4f).

The spatial variation obtained by QM was similar to the spatial variation of the reanalysis air temperature as shown in Figure 4.7b due to the assumption of a perfect dependence between variables in QM. The visual comparison of the spatial variation of the elevation (see Chapter 3, Figure 3.3) with the spatial variation of the map obtained by MCQM-I (Figure 5.4c) revealed that this method was able to describe the covariability between the air temperature and the elevation. Mean absolute bias was equal to 2.83ºC at this day. Mean absolute error and mean absolute prediction error were equal to 2.07ºC and 1.55ºC for MCQM-I, whereas for QM they were equal to 2.62ºC and 1.93ºC, respectively.

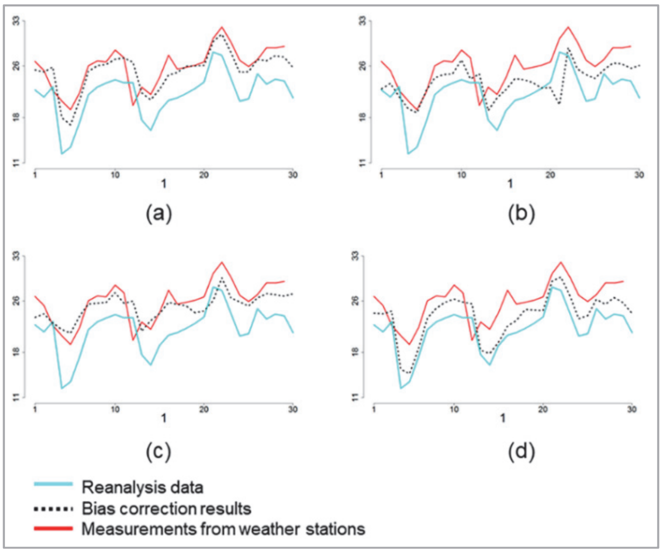


Figure 5.3 Time-series of the daily mean air temperature obtained from: weather stations, ECMWF ERA-interim reanalysis data, and bias correction methods at the first station in June 2014. a) MCQM-I, b) MCQM-II, c) MCQM-III, and d) QM. The vertical axis is daily mean air temperature.

Time-series of the bias-corrected values obtained by MCQM-II at the first station (Figure 5.3b) showed that MCQM-II successfully corrects for bias at most days except for days with extreme temperature in comparison with time-series obtained by MCQM-I and QM (Figure 5.3a and Figure 5.3d). Mean absolute error and mean absolute prediction error were equal to 2.62ºC and 2.67ºC for MCQM-II at this station, whereas for QM they were equal to 2.84ºC and 2.82ºC, respectively. MCQM-II resulted in a more heterogeneous map at June 1st, 2014 (Figure 5.4d) than the maps obtained by MCQM-I and QM (Figure 5.4c and Figure 5.4f). Mean absolute error and mean absolute prediction error were equal to 2.66ºC and 2.15ºC for MCQM-II at this day, whereas for QM they were equal to 2.62ºC and 1.93ºC, respectively.

Time-series of the bias-corrected values obtained by MCQM-III (Figure 5.3c) at the first station showed that MCQM-III performed better than MCQM-I

(Figure 5.3a) in correcting for bias at most days except for the days with extremes. Mean absolute error and mean absolute prediction error were equal to 1.77ºC and 1.68ºC for MCQM-III at this station, whereas for QM they were equal to 2.84ºC and 2.82ºC, respectively. The Figure 5.4e showed that MCQM-III resulted in a heterogeneous map as compared with the maps obtained by other methods at June 1st, 2014. Mean absolute error and mean absolute prediction error were equal to 2.36ºC and 1.84ºC for MCQM-III at this day, whereas for QM they were equal to 2.62ºC and 1.93ºC, respectively.

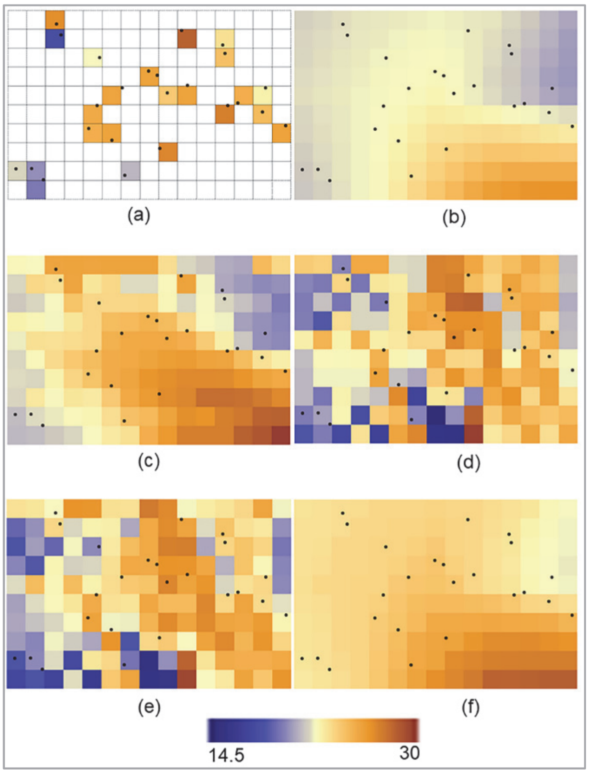


Figure 5.4 Daily mean air temperature obtained from: a) weather stations, b) ECMWF ERA-interim reanalysis data, and the bias correction methods at June 1st 2014; c) MCQM-I, d) MCQM-II, e) MCQM-III, and f) QM. For experimentation in our study, a sample subset of 10 × 15 grid cells of ECMWF dataset is selected at a spatial resolution of 0.125º Lat/Long.

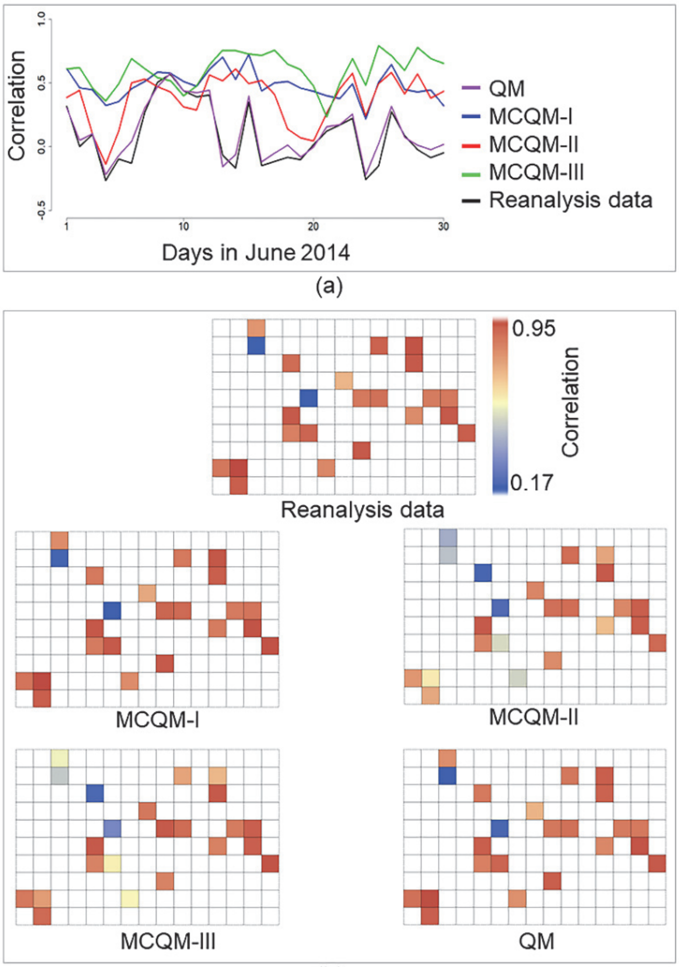
5.4.3 Evaluation and comparison

Leave-*K*-out cross-validation was carried out where *K* has values in the range of one to 11 denoting the number of measurements from a weather station at one day for 11 years. MCQM-III was superior to MCQM-I, MCQM-II, and QM as shown by MAE (Table 5.3). The average of absolute bias was equal to 3.6°C whereas MAE were slightly above 2°C. SES showed that MCQM-I resulted in more precise predictions in time, i.e., 30 days in June (Table 5.3, second column) whereas TES indicated that MCQM-III resulted in more precise predictions in space (Table 5.3, third column). To extend the evaluation of the bias correction methods beyond the cross-validation, we can perform a random split sampling validation in a well-monitored study area. It allows potentially more reliable uncertainty assessments. It is, however, beyond the scope of this paper. We treated the available measurements as benchmarks during the cross-validation. The horizontal distances, height differences and differences in land cover between the location of a station and the centre of a grid cell is associated with uncertainties.

Table 5.3 Total mean absolute error (MAE), spatial error scores (SES), temporal error scores (TES), spatial correlation scores (SCS), and temporal correlation scores (TCS), obtained by the quantile mapping (QM), and the multivariate quantile mappings (MCQM-I, MCQM-II and MCQM-III). The underlined values denote the best method.

Method	MAE	SES	TES	SCS	TCS
MCQM-I	2.23	51	58	77	85
MCQM-II	2.40	63	88	46	65
MCQM-III	2.13	54	38	61	112
QM	2.68	72	116	56	38

MCQM-I resulted in stronger correlations in time as shown by SCS (Table 5.3, fourth column) and correlations r_i (Figure 5.5b) whereas MCQM-III resulted in more strong correlations in space as shown by TCS (Table 5.3, last column) and correlations r_j (Figure 5.5a). A comparison based upon TCS showed that the new methods perform better than QM in correcting reanalysis air temperature at unvisited locations in a data-scarce area. It further revealed that MCQM-II including only one nearest neighbour was unable to represent the spatial variation of daily air temperature. In order to do so, MCQM-II needs to be extended towards more nearest neighbours allowing the use of a correlogram. A correlogram, however, faces the balancing issue between the number of spatial bins and the number of observations. The effect of the number of nearest neighbours on MCQM-II needs to be further investigated in a well-monitored area. Correlations r_j and r_i between the measurements and bias-corrected values obtained by QM were close to the correlations between the measurements and the reanalysis values (Figure 5.5a and Figure 5.5b). This was expected because of the assumption of a perfect dependence between variables in QM (see Section 5.2.3).



(a)

(b)

Figure 5.5 Correlations r_i and r_j that indicate temporal and spatial dependences between measurements and bias-corrected values, and between measurements and ECMWF ERA-interim reanalysis data. a) r_j at each day in June 2014, b) r_i at each weather station.

The previous comparisons showed the performance of the methods based upon an individual criterion. To evaluate the performance based upon all criteria, we ranked the methods in each column of Table 5.3 where the lowest rank value denotes the best method (Table 5.4). Then, the overall score based upon the sum of the rank values showed that MCQM-I, MCQM-II, and MCQM-III reduced bias with 58%, 16% and 63%, respectively as compared with QM (Table 5.4).

A practical advantage of MCQM-III is that it predicts the spatial variation of the bias-corrected air temperature maps in a data-scarce area (Appendix 5.2, Figure 5.8). The use of MCQM-III, however, is limited to the availability of the covariate at unvisited locations. We applied MCQMs to correct for bias in reanalysis air temperature, highlighting the potential of the methods for other weather data. Further comparison to other bias correction methods e.g. triple collocation analysis (Stoffelen 1998) might help to assess the performance of MCQMs.

Table 5.4 Overall score based upon Table 5.3. The methods are ranked based upon each criterion, i.e., each column in Table 5.3 where the lowest rank value denotes the best method. Then, an overall score based upon the sum of the rank values is obtained for each method. The underlined value denotes the best method.

Method	**Rank value based on**					**Overall score**
	MAE	**SES**	**TES**	**SCS**	**TCS**	
MCQM-I	2	1	2	1	2	8
MCQM-II	3	3	3	4	3	16
MCQM-III	1	2	1	2	1	7
QM	4	4	4	3	4	19

5.5 Conclusions

This study addressed bias correction in ECMWF reanalysis air temperature using its covariates in a data-scarce area. We developed three multivariate copula quantile mappings to do so. We concluded the following:

- The new methods are beneficial for the local refinement of reanalysis weather data at grid cells without weather station measurements.
- The new methods are advantageous as they can treat covariability, i.e., both weather data and covariates, and hence increase the precision of the mapping.

We see two ways to further extend the current study. First, we selected the number and type of covariates based upon our experience. A more general sensitivity analysis might help to show the effects of other covariates, e.g., land surface temperature and land cover. Second, it might be of interest to study the ability of the new methods to reproduce other statistical moments of the theoretical marginal distribution of air temperature. This could help to further model extremes in air temperature.

Appendix 5.1 Evaluating the stationarity assumption

To test for the assumption of second-order stationarity, we considered the null hypothesis H_0 as:

$$E[X_i] = \mu, \tag{5.16}$$

where X_i is a random variable at location i and *E[]* denotes the mathematical expectation. The alternative hypothesis H_1 is that there is a trend of degree one as:

$$E[X_i] = \beta_0 + \beta_1 . x_i' + \beta_2 . y_i', \tag{5.17}$$

where x_i' and y_i' are coordinates of location i and the β_j denote regression parameters. The parameters are estimated using a generalised linear model followed by their *p* values from a *t* test. We applied this trend to the measurements from 24 weather stations at each day of June 2014. The values of β_1 and β_2 were found to be not significantly different from zero, with their *p* values above 0.05 at most of the days (Figure 5.6). At the six days (out of 30 days) when the *p* value was below 0.05, it was still above 0.01. Based on this evidence, and the limited effects of including non-stationarity, we felt confident to assume stationarity.

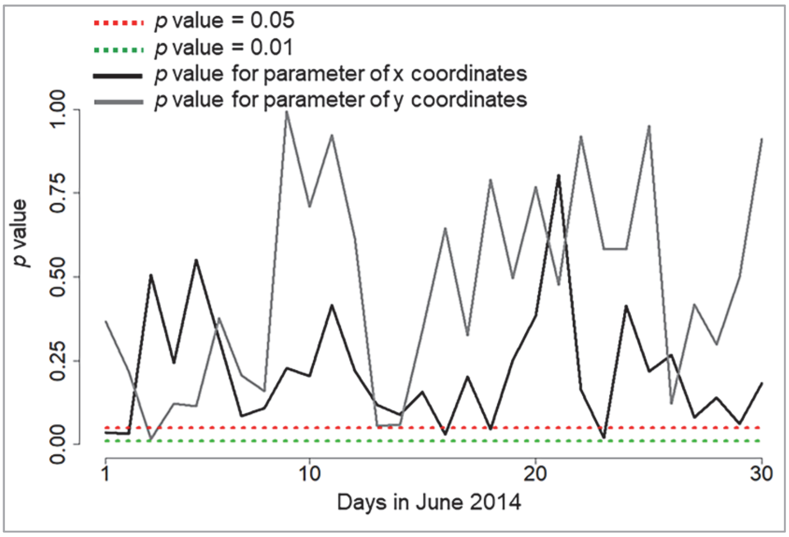
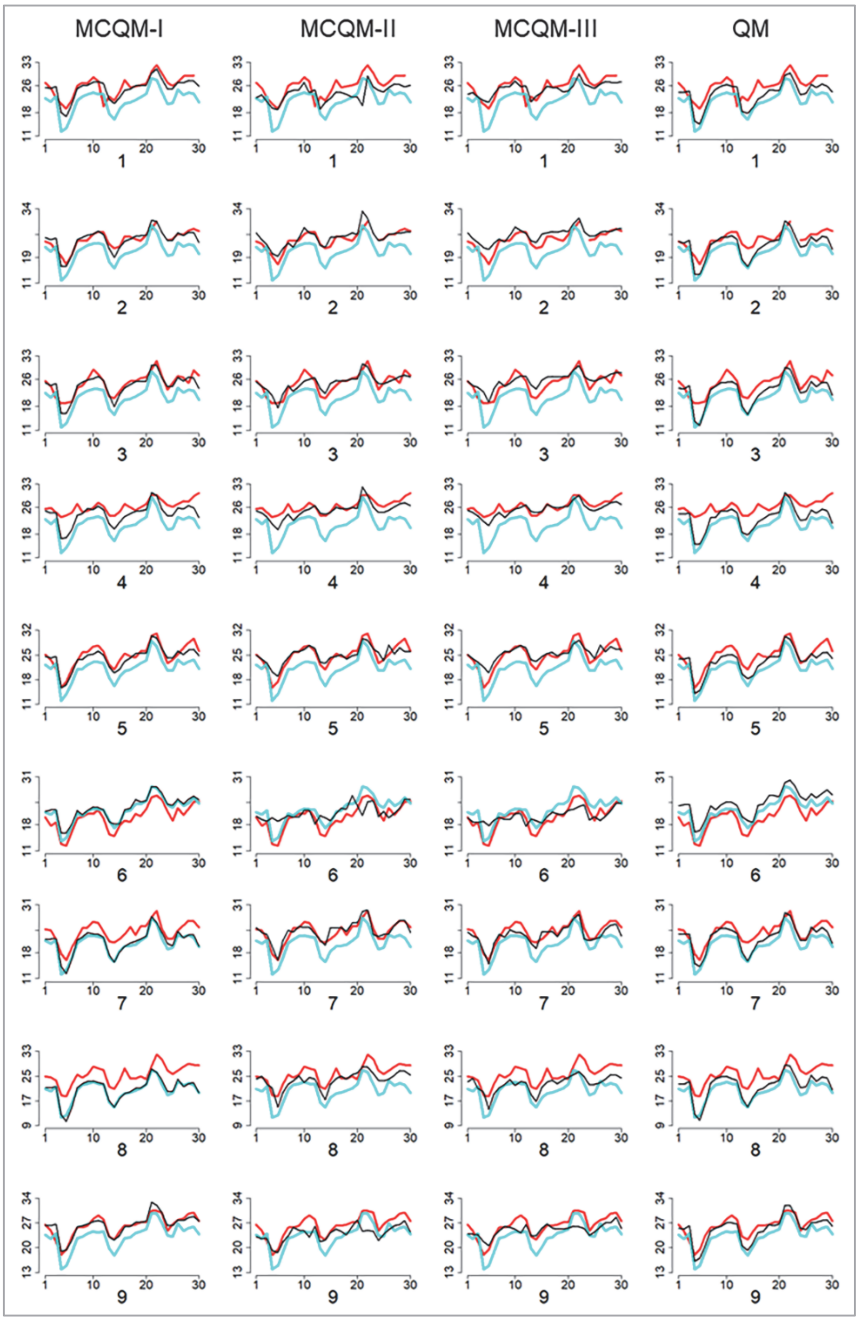
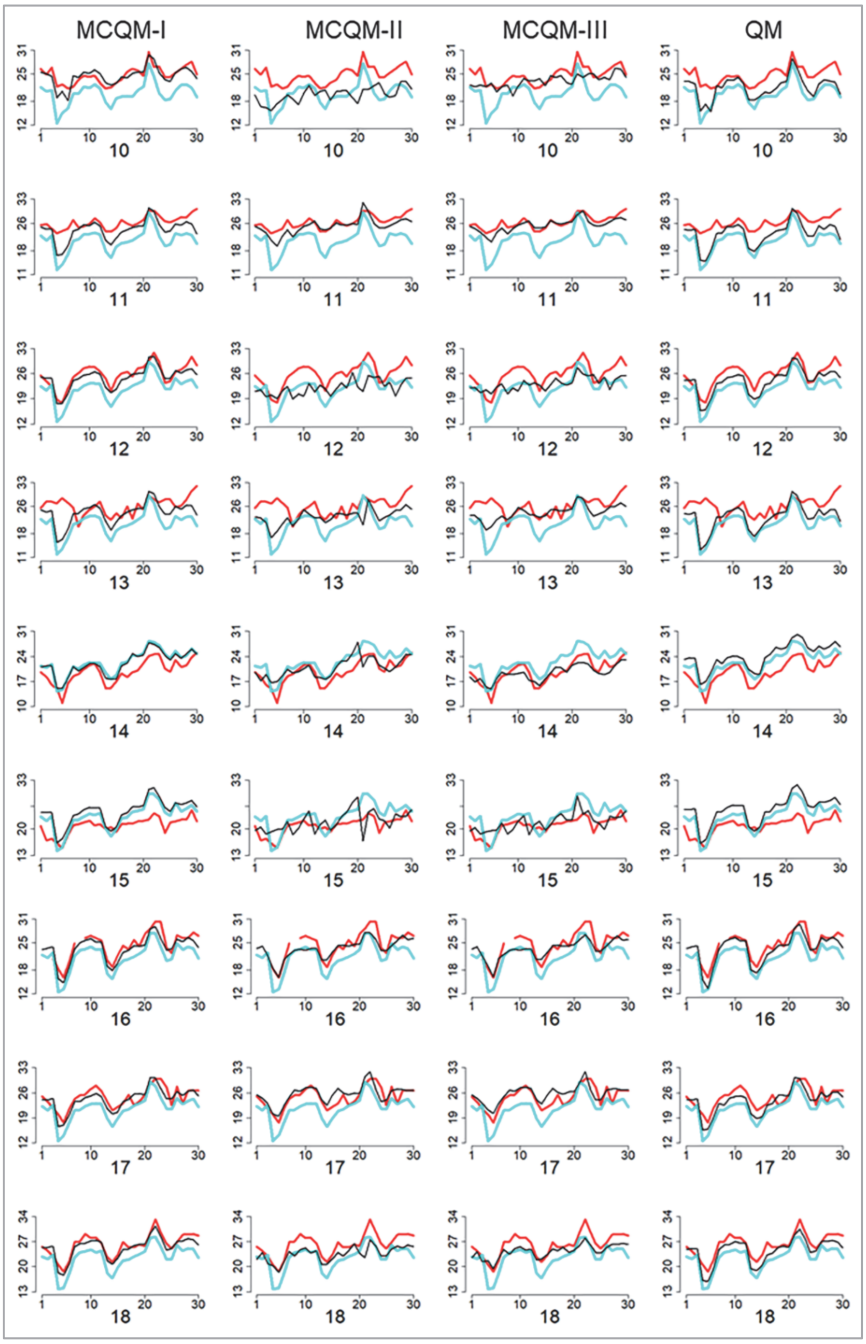


Figure 5.6 *p* values of the mean parameter in the trend analysis.

Appendix 5.2 Predictions in time and space




MCQM-I
MCQM-II
MCQM-III
QM
12 18 25 31
1 10 20 30
10
11 18 26 33
11
12 19 26 33
12
11 18 26 33
13
10 17 24 31
14
13 20 33
15
12 18 25 31
16
12 19 26 33
17
13 20 27 34
18

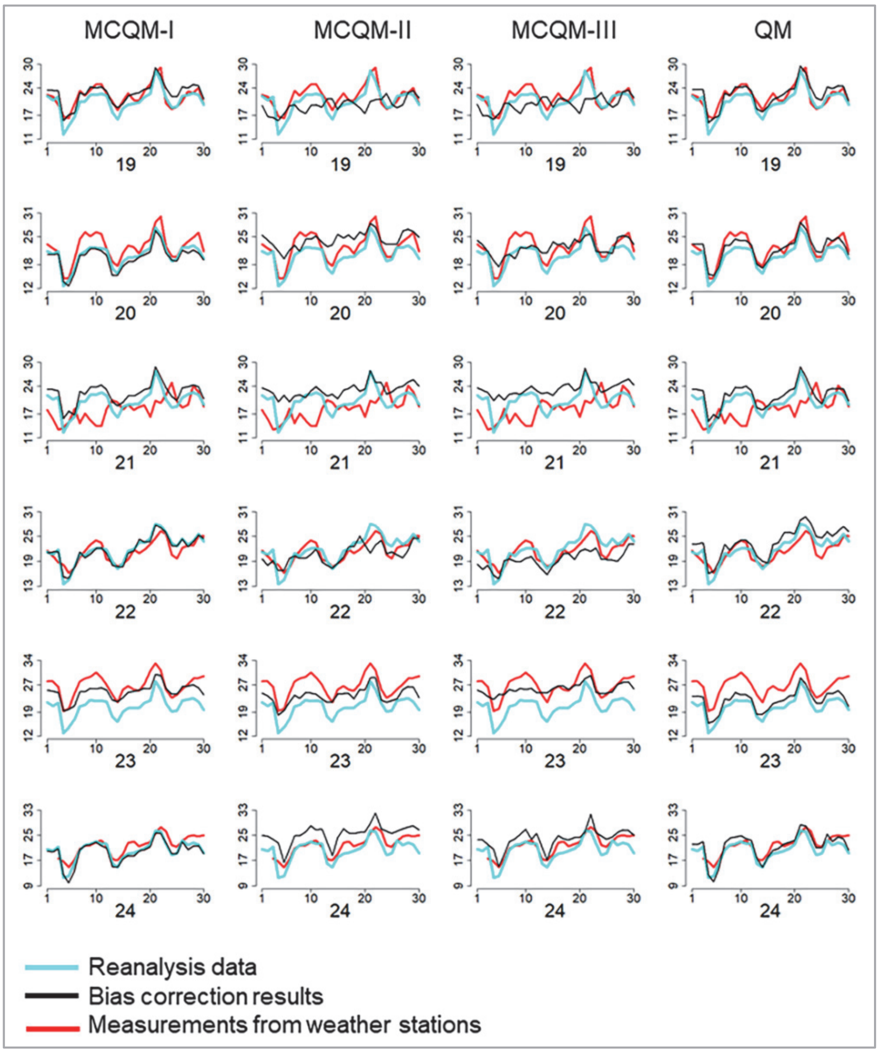
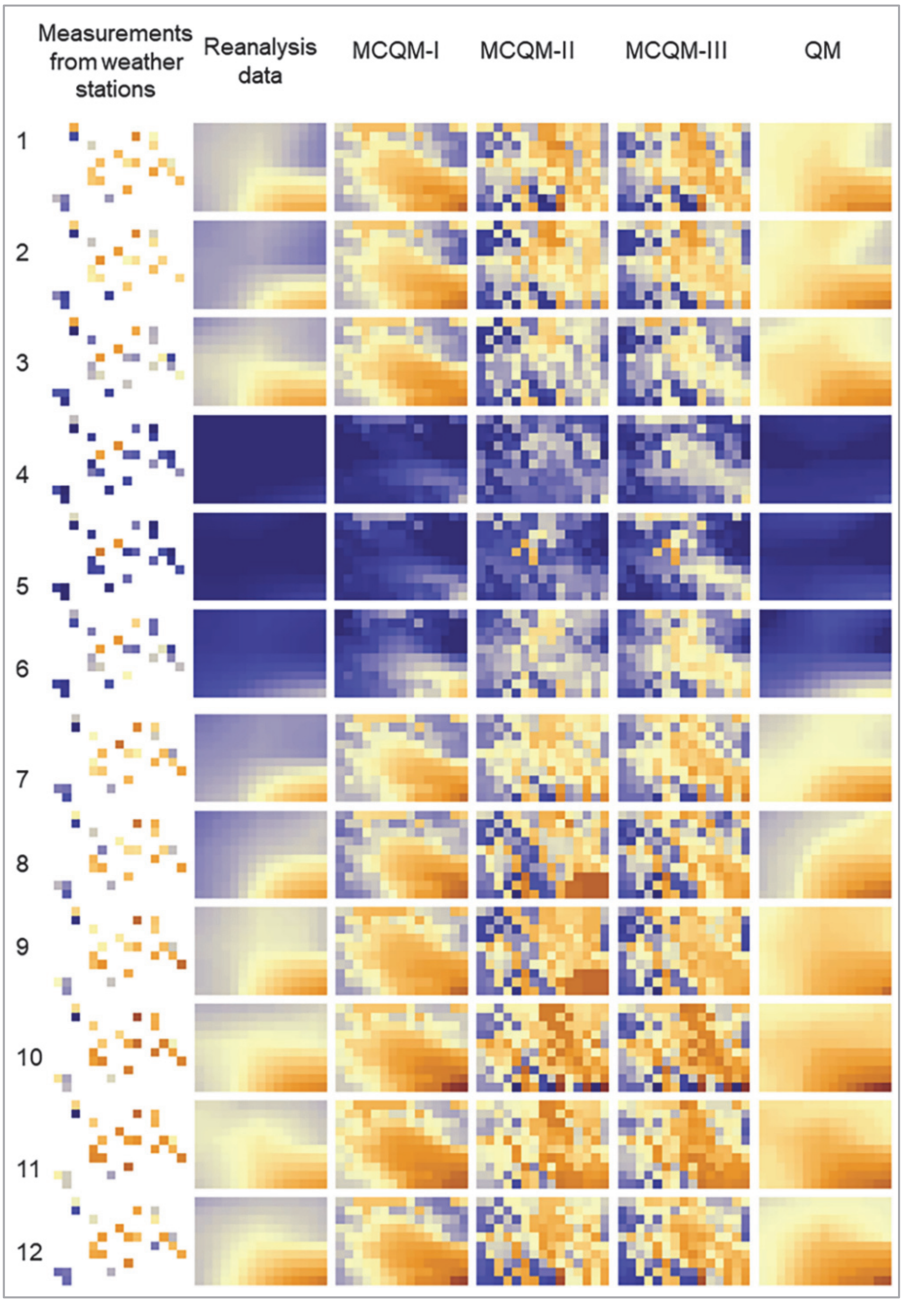
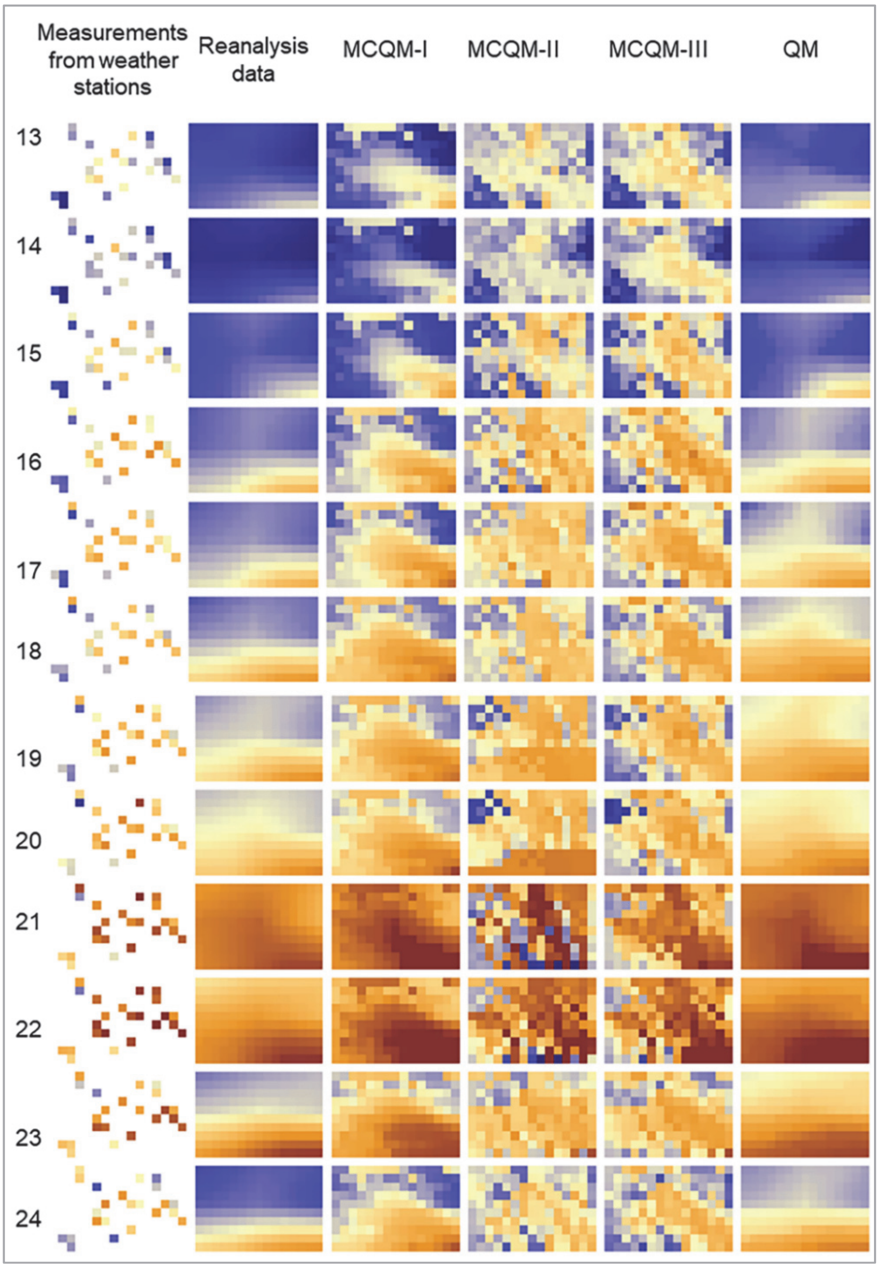


Figure 5.7 Time-series of the daily air temperature obtained from: weather stations, ECMWF ERA-interim reanalysis data, and bias correction methods, at each station in June 2014. The vertical axis is the daily mean air temperature. The number on each graph denotes the weather station number.


Measurements from weather stations
Reanalysis data
MCQM-I
MCQM-II
MCQM-III
QM
1
2
3
4
5
6
7
8
9
10
11
12


Measurements from weather stations
Reanalysis data
MCQM-I
MCQM-II
MCQM-III
QM
13
14
15
16
17
18
19
20
21
22
23
24

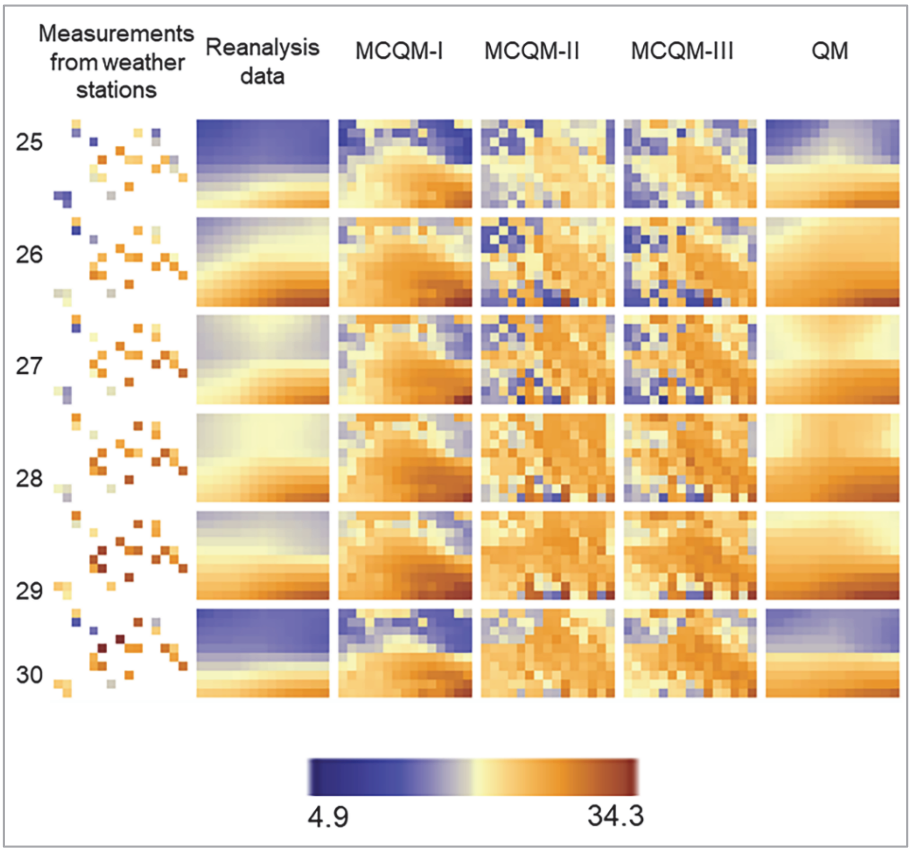
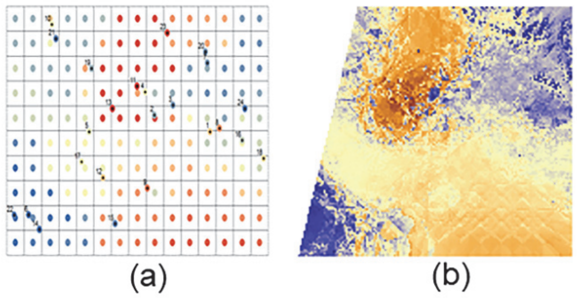


Figure 5.8 Daily mean air temperature obtained from: weather stations, ECMWF ERA-interim reanalysis data, and bias correction methods, at each day in June 2014. For experimentation in this study, a sample subset of 10 × 15 grid cells of ECMWF dataset is selected at a spatial resolution of 0.125º Lat/Long.

Chapter 6: Copula-based interpolation methods using collocated covariates



a) Observations
b) Interpolation results

This chapter is published as: Alidoost F., Stein A., Su Z. (2018). Copula-based interpolation methods for air temperature data using collocated covariates. *Spatial Statistics*, https://doi.org/10.1016/j.spasta.2018.08.003.

Abstract

This paper introduces two copula-based interpolation methods to produce air temperature maps in a data-scarce area: a spatial copula interpolator including covariates, and a mixed copula interpolator. The methods allow a construction of the conditional distribution of air temperature given the collocated covariates. Our study compared the new methods with the spatial copula interpolator, the ordinary kriging predictor and the co-kriging predictor. Daily mean air temperature was used from weather stations and ERA_Interim reanalysis weather data at 174 locations in the Qazvin Plain, Iran. Spatial copula interpolator including covariates resulted in more precise predictions as shown by leave-two-out cross-validation. Visual inspection of air temperature maps demonstrated that the new methods well represented spatial variability of air temperature at a spatial resolution of 1 km. The results showed an improved performance of the new methods to describe both spatial variability and covariability between variables. The methods are potentially useful for other sparsely and irregularly distributed weather data.

Keywords

copula, interpolation, covariate, data scarce, air temperature

Author contributions

F.A. conceived and designed the analysis, collected and processed the data, developed tools, performed the analysis, wrote the manuscript, is the corresponding author.

A.S. supervised the findings of this work, verified the analytical methods, encouraged A.F. to investigate copulas for interpolation, improved the English wording.

Z.S. supervised the findings of this work.

All authors contributed to the interpretation of the results, and commented on the final manuscript.

Structure of the chapter

After an introduction in section 6.1, the copula-based interpolation methods are presented in section 6.2, the study area and data are introduced in section 6.3, the results are described in section 6.4, and the discussion and conclusion are in section 6.5.

6.1 Introduction

A copula is a multivariate joint distribution that describes the dependence structure between variables (Nelsen, 2006). The joint distribution is estimated using a distribution family that can be different from the family of the marginal distributions of the involved variables. An appealing property of a copula in describing spatial dependences is that its parameter is estimated by means of a correlogram that describes dependences based upon the correlation between marginals (Oden, 1984; Gräler and Pebesma, 2011). The purpose of copula-based interpolation methods is to predict the marginal probability at an unvisited location given the marginal probabilities of the nearest neighbours.

Recently, several studies have assessed the performance of copula-based interpolation methods and compared with kriging methods (Bárdossy and Li, 2008; Haslauer et al., 2016; Heißerer et al., 2016; Durocher et al., 2016). Gräler and Pebesma (2011) investigated the application of a multivariate copula that models spatio-temporal random fields using vine structures in interpolation of daily mean PM_{10} measurements. Gräler (2014) demonstrated the potential of spatio-temporal copula interpolation with a single covariate.

This study focuses on mean air temperature. The use of these data in hydrological models, e.g., crop growth simulations for assessing crop water requirement has been the key to support irrigation management. Application of hydrological models at unvisited locations remains a challenge because weather stations are usually sparse and located at irregular positions. A solution to this problem is to use gridded air temperature data from a weather forecast system. The coarse spatial resolution of those data, however, is a source of uncertainty because of the spatial variability at local scales (Aalto et al., 2013). Hence, interpolation has to take place to predict air temperature at unvisited locations.

Kriging is a well-established interpolation method (Cressie, 1993). Since the twentieth century, a variety of methods has been developed. To predict air temperature values at a spatial resolution of 1 km, typical examples are spatio-temporal regression-kriging with incorporation of remote sensing images (Hengl et al., 2012), generalized additive models (Aalto et al., 2013), and residual kriging and regression kriging methods using auxiliary maps (Wu and Li, 2013; Kilibarda et al., 2014). Parmentier et al. (2015) compared universal kriging, generalized additive models and geographically weighted regression. Kilibarda et al. (2014) showed the effect of daily land surface temperature on both minimum and maximum air temperature variability.

With the aim to improve prediction of air temperature in a data-scarce area, we present two interpolation methods based upon copulas: a spatial copula interpolator including covariates, and a mixed copula interpolator. The first method considers two types of dependences: spatial dependences of air temperature at a single location and its nearest neighbours, and non-spatial dependences between air temperature and its collocated covariates at that location. The second extends the first method by including the non-spatial dependences of air temperature and its collocated covariates at the nearest neighbours. The two methods are compared with the spatial copula interpolator, the ordinary kriging and co-kriging predictors.

6.2 Interpolation methods

6.2.1 Spatial copula interpolator

Let $f(X|X = x_1, \dots, X = x_n)$ be the conditional density distribution of the variable X at an unvisited location conditioned on its n nearest neighbours. The conditional expectation is the optimal predictor to derive the value of the variable X at an unvisited location, denoted by $\hat{x}_0$. It can be shown that it minimizes the Bayes risk (Cressie, 1993). The conditional expectation can be either linear or nonlinear in X and it can be written using the conditional copula density function $c(U|U = u_1, \dots, U = u_n)$ as:

$$\hat{x}_0 = E[X|X = x_1, \dots, X = x_n] = \int_x x \cdot f(X|X = x_1, \dots, X = x_n)dx$$
$$= \int_0^1 F^{-1}(u) \cdot c(U|U = u_1, \dots, U = u_n)du\,, \tag{6.1}$$

where F is the marginal cumulative distribution function, i.e., $u = F(x)$ (Bárdossy and Li, 2008). The predictor has two main parts: a marginal distribution $F(.)$ and a multivariate copula $c(.|.)$. The last equality in (1) can be proven by:

$$f(X|X = x_1, \dots, X = x_n) = \frac{f(x_0, \dots, x_n)}{f(x_1, \dots, x_n)} = \frac{c(u_0, \dots, u_n) \cdot f(x_0) \cdot \dots \cdot f(x_n)}{c(u_1, \dots, u_n) \cdot f(x_1) \cdot \dots \cdot f(x_n)} =$$
$$c(U|U = u_1, \dots, U = u_n) \cdot f(x_0) = c(U|U = u_1, \dots, U = u_n) \cdot \frac{du}{dx}, \tag{6.2}$$

where $c(u_0, \dots, u_n)$ and $c(u_1, \dots, u_n)$ are the copula density functions. The choice of a Gaussian distribution for f in (6.1) leads to a linear predictor that is the equivalent to the simple kriging predictor (Cressie, 1993). Such a predictor is able to capture extremes if it is based upon local nearest neighbours rather than a large set of neighbouring observations.

Following section 2.2, the joint density function $c(u_0, \dots, u_n)$ with *m=n+1* variables, it is decomposed into *m.(m-1)/2* bivariate copulas (Gräler, 2014) based on a canonical vine or C-vine structure (Aas et al., 2009). For *m*=3, $c(u_0, u_1, u_2)$ is decomposed as

$$c(u_0, u_1, u_2) = c(u_0, u_1) \times c(u_0, u_2) \times c(C(u_1|u_0), C(u_2|u_0)), \tag{6.3}$$

where $C(.|.)$ is the conditional copula. The first tree in the vine structure consists of spatial bivariate copulas, e.g., $c(u_0, u_1) \times c(u_0, u_2)$, taking the influence of the neighbours into account. The parameter of the spatial bivariate copula is obtained from the correlogram obtained with binned data pairs (Gräler, 2014). Pairs with distances larger than the distance in last spatial bin are considered independent and are described by the Product copula family (Nelsen, 2006). A polynomial of degree two fitted to Kendall's τ values estimates the correlation function. The remaining trees in the vine structure consist of non-spatial bivariate copulas, e.g., $c(C(u_1|u_0), C(u_2|u_0))$.

6.2.2 The spatial copula interpolator including covariates

To introduce the spatial copula interpolator including covariates, we consider one variable X and two covariates, e.g., Y and Z. The aim is to predict $\hat{x}_0$ with a finite sample of X. Samples of Y and Z are available at all locations. The conditional copula density function in (6.1) is then written as $c(U|U = u_1, \dots, U = u_n, V = v_0, W = w_0)$, where $v_0 = F_Y(y_0)$, $w_0 = F_Z(z_0)$, $_0$ denotes an unvisited location, F_Y and F_Z are marginal distribution functions of the covariates. The mean predictor in (6.1) equals:

$$\hat{x}_0 = \int_0^1 F^{-1}(u) \cdot c(U|U = u_1, \dots, U = u_n, V = v_0, W = w_0) du \,. \qquad (6.4)$$

By conditioning on V and W, the collocated covariates at an unvisited location, i.e., v_0 and w_0 are incorporated to the predictor. The conditional distribution can be extended to higher dimensions by including more than two covariates in a straightforward way.

In this study, we will use the empirical marginal probability u_i at location i is defined using the following rank-order-transformation $u_i = \frac{\mathrm{rank}(x_i)}{N+1}$, where N denotes the total number of observations. A similar transformation is also applied to y_i and z_i. The empirical marginal distribution avoids using the theoretical marginal distributions that might affect the estimation of copula parameter. By means of kernel density estimation, a continuous approximation of the marginal distribution F is obtained under the assumption of stationary (Silverman 1986). Note that the empirical probabilities are limited to observations and therefore, the interpolation methods are unable to predict extreme values outside the range of the observations.

6.2.3 Mixed copula interpolator

Next we introduce the second method, the mixed copula interpolator. The conditional distribution of X conditioned on Y and Z at location i is equal to $C(U|V = v_i, \mathrm{W} = w_i)$, where C is a conditional copula, $v_i = F_Y(y_i)$ and $w_i = F_Z(z_i)$. The conditional probability p_i equals:

$$p_i = C(u_i | V = v_i, \mathrm{W} = w_i). \tag{6.5}$$

The conditional copula C is estimated in a similar way for the spatial copula interpolator including covariates. The conditional probability p_i is used as the probability of nearest neighbour i for copula in (6.4) and the final form of the predictor equals:

$$\hat{x}_0 = \int_0^1 F^{-1}(u) \cdot c(U | U = p_1, \dots, U = p_n, V = v_0, W = w_0) du\,. \tag{6.6}$$

Hence, the collocated covariates at the nearest neighbour, i.e., y_i and z_i are incorporated into the predictor. The conditional distribution can be extended to higher dimensions for including more than two covariates.

6.2.4 Comparison and evaluation of the interpolation methods

We compare the spatial copula interpolator including covariates (6.4) and the mixed copula interpolator (6.6), with the spatial copula interpolator (6.1), the ordinary kriging predictor (Cressie, 1993) and the co-kriging predictor (Stein and Corsten, 1991). We treat available observations from *n* weather stations as benchmarks for leave-k-out cross-validation to quantify the performance of the interpolation methods. To this end, *k* stations are removed from the *n* weather stations and predictions $\hat{x}_i, i = 1,..,k$ are obtained using observations from the reminder of the stations. Each interpolator is then applied on $m = \frac{n!}{k!(n-k)!}$ replications of dependence structures. The mean absolute error (MAE) and error score (ES) (Durai and Bhradwaj 2014) are determined as:

$$MAE = \frac{1}{m} \sum_{j=1}^{m} \left(\frac{1}{k} \sum_{i=1}^{k} |x_{ij} - \hat{x}_{ij}| \right), \tag{6.7}$$

$$ES = rank(MAE), \tag{6.8}$$

for each method. The smallest ES indicates the best interpolator. The overall prediction quality depends upon a good estimation of the copula and the marginal distributions as well as the number of the observations.

The coverage of 90%, 95% and 99% prediction intervals from the conditional distributions $F(X|.)$ are investigated at each weather station. The number of observed values that fall in the intervals provides insight into the performance of the copula-based methods. This should be interpreted with care, because the type and number of covariates can be different in the copula-based methods. In addition, spatial variation of mean and standard deviation of the conditional distributions are compared at each weather station.

A 95% prediction interval width (PIW) at an unvisited location is obtained as $PIW_0 = F^{-1}\big(C^{-1}(0.975|.)\big) - F^{-1}\big(C^{-1}(0.025|.)\big)$, describing the uncertainty of the predictions (Li, 2010). The kriging methods result in the prediction error variance s_0^2 (Cressie, 1993; Kutner et al., 1996). A 95% PIW at unvisited location under the assumption of a Gaussian joint distribution is obtained as $PIW_0 = (\hat{x}_0 + 1.96 \cdot s_0) - (\hat{x}_0 - 1.96 \cdot s_0)$.

The methods were implemented in R using the packages gstat (Pebesma, 2004), copula (Kojadinovic and Yan, 2010), spcopula (Gräler and Pebesma, 2011), and VineCopula (Brechmann and Schepsmeier, 2013). We contributed to spcopula and VineCopula packages in R to interpolate the random field spatially including more than one covariate.

6.3 Application: mean air temperature in Iran

We applied the interpolation methods to mean air temperature in the Qazvin plain, Iran on June 6th and 22nd 2014 denoted by d_6 and d_{22}, respectively (see Section 3.1). These two days were selected as these were Landsat 8 overpass days and thus provided three covariates for the 19 of the 24 weather stations: land surface temperature (LST), leaf area index (LAI) and SRTM elevation (see Section 3.1). Five weather stations were outside the coverage of Landsat 8 images (Zanter, 2016). Investigating the correlations (Table 6.1), we ignored LST as a covariate at d_{22}. The covariates are at different spatial resolutions. Throughout we maintained a resolution of 1 km that represents spatial variation of air temperature (Figure 6.1).

Table 6.1 Correlations between mean air temperature and its covariates on d_6 and d_{22}. The temperature values are the combination of bias-corrected values and measurements from weather stations. The covariates are elevation, land surface temperature (LST) and leaf area index (LAI).

	Elevation	LST	LAI
Mean air temperature on d_6	-0.25	0.24	-0.23
Mean air temperature on d_{22}	-0.26	-0.02	-0.23

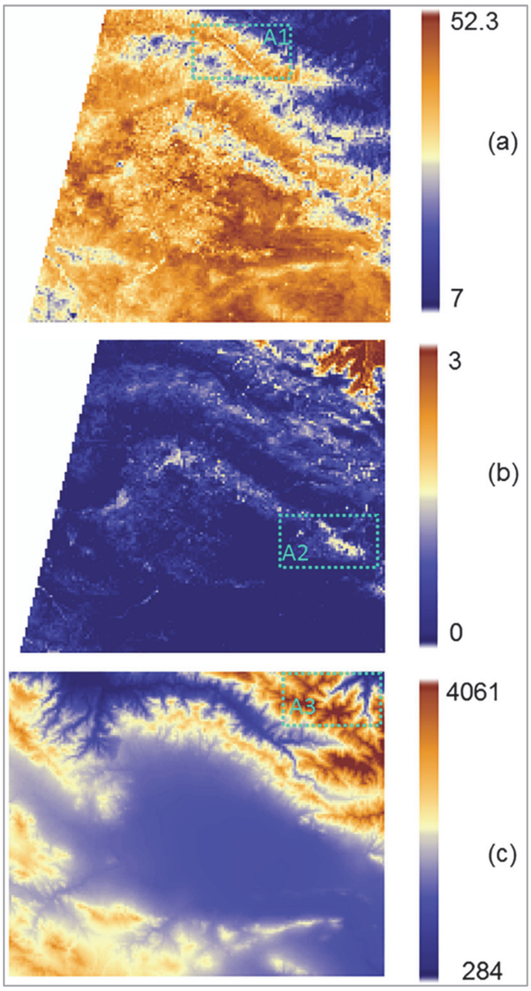


Figure 6.1 Three covariates for air temperature at a resolution of 1km. a) LST and b) LAI are obtained using Landsat 8 bands. c) Surface elevation is obtained from the SRTM dataset. The areas A1, A2 and A3 are selected to investigate the covariability of the air temperature.

We defined bias as a systematic overestimation and underestimation of reanalysis weather data with respect to measurements (Persson, 2013; Mao et al., 2015). The average bias for all stations equals 3.9°C and 3.4°C at d_6 and d_{22}, respectively. We applied a bias correction method to obtain bias corrected values (Alidoost and Stein, 2016). A two-sample Kolmogorov-Smirnov test was performed of the null hypothesis that measurements and bias corrected values are drawn from the same distribution. The p values were equal to 0.22 and 0.65 at d_6 and d_{22}, respectively, did not reject the null hypothesis. Based upon these results, we used a combination of measurements and the bias-corrected values as observations for fitting purposes in the interpolation methods (Figure 6.2).

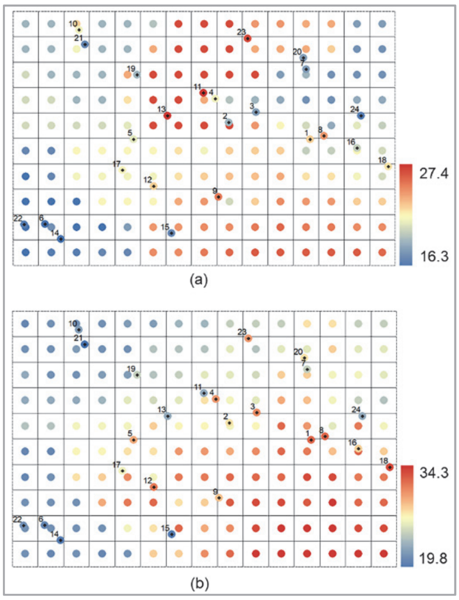


Figure 6.2 Spatial variation of mean air temperature at 174 locations from the weather stations and the bias corrected reanalysis weather data on d_6 (a) and d_{22} (b).

6.4 Results

6.4.1 Distribution of the observations

The empirical marginal distribution is shown in Figure 6.3. The number of observations in the tails of the distributions was low, i.e., there were two extremes in the upper tail at d_6 and one in the lower tail at d_{22} (Figure 6.3). For copula-based interpolators, in contrast to kriging predictors, it is a challenge to estimate a skewed marginal distribution with two extreme values out of 174 observations. They are not able to predict the extremes in leave-k-out cross validation for $k \geq 2$. Hence, the marginal distribution function has to be well estimated.

Figure 6.4 shows the fit to Kendall's τ values in the correlogram for six and five spatial bins at d_6 and d_{22}, respectively. Apparently, the correlogram changes over the range of [-0.2, 0.7] describing the positive and negative dependences. The Student t and Clayton copulas are selected according to the lowest AIC values at each bin at d_6 whereas Student t and Gumbel copulas are selected at d_{22}.

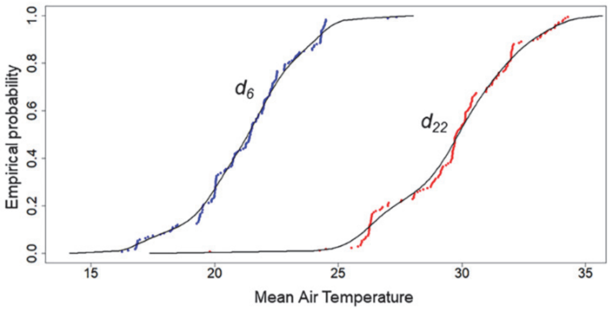


Figure 6.3 Empirical marginal probabilities obtained on *d₆* and *d₂₂*. The empirical marginal distribution function is obtained using kernel density estimation.

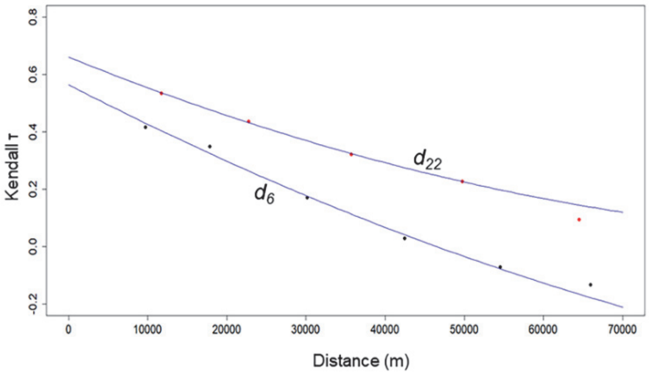


Figure 6.4 Kendall's τ is obtained using observations at 174 locations on d_6 and d_{22} . A polynomial function is fitted to obtain τ at each distance. The parameters of five spatial bivariate copulas are then estimated by maximum likelihood. The best fitting copula is selected according to the lowest AIC values at each bin.

The multivariate distributions were estimated using the C-vine structures and the conditional cumulative probabilities $F(X|.)$ for 19 weather stations are shown in Figure 6.5. The number of observed values that fall within the 90%, 95% and 99% prediction intervals for spatial copula interpolator using covariates were equal to 15, 17 and 19 whereas for mixed copula interpolator were equal to 14,17 and 18, respectively (Table 6.2). Hence, it showed a good performance of the methods in fitting of the distributions.

For the ordinary kriging, the variogram is obtained for the same number of spatial bins as the correlogram, followed by fitting a Gaussian variance function to the variogram of the mean air temperature (Figure 6.6). We applied the co-kriging in this study based upon the proportional model using the same variance and covariance functions. Gaussian covariance functions were fitted to cross variograms obtained for air temperature and its covariates.

Table 6.2 The number of observed values that fall in the 90%, 95% and 99% prediction intervals of the conditional cumulative probabilities $\boldsymbol{F}(\boldsymbol{X}|.)$ for 19 weather stations on d_6. The observed values are the measurements from weather stations. The covariates are elevation, land surface temperature (LST) and leaf area index (LAI).

	Prediction interval		
	90	**95**	**99**
Spatial copula interpolator using covariates	15	17	19
Mixed copula interpolator	14	17	18
Spatial copula interpolator	15	19	19

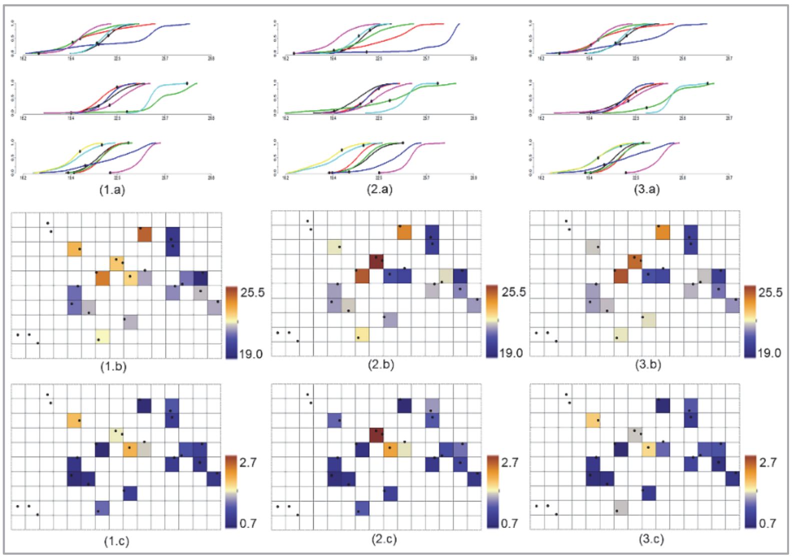


Figure 6.5 a) The conditional cumulative probabilities $\boldsymbol{F}(\boldsymbol{X}|.)$ for 19 weather stations and the spatial variation of b) mean and c) standard deviation of the conditional distributions of the predictions on d_6. The observed values in the conditional cumulative distributions are denoted by black dots.

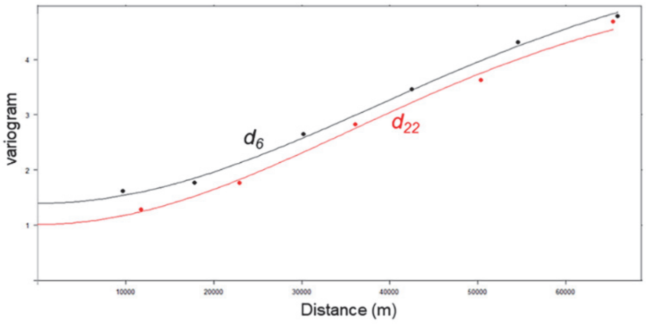


Figure 6.6 The variogram obtained on d_6 and d_{22} for the same number of spatial bins as the correlogram. A Gaussian variogram model is fitted.

6.4.2 Evaluation and comparison

For the leave-*k*-out cross-validation, we took *k=2* due to low number of the weather stations. The spatial copula interpolator including covariates resulted into the lowest MAE using different covariates at the two days (Table 6.3). The ES shows that the spatial copula interpolator including covariates improved predictions of the mean air temperature with 58% comparing with the co-kriging predictor (Table 6.3). In addition, cross-validation showed that the use of LAI as a covariate resulted into more precise predictions.

Table 6.3 Cross-validation expressed as the mean absolute error (MAE) obtained by the spatial copula interpolator using covariates, the mixed copula interpolator, the spatial copula interpolator, the ordinary kriging predictor, and the co-kriging predictor. The leave-two-out cross-validation is done for combinations of the covariates, i.e., elevation (E), land surface temperature (LST) and Leaf area index (LAI) at two days. To compare the five methods, an error score (ES) is obtained based upon MAE for each method. The smallest ES indicates the best interpolator.

Day	Covariate	Spatial copula interpolator using covariates	Mixed copula interpolator	Spatial copula interpolator	Ordinary kriging	Co-kriging
6	E	1.550	1.669	1.555	1.597	1.598
	LAI	1.503	1.525	1.555	1.597	1.595
	LST	1.557	1.611	1.555	1.597	1.599
	E, LAI	1.480	1.633	1.555	1.597	1.596
	E, LST	1.529	1.654	1.555	1.597	1.598
	LST, LAI	1.531	1.735	1.555	1.597	1.599
	E, LST, LAI	1.551	1.833	1.555	1.597	1.597
22	E	1.390	1.365	1.378	1.328	1.330
	LAI	1.291	1.308	1.378	1.328	1.328
	E, LAI	1.301	1.360	1.378	1.328	1.331
ES		**15**	41	28	30	36

6.4.3 Prediction

For making predictions, we considered n = 8 nearest neighbours and three covariates at d_6. The mixed copula interpolator was able to capture extremes (Figure 6.7.b) in contrast to the spatial copula interpolator including covariates (Figure 6.7.a). The vegetated and non-vegetated areas based upon LAI, and highest and lowest elevated areas (Figure 6.1) are interesting areas to consider to which degree the new interpolation methods take the spatial variation of covariates into account. The new methods resulted in the most heterogeneous map and, visually, more realistic spatial patterns than the kriging methods and the spatial copula interpolator, whereas the latter was more heterogeneous (Figure 6.7.c) than the map obtained by the kriging predictors (Figure 6.7.d and Figure 6.7.e). We further note that a low number of spatial bins leads to unrealistic spatial patterns as shown in Figure 6.7.

The boxplots (Figure 6.8) shows that copula-based methods well represent the mean values of the observations. The issue of failing to represent extremes using copula-based methods is related to the low number of observations in the tails of the marginal distribution. The ranges of 95% PIW for copula-based methods are equal to [0.6, 12.6]°C whereas for kriging methods are equal to [4.9, 6.3] °C (Figure 6.9).

The spatial variation of the mean and standard deviation of the conditional distributions (Figure 6.5) shows that there is no reason to assume any lack of homogeneity. In fact, any pattern in the standard deviations would indicate such lack of homogeneity. Values, however, are relatively low as compared to the standard deviation of the observations (2.9°C). A few relatively high values occur in the centre of the study area. These are caused by the presence of extreme values at locations covered by the same pixels.

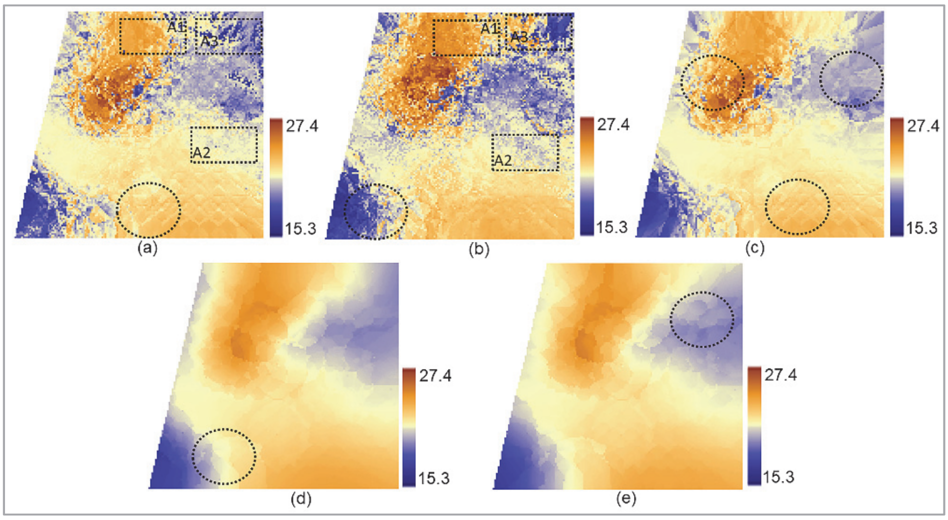


Figure 6.7 Daily mean air temperature predicted at a spatial resolution of 1 km on d_6 based upon a) the spatial copula interpolator including covariates, b) the mixed copula interpolator, c) the spatial copula interpolator, d) the ordinary kriging predictor, e) the co-kriging predictor for a neighbourhood of eight locations. The circled areas denote squares as artefacts that represent the unrealistic spatial patterns. The areas A1, A2 and A3 as shown in Figure 6.1, are examples where the covariability becomes apparent in the results of the new methods.

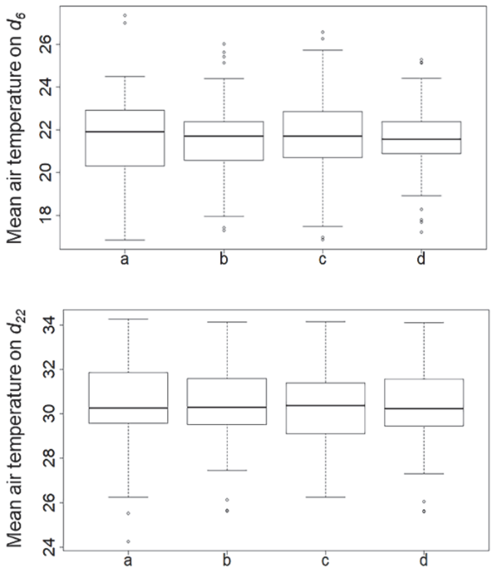


Figure 6.8 Boxplots comparing the observations (a) with predicted values by: b) the spatial copula including three covariates, c) the mixed copula, and d) the spatial copula interpolator. Here, observations are a combination of bias-corrected values and measurements from the weather stations on d_6 and d_{22}.

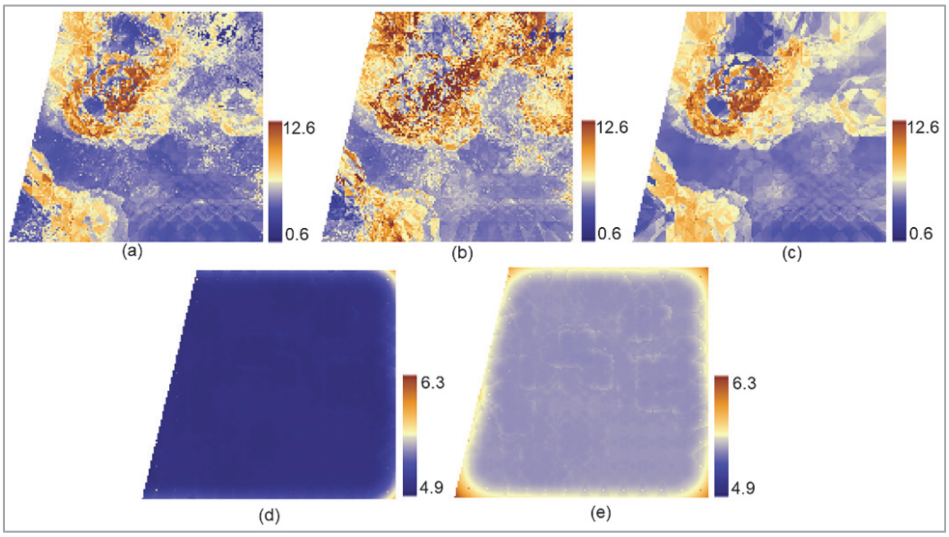


Figure 6.9 95% prediction interval widths (PIW) for each interpolation method on d_6, a) the spatial copula including three covariates, b) the mixed copula, c) the spatial copula, d) the ordinary kriging, and e) the co-kriging. The spatial copula interpolator resulted in the lowest uncertainty among copula-based methods. Ordinary kriging has smaller PIWs and is based upon assuming a Gaussian joint distribution.

6.5 Discussion and conclusion

Two dependences were characterized in spatial interpolation of a weather variable: spatial variability and dependency with other variables, i.e., covariability. We developed two methods based upon spatial copulas of air temperature, and non-spatial dependences between air temperature and its collocated covariates. The multivariate distributions are decomposed into bivariate copulas using vine structures that are generally well understood and can be estimated in a straightforward way.

The new methods provide more information about the uncertainty when interpreting the spatial variability of the PIW. We proposed to estimate the empirical marginal distribution that describes the statistical properties of daily air temperature without the knowledge of the theoretical form of the family's distribution function. The marginal distribution is, however, still assumed to be stationary. The local marginal distribution at an unvisited location (Heißerer et al., 2016) might help to improve the prediction as well as the PIW.

We treated the available observations from weather stations as benchmarks during cross-validation, but we realized that the quality of measurements differs at each station. For example, stations 11 and 13 represent high extremes relative to other stations at the same day. A time-series analysis of the air temperature (not shown) revealed that the quality of measurements at those stations is low. In particular, the correction for bias in the reanalysis weather data and the retrieval of covariates from remote sensing images are uncertain. A hierarchical model may be further explored to include uncertainty aspects of those observations.

We used the AIC to select the suitable copula family. The selection of families, however, depends upon the number of observations and further research is needed to develop strategies for selection optimization. Although several copula families can be found in the literature, we use five families because obtaining the inverse of the conditional copula distribution may lead to computational limitations. In addition, as all five families were symmetric, alternative families can be investigated.

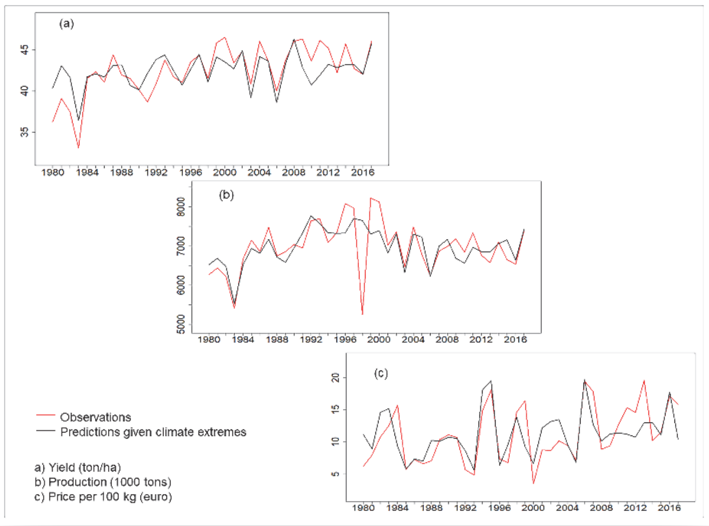
For the interpolation of air temperature in a data-scarce area, we selected three covariates LST, LAI and elevation that were retrieved from remote sensing images. Our study has shown that the covariates can easily be included as additional information in estimating the joint distribution, thus allowing for a richer dependence structure. The copulas are generally able to describe both spatio-temporal and non-spatial dependences. A practical advantage of our methods is that we can analyse the joint behaviour of more than one covariate and their effects on the spatial variability of the daily air temperature locally. The availability of bias-corrected values and the covariates derived from remote sensing images are, however, limitations for applying the methods on daily scales.

In order to provide a scenario that can be used to evaluate the new methods with less likely uncertain observations, we set up an experiment using the Meuse dataset (Pebesma, 2004). The leave-*k*-out cross validation showed that the average MAE values for mixed copula interpolator using Meuse variables zinc, lead, copper and cadmium were equal to 95.3, 33.7, 7.4 and 1.0, whereas for the co-kriging predictor they were equal to 173.2, 55.8, 12.8 and 1.8, respectively. Further applications of the new methods in other case studies including simulation-based information should provide more insight on these methods in the future.

We see several ways to further extend the current study. First, we applied the new methods in a data-scarce area, and we aimed to highlight the potential and the use of the methods for a larger dataset as well. Further comparison to other interpolation methods (Kilibarda et al., 2014) might help to assess the performance of the new methods. Second, essentially, we used the combination of the reanalysis weather data with a coarse spatial resolution and measurements from weather stations to predict mean air temperature at a higher spatial resolution. Such integration of bias correction and interpolation can be further investigated as a copula-based downscaling method. Third, in this study we selected the number and type of covariates, number of nearest neighbours, and number of spatial bins in both variogram and correlogram based upon our experience. A more generally applicable sensitivity analysis might help to show the effects of these parameters on the results.

Based upon the cross-validation, width of the prediction interval and visual inspection, we conclude that new methods allow describing both spatial variability and covariability between weather variables and covariates using multivariate joint distributions. In addition, the use of LAI as covariate in the interpolation of the mean air temperature reduces the uncertainties.

Chapter 7: Evaluating the effects of climate extremes on crop variables using copulas



This chapter is submitted as: Alidoost F., Stein A., Su Z. Evaluating the effects of climate change on crop yield, production and price using multivariate distributions: a new copula application. *Journal of Weather And Climate Extremes.*

Abstract

Climate change poses risks to agriculture and food security. To assess the impacts, this paper models the complex dependences of climate extreme indices and the crop-related variables: yield, production, and price of a crop. Using a comprehensive copula-based analysis, the conditional distributions of the crop-related variables given extremes of air temperature and precipitation are estimated. We used potatoes in the Netherlands as a case study. Weather data were obtained from 33 weather stations and ECMWF ERA-interim archive during the period 1980-2017. A joint behavior analysis predicted the yield, the production and the price with the relative mean absolute error equal to 5.4%, 3.6%, and 27.9%, respectively. The study showed that copulas adequately describe the multivariate dependences. Those in turn are able to quantify the impact of climate extremes, including their uncertainties.

Keywords

Climate change, Copulas, Crop, Multivariate distributions, Weather extremes.

Author contributions

F.A. conceived and designed the analysis, collected and processed the data, developed tools, performed the analysis, wrote the manuscript, is the corresponding author.

Z.S. supervised the findings of this work, verified the analytical methods, encouraged A.F. to investigate copulas for studying the impact of weather extremes on crop.

A.S. supervised the findings of this work, verified the analytical methods, encouraged A.F. to investigate copulas, improved the English wording.

All authors contributed to the interpretation of the results, and commented on the final manuscript.

Structure of the chapter

After an introduction in section 7.1, the application is introduced in section 7.2, climate extreme indices and joint behavior analyses are presented in section 7.3, the results are presented in section 7.4, followed by discussions and conclusion in section 7.5.

7.1 Introduction

Many complex processes and interactions determine crop responses to climate changes (Challinor et al. 2009a). Efforts have been made mainly to evaluate the impacts of the changes on crop yield (Challinor et al. 2013; Pirttioja et al. 2015; Gaupp et al. 2017; Nguyen-Huy et al. 2018). Little attention has been given to understand the impacts of climate change on crop production and production's price. Those are, however, important if say agricultural insurance should support farmers against the impacts and economic changes or climate information should answer stakeholders about total revenue (Dinku et al. 2011; Partridge and Wagner 2016; Anderson 2017).

An objective in local climate change studies is to quantify the changes in air temperature and precipitation extremes as they may result in a variety of climate-related crop stresses. Temperature affects the duration of the crop growing season, rates of photosynthesis, respiration, grain filling and thus the crop yield and production. Drought increases crop water stress and on the other hand, intensive rainfall may cause a flood and waterlogged soils (Lobell and Gourdji 2012). The assessment of impacts is primarily based upon extremes obtained from a long time-series of data from weather stations which are, however, sparse at local scales (Sarma 2005). Global assessments of crop production easily ignore variation at local scales (Lobell and Gourdji 2012). Therefore, additional spatially distributed data are needed for the assessment at those scales.

Weather data generated by the European Centre for Medium-range Weather Forecasts (ECMWF) are retrieved on spatial grids with coarse resolutions, typically in the order of ten kilometers. In addition, they are prone to uncertainty and their over- or underestimation compared to data from weather stations is often large (Hannah and Valdes 2001; Dee et al. 2011; Durai and Bhradwaj 2014). Hence, there is a challenge for the assessment of impacts when using weather data from ECMWF (Challinor et al. 2009b).

Analyzing changes in climate extremes requires long-term daily data that are not readily available in many parts of the world. The Expert Team on Climate Change Detection and Indices (ETCCDI) defined a total of 27 indices, which focus primarily on extremes (Sillmann et al. 2013). The study of extreme indices has become increasingly important due to their significant impacts on natural processes. Climate change consists of variations in several climate extremes and their impacts usually affect several agricultural variables. Therefore, multivariate joint distributions play an essential role in describing joint behaviors (Miao et al. 2016). The extension of a joint distribution to a d-dimensional distribution, $d > 2$, however, is often not straightforward (Salvadori et al. 2007). In this context, copulas help to construct multivariate distributions of related variables (Sklar 1973). While weather data are generally measured at a daily scale, climate indices describe the extremes at a yearly scale, for instance, the number of cold days in a year. Crop-related data are often recorded as seasonal and annual time-series. The use of extreme indices thus facilitates the estimation of the joint distribution of crop and weather variables. Applications of copulas include various practices, whereas there is a vast literature in geostatistics and hydrology (Bárdossy and Li 2008; Gräler and Pebesma 2011; Alidoost et al. 2018), meteorology and climate research (Scholzel and Friederichs 2008), and risk assessment (Renard and Lang 2007). Copula-based methods so far employed bivariate and trivariate joint return periods to analyze the dependencies between extremes indices (Miao et al. 2016; Zscheischler et al. 2017).

With the aim to assess the impact of climate changes on crop, we analyze the joint behavior of climate extreme indices with crop-related variables, e.g., yield, production, and price using multivariate distributions. We selected seven climate extreme indices, which are related to extremes in air temperature and precipitation. Previous studies in the literature have investigated the effect of only two or three indices on a single crop (Miao et al. 2016; Zscheischler et al. 2017). Our assessment applied on measurements from weather station is compared with the one applied on ECMWF weather data. Our study focuses on the use of copulas for the construction of multivariate distribution functions. Both good description of copulas and the main theorems are available in the literature (Nelsen 2006).

7.2 Study site and data in the Netherlands

We chose 33 KNMI stations where both rainfall and temperature measurements are available in the Netherlands during the period 1980-2017 (see Section 3.2). We selected 33 nearest grid points to the chosen KNMI stations from the ECMWF data. Daily minimum and maximum air temperatures are obtained using the minimum and maximum values of the hourly data, and daily precipitations are obtained using the sum of the hourly data for both weather datasets. For comparison purposes, we define the bias as systematic differences between ECMWF weather data and weather station measurements. Note that the source of bias lies in the different number of the measurements and the uncertainty in ECMWF weather data. The mean absolute bias was equal to 0.79ºC, 0.44ºC, and 1.84 mm for the daily minimum and maximum temperature, and the daily precipitation, respectively.

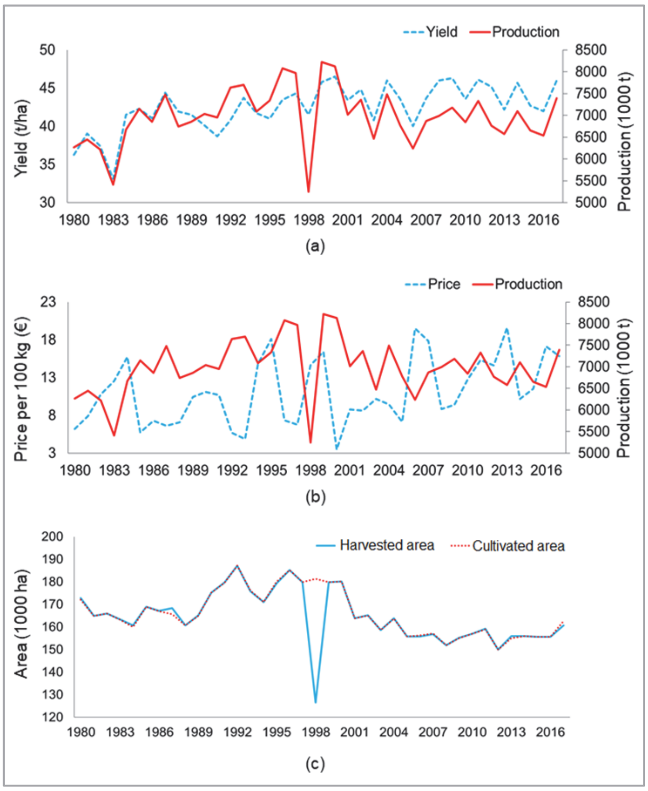


Figure 7.1 Temporal trends in the crop-related variable: a) yield and production, b) price and production, c) cultivation and harvested areas of potatoes.

We considered potato yield, production and price in the Netherlands during the period 1980-2017 (see Section 3.2). The yield and the production are naturally spatio-temporal variables, but their data are available either per province or country. There is, however, one price value for the country at each year. Regarding the variations in the crop-related variables, there is a significant drop in the production but not in the yield in the year 1998 (Figure 7.1b). Note that the production is yield × area. Comparing the cultivation and harvested areas of potatoes (in 1000 ha) revealed that the drop in the production was related to a drop in the harvested area (Figure 7.1.c). In the following, we investigated the climate event related to the drop.

7.3 Copula-based methods

In the following, a marginal distribution, i.e., the cumulative distribution of a variable is estimated by fitting an empirical distribution to data. In a multivariate case, a joint cumulative distribution is estimated by fitting copulas to data. The estimation methods of copulas are explained in Section 2.2.

7.3.1 Joint behavior analysis

A climate extreme index x_{ij} at location i and year j, is obtained for $N = 33$ locations and $M = 38$ years based upon the definitions provided by the Expert Team on Climate Change Detection and Indices (Table 7.1, Sillmann et al. 2013). There are both spatial and temporal dependences between the x_{ij}. A crop-related variable Y is sampled by y_j for M years. Based upon the estimates of autocorrelation function, we consider the y_j to be independent. We analyse the joint behaviour using the conditional distribution of Y given x, i.e., $F(Y|X = x)$, where x is a climate extreme index at year j.

To retrieve x from the spatio-temporal data x_{ij}, a marginal distribution is estimated using $x_{ij} > 0$ at year j for N locations, after testing for spatial stationarity of the mean. To do so, we evaluated the second order spatial stationarity assumption regarding the mean value using linear regression (Cressie 1993). Then, x can be obtained as either the median, i.e., the 50th percentile in the distribution, the mean or the mode. We conducted a cross-validation to compare the performance of those three predictors in selecting the dominant driving index. Based upon the results (not shown), we chose the median as the optimal predictor, as it minimizes the mean absolute prediction errors (Journel 1984; Cressie 1993). In the case that the standard deviation of $x_{ij} > 0$ for N locations is zero, the average of $x_{ij} > 0$ is used as x. Hence, we reduce the dimensionality of a spatio-temporal variable X from space-time to time. This provides a simple, but statistically sound method when we can select a dominant driving climate index in practical applications.

The conditional distribution $F(Y|X = x)$ is determined using the joint distribution $F(X, Y)$. The distribution can be extended to a d-dimensional distribution, $d > 2$, using either more than one climate extreme index or more than one crop-related variable. In our study, we choose seven climate extreme indices (Table 7.1) and analyse three joint behaviours using the distribution of: yield given seven indices $F(yield|X = x_1, \dots, X = x_7)$, production given seven indices $F(production|\, X = x_1, \dots, X = x_7)$, price given the production and seven indices $F(price|production, X = x_1, \dots, X = x_7)$. Here x_i is a climate index and i is mentioned in (Table 7.1). Different combinations of indices represent different climate conditions. In our study, the climate extreme indices are grouped as four events: 1) cold days, cold nights and very wet days indicated by x_1, x_2, x_5; 2) cold days, cold nights and consecutive wet days indicated by x_1, x_2, x_7; 3) warm days, warm nights and consecutive dry days indicated by x_3, x_4, x_5; and 4) all the seven indices indicated by $x_1, \dots, x_7$. The results of a joint behaviour analysis applied to the weather station measurements (dataset 1) will be compared with those applied to the ECMWF reanalysis weather data (dataset 2).

Table 7.1 Seven climate indices based upon daily temperature and precipitation used in this study. The Expert Team on Climate Change Detection and Indices (ETCCDI) provides the definitions.

Index ID	Index name	Label	Index definition
1	Cold days	TX10p	Number of days per each year during the reference period when $T_{dj} < T_{10p}$. T_{dj} is the daily maximum temperature on day *d* in year *j*. A cumulative distribution is determined using daily maximum temperatures in a five days window centered on *d* during the reference period. T_{10p} is the daily maximum temperature with 10th percentile in the distribution.
2	Cold nights	TN10p	Number of days per each year during the reference period when $T_{dj} < T_{10p}$. T_{dj} is the daily minimum temperature on day *d* in year *j*. A cumulative distribution is determined using daily minimum temperatures in a five days window centered on *d* during the reference period. T_{10p} is the daily minimum temperature with 10th percentile in the distribution.
3	Warm days	TX90p	Number of days per each year during the reference period when $T_{dj} > T_{90p}$. T_{dj} is the daily maximum temperature on day *d* in year *j*. A cumulative distribution is determined using daily maximum temperatures in a five days window centered on *d* during the reference period. T_{90p} is the daily maximum temperature with 90th percentile in the distribution.
4	Warm nights	TN90p	Number of days per each year during the reference period when $T_{dj} > T_{90p}$. T_{dj} is the daily minimum temperature on day *d* in year *j*. A cumulative distribution is determined using daily minimum temperatures in a five days window centered on *d* during the reference period. T_{90p} is the daily minimum temperature with 90th percentile in the distribution.
5	Very wet days	R95p	Number of days per each year during the reference period when $PR_{dj} > PR_{95p}$. PR_{dj} is the daily precipitation amount on wet day *d* in year *j*. On a wet day PR > 1mm. A cumulative distribution is determined using daily precipitation on wet days during the reference period. PR_{95p} is the daily precipitation with 95th percentile in the distribution.
6	Consecutive dry days	CCD	Largest number of consecutive days per each year during the reference period when $PR_{dj} \leq 1$ mm.
7	Consecutive wet days	CWD	Largest number of consecutive days per each year during the reference period when $PR_{dj} > 1$ mm.

7.3.2 Marginal and joint distributions estimation

We use the empirical marginal probability u_i where $i = 1, \ldots, n$ and n denotes the total number of observations of the variable of interest Z. Following rank-order-transformation $u_i = \frac{\mathrm{rank}(z_i)}{n+1}$, a continuous approximation of the marginal distribution of Z is obtained by means of kernels density estimation (Silverman 1986).

The joint distribution function $F(X,Y)$ is determined using a copula $C(U,V)$, where U and V are uniformly distributed random variables (Sklar 1973); (Nelsen 2006). According to Sklar's theorem, the joint probability $F(x,y)$ is equal to $C(u,v)$ and the joint density $f(x,y)$ is equal to $c(u,v) \times f_X(x) \times f_Y(y)$, where $u = F_X(x)$, $v = F_Y(y)$, and c is the copula density function (see Section 2.1). A multivariate copula describes dependences between three or more variables. In the first two analyses, the joint distribution is an 8-dimensional function whereas it is a 9-dimensional function in the last analysis. Following section 2.2, the conditional distribution $F(Y|.)$ is determined using a C-vine structure and five copula families.

7.3.3 Prediction and cross-validation

Since the conditional distribution $F(Y|.)$ is estimated, any *p*th percentile in the distribution can be used to predict $\hat{y}$, e.g.,:

$$\hat{y}_{mean} = E[Y|.] = \int_y y \cdot f(Y|.)dy, \tag{7.1}$$

$$\hat{y}_p = F^{-1}(p|.), \qquad p \in [0,1], \tag{7.2}$$

where f is the joint density function (Bárdossy and Li 2008). We select the mean predictor in (7.1), being the optimal predictor, as it minimizes the mean squared prediction error (Cressie 1993); (Journel 1984). The relative mean absolute error (RMAE) in percentage for M years equals:

$$RMAE = 100 \times \frac{1}{M}\sum_{j=1}^{M}\left(\frac{|y_j - \hat{y}_{j,mean}|}{y_j}\right). \tag{7.3}$$

We use RMAE to determine whether the different weather datasets produce statistically different predictions due to the uncertainty in ECMWF weather data. With a leave-*one*-out cross-validation, we assess the quality of the predictions. To do so, one observation y_j is removed and $\hat{y}_{j,mean}$ is predicted using the remainder of the observations. The RMAE in percentage for M years is then obtained in (7.3).

7.3.4 Validation

To evaluate the performance of the joint behaviour analyses, we conduct a leave-k-out validation. To do so, first, k observations $\mathrm{y}_j, \mathrm{y}_{j+1}, \ldots, \mathrm{y}_M$ are removed at year j, where $j = M - m + 1, \ldots, M$, and in our study m is 25% of the M years. Then, $\hat{y}_{j,mean}$ is predicted using the observations $\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_{j-1}$, i.e., without any information from the future, as is natural. The RMAE in percentage for $m \ll M$ years is then obtained as:

$$RMAE = 100 \times \frac{1}{m} \sum_{j=M-m+1}^{M} \left(\frac{|y_j - \hat{y}_{j,mean}|}{y_j} \right). \tag{7.4}$$

We perform an additional successive validation for the price. Let us consider that the year j is a target that its climate extreme indices are available. We want to predict both the production and the price at the target year j, where the observations are available at the years $1, 2, \ldots, \mathrm{j} - 1$. The target production is predicted in the second joint behavior analysis, followed by a prediction of the target price in the last joint behaviour using the target production. The mean relative error is then obtained in (7.4).

7.3.5 Assessment of the impact of climate extremes on crop

To assess the effect of climate extremes on the crop-related variable, we consider $(x_1, \ldots, x_7)$ as a climate extreme event. The event is characterized by a joint density $f(x_1, \ldots, x_7)$, where multivariate density function $f(.)$ is estimated using copulas. The joint return period T of the climate extreme indices corresponds to the probability of $P[X_1 > x_1, \ldots, X_7 > x_7]$ is obtained as:

$$T = \frac{\mu_T}{P[X_1 > x_1, \ldots, X_7 > x_7]}, \tag{7.5}$$

where $\mu_T = 1$ year (Salvadori et al. 2007), and in our study $T = 10, 50, 100$. As no closed form exists for $P[X > x_1, \ldots, x_7]$, the probabilities $P[.]$ of 100000 simulated values of $(x_1, \ldots, x_7)$ are obtained numerically using copulas and the addition rules in the probability theory (Stirzaker 2003) as $P[X_1 > x_1, \ldots, X_7 > x_7] = 1 - P[X_1 \le x_1 \; or \; \ldots or \; X_7 \le x_7] = 1 - \left(\sum_{i=1}^{7} F_i\,(x_i) - \sum_{i,j=1}^{7} F(x_i, x_j) + \sum_{i,j,k=1}^{7} F(x_i, x_j, x_k) + \cdots + (-1)^{7-1} F(x_1\,, \ldots, x_7) \right)$. The events with $9.9 \le T \le 10.1$ years return period are selected as representative events for the events of a $T = 10$ years return period. The same procedure is applied to select the events with 50 and 100 years return period. The variable Y given those events is then predicted using the mean predictor explained in (7.1). We illustrate the procedure using an example for the first event; the probabilities $P[X_1 > x_1, X_2 > x_2, X_5 > x_5]$ of 100000 simulated values of (x_1, x_2, x_5) are obtained. Then, the return levels x_1, x_2, and x_5 with $\frac{1}{10.1} \le P[X_1 > x_1, X_2 > x_2, X_5 > x_5] \le \frac{1}{9.9}$

are selected. Then, yield values are predicted using the conditional distribution $f(Y|x_1, x_2, x_5)$ in (7.1).

7.4 Results

7.4.1 Joint behaviour analysis

Climate extreme indices are obtained using the daily weather data in the growing season at 33 stations for 38 years, where the spatial domain is a country for joint behaviour analyses. Figure 7.2 shows the time-series of the dominant climate extreme index. The highest number of cold days were equal to 26 and 24 days in the years 1984 and 1986, respectively, and the highest number of warm days was equal to 31 days in the year 2006 which is related to the heatwave in 2006 (KNMI 2006). The highest number of consecutive wet days was equal to 11 days in the year 1998 (Figure 7.2), related to the flood on 16 September 1998 caused by El Niño (ESA/ESRIN 2018). The flood was responsible for a large drop in the harvested area and the production at that year.

Comparing climate extremes indices retrieved from both weather datasets denotes that the bias in the precipitation data resulted in a mean absolute bias of two, six and five days in the very wet days, the consecutive dry days, and the consecutive wet days, respectively. Figure 7.3 shows the empirical marginal distributions of the involved variables in the joint behavior analyses. The bias in the last three climate extreme indices is comparatively large. In the following, we investigate the performance of the joint behavior analyses when using ECWMF weather data.

Focusing on potatoes, we used the weather data in the growing season to obtain extreme indices. Restricting the data to the growing season is prone to uncertainty because the harvested area, hence the production and consequently the price, can be affected by for example a flood before the growing season.

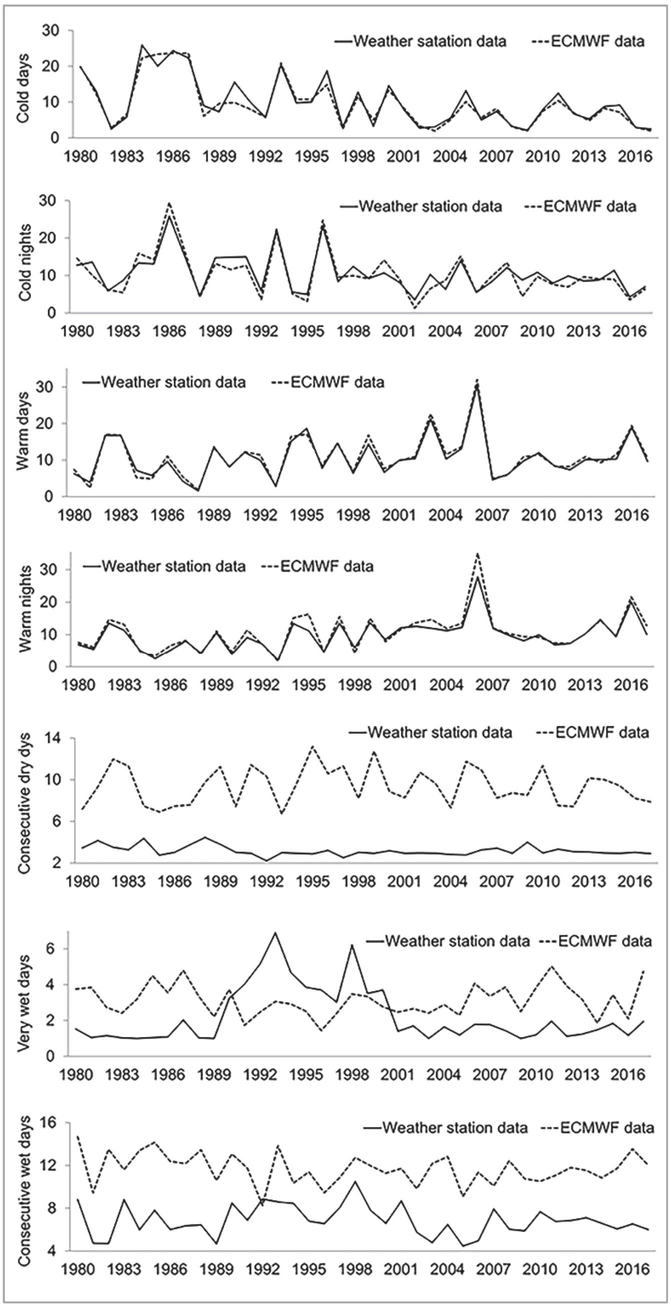


Figure 7.2 Time-series of the dominant climate extreme index. Climate extreme indices are obtained using the air temperature and precipitation data, retrieved from the weather stations and the ECMWF ERA-interim archive, in the growing season of potatoes at 33 stations for 38 years.

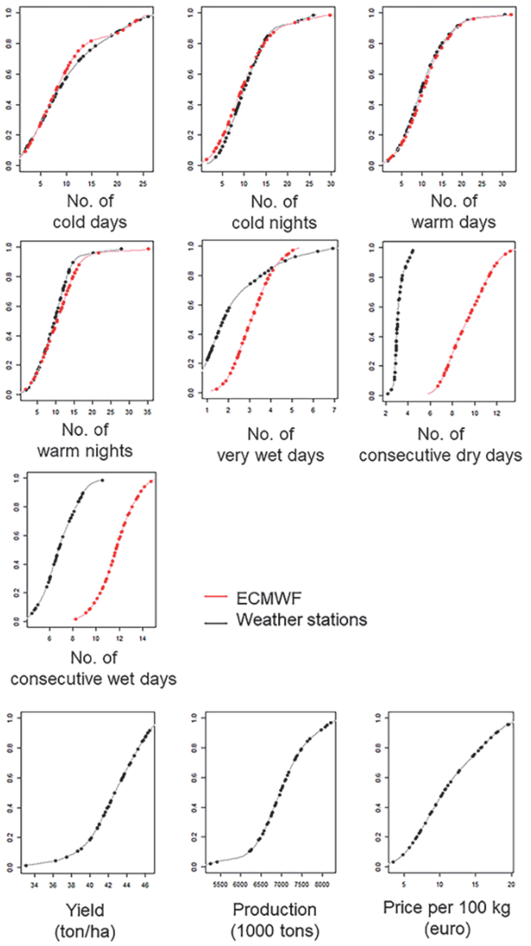


Figure 7.3 Empirical marginal distributions of the involved variables in the joint behavior analyses. The involved variables are: the seven dominant climate extreme indices, yield, production, and price. The vertical axis denotes the empirical cumulative probability.

7.4.2 Prediction and cross-validation

Boxplots in Figure 7.4 show the predictions of the crop-related variables, where p varies in the range of $[0,1]$ in (2), from 1980 to 2017. All observations fall in the prediction intervals except the low production value in the year 1998. Hence, it denotes a good performance of the joint behaviour analyses in estimating the joint distributions. In addition, comparing the predictions obtained by the mean predictor and the observations indicates that the joint behaviour analyses well represents the temporal variation of the crop-related variables. The relative mean absolute errors (RMAEs) were equal to 3.6%, 4.5% and 23.7% for the three joint behaviour analyses, where $M = 38$ in (3) and the climate extreme indices are obtained by the dataset 1 (Table 7.2). The RMAE values obtained by leave-*one*-out cross-validation, were equal to 5.0%, 6.1% and 40.2% for the joint behaviour analyses. As can be seen, the errors for the price were comparatively large.

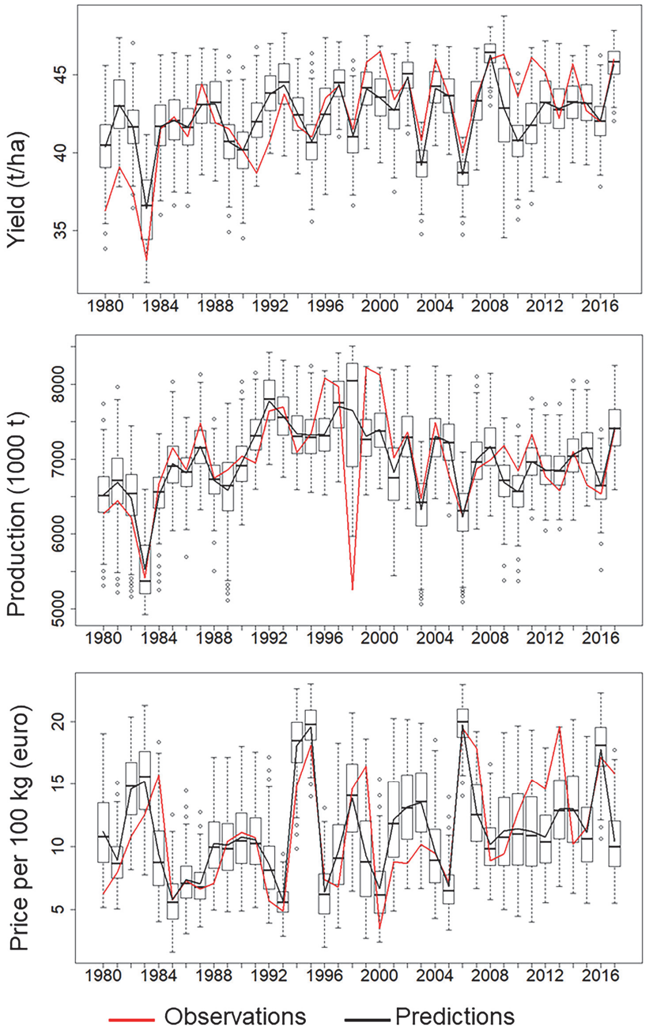


Figure 7.4 Predictions, shown as boxplots, of production, yield and price given the climate extreme indices obtained by the measurements dataset. The black line indicates the predictions obtained by the mean predictor whereas the red line indicates the observations.

For the three joint behavior analyses applied to the dataset 2, the RMAE values by the mean predictor were equal to 3.3%, 6.3%, and 23.6%, whereas by leave-*one*-out cross-validation, were equal to 4.9%, 9.8% and 38.4% (Table 7.2). Comparing the results with those obtained by weather station measurements showed that the quality of the predictions of the yield and the price is rather good in the presence of bias.

Table 7.2 Relative mean absolute error (RMAE) in percentage. Dataset 1 denotes weather station measurements, whereas dataset 2 denotes ECMWF weather data.

		Yield	Production	Price
Dataset 1	Prediction	3.6	4.5	23.7
	Leave-one-out cross-validation	5.0	6.1	40.2
	Leave-*k*-out cross-validation	5.4	3.6	27.9
	Successive validation	-	-	26.4
Dataset 2	Prediction	3.3	6.3	23.6
	Leave-one-out cross-validation	4.9	9.8	38.4
	Leave-*k*-out cross-validation	3.7	5.7	23.9
	Successive validation	-	-	17.9

Note that the joint behavior analyses in this study were only applied to the seven indices indicating the frequency of the weather extremes. The question can be posed whether considering a subset of the indices can improve the predictions. To answer this question, we conducted a sensitivity analysis using three subsets: 1) a cold event containing cold days, cold nights and very wet days, 2) a cold event containing cold days, cold nights and consecutive wet days, 3) a heat event containing warm days, warm nights and consecutive dry days. The RMAE values obtained by mean predictor (not shown) revealed that no improvements in the predictions were achieved. The low production value in the year 1998, however, falls in the prediction intervals of the production given the second subset. Considering other climate indices which are responsible for the intensity and the duration of extremes, should thus provide more insight on the predictions.

7.4.2 Validation

Validation was carried out, where $m = 9$ in (4), i.e., 25% of the 38 years. We could not further increase m, because it is important to use a reasonable number of data, here 75% of the 38 years, for estimation purposes. Using dataset 1, the RMAE values were equal to 5.4%, 3.6% and 27.9% for the three joint behaviour analyses, whereas the RMAE value was equal to 26.4% for the price in a successive validation (Table 7.2). Except for the production errors, the RMAE values of the dataset 2 are lower than dataset 1, because the number of data in dataset 2 is higher than that in dataset 1. In all the three joint behaviour analyses, the RMAE values were relatively low showing that the presented copula-based analysis was able to well represent the complex dependences. The low number of m implies, however, a limitation on the validation.

7.4.3 The impact of climate extremes on crop

The effect of climate extremes on the crop-related variables is assessed in two steps: first, the determination of plausible weather extreme indices associated with a joint return period, e.g. 10, 50 or 100 years in my study; second, the prediction of the crop-related variables e.g. yield, production and price conditioned on those extremes indices. Boxplots in Figure 7.5 show the predictions of yield, production and price given climate events with 10, 50, and 100 years joint return periods. For example, the first boxplot in the first row indicates yield variations ranging from 39 to 48 t.ha^{-1} because of the first event. Note that the predictions are the mean values obtained from the equation (7.1). We compared the lowest values of the predicted yield and production and the highest values of the predicted price with the average of their observations. It revealed that the event four with 50 years joint return period resulted into the largest variation among different events with different joint return periods: 21.0% and 28.5% decreases in yield and production, respectively, and 92% increases in price (see Figure 7.5). Note that event four contains all the seven extreme indices. Possible source of this variation lies in both complexity and flexibility in dependence structures: uncertainty either increases due to the larger number of the indices in joint distributions, or it decreases due to the larger number of indices where the joint distribution can well represent the dependence structures. As mentioned in section 4.2, the RMAE values obtained by cross-validation revealed that no improvements in the predictions were achieved using events one to three. It illustrates that event four allows for a good representation of the dependence structures. A high dimensionality of the distribution corresponded with an advantage of using the joint behavior analysis: using event four more information is obtained than using events one to three by selecting a subset of the indices. Due to the high dimension of joint distributions, the computational cost of the return periods is considerably high. The source of this cost and, consequently, the uncertainty in return periods lie in generating simulated values of the climate indices through their joint distribution, in numerical evaluation of joint probabilities using empirical copulas, and in successive procedures of the addition rules in probability.

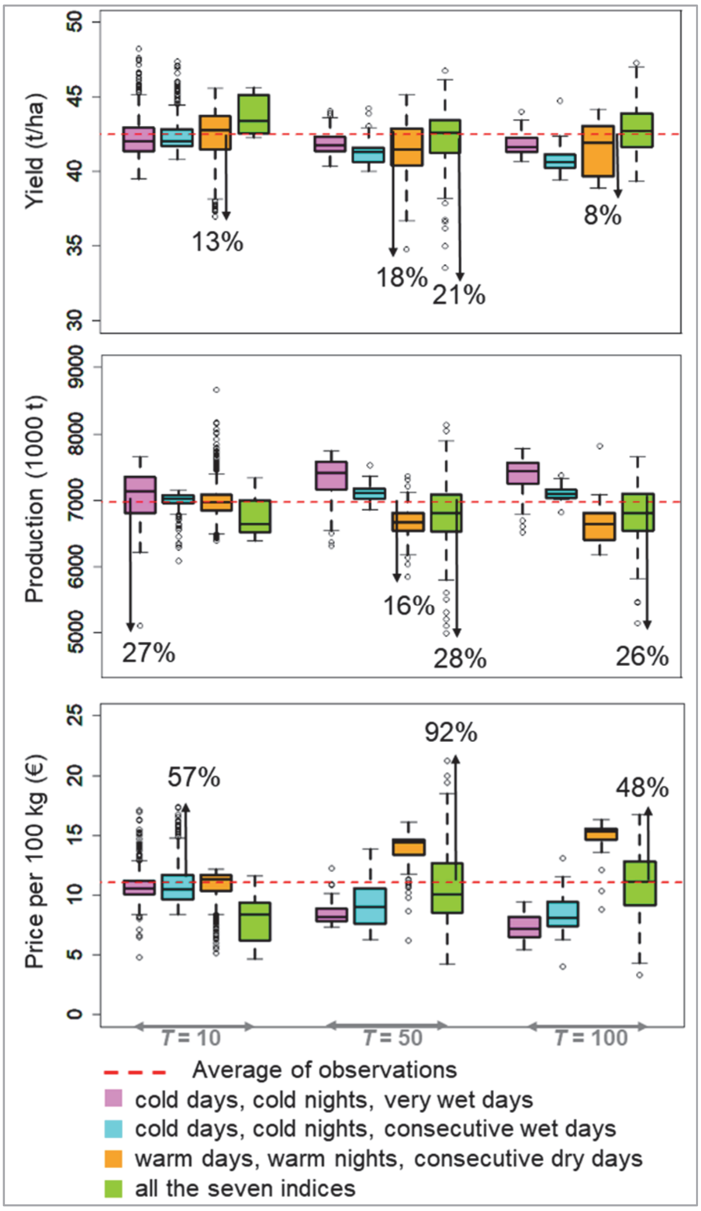


Figure 7.5 Predicted yield, production and price given climate extremes with T= 10, 50 and 100 years joint return periods. The boxplots show the predictions given simulated climate extreme indices. The joint distributions are estimated using the measurements dataset. The colors of boxplots indicate the events as magenta: cold days, cold nights, very wet days, blue: cold days, cold nights, consecutive wet days, orange: warm days, warm nights, consecutive dry days, green: all the seven indices. The horizontal red line denotes the average values of the observations.

7.5 Discussion and conclusion

We provided copula-based joint analyses to assess the impact of climate extreme indices on the yield, the production and the price of potatoes. The results of the predictions, leave-one-out cross-validations, and leave-*k*-out validations showed the practical advantage of copulas in estimating high dimensional joint distributions that describe the complex dependences. The use of C-vine structures in estimating multivariate distributions was beneficial as it allows for a huge degree of flexibility in describing the dependences because the involving bivariate copulas are estimated using five copula families. In addition, the conditional distributions are useful for a comprehensive uncertainty assessment using confidence intervals widths.

We conducted cross-validation to compare the performance of the median, mean and mode in selecting the dominant driving index, and therefore, reducing the dimensionality of climate extreme indices from space-time to time (not shown). The other percentiles in the distribution should be further explored. In addition, a sensitivity analysis might help to explain the effects of other estimation methods of marginal distribution on the results. For validation purposes, we chose the mean predictor. Further research will be necessary to investigate the use of other predictors in (2) to obtain the predictions.

The presented joint behavior analyses are general and could be applied on a different spatial domain, e.g., a province, where the price is assumed to be invariant between provinces. A limitation of decreasing the spatial domain from a country to province is that the number of weather stations is low to obtain the dominant climate extreme index. This limitation can be overcome using gridded ECMWF data which is, however, out of the scope of our paper. A bias correction method can be further investigated to correct for bias in the indices.

We see two ways to extend the current study:

We selected the seven climate extreme indices related to the air temperature and precipitation. The question can thus be posed whether other weather variables like the humidity and the wind produce statistically different predictions. In addition, whether other climate extreme indices can improve the predictions, for example, the indices for intensity and duration of the extreme precipitation. Due to the complexity of dependences, a challenge is to decide which climate extreme indices are important to be included in the joint behavior analyses. A sensitivity analysis needs to be further implemented to address these issues. In addition, the joint behavior of climate extreme indices and other crops like maize and wheat can be analyzed.

We neglected the effect of the conditions such as social-economic conditions, climate change adaptation scenarios and technologies on the yield, the production, and the price. Additional knowledge may lead to an improvement of the predictions, for example, the joint behaviour analysis of the price can be extended to include social-economic information.

Chapter 8: Synthesis

This chapter summarizes the study’s findings and synthesizes the research to point out significant results, obstacles, prospects, and limitations.

8.1 Findings

The research described in this thesis investigates new, innovative copula-based methods for improving the availability of climate and weather information in data-scarce environments. The performance of the new methods was evaluated by comparing them with several methods commonly applied for improving data availability. The comparison revealed a number of theoretical and practical issues in representing spatial variability and its associations, and when assessing the uncertainty. The following paragraphs describe the overall findings.

Two copula-based methods for correcting bias were used to correct daily reanalysis air temperatures for bias in an agricultural area in Iran. The copulas described the dependencies between two sources of the air temperature data: an ECMWF archive and a network of weather stations. After estimating joint distributions, new predictors based upon Conditional Probabilities (CP) were defined to obtain air temperatures at locations within the agricultural area. The two methods for bias correction, CP-I and CP-II, performed better than methods commonly applied (i.e. conditional expectation and conditional median predictors) in representing the spatial and temporal variation of the bias-corrected air temperatures.

Three new Multivariate Copula Quantile Mappings (MCQM-I, MCQM-II, and MCQM-III) were used to study two types of dependencies: the spatial variability of air temperature, and its association with elevations. The MCQMs were able to accommodate those dependencies, thereby improving the precision of the one-dimensional quantile mapping in predicting bias-corrected air temperatures.

Among the new bias correction methods, both CP-II and MCQM-III could improve air temperatures retrieved from ECMWF in a data-scarce environment. The evaluation criteria showed that CP-II was superior to MCQM-III, albeit at a higher computational cost.

A comparison of the conditional expectation and conditional median predictors with the one-dimensional quantile mapping revealed that copula-based methods performed better in correcting bias. This is in line with previous studies, although my results demonstrated that there was a similarity between these three methods. The spatial variation of the bias-corrected air temperatures was equal to the variation of the ECMWF air temperatures but not to that of the weather station measurements.

Two copula-based interpolators were introduced to produce weather maps in a data-scarce environment. The methods allowed the description of two types of the dependencies: spatial dependencies of air temperatures, and its associations with land variables. The interpolators were compared with the ordinary kriging and co-kriging predictors. The spatial copula interpolator including covariates, and the mixed copula interpolator describe both the spatial variability of air temperatures and its association with land variables obtained from remote sensing products. The copula-based interpolators are potentially useful for other sparsely and irregularly distributed weather data.

The mixed copula interpolator allowed the inclusion of additional variables in the modeling of spatial random fields using multivariate distributions. Hence, the joint distribution contained three types of dependencies: spatial dependencies between the variable of interest at a single location and its nearest neighbors, non-spatial dependencies between the variable of interest and its collocated covariates at that location; and the non-spatial dependencies of the variable of interest and its collocated covariates at its nearest neighbors.

In the comprehensive copula-based analyses, the conditional distribution of a crop-related variable given climate extreme indices was estimated. Then, the distribution was employed to predict the variable under climate change. The analyses were applied to two datasets: weather station measurements and ECMWF weather data. The copula-based analyses helped in modeling dependencies between the climate extreme indices and the crop-related variables using high-dimensional multivariate distributions. This suggests that, given climate extremes indices, the conditional distribution of a crop-related variable is advantageous for quantifying the impacts of climate extremes, including their uncertainties.

8.2 Significance

The research described in this thesis is unique in several aspects particularly related to the use of copulas, Earth observation data, and the developed functions.

The findings on the application of copulas in describing the dependencies between several variables indicate that copulas can estimate any multivariate distribution. A copula is neither a method nor a model, rather it is a joint distribution function. This definition is not dependent on the underlying statistical process and thus allows pragmatic application in agricultural and hydrological studies. My research therefore rebuts the assertion that "Generally, copulas are used only if Gaussian assumptions fail, e.g. fat-tailed volatility in financial markets."; this comment was made by an anonymous reviewer of a paper I submitted to the Journal of Spatial Science.

This thesis delivers an important message relating to the difference between estimation and prediction (Kutner et al. 1996). Initially, the joint distribution is fitted to the data, and the goodness of fit is tested using statistical tests. Next, a predictor is selected to predict the variable of interest. The choice of predictor is not related to how good estimation is, but rather to the loss functions. For instance, two conditional quantiles, the mean and the median, have been identified in the literature as the predictors that minimize squared error loss and absolute error loss (see Section 2.3). These predictors produce smooth maps where spatial stationarity is assumed for estimating bivariate joint distributions. To improve spatial predictions, however, the predictors, CP-I and CP-II were defined based upon several varying conditional probabilities. Flexibility in selecting predictors that are different from the conventional mean and median is a practical advantage that copulas when estimating distributions.

The findings of my research demonstrate the advantages of using Earth observation data in data-scarce environments. ECMWF ERA-interim archive, MODIS products, Landsat 8 data, and the SRTM dataset are a few examples used in my study. The results demonstrate that embedding satellite products in multivariate distributions leads to improvements in predicting weather data.

Several salient aspects were revealed by the extensive literature review included in the study. For instance, previously it had been reported that a Gaussian distribution is often assumed to be suitable for estimating distribution functions of air temperature, whereas a gamma distribution is assumed for precipitation. Those estimation procedures are usually based upon weather time-series (see Chapter 5). Hence, this may give the impression that air temperature always follows a Gaussian distribution, irrespective of its domain of distribution, i.e. spatial, spatio-temporal or non-spatial. The findings of my study confirm that those assumptions are stochasticity assumptions and not the property of physical processes similar to stationarity and ergodicity. In practice, a finite sample of a random variable is observed in space and time without replication. Most of us would make inferences about the joint distribution from those observations. Making assumptions should, of course, not be a concern, but rather encourage a dedicated choice with which to proceed with the investigations.

One other practical advantage of the research is that the new methods are generic: application of the new copula-based methods in other case studies should provide more insight into the nature of these methods. Since R programing software is an open source environment that has been increasingly used for the implementation of statistical operations, this new research has contributed to the spcopula (Gräler, 2011), and VineCopula (Brechmann and Schepsmeier, 2013) packages available in R. The functions developed in this research will also be available on the GitHub software development platform after publication of my research results. The availability of these functions will assist:

- Spatial interpolation of a random field that includes more than one covariate;

- Definition of different predictors based upon multivariate distributions;
- Inverse transformation of conditional distributions;
- Calculation of conditional probability based upon multivariate distributions;
- Calculation of high-dimensional joint probabilities based upon addition rules in probability theory; and
- production of a cross-correlogram which that has a definition similar to that of a cross-variogram.

8.3 Obstacles

During the research some problems arose related to the uncertainty, operational goals, and temporal characteristics of some of the bias correction methods.

One concern about the findings of this research was that the available weather station measurements in Iran were treated as benchmarks in the bias correction and interpolation methods. Depending upon the instrument used to measure air temperature and the temporal frequency of measurement, the weather stations were categorized as one of three types: synoptic, climatology type1, and climatology type2. Time-series of air temperature at climatology type2 stations revealed that the quality of their measurements is low. As a result, the degree to which the results are affected by the varying quality of data is a source of uncertainty. Another source of uncertainty is the satellite products, in particular covariates such as land surface temperature and leaf area index, which were retrieved through several procedures. To my knowledge, there is no closed-form mathematical expression for calculating the propagation of the uncertainty in such a way that it could be used to develop new methods. Implementation of the methods for simulation-based information may provide an alternative approach to this issue.

Regarding operational applications that include extensive datasets, the present research faces a limitation that emerges from finding the optimum parameters: For instance, the use of other copula families in the C-vine structure; the number and type of covariates; and the number of neighbors for CP-II and the interpolators. In addition, the computational cost of copula-based methods when working with high-dimensional distributions is relatively great. The use of parallel evaluations, powerful processors, and comprehensive sensitivity analysis might help to deal with this. An additional limitation is related to the availability of remote sensing data when the new interpolation methods need to be applied on daily scales.

This study shows that both CP-II and MCQM-III were less successful than CP-I and MCQM-I in representing the temporal variation of biased-corrected values. This may raise concerns about their application to predictions in time. The use of the spatio-temporal information can improve the methods, requiring that the developed functions need to be extended to include spatio-temporal data frames.

8.4 Prospects

There are four new areas to be explored that relate to the utilization of copulas: methodological development, promising applications, types of problems at hand, and education.

In this thesis I present the initials steps in developing methods for combining copula-based bias correction and interpolators for downscaling. In addition, the idea of the estimating high-dimensional multivariate distributions that include covariates opens a new approach in data/information fusion. Another opportunity arising from the use of copulas takes advantage of spatially varying conditional probabilities (e.g. CP-II) in Bayes classifiers and machine-learning environments. I also identify another route for future research: that of implementing copula-based methods of bias correction to predict bias-corrected values at an unvisited moment in time, instead of spatially.

For the applications concerning different types of datasets, the methods of bias correction presented can be applied to other weather parameters (such as precipitation) obtained from weather stations and ERA-I reanalysis. A promising, novel application of the bias correction methods is the local improvement of the land surface parameters retrieved from the European Centre for Medium-range Weather Forecasts (ECMWF). For instance, the daily evapotranspiration (ET) at local scales is important information in determining crop water requirements for use in an advisory system for irrigation. It is of interest, therefore, to develop a copula-based method of bias correction based upon MCQMs and apply it to evapotranspiration data obtained from Landsat products and ECMWF data so that temporal gaps can be filled. With respect to joint behavior analysis, a copula-based procedure can include socio-economic information to study the effects of climate change on urban areas.

The use of new copula-based methods should be further explored for different types of problems. In this context, estimation of high-dimensional distributions using C-vine structures has great potential for describing complex dependencies found in, for example, wicked problems such as climate change, heat waves, and frequent wildfires (Aerts et al. 2016).

As I mention in Chapter 1, the exploitation of copulas in geostatistics, as well as climate studies, is still relatively new. With this in mind, the case studies and the functions developed through this research, together with recent copula-based studies (Gräler 2014), could have a role to play in education in geostatistics. The main prerequisites for those who want to learn more are a good understanding of the geo/mathematical statistics and the basic theories in probability.

8.5 Limitations

There are a number of issues arising from my research that will need to be addressed in the future: numerical evaluations, non-stationarity, the First Law of Geography, deterministic approaches, and climate impact studies.

Numerical evaluations concern the implementations of some mathematical/statistical operations for copula families, such as partial derivatives and inverse transformations. These are still at an experimental stage and are subject to change during the development of functions to implement copula-based methods. In addition, a *d*-dimensional joint probability, $d > 2$, is obtained using the numerical evaluations and simulations that are associated with uncertainty. In the case of high-dimensional distributions, however, the joint probability is close to zero.

Stationarity was assumed throughout the estimation of joint distributions. This assumption was justified through either some statistical tests or a test scenario (see Chapters 4 and 5). It is notable that those assumptions are justifiable in the case that one fails to reject them (Cressie 1993). The degree to which degree my findings alter in a non-stationary case study remains unanswered.

The First Law of Geography (Tobler's First Law) states that "All things are related, but nearby things are more related than distant things." It is by now generally accepted that geostatisticians exploit this law not only for spatial modeling but also for spatio-temporal modeling. Among the methods I have researched for this thesis, those considering their spatially nearest neighbors in their formulations implicitly acknowledge the Tobler's First Law: i.e. CP-II, MCQM-II, MCQM-III and the interpolators. It is for this reason that much attention should be given to the probabilistic nature of the desired variables. For example, crop production and price in a given year do not impact those in the following year. This explains why temporally varying predictor similar to CP-II could not be used in the joint behavior analysis presented in Chapter 7.

For a deterministic approach, the main theories of copulas that are based upon probabilistic explanations need to be extended. There is, however, a definition of copulas that uses Geometric methods, without any reference to distribution functions or random variables (Nelsen 2006). In future research it would be interesting to find out whether the new methods I present in this thesis are applicable in a deterministic setting.

With respect to climate impact studies, the copula-based methods assume that weather can be defined as a stochastic or random process. Analytical skills, therefore, are necessary to interpret the statistical characteristics of weather data. In addition, statistical methods are data-driven i.e. the methods are applied to historical data and they give the desired output. In this thesis, I used measured data retrieved from weather stations to estimate the distribution functions and to validate the results. The cost and availability of measured data may hinder the application of the methods.

Bibliography

Aalto, J., Pirinen, P., Heikkinen, J., and Venäläinen, A. (2013). Spatial interpolation of monthly climate data for Finland: comparing the performance of kriging and generalized additive models, *Theoretical and Applied Climatology*, 112(1-2), 99-111.

https://doi.org/10.1007/s00704-012-0716-9.

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

https://doi.org/10.1016/j.insmatheco.2007.02.001.

Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control,* 19(6), 716–723.

Alidoost, F., Stein, A. (2016). Correction of daily ECMWF air temperature data based on copula concept, *'SAIL35, Eye on Foliage' international symposium*.

Alidoost, F., Stein, A. and Su, Z. (2018). Copula-based interpolation methods for air temperature data using collocated covariates. *Spatial Statistics*, 28, 128-140. https://doi.org/10.1016/j.spasta.2018.08.003.

Allen, R. G., Pereira, L. S., Raes, D., and Smith, M. (1998). FAO Irrigation and Drainage Paper No. 56.

Anderson, K., Ryan, B., Sonntag, W., Kavvada, A., and Friedl, L. (2017). Earth observation in service of the 2030 Agenda for Sustainable Development. *Geo-spatial Information Science*, 20(2), 77-96.

https://doi.org/10.1080/10095020.2017.1333230.

Bárdossy, A. and Li, J. (2008). Geostatistical interpolation using copulas. *Water Resources Research,* 44(7), 15. https://doi.org/10.1029/2007WR006115.

Beukema, H. P. and van der Zaag, D. E., (1990). *Introduction to Potato Production*. Pudoc at Wageningen.

Brechmann, E. C. and Schepsmeier, U. (2013). Modeling Dependence with C- and D-Vine Copulas: The R Package CDVine. *Journal of Statistical Software,* 52(3), 1–27. DOI: 10.18637/jss.v052.i03.

CBS (2018). CBS open data services. In: Central Bureau for Statistics, the Netherlands.

Challinor, A. J., Ewert, F., Arnold, S., Simelton, E., and Fraser, E. (2009a). Crops and climate change: progress, trends, and challenges in simulating impacts and informing adaptation. *Journal of Experimental Botany*, 60(10), 2775–2789.

Challinor, A. J., Osborne, T., Morse, A., Shaffrey, L., Wheeler, T., Weller, H., and Vidale, P. L. (2009b). Methods and resources for Climate impacts research Achieving Synergy. *Bulletin of the American Meteorological Society*, 90(6), 836-848. https://doi.org/10.1175/2008BAMS2403.1.

Challinor, A. J., Smith, M. S. and Thornton, P. (2013). Use of agro-climate ensembles for quantifying uncertainty and informing adaptation. *Agricultural and Forest Meteorology*, 170, 2–7.

https://doi.org/10.1016/j.agrformet.2012.09.007.

Chambers J., Hastie T., Pregibon D. (1990). *Statistical Models in S*. In: Momirović K., Mildner V. (eds) Compstat. Physica-Verlag HD. 317-321.

https://doi.org/10.1007/978-3-642-50096-1_48.

Conover, W. J. (1971). *Practical Nonparametric Statistics*. New York: John Wiley & Sons. 295–314.

Cressie, N. (1993). *Statistics for Spatial Data.* Canada: John Wiley & Sons, 105–110.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Holm, E. V., Isaksen, L., Kallberg, P., Kohler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J. J., Park, B. K., Peubey, C., Rosnay, P. d., Tavolato, C., Thepaut, J. N., and Vitart, F. (2011). The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society,* 137(656), 553–597.

Demarta, S. and McNeil, A. J. (2005). The *t* copula and related copulas. *International Statistical Review/Revue Internationale de Statistique,* 73(1), 111-129.

Dinku, T., Asefa, K., Hilemariam, K., Grimes, D., and Connor, S. (2011). Improving Availability, Access and Use of Climate Information. *WMO Bulletin*, 60(2), 80-86.

Dodds, P. G., Huijsmans, C. B. and DePagter, B. (1990). Characterizations of Conditional Expectation-Type operators. *Pacific Journal of Mathematics,* 141(1). 55-77.

https://projecteuclid.org/euclid.pjm/1102646774.

Durai, V. R., and Bhradwaj, R. (2014). Evaluation of statistical bias correction methods for numerical weather prediction model forecasts of maximum and minimum temperatures. *Natural Hazards,* 73(3), 1229-1254. https://doi.org/10.1007/s11069-014-1136-1.

Durocher, M., Chebana, F., and Ouarda, T. B. M. J. (2016). On the prediction of extreme flood quantiles at ungauged locations with spatial copula, Journal of Hydrology., 533, 523–532, 2016.

https://doi.org/10.1016/j.jhydrol.2015.12.029.

ESA/ESRIN (2018). Earth Watching-Natural disasters. European Space Agency. Available from: https://earth.esa.int/web/earth-watching/natural-disasters.

Eurostat (2018). European Statistics database.

Fritsch, F. N., and Carlson, R. E. (1980). Monotone piecewise cubic interpolation, SIAM *Journal on Numerical Analysis*, 17(2), 238–246.

https://doi.org/10.1137/0717021.

Gaupp, F., Pflug, G., Hochrainer-Stigler, S., Hall, J., and Dadson, S. (2017). Dependency of Crop Production between Global Breadbaskets: A Copula Approach for the Assessment of Global and Regional Risk Pools. *Risk Analysis*, 37(11). https://doi.org/10.1111/risa.12761.

Genest, C. and Favre, A.C. (2007). Everything You Always Wanted to Know about Copula Modeling but Were Afraid to Ask. Journal of Hydrologic Engineering, 12(4), 347-368.

https://doi.org/10.1061/(ASCE)1084-0699(2007)12:4(347).

Genest, C., Remillard, B. and Beaudoin, D. (2009). Goodness-of-fit tests for copulas: A review and a power study. *Insurance: Mathematics and Economics,* 44(2), 199–213.

https://doi.org/10.1016/j.insmatheco.2007.10.005.

Gräler, B. (2014). *Developing spatio-temporal copulas* (Doctoral). Westfälische Wilhelms-Universität Münster.

Gräler, B., and Pebesma, E. (2011). The pair-copula construction for spatial data: a new approach to model spatial dependency, *Procedia Environmental Sciences*, 7, 206-211.

https://doi.org/10.1016/j.proenv.2011.07.036.

Hannah, E. and Valdes, P. (2001). Validation of ECMWF (re)analysis surface climate data, 1979-1998, for Greenland and implications for mass balance modelling of the Ice Sheet. *International Journal of Climatology,* 21(2), 171–195. https://doi.org/10.1002/joc.609.

Haslauer, C. P., Heißerer, T., and Bárdossy, A. (2016). Including land use information for the spatial estimation of groundwater quality parameters – 2. Interpolation methods, results, and comparison, Journal of Hydrology, 535, 699–709.

http://dx.doi.org/10.1016/j.jhydrol.2016.01.054.

Heißerer, T., Haslauer, C. P., and Bárdossy, A. (2016). Including land use information for the spatial estimation of groundwater quality parameters – 1. Local estimation based on neighbourhood composition, Journal of Hydrology, 535, 688–698.

http://dx.doi.org/10.1016/j.jhydrol.2015.12.049.

Hengl, T., Heuvelink, G. B. M., Tadić, M. P., and Pebesma, E. J. (2012). Spatio-temporal prediction of daily temperatures using time-series of MODIS LST images, *Theoretical and Applied Climatology*, 107(1-2), 265-277.

https://doi.org/10.1007/s00704-011-0464-2.

Huang, W. and Prokhorov, A. (2014). A Goodness-Of-Fit test for copulas. *Econometric Reviews,* 33(7), 751–771.

https://doi.org/10.1080/07474938.2012.690692.

Ines, A. V. M. and Hansen, J. W. (2006). Bias correction of daily GCM rainfall for crop simulation studies. *Agricultural and forest meteorology,*138(1-4), 44–53. https://doi.org/10.1016/j.agrformet.2006.03.009.

Jarvis, A., Reuter, H. I., Nelson, A., and Guevara, E. (2008). Hole-filled seamless SRTM data V4, *International Centre for Tropical Agriculture* (CIAT).

Jiménez-Muñoz, J. C., Sobrino, J. A., Skoković, D., Mattar, C., and Cristóbal, J. (2014). Land Surface Temperature Retrieval Methods From Landsat-8 Thermal Infrared Sensor Data, *IEEE Geoscience and Remote sensing Letters*, 11(10), 1840-1843. DOI: 10.1109/LGRS.2014.2312032.

Joe, H. (1993). Parametric families of multivariate distributions with given margins. *Journal of Multivariate Analysis,* 46(2), 262-282.

https://doi.org/10.1006/jmva.1993.1061

Journel A.G. (1984). mAD and Conditional Quantile Estimators. In: Verly G., David M., Journel A.G., Marechal A. (eds) *Geostatistics for Natural Resources Characterization*. Springer, Dordrecht, 261-270.

Kilibarda, M., Hengl, T., Heuvelink, G. B. M., Gräler, B., Pebesma, E., Tadić, M. P., and Bajat, B. (2014). Spatio-temporal interpolation of daily temperatures for global land areas at 1 km resolution, *Journal of Geophysical Research: Atmospheres*, 119(5), 2294-2313. https://doi.org/10.1002/2013JD020803.

Klimatologie weather data from the Netherlands (2018). Royal Netherlands Meteorological Institute.

KNMI, 2006. Available from: https://web.archive.org/web/20070207033312/http://www.knmi.nl/klimatologie/maand_en_seizoensoverzichten/maand/jul06.html [Accessed 2018].

Kojadinovic, I. and Yan, J. (2010). Modeling Multivariate Distributions with Continuous Margins Using the copula R Package. *Journal of Statistical Software,* 34(9), 1–20. http://hdl.handle.net/10.18637/jss.v034.i09.

Kuipers, L. and Niederreiter, H. (2012). *Uniform Distribution of Sequences.* Canada: John Wiley & Sons, 1–4.

Kum, D., Lim, K. J., Jang, C. H., Ryu, J., Yang, J. E., Kim, S. J., Kong, D. S. and Jung, Y. (2014). Projecting Future Climate Change Scenarios Using Three Bias-Correction Methods. *Hindawi Publishing Corporation Advances in Meteorology,* 1–13.

http://dx.doi.org/10.1155/2014/704151.

Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. (1996). *Applied Linear Statistical Models*, edited by 5th, McGraw-Hill/Irwin, New York.

Lafon, T.*,* Dadson, S., Buys, G. and Prudhomme, C. (2013). Bias correction of daily precipitation simulated by a regional climate model: a comparison of methods. *International Journal of Climatology,* 33(6), 1367–1381.

https://doi.org/10.1002/joc.3518

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

Lenderink, G., Buishand, A. and van Deursen, W. (2007). Estimates of future discharges of the river Rhine using two scenario methodologies: direct versus delta approach. *Hydrology and Earth System Sciences,* 11(3), 1145–1159. https://doi.org/10.5194/hess-11-1145-2007

Li, J. (2010). *Application of Copula as a New Geostatistical Tool*, PhD, Institut für isserbau der Universität Stuttgart, Universität Stuttgart, 161 pp.

Lobell, D. B. and Gourdji, S. M. (2012). The Influence of Climate Change on Global Crop Productivity. *Plant Physiology*, 160, 1686–1697. https://doi.org/10.1104/pp.112.208298

Manner, H. (2007). *Estimation and Model Selection of Copulas with an Application to Exchange Rates.* Universiteit Maastricht: Maastricht research school of Economics of TEchnology and ORganizations.

Mao, G., Vogl, S., Laux P., Wagner S. and 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.

Miao, C., et al. (2016). Joint analysis of changes in temperature and precipitation on the Loess Plateau during the period 1961–2011. Climate Dynamics, 47, 3221-3234.

Mulla, D. J. (2013). Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps. *Biosystems Engineering*, 114(4), 358-371.

https://doi.org/10.1016/j.biosystemseng.2012.08.009.

Nelsen, R. (2003). Properties and applications of copulas: A brief survey. *In:* Dhaene, J., Kolev, N. and Morettin, P. eds. *Proceedings of the First Brazilian Conference on Statistical Modeling in Insurance and Finance.* University Press USP: Sao Paulo.

Nelsen, R. B. (2006). *An Introduction to Copulas.* United States of America: Springer.

Nguyen-Huy, T., Deo, R. C., Mushtaq, S., An-Vo, D. A., and Khan, S. (2018). Modeling the joint influence of multiple synoptic-scale, climate mode indices on Australian wheat yield using a vine copula-based approach. *European Journal of Agronomy*, 98, 65–81. https://doi.org/10.1016/j.eja.2018.05.006

Oden, N. L. (1984). Assessing the significance of a spatial correlogram, *Geographical analysis*, 16(1), 1–16.

https://doi.org/10.1111/j.1538-4632.1984.tb00796.x.

Parmentier, B., McGill, B. J., Wilson, A. M., Regetz, J., Jetz, W., Guralnick, R., Tuanmu, M.N., and Schildhauer, M. (2015). Using multi-timescale methods and satellite-derived land surface temperature for the interpolation of daily maximum air temperature in Oregon, *International Journal of Climatology*, 35(13), 3862–3878.

DOI:10.1002/joc.4251.

Partridge, A. G., and Wagner, N. J. (2016). Risky Business: Agricultural Insurance in the Face of Climate Change. *Elsenburg Journal*, 13(3), 49-53.

Pebesma, E. J. (2004). Multivariable geostatistics in S: the gstat package. *Computers & Geosciences,* 30(7), 683–691.

https://doi.org/10.1016/j.cageo.2004.03.012

Persson, A. (2013). *User guide to ECMWF forecast products.*UK: ECMWF.

Pirttioja, N., Carter, T. R., Fronzek, S., Bindi, M., Hoffmann, H., Palosuo, T., Ruiz-Ramos, M., Tao, F., Trnka, M., Acutis, M., Asseng, S., Baranowski, P., Basso, B., Bodin, P., Buis, S., Cammarano, D., Deligios, P., Destain, M.-F., Dumont, B., Ewert, F., Ferrise, R., François, L., Gaiser, T., Hlavinka, P., Jacquemin, Kersebaum, I., Kollas, C., Krzyszczak, J., Lorite, I. J., Minet, J., M. I. Minguez, K. C., Montesino, M., Moriondo, M., Müller, C., Nendel, C., Öztürk, I., Perego, A., Rodríguez, A., Ruane, A. C., Ruget, F., Sanna, M., Semenov, M. A., Slawinski, C., Stratonovitch, P., Supit, I., Waha, K., Wang, E., Wu, L., Zhao, Z., and Rötter, R. P. (2015). Temperature and precipitation effects on wheat yield across a European transect: a crop model ensemble analysis using impact response surfaces. Climate Research, 65, 87–105. https://doi.org/10.3354/cr01322.

Renard, B. and Lang, M. (2007). Use of a Gaussian copula for multivariate extreme value analysis: Some case studies in hydrology. *Water Resources Research*, 30, 897–912. https://doi.org/10.1016/j.advwatres.2006.08.001

Salvadori, G., De Michele, C., Kottegoda, N.T. and Rosso, R. (2007). *Extremes In Nature: An Approach Using Copulas.* Dordrecht, The Netherlands: Springer.

Sarma, A. A. L. N. (2005). Scales of Climate, in: Encyclopedia of World Climatology, edited by: Oliver, J. E., *Encyclopedia of Earth Sciences Series Springer Netherlands*, 637-639.

Scholzel, C. and Friederichs, P. (2008). Multivariate non-normally distributed random variables in climate research – introduction to the copula approach. *Nonlin. Processes Geophys*, 15, 761–772. https://doi.org/10.5194/npg-15-761-2008

Sharifi, M. (2013). *Development of planning and monitoring system supporting irrigation management in the Ghazvin irrigation network.* Tehran, Iran: SAJ Co.

Sillmann, J., Kharin, V. V., Zhang, X., Zwiers, F. W., and Bronaugh, D. (2013). Climate extremes indices in the CMIP5 multimodel ensemble: Part 1. Model evaluation in the present climate. *Journal of Geophysical Research: Atmospheres*, 118(4), 1716–1733.

https://doi.org/10.1002/jgrd.50203.

Silverman, B. W. (1986). Density estimation for statistics and data analysis. UK Chapman and Hall/CRC.

Sklar, A. (1973). Random variables, joint distribution functions, and copulas. *Kybernetika,* 9(6), 449-460. DOI:10.1214/lnms/1215452606.

Stein, A., and Corsten, L. C. A. (1991). Universal kriging and cokriging as a regression procedure. *Biometrics*, 47(2), 575-587.

Stirzaker, D. (2003). *Elementary Probability*. Cambridge University Press.

Stoffelen, A. (1998). Toward the true near-surface wind speed: Error modeling and calibration using triple collocation. *Journal of Geophysical Research,* 103(C4), 7755-7766.

Teutschbein, C. and Seibert, J. (2012). Bias correction of regional climate model simulations for hydrological climate-change impact studies: Review and evaluation of different methods. *Journal of Hydrology,* 456–457, 12–29. https://doi.org/10.1016/j.jhydrol.2012.05.052

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., Dumedah, G., and 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(6), 3507–3521. DOI:10.1109/TGRS.2014.2378913.

Vogl, S., Laux, P., Qiu, W., Mao, G. and 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/hessd-9-937-2012.

Wu, T., and Li, Y. (2013). Spatial interpolation of temperature in the United States using residual kriging, *Applied Geography*, 44, 112-120.

https://doi.org/10.1016/j.apgeog.2013.07.012

Zanter, K. (2016). *Landsat 8 (L8) Data Users Handbook*, Department of the Interior U.S. Geological Survey, EROS Sioux Falls, South Dakota.

Zscheischler, J., Orth, R. and Seneviratne, S. I. (2017). Bivariate return periods of temperature and precipitation explain a large fraction of European crop yields. *Biogeosciences*, 14, 3309–3320. https://doi.org/10.5194/bg-14-3309-2017.

Fakhereh (Sarah) Alidoost is a researcher/analyst in Geomatics and Remote Sensing (RS). Her research interest is the use of Earth observation and geostatistics in understanding the interaction between weather, land, and water.

Majoring in Geodesy & Geomatics - Surveying Engineering, she learned about Earth observation systems and environmental science.

K.N. Toosi University of Technology, Iran, 2005- 2009

During her Master's study in RS, she acquired the knowledge of geostatistics, RS retrieval techniques, photogrammetry, and data integration/analysis. She graduated *with distinction*.

K.N. Toosi University of Technology, Iran, 2009- 2012

Her interests for operational applications of RS gave her the chance to join an international research team in Iran. She contributed to two national projects: rice production monitoring system in the Caspian plain, and improvement of water productivity and demand management in Qazvin irrigation network.

SAJ, Iran, 2012- 2014

Her passionate curiosity for learning helped her to move to the Netherlands as a *highly skilled migrant*. She, as an *AIO* accomplished her Ph.D. research that constitutes a relatively new area based upon copulas for describing the dependencies in weather variables and non-climatic variables, considering ancillary RS data- the present dissertation.

Departments of EOS and WRS, the faculty of ITC, University of Twente, the Netherlands, 2014-2018

Next to her research activities, she prepared new education material for data assimilation and copulas.

the faculty of ITC, University of Twente, Nov. - Dec. 2018

Her research has produced the following academic output besides the present dissertation:

Alidoost, F., Stein, A. Su, Z., and Sharifi, M. A., 2019. Multivariate copula quantile mapping for bias correction of reanalysis air temperature data, Journal of Spatial Science, https://doi.org/10.1080/14498596.2019.1601138 (In Press).

Alidoost, F., Stein, A. and Su, Z., 2018. Copula-based interpolation methods for air temperature data using collocated covariates, *Spatial Statistics*, (https://doi.org/10.1016/j.spasta.2018.08.003).

Alidoost, F., Sharifi, M. A., and Stein, A., 2014. Region- and pixel-based image fusion for disaggregation of actual evapotranspiration, *International Journal of Image and Data Fusion*, (https://doi.org/10.1080/19479832.2015.1055834).

Alidoost, F., Stein, A. and Su, Z., The use of bivariate copulas for bias correction of reanalysis air temperature data, *Journal of PLOS ONE*, (Under review).

Alidoost, F., Su, Z. and, Stein, A., Evaluating the effects of climate changes on crop yield, production and price using multivariate distributions: a new copula application, *Weather and climate extremes*, (Under review).

Chhipa, V., Shankar, H., Stein, A., Kallambukattu, J., Alidoost F., 2019. Assessing spatial variability of soil health related variables in a hilly terrain, *Geoderma,* 343, 130–138, (https://doi.org/10.1016/j.geoderma.2019.02.018)

Alidoost, F., Stein, A., 2016. Correction of daily ECMWF air temperature data based on copula concept, *SAIL35: Eye on Foliage, symposium*, Enschede, The Netherlands.

Alidoost, F., Stein, A., 2017. Interpolation of daily mean air temperature data via spatial and non-spatial copulas, *Spatial Statistics: One world, one health*, Lancaster, UK.

Bostan, P., Stein, A., Alidoost, F., Osei, F., 2019, Minimum temperature mapping with spatial copula interpolation, *Spatial Statistics 2019: Towards Spatial Data Science*, Sitges, Spain (submitted).

Stein, A., Alidoost, F., van Zoest, V., 2019, Spatial interpolation of extreme PM1 values using copulas, *first International Symposium on Computational and Methodological Statistics and Biostatistics*, University of Pretoria, South Africa (submitted).

SENSE

Netherlands Research School for the
Socio-Economic and Natural Sciences of the Environment

DIPLOMA

For specialised PhD training

The Netherlands Research School for the
Socio-Economic and Natural Sciences of the Environment
(SENSE) declares that

Fakhereh Alidoost

born on 8 June 1986 in Esfahan, Iran

has successfully fulfilled all requirements of the
Educational Programme of SENSE.

Enschede, 24 April 2019

The Chairman of the SENSE board	the SENSE Director of Education
Prof. dr. Martin Wassen	Dr. Ad van Dommelen

The SENSE Research School has been accredited by the Royal Netherlands Academy of Arts and Sciences (KNAW)

KONINKLIJKE NEDERLANDSE
AKADEMIE VAN WETENSCHAPPEN

The SENSE Research School declares that **Fakhereh Alidoost** has successfully fulfilled all requirements of the Educational PhD Programme of SENSE with a work load of 40.8 EC, including the following activities:

SENSE PhD Courses

- Research in context activity: 'Active participation in Coaching Bootcamp (including intercultural sensitivities, academic publishing and presenting and coaching) and writing of accessible press release on publication of PhD dissertation' (2018)

Other PhD and Advanced MSc Courses

- Geo-statistics and Advanced Geo-statistics, University of Twente/ ITC (2015)
- Write to publicize, University of Twente (2016)
- Earth observation and quantification of water cycle components, University of Twente/ ITC (2017)
- Atmosphere-Vegetation-Soil Interaction, University of Twente/ ITC (2017)
- Technical Writing and Editing, University of Twente (2017)
- Supervising Students course, University of Twente (2018)

External training at a foreign research institute

- Parametrization of sub-grid physical processes, European centre for medium range weather forecast, United Kingdom (2017)

Management and Didactic Skills Training

- Supervising MSc student with thesis entitled 'Comparison of deterministic and stochastic interpolation methods by assessing spatial variability in soil properties in a hilly terrain' (2017)
- Teaching copulas and supervising practical session of R programming in the MSc course 'Advanced geo-statistics' (2018)
- Supervising an intern (2018)

Oral Presentations

- *Interpolation of daily mean air temperature data via spatial and non-spatial copulas*. Spatial Statistics 2017, 4-7 July 2017, Lancaster, United Kingdom
- *Copulas for weather data, two applications in a data scarce area: bias correction and interpolation*. KNMI meetings, 12 July 2017, Utrecht, the Netherlands
- *Weather forecast data and statistical methods to interpolate daily air temperature*. Weather Research and Forecasting meso-scale modelling group meeting, 22 November 2016, Wageningen, The Netherlands

SENSE Coordinator PhD Education

Dr. Peter Vermeulen