Dipl.-Ing. Thomas Mendlik

Statistical Tools to Quantify Uncertainty of Dependent Climate Change Projections from Multi-Model Ensembles

PHD THESIS

written to obtain the academic degree of a
Doctor of Engineering Sciences

Doctoral studies of Engineering Sciences at the doctoral school
“Mathematics and Scientific Computing”

Graz University of Technology

Supervisor:
Univ.-Prof. Dipl.-Ing. Dr.techn. Ernst Stadlober

Institut für Statistik

Graz, September 2016

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Graz, am . . . .

(Unterschrift)

Statutory Declaration

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

. . . .

date (signature)

Acknowledgements

During my PhD studies I was lucky to have the chance to work in large projects, where besides engaging with inspiring people, I got the opportunity to be a part in solving large scale scientific questions. Severeal of those questions asked do now form the backbone of this thesis. For this opportunity I would like to thank our former research team leader Dr. Andreas Gobiet, who worked hard to aquire all these projects, which for the biggest part also funded my research.

Financially I would like to thank the EU-FP7 framework for funding the project IMPACT2C and therefore making this research possible. My thanks also go to the Austrian Science Fund (FWF) for funding the project “Undersanding contrasts in high mountain hydrology in Asia” (UNCOMUN). Thanks to the Federal Ministry of Science, Research and Economy (BMWFW) for funding the OeAD grant Marietta Blau, which enabled my research visit abroad. I further acknowledge the World Climate Research Programme’s Working Group on Coupled Modelling, which is responsible for CMIP, and I thank the climate modeling groups for producing and making available their model output.

I would especially like to thank all of my collegues at the Wegener Center for all these years of wonderful collaboration. This work would not be what it is without all the skills I learned and the fun I had at this Institute. My special thanks here go to Armin Leuprecht, Georg Heinrich and Martin Jury, for all the programing knowledge I could pick up and for all the hours of inspiring discussions and good talks we had. My thanks also go to Sabine Tschürtz, Bettina Schlager and the entire administration, who with their uplifting spirits and their commitment keep this Institute going. I also want to thank Michael Pock for providing the LaTeX-environment used in this thesis.

My sincere gratitude goes to my supervisor Univ.-Prof. Dipl.-Ing. Dr.techn. Ernst Stadlober, who always took plenty of time whenever I needed anything and was prompt in answering any mail with any request I sent. I highly appreciate that. Also the thoroughness and detail with which he reviewed this thesis was beyond the scope of PhD supervision. I would also like to thank Prof Mat Collins for taking care of me and integrating me in the Exeter Climate Systems group during my visit in the UK.

And at last I want to thank all of my friends and my brother Mischa for their support and for all the fun I had with you. All those activities and new hobbies certainly helped to keep my sanity. Thanks to my parents who always care about me and who encouraged me in times I felt uncertain about my path. And thank you Barbara for your love, your good spirit and your patience with me - especially during the finishing phase.

Abstract

A global rise of surface temperature has been observed since the late 19th century and has been accelerating since the last couple of decades. As such a warming will undoubtedly have large impacts on humanity, the projection of a future climate still remains a challenging task. Climate model outcomes serve as the most detailed basis for such a future climate change analysis, but at the same time these models comprise of large uncertainties. This is due to simplifications of physical processes, model errors and the unknown future evolution of greenhouse gas concentrations. Dealing with these uncertainties is one of the major topics in modern climate research.

In this work we address some of the current scientific questions of interest, starting with the data processing of huge data sets of climate model ensembles and ending with the statistical inference on possible climate changes. The questions to be answered mostly arose within projects such as the EU-FP7 large scale project IMPACT2C[1], where the ultimate aim was to understand the impacts of climate change on sectors important to humanity such as water, energy, and agriculture, while accounting for the climate model uncertainties and communicating them to the public. During such projects the following three main topics for the thesis were developed:

In the first step, we present the R tool wux which enables automated retrieval and processing of entire ensembles of climate simulations. Besides generating datasets for extensive statistical analysis, this tool also provides routines for simple exploratory data analysis of such ensembles.

Second, we present a method to select a subset of representative climate models from such ensembles, a procedure often needed for further climate impact research like hydrological modelling. This method detects and accounts for model inter-dependencies and tries to conserve the information content of the entire ensemble.

And third, as a last point in a climate change analysis, we show how to extract all the information available in such an ensemble with a novel way for a statistical uncertainty analysis. The innovative part in this method is the explicit formulation of the inter-model dependencies as well as the assessment of the non-normally distributed nature of projected climate change, which so far has not been performed in literature. Using a hierarchical model approach, it is also possible to quantify the relative importance of the individual sources of uncertainty and thereby to account for natural internal variability on different time scales, as well as uncertainties induced by the climate models.

[1] www.atlas.impact2c.eu

Zusammenfassung

SEIT dem 19ten Jahrhundert kann man einen globalen Temperaturanstieg beobachten, der sich in den letzten Jahrzehnten zudem bescheunigt hat. Dieser Anstieg wird zweifellos Auswirkungen auf die Gesellschaft haben, allerdings ist es immer noch schwierig den Klimawandel genau zu modellieren. Die bislang präziseste Methode diesen zu beschreiben, ist mit Hilfe von globalen Klimamodellen. Diese Klimamodelle beinhalten allerdings diverse Vereinfachungen um physikalische Prozesse darzustellen und haben inherente Modellfehler, woraus sich bei unterschiedlichen Modellen unterschiedliche Klimawandelszenarien ergeben. Diese Unsicherheiten zu verstehen und zu beschreiben, ist eines der zentralen Ziele moderner Klimaforschung. Diese Arbeit beschreibt einige dieser Themen, welche sich zum Großteil aus Projekten, wie dem EU-FP7 Projekt IMPACT2C[2], ergeben haben. Folgende drei Punkte werden in dieser Arbeit präsentiert:

Als Erstes präsentieren wir das R Paket wux, welches gleich eine Vielzahl an Klimamodelldaten herunterladen kann um diese dann für eine statistische Datenanalyse vorzuprozessieren. Als zweiten Punkt dieser Arbeit leiten wir eine Methode zur geeigneten Auswahl von Klimamodellen her. Diese werden dann als Eingangsdaten für Modelle hergenommen, welche zusätzlich Auswirkungen des Klimawandels simulieren, wie zum Beispiel die Vorhersage von Wasserverorgungen in risikoreichen Regionen. Dabei ist es wesentlich, die Unsicherheitsspanne aller vorhandenen Klimamodelle zu berücksichtigen. Als letzten Punkt dieser Arbeit leiten wir ein statistisches Modell zur Schätzung der Verteilung von unterschiedlichen Klimaszenarien her. Neuartig an dieser Methode ist die Einbeziehung von Abhängigkeitsstrukturen unterschiedlicher Klimamodelle, welche in bisherigen Studien noch wenig Berücksichtigung fand. Auch die Standardannahme der Normalverteilung wird untersucht und durch eine schiefe Verteilung ersetzt. Dabei werden unterschiedliche Unsicherheitskomponenten mit Hilfe eines hierarchischen Ansatzes geschätzt und mit natürlicher Klimavariabilität verglichen.

[2] www.atlas.impact2c.eu

Contents

List of Figures

List of Tables

List of R Code

Acronyms

Symbols

iid independent and identically distributed. 75

wux Wegener Center uncertainty explorer. 25, 37

A

AL Alpine Region. 4, 41, 73, 111, 114, 133, 141

ANOVA analysis of variance. 33

C

CDO Climate Data Operators. 26, 53

CI confidence interval. 4, 73, 107, 135, 136, 142

CMIP3 Coupled Model Intercomparison Project Phase 3. 19, 21

CMIP5 Coupled Model Intercomparison Project Phase 5. 4, 13, 17, 38, 47, 73, 111, 118, 124, 136, 141, 142

CORDEX COordinated Regional Climate Downscaling EXperiment. 13

CP centred parameter. 95, 96

D

DJF winter. 63, 64, 133, 136

DP direct parameter. 95–97

E

EB Empirical Bayes. 118–120

ENSO El-Niño Southern Oscillation. 15

ESGF Earth System Grid Federation. 13, 39

G

GAR Greater Alpine Region. 45

GCM General Circulation Model. 3, 9–11, 13, 17, 19, 21, 22, 27, 37, 38, 40, 46, 47, 49, 50, 63, 64, 67, 69, 70, 81, 83, 111, 114, 117–120, 123, 124, 129, 133, 135, 136, 140–142

GHG greenhouse gas. 8–11

H

HURS relative humidity. 63, 64

I

IP Iberian Peninsula. 4, 73, 111, 114, 133, 136, 141

IPCC Intergovernmental Panel on Climate Change. 7, 11

J

JJA summer. 63, 111, 117

L

LM linear model. 133

LMM linear mixed-effects model. 76, 80, 140

LR likelihood ratio. 136

LRT likelihood ratio test. 4, 107, 147

LS least-squares. 118, 119

M

MAM spring. 63, 64

MLE maximum likelihood estimate. 95, 107–110, 130, 135

MME multi-model ensemble. 3, 4, 16, 17, 19, 27–29, 57, 63

MSE mean squared error. 20

N

NCL The NCAR Command Language. 26, 53

NetCDF Network Common Data Form. 3, 25, 26, 33, 37–39, 114, 147

O

OLS ordinary least-squares. 75, 76

P

PC principle component. 57, 59–62, 64, 66

PCA principle component analysis. 3, 57, 60, 61, 64

PDF probability density function. 4, 29, 30, 93, 94, 110, 142

PPE perturbed physics ensemble. 15–17

PR precipitation amount. 3, 63

R

RCM Regional Climate Model. 3, 4, 11, 13, 17, 45–47, 57, 63, 64, 66, 67, 69, 70

RCP Representative Concentration Pathway. 10, 11, 13, 47

RSDS global radiation. 63, 64

RV Random Variable. 75, 95

RVC relative variance change. 123, 124, 140

S

SC Scandinavian Region. 4, 73, 111, 114, 133, 136, 141

SN skew-normal. 100, 102–104, 130, 147

SN-LMM skew-normal linear mixed-effects model. 4, 97, 98, 142

SON autumn. 63, 64

T

TAS mean air temperature. 3, 63

U

UNFCCC United Nations Framework Convention on Climate Change. 7

V

VC variance component. 77, 81, 123, 124, 133, 135, 136, 140–142

W

WI Woodbury identity. 98, 100

WMO World Meteorological Organization. 7

WSS wind speed. 63

Preface

The following two publications emerged during the work of this thesis and play a central role in this document. Text passages in this thesis have also been used in those two publications:

Mendlik, T. and A. Gobiet (2016). 'Selecting climate simulations for impact studies based on multivariate patterns of climate change'. In: *Climatic Change* 135.3, pp. 381–393. DOI: 10.1007/s10584-015-1582-0.

Mendlik, T., G. Heinrich, A. Gobiet and A. Leuprecht (2016). 'From climate simulations to statistics - Introducing the wux package'. In: *Austrian Journal of Statistics* 45, pp. 81–96. DOI: 10.17713/ajs.v45i1.98.

During the work on this thesis following additional publication emerged, which however do not directly contribute to the content of the PhD thesis:

Fox Maule, C., T. Mendlik and O. B. Christensen (2016). 'The effect of the pathway to a two degrees warmer world on the regional temperature change of Europe.' In: *Climate Services*. DOI: dx.doi.org/10.1016/j.cliser.2016.07.002.

Heinrich, G., A. Gobiet and T. Mendlik (2014). "Extended regional climate model projections for Europe until the mid-twentyfirst century: combining ENSEMBLES and CMIP3". In: *Climate Dynamics* 42.1, pp. 521–535. DOI: 10.1007/s00382-013-1840-7.

Ravazzani, G., M. Ghilardi, T. Mendlik, A. Gobiet, C. Corbari and M. Mancini (2014). "Investigation of climate change impact on water resources for an Alpine basin in northern Italy: Implications for evapotranspiration modeling complexity". In: *PLoS ONE* 9.10. Ed. by J. M. Dias, e109053. DOI: 10.1371/journal.pone.0109053.

Stoffel, M., T. Mendlik, M. Schneuwly-Bollschweiler and A. Gobiet (2014). "Possible impacts of climate change on debris-flow activity in the Swiss Alps". In: *Climatic Change* 122.1, pp. 141–155. DOI: 10.1007/s10584-013-0993-z.

Vautard, R., A. Gobiet, S. Sobolowski, E. Kjellström, A. Stegehuis, P. Watkiss, T. Mendlik, O. Landgren, G. Nikulin, C. Teichmann and D. Jacob (2014). 'The European climate under a 2 °C global warming'. In: *Environmental Research Letters* 9.3, p. 034006. DOI: 10.1088/1748-9326/9/3/034006.

Wilcke, R. A. I., T. Mendlik and A. Gobiet (2013). “Multi-variable error correction of regional climate models”. In: *Climatic Change* 120.4, pp. 871–887. DOI: 10.1007/s10584-013-0845-x.

1 Introduction

A changing climate causes changes in systems of direct value to humanity. It is likely, for example, that regional decrease of precipitation will cause a decrease in nearby river run-offs, which, as a consequence, will change the crop yield of surrounding wheat fields.

Such processes can be simulated with a magnitude of numerical models. The most detailed information on future climate change is provided by General Circulation Models (GCMs), which simulate the behaviour of Earth's atmosphere and ocean. These models provide information on a global change of climate parameters such as mean air temperature (TAS) and precipitation amount (PR). To obtain more regionalised information on climate change, these GCMs are often refined with Regional Climate Models (RCMs) and empirical-statistical post-processing methods. The outcome of these models can be fed into climate impact models, which take meteorological parameters (e.g. temperature and precipitation) as input and give e.g. crop-yield as an output. Future projections like these can then be used as a basis for political and economical long-term decisions.

However, GCMs and RCMs are subject to considerable uncertainties (e.g. Tebaldi and Knutti 2007) originating from the chaotic behaviour of the climate system and the unknown future evolution of greenhouse gas concentrations and other forcing agents of the climate system, as well as simplifications and errors in climate models. Those inherent uncertainties are often investigated using multi-model ensembles (MMEs), which also challenges climate change impact assessments to base their investigations on multi-model climatological input.

In this thesis we present three tools to handle uncertainty in climate research:

1. The R package wux can download entire MMEs in Network Common Data Form (NetCDF) format and aggregate each climate model to the desired spatial and temporal resolution to obtain a data.frame for further statistical analysis (Mendlik, Heinrich, A. Gobiet et al. 2016; Mendlik, Heinrich and Leuprecht 2015).

2. A tool to select climate models from a MME in order to run impact models, while accounting for the climate model uncertainty and dependencies (Mendlik and A. Gobiet 2016). To do that, the method first detects relevant patterns of climate change with a principle component analysis (PCA). Based on these patterns, groups of similar climate simulations are found with a hierarchical clustering method. We present an example application used in the EU-FP7 project

IMPACT2C[1] where a subset of RCMs had to be selected as an input for several climate impact models, based on a multitude of climate parameters across the entire European continent.

3. A tool to estimate the probability density function (PDF) for the projected climate change of a MME, while accounting for climate model dependencies, the unbalanced data structure and non-normality. The skew-normal linear mixed-effects model (SN-LMM), which is used to model the data, also allows for explicit estimation of the individual sources of uncertainty of the MME. We derive confidence intervals (CIs) for each parameter estimate with a second-order Wald approximation, with likelihood ratio test (LRT) statistics, and with non-parametric block-bootstrap and parametric bootstrap techniques. As the last part, we present a case study to quantify the projected seasonal temperature climate change of the Coupled Model Intercomparison Project Phase 5 (CMIP5) ensemble over three different European regions: The Alpine Region (AL), the Iberian Peninsula (IP) and the Scandinavian Region (SC).

The thesis is split into four major Parts: Part I introduces the term climate change and briefly explains climate models (Chapter 2). In Chapter 3 the role of statistics and the challenges to analyse ensembles of these climate models is discussed in-depth. At last, Chapter 4 in Part I introduces the three tools presented above in more depth and provides a literature review. Part II of the thesis explains the functionality of the wux package, which has been used to pre-process all data of this work (Chapter 6-8). The text in this part has also been used in a peer-reviewed publication (Mendlik, Heinrich, A. Gobiet et al. 2016). Part III introduces the method to sub-select climate models from a MME to run impact models. Chapter 9 explains the underlying idea behind the method and Chapter 10 presents a European model-selection case study of an RCM ensemble. Also in this part the text has been used for a publication in a peer-reviewed journal (Mendlik and A. Gobiet 2016). Finally, Part IV presents the SN-LMM to estimate the expected climate change and the uncertainties in a MME. Chapter 12 derives a class of hierarchical models to address dependencies and which is then extended in Chapter 13 to fit skewed data. Chapter 14 then introduces different methods on how to obtain CIs for the estimates. And finally, Chapter 15 shows a case study of uncertainty quantification of the CMIP5 ensemble. The R code for this uncertainty analysis is outlined and summarised in Appendix B. Some of the more technical proofs can be found in Appendix A.

[1] www.atlas.impact2c.eu

Part I

From Climate to Statistics

2 Climate and Climate Change

2.1 Climate, Climate Variability and Climate Change

Climate, which can be understood as the "average weather" is defined by the Intergovernmental Panel on Climate Change (IPCC) as:

> *Climate in a narrow sense is usually defined as the average weather, or more rigorously, as the statistical description in terms of the mean and variability of relevant quantities over a period of time ranging from months to thousands or millions of years. The classical period for averaging these variables is 30 years, as defined by the World Meteorological Organization (WMO). The relevant quantities are most often surface variables such as temperature, precipitation and wind. Climate in a wider sense is the state, including a statistical description, of the climate system* (IPCC 2013).

The climate system is an interactive system consisting of the atmosphere, land surface, snow and ice, oceans and other bodies of water and living things. Due to internal dynamics and due to external factors, the climate system keeps on constantly changing. And so does the climate. The IPCC therefore defines *climate variability* as:

> *Climate variability refers to variations in the mean state and other statistics (such as standard deviations, the occurrence of extremes, etc.) of the climate on all spatial and temporal scales beyond that of individual weather events. Variability may be due to natural internal processes within the climate system (internal variability), or to variations in natural or anthropogenic external forcing (external variability).*

In this work we differentiate between *natural climate variability* and *climate change*, which is attributable purely to human activity. We therefore consider the definition of the United Nations Framework Convention on Climate Change (UNFCCC) defining climate change as

> *...a change of climate which is attributed directly or indirectly to human activity that alters the composition of the global atmosphere and which is in addition to natural climate variability observed over comparable time periods.*

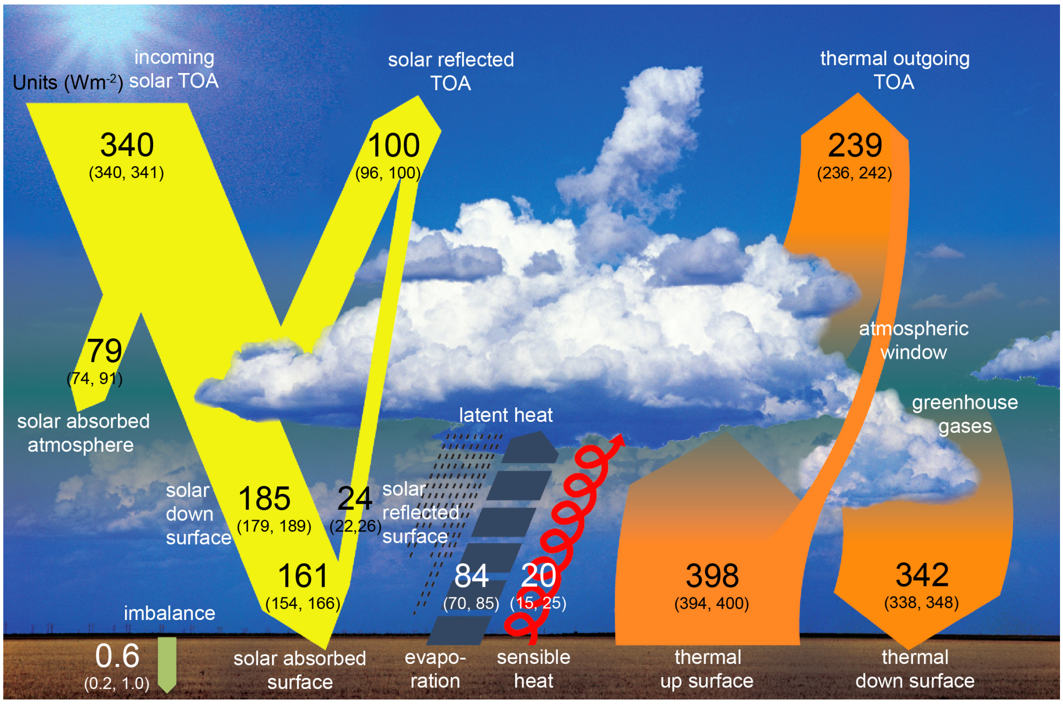


Figure 2.1: Global mean energy budget under present-day climate conditions. The numbers represent the energy fluxes in W/m^2 with uncertainty ranges. Source: Hartmann et al. 2013.

2.2 The Greenhouse Effect

The climate system is powered by solar radiation. The incoming solar radiation during daytime on top of the Earth's atmosphere is about 1370 W/m^2. Due to to planet's spherical shape, the amount of energy averaged on the entire planet is about 1/4 of that, being about 340 W/m^2 as shown in Figure 2.1. About 30 % of this radiation is reflected directly back into space due to the albedo effect of the clouds and light coloured earth surfaces such as ice, snow and deserts. The remaining radiation (about 240 W/m^2) is absorbed by Earth's surface and atmosphere. The same amount of energy must be radiated back in form of longwave radiation. Emitting this amount of energy, the Earth would have about −19 °C, but the planet has roughly 14 °C on average.

The reason for such a warm earth is the greenhouse effect caused by greenhouse gases (GHGs) which create a blanket around the planet which absorbs outgoing longwave radiation and emits parts back to the surface and parts into space. The most potent GHGs are water vapour and CO_2, whereas the two gases occurring most often, namely nitrogen and oxygen, have no such effect. Clouds have a blanket effect as well (which can be felt with warm cloudy nights), but due to the albedo effect, they also reflect substantial amount of incoming radiation from the sun.

Releasing additional GHGs enhances this blanketing effect. Human activity has increased the amount of CO_2 in the atmosphere by more than 35% since the industrial era, primarily by combustion of fossil fuels and by removing forests (Solomon 2007). These human-made changes in the atmospheric composition have major consequences for the climate.

2.3 Climate Models

Climate models are mathematical representations of the climate system and they investigate its response to various forcings. The complexity of such models range from simple energy-balance models (Section 2.2) to General Circulation Models (GCMs) representing more detailed physical processes and simulating the interactions and feedbacks in the atmosphere, ocean, cryosphere and land surface.

The Simplest Climate Model

The very simple zero-dimensional climate model (Saha 2008) represents the radiative balance from Section 2.2 and can be written as

$$(1-a)S\pi r^2 = 4\pi r^2 \epsilon \sigma T^4$$

with S being the solar constant (around $1370\,\mathrm{W/m^2}$), a being the planets average albedo ($30\,\%$), r being Earth's radius, ϵ is the effective emissivity of the planet (estimated to be around 0.612) and σ being the Stefan-Boltzmann constant, being about $5.67 \times 10^{-8}\ \mathrm{J\ K^{-4}\ m^{-2}\ s^{-1}}$. The left side of the equation shows the incoming solar energy and the right side the outgoing energy from Earth. T is the equilibrium temperature of Earth. Solving for this temperature yields 288 K (being around 15 °C) as the planet's average temperature. The emissivity and albedo parameter account for the greenhouse effect as described in Section 2.2.

General Circulation Models (GCMs)

GCMs, in contrast, mathematically describe the general circulation of the atmosphere and ocean. They are based on fundamental principles of physics, being the conservation of mass, energy and momentum. The conservation of mass says that inside a given volume, no mass can be generated out of nowhere, which means that mass can change only by in- and outflows. Conservation of energy (first law of thermodynamics) means that the total energy remains constant and can only be transformed from one form to another. The conservation of momentum is modelled with the Navier-Stokes equations of fluid motion and accounts for magnitude and direction of velocity in reference to

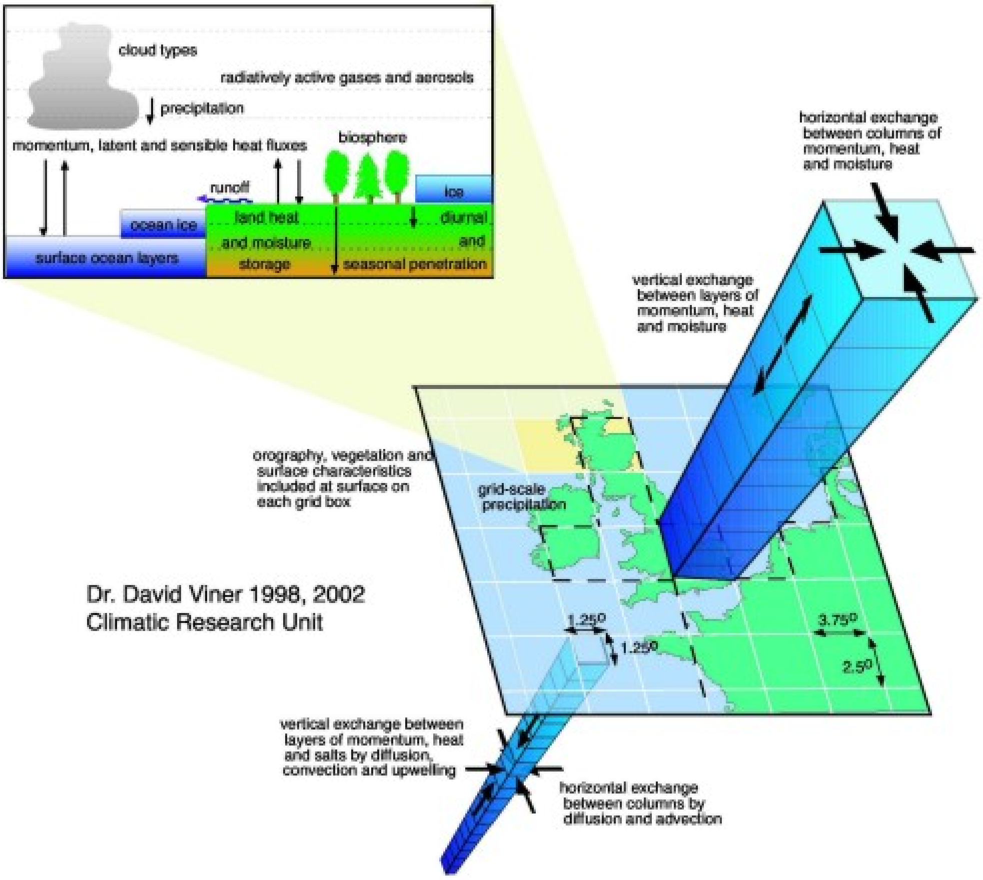

cloud types
radiatively active gases and aerosols
precipitation
momentum, latent and sensible heat fluxes
biosphere
runoff
ice
ocean ice
land heat and moisture storage
diurnal and seasonal penetration
surface ocean layers
horizontal exchange between columns of momentum, heat and moisture
vertical exchange between layers of momentum, heat and moisture
orography, vegetation and surface characteristics included at surface on each grid box
grid-scale precipitation
Dr. David Viner 1998, 2002
Climatic Research Unit
1.25º
1.25º
3.75º
2.5º
vertical exchange between layers of momentum, heat and salts by diffusion, convection and upwelling
horizontal exchange between columns by diffusion and advection

Figure 2.2: Schematic description of a GCM. Source: www.ipcc-data.org.

present forces such as the pressure gradient, gravity and the Coriolis force caused by a rotating Earth. These three fundamental principles form the primitive equations of the model, forming a set of nonlinear differential equations. These equations need to be simplified and solved numerically by discretising Earth into a 3-dimensional longitude-latitude-height grid with discrete time steps (Figure 2.2). A typical horizontal resolution of a GCM can be between 100 km and 500 km.

The higher the resolution, the more processes can be included into the model. Processes which cannot be represented explicitly, either due to this lack of resolution (e.g. cloud processes and turbulence) or due to their complexity (like biochemical processes in vegetation) need to be parameterized. These parametrizations are one of the reasons why different GCMs yield different future climate projections (see Section 3.1).

In order to produce long-term climate change projections, socioeconomic scenarios reflecting human behaviour (like GHG emission and land-use changes) are defined and fed into the GCMs as the anthropogenic driving element. Currently, these scenarios are the Representative Concentration Pathways (RCPs), being four consistent sets of time-dependent forcing projections that could potentially be realised with more than one underlying socioeconomic scenario. They are named after their approximate value of radiative forcing (in $\mathrm{W/m^2}$) at 2100 relative to the pre-industrial era. RCP2.6 (the lowest of the four) peaks at $3.0\,\mathrm{W/m^2}$ and then declines to $2.6\,\mathrm{W/m^2}$ in 2100, RCP4.5 (medium-low) and RCP6.0 (medium-high) stabilise after 2100 at 4.2 and $6.0\,\mathrm{W/m^2}$ respectively, while RCP8.5 (highest) reaches $8.3\,\mathrm{W/m^2}$ in 2100 on a rising trajectory (Collins, Knutti et al. 2013).

Regional Climate Models (RCMs)

As the resolution of GCMs is too coarse to get regional and local climate information, one can run Regional Climate Models (RCMs) to obtain a more detailed view. RCMs are dynamically downscaled based on large-scale climate conditions often provided by a GCM. In contrast to GCMs, RCMs do not model the entire planet, but rather a specific region nested within a GCM. Modern RCMs have a typical horizontal resolution of 12.5 km to 50 km and can resolve processes which are not represented in GCMs.

Climate Models Ensembles

The climate system is always in motion because of the dynamic interactions between its components. These natural fluctuations (natural climate variability) make it difficult to assess for near-term climate change as they may overshadow the effect of external forcings like the increase of GHGs. For example, if we observe a new temperature record this summer, it is difficult to attribute this anomaly to naturally occurring climate variability or to climate change, as we only have a single observed "realisation" of climate. Simulating several climate models with different external forcings helps to understand the effect of human behaviour on the climate system. For the next decades, however, these climate model ensembles project that external forcings will dominate internal variability. The choice of scenarios of future GHG emissions (RCPs), has a huge impact on the long-term change of mean global temperature (IPCC 2013).

However, different climate models project different climate changes, as there are various alternative and equally plausible ways to numerically represent and approximate the climate system. We discuss the consequences and reasons for these uncertainties in Section 3.1. This model diversity is another important reason to consider ensembles of climate models and so to quantify climate change uncertainty. However, the contributions of these models happen on a voluntary basis and are neither systematic nor comprehensive. In addition, certain model inadequacies are common to all models and different models have different strengths and weaknesses.

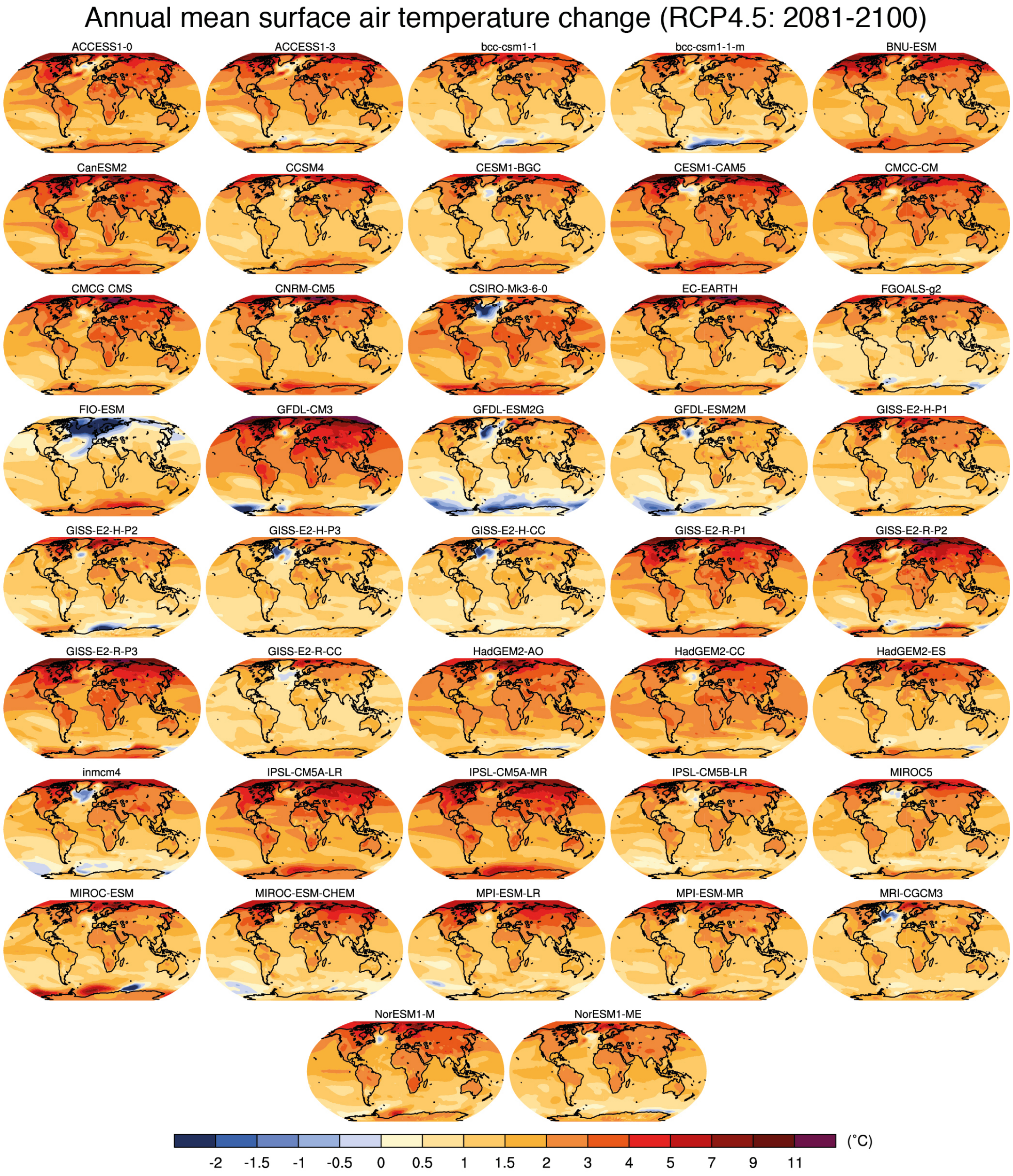


Figure 2.3: Projected CMIP5 temperature climate changes from 1981-2000 to 2081-2100, as presented in the fifth assessment IPCC report. The individual GCMs are driven with the RCP4.5 external forcing. Source: Collins, Knutti et al. 2013.

The current state-of-the-art of GCM ensemble has been established as part of the Coupled Model Intercomparison Project Phase 5 (CMIP5) project, (Taylor, Stouffer and Meehl 2012), which has been used throughout the current IPCC report (IPCC 2013). This project involves a worldwide coordination of GCM experiments, to synchronise the model inputs, model diagnostics and the distribution in data archives with the Earth System Grid Federation (ESGF). Figure 2.3 shows the different projected temperature climate change signals at the end of the century of 42 GCMs from the Coupled Model Intercomparison Project Phase 5 (CMIP5) ensemble under the medium-low scenario RCP4.5.

Ensembles like these can be used to interpret large-scale phenomena like temperature change, however, for more local and more complex processes like regional precipitation, ensembles of RCMs might be more accurate. The current state-of-the-art RCM ensembles are coordinated with the COordinated Regional Climate Downscaling EXperiment (CORDEX) (Jacob et al. 2013), where each continent is modelled separately, for example Europe within the EURO-CORDEX project. Different GCMs are dynamically downscaled to account for the uncertainty range of the CMIP5 experiments (see Figure 3.1). The European predecessor RCM ensemble was the ENSEMBLES project Hewitt and Griggs 2004.

Regional climate modelling, however, introduces another source of uncertainty, as different RCMs potentially yield different results even when driven by the same GCM.

2.4 From Climate Models to Climate Impacts

To make decisions for climate change mitigation or adaptation, it might not be enough to consider climate change signals of temperature or other meteorological parameters, but rather to inspect response of climate change on parameters of direct value for humanity. A decision maker might be interested whether to build a hydro-power plant in a certain region or not, and would therefore like to know if the river runoff is likely to change or not. This is where climate impact models come to play. These models usually take climate model output (usually regionalised, post-processed to fit the user needs) as an input variable to simulate the impacts of a changing climate (like the hydrological impacts in the Alpine region described in Ravazzani et al. 2014). However, it is often difficult to make such decision because of the inherent uncertainty. Each step in climate research is a potential source of this uncertainty, forming a cascade of uncertainty (Figure 2.4, Mearns, Giorgi et al. 2003). This usually consists of:

1. Specifying alternative future emissions reflecting human behaviour (e.g. CO_2)

2. Converting emissions to concentrations

3. Converting concentrations to climate forcing (e.g. RCP)

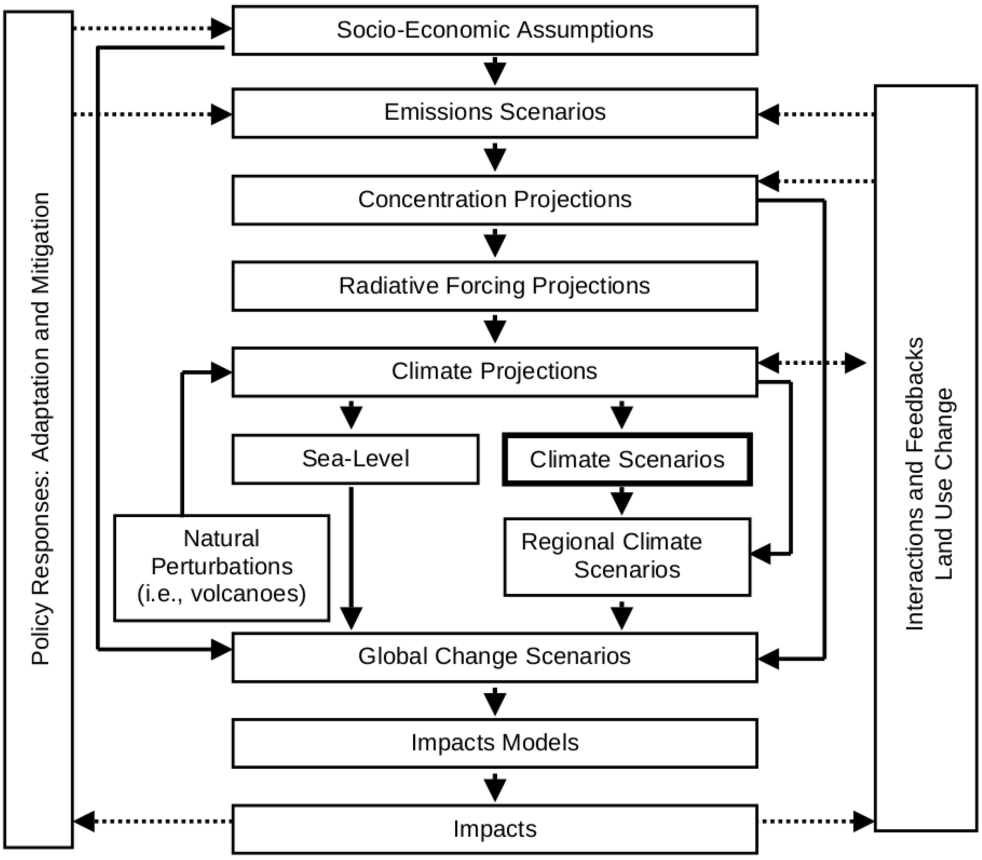

Socio-Economic Assumptions
Emissions Scenarios
Concentration Projections
Radiative Forcing Projections
Climate Projections
Sea-Level
Climate Scenarios
Natural Perturbations (i.e., volcanoes)
Regional Climate Scenarios
Global Change Scenarios
Impacts Models
Impacts
Policy Responses: Adaptation and Mitigation
Interactions and Feedbacks
Land Use Change

Figure 2.4: Cascade of uncertainty. Source: Mearns, Giorgi et al. 2003.

4. Modelling the climate response to a given forcing (e.g. RCMs or GCMs)
5. Converting the model response into inputs for impact studies
6. Modelling impacts

It is difficult to manage this cascade for impact studies, as only small subsets of potential pathways through the cascade will usually be explicitly modelled. The aim therefore is to find and develop techniques which consider a representative range of climates and to find probabilistic techniques to manage the large ranges of possible climate changes (Mearns, Giorgi et al. 2003).

3 Quantify Uncertainty: The Role of Statistics

3.1 Where does Climate Model Uncertainty come from?

The outcomes between climate models can vary quite substantially, even though the underlying physical principles might be the same. Such differences do occur when modelling the present climate as well as when making future projections of the climate system. This discrepancies are usually referred to as "uncertainties". The aim of this chapter is to shed light into the reasons behind those uncertainties.

Basically, when considering a climate simulation, there are three big sources of uncertainty in an ensemble when considering climate change: Natural internal variability, model uncertainty and scenario uncertainty.

1. **Natural internal variability** is the part of uncertainty which occurs due to internal processes in the climate system in the absence of any radiative forcing of the planet. Processes in the atmosphere, for example, can cause variability on a nearly instantaneous time scale (e.g. cloud formation) up to variability over years (e.g. troposphere-stratosphere exchange). Such natural variabilities can be on even longer time scales when considering processes in the ocean or large ice-sheets. Coupling of such components can also lead to strong internal variability, such as the El-Niño Southern Oscillation (ENSO). Such internal variabilities are important, as they can have the potential to reverse (for a decade or so) the longer-term trends that are associated with anthropogenic climate change (Hawkins and Sutton 2009). In a climate model ensemble, such variabilities can be assessed when changing the initial conditions of a particular climate model. In literature this variability is often addressed as *initial condition uncertainty* (Collins, Tett and Cooper 2001).

2. **Model uncertainty** is due to the fact that climate models are only approximations of the real climate system and therefore induce errors. Two main components are the parametric and the structural uncertainty.

   Parametric uncertainty arises as not all processes can be resolved explicitly on the model grid, so their influence on larger-scale processes must be parameterized empirically or by using expert judgement. Prominent processes are for example clouds which need to be parameterized. Without describing the actual cloud element, physical theories are used to describe the statistics of the cloud fields (e.g. the fractional cloudiness or the area-averaged precipitation rate, Solomon 2007). Parameterization uncertainties can be explored using perturbed physics ensembles (PPEs), where the parametrizations of a single climate model are changed systematically. However, the PPE approach does not reflect the full range of uncertainties as using only one model does not account for structural uncertainty (see next paragraph).

Structural uncertainty emerges because it is simply not feasible to describe all processes in a numerical climate model. It has to be decided which physical processes should be explicitly modelled and what parameterization schemes should be used. Usually, model uncertainty which does not emerge from different parameter values (i.e. not being part of parametric uncertainty), is said to be structural uncertainty (Tebaldi and Knutti 2007). It is not possible for one climate model to describe this structural uncertainty, regardless of the range of parameter values. Instead, several, ideally independently developed models, have to be considered simultaneously. This is the motivation for using multi-model ensembles (MMEs). The difficulty with this type of uncertainty is designing MMEs, as processes and parameterization schemes cannot be as easily perturbed as parameter values in a PPE.

3. **Scenario uncertainty** reflects the unknown future behaviour of humans (as world economic and social development) which leads to different possible “pathways” of future greenhouse gas emissions leading to a different radiative forcing of the climate system. This is a major part of the projected uncertainty in climate change analysis, but the difficulty is that it remains impossible to link likelihoods or probabilities to different pathways. Therefore climate change analysis is often performed conditional on a specific pathway.

Studies like Hawkins and Sutton 2009 tried to quantify the relative importance of these three sources of uncertainty in a multi model ensemble and so deriving a signal-to-noise ratio of the projected climate change signal. Also an important field of research is the so-called climate detection and attribution. There the aim is to understand the anthropogenic causes and natural external forcings (e.g. changes in solar radiation, volcanism) of climate change and distinguish those from changes due to internal climate system processes.

3.2 Statistical Challenges

In the previous chapters it has been shown that it requires a MME to quantify the components of climate change uncertainty. However, climate models in such a multi-model ensemble cannot be regarded as a simple random draw from common distributions. One reason is the lack of statistical experimental design: Models are developed voluntarily from institutions worldwide, hence model components such as parametrizations are not systematically changed. Also, certain institutions provide more models than others, leading to an overall unbalanced design. MMEs are therefore often called *ensemble of opportunity*. In addition, climate models (even across developing institutions) are known to share certain components which leads to inter-model dependencies, which makes it difficult to justify the independence assumption when quantifying the uncertainty of such an ensemble with a statistical model. Because of these dependencies, the multi-model ensemble can be expected to be even more unbalanced. Another challenge when interpreting such an ensemble is the explicit formulation of the statistical assumptions of the modelled and the observed climate system: Weather climate models are supposed to be distributed around the observed *truth*, or weather earths climate should be interpreted as one draw from all possible climates. Also there have been only very few studies which explicitly formulated the model inter-dependencies in a statistical framework and account for the imbalanced data structure. The following Chapters discuss those topics in more detail.

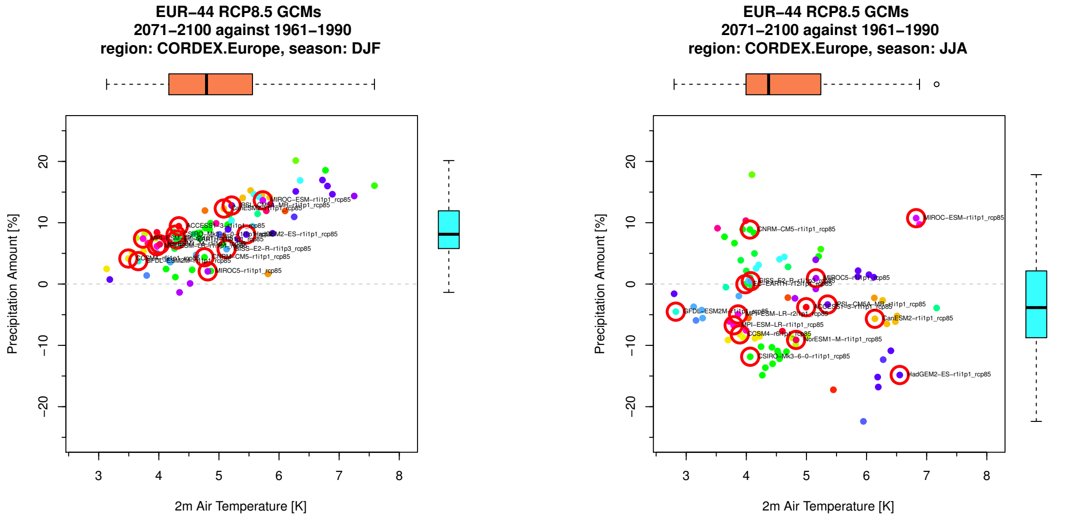
Lack of Experimental Design

A multi-model ensemble is often called an *ensemble of opportunity* (Tebaldi and Knutti 2007), as climate simulations are contributed from anyone who is willing to do so. There is no underlying experimental design in a statistical sense where the components of uncertainties, like parametrizations or structural components (Section 3.1) are sampled in a systematic way.

Another problem is that such a multi-model ensemble will probably not span the full range of behaviour or uncertainty. The reason is that usually climate simulations are tuned to match the observed climate. Once a best setup is found, the model is submitted to the ensemble. It is rarely the case that other parameterization settings are sought, which would also yield a satisfactory observation agreement but rather project a different future. In contrast to a MME, PPEs (Collins, Booth et al. 2006, Murphy et al. 2007) do sample this uncertainty component.

As a consequence it can be argued that any quantification of uncertainty will yield wrong estimates. Alternatively one can analyse the ensemble in a more qualitative way, by finding several representative climate scenarios, and work on a case-study basis (Whetton et al. 2012). Recently, several methods have emerged on how to select such a sub-ensemble (Cannon 2015, Mendlik and A. Gobiet 2016, Zubler et al. 2015).

The problem of a lack of an experimental design is even amplified when designing an ensemble of physically downscaled RCMs. An RCM is driven by a GCM and this forcing has a huge effect on the downscaled simulation (Heinrich, A. Gobiet and Mendlik 2014). Because of computational restrictions it is not possible to run all combinations of RCMs

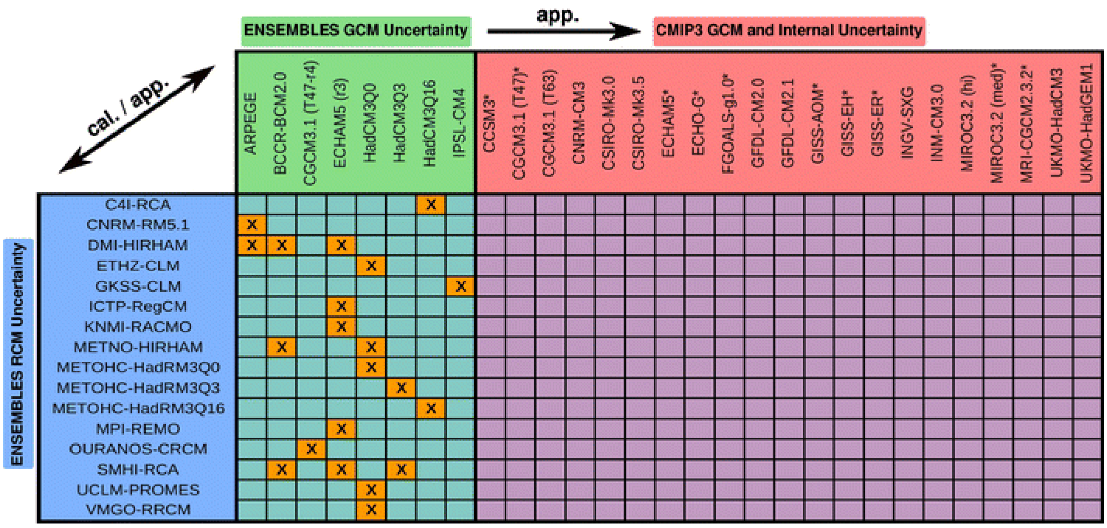


Copyright © 2016 by Wegener Center

Copyright © 2016 by Wegener Center

Figure 3.1: Climate Change signals for GCM of the CMIP5 ensemble over European domain (left: winter, right: summer). The marked simulations were selected for downscaling with an RCM within the current EURO-CORDEX project.

with every GCM (Figure 3.2), so a selection of representative GCMs should be applied here. This approach is pursued in the current coordinated downscaling experiment EURO-CORDEX (Jacob et al. 2013), where driving GCMs are selected to sample the extremes in temperature and precipitation changes (Figure 3.1). In a similar way the Australian regional ensemble NARCliM has been designed to run 12 simulations, so that the driving GCMs span the uncertainty range of projected future temperature and precipitation changes (Evans, Ji, C. Lee et al. 2013). For the North-American regional ensemble NARCCAP, four driving GCMs have been selected which should be downscaled with 6 RCMs (Mearns, Gutowski et al. 2009). In order not to run all possible 24 combinations they used a balanced fractional factorial design to reduce the amount to 12 simulations. Though this design was explicitly chosen to quantify uncertainty using a Bayesian probabilistic approach (Tebaldi, R. L. Smith et al. 2005), it is not clear whether the range of possible climate changes has been sampled as in EURO-CORDEX or in NARCliM. The predecessor project from EURO-CORDEX, the ENSEMBLES project (Hewitt and Griggs 2004) did not use any type of sampling design, so the resulting 21 simulations were unbalanced with respect to the driving GCM. Some methods have been developed (Déqué, Rowell et al. 2007, Déqué, Somot et al. 2011) to account for the unbalance and fill up the missing simulations from the RCM-GCM matrix (Figure 3.2). As those approaches deal with the unbalanced nature of the ensemble dataset, there


ENSEMBLES GCM Uncertainty
app.
CMIP3 GCM and Internal Uncertainty
cal. / app.
ENSEMBLES RCM Uncertainty

	ARPEGE	BCCR-BCM2.0	CGCM3.1 (T47-r4)	ECHAM5 (r3)	HadCM3Q0	HadCM3Q3	HadCM3Q16	IPSL-CM4	CCSM3*	CGCM3.1 (T47)*	CGCM3.1 (T63)	CNRM-CM3	CSIRO-Mk3.0	CSIRO-Mk3.5	ECHAM5*	ECHO-G*	FGOALS-g1.0*	GFDL-CM2.0	GFDL-CM2.1	GISS-AOM*	GISS-EH*	GISS-ER*	INGV-SXG	INM-CM3.0	MIROC3.2 (hi)	MIROC3.2 (med)*	MRI-CGCM2.3.2*	UKMO-HadCM3	UKMO-HadGEM1
C4I-RCA							X																						
CNRM-RM5.1	X																												
DMI-HIRHAM	X	X		X																									
ETHZ-CLM					X																								
GKSS-CLM								X																					
ICTP-RegCM				X																									
KNMI-RACMO				X																									
METNO-HIRHAM		X			X																								
METOHC-HadRM3Q0					X																								
METOHC-HadRM3Q3						X																							
METOHC-HadRM3Q16							X																						
MPI-REMO				X																									
OURANOS-CRCM			X																										
SMHI-RCA		X		X		X																							
UCLM-PROMES					X																								
VMGO-RRCM					X																								

Figure 3.2: GCM-RCM matrix for ENSEMBLES RCMs. The orange coloured cells marked with X's indicate the available simulations and empty cells represent the missing GCM–RCM combinations. The models spanning the RCM and GCM uncertainty of ENSEMBLES are highlighted in blue and green, respectively. Additional uncertainty due to the CMIP3 GCMs is displayed in red.Source: Heinrich, A. Gobiet and Mendlik 2014.

still remains the problem of a possible underestimation of the range of climate changes, as an arbitrary subset of GCMs has been selected out of an ensemble of GCMs. The method developed by Heinrich, A. Gobiet and Mendlik 2014 accounts for this problem, as it extends the methods of filling the RCM-GCM matrix to GCMs which have not been downscaled at all (Figure 3.2). This way one can account for the possible underestimation of climate changes by neglecting extreme GCMs.

Statistical Frameworks to interpret MME

In order to properly quantify the uncertainty of a changing climate, it is important to understand the sampling scheme of the MME. Understanding the scheme and the role of the observed climate allows to explicitly formulate a statistical framework to interpret the climate model ensemble.

One paradigm, which has been the basis of many studies, assumes that all climate models stem from a distribution centred around the "true" (observed) climate. In such a *truth plus error* framework, the error (i.e. model minus observation) would converge to zero when adding independent simulations to the ensemble. It has been shown in several studies that this paradigm is not defensible (Abramowitz and Bishop 2015). There are mainly three reasons for this drawback:

1. In certain regions, climate models are known to have a common discrepancy from the observed climate due to being only approximations to the real climate (Knutti 2010). Adding new models to the ensemble makes the average error converge to this shared bias and not to zero. Because of this common bias, when looking at the marginal distribution, the climate models are also not independent any more.

2. Even when not accounting the model errors, climate simulations have been shown not to be statistically independent due to shared components among models (see next section, Masson and Knutti 2011).

3. Even if the scientist would be able to create an ensemble of perfect simulations (not violating 1.) which are all statistically independent (not violating 2.), the "truth plus error" paradigm would assume that our observed climate is deterministic because it does not account for the influence of internal variability such as El Niño-Southern Oscillation (see Section 3.1): Due to the chaotic nature of the climate system, an "alternative" planet earth would likely have a slightly different observed climate. This internal variability will be common to all climate simulations when considering model-minus-error, so the error mean will again not converge to zero, again inducing marginally dependent climate simulation errors.

These issues have led to alternative frameworks such as the paradigm of *exchangeability*: the observed climate and climate models are treated as being random variables stemming from the same distribution and therefore accounting for internal variability and hence accounting for 3). In this case, increasing the amount of simulations would not decrease uncertainty indefinitely (Knutti, Abramowitz et al. 2010, Annan and Hargreaves 2010, Rougier, Goldstein and House 2013). The study of Chandler 2013 extended this approach by adding the concept of conditional exchangeability: The simulations and the observed climate do not stem from the same distribution, rather the simulations given the observed climate do. This way, shared discrepancies (1) are accounted for.

The study from Bishop and Abramowitz 2012 coined the term *replicate earth* with the same idea that observations and simulations are drawn from the same distribution. If simulations would be perfect representations of the real climate, the quantified uncertainty would purely reflect internal variability (3). They present a method to find optimal weights to decrease the mean squared error (MSE) of bias-corrected simulations (accounting for 1) to the observed climate accounting for correlations of the model errors (accounting for 2).

One big problem when incorporating the observed climate to the ensemble of simulations is that the relation of observed climate and simulations may not be the same in future projections. So usually some stationarity assumptions have to be made. Chandler 2013 incorporated a parameter to account for the future relation of the observed (yet unknown) climate to the simulations in his Bayesian model. Abramowitz and Bishop 2015 checked the method of Bishop and Abramowitz 2012 for future relations of observed and simulated climate by incorporating the so called “perfect-model” approach: They assume a particular simulation as being the observation which enables checking for future relationships.

Violation of Independency Assumption

In current statistical analysis of climate model ensembles, the simulations are considered as independent in the sense that every model contributes additional information (e.g. Tebaldi, R. L. Smith et al. 2005, Tebaldi and Knutti 2007, Buser, Künsch and Weber 2010, Fischer et al. 2012). However, if the simulations make same simplifications in parameterizing processes or share the numerical schemes to describe processes, their deviations from the true climate system or from other simulations will be similar. For example, simulations sharing the same computer codes to describe atmospheric processes will tend to project similar climates in contrast to simulations using a complete different scheme. The violations of independence lead to an underestimation of the climate change uncertainty as well as to a biased estimate of the expected change in case of unbalanced ensembles. This problem has been emphasised in several studies (e.g. Knutti 2010, Knutti, Furrer et al. 2010, Knutti, Abramowitz et al. 2010, Mearns 2010, Pirtle, Meyer and Hamilton 2010, Storch and Zwiers 2013), but so far only a few statistical methods have been proposed in literature to tackle the dependency issue.

The work by Pirtle, Meyer and Hamilton 2010 discusses model dependency and their causes in detail. It raises the concern about “robustness” when several models coincide, as there is no metric of dependence so far. They argue that the attention of GCM developers should be devoted much more on model independence, and finding methods to understand model agreement is a crucial step in climate research. At the same time another discussion paper (Abramowitz 2010) emerges, which stresses that model independence and model performance are two unrelated properties of model projections Therefore, model agreement, as a desired property of a “well performing” ensemble might not be that desirable at all, as this might just mean that the models are strongly dependent. This is particularly the case in the study of Tebaldi, R. L. Smith et al. 2005, where GCMs have been up-weighted if they agree in their climate projection, as this agreement is interpreted as good model performance. However, from the model dependency point of view, the approach should be the other way around: models which agree might just be highly dependent and should therefore be down-weighted because of double-counting similar information.

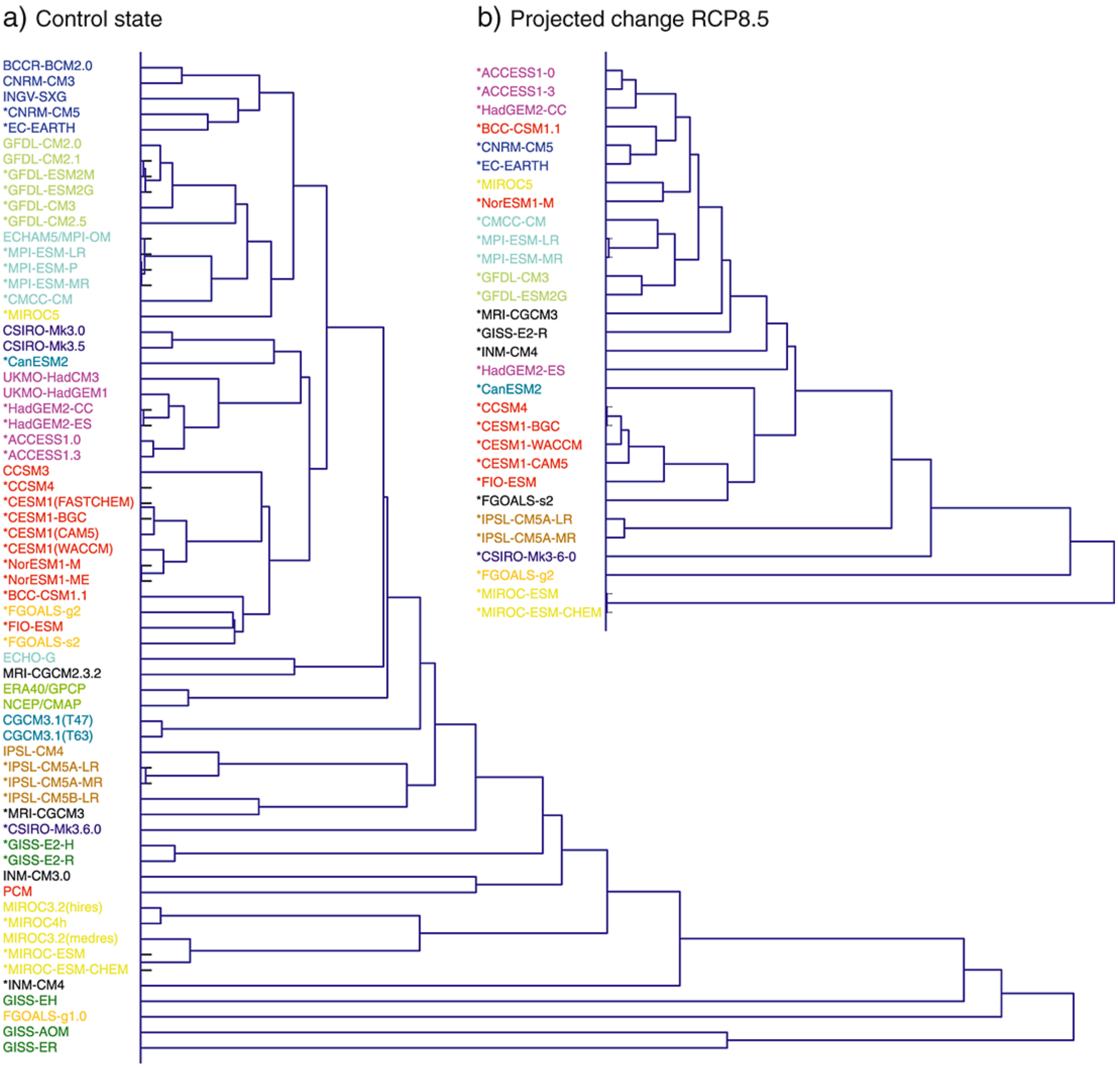
The article of Pennell and Reichler 2010 is one of the first published attempts to identify model dependence. They calculate the effective number of independent GCMs based on correlation of model errors (i.e. climate model minus observation). In their analysis of the CMIP3 multi-model ensemble, the effective amount of models reduces from 24 total to only 9. This is in agreement with the study of Knutti, Furrer et al. 2010. They argue, that the bias of GCMs does not converge towards 0 as fast as it should if the simulations were independent. As a result, uncertainty estimates in current literature seem to be too small.

The work of Masson and Knutti 2011 showed this kind of climate dependency by analysing the temperature and precipitation outputs for the historical period with a hierarchical cluster analysis. They conclude that models developed at the same institution show the most striking similarities. Also strong dependencies can be found between models that use the same atmosphere model or different versions thereof. The follow-up study of Knutti, Masson and Gettelman 2013 extends the analysis to finding strong similarities also for the future climate change projection (Figure 3.3). They categorise the ensemble into clusters of similar behaviour.

One of the first attempts to actually quantify the model dependence in order to accordingly weight individual models while quantifying climate uncertainty, has been developed by Bishop and Abramowitz 2012. They seek a linear combination of climate models which minimises the mean squared error to the observed climate. This weighting scheme accounts for both model performance and model independence. In a follow-up study, Abramowitz and Bishop 2015 use the same method to quantify uncertainty in the CMIP5 multi-model ensemble. Their method shows a decrease of uncertainty of the projected future. At a first glance this might seem counter-intuitive, as dependent models contain less information and therefore from a statistical point of view the true variance should be rather underestimated. However, their weighting scheme is based on minimising the MSE with respect to the observed climate. This forcing of the linear combination of the models to the observed climate reduces the variability.

The recent study of Zubler et al. 2015 pursue a similar path as Masson and Knutti 2011. First, they cluster GCMs in the historical period to identify similarities among models. Then they simply down-weight GCMs by the inverse number of GCMs in the same cluster and calculate the (weighted) quantiles as a measure of uncertainty. Similar ideas already exist. The work of Evans, Ji, Abramowitz et al. 2013 stresses that the true information content of an ensemble is smaller due to inter-dependencies and so when selecting a smaller subset of the ensemble (for e.g. impact studies), this true information content can be preserved when accounting for independence. They measure dependency based on the method by Bishop and Abramowitz 2012 mentioned above, namely based on the covariances of the model errors in the historical period. Basically the same line of argumentation is followed by Mendlik and A. Gobiet 2016, who reduce the ensemble size while retaining the full characteristics of the ensemble. In contrast to Evans, Ji, Abramowitz et al. 2013, their measure of dependency is based on similarities of the projected climate change to find clusters of dependent simulations, similarly as shown by Knutti, Masson and Gettelman 2013.

Based on a similar idea to categorise dependent models based on clusters, the study of Steinschneider et al. 2015 go further and analyse the climate change signals within a



a) Control state

b) Projected change RCP8.5

Figure 3.3: Model “family trees” shown as a dendrogram. Models with obvious similarities in code or produced by the same institution are marked with the same colour. (a) Similarity based on control climate from CMIP3 and CMIP5 (marked with asterisks) plus observations (ERA40/GPCP and NCEP/CMAP). (b) Similarity based on the predicted change in temperature and precipitation fields for the end of the 21st century in the RCP8.5 scenario relative to the control. Source: Knutti, Masson and Gettelman 2013.

fully Bayesian framework. They assume constant correlation between GCMs within the same cluster which have been identified by Knutti, Masson and Gettelman 2013. As an expected result, uncertainty is broader than assuming independent GCMs, which further significantly alters the quantification of risk of subsequent climate impact studies.

Model Performance

Another challenge is whether climate models should be weighted according to their skill to represent the current climate. Ideally, a climate model should be weighted based on the ability to represent the "true" climate change, but clearly this is not possible as projections of climate change relate to a state never before observed.

In fact, studies have shown that the past performance of climate models (based on global present-day diagnostics) do not correlate with their projected climate change signal (Knutti, Furrer et al. 2010). This means that if we would weight simulations based on their agreement with observations, or select a subset thereof, then the uncertainty of the future projection will not be better constrained - the model spread remains rather similar. This also applies to newly developed climate simulations, which tend to model the observed climate better without reducing the uncertainty of projected future climate change.

One particular problem with performance weighting is that often the same observational data have been used to tune the models (for example their parametrizations). Therefore there is the risk of double-counting information, over-confidence and circular logic when using the same data sets as used for model development (Knutti, Furrer et al. 2010).

Another problem arises when defining the performance of a model. The amount of metrics is huge, so whether a model is good or bad depends on the question asked. There is no model which performs best with regard to all variables (Tebaldi and Knutti 2007).

4 Author's Contributions

Climate model outcomes are the basis for future climate change analysis. The work presented in this thesis encompasses such analyses: Starting from the processing of entire ensembles of climate models and ending with the statistical interpretation of possible climate changes, while bearing in mind the complex structure of the data set. In this chapter we introduce these concepts and integrate them into the state-of-the-art research.

4.1 R Package for Climate Data Analysis

The R package Wegener Center uncertainty explorer (wux) (Mendlik, Heinrich, A. Gobiet et al. 2016) is a toolbox which enables multi-model handling for statistical analysis of climate scenarios. It is intended to be used to interpret climate model output and provides uncertainty information for the end-user of the climate simulations. Having in mind the heterogeneous target audience, we want this tool to perform following tasks:

1. Enable easy statistical *descriptive analysis* of user-defined climate model ensembles.

2. Be *expandable* to any kind of statistical analysis (to push the development of new statistical methods for climate multi-model analysis).

3. Easily *process climate simulations* to a common data format usable for statistical analysis. This enables reproducing data for any analysis needed.

Descriptive statistics of climatic changes from ensembles (point 1) are crucial to understand the underlying data. In practice people sometimes tend to forget this important step and prefer to directly address their complex research questions without having an overview of the data beforehand. A lot of valuable information lies in this analysis. Having some ready-to-use tools already implemented in wux should encourage users to perform this sort of analysis more often.

However, such a tool should not restrict the user to a pre-defined set of standard methods, on the contrary, development of new methods for statistical inference on climate simulations should be strongly supported, as this is still ongoing research (Knutti, Furrer et al. 2010). Having set up this tool directly in R, allows to explore an extremely broad pool of ready-to-use methods, also from other disciplines using different approaches (point 2).

One of the most time consuming and frustrating tasks when analysing climate simulations can be the step of processing data (point 3). The user of this tremendously big amount of datasets will find him-/herself challenged, when trying to aggregate them to the desired format (typically some sort of data frame) or get the desired statistics of the ensemble for certain geographical regions of interest. The challenge here is definitely a technical one: Processing ensembles of data in a binary-format usually requires dedicated programming work. The upside is that the data comes in the handy Network Common Data Form (NetCDF) file format[1], where a lot of meta-information about the data is stored in its header, however, life is more complicated in practice. Quite often it happens that meta-information between individual climate simulation output files differ substantially. For this reason it quickly becomes a nuisance when treating large samples of these files in an automated way. Up to now, no such tool is available which processes user-defined climate simulations in an automated way and which allows sophisticated statistical analysis. Furthermore, it is very difficult to reproduce statistical analysis from the scientific community when either the data set from the publication is not available, or the user wishes to apply the method with his/her own climate data. Providing a software which takes this burden, allows the user to solely focus on the interpretation of the climate model output without spending too many resources on technicalities. We consider it a great strength of this package to perform this task in an automated way.

Several powerful tools already exist to process climate model outputs, such as Climate Data Operators (CDO)[2], NCO (Zender 2008), climate explorer[3] (Oldenborgh et al. 2009) and The NCAR Command Language (NCL)[4]. All of those tools are designed to perform some sort of descriptive analysis and/or process the data to a desired format, however, none of those tools combines both easy multi-model handling and flexibility in statistical analysis. For example the climate explorer allows very straight forward processing of multi-model ensembles without any programming work. The user specifies what climate models to analyse simply by clicking on their names and the desired statistics. Such web-based tools however, being simple to use, lack of flexibility for a real programming interface. In addition it is not possible to extend those tools for own climate simulations which are not implemented. Also, statistical analysis is restricted to available methods. More programming-oriented tools like CDO and NCO also provide possibilities to analyse ensembles of climate simulations. However, the user has to specify the location of the data each time when calling a function and the data have to be pre-formatted for

[1] http://www.unidata.ucar.edu/software/netcdf/

[2] CDO 2014: Climate Data Operators. Available at: https://code.zmaw.de/projects/cdo

[3] http://climexp.knmi.nl

[4] The NCAR Command Language (Version 6.2.1) [Software]. (2014). Boulder, Colorado: UCAR/NCAR/CISL/VETS. http://dx.doi.org/10.5065/D6WD3XH5

the program to understand its meaning. Changing local NetCDF files too much is a restriction to reproducible research. Even though programming is possible, we are restricted to pre-defined CDO statistics operators. The main difference of wux compared to those tools is the easy way it can read in a multitude of climate simulations and simply the fact that this tool is embedded in R, which allows to apply a very broad range of sophisticated statistical tools and is not restricted only by methods implemented in the toolbox itself.

4.2 Climate Model Selection

Studies like Whetton et al. 2012 recommend to shrink the ensemble to a set of representative simulations which capture certain characteristics of the whole sample. This subset should then be used as a consistent forcing for various impact models. A sensible selection of climate simulations as input for climate change impact studies is needed in any case, either to limit computational demand and/or to mitigate biases in the ensemble statistics. Currently, such selection is often done “by opportunity” based on the ease of access to climate simulations or by subjective criteria.

However, MMEs have several issues, as systematical biases, inter-model dependencies, imbalance and lack of experimental design, which makes an uncertainty analysis of projected climate changes difficult (see Section 3.2). When selecting a subset of climate simulations for impact studies (as in Section 2.4) one has to account for this complex data structure. This work shows one such method to select a model subset especially accounting for model inter-dependencies and the unbalanced nature of MME while conserving the spread of the full ensemble.

Several studies aim to select such a representative subset. One of the first published approaches to tackle model selection with formal criteria, stems from J. B. Smith and Hulme 1998. They propose several criteria such as vintage (considering the latest generation of climate simulations only), resolution (the higher the resolution, the better), validity (model performance in the past) and representativeness (picking simulations from the high and low end of the range of climate change signals of temperature and precipitation to obtain a representative sub-sample). This method has been adopted by the IPCC guidelines for climate scenarios IPCC-TGICA 2007. Such a selection of GCMs has been applied by e.g. Murdock and Spittlehouse 2011 focusing on the region of British Columbia by analysing the models based on the spread of change in temperature and precipitation. A discussion on sub-selecting climate simulations for hydrology studies has been published by Salathe, Mote and Wiley 2007. They propose to sample driving climate simulations by considering the projected model spread for hydrology relevant parameters (temperature and precipitation change) to find groups of similar simulations and to select representative climate models. A generalisation to a multivariate setup has recently been presented by Cannon 2015. His proposed method maximises model diversity by selecting the most extreme simulations. All those studies have a non-probability sampling scheme in common: Instead of assigning probabilities to the simulations and sampling them randomly, the selection is based on qualitative characteristics which are relevant for the researcher (Mays and Pope 1995). The aim is to maximise diversity of these characteristics.

The good practice guide on assessing multi model climate projections of Knutti, Abramowitz et al. 2010 gives some more recent recommendations for model selection, also addressing the issue of model dependence. Knutti, Abramowitz et al. 2010 argue that agreement between models may arise due to the fact that models use similar simplifications and may feature similar errors. This means that models do not represent independent information and should be down-weighted in order to avoid biases in the statistical analysis of the ensemble, which are induced by double-counting similar models (Pirtle, Meyer and Hamilton 2010). Model selection can be regarded as a binary 0-1 weighting which should address these issues. Several impact studies address this problem of double-counting (e.g. Finger et al. 2012). Evans, Ji, C. Lee et al. 2013 presents a selection method taking into account model performance and independence in climate change signals. This method selects models which are most independent from the rest of the entire ensemble.

In the literature, models are often selected based only on their performance in the past, without regarding spread in the climate change signals, with the aim to use only the “best” models. However, correlation between past performance and future climate change signals are known to be very weak (Knutti, Furrer et al. 2010), which means that there is no clear indication that the best performing models in the past are most realistic with regard to the climate change signal. In addition, the ranking of models with regard to performance in the past is highly dependent on the definition of the performance measure (e.g. Jury et al. 2015), which leads to a very subjective ranking. Therefore it seems reasonable that model performance in the past should rather be used to detect and remove few severely unrealistic models which cannot be trusted in their future projections for some clearly argued reasons, but not to select a few “best performing” models, since there is no indication that they are more realistic in their future projections than other reasonably performing models.

This leads to a further model selection criterion, namely the conservation of statistical properties of the climate change signals - the sub-sample of the selected simulations should properly represent uncertainties. Recently, methods have been published based on this idea, partly combined with some pre-selection based on model performance (e.g. Bishop and Abramowitz 2012; McSweeney, Jones and Booth 2012).

Our method generalises the pragmatic approach of finding the model spread of climate change signals of, say, temperature against precipitation, as done in most studies (IPCC-TGICA 2007). It allows for simultaneous analysis of an arbitrary amount of meteorological parameters over several spatial regions of interest, and brings forth dominating patterns of climate change. Model similarities are detected based entirely on those patterns of projected change. This stands in contrast to most other studies, which find similarities in the 20th century historical runs (e.g. Abramowitz and Gupta 2008; Bishop and Abramowitz 2012; Pennell and Reichler 2010).

4.3 Quantification of Climate Change Uncertainty

The ultimate goal in climate research is to deliver some metric of expected climate change with some measure of expected uncertainty. The complications arise when considering the problematic design of the data structure: Inter-model dependencies, differing model qualities, systematic biases, unbalanced design and the lack of an experimental design are considered the main flaws of multi-model ensembles (see Section 3.2). There is no trivial solution for those problems. Several studies addressed certain issues when quantifying climate change and its uncertainties, however certain problems, like model inter-dependencies, began to receive attention only recently. In probabilistic interpretations of MMEs, in almost the entire climate literature climate models are treated as independent random variables (e.g. Tebaldi, R. L. Smith et al. 2005). Several studies showed that such independence assumptions do not hold (Knutti, Masson and Gettelman 2013; Masson and Knutti 2011; Mendlik and A. Gobiet 2016). Violation of the independence assumption means that the information content is overestimated leading to an overconfidence of the model variance (see Section 12.1 for a mathematical explanation).

Another problem arising from model inter-dependencies is the unbalanced nature of MMEs. Unbalanced in the sense, that some modelling groups provide more and some provide less simulations to the ensemble. Hence not only leads an invalid independence assumption to an underestimation of the scale estimate, it can also lead to a biased estimate of the mean climate change signal.

So far there are almost no studies presenting a suitable statistical model for this problem. Particularly, two methods only have been published so far: Bishop and Abramowitz 2012 with a follow up study by Abramowitz and Bishop 2015 present a method with explicitly computed model weights which account for historical model similarity. Those weights are based on the MSE to the observed climate. The same weights are then used to estimate the uncertainty of future projections. A different approach which is more similar to the proposed methodology of this study has recently been published by Steinschneider et al. 2015. They also use the model similarity clustering results from Knutti, Masson and Gettelman 2013 to define a blocking structure in their statistical model. Also entirely omitting the role of the observed climate, they derive a Bayesian model accounting for model similarities in the historical period and use again the same correlation estimate for future projections. For both changes in precipitation and changes in temperature they assume a multivariate Gaussian distribution. Another study by Zubler et al. 2015 quantifies the uncertainty using simple weighted quantiles of the model ensemble, by accounting the clustered structure of the data obtained from

Knutti, Masson and Gettelman 2013. While accounting for the unbalanced structure of the ensemble, their approach has no underlying distributional assumptions. However, they do not distinguish between the underlying sources of uncertainty.

An important point here is the interpretation of the estimated pdf/uncertainty, as there are several sources of such a model spread (see Section 3.1): internal (natural) variability, external (anthropogenic) forcings and climate model uncertainty. Simply estimating a probability density function (PDF) among all available data would mix up all the uncertainties, which have very different implications on climate change. Most studies have in common to calculate the uncertainty conditional on a particular human behaviour, thus not accounting on this source of uncertainty. The method proposed by Bishop and Abramowitz 2012 eliminates all model uncertainty by forcing the models to spread around the observed climate. Their estimated PDF can therefore be interpreted as uncertainty created by *internal variability*. As this is the only source of uncertainly they account for, their estimated spread is smaller than in other studies. The method presented in Steinschneider et al. 2015 does not intend to reduce any uncertainties and can be seen on the other side of the spectrum by purely estimating the climate model uncertainty: Their method does not account the internal natural variability in any sense. Internal variability is usually estimated by the spread of the year-to-year variations of models (Hawkins and Sutton 2009) and by estimating the distribution of initial condition ensembles (Collins, Tett and Cooper 2001).

Our work presents a novel approach to estimate such a PDF while accounting for model inter-dependencies. In contrast to the studies described above, our method explicitly accounts for internal climate model variability, by considering natural year-to-year variations as well as interpreting the outcome of initial condition ensembles, accounting for a long-term internal variability. In addition, our method yields estimates of different sources of uncertainty, by implementing a multi-level regression model, and can therefore relate to studies like Hawkins and Sutton 2009. Bearing this in mind, this method accounts for the unbalanced structure of the dataset, without the need to fill-up “missing” climate simulation outcomes as in Déqué, Somot et al. 2011 or Heinrich, A. Gobiet and Mendlik 2014. Further, by using bootstrap samples, the uncertainty of the individual estimates (i.e. uncertainty of the uncertainty) is quantified as well. Also, in contrast to almost any study estimating average fields of temperature and precipitation changes, we rigorously check for the underlying statistical assumptions. We find that in most cases the climate change signals are not normally distributed, but seem to stem rather from a skewed distribution.

Part II

Processing Climate Data - The “wux” Package

5 The Package

5.1 Package Overview

The wux package is meant to be an interfacing toolbox for scientists performing statistical analysis on climate models. Its focus is to provide a simple data frame for the user to make statistical inference on the ensemble. In particular, this package performs following actions, which are depicted in Figure 5.1 and described in Table 5.1:

Climate data processing. The function models2wux reads output of climate model simulations from Network Common Data Form (NetCDF) files, extracts subregions of interest, and writes climate change signals or time series to a data frame. Specific meta-information, like file locations, are stored in a modelinput input argument, which allows simple processing of the simulations. For any new climate simulation it is enough to specify those meta-information without having to actually program a new input routine.

Statistical analysis of climate change signals. Based on the data frame returned by models2wux, we implemented various plotting options and summarizing utilities for a descriptive analysis of the projected climate change signals (e.g. scatterplots of temperature and precipitation). In addition, reconstruction tools allow to fill up missing climate simulations by multiple imputation methods. Based on such a reconstructed data frame (here termed as rwux.df), the user can assess for variance components via the implemented analysis of variance (ANOVA) tools or perform exploratory data analysis.

I. Climate Data Processing (Chapter 6)

Function	Input	Output	Description
models2wux	NetCDF files	wux.df	Reads NetCDF climate model output, processes it, and writes the results to a data frame which is the backbone of all further wux analyses.
read.wux.table	wux.df files	wux.df	Reads data frame files produced by models2wux.

II. Statistical Analysis of Climate Change Signals (Chapter 7)

Function	Input	Output	Description
a) Descriptive analysis (Section 7.1)			
summary	wux.df/rwux.df	summary statistics	Summary statistics of the wux data frame (wux.df object).
plot	wux.df/rwux.df	figure	Scatter plot
plotAnnualCycle	wux.df/rwux.df	figure	Annual cycle plot
hist	wux.df/rwux.df	figure	Density plot
b) Reconstruction tools (Section 7.2)			
reconstruct	wux.df	rwux.df	Filling missing values of an unbalanced climate model design matrix in order to avoid biased ensemble estimates. Currently, the underlying reconstruction technique is based on an ANOVA using various methods for estimation. Returns reconstructed wux data.frame of class rwux.df.
c) Analysis of variance components (Section 7.2)			
aovWux	rwux.df	wux.aov	Extracts variance components of multiple climate model simulations using an ANOVA. Data must be balanced, so a reconstruction preprocessing is necessary.
plot	wux.aov	figure	Barchart for aovWux output.

Table 5.1: Most important functionalities of the wux package. (Source: Mendlik, Heinrich, A. Gobiet et al. 2016).

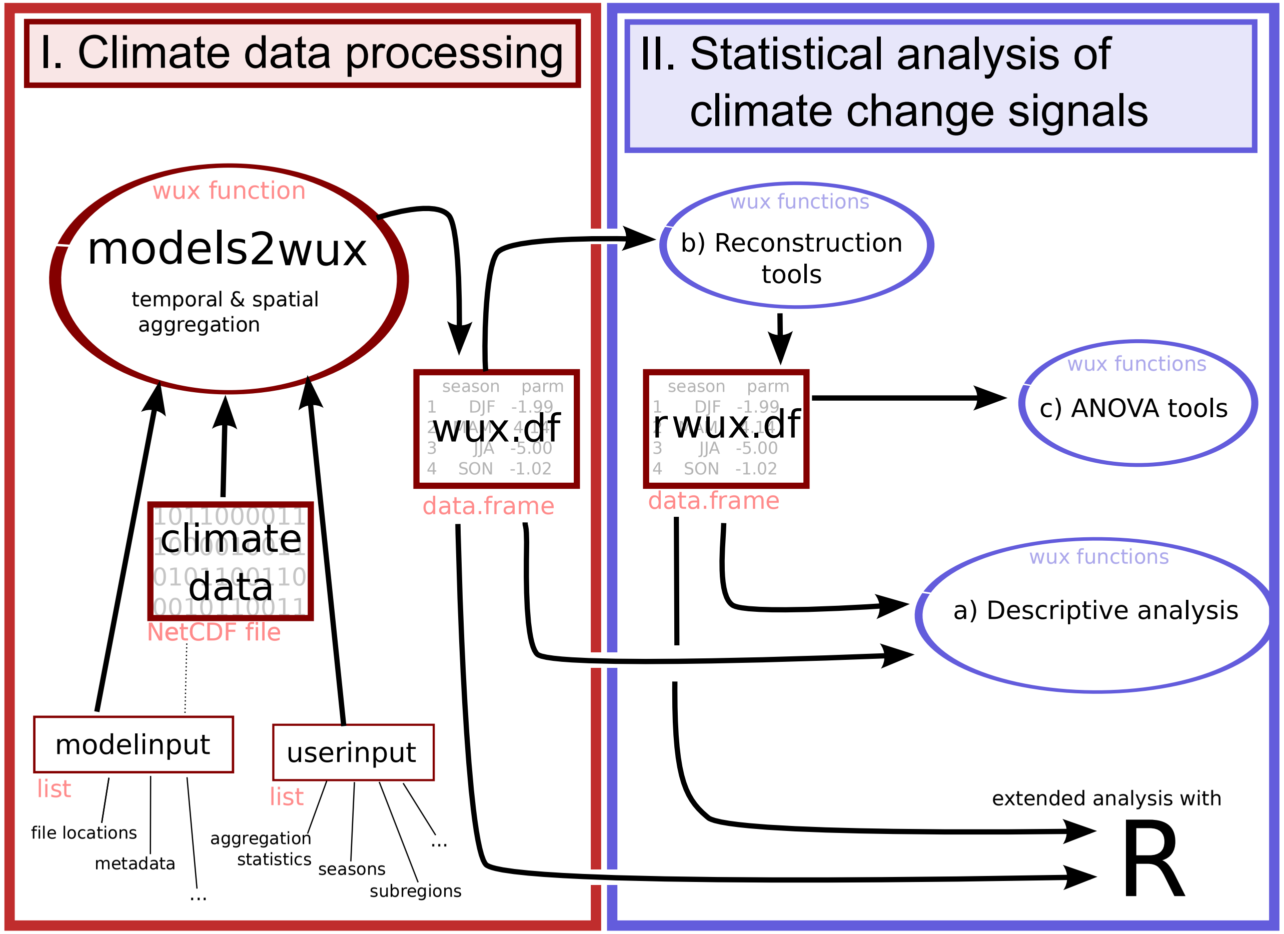


Figure 5.1: Basic functionalities of the wux package. (Source: Mendlik, Heinrich, A. Gobiet et al. 2016).

6 Climate Data Processing

The central role of the Wegener Center uncertainty explorer (wux) package is to automatically read in binary climate model output data from NetCDF files and process them to a data frame for statistical analysis. This task is performed by the function models2wux. The resulting data frame (further called wux.df, as it is technically a wux.df object) contains the climate change signals for user-specified periods, regions, seasons, and parameters for each of the climate models. One example wux.df is shown at the end of Section 6.1. Alternatively, also time series data can be obtained.

6.1 From Climate Model Output to wux data frame

This is what models2wux is doing for each specified climate model:

1. Read in a three dimensional array (longitude, latitude, time) from binary climate model output.

2. *Temporal aggregation* of the fields according to user-specified climate periods and seasons. Aggregation statistics can also be specified by the user.

3. *Spatial aggregation* (arithmetic mean) over geographical domain.

4. Computing climate change signals for specified periods.

The resulting climate change signals for each climate model are returned to a data frame.

Temporal aggregation can be performed several times serially, going from fine temporal resolution to coarser resolution, each time using another statistic for aggregation. For example, daily temperature of a climate model output could first be aggregated to monthly resolution using the mean function and as a second step the warmest month in the year can be calculated with max. This would result in a climate change signal of the warmest monthly averages. We can thus calculate a vast amount of sufficient statistics to explore the climate data. Also, the user has the possibility to retrieve the full time-series of the climate model instead of the climate change signal. This can, however, result in quite a large data frame. The lowest time resolution currently implemented for time-series data is on a monthly basis.

Being able to flexibly perform spatial aggregation over a specified domain is one of the key strengths of this program. Several ways exist for the user to identify the region of interest. For example a rectangular region defined by the longitude-latitude corners can be specified. For more flexibility, polygons can be defined using ESRI shapefiles[1] to cut out and aggregate over the desired subregion domain. The spatial aggregation is always performed using the arithmetic mean over geographical regions of any complexity. However, this process is not as trivial as it first may seem. One problem lies in the geographical projection of the climate model. Averaging over pixels of a model on a Mercator projection (angle preserving) will result in a different value than averaging over pixels in an area-preserving projection. General Circulation Models (GCMs) usually do not come on an area-preserving projection. Therefore, the pixels should be weighted by the cosine of their latitudes, otherwise areas near the poles would gain much more weight then areas near the equator. When aggregating over a certain subregion, another problem arises from the gridpoints which are associated with the subregion. Instead of either considering a gridpoint to be within a region or not (0 and 1 weight), we may want to weight all the model cells that contribute even partly to the considered subregion, i.e. seize the fraction of the cell corresponding to the area covered by the subregion.

6.2 Setting up models2wux

To process a climate multi-model ensemble of your choice, models2wux needs two input arguments userinput and modelinput, each being a named list object or a file containing a named list.

modelinput stores general information about your climate data, i.e. the locations of the NetCDF files and their filenames. It also saves certain meta-information for the specific climate simulations (e.g. a unique acronym for the simulation, the developing institution, the radiative forcing). Usually the modelinput information should be stored in a single file on your system and should be updated when new climate simulations come in. It is advisable to share this file with your colleagues if you work with the same NetCDF files on a shared IT infrastructure.

The second input argument, userinput, defines which meteorological parameters of which climate simulations defined in modelinput should be analysed. This is simply done by calling the models acronym, as all meta-information is already stored in the modelinput file. Also the geographical regions of interest and the temporal statistics are specified in this file. This file typically changes depending on the type of analysis performed.

6.3 Getting Started

We explain models2wux in more detail by considering an example of a typical workflow for climate data processing. We start with downloading a couple of GCMs from the

[1] http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf

Coupled Model Intercomparison Project Phase 5 (CMIP5) project (Taylor, Stouffer and Meehl 2012), then we specify their meta-information and the output statistics and finally we run models2wux to process the binary data to an object of class wux.df.

To obtain CMIP5 climate simulations you can get started with downloading some example NetCDF files directly from an Earth System Grid Federation (ESGF) node[2] or using the CMIP5fromESGF function from the wux package (Linux only).

R Code 6.1: wux package: Automatic data downloader for CMIP5 models.

```
> ## I) Load wux functions and example datasets...
> library("wux")

> ## II) obtain some climate simulations
> CMIP5fromESGF(save.to = "~/tmp/CMIP5/",
                models = c("NorESM1-M", "CanESM2"),
                variables = c("tas", "pr"),
                experiments= c("historical", "rcp85"))
```

Here, we download the 2 m air temperature and surface precipitation files (tas and pr) from two simulations NorESM1-M and CanESM2 for the historical period (here 1850–2005) and the future projection (2006–2100), assuming a strong change in future radiative forcing (rcp85, see Taylor, Stouffer and Meehl 2012). The data will be downloaded into a temporary directory /tmp/CMIP5/ which can take a while. You need a valid account at any ESGF node for this function to run.

In order to run models2wux, you need to specify the two input arguments explained above: A modelinput file to define which climate simulations you have on your hard-disk and a userinput file which controls models2wux itself. An example for the model specification can be obtained in the package itself:

R Code 6.2: wux package: Meta-data example for climate model CanESM2.

```
> ## III) Meta-information on downloaded data for models2wux.
> data(modelinput_test)
> str(modelinput_test)
List of 2
 $ CanESM2-r1i1p1_rcp85  :List of 11
 ..$ rcm                : chr ""
 ..$ gcm                : chr "CanESM2"
 ..$ gcm.run            : num 1
 ..$ institute          : chr "CCCma"
 ..$ emission.scenario: chr "rcp85"
 ..$ file.path.alt    :List of 2
 .. ..$ air_temperature :List of 2
 .. .. ..$ historical     : chr "~/tmp/CMIP5/CanESM2/historical"
 .. .. ..$ scenario       : chr "~/tmp/CMIP5/CanESM2/rcp85"
 .. ..$ precipitation_amount:List of 2
 .. .. ..$ historical     : chr "~/tmp/CMIP5/CanESM2/historical"
 .. .. ..$ scenario       : chr "~/tmp/CMIP5/CanESM2/rcp85"
 ..$ file.name          :List of 2
 .. ..$ air_temperature  :List of 2
```

[2] e.g. from the data node http://pcmdi9.llnl.gov

```
.. .. ..$ historical      : chr "tas_Amon_CanESM2_historical_r1i1p1_
    185001-200512.nc"
.. .. ..$ scenario        : chr "tas_Amon_CanESM2_rcp85_r1i1p1_
    200601-210012.nc"
.. ..$ precipitation_amount:List of 2
.. .. ..$ historical      : chr "pr_Amon_CanESM2_historical_r1i1p1_
    185001-200512.nc"
.. .. ..$ scenario        : chr "pr_Amon_CanESM2_rcp85_r1i1p1_200601-210012.
    nc"
..$ gridfile.path    : chr "~/tmp/CMIP5/CanESM2/historical"
..$ gridfile.filename: chr "tas_Amon_CanESM2_historical_r1i1p1_
    185001-200512.nc"
..$ resolution       : chr ""
..$ what.timesteps   : chr "monthly"
$ NorESM1-M-r1i1p1_rcp85:List of 11
...
```

This input specifies the simulations which have just been downloaded. It is a named list with the name being an unique acronym of the climate simulation. The example input here specifies two simulations, but for the sake of brevity we only display the first one, being the CanESM2-r1i1p1-rcp85 model. As this is a GCM, the rcm tag has no entry. The other tags specify the model in more detail: This simulation is run number 1 of the GCM CanESM2 and has been developed by the CCCma institution[3]. The corresponding anthropogenic forcing is rcp85. file.path.alt defines the file locations for both temperature and precipitation files as well as for historical runs and future scenario projections. In this case the historical and the future scenario runs are located in different directories, whereas both meteorological parameters are saved in the same path. file.name gives information for the corresponding file names. The files which are necessary to define the geographical longitude and latitude information are specified in gridfile.path and gridfile.filename. The data is on a monthly timescale, which is defined in what.timesteps, and the horizontal resolution is not specified here as it is optional.

It is advisable to store this list as a single file on your system. You should share this file with colleagues using the same IT infrastructure to use synergies. Such a file can also be created in an automated way using the function CMIP5toModelinput, for data obtained with CMIP5fromESGF (see the manual for more details).

Next, we want to tell models2wux to get climate change signals of both simulations we just defined above. In this example we are specifically interested in the temperature changes for the Alpine area at the end of the 21st century. Therefore we specify a user input file which contains a named list with all the necessary information:

R Code 6.3: wux package: Example config file to retrieve seasonal temperature climate change signals of two GCMs.

```
> ## IV) Input argument controlling models2wux.
> data(userinput_CMIP5_changesignal)
> str(userinput_CMIP5_changesignal)
```

[3] Canadian Centre for Climate Modelling and Analysis (www.ec.gc.ca/ccmac-cccma)

```
List of 9
 $ parameter.names     : chr "air_temperature"
 $ area.fraction       : logi TRUE
 $ reference.period    : chr "1971-2000"
 $ scenario.period     : chr "2071-2100"
 $ temporal.aggregation:List of 1
  ..$ stat.level.1:List of 3
  .. ..$ period      :List of 4
  .. .. ..$ DJF: chr [1:3] 12 1 2
  .. .. ..$ MAM: chr [1:3] 3 4 5
  .. .. ..$ JJA: chr [1:3] 6 7 8
  .. .. ..$ SON: chr [1:3] 9 10 11
  .. ..$ statistic  : chr "mean"
  .. ..$ time.series: logi FALSE
 $ subregions          :List of 1
  ..$ AL: num [1:4] 5 15 48 44
 $ plot.subregion      :List of 4
  ..$ save.subregions.plots: chr "/tmp/"
  ..$ xlim                 : num [1:2] 0 20
  ..$ ylim                 : num [1:2] 40 50
  ..$ cex                  : num 10
 $ save.as.data     : chr "/tmp/wuxexample"
 $ climate.models   : chr [1:2] "CanESM2-r1i1p1_rcp85", "NorESM1-M-r1i1p1_
     rcp85"
```

This userinput input argument tells models2wux to process air-temperature (parameter.names) for both models CanESM2-r1i1p1-rcp85 and NorESM1-M-r1i1p1-rcp85 (climate.models tag). We define our 30 years base period (tag reference.period) to be 1971–2000 and the projected future period of interest (tag scenario.period) for the climatic change to be the 30 years of 2071–2100. We want the data to be aggregated to seasons summer (June, July, August: JJA), autumn (SON), winter (DJF) and spring (MAM). For each of those seasons models2wux returns the climate change signal defined by the user by calculating scenario.period minus reference.period (for precipitation, changes are in addition calculated relative to reference.period). When setting the attribute time.series to TRUE, the output is a transient time series instead of climate change.

We want to aggregate over the spatial extend of the Alpine Region (AL), (see J. H. Christensen and O. B. Christensen 2007), which is defined in the subregions tag. Here it is a named vector of longitude and latitude coordinates and it defines a rectangular region (western, eastern, northern and southern coordinates of the corners). There are plenty of other ways to define a subregion, like reading in shapefiles. To analyze which model grid cells lie within the specified region, we can specify plot.subregion (see Figure 6.1). We usually want to aggregate all model cells which lie within the specified region, however, sometimes we would like to down-weight those cells which only partly contribute to the considered region. Setting area.fraction as TRUE weights the cells corresponding to the area covered by the subregion (Figure 6.1). Furthermore, area.fraction TRUE is necessary, if the size of the subregion is in the same order of magnitude as the grid cell. Such cases should be handled with care, since the grid point interpretation of climate models is problematic. In most cases, the analysed subregions should be much larger than the grid size of the models and the error produced by setting area.fraction to FALSE is negligible and processing gains a massive speed up. The data frame will also be saved as a comma-separated file to /tmp/wuxexample.

Finally we run models2wux with the input arguments explained above to obtain the temperature climate change signals (delta.air-temperature) for both simulations aggregated over the Alpine region and four seasons. Columns besides subreg, season and the temperature change parameter are meta-information of the climate data and derived from the modelinput input argument.

R Code 6.4: wux package: A data.frame of processed seasonal temperature climate change signals of two GCMs.

```
> ## V) Process NetCDF files
> climchange.df <- models2wux(userinput = userinput_CMIP5_changesignal,
>                             modelinput = modelinput_test)
> climchange.df
   subreg season                 acronym institute       gcm gcm.run em.scn
1      AL    DJF    CanESM2-r1i1p1_rcp85     CCCma   CanESM2       1  rcp85
2      AL    JJA    CanESM2-r1i1p1_rcp85     CCCma   CanESM2       1  rcp85
3      AL    MAM    CanESM2-r1i1p1_rcp85     CCCma   CanESM2       1  rcp85
4      AL    SON    CanESM2-r1i1p1_rcp85     CCCma   CanESM2       1  rcp85
13     AL    DJF  NorESM1-M-r1i1p1_rcp85       NCC NorESM1-M       1  rcp85
14     AL    JJA  NorESM1-M-r1i1p1_rcp85       NCC NorESM1-M       1  rcp85
15     AL    MAM  NorESM1-M-r1i1p1_rcp85       NCC NorESM1-M       1  rcp85
16     AL    SON  NorESM1-M-r1i1p1_rcp85       NCC NorESM1-M       1  rcp85
      period ref.per resolution corrected delta.air_temperature
1  2071-2100      no         NA        no              4.066630
2  2071-2100      no         NA        no              8.041165
3  2071-2100      no         NA        no              4.261498
4  2071-2100      no         NA        no              5.686222
13 2071-2100      no         NA        no              3.336806
14 2071-2100      no         NA        no              5.378479
15 2071-2100      no         NA        no              3.922325
16 2071-2100      no         NA        no              3.787082
```

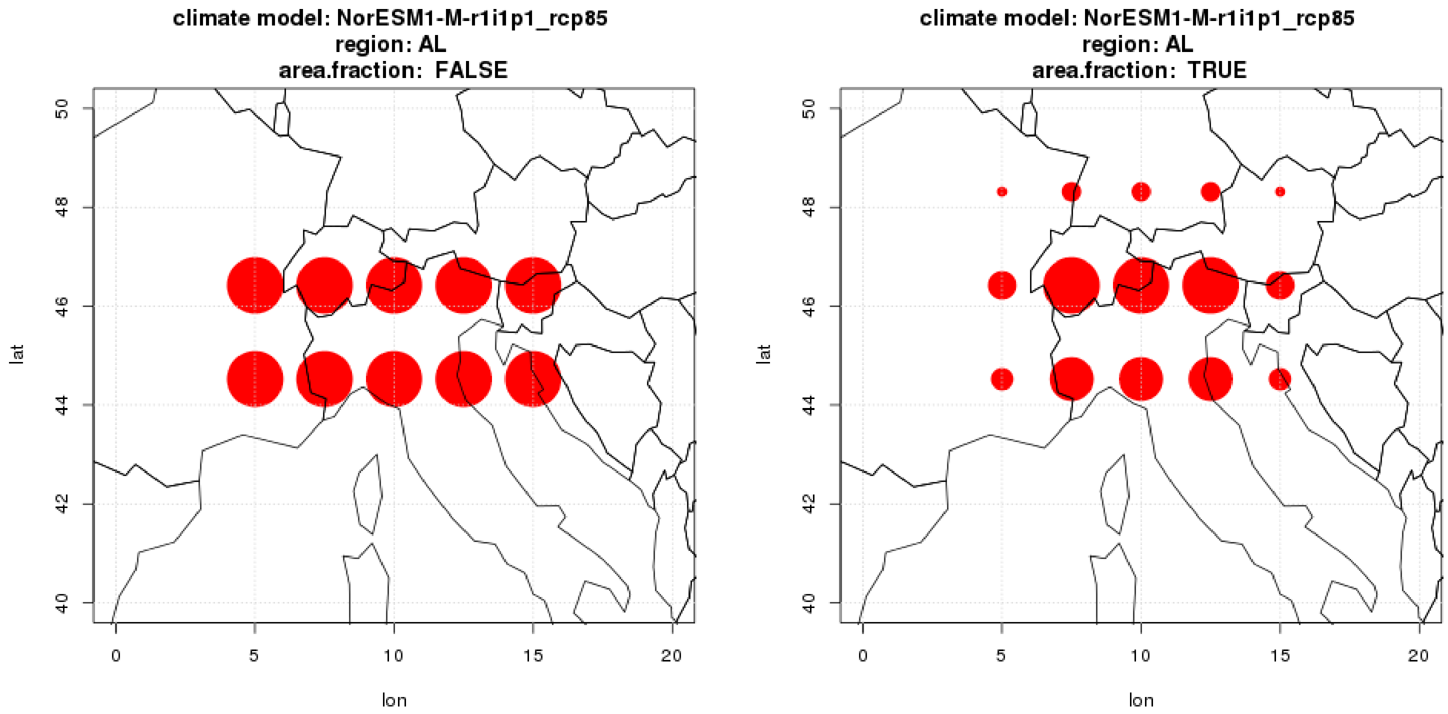


Figure 6.1: Grid cells of the NorESM1-M climate model being aggregated. On the left figure area.fraction is switched off, taking all cells with their centroids lying within the Alpine Region (AL) and weight them equally. The right figure has area.fraction on: The smaller the circles, the smaller the coverage of the model cells and the smaller their weight. (Source: Mendlik, Heinrich, A. Gobiet et al. 2016).

7 Statistical Analysis of Climate Change Signals

Several functions are available to analyze the processed climate change signals created by models2wux.

7.1 Descriptive Analysis

The summary function gives a descriptive overview of the climate model ensemble which has been processed. On the one hand it calculates categorical statistics (counting climate models, emission scenarios, RCM-GCM cross-tables, . . .) and on the other hand it returns statistics of continuous climate change signals (mean, standard deviation, coefficient of variation and quantiles) split by season, emission scenario, meteorological parameters and subregions. Let us consider the climate change signals from 1961–1990 until 2021–2050 in the Greater Alpine Region (GAR) of a multi-model ensemble consisting of 22 Regional Climate Models (RCMs) from the ENSEMBLESmo project (Linden and Mitchell 2009).

R Code 7.1: wux package: Summary of a wux.df object.

```
> ## VI b) Analyze climate change data - summary statistics
> data(ensembles)
> # consider Greater Alpine Region (GAR) only
> wuxtest.df <- droplevels(subset(ensembles, subreg == "GAR"))
> ## summary statistics
> summary(wuxtest.df)
    ----------------------------------------------------------------------
    ----------------------- FREQUENCIES BY SCENARIO ----------------------
    ----------------------------------------------------------------------
A1B:
  8 GCMs (disregarding runs)
  22 models total
  Number of GCMs used:
       ARPEGE   BCCR-BCM2.0         CGCM3 ECHAM5/MPI-OM      HadCM3Q0
            3             3             1             5             5
    HadCM3Q16     HadCxM3Q3      IPSL-CM4
            2             2             1
  Number of RCM runs:
   CLM   CRCM HIRHAM HadRM3 PROMES  RACMO    RCA   RCA3   REMO  RM4.5  RM5.1
     2      1      5      3      1      1      3      1      1      1      1
  RRCM  RegCM
     1      1
  Number of RCMs: 13
```

```
    ----------------------------------------------------------------------
    ----------------- CLIMATE MODEL STATISTICS BY SUBREGION ----------------
    ----------------------------------------------------------------------

------------ GAR ------------
perc.delta.precipitation_amount:
 [A1B]
       n       mean    sd      coefvar min     max     med     q25     q75
  DJF: 22      2.88    5.09    1.77    -8.96   10.25   3.81    1.54    5.8
  JJA: 22      -2.82   6.87    2.44    -12.42  10.71   -3.7    -7.19   1.61
  MAM: 22      -0.64   4.99    7.83    -9.41   6.61    0.7     -5.52   2.87
  SON: 22      0.76    5.7     7.51    -12.16  12.46   0.77    -2.09   3.65

delta.air_temperature:
 [A1B]
       n       mean    sd      coefvar min     max     med     q25     q75
  DJF: 22      1.66    0.51    0.31    0.92    2.41    1.56    1.19    2.13
  JJA: 22      1.7     0.65    0.38    0.47    2.79    1.88    1.31    2.18
  MAM: 22      1.25    0.53    0.43    -0.02   2.26    1.21    0.91    1.55
  SON: 22      1.57    0.55    0.35    0.61    2.88    1.64    1.27    1.8
```

For the sake of brevity, we do not show all parts of the output. The FREQUENCIES output shows that $n = 22$ climate simulations driven by 8 GCMs forced with one emission scenario (A1B) have been processed and shows the count of the specific RCMs and GCMs used in the analysis. The CLIMATE MODEL STATISTICS output shows a descriptive analysis of the continuous variables in the data set based on all $n = 22$ climate simulations available. In this case the continuous variables are the relative change of precipitation (perc.delta.precipitation-amount) in percent and the absolute change of temperature (delta.air-temperature) in °C. The precipitation change in the GAR is not significant for either season, but there is a tendency in DJF for a slight increase of total precipitation. In contrast to that, the change signal for temperature is significant for all seasons showing quite an uniform warming, where MAM seems to have the smallest trend.

Also, functions for a graphical overview of the climate model ensemble are available in wux. The method plot for a wux.df object draws one or more scatterplots containing climate change signals of selected meteorological parameters.

R Code 7.2: wux package: Calling a scatterplot of temperature and precipitaion climate change signals for an RCM ensemble.

```
> ## VI b) Analyse climate change data - scatterplots
> plot(ensembles, "perc.delta.precipitation-amount",
>      "delta.air_temperature", boxplots = TRUE,
>      xlim = c(-40,40), ylim = c(0, 4),
>      xlab = "Precipitation Amount [%]", ylab = "2-m Air Temperature [K]",
>      main = "Scatterplot", subreg.subset = c("GAR"))
```

This draws a simple scatterplot which accounts for certain meta-information of the climate change data frame and allows to highlight certain models. One of the scatterplots produced by this call is shown on the left hand side of Figure 10.2. This is a very useful plot as it gives a good overview on the model behaviour and the climate change uncertainty. In our example, some models project an increase in precipitation change, whereas some project a decline. No correlation between temperature and precipitation change is visible on this small spatial scale.

7.2 Data Reconstruction Methods

Due to limited computational capacities, even in large-scale climate modelling projects such as CMIP5 or EURO-CORDEX (Jacob et al. 2013) only a limited number of climate simulations can be realised and it is a question of the experimental design which uncertainty components are primarily tackled within the ensemble. Therefore, missing realisations within climate projection ensembles are a common problem and even simple ensemble estimates such as mean and variability for e.g. temperature changes are potentially biased due to unequal sampling of the uncertainty components. In order to avoid such biases, Déqué, Rowell et al. 2007 introduced an iterative data reconstruction method which assumes additivity between uncertainty components in order to estimate the missing climate change signals. This reconstruction method was further applied in several studies in order to obtain a balanced design for the analysis of variance components (Déqué, Rowell et al. 2007; Déqué, Somot et al. 2011; Heinrich, A. Gobiet and Mendlik 2014; Mendlik and A. Gobiet 2016; Prein, A. Gobiet and Truhetz 2011). In wux, we implemented the method of Déqué, Rowell et al. 2007 for a two-factorial design (reconstruct) such as realised in the ENSEMBLES project (Linden and Mitchell 2009). In ENSEMBLES, a set of 21 high resolution RCM simulations with a horizontal grid spacing of about 25 km was produced. The ensemble consists of 8 GCMs and 16 RCMs only forced by the A1B emission scenario, but due to limited computational resources, only a small fraction (16.4 % of the possible GCM-RCM combinations) could be realised. The result of such a reconstruction is shown in Figure 10.2. In that case, filling up the missing GCM-RCM combinations does not alter the distribution of temperature and precipitation change. However, as the method relies on an implicit formulation of the uncertainty components, it cannot be used to extend the ensemble to GCMs that have not been used as driver for any RCM in the ensemble. Further reconstruction methods which are able to extend the ensemble to GCMs outside of the original design are investigated in Heinrich, A. Gobiet and Mendlik 2014.

7.3 Example: Further Statistical Analysis

It is one of the key strengths of this package to be directly implemented in R and for that reason to have direct access to a huge magnitude of statistical methods to analyse climate data. We provide an example application in this chapter to show possible extensions based fully on the wux.df. We use a linear mixed effects model from the lme4 package

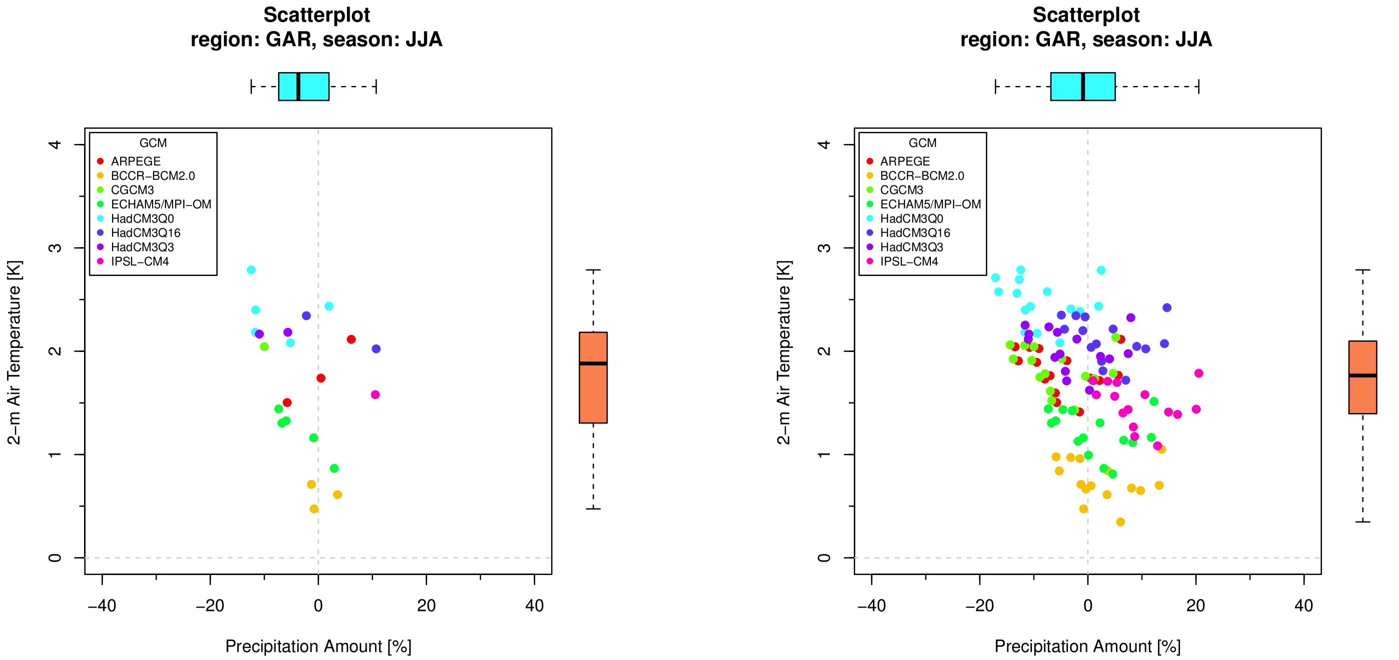


Figure 7.1: Projected changes of summer precipitation and temperature of the ENSEMBLES models from 1961–1990 to 2021–2050 in the Greater Alpine Region. The left plot shows the originally available 22 RCMs, whereas the right plot depicts a reconstructed dataset filled up with the function reconstruct. (Source: Mendlik, Heinrich, A. Gobiet et al. 2016).

(Bates et al. 2014) to estimate the average summer temperature trend over the Greater Alpine Region based on individual time-series of 16 GCMs from the CMIP5 ensemble under a moderate stabilisation scenario (Representative Concentration Pathway (RCP) 4.5).

To generate the appropriate wux.df, the timeseries tag in the userinput file was set TRUE (see Chapter 6). The aim here is to get an average linear trend while accounting for the unbalanced model design. Several of the GCMs were run a couple of times (up to 10 times) with different initial conditions, which induces a dependency structure in the data set. We assess for this dependency by putting random effects in the linear model:

$$Y_{ijk} = \beta_0 + \beta_1 \text{year}_{jk} + b_{0i} + b_{1i}\text{year}_{jk} + \epsilon_{ijk}$$

where Y_{ijk} is the average summer temperature projected by $i = 1, \ldots, 16$ GCMs with $j = 1, \ldots, n_i$ runs per GCM and $k = 1, \ldots, 130$ yearly time steps. The random effects are defined as

$$\begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{gcm} & 0 \\ 0 & \sigma^2_{gcm.t} \end{pmatrix} \right) \quad \text{and} \quad \epsilon_{ijk} \overset{iid}{\sim} N\left(0, \sigma^2_y\right).$$

We use the lmer function from the lme4 package for our analysis to estimate the fixed effects $\hat{\beta}_0, \hat{\beta}_1$ and to predict the individual random effects $\hat{\boldsymbol{b}}_0 = (\hat{b}_{0,1}, \ldots, \hat{b}_{0,16})', \hat{\boldsymbol{b}}_1 = (\hat{b}_{1,1}, \ldots, \hat{b}_{1,16})'$. The time-series data and the trends are shown in Figure 7.2 plotted with the lattice package (Sarkar 2008).

R Code 7.3: wux package: Example for an uncertainty analysis of CMIP5 multi-model ensemble.

```
> data(alpinesummer)
> ## pick just a few GCMs for this example - for a more compact display
> gcms.sub <- c("ACCESS1-3", "BCC-CSM1-1", "CESM1-CAM5", "CMCC-CM",
>               "CNRM-CM5", "CSIRO-Mk3-6-0", "EC-EARTH", "FGOALS-g2",
>               "GFDL-CM3", "HadGEM2-ES", "INM-CM4", "IPSL-CM5A-LR",
>               "MIROC5", "MPI-ESM-LR", "MRI-CGCM3", "NorESM1-M")
> alpinesummer.sub <- droplevels(subset(alpinesummer, gcm %in% gcms.sub))
> ## transform for better convergence
> alpinesummer.sub$time <- alpinesummer.sub$year - 1971
> lmm.fit <- lmer(air_temperature ~ 1 + time  + (1 |gcm) + (0 + time|gcm),
+                   data = alpinesummer.sub)
> summary(lmm.fit)
Linear mixed model fit by REML ['lmerMod']
Formula: air_temperature ~ 1 + time + (1 | gcm) + (0 + time | gcm)
   Data: alpinesummer.sub

REML criterion at convergence: 16472.2

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.0410 -0.6150 -0.0321  0.5766  4.5612

Random effects:
 Groups   Name        Variance  Std.Dev.
 gcm      (Intercept) 2.5671124 1.60222
 gcm.1    time        0.0001318 0.01148
 Residual             1.2482244 1.11724
Number of obs: 5330, groups:  gcm, 16

Fixed effects:
            Estimate Std. Error t value
(Intercept) 16.49168    0.40257   40.97
time         0.03443    0.00292   11.79

Correlation of Fixed Effects:
     (Intr)
time -0.016
> ## prints the first random effects
> head(coef(lmm.fit)$gcm)
              (Intercept)       year
ACCESS1-3         18.53855 0.03755274
BCC-CSM1-1        17.26063 0.02928145
CESM1-CAM5        16.11973 0.03574971
CMCC-CM           13.62953 0.03733811
CNRM-CM5          16.25376 0.03042184
CSIRO-Mk3-6-0     16.55908 0.04872848
```

The average slope $\hat{\beta}_1 = 0.34\,°\mathrm{C/decade}$ (0.034 °C/y) is highly significant and the individual slopes of the GCMs reach from slowly warming simulations $\hat{b}_{1,1} = 0.16\,°\mathrm{C/decade}$ to very sensitive simulations $\hat{b}_{1,16} = 0.56\,°\mathrm{C/decade}$ (not visible in this output) assuming linear temperature evolution over 130 years from 1971–2100. The residual standard deviation is $\hat{\sigma}_y = 1.12\,°\mathrm{C}$, which in this case can be interpreted as the average year-to-year natural variability.

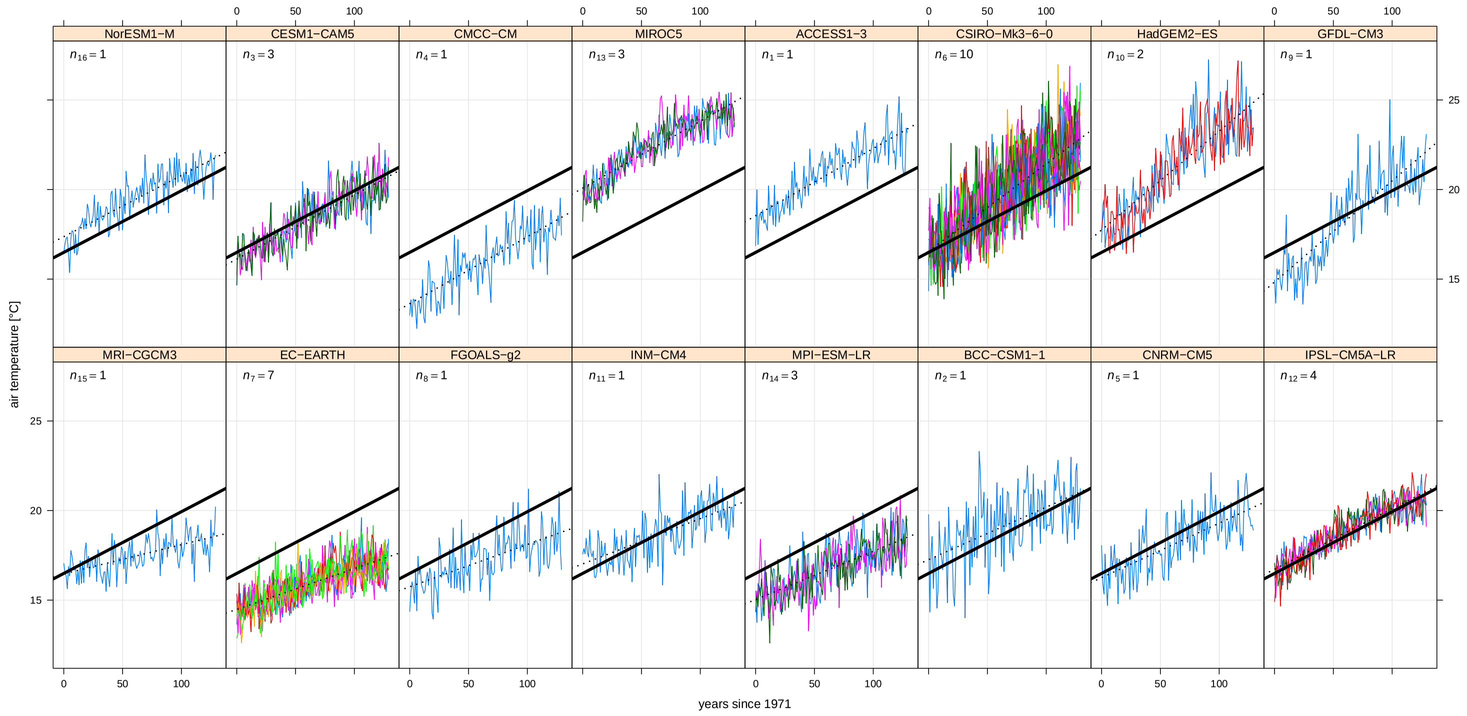


Figure 7.2: Time-series of GCMs from the CMIP5 ensemble for summer temperature in the Alpine region. The estimated average trend $\hat{\beta}_1$ is shown as a bold line, the predicted random effects trends are shown as a dashed line. The simulations are ordered from low trend (lower left panel) to high trend (upper right panel). (Source: Mendlik, Heinrich, A. Gobiet et al. 2016).

8 Conclusion

It is crucial in climate research not only to analyse outcomes of single climate models, but to consider entire multi-model ensembles, as it is virtually demanded in every climate impact related study to assess the associated uncertainties of the projected changes. There is, however, definitely a technical challenge to process large amounts of climate simulations at once, and not many tools exist to assess this problem. Another more general problem arises measuring the uncertainty in multi-model ensembles. It is somewhat uncomfortable to make statistical inference on multi-model ensembles, as they do not stem from a designed experiment (Knutti, Furrer et al. 2010), are utterly unbalanced (Déqué, Rowell et al. 2007), and are known to be biased (Maraun et al. 2010; Themeßl, A. Gobiet and Leuprecht 2011).

The focus here is not to show solutions for sophisticated statistical analyses of climate datasets, but merely to present a flexible and easy-to-use tool which is able to preprocess the datasets for further statistical analysis. This way, the user can focus on solving the grand challenges of statistical inference of multi-model datasets and does not need to spend valuable resources on technical data issues. The function models2wux fulfills exactly this task by processing magnitudes of binary climate model data to a R data frame of climate change signals. Subsequently, the user can take advantage of the vast amount of methods available in R, to analyse this data set.

However, this package also provides some functions for a first exploratory data analysis, as e.g. a summary function and some plotting routines. Such simple analysis provide very valuable information on the multi-model ensemble. In addition, we also provide a couple of methods to address the issue of unbalanced experimental designs. Several methods from literature are implemented to fill up the incomplete data matrix (Déqué, Rowell et al. 2007; Heinrich, A. Gobiet and Mendlik 2014).

It should be kept in mind, that also other software packages exist which partly fulfill similar tasks (e.g. climate explorer, Climate Data Operators (CDO), The NCAR Command Language (NCL)). The climate explorer can be a very convenient way to have a quick descriptive analysis of a multi-model ensemble. It is easy to use, but it is also restricted to a non-programming environment. Also, one can analyse only models which are implemented in the system, and the statistical methods are restricted as well. It should be noted, that no spatial analysis is currently possible within wux, as the emphasize lies on averaged domains. For spatial maps, tools as CDO or NCL are far better suited. Another limitation can be the hardware needed to process large datasets. R is not the most memory-efficient environment and one can run into trouble when reading climate simulations with a very high spatial resolution.

To sum it up, wux is a very flexible tool dealing with different aspects of climate model uncertainty in climate change impact investigations and enables a quick analysis of climate scenario uncertainty, which typically demands a considerable technical effort as well as fundamental knowledge about climate modelling. It can be used to achieve a quick overview on the involved uncertainties to identify the most important sources of uncertainty or to select representative sub-ensembles to be used as input for impact studies. wux is fully flexible regarding the meteorological parameter and region under consideration and is able to assess uncertainties based on multiple user-defined parameters.

Part III

Climate Model Selection

9 Statistical Methods

The following model selection method samples one representative climate simulation out of groups of models with similar characteristics, to obtain a sub-set of independent simulations which cover the multi-model ensemble (MME) spread. Those similar groups are found using clustering techniques. The model spread and the similarity measure can be defined upon an arbitrary amount of climate parameters and indicators. In addition, several spatial regions and several seasons of interest can be freely defined. Therefore this method is not limited to the commonly used temperature and precipitation changes of a single region, it is rather a multivariate extention.

Having such a complex set-up, it is necessary to decrease the dimensionality of the climate parameters to eliminate collinearities and to reduce random noise. This is done by using a principle component analysis (PCA) to identify *patterns of climate change* as step (1) (Jolliffe 2002). Step (2) finds model similarities with a hierarchical clustering algorithm (Huth et al. 2008) and finally, step (3) involves sampling of the simulations out of each cluster detected. We assume that unrealistic simulations have been sorted out in advance of the study. The next sections explain those steps in more detail.

9.1 Common patterns of climate change: PCA

With a PCA for each simulation, we transform the climate change signals of the meteorological parameters (like temperature and precipitation) to a linear combination of those variables. Those transformed meteorological variables are formally called principle components (PCs) and they form the *common patterns of climate change* (see Fig. 9.1). The transformations, which are the coefficients of the linear combinations, are called *loadings* and they describe which meteorological variables are combined to a particular pattern of climate change. The PCs of each simulation (i.e. patterns of climate change) are treated in the same way as meteorological variables, but they differ in being stochastically independent of each other, which is necessary for the subsequent cluster analysis.

In the next step the most dominant patterns of climate change signals have to be detected in order to reduce noise and make the subsequent cluster analysis more robust. We use the broken-stick method given in Jolliffe 2002, which compares the variances of individual PCs of the used dataset with a randomly generated dataset. If those random variances are equal or larger than the observed ones, the corresponding PC can be regarded as noise and should be excluded (Fig. 9.2).

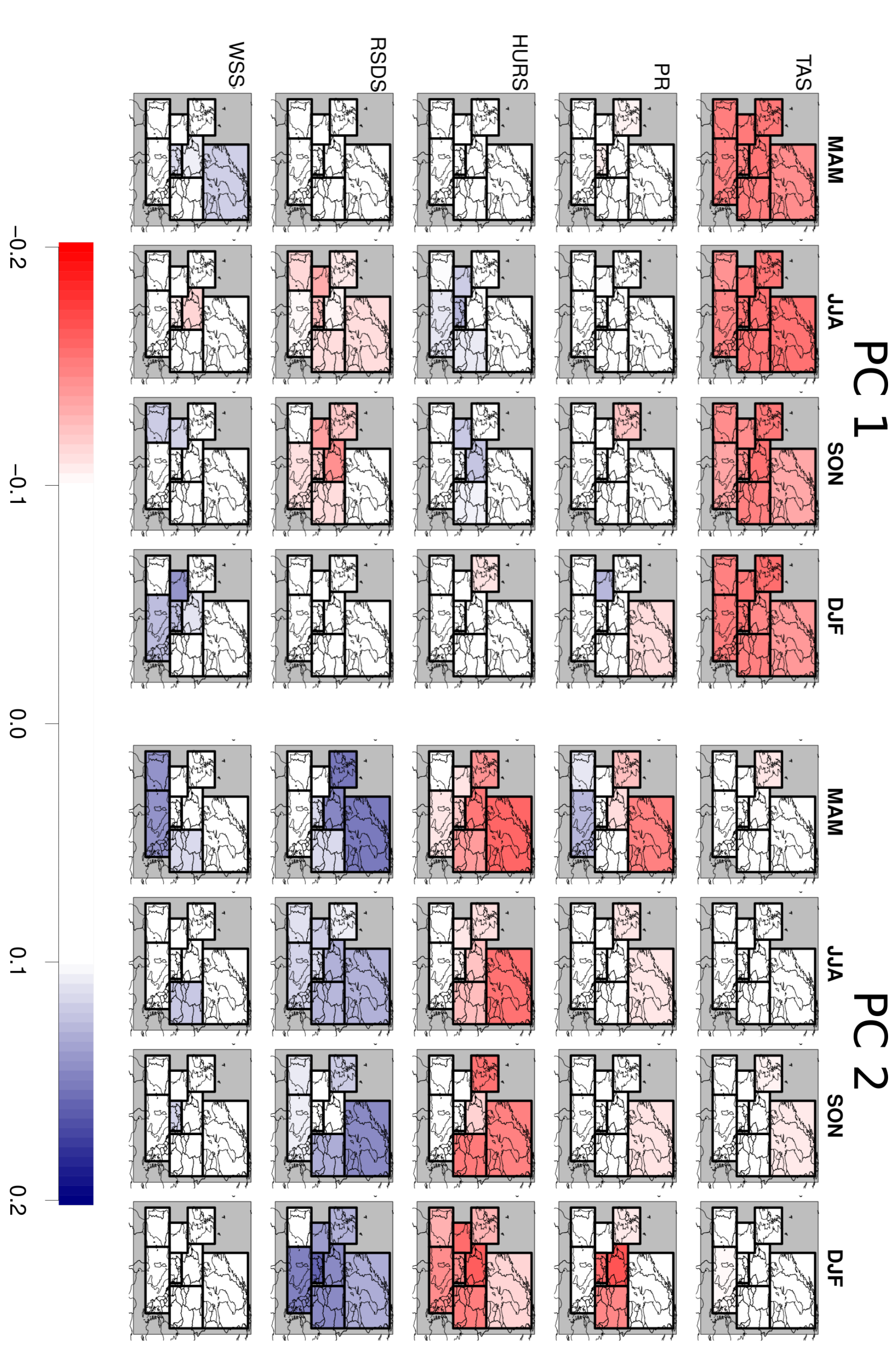


Figure 9.1: The coefficients of the linear combinations (loadings) yielding the first 2 dominant patterns of climate change (PCs) of the ENSEMBLES RCMs across Europe. Blue boxes indicate increase and red boxes decrease of the corresponding parameter. (Source: Mendlik and A. Gobiet 2016).

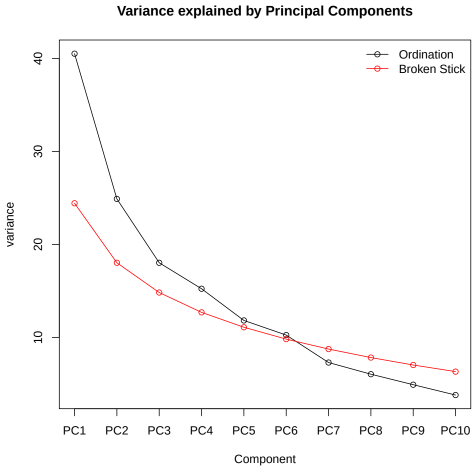


Figure 9.2: Scree plot depicting the variances explained by the individual PCs (patterns of climate change, black line). The red line depicts the variance of a randomly generated dataset. PCs with variances close to this red line can be regarded as noise and are excluded from the analysis. (Source: Mendlik and A. Gobiet 2016).

Population PC

Let $\boldsymbol{x} = (x_1, \ldots, x_p)' \in \mathbb{R}^p$ be the vector of variables. We are looking for a linear combination of $\boldsymbol{x}$ which maximises variance. The first *principle component* (PC1) is then

$$z_1 = \boldsymbol{\alpha}_1'\boldsymbol{x} = \alpha_{11}x_1 + \cdots + \alpha_{1p}x_p \in \mathbb{R} \tag{9.1}$$

with unknown coefficients (*loading vector*) $\boldsymbol{\alpha}_1 \in \mathbb{R}^p$. We seek for

$$\operatorname{Var}(z_1) = \operatorname{Var}(\boldsymbol{\alpha}_1'\boldsymbol{x}) = \boldsymbol{\alpha}_1' \operatorname{Var}(\boldsymbol{x})\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_1'\Sigma\boldsymbol{\alpha}_1 \to \max,$$

with $\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1 = 1$. The maximisation problem can be expressed as Lagrangian expression

$$\phi_1 = \boldsymbol{\alpha}_1'\Sigma\boldsymbol{\alpha}_1 - \lambda(\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1 - 1).$$

Setting $\partial\phi_1/\partial\boldsymbol{\alpha}_1 = 0$ we obtain

$$\begin{aligned} \Sigma\boldsymbol{\alpha}_1 - \lambda\boldsymbol{\alpha}_1 &= 0 \\ (\Sigma - \lambda I_p)\boldsymbol{\alpha}_1 &= 0. \end{aligned}$$

The loading vector $\boldsymbol{\alpha}_1$ is the eigenvector of Σ and λ is the corresponding eigenvalue. We then get

$$\mathrm{Var}(z_1) = \boldsymbol{\alpha}_1' \Sigma \boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_1' \lambda_1 \boldsymbol{\alpha}_1 = \lambda_1 \boldsymbol{\alpha}_1' \boldsymbol{\alpha}_1 = \lambda_1.$$

For PC2 we maximise $\mathrm{Var}(z_2) = \mathrm{Var}(\boldsymbol{\alpha}_2' \boldsymbol{x})$ with the constraints $\boldsymbol{\alpha}_2' \boldsymbol{\alpha}_2 = 1$ and orthogonality to the previous PC $\boldsymbol{\alpha}_1' \boldsymbol{\alpha}_2 = 0$ and get analogous coefficients.

Since eigenvetors and their corresponding eigenvalues are arranged in decreasing order, the variance of the PCs decrease with higher order.

In matrix notation we can write the PCA as a base transformation

$$\begin{aligned}
\begin{pmatrix} z_1 \\ \vdots \\ z_p \end{pmatrix} &= \begin{pmatrix} \alpha_{11} \\ \vdots \\ \alpha_{p1} \end{pmatrix} x_1 + \dots + \begin{pmatrix} \alpha_{1p} \\ \vdots \\ \alpha_{pp} \end{pmatrix} x_p \\
\begin{pmatrix} z_1 \\ \vdots \\ z_p \end{pmatrix} &= \begin{pmatrix} \alpha_{11} & \dots & \alpha_{1p} \\ \vdots & & \vdots \\ \alpha_{p1} & \dots & \alpha_{pp} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix} \\
\boldsymbol{z} &= A\boldsymbol{x} \in \mathbb{R}^p,
\end{aligned}$$

with the vector of principle components $\boldsymbol{z} = (z_1, \dots, z_p)' \in \mathbb{R}^p$, the loading matrix $A = (\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_p)' \in \mathbb{R}^{p \times p}$ containing rows of eigenvectors and the variables $\boldsymbol{x} = (x_1, \dots, x_p)' \in \mathbb{R}^p$

Sample PC

In reality the population covariance matrix $\mathrm{Var}\left((x_1, \dots, x_p)'\right) = \Sigma$ is unknown and has to be estimated using a sample for each variable x_i. Denote $\boldsymbol{x_i} = (x_{i1}, \dots, x_{in})'$ the sample of variable x_i with n observations. We write $X = (\boldsymbol{x_1}, \dots, \boldsymbol{x_p})' \in \mathbb{R}^{p \times n}$ as the sample variable matrix. The first *sample PC* is

$$\boldsymbol{z}_1 = \boldsymbol{\alpha}_1' X = \alpha_{11} \boldsymbol{x}_1 + \dots + \alpha_{1p} \boldsymbol{x}_p \in \mathbb{R}^n. \tag{9.2}$$

We define the *score* of observation i in the first PC as

$$z_{1i} = \boldsymbol{\alpha}_1' \boldsymbol{x_i} = \alpha_{11} x_{1i} + \dots + \alpha_{1p} x_{pi} \in \mathbb{R},$$

or in matrix notation, for $r = \mathrm{rank}(\mathrm{Cov(X)})$ we can write $Z = AX$, with $Z = (\boldsymbol{z}_1, \dots, \boldsymbol{z}_r) \in \mathbb{R}^{n \times r}$ and $A = (\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_r)'$, or in more detail:

$$\begin{pmatrix} z_{11} & \dots & z_{1i} & \dots & z_{1n} \\ \vdots & & \vdots & & \vdots \\ z_{r1} & \dots & z_{ri} & \dots & z_{rn} \end{pmatrix} = \begin{pmatrix} \alpha_{11} & \dots & \alpha_{1p} \\ \vdots & & \vdots \\ \alpha_{r1} & \dots & \alpha_{rp} \end{pmatrix} \cdot \begin{pmatrix} x_{11} & \dots & x_{1i} & \dots & x_{1n} \\ \vdots & & \vdots & & \vdots \\ x_{p1} & \dots & x_{pi} & \dots & x_{pn} \end{pmatrix}$$

with the rows of Z being the principle components and its columns the scores.

The X matrix can be reconstructed from the PCA scores, Z. Due to the orthogonality of A (i.e. $A \cdot A' = A' \cdot A = I$), its inverese is the transposed matrix A' with $X = A'Z$ or

$$\begin{pmatrix} x_{11} & \dots & x_{1i} & \dots & x_{1n} \\ \vdots & & \vdots & & \vdots \\ x_{p1} & \dots & x_{pi} & \dots & x_{pn} \end{pmatrix} = \begin{pmatrix} \alpha_{11} & \dots & \alpha_{1r} \\ \vdots & & \vdots \\ \alpha_{p1} & \dots & \alpha_{pr} \end{pmatrix} \cdot \begin{pmatrix} z_{11} & \dots & z_{1i} & \dots & z_{1n} \\ \vdots & & \vdots & & \vdots \\ z_{r1} & \dots & z_{ri} & \dots & z_{rn} \end{pmatrix}$$

so the k-th observation vector of X can be reconstructred using the k-th score (column of Z):

$$\begin{pmatrix} x_{1k} \\ \vdots \\ x_{pk} \end{pmatrix} = \begin{pmatrix} \alpha_{11} \\ \vdots \\ \alpha_{p1} \end{pmatrix} z_{1k} + \dots + \begin{pmatrix} \alpha_{1r} \\ \vdots \\ \alpha_{pr} \end{pmatrix} z_{rk}, \tag{9.3}$$

which is a linear combination of the eigenvectors $\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_r \in \mathbb{R}^p$ with coefficents $z_{1k}, \dots, z_{rk} \in \mathbb{R}$ being the corresponding scores.

When using PCA as a dimension reduction technique, i.e. consider only the first $q \leq r$ PCs, the reconstruction is done using only the first q scores and loading vectors.

$$\begin{pmatrix} x^*_{1k} \\ \vdots \\ x^*_{pk} \end{pmatrix} = \begin{pmatrix} \alpha_{11} \\ \vdots \\ \alpha_{p1} \end{pmatrix} z_{1k} + \dots + \begin{pmatrix} \alpha_{1q} \\ \vdots \\ \alpha_{pq} \end{pmatrix} z_{qk} + 0 + \dots + 0, \tag{9.4}$$

with x^*_{jk} being approximated, noiseless observations.

9.2 Model Similarity: Cluster Analysis

The aim here is to find groups of simulations based on their behaviour regarding the common climate change patterns obtained from the PCA. Those groups of simulations are found based on a hierarchical clustering algorithm which works like this: First, each simulation is assigned to its own cluster, and then the algorithm proceeds iteratively joining the two closest clusters in each agglomeration step until one cluster remains. The measure of distance is based on Ward's criterion, which finds new clusters with minimal variance. This procedure tends to find compact and spherical groups of data. Hierarchical clustering results in a tree-like similarity structure, which is meaningful if we believe that some particular clusters might be more closely related to other clusters.

Having obtained a tree-like structure of the dependence of the simulations (Fig. 10.1), the open question of how many simulations to actually select from the ensemble still remains. There is no unique and best solution to this problem, but there are some criteria on how to obtain an optimal amount of clusters. Our approach is to consider the distance criterion for each agglomeration step, which in case of Ward's criterion is the variance increase of the newly merged two clusters. We cut the tree where this increase of variance does not change considerably.

9.3 Model Selection: Sampling

We extend the idea of non-probability sampling by sampling one simulation out of each group of similar models obtained from the cluster analysis. This approach is also known as quota sampling, where one selects members out of each group with key information/characteristics relevant for the phenomenon being studied. This can be done by picking simulations from the scatterplot on the PCs. Another way would be to look at the distribution of the PCs for each simulation individually, which can be displayed with a bar-chart denoting their location within the scatterplots. Starting from an average simulation (all bars/PCs being close to 0) and then selecting one simulation with distinct extreme characteristics from each cluster. We will elaborate these methods in the next section.

10 Case Study: Model Selection for European Domain

We apply the methods explained above within an exemplary case study to select regional climate models for climate impact studies. The study is motivated by the EU-FP7 project IMPACT2C (e.g. Vautard et al. 2014) to seek driving data for multiple diverse impact models spread across the whole European continent. The large amount of different impact modelling groups drastically increases the number of meteorological variables which have to be considered. The analysis for this example, including the R code and the data can be found in the electronic supplementary material online.

10.1 Climate Data

The MME used here is the multi-model dataset from the ENSEMBLES project. In total there are 27 ENSEMBLES regional climate model simulations which are driven by 10 different General Circulation Models (GCMs) (all forced by the SRES A1B emission scenario Nakicenovic, Alcamo and Davis 2000), where the *ECHAM5* GCM appears three times using different initialisation and the *HadCM3* GCM also shows up with three different parametrization schemes (*Q0*, *Q3* and *Q16*). One RCM (*KNMI*) has been forced by all three *ECHAM5* realisations, however with a rougher horizontal resolution of 50km. Also, one additional simulation has been driven with this resolution, the remaining 25 simulations have a 25km resolution. As one simulation (*GKSS-CCLM4.8-IPSL*) lacks the variables relative humidity (HURS) and global radiation (RSDS), it has been omitted in this study, leading to 9 different driving GCMs. In addition one simulation (*OURANOS-CRCM-CGCM3*) shows very noticeable biases and has been excluded as well, leaving $n = 25$ regional climate simulations for the analysis driven by 8 different GCMs. The baseline period for the climate change signal is the 30 years average of 1971 to 2000. The future scenario period to determine the climate change signal is chosen to be 2021 to 2050.

The most important meteorological drivers for climate change impacts in the European study have been defined on the basis of a user survey among project partners, and experience from previous projects. In total, $p_{par} = 5$ parameters have been selected, being mean air temperature (TAS), precipitation amount (PR), HURS, RSDS and wind speed (WSS). The climate change signals of these variables are analyzed in subregions of Europe (as in J. H. Christensen and O. B. Christensen 2007) by aggregating spatially over $p_{spat} = 8$ domains for the $p_{seas} = 4$ seasons summer (JJA), winter (DJF), spring (MAM) and autumn (SON). In total, this gives 160 different parameters. We obtained the climate change signals using the R package *wux* Mendlik, Heinrich, A. Gobiet et al. 2016.

10.2 Common patterns of climate change

We use PCA to reduce the dimension space $p = p_{par} \times p_{seas} \times p_{spat} = 5 \cdot 4 \cdot 8 = 160$ of the $n = 25$ climate models. The broken-stick method detects 3-4 robust PCs (Fig. 9.2), excluding the remaining PCs as being random noise. We decided to reduce the dimensionality to $p_{red} = 3$ PCs.

The most dominant climate change pattern (PC1) is the temperature change along all four seasons for all European subregions. PC2 shows a negative relationship between HURS and RSDS. This means that simulations projecting a higher change in HURS than others tend to project a lower change in RSDS. This anti-correlation seems to hold for the entire European region in DJF and for the northern and eastern parts of Europe in the remaining seasons, especially in MAM and SON. A positive correlation of humidity and precipitation can be detected for the Scandinavian region over the whole year and in winter for mid- and eastern Europe and the Alpine region. PC3 shows a humidity-precipitation pattern for the southern regions for MAM (not shown).

10.3 Model similarity

Based on the first $p_{red} = 3$ PCs we performed a hierarchical cluster analysis as described in Section 9.2.

The tree-like dependency structure is visualised by a dendrogram in Figure 10.1. The height of the branches depict the measure of dissimilarity between simulations and clusters regarding the common patterns of climate change. The heights are used to detect the optimal number of clusters: Starting from one cluster with the highest height we increase the amount of clusters until there is no substantial change in heights any more. The heights are shown as a barplot in Figure 10.1.

We show partitions with 5 clusters to visualise the range of reasonable clustering. Notably, simulations driven by the lateral boundary conditions of the GCMs *ECHAM5-r3*, *BCM* and *ARPEGE* show very strong GCM specific clustering, meaning that those RCMs driven by the same GCM behave rather similar regarding the common patterns of climate change. Further, *ECHAM5-r3* and *BCM* driven simulations tend to be more similar than *ARPEGE* models. Interestingly, the 50km versions of the RCM *KNMI* driven by *ECHAM5-r1* and by *ECHAM5-r2* behave rather different than the *ECHAM5-r3* driven version and they are spread among different clusters. The simulations *KNMI* and *SMHI* both have identical set-ups using a 25km and a 50km resolution. In each

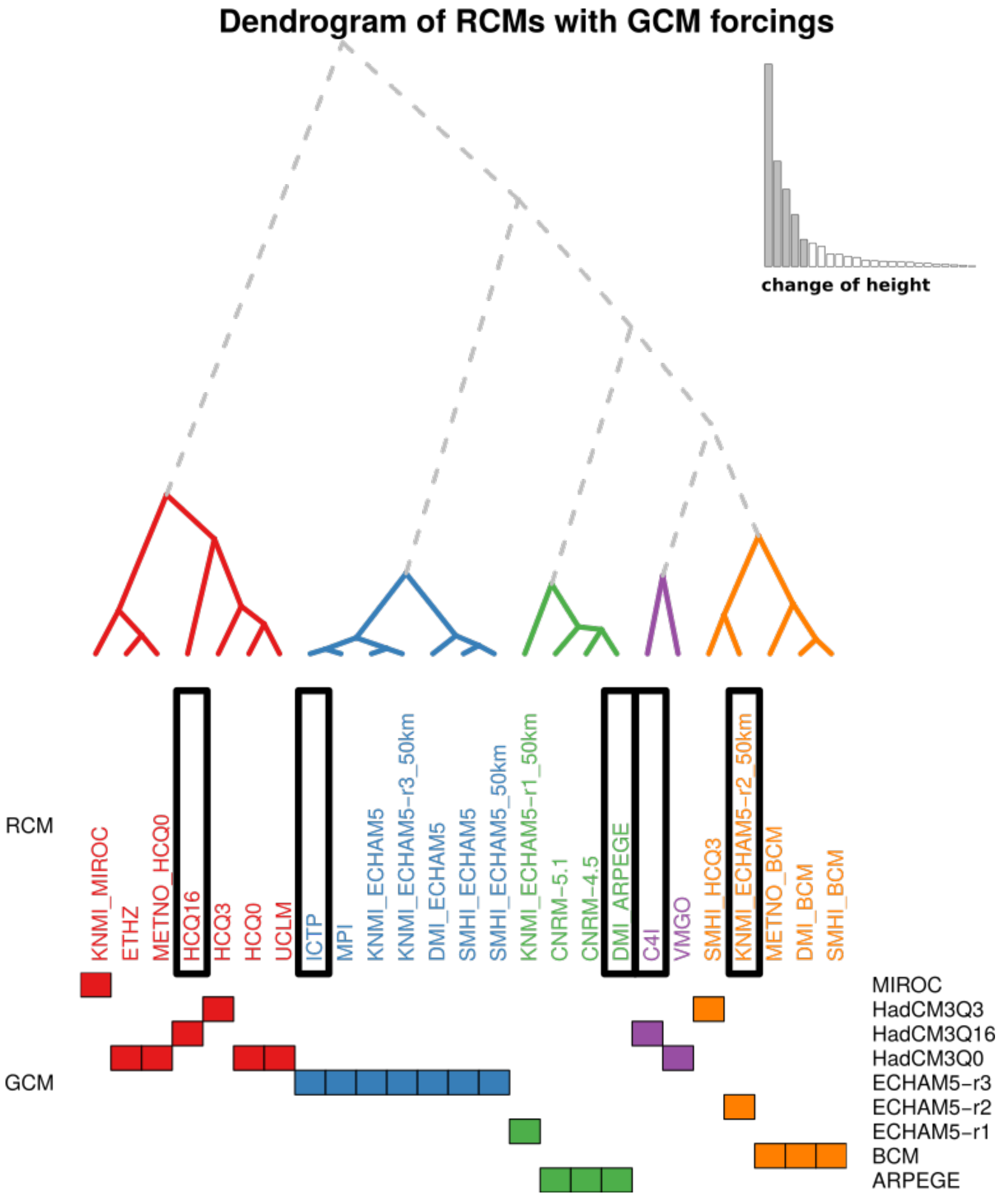


Figure 10.1: Cluster dendrogram for the first 3 PCs showing 5 clusters.The boxes on the bottom show the driving GCM of the corresponding RCM. The barplot on the top right shows the distance criterion (change of height within the dendrogram) for each agglomeration step when merging clusters. (Source: Mendlik and A. Gobiet 2016).

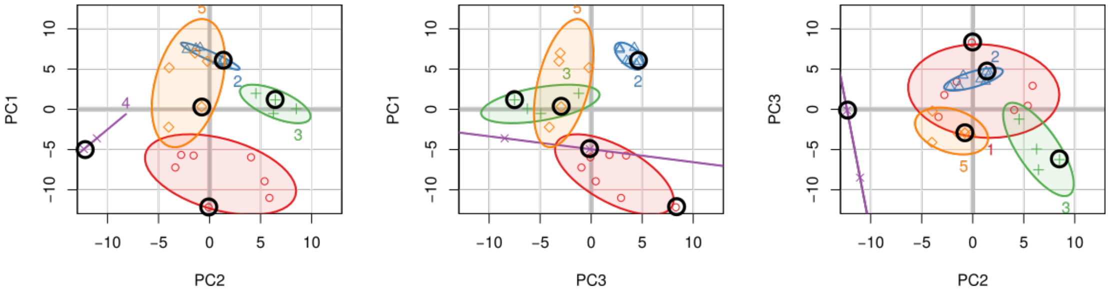


Figure 10.2: Climate change signals of the ENSEMBLES RCMs within the principal component space, showing the 75% variance ellipsoid within each cluster. The selected models are highlighted. (Source: Mendlik and A. Gobiet 2016).

case, the similarity is very high. On the other hand, RCMs driven by the three GCMs *HadCM3* show quite some heterogeneity: On the one hand they are split among two clusters of different sizes. On the other hand, RCMs driven by different *HadCM3* GCMs can be found in either cluster. Also the *MIROC* driven *KNMI* does not form a cluster of its own but behaves similar to *HadCM3* driven RCMs.

10.4 Model selection

Our model selection approach identifies 5 groups of similar simulations. As shown in Fig. 10.1, these groups also show dependencies, some more than others, but much weaker than between the individual simulations. By selecting one simulation of each cluster, we definitely reduced the model dependency and obtained a more independent ensemble. We decided to select an average climate simulation and 4 extreme simulations to span the uncertainty range.

Figure 10.2 depicts the climate change signals of the regional climate simulations on the principal component space with regard to the first three PCs. Simulations close to 0 can be interpreted as having an “average pattern” of the climate change induced by the corresponding principal component. The sign and order of magnitude in the scatterplot (Fig. 10.2) corresponds to the pattern described in the corresponding PC from the loadings plot in Figure 9.1. For PC1 (warming pattern), simulations within cluster 1 show highest changes, whereas cluster 2 and 5 tend to have cooler projections and cluster 3 is average. For PC2 (humidity pattern) cluster 4 and cluster 3 show distinct behaviour. PC3 (precipitation and humidity pattern) mostly distinguishes between cluster 2, 3 and 5.

The individual locations within the scatterplots can be visualised with barplots (Fig. 10.3), which makes sampling easier. We started by choosing one average simulation, being closest to 0 within all PC spaces. Then, out of each cluster, we took one ex-

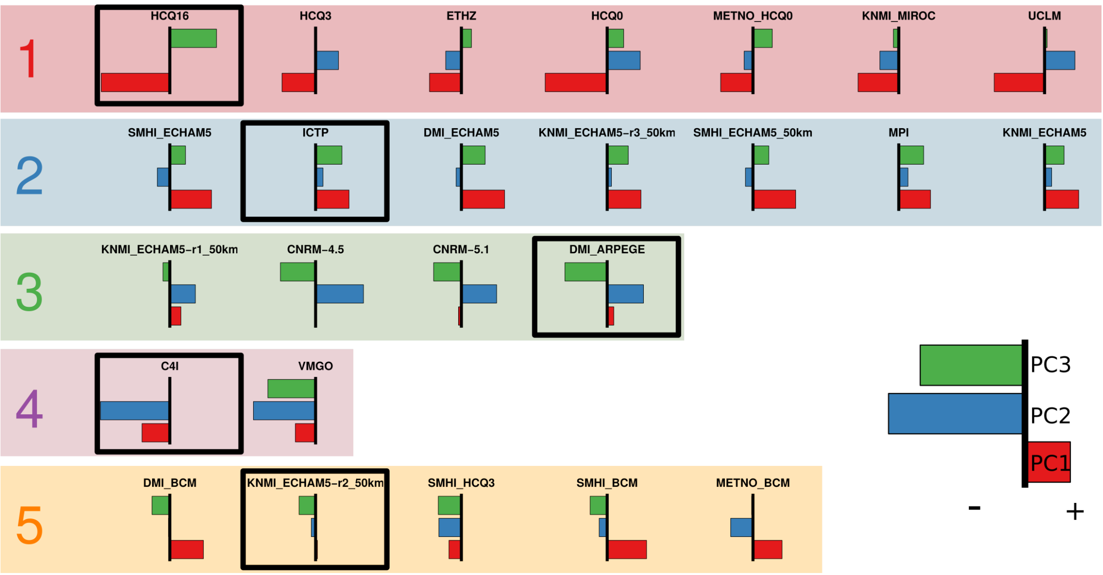


Figure 10.3: Climate change signals for each RCM simulation in the principal component space. The simulations are split according to their clustering. The suggested selections are highlighted. (Source: Mendlik and A. Gobiet 2016).

treme representative to obtain maximum diversity of our sub-sample. One possible model selection could be the following: *KNMI-ECHAM5-r2-50km* (average behavior), *C4I-HadCM3Q16* (low PC2), *DMI-ARPEGE* (high PC2, low PC3), *ICTP-ECHAM5-r3* (high PC1, high PC3) and *HCQ16-HadCM3Q16* (low PC1, high PC3). This selection is marked in Figures 10.1, 10.2 and 10.3.

Here, some driving GCMs appear two times, such as *ECHAM5* and *HadCM3Q16*, as the corresponding RCM does project a very different pattern of climate change. However, also other constellations could be possible, still capturing the extreme characteristics.

11 Summary and Discussion

In order to provide sound meteorological input for climate change impact studies, it is important to address the uncertainty induced by climate simulations. We present a simple tool to aid the user to select appropriate climate simulations (either GCMs or RCMs) as input for their studies. The aim of the proposed method is to sample the climate model uncertainty, find model similarities and sub select models being as independent as possible while conserving the spread of the full ensemble.

Our method generalises the pragmatic approach of finding the model spread of climate change signals of, say, temperature against precipitation (IPCC-TGICA 2007). It allows for simultaneous analysis of an arbitrary amount of meteorological parameters over several spatial regions of interest, and brings forth dominating patterns of climate change. Model similarities are detected based entirely on those patterns of projected change. This stands in contrast to most other studies, which find similarities in the 20th century historical runs (e.g. Abramowitz and Gupta 2008; Bishop and Abramowitz 2012; Pennell and Reichler 2010). An interesting aspect for further research would be the question of how model similarities in historical runs and dependencies in future projections relate over time.

The selected simulations in the sub-ensemble conserve the main climate change characteristics of the entire ensemble, but it might not share identical statistical properties like mean and standard deviation. This is a desirable property, as unbalanced ensemble designs often lead to biased estimates due to double-counting induced by model dependencies. For balanced and thus unbiased ensembles, like for reconstructed datasets (Heinrich, A. Gobiet and Mendlik 2014), the statistical properties are conserved, as the sample size is equally decreased in each cluster.

We do not discuss model selection based on performance, as different models show different strengths and weaknesses depending on the metric (Knutti, Abramowitz et al. 2010). We tend to the pragmatic approach of excluding few simulations with severe and clearly demonstrated deficiencies and keeping as many simulations as possible as input for the model selection procedure.

It should be noted that the proposed method does not deliver one single and unique subset. Instead the user has to decide on how to select one simulation out of each cluster. This can be done with probabilistic (random) sampling or non-probabilistic sampling. We do not recommend any type of random sampling as it is vulnerable to random sampling error: The randomness of the selection can result in a subset which is not representative for the ensemble. The probability of such a misspecification increases with decreasing sample size. For such small sample sizes it is more advisable to take most extreme simulations to sample the entire model-spread.

Cannon 2015 proposes a very interesting alternative model selection algorithm, addressing the same problems as presented in this work (multivariate set-up and model dependency). However, in contrast to the method proposed in this work, selected models are uniquely identified. This surely makes model selection simpler, but there is no flexibility for the user to add some subjective selection criteria when sampling, like the inclusion of an extremely well-performing simulation.

We demonstrate the presented model selection procedure with the ENSEMBLES multi-model dataset (Chapter 10). Our results show that the first two most dominating patterns of change relate to temperature and humidity and that the dataset can be split into 5 groups of similar simulations. A dominant factor for model similarity in this setting is the GCM forcing of the RCM. Interestingly, some GCM forcings lead to very dense clusters (*ECHAM5-r3*), while others are very heterogeneous and may even be split among different groups of similarity. This is particularly the case for the different initial conditions of *ECHAM5* (*r1*, *r2* and *r3*), each inducing a distinct behaviour of the RCMs. On the other hand, some driving GCMs do not create own clusters at all (e.g *MIROC*). In our example application this leads to a selection where two GCMs appear twice, whereas others are omitted entirely. Selecting simulations from each GCM would not necessarily span the entire uncertainty range.

However, our method is not restricted to the selection of RCMs, as a matter of fact it can also be used to select suitable GCMs.

To sum it up, we present a flexible method to select models from an ensemble of simulations conserving the model spread and accounting for model similarity. This reduces computational costs for climate impact modelling and enhances the quality of the ensemble at the same time, as it prevents double-counting of dependent simulations which would lead to biased statistics.

Part IV

Quantify Climate Uncertainty

The aim of this part is to estimate the expected climate change and provide a corresponding measure of uncertainty, representing the spread of future climate projections. We focus on three case study regions in Europe which are known to have very different impacts on climate change, namely the Alpine Region (AL), the Iberian Peninsula (IP) and the Scandinavian Region (SC).

The method presented here explicitly accounts for several sources of uncertainty (Section 3.1) by incorporating the entire state-of-the art multi-model ensemble Coupled Model Intercomparison Project Phase 5 (CMIP5) as described in Section 2.3, preprocessed with the "wux" package (Part II). The climate change uncertainty is quantified with a hierarchical regression model, which best accounts for the problematic structure of the data set (e.g. being massively imbalanced, see Section 3.1). The novelty of this approach is the explicit implementation of the dependence between climate models which share core components simulating the climate system in a similar way (Masson and Knutti 2011, Knutti, Masson and Gettelman 2013). Further this study shows that the common conception of Gaussian distribution (e.g. Zubler et al. 2015) is severely wrong and has a huge impact on the uncertainty estimates. We therefore implement a skewed distribution which better captures the data. Also, we provide $(1 - \alpha)$-confidence intervals (CIs) for each estimate of the hierarchical regression model and compare them using different methods. These CIs can be regarded as "uncertainties of the uncertainty estimates".

We therefore derive the Maximum-Likelihood estimate of a skew-normal mixed model and directly maximise the objective function with a numerical solver. Here, the EM-algorithm proposed in the literature does not work very well, it is in fact converging slower than brute-force maximisation.

12 Multilevel Regression Models

12.1 Motivation: Violating Independence Assumption

Let $Y_1, \ldots, Y_n$ be Random Variables (RVs) which can be formulated as

$$Y_i = \beta_0 + \sum_{j=1}^{p-1} \beta_j z_{ij} + \epsilon_i \tag{12.1}$$

with ϵ_i being a independent and identically distributed (*iid*) RV with

$$\epsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

We can write this formula as

$$Y_i = \boldsymbol{x}_i' \boldsymbol{\beta} + \epsilon_i, \quad i = 1, \ldots, n \tag{12.2}$$

with $\boldsymbol{x}_i' = (1, z_{i1}, \ldots, z_{ip-1})$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{p-1})'$. The n equations can be written more compactly as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{12.3}$$

with $\boldsymbol{Y} = (Y_1, \ldots, Y_n)' \in \mathbb{R}^n$ and $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)' \in \mathbb{R}^{n \times p}$ and unobservable errors $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)' \in \mathbb{R}^n$ with

$$\mathrm{E}(\boldsymbol{\epsilon}) = \boldsymbol{0} \text{ and } \mathrm{Var}(\boldsymbol{Y}) = \mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2 I_n. \tag{12.4}$$

To estimate the parameter vector, we usually use the ordinary least-squares (OLS) estimate

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \tag{12.5}$$

which we also obtain by maximising the likelihood function. The uncertainty (precision) of this parameter estimate is then

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}) = \mathrm{Var}\left((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}\right) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' \, \mathrm{Var}(\boldsymbol{Y})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} \tag{12.6}$$

which can be estimated with

$$\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}}) = \widehat{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

Introducing Dependence So far we assumed the data being independent, for example stemming from an *iid* normal distribution. But what happens to the estimate (12.5) if we violate this assumption?

Let's assume the data is dependent with $\mathrm{Var}(\boldsymbol{Y}) = V$. Then the expected value of the OLS estimate (12.5) is still

$$\mathrm{E}(\widehat{\boldsymbol{\beta}}_{dep}) = \mathrm{E}\left((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}\right) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\,\mathrm{E}(\boldsymbol{Y}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta},$$

meaning we have an unbiased estimate. We write $\widehat{\boldsymbol{\beta}}_{dep}$ to emphasise that the underlying data are dependent. However, the uncertainty of the estimate $\widehat{\boldsymbol{\beta}}_{dep}$ is now

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}_{dep}) = \mathrm{Var}((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'V\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}. \tag{12.7}$$

It can be shown that the diagonal entries of the covariance matrix (12.7) are always larger than the diagonal entries of the covariance matrix (12.6) of the estimate under *iid*.

If $V = \sigma^2 I_n$, the uncertainty (12.7) attains the Cramér-Rao lower bound $\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$. And because this is exactly the uncertainty of the *iid* case $\mathrm{Var}(\widehat{\boldsymbol{\beta}}_{iid}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$ we get

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}_{dep}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'V\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \geq I^{-1}(\sigma^2) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} = \mathrm{Var}(\widehat{\boldsymbol{\beta}}_{iid}). \tag{12.8}$$

This means that dependence between observed data **increase the uncertainty** of the estimate. Or in other words, if the model is mis-specified as *iid*, the true uncertainty will be underestimated.

12.2 2-stage Mixed Regression Model

When dealing with dependent, clustered data, assume $Y_{ij} \in \mathbb{R}$ is a response of the j-th member within the i-th cluster. Assume we have $i = 1, \dots, m$ clusters in total and within each cluster we have $j = 1, \dots, N_i$ observations Y_{ij}. Then we can describe this dependency structure with a linear mixed-effects model (LMM)

$$Y_{ij} = \sum_{l=1}^{p} x_{ijl}\beta_l + \sum_{k=1}^{d} z_{ijk}b_{ik} + \epsilon_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{b}_i + \epsilon_{ij}, \tag{12.9}$$

with fixed but unknown $\boldsymbol{\beta} \in \mathbb{R}^p$ and random $\boldsymbol{b}_i \in \mathbb{R}^d$. In a more compact notation the LMM (12.9) can be written as

$$\begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{iN_i} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}'_{i1} \\ \vdots \\ \boldsymbol{x}'_{iN_i} \end{pmatrix} \boldsymbol{\beta} + \begin{pmatrix} \boldsymbol{z}'_{i1} \\ \vdots \\ \boldsymbol{z}'_{iN_i} \end{pmatrix} \boldsymbol{b}_i + \begin{pmatrix} \epsilon_{i1} \\ \vdots \\ \epsilon_{iN_i} \end{pmatrix} \tag{12.10}$$

or

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \tag{12.11}$$

with $\boldsymbol{Y}_i = (Y_{i1}, \dots, Y_{iN_i})' \in \mathbb{R}^{N_i}$ the responses within cluster i, with design matrices $\boldsymbol{X}_i = (x'_{i1}, \dots, x'_{iN_i})' \in \mathbb{R}^{N_i \times p}$ and $\boldsymbol{Z}_i = (z'_{i1}, \dots, z'_{iN_i})' \in \mathbb{R}^{N_i \times d}$ and with

$$\boldsymbol{b}_i \sim N_d(\boldsymbol{0}, D), \quad \boldsymbol{\epsilon}_i \sim N_{N_i}(\boldsymbol{0}, \Sigma_i), \tag{12.12}$$

$$\boldsymbol{b}_1, \dots, \boldsymbol{b}_m, \boldsymbol{\epsilon}_1, \dots, \boldsymbol{\epsilon}_m \text{ independent}, \tag{12.13}$$

with unknown variance component (VC) $D \in \mathbb{R}^{d \times d}$ and $\Sigma_i \in \mathbb{R}^{N_i \times N_i}$.

Vectorization We can vectorize the responses of the $i = 1, \dots, m$ clusters to a more compact notation

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}$$

with the total amount of observations $N = \sum_{i=1}^{m} N_i$

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_1 \\ \vdots \\ \boldsymbol{Y}_m \end{pmatrix} \in \mathbb{R}^N, \quad \boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_m \end{pmatrix} \in \mathbb{R}^{N \times p}, \tag{12.14}$$

and the design matrix for the random effects defining the dependency structure of the clustered dataset with

$$\boldsymbol{Z} = \begin{pmatrix} \boldsymbol{Z}_1 & 0_{N_1 \times d} & \dots & 0_{N_1 \times d} \\ 0_{N_2 \times d} & \boldsymbol{Z}_2 & & \\ \vdots & & \ddots & \\ 0_{N_m \times d} & & & \boldsymbol{Z}_m \end{pmatrix} \in \mathbb{R}^{N \times q}, \text{ with } q = d \cdot m, \tag{12.15}$$

where q equals the dimensionality d of the random effects multiplied by the amount of clusters m and further

$$0_{N_i \times d} = \begin{pmatrix} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{pmatrix} \in \mathbb{R}^{N_i \times d}. \tag{12.16}$$

The random quantities become

$$\boldsymbol{\epsilon} = \begin{pmatrix} \boldsymbol{\epsilon}_1 \\ \vdots \\ \boldsymbol{\epsilon}_m \end{pmatrix} \in \mathbb{R}^N, \quad \boldsymbol{b} = \begin{pmatrix} \boldsymbol{b}_1 \\ \vdots \\ \boldsymbol{b}_m \end{pmatrix} \in \mathbb{R}^q, \tag{12.17}$$

Dimension	Description
Random Effects	
$q = d \cdot m$	vector length of random effect $\boldsymbol{b}_i$
d	dimension of random effect $\boldsymbol{b}_i$
m	number of clusters
Fixed Effects	
p	dimension of fixed effect $\boldsymbol{\beta}$
Observations	
N_i	number of observations in cluster i
$N = \sum_{i=1}^{m} N_i$	number of total observations

Table 12.1: Indices for 2-stage random effects model.

Observations	Fixed Effects		Random Effects	
y_{ij} (1×1)	$\boldsymbol{x}'_{ij}$ $(1 \times p)$	$\boldsymbol{\beta}$ $(p \times 1)$	$\boldsymbol{z}'_{ij}$ $(1 \times d)$	$\boldsymbol{b}_i$ $(d \times 1)$
$\boldsymbol{y}_i$ $(N_i \times 1)$	$\boldsymbol{X}_i$ $(N_i \times p)$	$\boldsymbol{\beta}$ $(p \times 1)$	$\boldsymbol{Z}_i$ $(N_i \times d)$	$\boldsymbol{b}_i$ $(d \times 1)$
$\boldsymbol{y}$ $(N \times 1)$	$\boldsymbol{X}$ $(N \times p)$	$\boldsymbol{\beta}$ $(p \times 1)$	$\boldsymbol{Z}$ $(N \times q)$	$\boldsymbol{b}$ $(q \times 1)$

Table 12.2: Dimensions of components at different vectorizations.

with the VCs

$$\boldsymbol{D} = \begin{pmatrix} D & & \boldsymbol{0} \\ & \ddots & \\ \boldsymbol{0} & & D \end{pmatrix} \in \mathbb{R}^{q \times q}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \Sigma_1 & & \boldsymbol{0} \\ & \ddots & \\ \boldsymbol{0} & & \Sigma_m \end{pmatrix} \in \mathbb{R}^{N \times N}. \tag{12.18}$$

The notation of the indices is summarised in Table 12.1, and Table 12.2 provides an overview over the dimensions for different vectorizations of the response Y_{ij}. The fixed effects $\boldsymbol{\beta}$, the variance components D and $\Sigma_1, \ldots, \Sigma_m$ can be estimated by iteratively solving *Henderson's mixed model equations* (Henderson 1982) which can be obtained by maximising the marginal likelihood of $\boldsymbol{Y}$.

Remark 12.1. If the random effects are normally distributed we obtain

$$\begin{pmatrix} \boldsymbol{\epsilon} \\ \boldsymbol{b} \end{pmatrix} \sim N_{N+q}\left(\begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D} \end{pmatrix} \right). \tag{12.19}$$

	$r_1(a)$	$r_1(b)$	$r_2(b)$	$r_3(b)$	$r_4(b)$	$r_1(c)$	$r_2(c)$	
a	1	.	.	.	.	.	.	$N_1 = 1$
GCM b	.	1	1	1	1	.	.	$N_2 = 4$
c	.	.	.	.	.	1	1	$N_3 = 2$
								$N = 7$

Table 12.3: Cross-table: Three climate simulations (a, b, c) having different amount of initialisation runs.

Example 12.2 (Dependency structure: Climate models with several initial runs)**.** Let the responses $Y_{ij} \in \mathbb{R}$ be the projected temperature climate change of global climate simulations $i = 1, \ldots, m$, $m = 3$. Each simulation has been started several times, each time with different initial conditions (see Section 3.1) so we have $j = 1, \ldots, N_i$ realisations with $N_1 = 1, N_2 = 4, N_3 = 2$. The design is summarised in Table 12.3.

We model the j-th run of the i-th simulation as

$$Y_{ij} = \beta + b_i + \epsilon_{ij},$$

in which case

$$\begin{pmatrix} Y_{11} \\ Y_{21} \\ Y_{22} \\ Y_{23} \\ Y_{24} \\ Y_{31} \\ Y_{32} \end{pmatrix} = \begin{pmatrix} \beta \\ \beta \\ \beta \\ \beta \\ \beta \\ \beta \\ \beta \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \\ b_2 \\ b_2 \\ b_2 \\ b_3 \\ b_3 \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \\ \epsilon_{24} \\ \epsilon_{31} \\ \epsilon_{32} \end{pmatrix}$$

which can be rewritten as

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\beta + \boldsymbol{Z}_i b_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, m$$

with $m = 3$ and $d = 1$ and

$$\boldsymbol{X}_1 = 1 \in \mathbb{R}^{N_1 \times 1}, \quad \boldsymbol{X}_2 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \in \mathbb{R}^{N_2 \times 1}, \quad \boldsymbol{X}_3 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \in \mathbb{R}^{N_3 \times 1},$$

and with same $\boldsymbol{Z}_i$

$$\boldsymbol{Z}_1 = 1 \in \mathbb{R}^{N_1 \times 1}, \quad \boldsymbol{Z}_2 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \in \mathbb{R}^{N_2 \times 1}, \quad \boldsymbol{Z}_3 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \in \mathbb{R}^{N_3 \times 1},$$

because the dimension of β is $p = 1$ and the sample sizes for each simulation being $N_1 = 1, N_2 = 4, N_3 = 2$. However, when putting together the different climate simulations (i.e. clusters of data), the model becomes

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon},$$

with

$$\boldsymbol{X} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \in \mathbb{R}^{N \times p}, \boldsymbol{Z} = \begin{pmatrix} \boldsymbol{Z}_1 & \boldsymbol{0}_{N_1 \times d} & \boldsymbol{0}_{N_1 \times d} \\ \boldsymbol{0}_{N_2 \times d} & \boldsymbol{Z}_2 & \boldsymbol{0}_{N_2 \times d} \\ \boldsymbol{0}_{N_3 \times d} & \boldsymbol{0}_{N_3 \times d} & \boldsymbol{Z}_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \in \mathbb{R}^{N \times q}, \boldsymbol{b} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} \in \mathbb{R}^q.$$

with total number of observations $N = 7$ and the dimensionality of the random effects vector being the amount of clusters ($m = 3$ climate simulation) times the amount of random effects ($d = 1$) being $q = d \cdot m = 3 \cdot 1 = 3$ and $\boldsymbol{Z}'$ being exactly the design Table 12.3.

Remark 12.3 (Hierarchical Model). The LMM can be written as a two level hierarchical model

$$\boldsymbol{Y}|\boldsymbol{b} \sim N(\boldsymbol{X\beta} + \boldsymbol{Zb}, \boldsymbol{\Sigma}) \tag{12.20}$$
$$\boldsymbol{b} \sim N(\boldsymbol{0}, \boldsymbol{D}) \tag{12.21}$$

The LMM is defined in terms of the conditional distribution, let us take a look at the marginal distribution of the responses.

Remark 12.4 (Marginal Distribution). The marginal distribution of the 2 level LMM $\boldsymbol{Y}$ is again a normal distribution with expected value

$$E(\boldsymbol{Y}) = E(E(\boldsymbol{Y}|\boldsymbol{b})) = E(\boldsymbol{X\beta} + \boldsymbol{Zb}) = \boldsymbol{X\beta} \tag{12.22}$$

and variance

$$Var(\boldsymbol{Y}) = Var(E(\boldsymbol{Y}|\boldsymbol{b})) + E(Var(\boldsymbol{Y}|\boldsymbol{b})) \tag{12.23}$$
$$= Var(\boldsymbol{X\beta} + \boldsymbol{Zb}) + \boldsymbol{\Sigma} \tag{12.24}$$
$$= \boldsymbol{ZDZ}' + \boldsymbol{\Sigma} \tag{12.25}$$

due to the law of conditional variance. The marginal distribution is thus

$$\boldsymbol{Y} \sim N(\boldsymbol{X\beta}, \boldsymbol{\psi}), \quad \boldsymbol{\psi} = \boldsymbol{ZDZ}' + \boldsymbol{\Sigma}. \tag{12.26}$$

The marginal distribution in Remark 12.4 allows us to interpret any 2-stage mixed model as an ordinary linear regression model with dependent error term

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}^*, \quad \boldsymbol{\epsilon}^* \sim N(\boldsymbol{0}, \boldsymbol{\psi}), \quad \boldsymbol{\psi} = \boldsymbol{ZDZ}' + \boldsymbol{\Sigma}.$$

Example 12.5 (Marginal Variance of Example 12.2)**.** Let us assume that the two-stage LMM

$$Y_{ij} = \beta + b_i + \epsilon_{ij}, \quad i = 1, \dots, m, \quad j = 1, \dots, N_i$$

has independent errors and random effects

$$\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2), \quad b_i \overset{iid}{\sim} N(0, \tau^2),$$

which leads to the variance matrices

$$Var(\boldsymbol{\epsilon}) = \boldsymbol{\Sigma} = \sigma^2 I_N, \quad Var(\boldsymbol{b}) = \boldsymbol{D} = \tau^2 I_q.$$

With $\boldsymbol{ZZ}'$ being the block-diagonal matrix

$$\boldsymbol{ZZ}' = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix} \in \mathbb{R}^{N \times N},$$

the marginal variance of the observations is

$$\begin{aligned} Var(\boldsymbol{Y}) &= \boldsymbol{\psi} = \tau^2 \boldsymbol{ZZ}' + \sigma^2 I_N \\ &= \begin{pmatrix} \tau^2+\sigma^2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \tau^2+\sigma^2 & \tau^2 & \tau^2 & \tau^2 & 0 & 0 \\ 0 & \tau^2 & \tau^2+\sigma^2 & \tau^2 & \tau^2 & 0 & 0 \\ 0 & \tau^2 & \tau^2 & \tau^2+\sigma^2 & \tau^2 & 0 & 0 \\ 0 & \tau^2 & \tau^2 & \tau^2 & \tau^2+\sigma^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \tau^2+\sigma^2 & \tau^2 \\ 0 & 0 & 0 & 0 & 0 & \tau^2 & \tau^2+\sigma^2 \end{pmatrix}. \end{aligned}$$

So the data-points are not independent marginally.

Example 12.6 (Marginal Variance Structure of climate models)**.** Consider the real data y_{ij} being the projected warming of GCMs $i = 1, \dots, 37$ in the Alpine area. Some models i have been run several times ($j = 1, \dots, N_i$) with slightly different initial conditions,

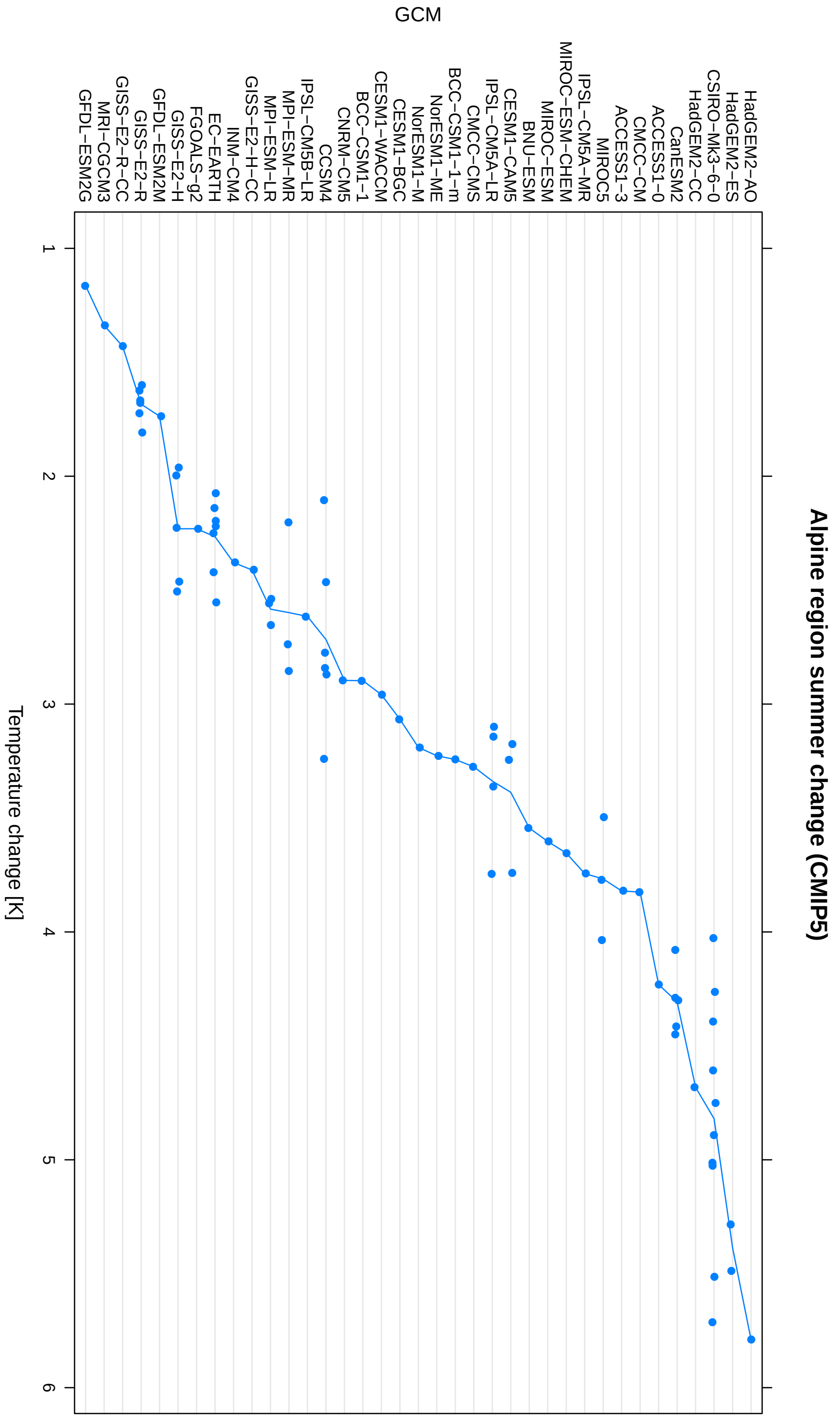


Figure 12.1: Temperature changes (x-axis) of individual GCMs (y-axis). The GCMs differ in their amount of initialisation runs: some have only one realisation where other have up to 10 ($1 \le N_i \le 10$) leading to an unbalanced design.

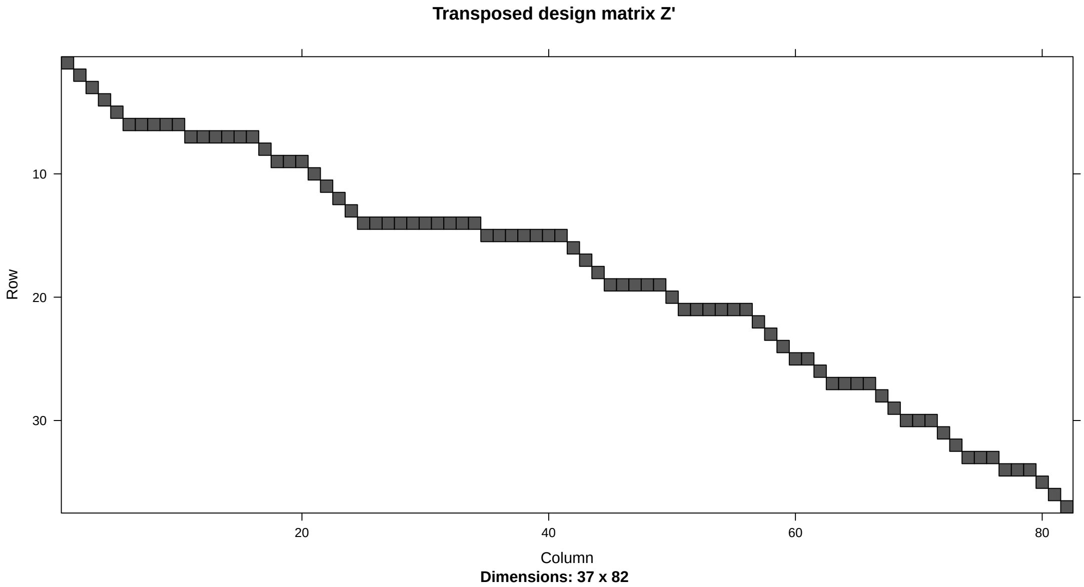


Figure 12.2: Design matrix $\boldsymbol{Z} \in \mathbb{R}^{N\times q}$ of random effects (here transposed). The dark boxes denote values with 1, where the rest is 0. We have $N = 82$ total observations and $q = m \cdot d = 37 \cdot 1$ with $m = 37$ clusters.

and we expect those to behave similarly. The responses can be seen in Figure 12.1, where we can clearly see clustering within GCMs. Lets take the same model as in Example 12.4

$$Y_{ij} = \beta + b_i + \epsilon_{ij}, \quad i = 1, \ldots, m, \quad j = 1, \ldots, N_i$$

with $m = 37$ and with the random GCM effect $b_i \overset{iid}{\sim} N(0, \tau^2)$ and the residuals $\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$. The VC τ^2 represents the spread between the GCMs and σ^2 the spread of the runs within a GCM. The design matrix $\boldsymbol{Z}$ in Figure 12.2 depicts the nested design of the data. We obtain the estimated VCs of $\hat{\sigma}^2 = 0.1$ and $\hat{\tau}^2 = 1.07$. The marginal covariance is thus $\hat{\sigma}^2 + \hat{\tau}^2 = 0.1 + 1.07 = 1.17$. The correlation estimate between the GCMs is here defined to be 0 and the correlation between the runs of the same GCM are $\hat{\tau}^2/(\hat{\sigma}^2 + \hat{\tau}^2) = .91$. The correlation structure between all available observations can be seen in Figure 12.3.

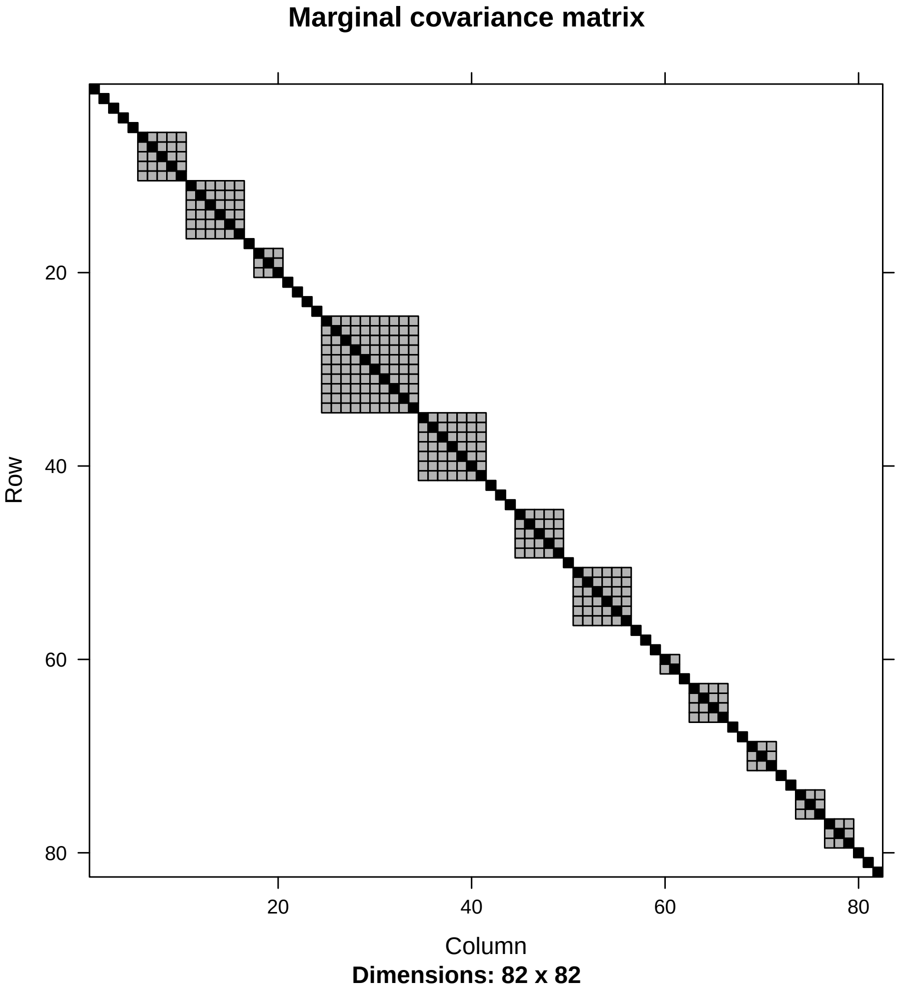


Figure 12.3: Marginal covariance structure of $N = 82$ climate simulations, which have been driven by $m = 37$ GCMs. The block-diagonal dependency structure has been induced by the random-effects design matrix $\boldsymbol{Z}$. The diagonal black entries is the variance of the runs $\widehat{\sigma}^2 + \widehat{\tau}^2 = 0.10 + 1.07 = 1.17$ where the off-diagonal grey entries denote the covariance $\widehat{\tau}^2 = 1.07$.

12.3 Multilevel Regression Model

We now generalise the classical definition of a hierarchical (2-stage mixed-effects) model by taking several levels of random effects into account, which we from now on call *multilevel model*. We first need following matrix operations:

Definition 12.7 (Direct sum)**.** The direct sum of two matrices $\boldsymbol{A} \in \mathbb{R}^{n\times m}$ and $\boldsymbol{B} \in \mathbb{R}^{p\times q}$ is defined as

$$\boldsymbol{A} \oplus \boldsymbol{B} = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{B} \end{pmatrix} \in \mathbb{R}^{(n+p)\times(m+q)},$$

with $\boldsymbol{0}$ being block-matrices with 0 entries.

Definition 12.8 (Kronecker product)**.** We define the Kronecker product of two matrices $\boldsymbol{A} \in \mathbb{R}^{n\times m}$ and $\boldsymbol{B} \in \mathbb{R}^{p\times q}$ as

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{pmatrix} a_{11}\boldsymbol{B} & \dots & a_{1m}\boldsymbol{B} \\ \vdots & \ddots & \vdots \\ a_{n1}\boldsymbol{B} & \dots & a_{nm}\boldsymbol{B} \end{pmatrix}.$$

Definition 12.9 (Multilevel Regression)**.** Let $\boldsymbol{Y}$ be a $N \times 1$ response vector. We want to model the N outcomes with a **L-level** random effects regression model

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Z}^{(1)}\boldsymbol{b}^{(1)} + \dots + \boldsymbol{Z}^{(L)}\boldsymbol{b}^{(L)} + \boldsymbol{\epsilon} \tag{12.27}$$

with $\boldsymbol{\beta} \in \mathbb{R}^{p\times 1}$ being the fixed effects and the random effects $\boldsymbol{b}^{(l)} \in \mathbb{R}^{q_l\times 1}$ of length q_l with $l = 1, \dots, L$ levels of random effects. The length of random effect of level l is

$$q_l = d_l \cdot m_l, \quad l = 1, \dots, L$$

with d_l being the **dimension of the distribution** and m_l the **number of clusters** of the level l random effect. This means the elements of $\boldsymbol{b}^{(l)}$ follow a known d_l-variate distribution (with unknown parameters). We further demand mutual independence among the random effects

$$\boldsymbol{b}^{(1)} \perp \dots \perp \boldsymbol{b}^{(L)} \perp \boldsymbol{\epsilon}.$$

We further define the number of total observations within a cluster i with N_i, and N_{ij} the amount of observations within cluster j which is nested within cluster i and so on. Similarly we define $m_{l(i)}$ to be the number of clusters in level l, nested within cluster i (of the lowest level L). $m_{l(ij)}$ would be the amount of level l clusters nested within cluster j in the second-lowest level nested within cluster i in the lowest level, and so on.

Dimension	Description
Random Effects	
L	number of levels of random effects (nesting depth)
I	number of clusters in lowest level L
J_i	number of clusters in level $L-1$ nested within cluster i
K_{ij}	number of clusters in level $L-2$ nested within cluster j and i
$\vdots$	
$m_{l(i)}$	number of clusters in level l nested within cluster i
$m_{l(ij)}$	number of clusters in level $L-2$ nested within cluster j and i
$\vdots$	
$q_l = d_l \cdot m_l$	vector length of vectorized random effect $\boldsymbol{b}^{(l)}$
d_l	dimension of random effect $\boldsymbol{b}_i^{(l)}$
m_l	number of total clusters at level l with
	$m_L = I$
	$m_{L-1} = \sum_{i=1}^{I} J_i$
	$m_{L-2} = \sum_{i=1}^{I} \sum_{j=1}^{J_i} K_{ij}$
	$\vdots$
$q = \sum_{l=1}^{L} q_l$	vector length of all random effects $\boldsymbol{b}$
$m = \sum_{l=1}^{L} m_l$	total amount of clusters of all levels
Fixed Effects	
p	dimension of fixed effect $\boldsymbol{\beta}$
Observations	
$N = \sum_{i=1}^{m} N_i$	total number of observations
$N_i = \sum_{j=1}^{J_i} N_{ij}$	number of observations in cluster i of lowest level L
$N_{ij} = \sum_{k=1}^{K_{ij}} N_{ijk}$	number of observations in cluster j (in $L-1$) within cluster i (in L)
$\vdots$	

Table 12.4: Notation for the indices and dimensions of the multilevel model.

Remark 12.10. The dependency structure of the data $\boldsymbol{Y}$ is defined by the random effects $\boldsymbol{b}^{(l)}$ for $l = 1, \ldots, L$ and its design matrices $\boldsymbol{Z}^{(l)}$.

Remark 12.11 (Index subscription). The notation of the single observation entries in $\boldsymbol{Y} \in \mathbb{R}^{N \times 1}$ depends on the amount of random effect levels L. For example, having no dependency structure ($L = 0$) as for the *linear regression model* we only need one index for each sample point Y_i

$$Y_i = \boldsymbol{x}_i' \boldsymbol{\beta} + \epsilon_i \tag{12.28}$$

where

$$\epsilon_i \overset{iid}{\sim} N(0, \sigma^2), \quad i = 1, \ldots, N.$$

Having one random effect ($L = 1$) as in the *two-stage regression model*, the single datapoints can be written as Y_{ij} with

$$Y_{ij} = \boldsymbol{x}_{ij}' \boldsymbol{\beta} + \boldsymbol{z}_{ij}' \boldsymbol{b}_i + \epsilon_{ij} \tag{12.29}$$

with $N = \sum_{i=1}^{I} N_i$ and

$$\begin{array}{lll} \text{Level 2} & \boldsymbol{b}_i \overset{iid}{\sim} N_d(\boldsymbol{0}, D), & i = 1, \ldots, I \\ \text{Level 1} & \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2), & j = 1, \ldots, N_i. \end{array}$$

In contrast, a 4-level multilevel model can be written as

$$Y_{ijkt} = \boldsymbol{x}_{ijkt}' \boldsymbol{\beta} + \boldsymbol{z}_{ijkt}^{(1)} \boldsymbol{b}_{ijk}^{(1)} + \boldsymbol{z}_{ijkt}^{(2)} \boldsymbol{b}_{ij}^{(2)} + \boldsymbol{z}_{ijkt}^{(3)} \boldsymbol{b}_{i}^{(3)} + \epsilon_{ijkt} \tag{12.30}$$

with $N = \sum_{i,j,k} N_{ijk}$ and

$$\begin{array}{lll} \text{Level 4} & \boldsymbol{b}_i^{(3)} \overset{iid}{\sim} N_{d_3}(\boldsymbol{0}, D_3), & i = 1, \ldots, I \\ \text{Level 3} & \boldsymbol{b}_{ij}^{(2)} \overset{iid}{\sim} N_{d_2}(\boldsymbol{0}, D_2), & j = 1, \ldots, J_i \\ \text{Level 2} & \boldsymbol{b}_{ijk}^{(1)} \overset{iid}{\sim} N_{d_1}(\boldsymbol{0}, D_1), & k = 1, \ldots, K_{ij} \\ \text{Level 1} & \epsilon_{ijkt} \overset{iid}{\sim} N(0, \sigma^2), & t = 1, \ldots, N_{ijk}. \end{array}$$

Remark 12.12. The 2-stage linear mixed effects model from Section 12.2 can be defined with $L = 1$, and so setting $\boldsymbol{b} = \boldsymbol{b}_l$ leading to

$$Y_{ij} = \boldsymbol{x}_{ij}' \boldsymbol{\beta} + \boldsymbol{z}_{ij}' \boldsymbol{b} + \epsilon_{ij}$$

where

$$\epsilon_{ij} \sim N(0, \sigma^2), \quad \boldsymbol{b}_i \sim N_d(\boldsymbol{0}, D)$$

with $\boldsymbol{b} = (\boldsymbol{b}_1', \dots, \boldsymbol{b}_m')'$ and $\boldsymbol{\epsilon} = (\epsilon_{11}, \dots, \epsilon_{mN_m})'$ leading to

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Zb} + \boldsymbol{\epsilon}$$

with

$$\boldsymbol{\epsilon} \sim N_N(\boldsymbol{0}, \sigma^2 I_N), \quad \boldsymbol{b} \sim N_q(\boldsymbol{0}, I_m \otimes D).$$

Remark 12.13. Any L-level random effects model can be written as

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Z}^{(1)}\boldsymbol{b}^{(1)} + \cdots + \boldsymbol{Z}^{(L)}\boldsymbol{b}^{(L)} + \boldsymbol{\epsilon}$$

where

$$\boldsymbol{Z} = (\boldsymbol{Z}^{(1)} \vdots \dots \vdots \boldsymbol{Z}^{(L)}) \quad \text{and} \quad \boldsymbol{b} = \begin{pmatrix} \boldsymbol{b}^{(1)} \\ \vdots \\ \boldsymbol{b}^{(L)} \end{pmatrix}.$$

Example 12.14 (4-Level nested hierarchical model)**.** Consider the hierarchical model

$$Y_{ijkt} = \beta_0 + a_{ijk} + b_{ij} + c_i + \epsilon_{ijkt}.$$

The data structure is shown in Figure 12.4. We have

$$i = 1, \dots, I, \quad j = 1, \dots, J_i, \quad k = 1, \dots, K_{ij}, \quad t = 1, \dots, N_{ijk}$$

where

$$\begin{array}{lc} \text{Level 4 :} & I = 2, \\ \text{Level 3 :} & J_1 = 2, J_2 = 1, \\ \text{Level 2 :} & K_{11} = 2, K_{12} = 3, K_{21} = 4, \\ \text{Level 1 :} & N_{ijk} \equiv 3. \end{array}$$

which defines the design matrix for the random effects when vectorizing the observations $\boldsymbol{Y} \in \mathbb{R}^N$, where $N = 27$ to

$$\boldsymbol{Y} = \boldsymbol{1}\beta_0 + \boldsymbol{Z}^{(1)}\boldsymbol{a} + \boldsymbol{Z}^{(2)}\boldsymbol{b} + \boldsymbol{Z}^{(3)}\boldsymbol{c} + \boldsymbol{\epsilon}.$$

From Definition 12.9 we get

$$\boldsymbol{a} = (a_{ijk})_{ijk} \in \mathbb{R}^{q_1}, \quad \boldsymbol{Z}^{(1)} \in \mathbb{R}^{N \times q_1}, \quad N = 27, \quad q_1 = d_1 \cdot m_1 = 9$$

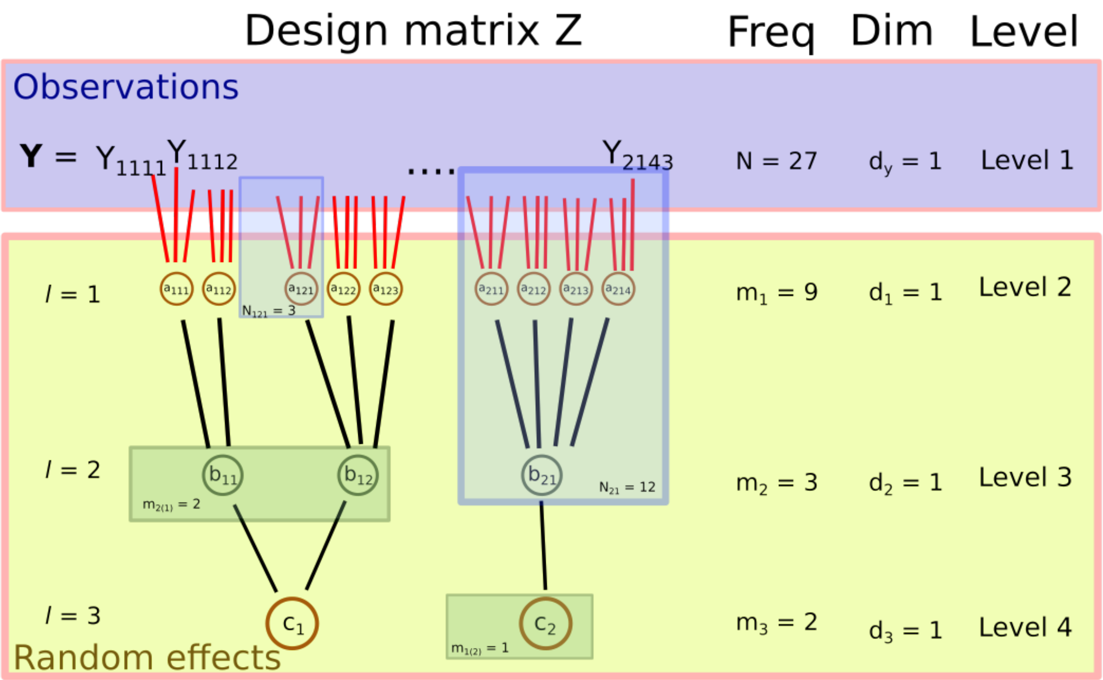


Figure 12.4: The hierarchical structure of a 4-Levels nested regression model having 3 levels of unobserved random effects a_{ijk}, b_{ij}, c_i and one level of observed data $\boldsymbol{Y}$. The blue boxes depict the number of observed data (e.g. N_{ij}) at different levels, whereas the green boxes depict the number of clusters (e.g. $m_{l(i)}$) at different levels.

and because every random effect a_{ijk} is repeated $N_{ijk} \equiv 3$ times, we get

$$\boldsymbol{Z}^{(1)} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \\ 0 & 1 & & 0 \\ 0 & 1 & & 0 \\ 0 & 1 & & 0 \\ & & \ddots & \\ 0 & \dots & 0 & 1 \\ 0 & \dots & 0 & 1 \\ 0 & \dots & 0 & 1 \end{pmatrix} = \begin{pmatrix} \mathbf{1}_3 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{1}_3 \end{pmatrix} = \mathbf{1}_3 \oplus \dots \oplus \mathbf{1}_3 = \boldsymbol{I}_9 \otimes \mathbf{1}_3$$

and similarly for the other random effects:

$$\begin{aligned} \boldsymbol{b} = (b_{ij})_{ij} \in \mathbb{R}^{q_2}, \quad \boldsymbol{Z}^{(2)} \in \mathbb{R}^{N \times q_2}, \quad N = 27, \quad q_2 = d_2 \cdot m_2 = 3 \\ \boldsymbol{c} = (c_i)_i \in \mathbb{R}^{q_3}, \quad \boldsymbol{Z}^{(3)} \in \mathbb{R}^{N \times q_3}, \quad N = 27, \quad q_3 = d_3 \cdot m_3 = 2 \end{aligned}$$

and

$$\begin{aligned} \boldsymbol{Z}^{(2)} &= \mathbf{1}_{N11} \oplus \mathbf{1}_{N12} \oplus \mathbf{1}_{N11} \in \mathbb{R}^{27 \times 3} \\ \boldsymbol{Z}^{(3)} &= \mathbf{1}_{N1} \oplus \mathbf{1}_{N2} \in \mathbb{R}^{27 \times 2}. \end{aligned}$$

We can write the model with a single random effect

$$\boldsymbol{Y} = \mathbf{1}\beta_0 + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon},$$

where

$$\boldsymbol{\gamma} = \begin{pmatrix} \boldsymbol{a} \\ \boldsymbol{b} \\ \boldsymbol{c} \end{pmatrix} \in \mathbb{R}^{q_1+q_2+q_3}, \quad \boldsymbol{Z} = (\boldsymbol{Z}^{(1)} \vdots \boldsymbol{Z}^{(2)} \vdots \boldsymbol{Z}^{(3)}) \in \mathbb{R}^{N \times (q_1+q_2+q_3)}$$

Remark 12.15 (Linearity of expected value). If the underlying distributions are linear with respect to the expected value (as it is the case for Gaussian distribution), any nested hierarchical (linear) regression model can be written as the model (12.27) with each random effect having zero mean. For example, consider

$$\begin{aligned} \boldsymbol{Y} | \boldsymbol{\gamma}^{(1)}, \boldsymbol{\gamma}^{(2)} &\sim N(\boldsymbol{X}^{(0)}\boldsymbol{\beta}^{(0)} + \boldsymbol{W}^{(1)}\boldsymbol{\gamma}^{(1)}, \boldsymbol{\Sigma}_y) \\ \boldsymbol{\gamma}^{(1)} | \boldsymbol{\gamma}^{(2)} &\sim N(\boldsymbol{\beta}^{(1)} + \boldsymbol{W}^{(2)}\boldsymbol{\gamma}^{(2)}, \boldsymbol{\Sigma}_1) \\ \boldsymbol{\gamma}^{(2)} &\sim N(\boldsymbol{\beta}^{(2)}, \boldsymbol{\Sigma}_2) \end{aligned}$$

which can be re-parameterized with $\boldsymbol{b}^{(1)} = \boldsymbol{\gamma}^{(1)} - \mathrm{E}(\boldsymbol{\gamma}^{(1)} | \boldsymbol{\gamma}^{(2)}) = \boldsymbol{\gamma}^{(1)} - (\boldsymbol{\beta}^{(1)} + \boldsymbol{W}^{(2)}\boldsymbol{\gamma}^{(2)})$ and $\boldsymbol{b}^{(2)} = \boldsymbol{\gamma}^{(2)} - \mathrm{E}(\boldsymbol{\gamma}^{(2)}) = \boldsymbol{\gamma}^{(2)} - \boldsymbol{\beta}^{(2)}$ leading to

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}^{(1)}\boldsymbol{b}^{(1)} + \boldsymbol{Z}^{(2)}\boldsymbol{b}^{(2)} + \boldsymbol{\epsilon}$$

with

$$\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_y), \quad \boldsymbol{b}^{(1)} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_1), \quad \boldsymbol{b}^{(2)} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_2), \quad \text{where } \boldsymbol{b}^{(1)} \perp \boldsymbol{b}^{(2)} \perp \boldsymbol{\epsilon}$$

and

$$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}^{(0)} \\ \boldsymbol{\beta}^{(1)} \\ \boldsymbol{\beta}^{(2)} \end{pmatrix}, \quad \boldsymbol{X} = (\boldsymbol{X}^{(0)} \vdots \boldsymbol{Z}^{(1)} \vdots \boldsymbol{Z}^{(2)}), \quad \boldsymbol{Z}^{(1)} = \boldsymbol{W}^{(1)}, \quad \boldsymbol{Z}^{(2)} = \boldsymbol{W}^{(1)}\boldsymbol{W}^{(2)},$$

because

$$\begin{aligned}
\boldsymbol{X}^{(0)}\boldsymbol{\beta}^{(0)} + \boldsymbol{W}^{(1)}\boldsymbol{\gamma}^{(1)} &= \boldsymbol{X}^{(0)}\boldsymbol{\beta}^{(0)} + \boldsymbol{W}^{(1)}(\boldsymbol{b}^{(1)} + \boldsymbol{\beta}^{(1)} + \boldsymbol{W}^{(2)}\boldsymbol{\gamma}^{(2)}) \\
&= \boldsymbol{X}^{(0)}\boldsymbol{\beta}^{(0)} + \boldsymbol{W}^{(1)}(\boldsymbol{b}^{(1)} + \boldsymbol{\beta}^{(1)} + \boldsymbol{W}^{(2)}(\boldsymbol{b}^{(2)} + \boldsymbol{\beta}^{(2)})) \\
&= \boldsymbol{X}^{(0)}\boldsymbol{\beta}^{(0)} + \boldsymbol{W}^{(1)}\boldsymbol{\beta}^{(1)} + \boldsymbol{W}^{(1)}\boldsymbol{W}^{(2)}\boldsymbol{\beta}^{(2)} + \boldsymbol{W}^{(1)}\boldsymbol{b}^{(1)} + \boldsymbol{W}^{(1)}\boldsymbol{W}^{(2)}\boldsymbol{b}^{(2)} \\
&= (\boldsymbol{X}^{(0)} \vdots \boldsymbol{W}^{(1)} \vdots \boldsymbol{W}^{(1)}\boldsymbol{W}^{(2)}) \begin{pmatrix} \boldsymbol{\beta}^{(0)} \\ \boldsymbol{\beta}^{(1)} \\ \boldsymbol{\beta}^{(2)} \end{pmatrix} + \boldsymbol{W}^{(1)}\boldsymbol{b}^{(1)} + \boldsymbol{W}^{(1)}\boldsymbol{W}^{(2)}\boldsymbol{b}^{(2)} \\
&= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}^{(1)}\boldsymbol{b}^{(1)} + \boldsymbol{Z}^{(2)}\boldsymbol{b}^{(2)}.
\end{aligned}$$

13 Skewed Regression

We now move further from normally distributed data to the class of skew-normal distributed data. The aim is to derive a multi-level regression model for this general class of distributions.

We denote $\phi_n(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\Phi_n(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ the probability density function (PDF) and the cumulative distribution function of the $N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution evaluated at $\boldsymbol{x} \in \mathbb{R}^n$. If $\boldsymbol{\mu} = \boldsymbol{0}$ and $\boldsymbol{\Sigma} = \boldsymbol{I}_n$ we write $\phi_n(\boldsymbol{x})$ and $\Phi_n(\boldsymbol{x})$.

13.1 Skew-Normal Distribution

We consider the skew-normal distribution as first defined in Azzalini 1985 and further extended in Azzalini and Dalla Valle 1996 and in Arellano-Valle, Bolfarine and Lachos 2005. We also introduce a reparameterization introduced by Arellano-Valle and Azzalini 2008 which results in a faster and more robust estimation as well as better interpretability.

Definition 13.1. The random variable Y follows the univariate skew-normal distribution with location parameter ξ, scale parameter ω^2 and skewness parameter λ when its PDF is

$$f_Y(y) = 2\phi_1(y|\xi, \omega^2)\Phi_1\left(\lambda\frac{y-\xi}{\omega}\right). \tag{13.1}$$

We write $Y \sim SN(\xi, \omega^2, \lambda)$ or $Y \sim SN(\lambda)$ if $\xi = 0$ and $\omega^2 = 1$.

Remark 13.2. If $\lambda = 0$, then $Y \sim N(\xi, \omega^2)$.

Proposition 13.3. *Let $Y \sim SN(\lambda)$, then*

$$Y = \delta|U| + (1-\delta^2)^{1/2}V, \tag{13.2}$$

where $\delta = \lambda/\sqrt{1+\lambda^2}$, $U \sim N(0,1)$, i.e. $|U| \sim HN(0,1)$, is independent of $V \sim N(0,1)$.

This mixture of Normal and Half-Normal distribution can be used to derive the first two moments.

Proposition 13.4. *Let $Y = \xi + \omega W$ where $W \sim SN(\lambda)$. Then*

$$Y \sim SN(\xi, \omega, \lambda),$$

and

$$\mathrm{E}(Y) = \mu = \xi + \sqrt{\frac{2}{\pi}}\omega\delta,$$
$$\mathrm{Var}(Y) = \mathrm{E}\left((Y-\mu)^2\right) = \sigma^2 = \omega^2(1 - \frac{2}{\pi}\delta^2),$$

with $\delta = \lambda/\sqrt{1+\lambda^2}$.

We derive those moments for the multivariate case below. The third standartised moment, the skewness is given by

$$\mathrm{Skew}(Y) = \mathrm{E}\left(\frac{(Y-\mu)^3}{\sigma^3}\right) = \gamma_1 = \frac{4-\pi}{2}\frac{(\delta\sqrt{2/\pi})^3}{(1-\delta^2 2/\pi)^{3/2}},$$

which is derived with the moment generating function (see Azzalini 1985). It is important to note that the skewness of the skew-normal distribution is limited to approximately $\gamma_1 \in (-.995, .995)$.

Definition 13.5. An n-dimensional random vector $\boldsymbol{Y}$ follows a skew-normal distribution with location vector $\boldsymbol{\xi} \in \mathbb{R}^n$, (positive definite) dispersion matrix $\boldsymbol{\Omega} \in \mathbb{R}^{n\times n}$ and skewness vector $\boldsymbol{\lambda} \in \mathbb{R}^n$ when its PDF is given by

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = 2\phi_n(\boldsymbol{y}|\boldsymbol{\xi},\boldsymbol{\Omega})\Phi_1(\boldsymbol{\lambda}'\boldsymbol{\Omega}^{-1/2}(\boldsymbol{y}-\boldsymbol{\xi})) \tag{13.3}$$

with $\boldsymbol{\Omega} = \boldsymbol{\Omega}^{1/2}\boldsymbol{\Omega}^{1/2}$. We write $\boldsymbol{Y} \sim SN_n(\boldsymbol{\xi},\boldsymbol{\Omega},\boldsymbol{\lambda})$.

Proposition 13.6. *Let $\boldsymbol{W} \sim SN_n(\boldsymbol{\lambda})$, then*

$$\boldsymbol{W} = \boldsymbol{\delta}|X_0| + (\boldsymbol{I}_n - \boldsymbol{\delta}\boldsymbol{\delta}')^{1/2}\boldsymbol{X}_1, \tag{13.4}$$

where $\boldsymbol{\delta} = \boldsymbol{\lambda}/\sqrt{1+\boldsymbol{\lambda}'\boldsymbol{\lambda}}$, $X_0 \sim N(0,1)$, is independent of $\boldsymbol{X}_1 \sim N_n(\boldsymbol{0},\boldsymbol{I}_n)$.

Proof. See Appendix A. □

Corollary 13.7. *Let $\boldsymbol{Y} = \boldsymbol{\xi} + \boldsymbol{\Omega}^{1/2}\boldsymbol{W}$, where $\boldsymbol{W} \sim SN_n(\boldsymbol{\lambda})$. Then*

$$\boldsymbol{Y} \sim SN_n(\boldsymbol{\xi},\boldsymbol{\Omega},\boldsymbol{\lambda}), \tag{13.5}$$

and

$$\mathrm{E}(\boldsymbol{Y}) = \boldsymbol{\mu} = \boldsymbol{\xi} + \sqrt{\frac{2}{\pi}}\boldsymbol{\Omega}^{1/2}\boldsymbol{\delta}, \tag{13.6}$$

$$\mathrm{Var}(\boldsymbol{Y}) = \boldsymbol{\Sigma} = \boldsymbol{\Omega} - \frac{2}{\pi}\boldsymbol{\Omega}^{1/2}\boldsymbol{\delta}\boldsymbol{\delta}'\boldsymbol{\Omega}^{1/2}. \tag{13.7}$$

with $\boldsymbol{\delta} = \boldsymbol{\lambda}/\sqrt{1+\boldsymbol{\lambda}'\boldsymbol{\lambda}}$.

Proof. Because $\boldsymbol{W} = \boldsymbol{\delta}|X_0| + (\boldsymbol{I}_n - \boldsymbol{\delta}\boldsymbol{\delta}')^{1/2}\boldsymbol{X}_1$, with $|X_0| \sim HN(0,1)$ it follows that $\mathrm{E}(|X_0|) = \sqrt{2/\pi}$ and $\mathrm{Var}(|X_0|) = 1 - 2/\pi$. We therefore obtain

$$\mathrm{E}(\boldsymbol{Y}) = \mathrm{E}(\boldsymbol{\xi} + \boldsymbol{\Omega}^{1/2}\boldsymbol{W}) = \mathrm{E}(\boldsymbol{\xi} + \boldsymbol{\Omega}^{1/2}\boldsymbol{\delta}|X_0| + \boldsymbol{\Omega}^{1/2}(\boldsymbol{I}_n - \boldsymbol{\delta}\boldsymbol{\delta}')^{1/2}\boldsymbol{X}_1,) = \boldsymbol{\xi} + \boldsymbol{\Omega}^{1/2}\boldsymbol{\delta}\sqrt{\frac{2}{\pi}}$$

and

$$\begin{aligned}\mathrm{Var}(\boldsymbol{Y}) &= \mathrm{Var}(\boldsymbol{\xi} + \boldsymbol{\Omega}^{1/2}\boldsymbol{\delta}|X_0| + \boldsymbol{\Omega}^{1/2}(\boldsymbol{I}_n - \boldsymbol{\delta}\boldsymbol{\delta}')^{1/2}\boldsymbol{X}_1,) \\ &= \boldsymbol{\Omega}^{1/2}\boldsymbol{\delta}(1 - 2/\pi)\boldsymbol{\delta}'\boldsymbol{\Omega}^{1/2} + \boldsymbol{\Omega}^{1/2}(\boldsymbol{I} - \boldsymbol{\delta}\boldsymbol{\delta}')\boldsymbol{\Omega}^{1/2} \\ &= \boldsymbol{\Omega} - \frac{2}{\pi}\boldsymbol{\Omega}^{1/2}\boldsymbol{\delta}\boldsymbol{\delta}'\boldsymbol{\Omega}^{1/2}.\end{aligned}$$

□

Remark 13.8. The expectation vector $\boldsymbol{\mu} = \mathrm{E}(\boldsymbol{Y})$ equals the location parameter vector $\boldsymbol{\xi}$ only if $\boldsymbol{\lambda} = \boldsymbol{0}$, i.e if $\boldsymbol{Y}$ is normally distributed. The same holds for the covariance matrix $\boldsymbol{\Sigma} = \mathrm{Var}(\boldsymbol{Y})$ which equals the dispersion matrix $\boldsymbol{\Omega}$ if $\boldsymbol{\lambda} = \boldsymbol{0}$.

Centred parameterization We can reparameterize the skew-normal RV $\boldsymbol{Y} = (Y_1, \dots, Y_n)'$ defined by its direct parameters (DPs) $(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\lambda})$ to its centred parameters (CPs) $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}_1)$ with $\boldsymbol{\mu} = \mathrm{E}(\boldsymbol{Y})$, $\boldsymbol{\Sigma} = \mathrm{Var}(\boldsymbol{Y})$ and $\boldsymbol{\gamma}_1 = (\gamma_{1,1}, \dots \gamma_{1,n})$ being the vector with Pearsons indices of skewness

$$\gamma_{1,i} = \frac{\mathrm{E}\left((Y_i - \mathrm{E}(Y_i))^3\right)}{\mathrm{Var}(Y_i)^{3/2}}$$

This is meaningful for two reasons: First, studies like Arellano-Valle and Azzalini 2008 show that the profiled DP likelihood function (as a function of the skewness $\boldsymbol{\lambda}$) has a problematic non-quadratic shape and a stationary point at $\boldsymbol{\lambda} = \boldsymbol{0}$. Also the estimated distributions of the maximum likelihood estimate (MLE) can be bi-modal. Second, it is difficult to directly interpret the location and dispersion parameters of $\boldsymbol{Y}$, it is more interesting to interpret the mean and variance of $\boldsymbol{Y}$.

To perform the reparameterization we first introduce the *normalised* variable

$$\boldsymbol{Z} = \boldsymbol{\omega}^{-1}(\boldsymbol{Y} - \boldsymbol{\xi}) \sim SN(\boldsymbol{0}, \bar{\boldsymbol{\Omega}}, \boldsymbol{\lambda}),$$

where $\boldsymbol{\omega}$ is a matrix consisting of the diagonal entries from $\boldsymbol{\Omega}$ and $\bar{\boldsymbol{\Omega}}$ being the normalised dispersion matrix

$$\bar{\boldsymbol{\Omega}} = \boldsymbol{\omega}^{-1}\boldsymbol{\Omega}\boldsymbol{\omega}^{-1}.$$

We also define $\bar{\boldsymbol{\delta}} = (1+\boldsymbol{\lambda}'\bar{\boldsymbol{\Omega}}\boldsymbol{\lambda})^{-1/2}\bar{\boldsymbol{\Omega}}\boldsymbol{\lambda}$ so that

$$\mathrm{E}(\boldsymbol{Z}) = \sqrt{\frac{2}{\pi}}\bar{\boldsymbol{\delta}} = \boldsymbol{\mu}_z, \quad \mathrm{Var}(\boldsymbol{Z}) = \bar{\boldsymbol{\Omega}} - \boldsymbol{\mu}_z\boldsymbol{\mu}_z' = \bar{\boldsymbol{\Omega}} - \frac{2}{\pi}\bar{\boldsymbol{\delta}}\bar{\boldsymbol{\delta}}' = \boldsymbol{\Sigma}_z.$$

The centred parameters are now given by $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}_1)$ where

$$\boldsymbol{\mu} = \mathrm{E}(\boldsymbol{Y}) = \boldsymbol{\xi} + \boldsymbol{\omega}\boldsymbol{\mu}_z, \quad \boldsymbol{\Sigma} = \mathrm{Var}(\boldsymbol{Y}) = \boldsymbol{\Omega} - \boldsymbol{\omega}\boldsymbol{\mu}_z\boldsymbol{\mu}_z'\boldsymbol{\omega} = \boldsymbol{\omega}\boldsymbol{\Sigma}_z\boldsymbol{\omega}$$

and by the skewness parameter where each component of $\boldsymbol{\gamma}_1 = (\gamma_{1,1}, \ldots \gamma_{1,n})'$ is

$$\gamma_{1,i} = \frac{4-\pi}{2}\frac{\mu_{i,z}^3}{(1-\mu_{i,z}^2)^{3/2}},$$

with $\boldsymbol{\mu}_z = (\mu_{1,z}, \ldots, \mu_{n,z})$.

Now for a given choice of admissible CP parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}_1)$ which corresponds to some point $(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\lambda})$ in the DP space, the log-likelihood function for the random sample $\boldsymbol{y} = (y_1, \ldots, y_n)'$ is given by evaluating the DP likelihood function at the corresponding point in the DP space

$$l_{CP}\left((\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}_1)\,; \boldsymbol{y}\right) = l_{DP}\left((\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\lambda}); \boldsymbol{y}\right)$$

For a given CP value $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}_1)$ the corresponding DP value is obtained as follows:

1. Calculate $\boldsymbol{\mu}_z = (\mu_{1,z}, \ldots, \mu_{n,z})$ with

   $$\mu_{i,z} = \frac{c_i}{\sqrt{1+c_i^2}}, \quad c_i = \left(\frac{2\gamma_{i,1}}{4-\pi}\right)^{1/3}.$$

   for each component of $\boldsymbol{\gamma}_1 = (\gamma_{1,1}, \ldots, \gamma_{n,1})'$.

2. Get $\bar{\boldsymbol{\delta}} = \sqrt{\pi/2}\boldsymbol{\mu}_z$.

3. Get the diagonal components $\sigma_{1,z}^2, \ldots, \sigma_{n,z}^2$ of $\boldsymbol{\Sigma}_Z$ with

   $$\sigma_{i,z} = \sqrt{1-\mu_{i,z}}.$$

4. Get the square roots σ_i of the diagonal of the given $\boldsymbol{\Sigma} = \mathrm{Var}(\boldsymbol{Y})$.

5. Get the location vector $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)'$ with the given expectation vector $\boldsymbol{\mu} = \mathrm{E}(\boldsymbol{Y})$

   $$\xi_i = \mu_i - \sigma_i\sigma_{i,z}^{-1}\mu_{i,z}.$$

6. Get the dispersion matrix $\boldsymbol{\Omega} \in \mathbb{R}^{n \times n}$ with

$$\omega_i = \sigma_i \sigma_{i,z}^{-1}, \quad \boldsymbol{\Omega} = \boldsymbol{\Sigma} + \boldsymbol{\omega} \boldsymbol{\mu}_z \boldsymbol{\mu}_z' \boldsymbol{\omega}'$$

   with $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)'$.

7. Get the skewness parameter vector $\boldsymbol{\lambda}$ with

$$\boldsymbol{\lambda} = \frac{1}{\sqrt{1 - \bar{\boldsymbol{\delta}}' \bar{\boldsymbol{\Omega}}^{-1} \bar{\boldsymbol{\delta}}}} \bar{\boldsymbol{\Omega}}^{-1} \bar{\boldsymbol{\delta}}, \quad \bar{\boldsymbol{\Omega}} = \boldsymbol{\omega}^{-1} \boldsymbol{\Omega} \boldsymbol{\omega}^{-1}.$$

This yields the DP parameters $(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\lambda})$ which can be plugged in to the likelihood to be maximised.

13.2 Skew-Normal Mixed Models

We discuss the skew-normal mixed models as defined in Arellano-Valle, Bolfarine and Lachos 2005 and Lin and J. C. Lee 2008.

Definition 13.9. The skew-normal linear mixed-effects model (SN-LMM) for response vector $\boldsymbol{Y}_i \in \mathbb{R}^{N_i}$ is defined as

$$\boldsymbol{Y}_i = \boldsymbol{X}_i \boldsymbol{\beta} + \boldsymbol{Z}_i \boldsymbol{b}_i + \boldsymbol{\epsilon}_i \tag{13.8}$$

$$\boldsymbol{b}_i \overset{iid}{\sim} SN_d(\boldsymbol{0}, \boldsymbol{\Omega}, \boldsymbol{\lambda}), \quad \boldsymbol{\epsilon}_i \overset{iid}{\sim} N_{N_i}(\boldsymbol{0}, \boldsymbol{\Sigma}_i), \tag{13.9}$$

with dimensions as shown in Table 12.2.

Remark 13.10. This leads to the hierarchical model

$$\begin{aligned} \boldsymbol{Y}_i | \boldsymbol{b}_i &\overset{ind}{\sim} N_{N_i}(\boldsymbol{X}_i \boldsymbol{\beta} + \boldsymbol{Z}_i \boldsymbol{b}_i, \boldsymbol{\Sigma}_i) \\ \boldsymbol{b}_i &\overset{iid}{\sim} SN_d(\boldsymbol{0}, \boldsymbol{\Omega}, \boldsymbol{\lambda}). \end{aligned}$$

Remark 13.11. For the correct interpretation of the model parameters it is important to consider expectation of the random vector

$$\mathrm{E}(\boldsymbol{Y}_i) = \boldsymbol{X}_i \boldsymbol{\beta} + \mathrm{E}(\boldsymbol{Z}_i \boldsymbol{b}_i) + \mathrm{E}(\boldsymbol{\epsilon}_i) = \boldsymbol{X}_i \boldsymbol{\beta} + \sqrt{\frac{2}{\pi}} \boldsymbol{Z}_i \boldsymbol{\Omega}^{1/2} \boldsymbol{\delta} \tag{13.10}$$

with $\boldsymbol{\delta} = \boldsymbol{\lambda}(1 + \boldsymbol{\lambda}' \boldsymbol{\lambda})^{-1/2}$ which follows from Corollary 13.7.

To obtain the marginal distribution of such mixed models we need the following Lemma.

Lemma 13.12. *For any* $\boldsymbol{Y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *and* $\boldsymbol{X} \sim N_d(\boldsymbol{\eta}, \boldsymbol{\Omega})$ *the mixture is*

$$\begin{aligned}\phi_p(\boldsymbol{y}|\boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{x}, \boldsymbol{\Sigma})\phi_d(\boldsymbol{x}|\boldsymbol{\eta}, \boldsymbol{\Omega}) = \phi_p(\boldsymbol{y}|\boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{\eta}, \boldsymbol{\Sigma} + \boldsymbol{A}\boldsymbol{\Omega}\boldsymbol{A}') \\ \times \phi_d(\boldsymbol{x}|\boldsymbol{\eta} + \boldsymbol{\Lambda}\boldsymbol{A}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\mu} - \boldsymbol{A}\boldsymbol{\eta}), \boldsymbol{\Lambda})\end{aligned}$$

where

$$\boldsymbol{\Lambda} = (\boldsymbol{\Omega}^{-1} + \boldsymbol{A}'\boldsymbol{\Sigma}^{-1}\boldsymbol{A})^{-1}.$$

Proposition 13.13 (Woodbury Identity)**.** *For any non-singular matrices* $\boldsymbol{A}$ *and* $\boldsymbol{B}$ *the Woodbury matrix identity is*

$$(\boldsymbol{A} + \boldsymbol{U}\boldsymbol{B}\boldsymbol{V})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{U}(\boldsymbol{B}^{-1} + \boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{U})^{-1}\boldsymbol{V}\boldsymbol{A}^{-1}.$$

Theorem 13.14. *For the SN-LMM model*

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i \tag{13.11}$$

$$\boldsymbol{b}_i \overset{iid}{\sim} SN_d(\boldsymbol{0}, \boldsymbol{\Omega}, \boldsymbol{\lambda}), \quad \boldsymbol{\epsilon}_i \overset{iid}{\sim} N_{N_i}(\boldsymbol{0}, \boldsymbol{\Sigma}_i) \tag{13.12}$$

with normally distributed Level-1 residuals, the marginal distribution for each $i = 1, \ldots, m$ *is*

$$f_{\boldsymbol{Y}_i}(\boldsymbol{y}_i|\boldsymbol{\theta}) = 2\phi_{N_i}(\boldsymbol{y}_i|\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\psi}_i)\Phi_1\left(\bar{\boldsymbol{\lambda}}\boldsymbol{\psi}_i^{-1/2}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})\right). \tag{13.13}$$

With

$$\bar{\boldsymbol{\lambda}} = \frac{\boldsymbol{\psi}_i^{-1/2}\boldsymbol{Z}_i\boldsymbol{\Omega}^{1/2}\boldsymbol{\lambda}}{\sqrt{1 + \boldsymbol{\lambda}'\boldsymbol{\Omega}^{-1/2}\boldsymbol{\Lambda}_i\boldsymbol{\Omega}^{-1/2}\boldsymbol{\lambda}}}$$

and

$$\boldsymbol{\psi}_i = \boldsymbol{\Sigma}_i + \boldsymbol{Z}_i\boldsymbol{\Omega}\boldsymbol{Z}_i', \quad \boldsymbol{\Lambda}_i = (\boldsymbol{\Omega}^{-1} + \boldsymbol{Z}_i'\boldsymbol{\psi}_i^{-1}\boldsymbol{Z}_i)^{-1}.$$

Proof. The marginal distribution of the observed data $\boldsymbol{y}_i \in \mathbb{R}^{N_i}$ depending on the parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\alpha}_1', \ldots, \boldsymbol{\alpha}_m', \boldsymbol{\eta}', \boldsymbol{\lambda}')'$ (writing $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}_i(\boldsymbol{\alpha}_i), \boldsymbol{\Omega} = \boldsymbol{\Omega}(\boldsymbol{\eta})$) is obtained by integrating out the random effects:

$$\begin{aligned}f(\boldsymbol{y}_i|\boldsymbol{\theta}) &= \int_{\mathbb{R}^d} f(\boldsymbol{y}_i|\boldsymbol{b}_i, \boldsymbol{\theta})f(\boldsymbol{b}_i|\boldsymbol{\theta})d\boldsymbol{b}_i \\ &= \int_{\mathbb{R}^d} \phi_{N_i}(\boldsymbol{y}_i|\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i, \boldsymbol{\Sigma}_i)2\phi_d(\boldsymbol{b}_i|\boldsymbol{0}, \boldsymbol{\Omega})\Phi_1(\boldsymbol{\lambda}'\boldsymbol{\Omega}^{-1/2}\boldsymbol{b}_i)d\boldsymbol{b}_i\end{aligned}$$

because of Lemma 13.12 (here abbreviated with L 13.12) we obtain

$$2\phi_{N_i}(\boldsymbol{y}_i|\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_i + \boldsymbol{Z}_i\boldsymbol{\Omega}\boldsymbol{Z}_i')\int_{\mathbb{R}^d} \phi_d(\boldsymbol{b}_i|\boldsymbol{\Lambda}_i\boldsymbol{Z}_i'\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}), \boldsymbol{\Lambda}_i)\Phi_1(\boldsymbol{\lambda}'\boldsymbol{\Omega}^{-1/2}\boldsymbol{b}_i)d\boldsymbol{b}_i.$$

because of the Woodbury identity (WI) 13.13 we obtain

$$\begin{aligned}
\boldsymbol{\Lambda}_i \boldsymbol{Z}_i' \boldsymbol{\Sigma}_i^{-1} &= (\boldsymbol{\Omega}^{-1} + \boldsymbol{Z}_i' \boldsymbol{\Sigma}_i^{-1} \boldsymbol{Z}_i)^{-1} \boldsymbol{Z}_i' \boldsymbol{\Sigma}_i^{-1} \\
&\overset{\text{WI}}{=} \left(\boldsymbol{\Omega} - \boldsymbol{\Omega} \boldsymbol{Z}_i' \left(\boldsymbol{\Sigma}_i + \boldsymbol{Z}_i \boldsymbol{\Omega} \boldsymbol{Z}_i' \right)^{-1} \boldsymbol{Z}_i \boldsymbol{\Omega} \right) \boldsymbol{Z}_i' \boldsymbol{\Sigma}_i^{-1} \\
&= \boldsymbol{\Omega} \boldsymbol{Z}_i' \left(\boldsymbol{I}_{N_i} - \left(\boldsymbol{\Sigma}_i + \boldsymbol{Z}_i \boldsymbol{\Omega} \boldsymbol{Z}_i' \right)^{-1} \boldsymbol{Z}_i \boldsymbol{\Omega} \boldsymbol{Z}_i' \right) \boldsymbol{\Sigma}_i^{-1} \\
&= \boldsymbol{\Omega} \boldsymbol{Z}_i' \left(\boldsymbol{\Sigma}_i + \boldsymbol{Z}_i \boldsymbol{\Omega} \boldsymbol{Z}_i' \right)^{-1} \left(\left(\boldsymbol{\Sigma}_i + \boldsymbol{Z}_i \boldsymbol{\Omega} \boldsymbol{Z}_i' \right) - \boldsymbol{Z}_i \boldsymbol{\Omega} \boldsymbol{Z}_i' \right) \boldsymbol{\Sigma}_i^{-1} \\
&= \boldsymbol{\Omega} \boldsymbol{Z}_i' \left(\boldsymbol{\Sigma}_i + \boldsymbol{Z}_i \boldsymbol{\Omega} \boldsymbol{Z}_i' \right)^{-1} \\
&= \boldsymbol{\Omega} \boldsymbol{Z}_i' \boldsymbol{\psi}_i^{-1}
\end{aligned}$$

with $\boldsymbol{\psi}_i = \boldsymbol{\Sigma}_i + \boldsymbol{Z}_i \boldsymbol{\Omega} \boldsymbol{Z}_i'$ and therefore

$$\begin{aligned}
f(\boldsymbol{y}_i|\boldsymbol{\theta}) &= 2\phi_{N_i}(\boldsymbol{y}_i|\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\psi}_i) \int_{\mathbb{R}^d} \phi_d(\boldsymbol{b}_i|\boldsymbol{\Omega}\boldsymbol{Z}_i'\boldsymbol{\psi}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}), \boldsymbol{\Lambda}_i)\Phi_1(\boldsymbol{\lambda}'\boldsymbol{\Omega}^{-1/2}\boldsymbol{b}_i)d\boldsymbol{b}_i \\
&= 2\phi_{N_i}(\boldsymbol{y}_i|\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\psi}_i) \int_{\mathbb{R}^d} \phi_d(\boldsymbol{b}_i|\boldsymbol{\Omega}\boldsymbol{Z}_i'\boldsymbol{\psi}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}), \boldsymbol{\Lambda}_i) \int_{-\infty}^{0} \phi_1(\alpha| - \boldsymbol{\lambda}'\boldsymbol{\Omega}^{-1/2}\boldsymbol{b}_i, 1)d\alpha d\boldsymbol{b}_i \\
&\overset{L13.12}{=} 2\phi_{N_i}(\boldsymbol{y}_i|\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\psi}_i) \int_{-\infty}^{0} \underbrace{\int_{\mathbb{R}^d} \phi_d(\boldsymbol{b}_i|\boldsymbol{\mu}_{b_i}, \boldsymbol{\Sigma}_{b_i})d\boldsymbol{b}_i}_{=1} \\
&\quad \times \phi_1(\alpha| - \boldsymbol{\lambda}'\boldsymbol{\Omega}^{-1/2}\boldsymbol{\Omega}\boldsymbol{Z}_i'\boldsymbol{\psi}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}), 1 + \boldsymbol{\lambda}'\boldsymbol{\Omega}^{-1/2}\boldsymbol{\Lambda}_i\boldsymbol{\Omega}^{-1/2}\boldsymbol{\lambda})d\alpha \\
&= 2\phi_{N_i}(\boldsymbol{y}_i|\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\psi}_i)\Phi_1(\boldsymbol{\lambda}'\boldsymbol{\Omega}^{1/2}\boldsymbol{Z}_i'\boldsymbol{\psi}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})|0, 1 + \boldsymbol{\lambda}'\boldsymbol{\Omega}^{-1/2}\boldsymbol{\Lambda}_i\boldsymbol{\Omega}^{-1/2}\boldsymbol{\lambda}) \\
&= 2\phi_{N_i}(\boldsymbol{y}_i|\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\psi}_i)\Phi_1\left(\bar{\boldsymbol{\lambda}}'\boldsymbol{\psi}_i^{-1/2}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})\right)
\end{aligned}$$

with

$$\bar{\boldsymbol{\lambda}} = \frac{\boldsymbol{\psi}_i^{-1/2}\boldsymbol{Z}_i\boldsymbol{\Omega}^{1/2}\boldsymbol{\lambda}}{\sqrt{1 + \boldsymbol{\lambda}'\boldsymbol{\Omega}^{-1/2}\boldsymbol{\Lambda}_i\boldsymbol{\Omega}^{-1/2}\boldsymbol{\lambda}}}$$

□

Remark 13.15. The marginal log-likelihood is then

$$\begin{aligned}
l(\boldsymbol{\theta};\boldsymbol{y}) \propto &-\frac{1}{2}\sum_{i=1}^{n}\log|\boldsymbol{\psi}_i| - \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})'\boldsymbol{\psi}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}) \\
&+ \sum_{i=1}^{n}\log\Phi_1\left(\bar{\boldsymbol{\lambda}}\boldsymbol{\psi}_i^{-1/2}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})\right),
\end{aligned}$$

with $\bar{\boldsymbol{\lambda}}$ as defined above and $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\alpha}_1', \ldots, \boldsymbol{\alpha}_m', \boldsymbol{\eta}', \boldsymbol{\lambda}')$ writing $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}_i(\boldsymbol{\alpha}_i)$, and $\boldsymbol{\Omega} = \boldsymbol{\Omega}(\boldsymbol{\eta})'$.

No explicit solution is available for the maximisation problem, so the likelihood function has to be maximised numerically.

13.3 Skew-Normal Multilevel Models

In this chapter we extend the Skew-Normal Mixed Models from Arellano-Valle, Bolfarine and Lachos 2005 and Lin and J. C. Lee 2008 to a more general multilevel setup, as is the case when modelling several nested random effects. As a result we obtain the marginal likelihood function, which in contrast to the mixed-effects models, cannot be written as a skew-normal distribution of the form defined by Azzalini and Dalla Valle 1996. This generalises the multilevel setup as defined for the normal case in Section 12.3.

Definition 13.16. Consider the multilevel skew-normal (SN) regression model for $\boldsymbol{Y} \in \mathbb{R}^N$ with L random effect levels

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}^{(1)}\boldsymbol{\gamma}^{(1)} + \dots + \boldsymbol{Z}^{(L)}\boldsymbol{\gamma}^{(L)} + \boldsymbol{\epsilon}, \tag{13.14}$$

with

$$\boldsymbol{\epsilon} \sim N_N(\mathbf{0}, \boldsymbol{\Sigma}).$$

Each level $l = 1, \dots, L$ of the random effect $\boldsymbol{\gamma}^{(l)}$ consists of m_l clusters each having dimension d_l, leading to a total length of the vector $\boldsymbol{\gamma}^{(l)}$ of $q_l = d_l \cdot m_l$. We write $\boldsymbol{\gamma}^{(l)} = \left(\boldsymbol{\gamma}_1^{(l)\prime}, \dots, \boldsymbol{\gamma}_{m_l}^{(l)\prime}\right)'$. The single clusters are denoted as $\boldsymbol{\gamma}_i^{(l)}$ and follow a skew-normal distribution

$$\boldsymbol{\gamma}_i^{(l)} \overset{iid}{\sim} SN_{d_l}(\mathbf{0}, \Omega_l, \boldsymbol{\lambda}_l).$$

This model can also be written as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon},$$

where

$$\boldsymbol{\gamma} = (\boldsymbol{\gamma}^{(1)\prime}, \dots, \boldsymbol{\gamma}^{(L)\prime})', \quad \boldsymbol{Z} = (\boldsymbol{Z}^{(1)} \vdots \dots \vdots \boldsymbol{Z}^{(L)}),$$

and the total length of $\boldsymbol{\gamma}$ is $q = \sum_{l=1}^{L} q_l$ and the overall amount of all clusters over all levels is $m = \sum_{l=1}^{L} m_l$.

Corollary 13.17. *The marginal density of model (13.14) from Definition 13.17 is*

$$f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\theta}) = 2^m \phi_N(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{\psi}) \Phi_m \left(\boldsymbol{\Gamma}\boldsymbol{\Omega}^{1/2}\boldsymbol{Z}\boldsymbol{\psi}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})|\mathbf{0}, \boldsymbol{I}_m + \boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1/2}\boldsymbol{\Lambda}\boldsymbol{\Omega}^{-1/2}\boldsymbol{\Gamma}'\right). \tag{13.15}$$

with

$$\begin{aligned} \boldsymbol{\psi} = \boldsymbol{\Sigma} + \boldsymbol{Z}\boldsymbol{\Omega}\boldsymbol{Z}', \quad \boldsymbol{\Lambda} &= (\boldsymbol{\Omega}^{-1} + \boldsymbol{Z}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Z})^{-1} \\ \boldsymbol{\Omega} = \boldsymbol{\Omega}_1 \oplus \dots \oplus \boldsymbol{\Omega}_L, \quad \boldsymbol{\Omega}_l &= \boldsymbol{I}_{m_l} \otimes \Omega_l \\ \boldsymbol{\Gamma} = \boldsymbol{\Gamma}_1 \oplus \dots \oplus \boldsymbol{\Gamma}_L, \quad \boldsymbol{\Gamma}_l &= \boldsymbol{I}_{m_l} \otimes \boldsymbol{\lambda}_l' \end{aligned}$$

where $\boldsymbol{I}_{m_l}$ denotes the $m_l \times m_l$ identity matrix, $\otimes$ is the Kronecker product (see Definition 12.8) and $\oplus$ the direct sum (see Definition 12.7).

Proof. We call the Woodbury identity as defined in Proposition 13.13 as WI. The marginal density follows from

$$\begin{aligned} f(\boldsymbol{y}) &= \int_{\mathbb{R}^q} f(\boldsymbol{y}|\gamma) f(\gamma) d\gamma \\ &= \int_{\mathbb{R}^{q_1}} \cdots \int_{\mathbb{R}^{q_L}} f(\boldsymbol{y}|\gamma) f(\boldsymbol{\gamma}^{(1)}) \cdots f(\boldsymbol{\gamma}^{(L)}) d\boldsymbol{\gamma}^{(1)} \dots d\boldsymbol{\gamma}^{(L)} \end{aligned}$$

where

$$\begin{aligned} f(\boldsymbol{\gamma}^{(l)}) &= \prod_{i=1}^{m_l} f(\boldsymbol{\gamma}_i^{(l)}) \\ &= \prod_{i=1}^{m_l} 2\phi_{d_l}(\boldsymbol{\gamma}_i^{(l)}|\boldsymbol{0}, \Omega_l)\Phi_1(\boldsymbol{\lambda}_l'\Omega_l^{-1/2}\boldsymbol{\gamma}_i^{(l)}) \\ &= 2^{m_l}\phi_{d_l \cdot m_l}(\boldsymbol{\gamma}^{(l)}|\boldsymbol{0}, \boldsymbol{I}_{m_l} \otimes \Omega_l)\Phi_{m_l}((\boldsymbol{I}_{m_l} \otimes (\boldsymbol{\lambda}_l'\Omega_l^{-1/2}))\boldsymbol{\gamma}^{(l)}) \\ &= 2^{m_l}\phi_{q_l}(\boldsymbol{\gamma}^{(l)}|\boldsymbol{0}, \boldsymbol{I}_{m_l} \otimes \Omega_l)\Phi_{m_l}((\boldsymbol{I}_{m_l} \otimes \boldsymbol{\lambda}_l')(\boldsymbol{I}_{m_l} \otimes \Omega_l^{-1/2})\boldsymbol{\gamma}^{(l)}) \\ &= 2^{m_l}\phi_{q_l}(\boldsymbol{\gamma}^{(l)}|\boldsymbol{0}, \boldsymbol{\Omega}_l)\Phi_{m_l}(\boldsymbol{\Gamma}_l\boldsymbol{\Omega}_l^{-1/2}\boldsymbol{\gamma}^{(l)}) \end{aligned}$$

with the vector length of the random effect level l of $q_l = d_l \cdot m_l$ and by setting $\boldsymbol{\Gamma}_l = \boldsymbol{I}_{m_l} \otimes \boldsymbol{\lambda}_l'\Omega_l^{-1/2}$ and $\boldsymbol{\Omega}_l = \boldsymbol{I}_{m_l} \otimes \Omega_l$ as a consequence of the independence of the random effects and because of the following fact:

$$\begin{aligned} \boldsymbol{I}_{m_l} \otimes \Omega_l^{-1/2} &= \left(\boldsymbol{I}_{m_l} \otimes \Omega_l^{-1/2}\right)^{1/2}\left(\boldsymbol{I}_{m_l} \otimes \Omega_l^{-1/2}\right)^{1/2} = \left(\left(\boldsymbol{I}_{m_l} \otimes \Omega_l^{1/2}\right)\left(\boldsymbol{I}_{m_l} \otimes \Omega_l^{1/2}\right)\right)^{-1/2} \\ &= \left((\boldsymbol{I}_{m_l}\boldsymbol{I}_{m_l}) \otimes \left(\Omega_l^{1/2}\Omega_l^{1/2}\right)\right)^{-1/2} = (\boldsymbol{I}_{m_l} \otimes \Omega_l)^{-1/2} = \boldsymbol{\Omega}_l^{-1/2}. \end{aligned}$$

Additionally we have

$$\begin{aligned} \Phi_p(\boldsymbol{Ax})\Phi_q(\boldsymbol{By}) &= \int_{-\infty}^{Ax} \phi_{p_1}(\boldsymbol{\alpha}_1|\boldsymbol{0}, \boldsymbol{I}_{p_1})d\boldsymbol{\alpha}_1 \int_{-\infty}^{By} \phi_{p_2}(\boldsymbol{\alpha}_2|\boldsymbol{0}, \boldsymbol{I}_{p_2})d\boldsymbol{\alpha}_2 \\ &= \int_{-\infty}^{0} \phi_{p_1}(\boldsymbol{\alpha}_1| - \boldsymbol{Ax}, \boldsymbol{I}_{p_1})d\boldsymbol{\alpha}_1 \int_{-\infty}^{0} \phi_{p_2}(\boldsymbol{\alpha}_2| - \boldsymbol{By}, \boldsymbol{I}_{p_2})d\boldsymbol{\alpha}_2 \\ &= \int_{-\infty}^{0} \phi_{p_1+p_2}\left(\begin{pmatrix}\boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2\end{pmatrix} | \begin{pmatrix}-\boldsymbol{Ax} \\ -\boldsymbol{By}\end{pmatrix}, \begin{pmatrix}\boldsymbol{I}_{p_1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{p_2}\end{pmatrix}\right) d\boldsymbol{\alpha}_1 d\boldsymbol{\alpha}_2 \\ &= \Phi_{p_1+p_2}\begin{pmatrix}\boldsymbol{Ax} \\ \boldsymbol{By}\end{pmatrix} \end{aligned}$$

$$
\begin{aligned}
&= \Phi_{p_1+p_2}\left(\begin{pmatrix}\boldsymbol{A} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{B}\end{pmatrix}\begin{pmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{pmatrix}\right)\\
&= \Phi_{p_1+p_2}\left((\boldsymbol{A}\oplus\boldsymbol{B})\begin{pmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{pmatrix}\right)
\end{aligned}
$$

and m times the direct sum of the same matrix A yields $A \oplus \cdots \oplus A = I_m \otimes A$.

The common distribution of the $l = 1, \ldots, L$ random effects with total length of $q = \sum_{l=1}^{L} q_l = \sum_{l=1}^{L} d_l m_l$ and total amount of clusters described by the random effects $m = \sum_{l=1}^{L} m_l$ is then

$$
\begin{aligned}
\prod_{l=1}^{L} f(\boldsymbol{\gamma}^{(l)}) &= \prod_{l=1}^{L} 2^{m_l}\phi_{q_l}(\boldsymbol{\gamma}^{(l)}|\boldsymbol{0},\boldsymbol{\Omega}_l)\Phi_{m_l}(\boldsymbol{\Gamma}_l\boldsymbol{\Omega}_l^{-1/2}\boldsymbol{\gamma}^{(l)})\\
&= 2^m\phi_q(\boldsymbol{\gamma}|\boldsymbol{0},\boldsymbol{\Omega}_1\oplus\cdots\oplus\boldsymbol{\Omega}_L)\Phi_m((\boldsymbol{\Gamma}_1\boldsymbol{\Omega}_1^{-1/2}\oplus\cdots\oplus\boldsymbol{\Gamma}_L\boldsymbol{\Omega}_L^{-1/2})\boldsymbol{\gamma})\\
&= 2^m\phi_q(\boldsymbol{\gamma}|\boldsymbol{0},\boldsymbol{\Omega})\Phi_m(\boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1/2}\boldsymbol{\gamma}).
\end{aligned}
$$

Now the marginal distribution becomes

$$
\begin{aligned}
f(\boldsymbol{y}|\boldsymbol{\theta}) &= \int_{\mathbb{R}^q} f(\boldsymbol{y}|\gamma,\boldsymbol{\theta})f(\gamma|\boldsymbol{\theta})d\boldsymbol{\gamma}\\
&= 2^m\int_{\mathbb{R}^q}\phi_N(\boldsymbol{y}|\boldsymbol{X\beta}+\boldsymbol{Z\gamma},\boldsymbol{\Sigma})\phi_q(\gamma|\boldsymbol{0},\boldsymbol{\Omega})\Phi_m(\boldsymbol{\Gamma\Omega}^{-1/2}\gamma)d\boldsymbol{\gamma}\\
&\overset{L13.12}{=} 2^m\phi_N(\boldsymbol{y}|\boldsymbol{X\beta},\boldsymbol{\Sigma}+\boldsymbol{Z\Omega Z}')\int_{\mathbb{R}^q}\phi_q(\gamma|\boldsymbol{\Lambda Z}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{X\beta}),\boldsymbol{\Lambda})\Phi_m(\boldsymbol{\Gamma\Omega}^{-1/2}\gamma)d\boldsymbol{\gamma}\\
&\overset{(*)}{=} 2^m\phi_N(\boldsymbol{y}|\boldsymbol{X\beta},\boldsymbol{\psi})\int_{\mathbb{R}^q}\int_{(\mathbb{R}^-)^m}\phi_q(\gamma|\boldsymbol{\Omega Z}'\boldsymbol{\psi}^{-1}(\boldsymbol{y}-\boldsymbol{X\beta}),\boldsymbol{\Lambda})\phi_m(\boldsymbol{\alpha}|-\boldsymbol{\Gamma\Omega}^{-1/2}\gamma,\boldsymbol{I}_m)d\boldsymbol{\alpha}d\boldsymbol{\gamma}\\
&\overset{L13.12}{=} 2^m\phi_N(\boldsymbol{y}|\boldsymbol{X\beta},\boldsymbol{\psi})\int_{(\mathbb{R}^-)^m}\overbrace{\int_{\mathbb{R}^q}\phi_q(\gamma|\boldsymbol{\mu}_\gamma,\boldsymbol{\Sigma}_\gamma)d\boldsymbol{\gamma}}^{=1}\times\\
&\quad\times\phi_m(\boldsymbol{\alpha}|-\boldsymbol{\Gamma\Omega}^{-1/2}\boldsymbol{\Omega Z}'\boldsymbol{\psi}^{-1}(\boldsymbol{y}-\boldsymbol{X\beta}),\boldsymbol{I}_m+\boldsymbol{\Gamma\Omega}^{-1/2}\boldsymbol{\Lambda\Omega}^{-1/2}\boldsymbol{\Gamma}')d\boldsymbol{\alpha}\\
&= 2^m\phi_N(\boldsymbol{y}|\boldsymbol{X\beta},\boldsymbol{\psi})\Phi_m(\boldsymbol{\Gamma\Omega}^{1/2}\boldsymbol{Z}'\boldsymbol{\psi}^{-1}(\boldsymbol{y}-\boldsymbol{X\beta})|\boldsymbol{0},\boldsymbol{I}_m+\boldsymbol{\Gamma\Omega}^{-1/2}\boldsymbol{\Lambda\Omega}^{-1/2}\boldsymbol{\Gamma}')
\end{aligned}
$$

with $\boldsymbol{\Lambda} = (\boldsymbol{\Omega}^{-1}+\boldsymbol{Z}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Z})^{-1}$ and with $(*)$ $\boldsymbol{\Lambda Z}'\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Omega Z}'\boldsymbol{\psi}^{-1}$ as in proof of Theorem 13.14. □

Definition 13.18. We discuss the multilevel SN regression model for an observed $\boldsymbol{Y} \in \mathbb{R}^N$

$$
\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Z}_\gamma\boldsymbol{\gamma} + \boldsymbol{Z}_b\boldsymbol{b} + \boldsymbol{\epsilon} \tag{13.16}
$$

consisting of both normal random effects $\boldsymbol{b} \in \mathbb{R}^{q_b}$ and a vector $\boldsymbol{\gamma} \in \mathbb{R}^{q_\gamma}$ with skew-normal random effect components where

$$\boldsymbol{b} \sim N_{q_b}(0, \boldsymbol{\Sigma}_b), \quad \boldsymbol{\epsilon} \sim N_N(\boldsymbol{0}, \boldsymbol{\Sigma}). \tag{13.17}$$

For the skew-normal part, we have L levels with

$$\boldsymbol{\gamma} = (\boldsymbol{\gamma}^{(1)'}, \dots, \boldsymbol{\gamma}^{(L)'})', \quad \boldsymbol{Z} = (\boldsymbol{Z}^{(1)} \vdots \dots \vdots \boldsymbol{Z}^{(L)}),$$

and the total length of $\boldsymbol{\gamma}$ being $q_\gamma = \sum_{l=1}^{L} q_l$ and the overall amount of all skew-normal clusters over all levels being $m = \sum_{l=1}^{L} m_l$. Each level $l = 1, \dots, L$ of the random effect $\boldsymbol{\gamma}^{(l)}$ consists of m_l clusters each having dimension d_l, leading to a total length of the vector $\boldsymbol{\gamma}^{(l)}$ of $q_l = d_l \cdot m_l$. We write $\boldsymbol{\gamma}^{(l)} = \left(\boldsymbol{\gamma}_1^{(l)'}, \dots, \boldsymbol{\gamma}_{m_l}^{(l)'}\right)'$. The single clusters are denoted as $\boldsymbol{\gamma}_i^{(l)}$ and follow a skew-normal distribution

$$\boldsymbol{\gamma}_i^{(l)} \overset{iid}{\sim} SN_{d_l}(\boldsymbol{0}, \Omega_l, \boldsymbol{\lambda}_l).$$

Corollary 13.19. *The marginal density of the multilevel SN regression model (13.16) is*

$$f(\boldsymbol{y}|\boldsymbol{\theta}) = 2^m \phi_N(\boldsymbol{y}|\boldsymbol{X\beta}, \boldsymbol{\psi}) \Phi_m(\boldsymbol{\Gamma\Omega}^{1/2} \boldsymbol{Z}_\gamma' \boldsymbol{\psi}^{-1}(\boldsymbol{y} - \boldsymbol{X\beta})|\boldsymbol{0}, \boldsymbol{I}_m + \boldsymbol{\Gamma\Omega}^{-1/2}\boldsymbol{\Lambda}\boldsymbol{\Omega}^{-1/2}\boldsymbol{\Gamma}') \tag{13.18}$$

with

$$\begin{aligned} \boldsymbol{\psi}_b = \boldsymbol{\Sigma} + \boldsymbol{Z}_b \boldsymbol{\Sigma}_b \boldsymbol{Z}_b', &\quad \boldsymbol{\psi} = \boldsymbol{\psi}_b + \boldsymbol{Z}_\gamma \boldsymbol{\Omega} \boldsymbol{Z}_\gamma \\ \boldsymbol{\Omega} = \boldsymbol{\Omega}_1 \oplus \dots \oplus \boldsymbol{\Omega}_L, &\quad \boldsymbol{\Omega}_l = \boldsymbol{I}_{m_l} \otimes \Omega_l \\ \boldsymbol{\Gamma} = \boldsymbol{\Gamma}_1 \oplus \dots \oplus \boldsymbol{\Gamma}_L, &\quad \boldsymbol{\Gamma}_l = \boldsymbol{I}_{m_l} \otimes \boldsymbol{\lambda}_l' \\ \boldsymbol{\Lambda} = (\boldsymbol{\Omega}^{-1} &+ \boldsymbol{Z}_\gamma' \boldsymbol{\psi}_b^{-1} \boldsymbol{Z}_\gamma)^{-1}, \end{aligned}$$

and with parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\alpha}_1', \boldsymbol{\alpha}_2', \boldsymbol{\eta}', \boldsymbol{\lambda}')'$ *writing* $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\alpha}_1)$, $\boldsymbol{\Sigma}_b = \boldsymbol{\Sigma}_b(\boldsymbol{\alpha}_2)$, *and* $\boldsymbol{\Omega} = \boldsymbol{\Omega}(\boldsymbol{\eta})'$. *This marginal density differs from the pure skew-normal marginal density from Corollary 13.17 by the term* $\boldsymbol{\psi}_b^{-1}$ *in* $\boldsymbol{\Lambda}$ *instead of* $\boldsymbol{\Sigma}^{-1}$.

Proof. From proof of Corollary 13.17 we obtain

$$
\begin{aligned}
f(\boldsymbol{y}|\boldsymbol{\theta}) &= \int_{\mathbb{R}^{q_b}} \int_{\mathbb{R}^{q_\gamma}} f(\boldsymbol{y}|\boldsymbol{b},\boldsymbol{\gamma},\boldsymbol{\theta}) f(\boldsymbol{b}|\boldsymbol{\theta}) f(\boldsymbol{\gamma}|\boldsymbol{\theta}) d\boldsymbol{\gamma} d\boldsymbol{b} \\
&= 2^m \int_{\mathbb{R}^{q_b}} \int_{\mathbb{R}^{q_\gamma}} \phi_N(\boldsymbol{y}|\boldsymbol{X\beta} + \boldsymbol{Z}_b\boldsymbol{b} + \boldsymbol{Z}_\gamma\boldsymbol{\gamma}, \boldsymbol{\Sigma}) \phi_{q_b}(\boldsymbol{b}|\boldsymbol{0},\boldsymbol{\Sigma}_b) \phi_{q_\gamma}(\boldsymbol{\gamma}|\boldsymbol{0},\boldsymbol{\Omega}) \Phi_m(\boldsymbol{\Gamma\Omega}^{-1/2}\boldsymbol{\gamma}) d\boldsymbol{\gamma} d\boldsymbol{b} \\
&\overset{L13.12}{=} 2^m \int_{\mathbb{R}^{q_\gamma}} \phi_N(\boldsymbol{y}|\boldsymbol{X\beta} + \boldsymbol{Z}_\gamma\boldsymbol{\gamma}, \boldsymbol{\Sigma} + \boldsymbol{Z}_b\boldsymbol{\Sigma}_b\boldsymbol{Z}_b') \overbrace{\int_{\mathbb{R}^{q_b}} \phi_{q_b}(\boldsymbol{b}|\boldsymbol{\mu}_b,\boldsymbol{\Lambda}_b) d\boldsymbol{b}}^{=1} \\
&\quad \times \phi_{q_\gamma}(\boldsymbol{\gamma}|\boldsymbol{0},\boldsymbol{\Omega}) \Phi_m(\boldsymbol{\Gamma\Omega}^{-1/2}\boldsymbol{\gamma}) d\boldsymbol{\gamma} \\
&= 2^m \int_{\mathbb{R}^{q_\gamma}} \phi_N(\boldsymbol{y}|\boldsymbol{X\beta} + \boldsymbol{Z}_\gamma\boldsymbol{\gamma}, \boldsymbol{\psi}_b) \phi_{q_\gamma}(\boldsymbol{\gamma}|\boldsymbol{0},\boldsymbol{\Omega}) \Phi_m(\boldsymbol{\Gamma\Omega}^{-1/2}\boldsymbol{\gamma}) d\boldsymbol{\gamma} \\
&\overset{L13.12}{=} 2^m \phi_N(\boldsymbol{y}|\boldsymbol{X\beta}, \boldsymbol{\psi}_b + \boldsymbol{Z}_\gamma\boldsymbol{\Omega}\boldsymbol{Z}_\gamma') \int_{\mathbb{R}^{q_\gamma}} \phi_{q_\gamma}(\boldsymbol{\gamma}|\boldsymbol{\Lambda}\boldsymbol{Z}_\gamma'\boldsymbol{\psi}_b^{-1}(\boldsymbol{y} - \boldsymbol{X\beta}), \boldsymbol{\Lambda}) \Phi_m(\boldsymbol{\Gamma\Omega}^{-1/2}\boldsymbol{\gamma}) d\boldsymbol{\gamma} \\
&\overset{(*)}{=} 2^m \phi_N(\boldsymbol{y}|\boldsymbol{X\beta}, \boldsymbol{\psi}) \int_{\mathbb{R}^{q_\gamma}} \int_{(\mathbb{R}^-)^m} \phi_{q_\gamma}(\boldsymbol{\gamma}|\boldsymbol{\Omega}\boldsymbol{Z}_\gamma'\boldsymbol{\psi}^{-1}(\boldsymbol{y} - \boldsymbol{X\beta}), \boldsymbol{\Lambda}) \phi_m(\boldsymbol{\alpha}| - \boldsymbol{\Gamma\Omega}^{-1/2}\boldsymbol{\gamma}, \boldsymbol{I}_m) d\boldsymbol{\alpha} d\boldsymbol{\gamma} \\
&\overset{L13.12}{=} 2^m \phi_N(\boldsymbol{y}|\boldsymbol{X\beta}, \boldsymbol{\psi}) \int_{(\mathbb{R}^-)^m} \overbrace{\int_{\mathbb{R}^{q_\gamma}} \phi_{q_\gamma}(\boldsymbol{\gamma}|\boldsymbol{\mu}_\gamma,\boldsymbol{\Lambda}_\gamma) d\boldsymbol{\gamma}}^{=1} \\
&\quad \times \phi_m(\boldsymbol{\alpha}| - \boldsymbol{\Gamma\Omega}^{-1/2}\boldsymbol{\Omega}\boldsymbol{Z}_\gamma'\boldsymbol{\psi}^{-1}(\boldsymbol{y} - \boldsymbol{X\beta}), \boldsymbol{I}_m + \boldsymbol{\Gamma\Omega}^{-1/2}\boldsymbol{\Lambda\Omega}^{-1/2}\boldsymbol{\Gamma}') d\boldsymbol{\alpha} \\
&= 2^m \phi_N(\boldsymbol{y}|\boldsymbol{X\beta}, \boldsymbol{\psi}) \Phi_m(\boldsymbol{\Gamma\Omega}^{1/2}\boldsymbol{Z}_\gamma'\boldsymbol{\psi}^{-1}(\boldsymbol{y} - \boldsymbol{X\beta})|\boldsymbol{0}, \boldsymbol{I}_m + \boldsymbol{\Gamma\Omega}^{-1/2}\boldsymbol{\Lambda\Omega}^{-1/2}\boldsymbol{\Gamma}'),
\end{aligned}
$$

with $(*)$ $\boldsymbol{\Lambda}\boldsymbol{Z}_\gamma'\boldsymbol{\psi}_b^{-1} = \boldsymbol{\Omega}\boldsymbol{Z}_\gamma'\boldsymbol{\psi}^{-1}$ as in the proof of Theorem 13.14. We further have $\boldsymbol{\psi}_b = \boldsymbol{\Sigma} + \boldsymbol{Z}_b\boldsymbol{\Sigma}_b\boldsymbol{Z}_b'$, $\boldsymbol{\psi} = \boldsymbol{\psi}_b + \boldsymbol{Z}_\gamma\boldsymbol{\Omega}\boldsymbol{Z}_\gamma' = \boldsymbol{\Sigma} + \boldsymbol{Z}_b\boldsymbol{\Sigma}_b\boldsymbol{Z}_b' + \boldsymbol{Z}_\gamma\boldsymbol{\Omega}\boldsymbol{Z}_\gamma'$ and $\boldsymbol{\Lambda} = (\boldsymbol{\Omega}^{-1} + \boldsymbol{Z}_\gamma'\boldsymbol{\psi}_b^{-1}\boldsymbol{Z}_\gamma)^{-1}$ and $\boldsymbol{\Gamma}$ as defined above. □

Remark 13.20. To solve the likelihood function (13.18) of model (13.16) we need to invert two large marginal matrices, namely

$$
\begin{aligned}
\mathrm{Var}(\boldsymbol{Y}|\boldsymbol{\gamma}) &= \boldsymbol{\psi}_b = \boldsymbol{\Sigma} + \boldsymbol{Z}_b\boldsymbol{\Sigma}_b\boldsymbol{Z}_b' \\
\mathrm{Var}(\boldsymbol{Y}) &= \boldsymbol{\psi} = \boldsymbol{\Sigma} + \boldsymbol{Z}_b\boldsymbol{\Sigma}_b\boldsymbol{Z}_b' + \boldsymbol{Z}_\gamma\boldsymbol{\Omega}\boldsymbol{Z}_\gamma',
\end{aligned}
$$

where $\mathrm{Var}(\boldsymbol{Y})$ is the overall marginal variance and $\mathrm{Var}(\boldsymbol{Y}|\boldsymbol{\gamma})$ the marginal variance given the SN random effect $\boldsymbol{\gamma}$. For normal multilevel models or skew-normal mixed models we only need to invert one such matrix.

However, due to the block-diagonal structure of the marginal matrices we can reduce the problem using the Woodbury identity (Proposition 13.13) with

$$
\boldsymbol{\psi}^{-1} = (\boldsymbol{\psi}_b + \boldsymbol{Z}_\gamma\boldsymbol{\Omega}\boldsymbol{Z}_\gamma')^{-1} = \boldsymbol{\psi}_b^{-1} - \boldsymbol{\psi}_b^{-1}\boldsymbol{Z}_\gamma(\boldsymbol{\Omega}^{-1} + \boldsymbol{Z}_\gamma'\boldsymbol{\psi}_b^{-1}\boldsymbol{Z}_\gamma)^{-1}\boldsymbol{Z}_\gamma'\boldsymbol{\psi}_b^{-1}.
$$

In which case we also need to invert two matrices ($\boldsymbol{\Omega}^{-1}$ has to be inverted in both cases, so we do not count that especially), but now the second matrix $\boldsymbol{\Omega}^{-1} + \boldsymbol{Z}'_{\gamma}\boldsymbol{\psi}_{b}^{-1}\boldsymbol{Z}_{\gamma}$ is usually of a much smaller dimension, which speeds up the calculations.

14 Confidence Intervals for Estimators

The reason why we take MLEs is its property of being *consistent* estimates (converging to the true parameter value with increasing sample size) and its property of being *asymptotically efficient* (the standard errors are the smallest possible). We elaborate the latter point and explain the estimation of the standard errors (being the square root of the asymptotic variance) in the following Section as described in Casella and Berger 2002.

Hence we know the asymptotic distribution of the MLE, so we can estimate $1-\alpha$ CIs as a measure of uncertainty. In this work, we present three different ways for estimation:

1. $1-\alpha$ CIs based purely on the estimated MLE variance (standard error) (*Wald procedure*)
2. $1-\alpha$ CIs based on the likelihood ratio test (LRT) statistic
3. $1-\alpha$ CIs based on bootstrap samples

14.1 Wald Procedure Interval

Definition 14.1. For an estimator T_n, suppose that $k_n(T_n - \tau(\theta)) \to N(0, \sigma^2)$ in distribution, with k_n being a sequence of normalising constants. The parameter σ^2 is called the *asymptotic variance*.

Definition 14.2. We call the square root of the asymptotic variance $\sqrt{\sigma^2}$ of the estimator T_n as defined above its *standard error*.

Definition 14.3. A sequence of estimators W_n is *asymptotically efficient* for a parameter $\tau(\theta)$ if $\sqrt{n}[W_n - \tau(\theta)] \to N(0, v(\theta))$ in distribution and

$$v(\theta) = \frac{[\tau'(\theta)]^2}{E_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right)}. \tag{14.1}$$

The asymptotic variance of W_n achieves the Cramer-Rao Lower Bound.

Theorem 14.4 (Asymptotic efficiency of MLEs)**.** *Let $X_1, X_2, \ldots$ be iid $f(x|\theta)$, and $\hat\theta$ denote the MLE of θ, and let τ be a continuous function of θ. Under the regularity conditions on $f(x|\theta)$ and hence, $L(\theta|\boldsymbol{x})$*

$$\sqrt{n}[\tau(\hat\theta) - \tau(\theta)] \to N(0, v(\theta)), \tag{14.2}$$

where $v(\theta)$ is the Cramer-Rao Lower Bound. That is, $\tau(\hat{\theta})$ is a consistent and asymptotically efficient estimator of $\tau(\theta)$.

Remark 14.5. We can now approximate the true variance of the MLE by calculating the Cramer-Rao Lower bound:

$$\mathrm{Var}_\theta(h(\hat{\theta})) \approx \frac{(h'(\theta))^2}{E_\theta\left(\left(\frac{\partial}{\partial\theta}\log L(\theta|\boldsymbol{X})\right)^2\right)} = \frac{(h'(\theta))^2}{-E_\theta\left(\frac{\partial^2}{\partial\theta^2}\log L(\theta|\boldsymbol{X})\right)}$$

where the denominator is termed as the *Fisher Information.* To estimate this variance we can approximate using

$$\widehat{\mathrm{Var}}_\theta(h(\hat{\theta})) \approx \frac{(h'(\theta))^2|_{\theta=\hat{\theta}}}{-\frac{\partial^2}{\partial\theta^2}\log L(\theta|\boldsymbol{X})|_{\theta=\hat{\theta}}}, \tag{14.3}$$

with the denominator called the *expected information number.* As the expected information is a consistent estimator of the Fisher information, $\widehat{\mathrm{Var}}_\theta(h(\hat{\theta}))$ is a consistent estimator for $\mathrm{Var}_\theta(h(\hat{\theta}))$.

In this work, we denote the standard error of a function $h(\hat{\theta})$ of the MLE of θ with $SE(h(\hat{\theta})) = \sqrt{\widehat{\mathrm{Var}}_\theta(h(\hat{\theta}))}$ as the square root of the estimated approximated variance of the MLE.

From Slutsky's theorem the normalised distribution of the function $h(\hat{\theta})$ converges to the standard normal distribution

$$\frac{h(\hat{\theta}) - h(\theta)}{SE(h(\hat{\theta}))} \to N(0,1),$$

which yields the approximate confidence interval

$$h(\hat{\theta})) - z_{1-\alpha/2}SE(h(\hat{\theta})) \leq h(\theta)) \leq h(\hat{\theta})) + z_{1-\alpha/2}SE(h(\hat{\theta})), \tag{14.4}$$

with $z_\alpha = \Phi^{-1}(\alpha)$ being the standard normal α-quantile.

14.2 Likelihood-Ratio Interval

A very useful method for complicated models such as those derived in the chapters above, is the *likelihood ratio* method to construct tests, which is:

$$\lambda(\boldsymbol{x}) = \frac{\sup\limits_{\Theta_0} L(\theta|\boldsymbol{x})}{\sup\limits_{\Theta} L(\theta|\boldsymbol{x})}. \tag{14.5}$$

This is a convenient statistic, as the two suprema of $L(\theta|\boldsymbol{x})$ over the sets Θ_0 and Θ can be calculated numerically.

To define a level α test, a constant must be chosen so that

$$\sup_{\theta \in \Theta_0} P_\theta(\lambda(\boldsymbol{X}) \leq c) \leq \alpha. \tag{14.6}$$

Theorem 14.6 (Asymptotic distribution of the LRT)**.** *For testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$, suppose $X_1, \ldots, X_n$ are iid $f(x|\theta)$, $\hat{\theta}$ is the MLE of θ, and $f(x|\theta)$ satisfies certain regularity conditions. Then under H_0, as $n \to \infty$,*

$$-2 \log \lambda(\boldsymbol{X}) \to \chi^2_1 \text{ in distribution,}$$

where χ^2_1 is a χ^2 random variable with 1 degree of freedom.

Proof. We expand $\log L(\theta|\boldsymbol{x}) = l(\theta|\boldsymbol{x})$ in a Taylor-series around the MLE $\hat{\theta}$, giving

$$l(\theta|\boldsymbol{x}) = l(\hat{\theta}|\boldsymbol{x}) + l'(\hat{\theta}|\boldsymbol{x})(\theta - \hat{\theta}) + l''(\hat{\theta}|\boldsymbol{x})\frac{(\theta - \hat{\theta})^2}{2!} + \cdots . \tag{14.7}$$

Substitute the expansion for $l(\theta_0|\boldsymbol{x})$ in $-2\log\lambda(\boldsymbol{x}) = -2l(\theta_0|\boldsymbol{x}) + 2l(\hat{\theta}|\boldsymbol{x})$, and get

$$-2 \log \lambda(\boldsymbol{x}) \approx -l''(\hat{\theta}|\boldsymbol{x})(\theta_0 - \hat{\theta})^2,$$

as $l'(\hat{\theta}|\boldsymbol{x}) = 0$. Since $-l''(\hat{\theta}|\boldsymbol{x})$ is the observed information $\hat{I}_n(\hat{\theta})$ and $\frac{1}{n}\hat{I}_n(\hat{\theta}) \to I(\theta_0)$ it follows that $-2\log\lambda(\boldsymbol{X}) \to \chi^2_1$. □

We can now invert the LRT statistic to obtain an approximate $1-\alpha$ confidence interval:

$$\left\{\theta : -2\log\left(\frac{L(\theta|\boldsymbol{x})}{L(\hat{\theta}|\boldsymbol{x})}\right) \leq \chi^2_{1,1-\alpha}\right\} \tag{14.8}$$

14.3 Bootstrap Interval

A very distinct approach, which does not require any normal theory, is *Bootstrapping* (Efron and Tibshirani 1994). The basic idea behind this approach is to learn about the distribution of a statistic by *resampling* the observed sample. The logic behind this idea is that as the observed sample represents the population in one way, a magnitude of samples should give us information about the characteristics of the population. Bootstrapping helps us learn about the sample characteristics by resampling (with replacement) and use this information to infer to the population. This way we get a measure of uncertainty for a statistic of interest.

For a sample $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ and an estimate $\hat{\theta}(x_1, x_2, \ldots, x_n) = \hat{\theta}$ we draw resamples with replacement and obtain $\boldsymbol{x}^* = (x_1^*, x_2^*, \ldots, x_n^*)$. This *nonparametric bootstrap* is performed B times and we can calculate

$$\mathrm{Var}^*(\hat{\theta}) = \frac{1}{B-1} \sum_{i=1}^{B} (\hat{\theta}_i^* - \overline{\hat{\theta}_i^*})^2, \tag{14.9}$$

where $\hat{\theta}_i^*$ is the estimator $\hat{\theta}(\boldsymbol{x}_i^*)$ obtained from the ith resample $\boldsymbol{x}_i^*$, $1 \leq i \leq B$ and obtain the overall average $\overline{\hat{\theta}_i^*} = \frac{1}{B} \sum_{i=1}^{B} \hat{\theta}_i^*$. This way we can approximate the standard error of any statistic without distributional assumptions.

Another type of bootstrapping is the *parametric bootstrap*, which is based on the estimated functional form for the population distribution function. For a sample $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ from a distribution with PDF $f(x|\theta)$ where θ can be a vector of parameters, we can estimate θ with its MLE $\hat{\theta}$, plug it into the distribution and draw random samples

$$X_1^*, X_2^*, \ldots X_n^* \sim f(x|\hat{\theta}).$$

We repeat this B times and estimate the variance of $\hat{\theta}$ with (14.9).

In case of clustered data, we perform a *block bootstrap*. We then sample (with replacement) entire blocks of dependent data. If the blocking structure is nested, this resampling is performed recursively.

The $(1 - \alpha)$ confidence interval for the estimate $\hat{\theta}$ is then calculated based on the bootstrap sample

$$\left\{\theta : \hat{\theta}^*_{(\alpha/2)} \leq \theta \leq \hat{\theta}^*_{(1-\alpha/2)}\right\},$$

with $\hat{\theta}^*_{(\alpha/2)}$ being the $\alpha/2$-quantile of the sample of bootstrapped coefficients $(\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_B^*)$.

15 Case Study: Regional Climate Change in Europe

In this Chapter the aim is to estimate the expected temperature climate change signal and its variance in three European hot-spot regions. In addition we quantify following components of uncertainty: internal natural variability (*initial-conditions uncertainty*), *model similarity* (based on pre-defined classes of models) and *structural uncertainty* of General Circulation Models (GCMs) (variation of GCMs within the same class of simulations). The uncertainties can be added together to obtain the overall uncertainty. The analysis is based on yearly temperature time-series which further allows to quantify year-to-year fluctuations, which are estimated as well, and therefore accounting natural sources of uncertainty in more detail. In addition the skewed nature of the data is accounted for by implementing a skew-normal distribution.

15.1 Data

We are interested in the projected seasonal temperature climate change signal (1971-2000 to 2071-2099) for summer and winter of 77 Coupled Model Intercomparison Project Phase 5 (CMIP5) climate simulations (see Section 2.3) over the Alpine Region (AL), the Iberian Peninsula (IP) and the Scandinavian Region (SC) (as defined in J. H. Christensen and O. B. Christensen 2007). The regions are shown in Figure 15.1. Every simulation in each of those regions is represented as a time series of a possible climate development as shown in Figure 15.2.

Each of those 77 climate simulations is a realisation of a GCM (Section 2.3) run with certain initial conditions. As discussed in Section 3.1, those GCMs again share certain components and/or have been developed by the same research groups. Therefore GCMs can also be classified to similar simulations (Knutti, Masson and Gettelman 2013).

This can be described in a hierarchical setup with 4 levels:

Level 1: The individual time series with a year-to-year temperature evolution of 129 years from 1971-2099. This describes the **natural climate variability**. We extract the last $N_{ijk} = 30$ years of climate anomaly (see next section).

subset: *nat*, **index**: $t = 1, \dots, N_{ijk}$

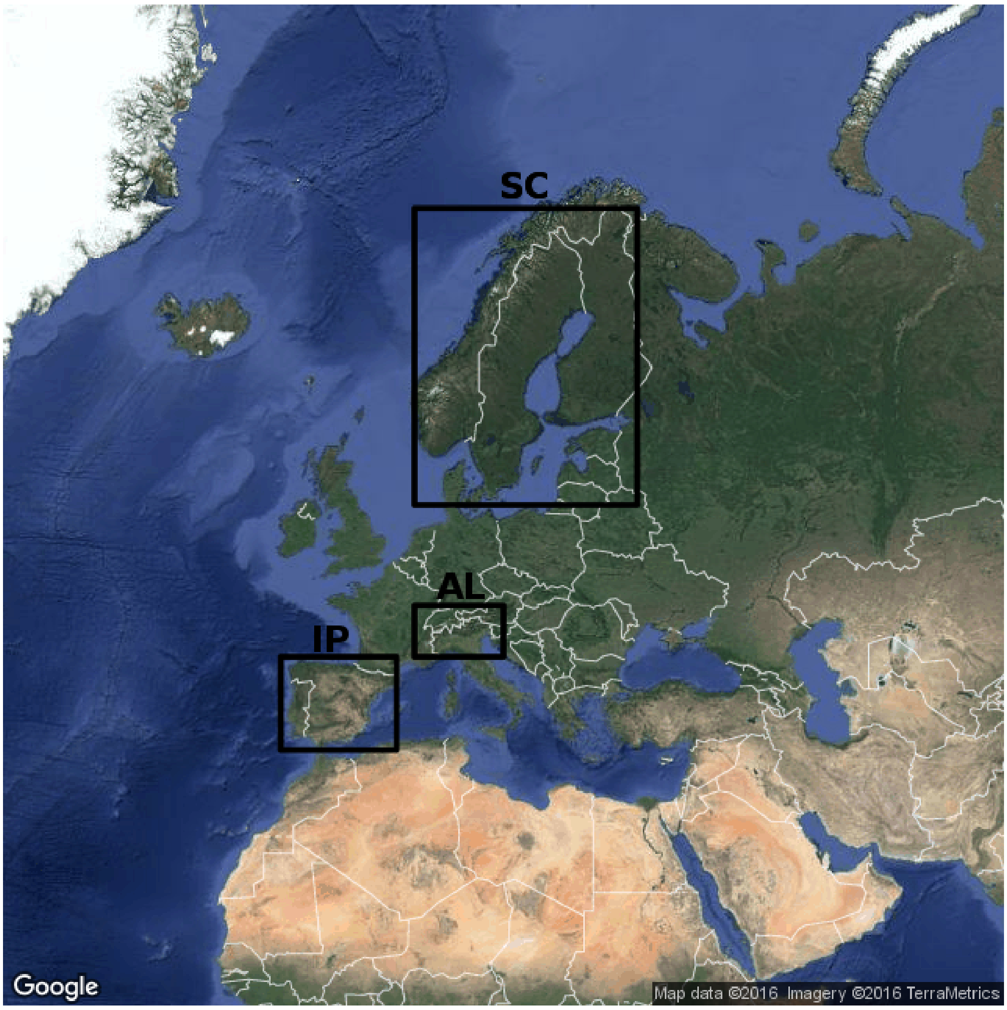


Figure 15.1: Study regions, Alpine Region (AL), Iberian Peninsula (IP) and Scandinavian Region (SC).

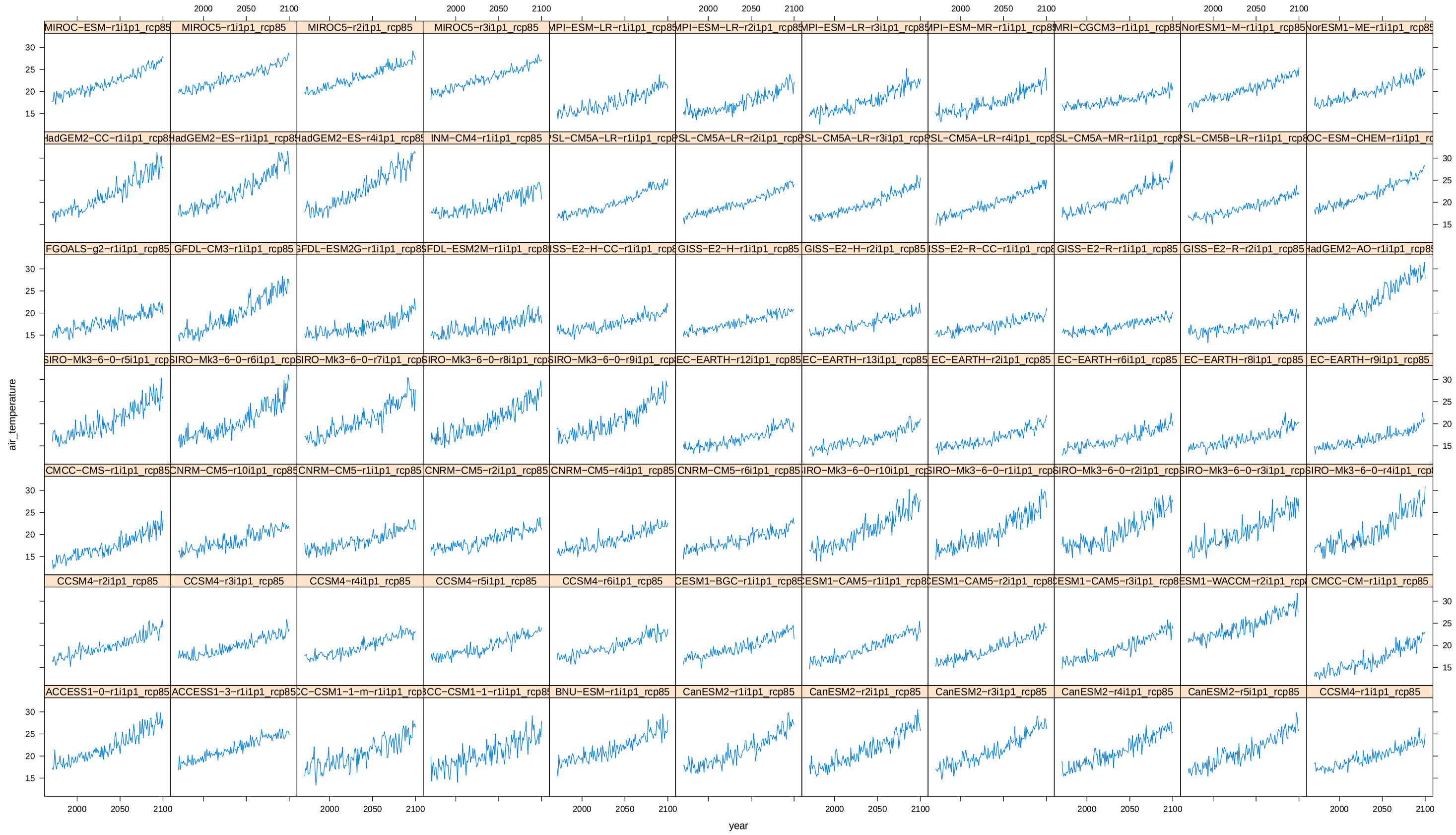


Figure 15.2: 77 complete time-series from 1971 to 2099 for temperature in the Alpine Region (AL) in summer (JJA) obtained from the CMIP5 ensemble.

Level 2: The $m_1 = 77$ time-series partly stem from the same GCM (being the same piece of computer code) but the simulations have been started with different climate conditions. Representing the **initial condition uncertainty**.

subset: *run*, **index**: $k = 1, \ldots, K_{ij}$

Level 3: The $m_2 = 38$ GCMs can again be assigned to a certain class of models. Within the same model similarity class, the GCM computer codes are partly shared among the GCMs. This represents parts of the **structural model uncertainty**.

subset: *gcm*, **index**: $j = 1, \ldots, J_i$

Level 4: The $m_3 = 15$ classes of different models as defined by Knutti, Masson and Gettelman 2013. The variation of the model similarity classes can also be interpreted as a part of **structural model uncertainty**. Time-series across different model similarity classes are considered statistically independent.

subset: *sim*, **index**: $i = 1, \ldots, I$

This hierarchical structure induces dependencies which would, if ignored, lead to an overestimation of estimator precision (see Section 12.1). Also, this hierarchy structure gives information about different types of variability in the dataset (also called uncertainties, see Section 3.1), which we would like to quantify as well. The hierarchical structure is schematically displayed in Figure 15.3. As we have now $N_{ijk} = 30$ observed years for each of the $m_1 = 77$ simulations, this yields the total amount of $77 \cdot 30 = 2310$ data points for each of the 3 regions (IP, AL, SC) for each of the two seasons (summer and winter).

Figure 15.4 depicts the individual climate change signals of all climate models grouped by their GCM and model similarity (here omitting the natural year-to-year variability). The clustering indicating a dependency structure is clearly visible.

Figure 15.4 also shows two potential problems for the analysis: Firstly, one particular GCM (*GFDL-CM3*) has a very strong climate change signal, whereas other *GFDL* simulations are on the cooler side. This outlier has to be investigated more closely in the analysis. Secondly, the internal variability of the *CSIRO-Mk3* GCM is higher than that of other models (like *EC-EARTH*). This can cause problems for the homoscedasticity assumption in linear regression.

From the CMIP5 simulations we further excluded the *FIO* model which does not seem trustworthy (Collins, Knutti et al. 2013).

Data Preprocessing

We consider a total of 77 global climate simulations which have been preprocessed from the binary Network Common Data Form (NetCDF) format to a dataframe of individual

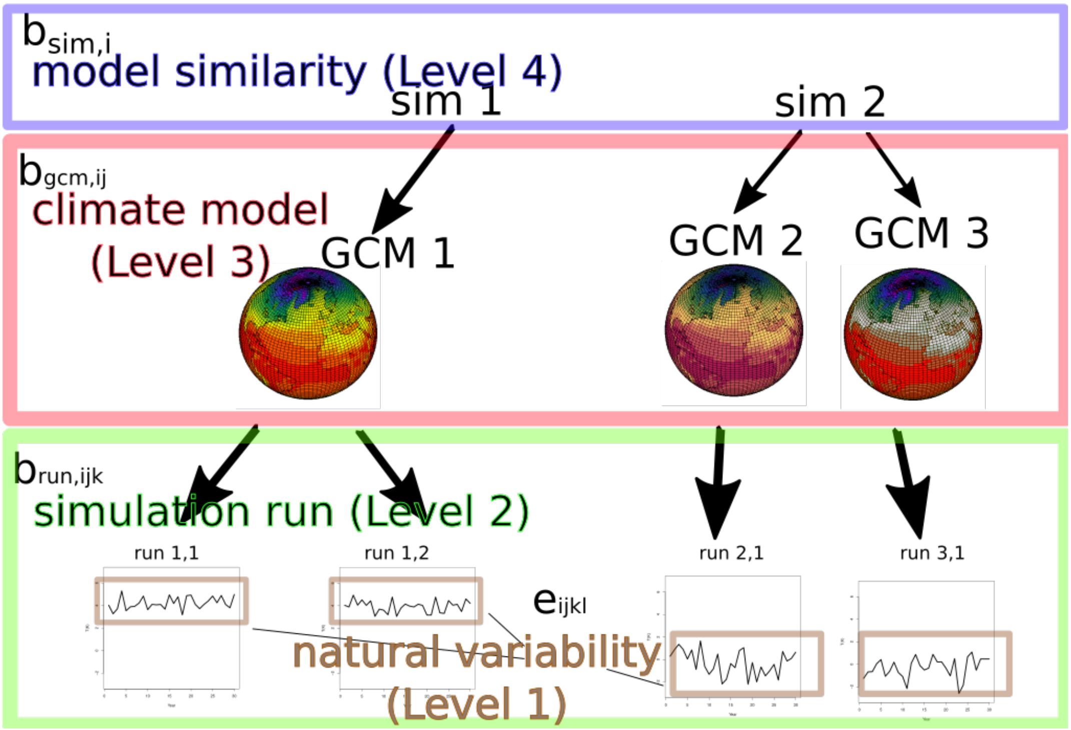


Figure 15.3: Random effects of the hierarchical regression model with 4 levels defining the major sources of dependencies: climate model similarity by sharing parts of codes ($b_{sim,i}$), different global climate models (GCMs, $b_{gcm,ij}$), multiple runs of the same GCM (termed *simulation*, $b_{run,ijk}$) and the natural variability being ϵ_{ijkl}.

time series with our *wux* package (Part II) for easy handling in R. But in order to fit the statistical model we need some further detailed preprocessing. In this study we are mainly interested in the climate change signals projected by each climate simulation, but we are not interested in the specific functional shape of the time series trend. We therefore transfer each of the 129 year trends (1971-2099, Figure 15.2) to 30 years (2070-2099) temperature anomalies. These anomalies are the deviations from the average climate state in the period 1971-2000 (Figure 15.5). This preprocessing is described in the following section.

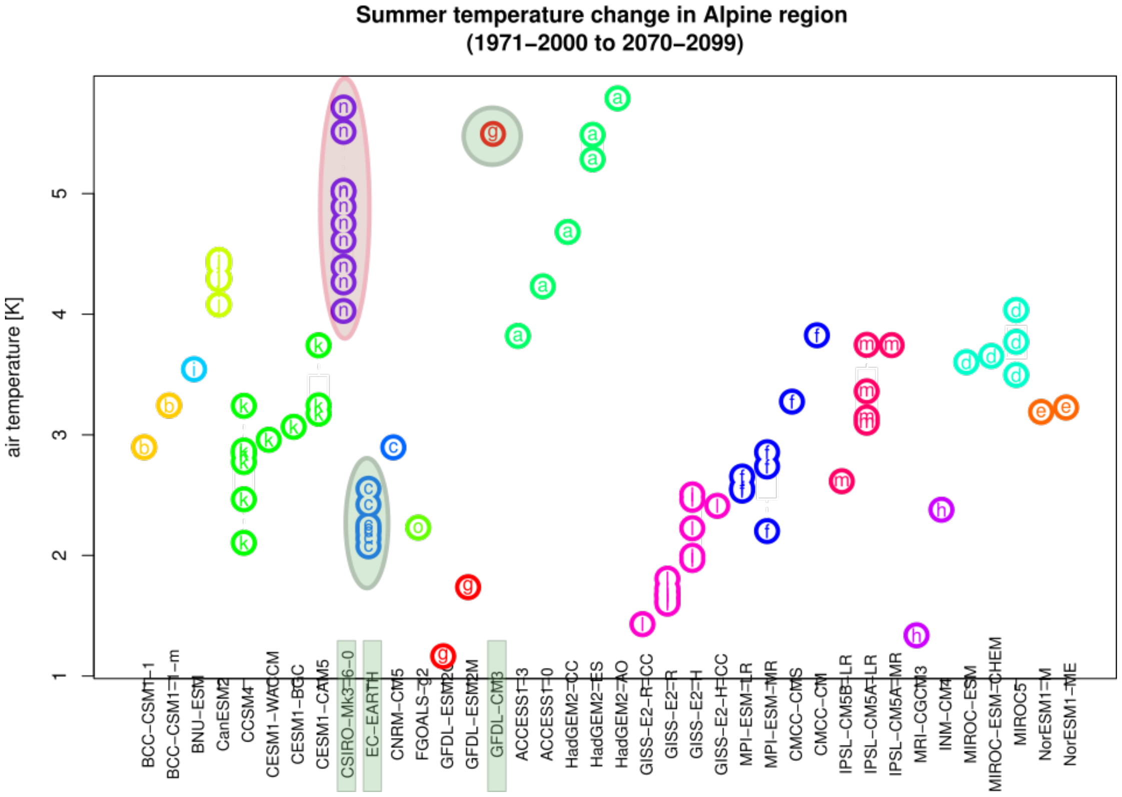


Figure 15.4: Comparing temperature changes (y-axis) of different GCMs (x-axis). The points represent individual GCM runs. The colours and letters mark models which have either been developed at the same institute, or share substantial parts of code. The three problematic GCMs are highlighted.

Obtaining Time-Series Anomalies

Individual time-series have distinct (non-linear) trends, and we are not interested in predicting the shape, but rather to catch the climate change and the year-to-year variability. An alternative solution to fit a non-linear mixed model is to subtract the non-linear shape from each time series individually before fitting the mixed model. This has been done using following steps:

1. Get the average temperature (i.e. climate) for each simulation's reference period (1971-2000) and the scenario period (2070-2099).

2. Calculate the climate change signal for each simulation (scenario period minus reference period).

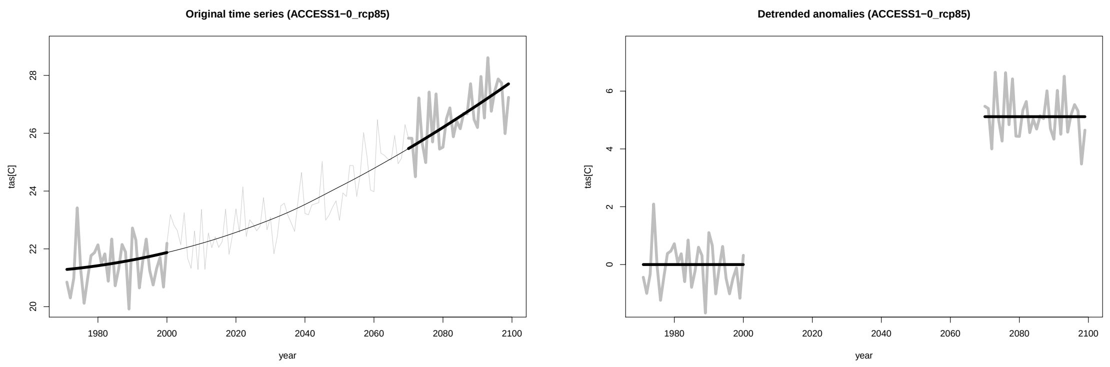


Figure 15.5: Temperature for IP in JJA. Left: Original, complete time-series from 1971 to 2099 with loess smoother. Right: Detrended time-series anomaly 1971-2000 and 2070-2099.

3. For each complete time-series from 1971-2099, fit a loess smoother (figure 15.5 - left).

4. Subtract the loess smoother from the data - all data is now centred around 0.

5. Take only data points of the reference period (1971-2000) and the scenario period (2070-2099), delete the residual data.

6. Add the individual climate change signals to the scenario period (2070-2099) part (figure 15.5 - right).

The result is given as two temperature time-series for each simulation: A zero-centred series representing the reference period (1971-2000) and a scenario anomaly time series centred around the average future (2070-2099) temperature climate. For the following statistical analysis we only take the future anomaly curve with 30 data points.

15.2 Statistical Model I: Normal Multilevel Model

Let $Y_{ijkt} \in \mathbb{R}$ (**Level1**) denote the temperature anomaly at time step $t = 1, \ldots, 30$ of the kth run (run, **Level2**) of the jth GCM (gcm, **Level3**) stemming from the ith model family (sim, **Level4**).

We consider a linear mixed effects model with y_{ijkt} being the yearly temperature anomalies in time-step $t = 1, \ldots, 30$ in the future scenario 2070-2099 for simulation number $k = 1, \ldots, K_{ij}$ of the GCM $j = 1, \ldots, J_i$ being in the similarity class $i =$

$1,\ldots,15$. The model to implement the data is as follows:

$$\begin{array}{llll}
\textbf{Level 1} & & Y_{ijkt}|b^{(1)}_{ijk} \overset{iid}{\sim} N(\beta_0 + b^{(1)}_{ijk}, \sigma^2) & t = 1,\ldots,N_{ijk} \\
\textbf{Level 2} & l = 1 & b^{(1)}_{ijk}|b^{(2)}_{ij} \overset{iid}{\sim} N(b^{(2)}_{ij}, \sigma^2_{b1}) & k = 1,\ldots,K_{ij} \\
\textbf{Level 3} & l = 2 & b^{(2)}_{ij}|b^{(3)}_{i} \overset{iid}{\sim} N(b^{(3)}_{i}, \sigma^2_{b2}) & j = 1,\ldots,J_{i} \\
\textbf{Level 4} & l = 3 & b^{(3)}_{i} \overset{iid}{\sim} N(0, \sigma^2_{b3}) & i = 1,\ldots,I.
\end{array}$$

This relationship can be written in a multilevel regression model as in Section 12.3

$$Y_{ijkt} = \beta_0 + b^{(1)}_{ijk} + b^{(2)}_{ij} + b^{(3)}_{i} + \epsilon_{ijkt}$$

with

$$b^{(1)}_{ijk} \overset{iid}{\sim} N(0, \sigma^2_{b1}), \quad b^{(2)}_{ij} \overset{iid}{\sim} N(0, \sigma^2_{b2}), \quad b^{(3)}_{i} \overset{iid}{\sim} N(0, \sigma^2_{b3}), \quad \epsilon_{ijkt} \overset{iid}{\sim} N(0, \sigma^2),$$

where we assume constant natural variability across all simulations.

However, for the sake of better readability and interpretability, we write

$$Y_{ijkt} = \beta_0 + b_{sim,i} + b_{gcm,ij} + b_{run,ijk} + \epsilon_{ijkt} \tag{15.1}$$

with the random effects of climate change $b_{sim,i}$ for *model similarity* i , $b_{gcm,ij}$ for *GCM* j and its k-th *run* of the GCM being $b_{run,ijk}$. ϵ_{ijkl} denotes the year-to-year temperature variability. We write

$$b_{mod,i} \overset{iid}{\sim} N(0, \sigma^2_{mod}), \quad b_{gcm,ij} \overset{iid}{\sim} N(0, \sigma^2_{gcm}), \quad b_{sim,ijk} \overset{iid}{\sim} N(0, \sigma^2_{sim})$$

and with

$$\epsilon_{ijkt} \overset{iid}{\sim} N(0, \sigma^2_{nat})$$

The hierarchical model is depicted in Section 15.3.

15.3 Diagnostics

In this section we check for the adequacy of the statistical model which fits the CMIP5 multi-model ensemble data. Also of interest is the detection of influential data points, which have a strong effect on the statistical model outcome. These data can be either actual observations (here temperature climate projections) or un-observed quantities, expressed with random effects. The diagnostics for this case study have been performed as explained in Loy and Hofmann 2014.

Model Assumptions (Residual Analysis)

To check whether the statistical model is suited for the underlying data, we will work with three types of residuals as discussed in Loy and Hofmann 2014.

Level-1 (conditional) residuals: The residuals of the actually observed data points being $\boldsymbol{\epsilon}_i = \boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{b}_i$. Those can be estimated by either plugging in the Empirical Bayes (EB) estimates of $\boldsymbol{b}_i$ or by performing a separate least-squares (LS) fit for each group. As using the EB leads to confounded residuals, the recommended analysis is based on the LS method, which we will perform as well. We will plot the semi-standardised residuals $\widehat{\boldsymbol{\epsilon}}_i^*$ defined as

$$\widehat{\boldsymbol{\epsilon}}_i^* = \widehat{\sigma}_i\widehat{\Delta}_i^{-1/2}\widehat{\boldsymbol{\epsilon}}_i$$

where $\widehat{\Delta}_i$ is the diagonal matrix with elements equal to the diagonal of $\mathrm{Var}(\widehat{\boldsymbol{\epsilon}}_i) = \sigma_i^2(\boldsymbol{1} - \boldsymbol{h}_i)$ with $\boldsymbol{h}_i$ being the vector containing the diagonal elements of the hat matrix $\boldsymbol{H}_i = \boldsymbol{X}_i(\boldsymbol{X}_i'\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i$ from the LS model fit and $\mathrm{Var}(\boldsymbol{\epsilon}_i) = \sigma_i^2\boldsymbol{I}_{N_i}$. With these semi- standardised residuals we both check the linearity of the fixed effects $\boldsymbol{\beta}$ and to check for the homoscedasticity assumption of the linear model. QQ-plots are then performed to check for the normality assumption.

Random effects residuals: The residuals of the un-observed groupings which are modelled with random effects being either $\boldsymbol{Z}_i\boldsymbol{b}_i$ or $\boldsymbol{b}_i$. Also here we can either predict the random effects $\hat{\boldsymbol{b}}_i = \hat{\boldsymbol{\beta}}_i - \boldsymbol{W}_i\hat{\boldsymbol{\gamma}}$ either with LS (with $\boldsymbol{W}_i\hat{\boldsymbol{\gamma}}$ being the average trend), or use the EB as $\hat{\boldsymbol{b}}_i = \mathrm{E}(\boldsymbol{b}_i|\boldsymbol{y}_i)$.

Marginal (composite) residuals: Sum of level-1 and random-effects residuals $\boldsymbol{\zeta}_i = \boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} = \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i$. This residual is heavily confounded as all residual types are mixed together, not giving relevant information on the reason of a possible misspecification.

Here we perform diagnostics at different levels of the model. Level 1, the ϵ residual, is in this case the natural year-to-year variability of the temperature. The random effects residuals are the internal variability from different initial conditions, model similarity and GCM variability .

Residual Diagnostics for Alpine Region (AL) in Summer (JJA)

Level 1 Diagnostics Level 1 residuals (year-to-year changes of mean temperature) are assumed to be normally distributed with zero mean and constant variance. The left-hand side of Figure 15.6 shows the L1 residuals against the fitted y values. As there is no visible trend or structure, the linearity assumption does not seem to be violated. It is not clearly visible weather the homoscedasticity assumption is violated. For this Figure 15.6

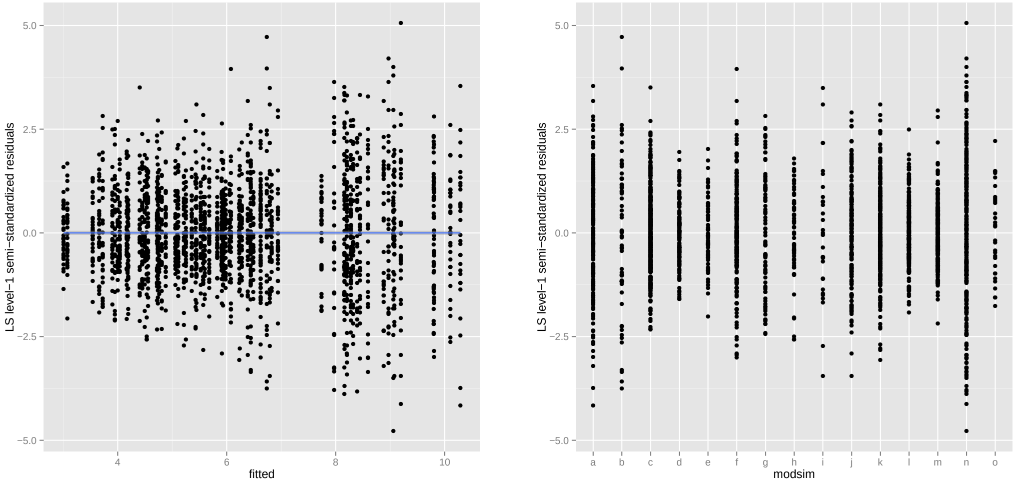


Figure 15.6: Semi-standardised L1 residuals against fitted values (left), and against the model similarity classes "sim" (right).

(right) depicts the residual spreads for the individual classes of simulation (sim). The natural year-to-year variance seems to be similar for most simulations, however, some simulations (e.g. class n with longest variance and d with smallest variance) have clearly a different natural variability and therefore violate the homogeneity assumption when assuming constant variance.

Checking the normality assumption leads to a similar picture: if assuming constant L1 variance, the residuals are clearly more heavy tailed than expected from Gaussian distributed data (Figure 15.7). This clearly underlines the need for non-constant variances. The L1 distributions for every single simulation are depicted in Figure 15.8 (left). There, the residuals seem to follow a Gaussian distribution. This seems also to be true when aggregating the L1 residuals into the same class (sim) of GCMs (Figure 15.8 right). This indicates it is enough to assume constant variance σ_i^2 for simulations stemming from the same model class i and let the year-to-year variance vary between different classes of simulations.

Random Effects Diagnostics We look for the distributional assumptions of the random effects residuals $b_{sim,i}, b_{gcm,ij}, b_{run,ijk}$, which are obtained by the EB residual estimates. Figure 15.9 depicts their predicted distributions. The distribution of the *run* RE $b_{run,ijk}$ (being the distribution of the same gcm ij with different initial conditions k) indicates a more heavy tailed shape than expected from a normal distribution (Figure 15.9, left). This is due to the fact that the spread of the initial conditions ensembles is not constant across GCMs (see Figure 15.4).The Level 3 RE distribution $b_{gcm,ij}$ of GCMs j of same

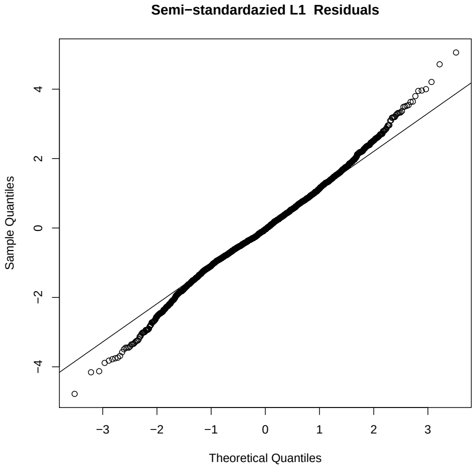


Figure 15.7: Checking of normality distribution of semi-standardised L1 residuals.

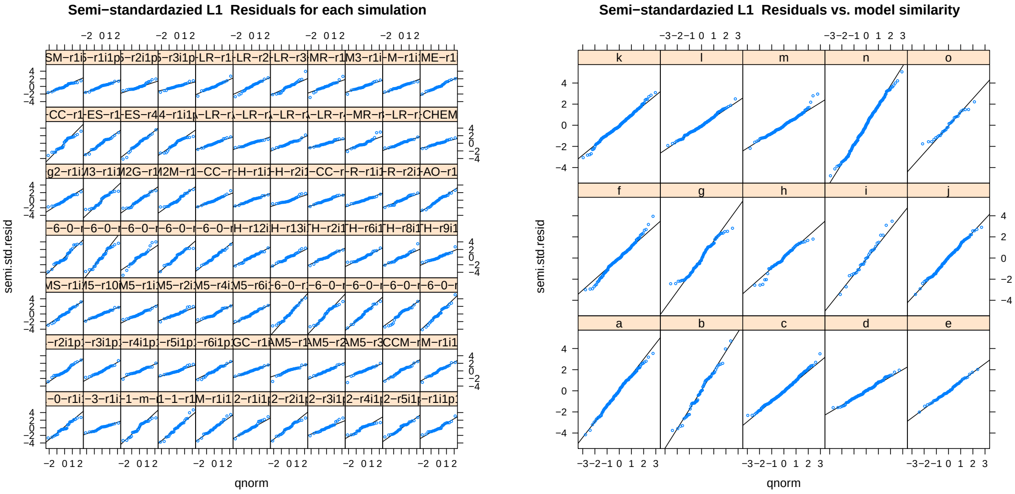


Figure 15.8: Checking of normality assumption of semi-standardised L1 residuals grouped by each simulation run (left) and by its similarity class (right).

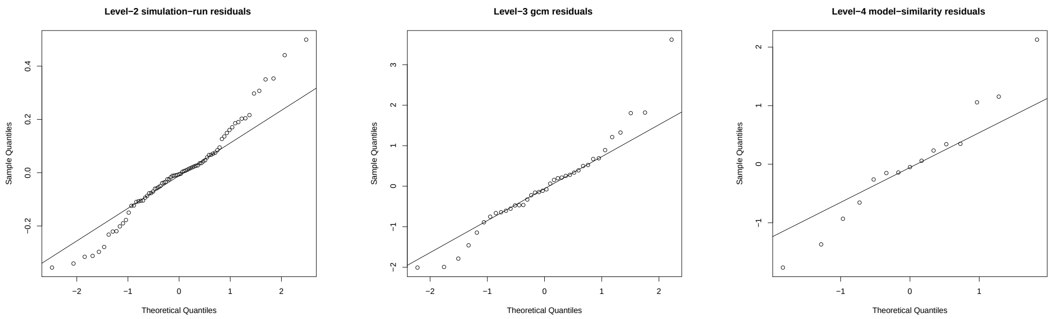


Figure 15.9: Random effects residual (Empirical Bayes) distribution for the Level 2 "run" effect (left), the Level 3 "gcm" effect (middle) and the Level 4 "sim" effect (right).

model class i, shows an outlying simulation (sim class g) and in general more of a skewed shape (Figure 15.9, middle). When fitting the residuals of the random slope of $b_{gcm,ij}$, the skewness estimate is significantly positive. The distribution of the predicted $b_{sim,i}$s (the spread of model classes) also shows a tendency towards a heavy tailed distribution, but the sample is quite small to give a clear conclusion (Figure 15.9, right).

Marginal Residuals The marginal distribution as the sum of all residuals is shown in Figure 15.10. A violation of the normality distribution is clearly visible as is the skewness of the Level 3 (gcm) effect.

Influence Diagnostics

Here we analyse the changes of the estimates and changes when deleting individual clusters (*random-effect deletion*). We omit a discussion on deletion based on individual observations (*level-1 deletion*) as the influence of single years to the overall fit is not too large.

For all influence diagnostics tools, we build our judgement based on visual identification of gaps in the empirical distribution. We therefore consider the outlier measure for boxplots to identify influential observations. So all data which exceed

$$Q_3 + 3 \times IQR$$

are marked and considered as outlier, where Q_3 is the $q_{0.75}$-quantile and $IQR = q_{0.75} - q_{0.25}$ is the interquartile range of the data.

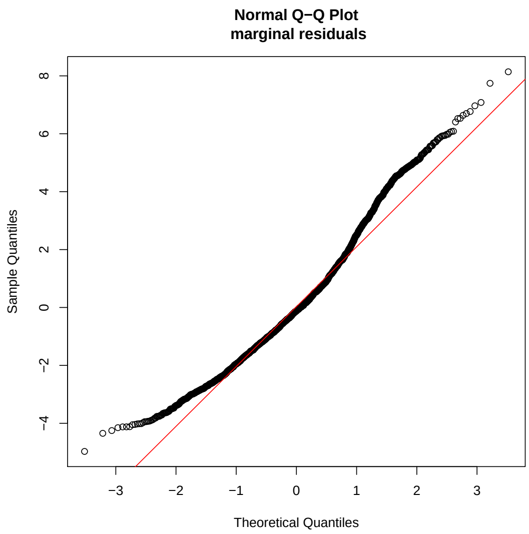


Figure 15.10: Marginal Level 1 residuals.

Diagnostics for Fixed Effects We use Cook's distance to assess for changes in the estimated fixed effects

$$C_i(\widehat{\boldsymbol{\beta}}) = \frac{1}{p}\left(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)}\right)' \widehat{Var(\widehat{\boldsymbol{\beta}})} \left(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)}\right),$$

with p being the rank of X. It describes the changes in the fixed parameter estimates normalised by the expected variability of the estimate. To account for changes of the parameter precision we can use the covariance trace

$$\mathrm{COVTRACE}_i(\widehat{\boldsymbol{\beta}}) = \left| tr\left(\widehat{Var(\widehat{\boldsymbol{\beta}})}^{-1} \widehat{Var(\widehat{\boldsymbol{\beta}}_{(i)})} \right) - p \right|$$

Diagnostics for Variance Components We can use Cook's distance and the covariance trace as for the fixed effects, however, calculation of covariance matrix of the random effects is expensive (with e.g. parametric bootstrap). Alternatively we can calculate the relative variance change (RVC) (Dillane 2005)

$$\mathrm{RVC}_i(\widehat{\theta}_l) = \frac{\widehat{\theta}_{l(i)}}{\widehat{\theta}_l} - 1, \tag{15.2}$$

where $\widehat{\theta}_{l(i)}$ is the estimate of the variance component when the ith unit is deleted.

Influence Diagnostics for Alpine Region (AL) in Summer (JJA)

Level 2 Influence We investigate the influence of excluding single data clusters on the estimate of the fixed effects $\widehat{\beta}$ and on the estimates of the variance components (VCs) $\sigma^2_{mod}, \sigma^2_{gcm}, \sigma^2_{sim}$ and $\sigma^2_{nat,i}$. Figure 15.11 shows Cook's distances for each L2 residual. The most striking influence on the fixed effects estimates stems from the outlying GCM *GFDL-CM3* (having only one run) and the GCM *CSIRO-Mk3* (which has several runs). Also the entire model similarity class *a* (being *HadGEM* and *ACCESS* models) seems to have an influence on the estimation of the fixed effects. Also, the GCM *CSIRO-Mk3* seems to have some influence on the standard error of the estimate of the fixed effect as well as the two simulation runs *FGOALS-g2* and *BNU-ESM* (Figure 15.12).

Both models *GFDL-CM3* and *CSIRO-Mk3* show the largest influence also on the VC estimates: As all the runs of the GCM *CSIRO-Mk3* have an higher-than-average natural year-to-year variability, an exclusion of all those runs decreases the σ_{nat} estimate by 25% (Figure 15.13, top-left and Figure 15.14, top-left). The same model also has an inflating effect on the variability estimate of σ_{run} (Figure 15.13, top-right), but on a much smaller scale ($< 10\%$). The same effect is induced by other GCMs as well. A deflating effect on the initial condition uncertainty is detected by the GCM *EC-EARTH* and others, again being minor (Figure 15.14, top-right). A large influence on the Level 3 random effect *gcm* can be clearly assigned to the GCM *GFDL-CM3*. An exclusion of this particular model, and even the exclusion of the entire *GFDL* model family (g) leads to a decrease of this VC by over 50% (Figure 15.13 and Figure 15.14, bottom left). However, at the same time when excluding this model, the VC of Level 4 (model similarity *sim*) increases the same amount (Figure 15.13 and Figure 15.14, bottom right). As both Level 3 and Level 4 explain the same structural uncertainty of the CMIP5 multi-model ensemble, this means the overall structural uncertainty is still not influenced too much by this particular model.

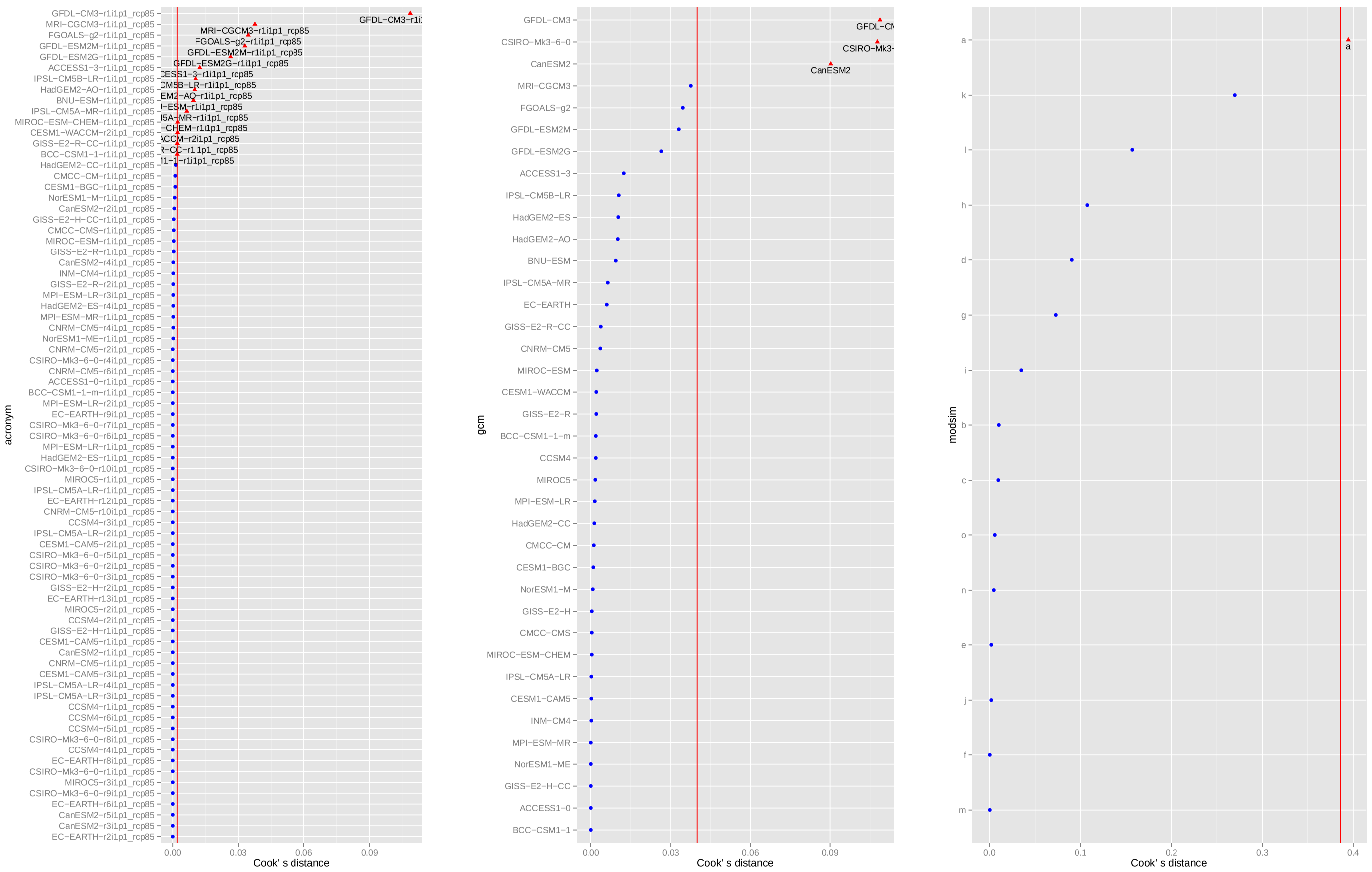


Figure 15.11: Influence on fixed effects (i.e. the average climate change signal): Cooks distance with case deletion for simulation runs (left), gcm (middle) and model similarity (right).

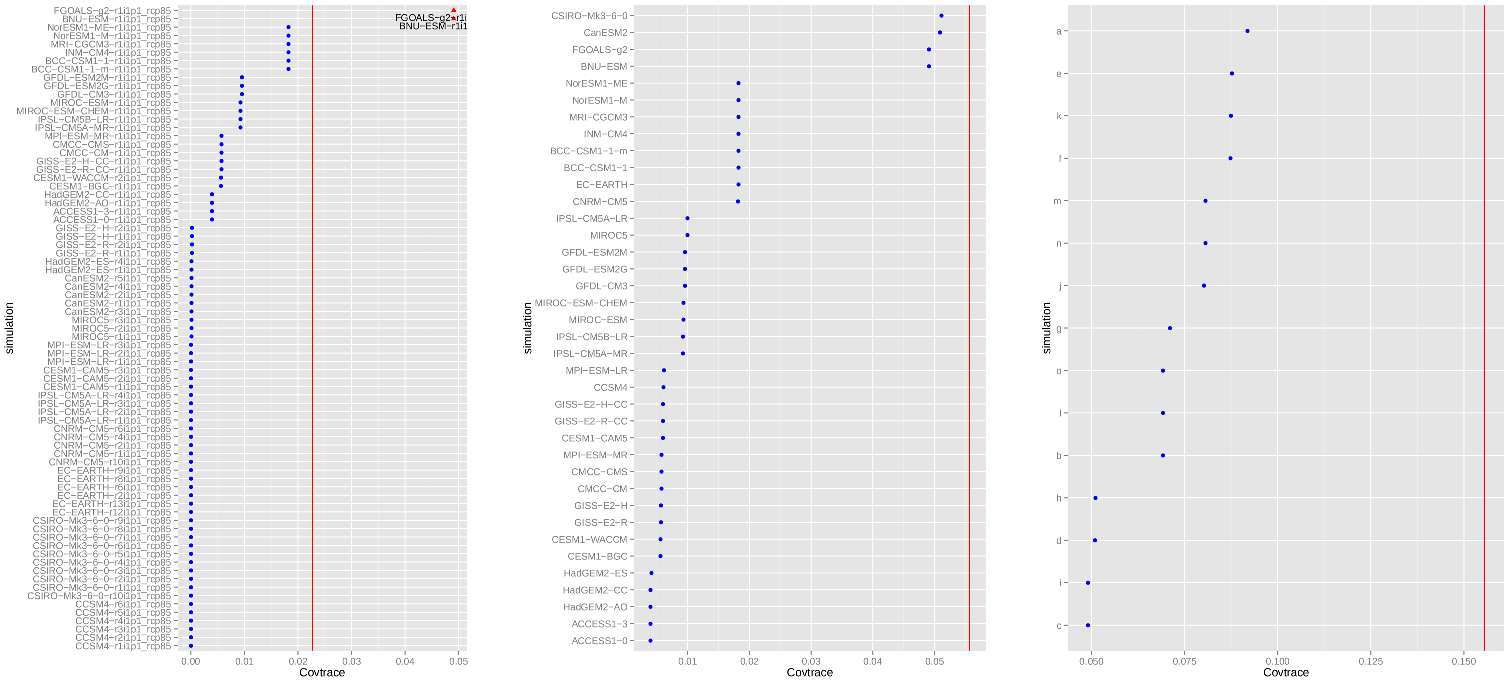


Figure 15.12: Influence on precision of the fixed effects (i.e. the standard error): Covtrace with case deletion for simulation runs (left), gcm (middle) and model similarity (right).

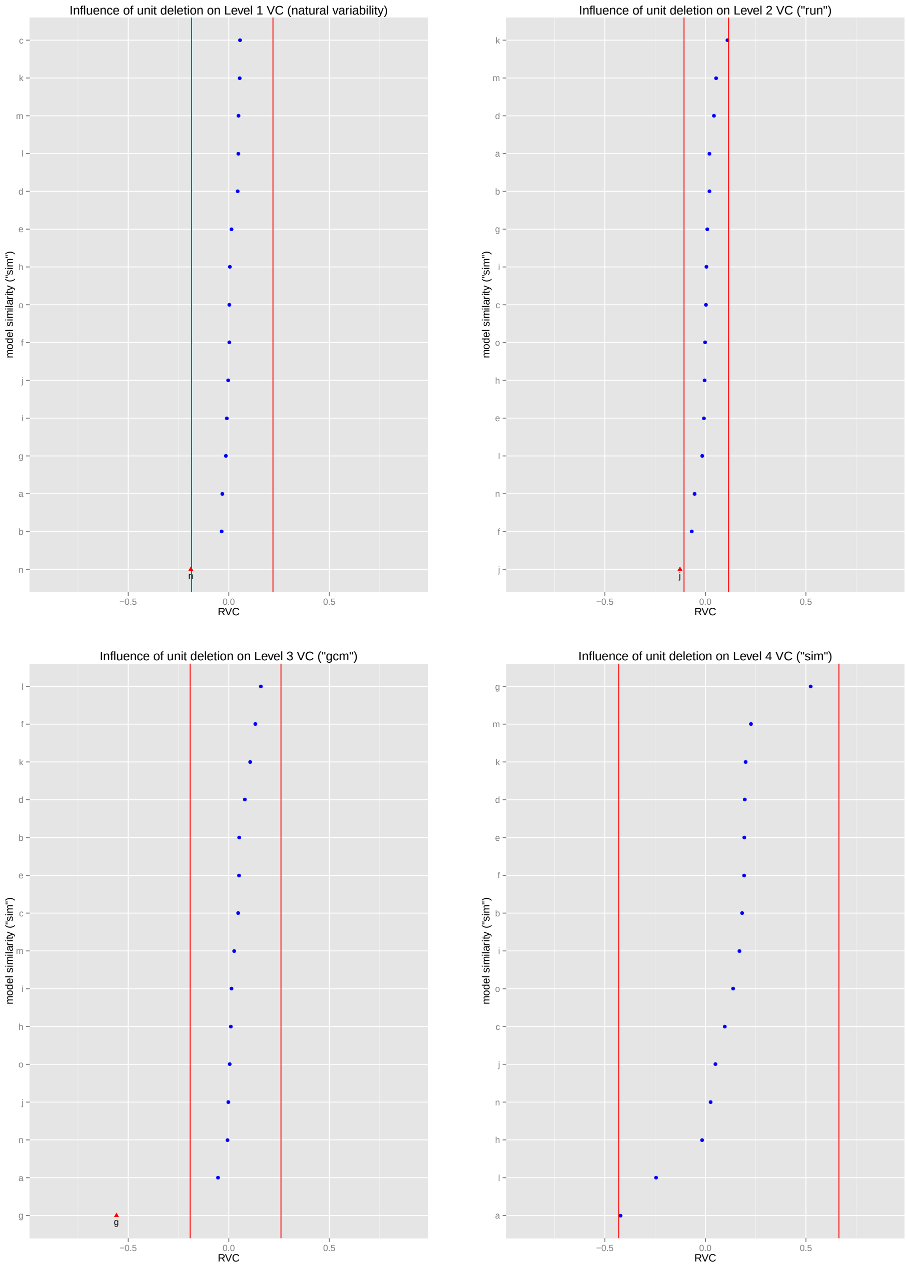


Figure 15.13: RVC when excluding entire clusters of the same model similarity class (Level4) for VC estimates on Level 1 (top left), Level 2 (top right), Level 3 (bottom left) and on Level 4 (bottom right).

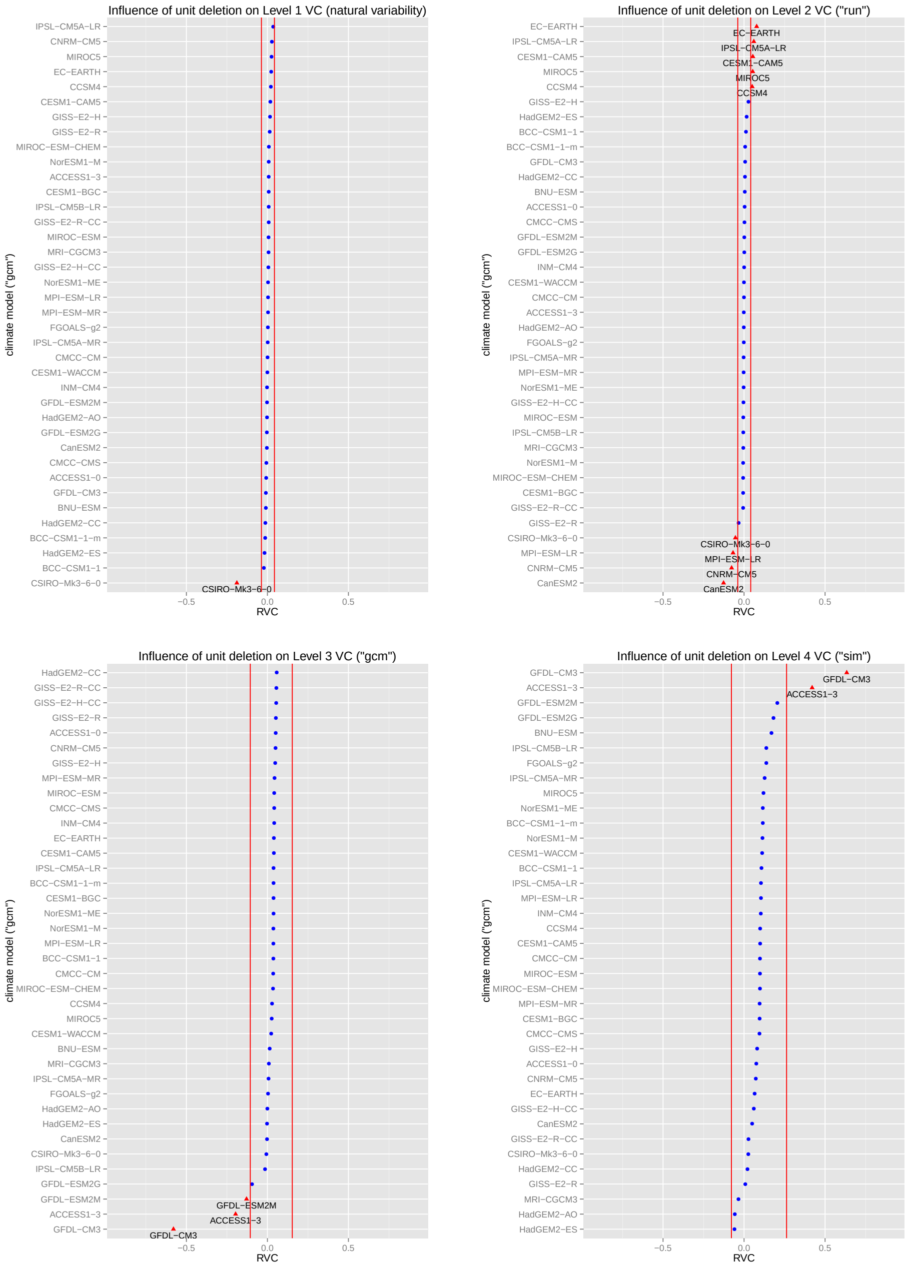


Figure 15.14: RVC when excluding entire clusters of the same *gcm* clusters (Level 3) for VC estimates on Level 1 (top left), Level 2 (top right), Level 3 (bottom left) and on Level 4 (bottom right).

15.4 Statistical Model II: Skew-Linear Multilevel Model

We extend the LMM be adding a skewness parameter for the GCM random effect (reason for this see Section 15.3):

$$b_{gcm,ij} \overset{iid}{\sim} SN(0, \tau^2_{gcm}, \lambda),$$

Likelihood Function

We extend the LMM be adding a skewness parameter for the GCM random effect as the corresponding residuals showed a non-symmetric behaviour. Also we do not assume overall constant natural variability at Level1 anymore, but rather assume this variability to be constant among simulations being from the same similarity class (Level4).

The distributional assumptions are now

Level 1		$Y_{ijkt} \vert b^{(1)}_{ijk} \overset{iid}{\sim} N(\beta_0 + b^{(1)}_{ijk}, \sigma^2_i)$	$t = 1, \ldots, N_{ijk}$
Level 2	$l = 1$	$b^{(1)}_{ijk} \vert \gamma^{(2)}_{ij} \overset{iid}{\sim} N(\gamma^{(2)}_{ij}, \sigma^2_{b1})$	$k = 1, \ldots, K_{ij}$
Level 3	$l = 2$	$\gamma^{(2)}_{ij} \vert b^{(3)}_i \overset{iid}{\sim} SN(b^{(3)}_i, \omega^2, \lambda)$	$j = 1, \ldots, J_i$
Level 4	$l = 3$	$b^{(3)}_i \overset{iid}{\sim} N(0, \sigma^2_{b3})$	$i = 1, \ldots, I.$

As before, this relationship can be written in a multilevel regression model

$$Y_{ijkt} = \beta_0 + b^{(1)}_{ijk} + \gamma^{(2)}_{ij} + b^{(3)}_i + \epsilon_{ijkt}$$

with

$$b^{(1)}_{ijk} \overset{iid}{\sim} N(0, \sigma^2_{b1}), \quad \gamma^{(2)}_{ij} \overset{iid}{\sim} SN(0, \omega^2, \lambda), \quad b^{(3)}_i \overset{iid}{\sim} N(0, \sigma^2_{b3})$$

being mutually independent and

$$\epsilon_{ijkt} \overset{iid}{\sim} N(0, \sigma^2_i)$$

where we assume constant natural variability across all simulations within one model similarity class.

In contrast to the normal distribution, the skew-normal distributions' expected value and variance cannot be directly interpreted reading the location- and scale parameter. Its expectation value and variance are

$$E(\gamma^{(2)}_{ij}) = \sqrt{\frac{2}{\pi}}\,\delta\omega,$$
$$Var(\gamma^{(2)}_{ij}) = \omega^2 - \frac{2}{\pi}\,\delta^2\omega^2,$$

with $\delta = \lambda/\sqrt{1+\lambda^2}$. The skewness of the random effect is

$$\gamma = \mathrm{E}\left(\frac{(\gamma_{ij}^{(2)} - \mathrm{E}(\gamma_{ij}^{(2)}))^3}{\mathrm{Var}(\gamma_{ij}^{(2)})^{3/2}}\right) = \frac{4-\pi}{2}\frac{(\delta\sqrt{2\pi})^3}{(1-2\delta^2/\pi)^{3/2}}, \tag{15.3}$$

where for the skew-normal (SN) class of distribution, the skewness is limited to $|\gamma| \leq .995$. The expected temperature climate change can be interpreted as

$$\mathrm{E}(Y_{ijkl}) = \beta_{0,dp} + \sqrt{\frac{2}{\pi}}\delta\omega =: \beta_0 \tag{15.4}$$

and with the overall (marginal) variance being

$$\mathrm{Var}(Y_{ijkl}) = \sigma^2_{run} + \sigma^2_{gcm} + \sigma^2_{sim} + \sigma^2_{nat,i} \tag{15.5}$$

with $\sigma^2_{gcm} := Var(\gamma_{ij}^{(2)}) = \omega^2 - (2/\pi)\delta^2\omega^2$.

We are therefore primarily interested in estimating the expected climate change parameter β_0 and its variance components $\sigma^2_{sim}, \sigma^2_{gcm}, \sigma^2_{run}$ and the variance component of the natural year-to-year variability $\sigma^2_{nat,i}$.

Those reparameterizations have been performed directly within the likelihood function, thus the maximum likelihood estimates (MLEs) obtained are directly interpretable (see Section 13.3) for details.

Vectorize to Level 2 We can vectorize the level 1 random effect and obtain the observed vector of i in j in k:

$$\boldsymbol{Y}_{ijk} = \mathbf{1}\beta_0 + \mathbf{1}b_{ijk}^{(1)} + \mathbf{1}\gamma_{ij}^{(2)} + \mathbf{1}b_i^{(3)} + \boldsymbol{\epsilon}_{ijk}$$

with $\mathbf{1} = \mathbf{1}_{N_{ijk}}$ and N_{ijk} is the amount of observed data in clusters k within j within i. Due to normality of L1

$$\boldsymbol{\epsilon}_{ijk} \sim N_{N_{ijk}}(\mathbf{0}, \boldsymbol{I}_{N_{ijk}} \otimes \sigma_i^2)$$

Vectorize to Level 3 If we further vectorize to $\boldsymbol{Y}_{ij}$ we get

$$\begin{pmatrix} b_{ij1}^{(1)} \\ \vdots \\ b_{ijK_{ij}}^{(1)} \end{pmatrix} \sim N_{K_{ij}}\left(\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{b1} & & 0 \\ & \ddots & \\ 0 & & \sigma^2_{b1} \end{pmatrix}\right),$$

which we can write as $\boldsymbol{b}_{ij}^{(1)} = (b_{ij1}^{(1)}, \ldots, b_{ijK_{ij}}^{(1)})'$ and so

$$\boldsymbol{b}_{ij}^{(1)} \sim N_{K_{ij}}(\mathbf{0}, \boldsymbol{I}_{K_{ij}} \otimes \sigma^2_{b1})$$

and we obtain the design matrix for this random effect as

$$\boldsymbol{Z}_{ij}^{(1)} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \dots & 0 \\ & \ddots & & \\ & & \ddots & \\ 0 & \dots & 0 & 1 \\ \vdots & & \vdots & \vdots \\ 0 & \dots & 0 & 1 \end{pmatrix} = \begin{pmatrix} \mathbf{1}_{L_{ij1}} & & 0 \\ & \ddots & \\ 0 & & \mathbf{1}_{L_{ijK_{ij}}} \end{pmatrix} = \mathbf{1}_{L_{ij1}} \oplus \dots \oplus \mathbf{1}_{L_{ijK_{ij}}} \in \mathbb{R}^{N_{ij} \times K_{ij}}$$

with

$$\boldsymbol{Y}_{ij} = \mathbf{1}\beta_0 + \boldsymbol{Z}_{ij}^{(1)} \boldsymbol{b}_{ij}^{(1)} + \mathbf{1}\gamma_{ij}^{(2)} + \mathbf{1}b_i^{(3)} + \boldsymbol{\epsilon}_{ij}$$

with length of each $\mathbf{1}$ being the amount of observed data given the first two levels ij being $N_{ij} = \sum_{k=1}^{K_{ij}} N_{ijk}$ and K_{ij} being the sample size of the random effects (number of clusters) $b_{ijk}^{(1)}$ given ij.

Vectorize to Level 4 Now vectorizing the data to $\boldsymbol{Y}_i$ we get

$$\boldsymbol{Y}_i = \mathbf{1}\beta_0 + \boldsymbol{Z}_i^{(1)} \boldsymbol{b}_i^{(1)} + \boldsymbol{Z}_i^{(1)} \boldsymbol{\gamma}_i^{(2)} + \mathbf{1}b_i^{(3)} + \boldsymbol{\epsilon}_i$$

with

$$\boldsymbol{b}_i^{(1)} = (\boldsymbol{b}_{i1}^{(1)\prime}, \dots, \boldsymbol{b}_{iJ_i}^{(1)\prime})', \quad \boldsymbol{\gamma}_i^{(2)} = (\gamma_{i1}^{(2)}, \dots, \gamma_{iJ_i}^{(2)})'$$
$$\boldsymbol{Z}_i^{(1)} = \boldsymbol{Z}_{i1}^{(1)} \oplus \dots \oplus \boldsymbol{Z}_{iJ_i}^{(1)}, \quad \boldsymbol{Z}_i^{(2)} = \mathbf{1}_{N_{i1}} \oplus \dots \oplus \mathbf{1}_{N_{iJ_i}}$$

and with the number of clusters at level l nested within cluster i we get $m_{l(i)} = \sum_{j=1}^{J_i} K_{ij}$ for $l = 1$, yielding

$$\boldsymbol{b}_i^{(1)} = (\boldsymbol{b}_{i1}^{(1)\prime}, \dots, \boldsymbol{b}_{iJ_i}^{(1)\prime})' \sim N_{m_{1(i)}}(\mathbf{0}, \sigma_{b1}^2 \boldsymbol{I}_{m_{1(i)}}).$$

However, the multivariate variable $\boldsymbol{\gamma}_i^{(2)} \in \mathbb{R}^{J_i}$ is not SN distributed anymore, as shown in Corollary 13.19.

Marginal Distribution at Level 4 We rewrite the model with

$$\boldsymbol{b}_i := \begin{pmatrix} b_i^{(3)} \\ \boldsymbol{b}_i^{(1)} \end{pmatrix}, \quad \gamma_i := \boldsymbol{\gamma}_i^{(2)} = \left(\gamma_{i1}^{(2)}, \dots, \gamma_{iJ_i}^{(2)}\right)'$$
$$\boldsymbol{Z}_i^b := \left(\mathbf{1}_{N_i} \vdots \boldsymbol{Z}_i^{(1)}\right), \quad \boldsymbol{Z}_i^\gamma := \boldsymbol{Z}_i^{(2)},$$

yielding

$$\boldsymbol{Y}_i = \boldsymbol{1}\beta_0 + \boldsymbol{Z}_i^{\gamma}\boldsymbol{\gamma}_i + \boldsymbol{Z}_i^{b}\boldsymbol{b}_i + \boldsymbol{\epsilon}_i,$$

as in Corollary 13.19 and as $m_3(i) \equiv 1$, the vector length of $\boldsymbol{\gamma}_i$ equals $m_3(i) + m_1(i) = m_1(i) + 1$ and we have

$$\boldsymbol{b}_i = \begin{pmatrix} b_i^{(3)} \\ \boldsymbol{b}_i^{(1)} \end{pmatrix} \sim N_{m_1(i)+1}\left(\boldsymbol{0}, \boldsymbol{\Sigma}_b\right), \quad \text{with} \quad \boldsymbol{\Sigma}_b = \begin{pmatrix} \sigma_{b3}^2 & 0 \\ 0 & \boldsymbol{I}_{m_1(i)}\sigma_{b1}^2 \end{pmatrix}$$

The number of levels L in the skew-normal part $\boldsymbol{\gamma}_i$ is 1 so we can skip the superscript (l) yielding $\boldsymbol{\gamma}_i = (\gamma_{i1}, \ldots \gamma_{iJ_i})'$ with

$$\gamma_{ij} \overset{iid}{\sim} SN(0, \omega^2, \lambda),$$

and we obtain the marginal distribution from Corollary 13.19 with $m_{2(i)}$ being the total amount of clusters of the $l = 2$ skew-normal random effect, nested within cluster i (length of $\boldsymbol{\gamma}_i$):

$$f(\boldsymbol{y}_i|\boldsymbol{\theta}) = 2^{m_{2(i)}}\phi_{N_i}(\boldsymbol{y}_i|\boldsymbol{1}\beta_0, \boldsymbol{\psi}_i)\Phi_{m_{2(i)}}(\lambda\omega\boldsymbol{Z}_i^{\gamma'}\boldsymbol{\psi}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{1}\beta_0)|\boldsymbol{0}, \boldsymbol{I}_{m_2(i)} + \lambda^2/\omega^2\boldsymbol{\Lambda}_i) \tag{15.6}$$

with $\boldsymbol{\theta} = (\beta_0, \sigma_1^2, \ldots, \sigma_I^2, \sigma_{b1}^2, \omega, \sigma_{b3}^2, \lambda)'$ and

$$\begin{aligned}
\boldsymbol{\Lambda}_i &= (\omega^{-2}\boldsymbol{I}_{J_i} + \boldsymbol{Z}_i^{b'}\boldsymbol{\psi}_{b,i}^{-1}\boldsymbol{Z}_i^{b})^{-1} \\
\boldsymbol{\psi}_{b,i} &= \sigma_i^2\boldsymbol{I}_{N_i} + \boldsymbol{Z}_i^{b}\boldsymbol{\Sigma}_b\boldsymbol{Z}_i^{b'} = \sigma_i^2\boldsymbol{I}_{N_i} + \sigma_{b3}^2\boldsymbol{1}_{N_i}\boldsymbol{1}_{N_i}' + \sigma_{b1}^2\boldsymbol{Z}_i^{(1)}\boldsymbol{Z}_i^{(1)'} \\
\boldsymbol{\psi}_i &= \boldsymbol{\psi}_b + \omega^2\boldsymbol{Z}_i^{b}\boldsymbol{Z}_i^{b'}.
\end{aligned}$$

The Likelihood Function We can now write the log-likelihood function to be maximised as

$$\begin{aligned}
l(\boldsymbol{\theta}; \boldsymbol{y}) &= \log\left(\prod_{i=1}^{I} f(\boldsymbol{y}_i|\boldsymbol{\theta})\right) = \sum_{i=1}^{I}\log f(\boldsymbol{y}_i|\boldsymbol{\theta}) \\
&= \sum_{i=1}^{I}\left(m_{2(i)}\log 2 - \frac{1}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{\psi}_{b,i}| - \frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{1}\beta_0)'\boldsymbol{\psi}_{b,i}^{-1}(\boldsymbol{y}_i - \boldsymbol{1}\beta_0)\right) \\
&\quad + \sum_{i=1}^{I}\left(\Phi_{m_{2(i)}}(\lambda\omega\boldsymbol{Z}_i^{\gamma'}\boldsymbol{\psi}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{1}\beta_0)|\boldsymbol{0}, \boldsymbol{I}_{m_2(i)} + \lambda^2/\omega^2\boldsymbol{\Lambda}_i)\right).
\end{aligned}$$

There is no analytic solution for this maximisation problem. We therefore optimise this likelihood function numerically with the native optimiser in R which allows box constraints (Byrd et al. 1995).

15.5 Results

Estimates of Parameters

Here we compare the estimates across the different statistical models, typically used in climate science with the normal multilevel models and the here newly developed skew-normal multilevel model.

Statistical Models The classical linear models (LMs) analysed here are often used in climate research to quantify climate model uncertainties. It is a well-known fact that different runs from the same GCM (here Level 2) project similar climate changes and thereby are highly dependent. In practice this is by-passed by either averaging over all runs of the same GCM or to pick one as representative and to perform the uncertainty analysis based on the reduced set. Also, the time series has to be averaged as well, as this would also yield an additional dependency component, so natural year-to-year variability cannot be analysed with this model. In Table 15.1 we indicate the deletion/averaging in the dataset by writing *-d-* in the corresponding column.

The linear mixed effect model *LMM* is the logical extension of the approach above when we do not eliminate the initial-condition uncertainty by keeping all runs from the GCMs. Also, this model can address the time-series structure and therefore estimates the natural variability as well.

The model *LMM-sim* adds another level to the multilevel model *LMM* by adding the model similarity component. This model (15.1) was described in the Section before.

The model *LMM-het* is the same as *LMM-sim*, but with heterogeneous Level 1 year-to-year variability in the time series.

Model *SN-LMM* is the skew-normal multilevel model and estimates the same parameters as model *LMM-het* but relaxes the normality assumption by allowing for skewness of the VC of Level 2 (GCM).

Interpretation Table 15.1 shows the estimates of the models described above. Firstly, the expected climate change signals $\widehat{\beta}_0$ are all very similar across the different methods. Also the corresponding standard errors are comparable. Interestingly in SC in winter, the standard error of the *SN-LMM* is quite smaller than that of its most similar model *LMM-het*. When considering the 3 Level model *LMM* which ignores the similarity structure, its estimated standard error tends to be smaller than for those models which explicitly account for model dependency. For the *SN-LMM* model, in IP in summer and in SC in winter (DJF), the Level 4 VC for similarity is quite a bit smaller compared to its normally distributed sibling *LMM-het*.

When interpreting the parameter estimates one can see some differences between the different seasons and the different regions. In IP and AL the summer climate change signals are much larger than in winter, which is the other way around for SC, where the

		summer						winter					
		β_0	σ_{nat}	σ_{run}	σ_{gcm}	σ_{sim}	γ	β_0	σ_{nat}	σ_{run}	σ_{gcm}	σ_{sim}	γ
IP													
	LM	5.20 (.22)	-*d*-	-*d*-	1.31	-	-	3.20 (.14)	-*d*-	-*d*-	.81	-	-
	LMM	5.18 (.21)	.78	.20	1.25	-	-	3.18 (.13)	.63	.21	.74	-	-
	LMM-sim	5.19 (.28)	.78	.20	.91	.87	-	3.15 (.16)	.63	.21	.53	.49	-
	LMM-het	5.19 (.29)	.82*	.20	.90	.92	-	3.14 (.17)	*.65	.22	.52	.51	-
	SN-LMM	5.23 (.22)	.80*	.19	.89	.67	.995	3.14 (.16)	*.65	.22	.49	.52	.16
AL													
	LM	6.07 (.31)	-*d*-	-*d*-	1.89	-	-	4.15 (.18)	-*d*-	-*d*-	1.10	-	
	LMM	6.05 (.30)	1.21	.29	1.80	-	-	4.12 (.17)	1.08	.32	1.01	-	-
	LMM-sim	6.03 (.39)	1.21	.29	1.28	1.20	-	4.07 (.22)	1.08	.33	.68	.68	-
	LMM-het	6.02 (.40)	*1.22	.28	1.28	1.26	-	4.07 (.23)	*1.15	.31	.69	.71	-
	SN-LMM	6.07 (.34)	*1.18	.28	1.31	1.02	.995	4.08 (.22)	*1.13	.31	.67	.68	.995
SC													
	LM	4.61 (.22)	-*d*-	-*d*-	1.30	-	-	6.04 (.29)	-*d*-	-*d*-	1.71	-	-
	LMM	4.57 (.20)	.97	.36	1.16	-	-	5.90 (.27)	1.37	.57	1.56	-	-
	LMM-sim	4.55 (.26)	.97	.36	.86	.77	-	5.93 (.33)	1.37	.57	1.25	.90	-
	LMM-het	4.55 (.27)	*.98	.35	87	.81	-	5.93 (.34)	*1.38	.55	1.26	.96	-
	SN-LMM	4.57 (.26)	*.96	.35	.85	.80	.995	5.97 (.24)	*1.33	.52	1.33	.52	.995

Table 15.1: Seasonal temperature climate change projections for 2070-2099 in the Iberian Peninsula (IP), the Alpine Region (AL) and the Scandinavian Region (SC). β_0 is the average climate change signal with corresponding standard error. The σs denote the 4 Levels of VCs. Model *LM* is the linear regression modelling for the climate change signal and picking one run (Level2) from each GCM (Level3) and $-d-$ denotes that the corresponding data have been eliminated before the analysis. *LMM* is a linear mixed model additionally assessing for the time series and the different runs. *LMM-sim* adds a model similarity Level. *LMM-het* implements varying natural variability (L1) for each similarity class (L4). *SN-LMM* relaxes the normality assumption by modelling a skewed distribution. The asterisk $*$ denotes the average variance of $\sigma^2_{nat,i}$ over all $i = 1, \ldots, 15$.

winter shows a larger change than the summer. Natural year-to-year variability seems to go along with this pattern. The largest natural variability can be detected in SC in winter and in AL in summer and winter. IP shows very little natural fluctuations. The initial condition uncertainty is the smallest component of all VCs - different initial conditions of the same GCM tend to yield similar results. The VCs for the structural climate model uncertainty σ_{gcm} and σ_{sim} are mostly of the same order of magnitude and are by far larger than the natural variability. All regions in both seasons (except for IP in winter) show a large right-skewed behaviour with γ_1 being at the parameter boundary of .995.

Estimates of Standard Errors

Standard errors of the parameters obtained by the skew-normal multilevel model are estimated with both non-parametric and parametric bootstrap techniques as well as using the likelihood ratio test statistic and Wald's test statistic (Chapter 14). The statistical model here is slightly different from the model above, as the time series has been averaged to a single climate change signal value for each simulation. This decreases the complexity of the dataset which is necessary for computational reasons, as obtaining the standard errors of the estimates often requires excessive re-evaluation of the model, which in our case, is very slow.

Methods We implemented the non-parametric bootstrap (Section 14.3) by sampling the data in the dataset from all levels (with replacement) starting from the lowest level (Level4). So first the algorithm samples from the model similarity classes (Level4), then within those classes remaining in the dataset, we sample the GCMs (Level3). From the remaining GCMs we sample the initial condition runs from the GCM (Level2). On Level 1 we perform no randomisation, in fact, we average the entire time-series to infer on climate change signals only, as an extensive bootstrap would not be computationally feasible otherwise. This approximation showed no effect on the resulting estimates.

The parametric bootstrap has been performed based on random number generation based on the MLE estimates. So at each level of the model, the random effects as well as the Level 1 observations have been artificially generated and re-estimated. The variability of the MLE estimates is then approximated by the variability of the parametric bootstrap sample.

The much faster likelihood ratio test statistic (Section 14.2) evaluates the likelihood with varying parameter values and compares it to the likelihood value of the MLE.

Wald's test statistic (Section 14.1) is based on the second order approximation of the MLE distribution and assumes symmetric behaviour (i.e. a quadratic likelihood function). We only need to calculate the Hessian matrix of the likelihood function at the MLE.

Results The results are shown in Figure 15.15 for the climate change parameter β_0, in Figure 15.16 for the three VCs and in Figure 15.17 for Pearson's skewness coefficient γ_1.

The uncertainty estimates for the climate change signal β_0 are all very similar across different estimation methods. For this parameter the simplest Wald standard error suffices. All climate change signals highly differ from 0.

The different methods to estimate the uncertainties of the VCs, show quite a similar picture for Level 3 VC (variability across GCMs) and for Level 2 VC (variability across runs of the same GCM). Both VCs are significantly different from 0 across all methods and regions/seasons. It seems quite surprising that Wald's approximation still seems to work pretty well. For the climate model similarity VC the behaviour between the bootstrap techniques and the likelihood based techniques (likelihood ratio (LR) and Wald) start to differ. For all regions and seasons the bootstrap CIs are quite large and always contain 0. Wald and LR are similar and tend to have a smaller CI spread indicating statistical significance.

However, for the skewness coefficient γ_1 Wald's method does not seem to work at all, as it highly underestimates the variability of the skewness estimate. In addition, the method has difficulties on the parameter space boundaries as the CIs can reach values which are outside the defined range. Beside of that the two bootstrap techniques perform similarly, where the LR CI estimate tends to be a bit narrower but still seems to be reasonable. Overall the skewness parameters tend to have large CIs mostly including 0. The parametric bootstrap CI always overlaps with 0, whereas the non-parametric bootstrap shows significance only in SC in winter, but being of the same order of magnitude as its parametric version. The LR CI indicates a significant skewness for IP in summer and Al in summer as well as SC in winter. IP in DJF seems to be the only case where no skewness is clearly prevalent.

15.6 Summary and Conclusions

In this Section we present a statistical model which is able to assess the problematic design structure of climate multi-model ensembles (CMIP5), such as imbalance and dependencies across climate simulations. In addition, rigorous model checking shows the need of a more general distribution than the Gaussian, which in this case is the skew-normal distribution.

Statistical Model

The multi-model ensembles can be nicely modelled by a hierarchical statistical model, accounting for each uncertainty component separately. Every hierarchy level is represented by a variance parameter (called variance component, VC) which is being estimated and can be directly interpreted as a source of uncertainty in the climate multi-model

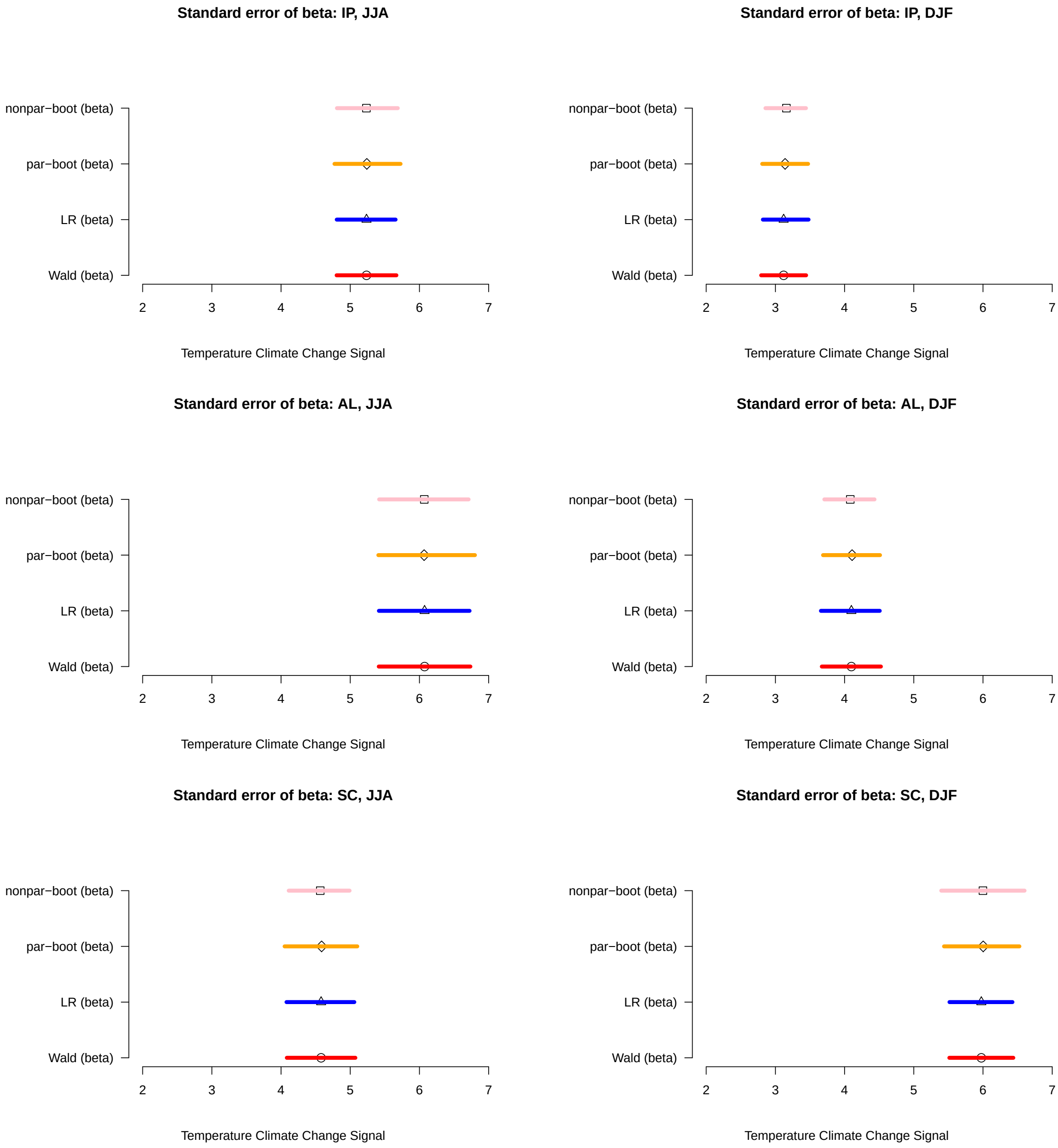


Figure 15.15: Estimated climate change signal: CIs for estimated fixed effect $\widehat{\beta}_0$.

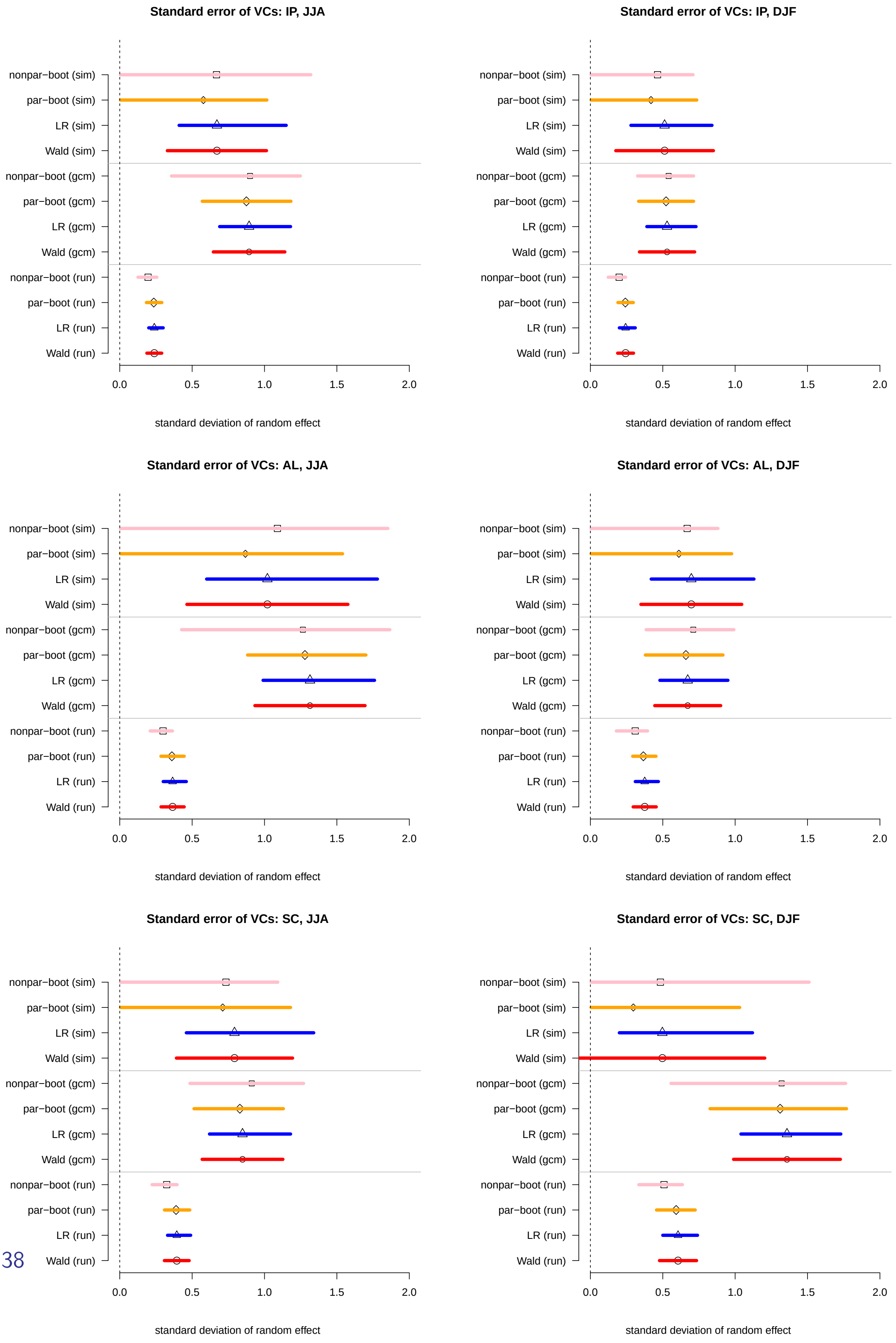


Figure 15.16: Estimated sources of climate uncertainties: CIs for estimates of VCs $\widehat{\sigma}_{sim}$ (*model similarity/structural uncertainty*), $\widehat{\sigma}_{gcm}$ (*structural uncertainty*) and $\widehat{\sigma}_{run}$ (*initial condition uncertainty*).

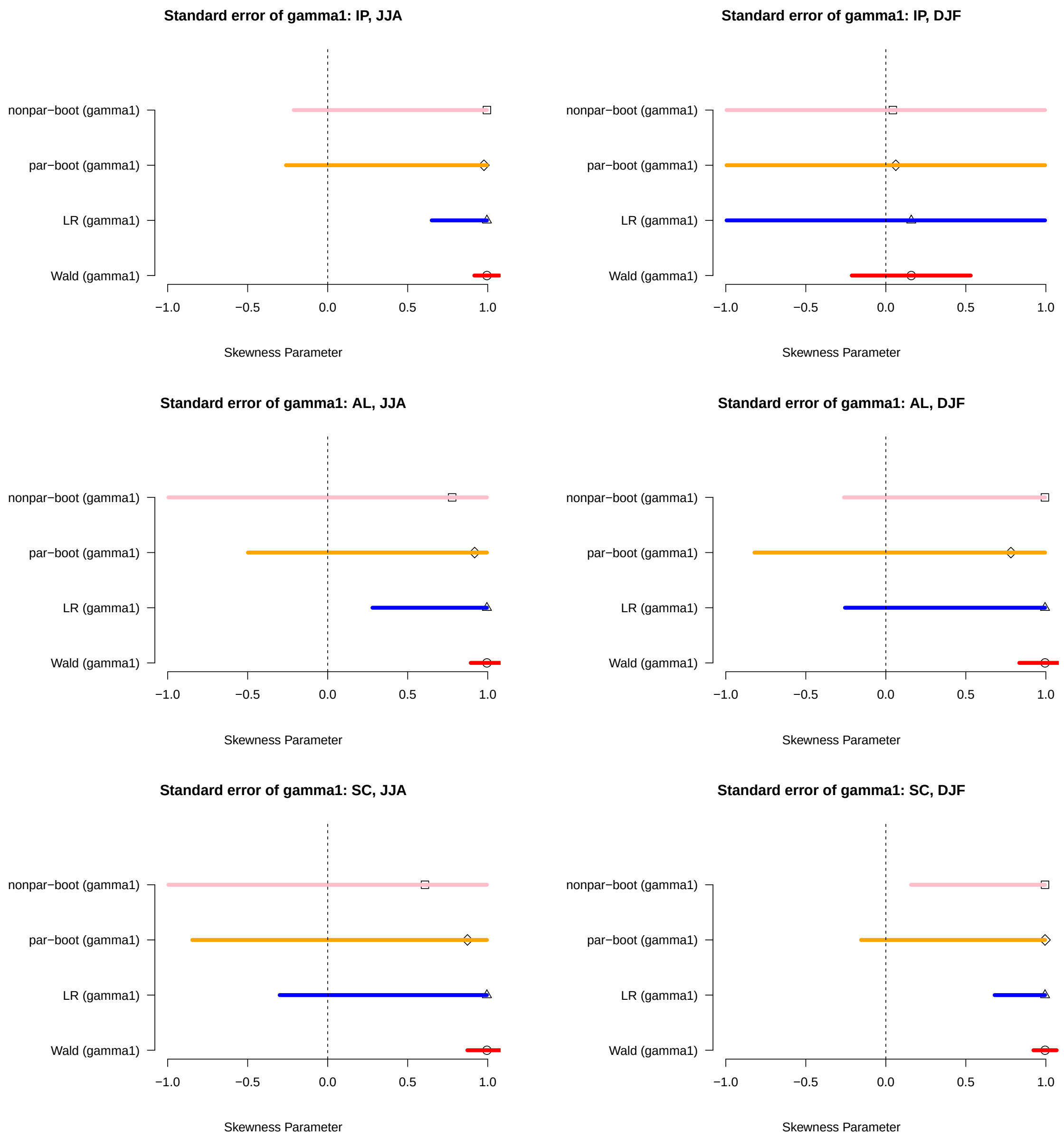


Figure 15.17: Estimated skewness of the pdf: CIs for skewness coefficient $\widehat{\gamma}_1$ ($|\gamma_1| \leq .995$).

ensemble. We start from the natural year-to-year variability described by a climate simulation (Level 1, *natural variability*), and go further to assess the uncertainty induced by different initial conditions of the same climate simulation (Level 2, *initial condition uncertainty*). Further, we account for the variation between climate models (GCMs) stemming from the same model family, either because they have been implemented by the same research group, or because they share key components in their computer code (Level 3, *structural uncertainty*). And at last we describe the variability of those model similarity classes (Level 4, *structural uncertainty*). Figure 15.3 depicts this hierarchical structure.

There are several tools available to implement such a statistical model when assuming normal distribution, such as the R packages *lme4* and *nlme* which run linear mixed-effects models (LMMs). However, when dealing with skewed data, the corresponding theory exist for 2-stage mixed models only (e.g Arellano-Valle, Bolfarine and Lachos 2005 and Lin and J. C. Lee 2008). In this work we derive the likelihood function for skewed multilevel frameworks exceeding 2 stages, as we present here for climate multi-model ensembles. The maximisation of the likelihood function is performed directly with an optimiser, which can get very slow when dealing with many data points as several large matrices have to be inverted.

Besides estimating the VCs (i.e. the sources of climate model uncertainty), another big aim in this study is to assess for the uncertainty of the estimates themselves, being the “uncertainty of the uncertainty”. We therefore implement several methods such as parametric- and non-parametric bootstrap techniques and the likelihood ratio and Wald confidence intervals.

Adequacy of Normal Multilevel Model

We check the influence of individual L1 observations on the estimates of the fixed effects and the random effects. This is done by re-fitting the LMM omitting individual observations. Figure 15.11 shows the influence on the estimates of the fixed effects (Cook’s distance) and the influence on the precision of the fixed effects (Covtrace). Excluding any of the *GFDL* simulations, especially *GFDL-CM3*, seems to alter the fixed effects. Two simulations have more influence on the standard error of the estimates than others.

The influence on the variance components is depicted with RVC plots. The exclusion of some of the *CSIRO* models shows some effect on the natural variability (Level 1) and the initial condition variability (Level 2) VC. This means that those *CSIRO* models inflate the estimate of the internal variability more than other simulations. This behaviour could already be seen in Figure 15.4. An even stronger influence shows the *GFDL-CM3* model on the estimate of the *gcm* VC. Excluding this single simulation decreases the estimated variance of the gcm random effect by more than 50%. At the same time the estimated variance of the *sim* random effect would increase by over 50%. A much weaker influence can be seen by other simulations as well.

Exclusion of level 2 observations (gcms and *sim*) reveal influence on the fixed effects (Figure 15.11) of the GCMs *GFDL-CM3* and the *CSIRO* models and of models of family "a" (being *HadGEM* and *ACCESS* models).

To detect changes of the variance components (VCs) when excluding L2 objects we refer to Figure 15.13. We can see a drastic change of the VC *gcm* when excluding the GCM *GFDL-CM3*, which we observed as being an outlier. The estimated variance decreases by more than 60%, and at the same time the variance for sim increases by the same amount. The same effect happens when excluding all *GFDL* simulations (i.e. excluding *sim* "g").

The specified LMM is also very sensitive to the GCM *CSIRO-Mk3-6-0*. Excluding it leads to a drastic decrease of the estimated L1 residual (by almost 50%), which in this case can be interpreted as the internal variability of the GCM (sensitivity of the climate change signal to changes in initial conditions). This means the *CSIRO* GCM exhibits a much stronger variability then the remaining simulations and this strongly influences the overall estimate of the VC.

And at last, one model family (*sim a*) seems to have a strong influence on the *sim* VC estimate. excluding this leads to a strong decrease of more than 60% in that VC estimate. This means, that GCMs within that *sim* appear to have a a much stronger variability than GCMs in other *sim* families. One could re-think if this *sim* category is meaningful.

Climate Change Uncertainty

The expected temperature climate change signal projected by the CMIP5 ensemble with the high emission scenario RCP8.5 is quite heterogeneous with regard to the region and season (Table 15.1). The highest temperature changes can be seen in the Scandinavian Region (SC), where summer temperatures will rise by 4.6 °C and winters seem to get warmer in order of magnitude of 6 °C. The Alpine Region (AL) depicts similar warming signals where the seasonality is the other way around, having a stronger warming in summer than in winter by almost 2 °C. The Iberian Peninsula (IP) shows the same seasonality pattern with higher changes in summer than in winter, but the projected changes are slightly weaker (5.2 °C in summer, 3.1 °C in winter).

In addition to obtaining the expected climate change, our proposed method also estimates the *natural climate variability* - the internal variations of the climate system which happen without anthropogenic influence (Section 2.1). Compared to the estimated climate change signals mentioned above, this variability is very small: On average, the projected climate change signal is around 5 times larger than the standard deviation of natural climate variability. From a climatological point of view, man-made climate warming is highly significant when considering naturally occurring climate variations.

Also in purely statistical terms, the expected climate change signals in every region in each season is highly significant - the expected climate change signal is around 20 times larger than the estimated standard errors. Interestingly these standard errors of the climate change signals are similar across different statistical methods: The multilevel models accounting for model dependencies do not yield smaller estimator uncertainties as we would have expected (see motivating Example 12.1). The same holds also for the other model estimates being quite similar across the different statistical methods.

The VCs obtained by the hierarchical statistical model can be interpreted as sources of the CMIP5 uncertainties. The largest part of the CMIP5 uncertainties can be attributed to the *structural uncertainty* σ_{gcm} and σ_{sim}. σ_{sim} denotes the variability between different classes of model similarity as derived for the CMIP5 by Knutti, Masson and Gettelman 2013. The higher this uncertainty component, the higher the induced climate model inter-dependencies as described in Section 3.2. σ_{gcm} represents the variability of the GCMs stemming from the same model-family. The *initial conditions uncertainty* σ_{run} for the projected temperature climate change is quite small in comparison. The uncertainty component σ_{run} is estimated using different starting conditions of the same GCM.

Our derived multilevel model also estimates the skewness of the climate change probability density function (PDF) mainly induced by outlying GCMs. The point estimates of the skewness parameter seems to be quite high almost everywhere. However, it has to be noted that the skewness coefficient modelled by the skew-normal linear mixed-effects model (SN-LMM) is limited by around $|\gamma_1| \leq 0.995$. Therefore, this model has its limits when modelling highly skewed data, which might be a limitation in this study, as the real PDF might be more skewed.

We inspected four different methods to assess for the statistical significance for each estimate described above, i.e. deriving uncertainty estimates of the statistical estimates. The uncertainty of the estimates for the average climate change signal (fixed effects) is similar when applying the four different methods and highly significant. The statistical uncertainty of the VCs are very small for the initial conditions uncertainty and largest for the model similarity VC. The four methods to estimate the parameter uncertainties yield surprisingly similar results for the VCs. However, for the skewness parameter this is not true anymore, as Wald's CI seems highly inadequate. All other methods yield quite large confidence intervals for this parameter.

A Appendix: Proofs

Proposition 13.13 (Woodbury Identity). For any non-singular matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ the Woodbury matrix identity is

$$(\boldsymbol{A}+\boldsymbol{UBV})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{U}(\boldsymbol{B}^{-1}+\boldsymbol{VA}^{-1}\boldsymbol{U})^{-1}\boldsymbol{VA}^{-1}$$

Proof. To proof the Woodbury Idendity, we show that

$$(\boldsymbol{A}+\boldsymbol{UBV})\left(\boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{U}(\boldsymbol{B}^{-1}+\boldsymbol{VA}^{-1}\boldsymbol{U})^{-1}\boldsymbol{VA}^{-1}\right) = I$$

$$\begin{aligned}
&(\boldsymbol{A}+\boldsymbol{UBV})\left(\boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{U}(\boldsymbol{B}^{-1}+\boldsymbol{VA}^{-1}\boldsymbol{U})^{-1}\boldsymbol{VA}^{-1}\right) = \\
&\quad = \boldsymbol{I} - \boldsymbol{U}(\boldsymbol{B}^{-1}+\boldsymbol{VA}^{-1}\boldsymbol{U})^{-1}\boldsymbol{VA}^{-1} + \boldsymbol{UBVA}^{-1} \\
&\qquad - \boldsymbol{UBVA}^{-1}\boldsymbol{U}(\boldsymbol{B}^{-1}+\boldsymbol{VA}^{-1}\boldsymbol{U})^{-1}\boldsymbol{VA}^{-1} \\
&\quad = \boldsymbol{I} + \boldsymbol{UBVA}^{-1} - \left(\boldsymbol{U}(\boldsymbol{B}^{-1}+\boldsymbol{VA}^{-1}\boldsymbol{U})^{-1} + \boldsymbol{UBVA}^{-1}\boldsymbol{U}(\boldsymbol{B}^{-1}+\boldsymbol{VA}^{-1}\boldsymbol{U})^{-1}\right)\boldsymbol{VA}^{-1} \\
&\quad = \boldsymbol{I} + \boldsymbol{UBVA}^{-1} - \left(\boldsymbol{U} + \boldsymbol{UBVA}^{-1}\boldsymbol{U}\right)\left(\boldsymbol{B}^{-1}+\boldsymbol{VA}^{-1}\boldsymbol{U})^{-1}\right)\boldsymbol{VA}^{-1} \\
&\quad = \boldsymbol{I} + \boldsymbol{UBVA}^{-1} - \boldsymbol{UB}\left(\boldsymbol{B}^{-1}+\boldsymbol{VA}^{-1}\boldsymbol{U}\right)\left(\boldsymbol{B}^{-1}+\boldsymbol{VA}^{-1}\boldsymbol{U})^{-1}\right)\boldsymbol{VA}^{-1} \\
&\quad = \boldsymbol{I} + \boldsymbol{UBVA}^{-1} - \boldsymbol{UBVA}^{-1} \\
&\quad = \boldsymbol{I}
\end{aligned}$$

□

Lemma 13.12. For any $\boldsymbol{Y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\boldsymbol{X} \sim N_d(\boldsymbol{\eta}, \boldsymbol{\Omega})$ the mixture is

$$\begin{aligned}
\phi_p(\boldsymbol{y}|\boldsymbol{\mu}+\boldsymbol{Ax}, \boldsymbol{\Sigma})\phi_d(\boldsymbol{x}|\boldsymbol{\eta}, \boldsymbol{\Omega}) = {}& \phi_p(\boldsymbol{y}|\boldsymbol{\mu}+\boldsymbol{A\eta}, \boldsymbol{\Sigma}+\boldsymbol{A\Omega A}') \\
& \times \phi_d(\boldsymbol{x}|\boldsymbol{\eta}+\boldsymbol{\Lambda A}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}-\boldsymbol{A\eta}), \boldsymbol{\Lambda})
\end{aligned}$$

where

$$\boldsymbol{\Lambda} = (\boldsymbol{\Omega}^{-1}+\boldsymbol{A}'\boldsymbol{\Sigma}^{-1}\boldsymbol{A})^{-1}.$$

Proof. We denote $\boldsymbol{z} = \boldsymbol{y} - \boldsymbol{\mu} - \boldsymbol{A}\boldsymbol{\eta}$ and $\boldsymbol{w} = \boldsymbol{x} - \boldsymbol{\eta}$ and first show that

$$
\begin{aligned}
&(\boldsymbol{z} - \boldsymbol{A}\boldsymbol{w})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{z} - \boldsymbol{A}\boldsymbol{w}) + \boldsymbol{w}'\boldsymbol{\Omega}^{-1}\boldsymbol{w} = \\
&\qquad = \boldsymbol{z}'\boldsymbol{\Sigma}^{-1}\boldsymbol{z} + (\boldsymbol{A}\boldsymbol{w})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{A}\boldsymbol{w}) - \boldsymbol{z}'\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\boldsymbol{w} - \boldsymbol{w}'\boldsymbol{A}'\boldsymbol{\Sigma}^{-1}\boldsymbol{z} + \boldsymbol{w}'\boldsymbol{\Omega}^{-1}\boldsymbol{w} \\
&\qquad = \boldsymbol{z}'\boldsymbol{\Sigma}^{-1}\boldsymbol{z} + \boldsymbol{w}'(\boldsymbol{\Omega}^{-1} + \boldsymbol{A}'\boldsymbol{\Sigma}^{-1}\boldsymbol{A})\boldsymbol{w} - \boldsymbol{z}'\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\boldsymbol{w} - \boldsymbol{w}'\boldsymbol{A}'\boldsymbol{\Sigma}^{-1}\boldsymbol{z} \\
&\qquad = \boldsymbol{z}'\boldsymbol{\Sigma}^{-1}\boldsymbol{z} - \boldsymbol{z}'\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\boldsymbol{\Lambda}\boldsymbol{A}'\boldsymbol{\Sigma}^{-1}\boldsymbol{z} + \boldsymbol{w}'\boldsymbol{\Lambda}^{-1}\boldsymbol{w} - \boldsymbol{z}'\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\boldsymbol{w} - \boldsymbol{w}'\boldsymbol{A}'\boldsymbol{\Sigma}^{-1}\boldsymbol{z} + \\
&\qquad\quad + \boldsymbol{z}'\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\boldsymbol{\Lambda}\boldsymbol{A}'\boldsymbol{\Sigma}^{-1}\boldsymbol{z} \\
&\qquad = \boldsymbol{z}'(\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\boldsymbol{A}\boldsymbol{\Lambda}\boldsymbol{A}'\boldsymbol{\Sigma}^{-1})\boldsymbol{z} + (\boldsymbol{w}' - \boldsymbol{z}'\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\boldsymbol{\Lambda})(\boldsymbol{\Lambda}^{-1}\boldsymbol{w} - \boldsymbol{A}'\boldsymbol{\Sigma}^{-1}\boldsymbol{z}) \\
&\qquad = \boldsymbol{z}'(\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\boldsymbol{A}(\boldsymbol{\Omega}^{-1} + \boldsymbol{A}'\boldsymbol{\Sigma}^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{\Sigma}^{-1})\boldsymbol{z} + \\
&\qquad\quad + (\boldsymbol{w} - \boldsymbol{\Lambda}\boldsymbol{A}'\boldsymbol{\Sigma}^{-1}\boldsymbol{z})'\boldsymbol{\Lambda}^{-1}(\boldsymbol{w} - \boldsymbol{\Lambda}\boldsymbol{A}'\boldsymbol{\Sigma}^{-1}\boldsymbol{z}) \\
&\qquad \overset{\text{Woodbury}}{=} \boldsymbol{z}'(\boldsymbol{\Sigma} + \boldsymbol{A}\boldsymbol{\Omega}\boldsymbol{A}')^{-1}\boldsymbol{z} + (\boldsymbol{w} - \boldsymbol{\Lambda}\boldsymbol{A}'\boldsymbol{\Sigma}^{-1}\boldsymbol{z})'\boldsymbol{\Lambda}^{-1}(\boldsymbol{w} - \boldsymbol{\Lambda}\boldsymbol{A}'\boldsymbol{\Sigma}^{-1}\boldsymbol{z})
\end{aligned}
$$

With this result and because $\boldsymbol{z} - \boldsymbol{A}\boldsymbol{w} = \boldsymbol{y} - \boldsymbol{\mu} - \boldsymbol{A}\boldsymbol{x}$ and as of $|\boldsymbol{\Sigma} + \boldsymbol{A}\boldsymbol{\Omega}\boldsymbol{A}'||\boldsymbol{\Lambda}| = |\boldsymbol{\Sigma}||\boldsymbol{\Omega}|$ we obtain

$$
\begin{aligned}
&\phi_p(\boldsymbol{y}|\boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{x}, \boldsymbol{\Sigma})\phi_d(\boldsymbol{x}|\boldsymbol{\eta}, \boldsymbol{\Omega}) = \\
&\qquad = (2\pi)^{-p/2}\,|\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu} - \boldsymbol{A}\boldsymbol{x})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\mu} - \boldsymbol{A}\boldsymbol{x})\right) \\
&\qquad\quad \cdot (2\pi)^{-d/2}\,|\boldsymbol{\Omega}|^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\eta})'\boldsymbol{\Omega}^{-1}(\boldsymbol{x} - \boldsymbol{\eta})\right) \\
&\qquad = (2\pi)^{-p/2}\,(2\pi)^{-d/2}\,|\boldsymbol{\Sigma}|^{-1/2}|\boldsymbol{\Omega}|^{-1/2} \\
&\qquad\quad \cdot \exp\left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu} - \boldsymbol{A}\boldsymbol{x})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\mu} - \boldsymbol{A}\boldsymbol{x}) + (\boldsymbol{x} - \boldsymbol{\eta})'\boldsymbol{\Omega}^{-1}(\boldsymbol{x} - \boldsymbol{\eta})\right) \\
&\qquad = (2\pi)^{-p/2}\,(2\pi)^{-d/2}\,|\boldsymbol{\Sigma} + \boldsymbol{A}\boldsymbol{\Omega}\boldsymbol{A}'|^{-1/2}|\boldsymbol{\Lambda}|^{-1/2} \\
&\qquad\quad \cdot \exp\left(-\frac{1}{2}\boldsymbol{z}'(\boldsymbol{\Sigma} + \boldsymbol{A}\boldsymbol{\Omega}\boldsymbol{A}')^{-1}\boldsymbol{z} + (\boldsymbol{w} - \boldsymbol{\Lambda}\boldsymbol{A}'\boldsymbol{\Sigma}^{-1}\boldsymbol{z})'\boldsymbol{\Omega}^{-1}(\boldsymbol{w} - \boldsymbol{\Lambda}\boldsymbol{A}'\boldsymbol{\Lambda}^{-1}\boldsymbol{z})\right) \\
&\qquad = \phi_p(\boldsymbol{y}|\boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{\eta}, \boldsymbol{\Sigma} + \boldsymbol{A}\boldsymbol{\Omega}\boldsymbol{A}') \times \phi_d(\boldsymbol{x}|\boldsymbol{\eta} + \boldsymbol{\Lambda}\boldsymbol{A}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\mu} - \boldsymbol{A}\boldsymbol{\eta}), \boldsymbol{\Lambda})
\end{aligned}
$$

□

Proposition 13.6. Let $\boldsymbol{Y} \sim SN_n(\boldsymbol{\lambda})$, then

$$\boldsymbol{Y} = \boldsymbol{\delta}|U| + (\boldsymbol{I}_n - \boldsymbol{\delta}\boldsymbol{\delta}')^{1/2}\boldsymbol{V}, \tag{A.1}$$

where $\boldsymbol{\delta} = \boldsymbol{\lambda}/\sqrt{1 + \boldsymbol{\lambda}'\boldsymbol{\lambda}}$, $U \sim N(0, 1)$ is independent of $\boldsymbol{V} \sim N_n(\boldsymbol{0}, \boldsymbol{I}_n)$.

Proof. Because $U \sim N(0,1)$, it follows that $T = |U| \sim HN(0,1)$ with density $f_T(t) = 2\phi_1(t|0,1)$ for $t \in [0,\infty)$. We further have $\boldsymbol{Y}|T \sim N_n\left(\boldsymbol{\delta} T, (\boldsymbol{I}_n - \boldsymbol{\delta}\boldsymbol{\delta}')\right)$. The density of $\boldsymbol{Y}$ is

$$
\begin{aligned}
f_{\boldsymbol{Y}}(\boldsymbol{y}) &= \int_0^\infty f_{\boldsymbol{Y}|T}(\boldsymbol{y}|t)dt \\
&= \int_0^\infty \phi_n(\boldsymbol{y}|\boldsymbol{0}, \boldsymbol{\delta} t, \boldsymbol{I}_n - \boldsymbol{\delta}\boldsymbol{\delta}') 2\phi_1(t|0,1)dt \\
&\overset{L13.12}{=} \int_0^\infty \phi_n(\boldsymbol{y}|\boldsymbol{0}, (\boldsymbol{I}_n - \boldsymbol{\delta}\boldsymbol{\delta}') + \boldsymbol{\delta}\boldsymbol{\delta}') 2\phi(t|\boldsymbol{\Lambda}\boldsymbol{\delta}'(\boldsymbol{I} - \boldsymbol{\delta}\boldsymbol{\delta}')^{-1}\boldsymbol{y}, \boldsymbol{\Lambda})
\end{aligned}
$$

because

$$
\boldsymbol{\Lambda} = \left(1 + \boldsymbol{\delta}'(\boldsymbol{I} - \boldsymbol{\delta}\boldsymbol{\delta}')^{-1}\boldsymbol{\delta}\right)^{-1} = 1 - \boldsymbol{\delta}'\left((\boldsymbol{I} - \boldsymbol{\delta}\boldsymbol{\delta}') + \boldsymbol{\delta}\boldsymbol{\delta}'\right)^{-1}\boldsymbol{\delta} = 1 - \boldsymbol{\delta}\boldsymbol{\delta}'
$$

and $(\boldsymbol{I} - \boldsymbol{\delta}\boldsymbol{\delta}')^{-1} = \boldsymbol{I} + \boldsymbol{\delta}(1 - \boldsymbol{\delta}'\boldsymbol{\delta})^{-1}\boldsymbol{\delta}'$ it follows that

$$
\begin{aligned}
f_{\boldsymbol{Y}}(\boldsymbol{y}) &= \phi_n(\boldsymbol{y}|\boldsymbol{0}, \boldsymbol{I}_n) \int_0^\infty 2\phi(t|\boldsymbol{\Lambda}\boldsymbol{\delta}'(\boldsymbol{I} - \boldsymbol{\delta}\boldsymbol{\delta}')^{-1}\boldsymbol{y}, \boldsymbol{\Lambda}) \\
&= \phi_n(\boldsymbol{y}) \int_0^\infty 2\phi\left(t|(1 - \boldsymbol{\delta}'\boldsymbol{\delta})\boldsymbol{\delta}'\left(\boldsymbol{I} + \boldsymbol{\delta}(1 - \boldsymbol{\delta}'\boldsymbol{\delta})^{-1}\boldsymbol{\delta}'\right)\boldsymbol{y}, 1 - \boldsymbol{\delta}'\boldsymbol{\delta}\right) \\
&= \phi_n(\boldsymbol{y}) \int_0^\infty 2\phi\left(t|\boldsymbol{\delta}'(1 - \boldsymbol{\delta}'\boldsymbol{\delta} + \boldsymbol{\delta}'\boldsymbol{\delta})\boldsymbol{y}, 1 - \boldsymbol{\delta}'\boldsymbol{\delta}\right) \\
&= \phi_n(\boldsymbol{y}) \int_0^\infty 2\phi\left(t|\boldsymbol{\delta}'\boldsymbol{y}, 1 - \boldsymbol{\delta}'\boldsymbol{\delta}\right).
\end{aligned}
$$

We set $u = \frac{t - \boldsymbol{\delta}\boldsymbol{y}}{\sqrt{(1 - \boldsymbol{\delta}'\boldsymbol{\delta})}}$ leading to $du = dt/\sqrt{1 - \boldsymbol{\delta}'\boldsymbol{\delta}}$ and so

$$
\begin{aligned}
\int_0^\infty \phi\left(t|\boldsymbol{\delta}'\boldsymbol{y}, 1 - \boldsymbol{\delta}'\boldsymbol{\delta}\right) dt &= \frac{1}{\sqrt{2\pi(1 - \boldsymbol{\delta}'\boldsymbol{\delta})}} \int_0^\infty \exp\left(-\frac{(t - \boldsymbol{\delta} y)^2}{2(1 - \boldsymbol{\delta}'\boldsymbol{\delta})}\right) dt \\
&= \frac{1}{\sqrt{2\pi(1 - \boldsymbol{\delta}'\boldsymbol{\delta})}} \int_{-\frac{\boldsymbol{\delta} y}{\sqrt{1 - \boldsymbol{\delta}'\boldsymbol{\delta}}}}^\infty \exp\left(-\frac{u^2}{2}\right) du\sqrt{1 - \boldsymbol{\delta}'\boldsymbol{\delta}} \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{\boldsymbol{\delta} y}{\sqrt{1 - \boldsymbol{\delta}'\boldsymbol{\delta}}}} \exp\left(-\frac{u^2}{2}\right) du \\
&= \Phi\left(\frac{\boldsymbol{\delta}\boldsymbol{y}}{\sqrt{1 - \boldsymbol{\delta}'\boldsymbol{\delta}}}\right) \\
&= \Phi\left(\boldsymbol{\lambda}\boldsymbol{y}\right)
\end{aligned}
$$

with $\boldsymbol{\lambda} = \boldsymbol{\delta}/\sqrt{1 - \boldsymbol{\delta}'\boldsymbol{\delta}}$, yielding

$$
f_{\boldsymbol{Y}}(\boldsymbol{y}) = \phi_n(\boldsymbol{y})\Phi\left(\boldsymbol{\lambda}\boldsymbol{y}\right)
$$

and so $\boldsymbol{Y} \sim SN_n(\boldsymbol{\lambda})$. □

B Appendix: R Functions

This Chapter summarises the most relevant R functions which have been implemented for this thesis. To handle the huge volume of climate data, we have published the R package wux (Mendlik, Heinrich, A. Gobiet et al. 2016), which has already been described in Part II. This package, including the entire documentation and code is publicly available at CRAN (Mendlik, Heinrich and Leuprecht 2015). The R code used in Part III has been published as an online supplementary material to the work of Mendlik and A. Gobiet 2016, and we therefore omit a listing here as well.

We therefore limit the description of the computer code in this chapter to Part IV, which has not been published so far. For the sake of readability we do not show the actual code, but rather describe the functions verbally.

We start from the binary NetCDF data of the CMIP5 multi-model ensemble. The functions used to process these data to a suitable data.frame is described in Section B.1. Inference on these data is performed with the help of the Skew-normal Multilevel Models (SN-LMM) derived in derived in Chapter 13.3. These methods are presented in Section B.2. The different methods to approximate the uncertainty of the resulting estimates, as derived in Chapter 14, is summarised in Section B.3.

B.1 Preprocessing Climate Data

R Code B.1: Function from the wux package to create data.frame from binary NetCDF files (models2wux).

```
models2wux <- function(input.filename, does.plot.subregions = FALSE) {
  ## Creates a dataframe containing climate change signals of climate models
  ## listed in user.input.
  ##
  ## Args:
  ##   input.filename: Filename from input created by the user.
  ##                   The file contains a
  ##                   general.input section with tuning parameters and
  ##                   a model.input section listing the climate models
  ##                   to be processed.
  ##   does.plot.subregions: Boolean. If TRUE, interactive plots will be
  ##                            displayed showing (I) The Shapes of the
  ##                            subregion files (shapefiles, rectangle, ...)
  ##                            (II) a rough pixel map with the cliped areal
  ##                            data of the NetCDF file.
  ##
  ## Returns:
```

Function	Description	Imports package
Preprocessing Climate Data		
models2wux	creates data.frame from binary NetCDF files (Chapter 6)	wux
detrending.preprocess	obtain climate data time-series anomalies (Table 15.5 in Section 15.1)	-
aggregate.ccs	aggregates the time-series anomalies data.frame to a data.frame of climate change signals	-
SN Multilevel Model		
snlmm.varsig.cp.lik	SN multilevel log-likelihood function (Section 15.4)	nlme, sn, lme4pureR, HLMdiag
getSNLMM.varsig.cp	get centered parameter (CP) estimates of the SN multilevel model (Section 13.3)	mnormt
Uncertainty of MLEs		
getHessian.ucty	Wald type $1-\alpha$ CI (Section 14.1)	numDeriv
getLR.ucty	likelihood ratio test (LRT) static to obtain 1-α CI (Section 14.2)	-
getNonParmBoot.ucty	non-parametric bootstrap yielding a 1-α CI (Section 14.3)	parallel
nonparm.boot.resample	hierarchical block-resampling	
getParmBoot.ucty	parametric bootstrap yielding a 1-α CI (Section 14.3)	parallel
parm.boot.resample	parametric block-resampling	sn

Table B.1: Overview of functions implemented for the uncertainty analysis of the CMIP5 MME in Part IV.

```
  ##   Dataframe containing climate change signals or time series for each
      model.
  ##
...
}
```

R Code B.2: Function to obtain climate data time-series anomalies (detrending.preprocess).

```
detrending.preprocess <- function(file, region, seas){
  ## Obtain climate data time-series anomalies. This function reads in
  ## the time series data.frame (years 1971-2099) obtained from the
  ## wux package and performs following steps:
  ## 1) fit overall loess
  ## 2) capture climate change signal from 1971-2000 to 2070-2099
  ## 3) detrending time series 2070-2099 (subtracting loess fit)
  ## 4) center time series 2070-2099 to 0 mean
  ## 5) add climate change signal to centered time series 2070-2099
  ##
  ## Input:
  ##   file: (char) filename of data.frame from the models2wux
  ##         function from the wux package. Regionalized climate
  ##         timeseries data
  ##   region: (char) subregion of rationalised data. Here either
  ##           "AL", "IP" or "SC".
  ##   seas: (char) season of time-series data to be analysed. Here
  ##          either "JJA" or "DJF".
  ##
  ## Returns:
  ##   data.frame of climate data anomalies
...
}
```

R Code B.3: Function to aggregates the time-series anomalies data.frame to a data.frame of climate change signals (aggregate.ccs).

```
aggregate.ccs <- function(cmip){
  ## Aggregates the anomalies data.frame to a data.frame of climate
  ## change signals obtained from the "detrending.preprocess" function
  ## with the "aggregate" function. This way, the hierarchical depth of
  ## the anomaly data.frame is reduced by one, speeding up calculations
  ## for the uncertainty analysis.
  ##
  ## Input:
  ##   cmip: data.frame of climate data anomalies returned from the
  ##         "detrending.preprocess" function.
  ##
  ## Returns:
  ##   Aggregated data.frame with climate change signals for each climate
  ## simulation.
...
}
```

B.2 Skew-normal Multilevel Model

R Code B.4: Function of skew-normal multilevel log-likelihood function described in Section 15.4 to fit the regionalized CMIP5 multimodel ensemble.

```
snlmm.varsig.cp.lik <- function(parms.cp, yis, Xis, Zis, RE.inds,
                                       n.ranef = 3, fix.gamma = NULL,
                                       verbose = FALSE){
  ## Implementation of skew-normal multilevel log-likelihood function
  ##  to fit the regionalized CMIP5 multi-model
  ## ensemble. The likelihood is defined in the DP space, but this
  ## function takes CP parameters as input for better convergence and
  ## better interpretability and transforms them into the DP
  ## space. The function performs following tasks:
  ## 1) Reading in the centered parameters (CP) and transform them to
  ##    DP as described in section ().
  ## 2) Loop over each model-similarity class member i (here 17
  ##    members) to calculate the components of the log-likelihood
  ##    function. The multivariate normal CDF probabilities are
  ##    obtained using the "pmvnorm" function from the "mnormt"
  ##    package. Matrix inversion takes the most time here.
  ## 3) Return log-likelihood function.
  ##
  ## Assumptions: - natural variability const for each modsim (heterogeneity
      )
  ##              - internal variability constant
  ##              - gcm-modsim variability constant
  ##
  ## Input:
  ##   parms.cp: vector of centered parameters (CP).
  ##   yis: list of vectors of response values for each
  ##        model-similarity class.
  ##   Xis: list of fixed-effects design matrices for each
  ##        model-similarity class.
  ##   Zis: list of random-effects design matrices for each
  ##        model-similarity class.
  ##  RE.inds: list for indices of the random effects levels. This is
  ##           necessary due to the multilevel nature of the model.
  ##  n.ranef: depth of the multilevel model (here 3 random effect
  ##          levels).
  ##   fix.gamma:  (numerical) fix the skewness parameter to a certain
  ##               value (needed for LRT statistics). Default is NULL.
  ##   verbose: (boolean) More detailed output while optimising the
  ##               likelihood function. Mainly for debugging purposes.
  ##
  ## Returns:
  ##   Deviance of the likelihood (-2 times the log-likelihood value)
  ##   for given CP parameter values.
...
}
```

R Code B.5: Function to get centered parameter (CP) estimates of the skew-normal multilevel model for the CMIP5 multimodel ensemble with heterogenous model-similarity variance components.

```
getSNLMM.varsig.cp <- function(cmip, fix.gamma = NULL, verbose = FALSE){
  ## Get centered parameter (CP) estimates of the skew-normal
  ## multilevel model for the CMIP5 multi-model ensemble with
  ## heterogenous model-similarity variance components. This function
  ## performs following tasks:
  ## 1) First obtain starting values of the fixed effect and the VCs
  ##    by fitting the data using a LMM ("lme" function from the
  ##    "nlme" package)
  ## 2) Obtain starting values for the skewness parameter by fitting
  ##    the LMM residuals (Level 2) with a skew normal modal with the
  ##    "selm" function from the "sn" package.
  ## 3) For each model-similarity class, obtain the design matrices
  ##    for the fixed effects (X) and the random effects (Z). Z is
  ##    obtained with the function "mkRanefStructures" from the
  ##    "lme4pureR" package. Obtain the vector of response variable y,
  ##    in this case this is the temperature time-series of all models
  ##    from the same model similarity class.
  ## 4) Optimize the likelihood function "snlmm.varsig.cp.lik" with
  ##    the "optim" function using the "L-BFGS-B" method to define the
  ##    0 boundaries for the VCs and the (-.995,.995) boundaries for
  ##    the skewness coefficient gamma. The likelihood function is fed
  ##    with the starting values, the design matrices X, Z and the
  ##    response vectors y.
  ##
  ## Input:
  ##   cmip:       data.frame of the climate data anomalies obtained from
  ##               the function "detrending.preproc".
  ##   fix.gamma:  (numerical) fix the skewness parameter to a certain
  ##               value (needed for LRT statistics). Default is NULL.
  ##   verbose: (boolean) More detailed output while optimizing the
  ##               likelihood function. Mainly for debugging purposes.
  ##
  ## Returns: Vector of centered parameter (CP) estimates of the skew
  ##          normal multilevel model.
...
}
```

B.3 Uncertainty of MLEs

R Code B.6: Function obtain the Wald type 1-α CI.

```
getHessian.ucty  <- function(cmip.ccs){
  ## Obtain the Wald type 1-alpha CI. Approximates the standard errors
  ## of the CP parameter MLE obtained from the "getSNLMM.cp.ccs"
  ## function. It first fits the MLEs and then calculates the hessian
  ## matrix of the log-likelihood function given the MLEs using the
  ## "hessian" function from the package "numDeriv". The standard error
  ## estimates are then obtained by taking the square root of the
  ## diagonal entries of the Hessian matrix. The symmetrical 1-alpha CIs
  ## are obtained by
  ##   MLE +/- 1.96 *hessian
  ##
  ## Input:
  ##   cmip.ccs: data.frame of climate change signals of the CMIP5
  ##             multimodel ensembles obtained from the function
  ##             "aggregate.ccs".
  ##
  ## Returns:
  ##   Data.frame of 1-alpha confidence intervals with second-order
  ##   standard error approximations for all MLEs.
...
}
```

R Code B.7: Function which approximates the 1-α CI for the SN-LMM estimates using the LR test static.

```
getLR.ucty <- function(cmip.ccs){
  ## Approximates the 1-alpha CI for the SN-LMM estimates using the LR
  ## test static. For each MLE the likelihood function is evaluated at
  ## all feasible values. For each value the LRT statistic is
  ## performed. The 1-alpha CI is obtained by discarding all parameter
  ## values for which the LRT exceeds the "qchisq(.95, 1)" value.
  ##
  ## Input:
  ##   cmip.ccs: data.frame of climate change signals of the CMIP5
  ##             multi-model ensembles obtained from the function
  ##             "aggregate.ccs".
  ##
  ## Returns:
  ##   Data.frame of 1-alpha confidence intervals for all MLEs using the
  ##   LRT test statistic.
...
}
```

R Code B.8: Function to obtain the nonparametric bootstrap 1-α conficence interval for each parameter of the SN-LMM model using 1000 resamples.

```
getNonParmBoot.ucty <- function(cmip.ccs){
  ## Gets the non-parametric bootstrap 1-alpha confidence interval for
  ## each parameter of the SN-LMM model using 1000 resamples. This
  ## function uses the function "parSapply" from the package "parallel"
  ## to parallelize the resampling process on several cores. This
  ## function draws a bootstrap sample using thef unction
  ## "nonparm.boot.resample" and then fits the SN-LMM with the new
  ## data.frame. The resampling is done recursively, so each hierarchy
  ## level (modsim, gcm and run) is resampled (block-bootstrapping). From
  ## the 1000 MLEs from the bootstrap samples, the q0.025 and q0.975 are
  ## taken as the lower and upper CI bound.
  ##
  ## Input:
  ##   cmip.ccs: data.frame of climate change signals of the CMIP5
  ##             multi-model ensembles obtained from the function
  ##             "aggregate.ccs".
  ##
  ## Returns:
  ##   Data.frame of 1-alpha confidence intervals obtained by
  ##   non-parametric block bootstrapping.
...
}
```

R Code B.9: Function to get the non-parametric bootstrap 1-α conficence interval for each parameter of the SN-LMM model using 1000 resamples.

```
nonparm.boot.resample <- function(dat, cluster, replace) {
  ## Recursively resamples the nested factors of a data.frame and
  ## creates a new data.frame with identical dimensions. This function
  ## is necessary for non-parametric block bootstrapping.
  ##
  ## Input:
  ##   dat: data.frame from which to resample the factors.
  ##   cluster: character vector specifying the factors to be
  ##            resampled. This argument defines the hierarchy of the
  ##            data.frame from top to bottom. In this case by
  ##            c("modsim", "gcm", "run").
  ##   replace: boolean vector to indicate weather or not sampling
  ##            should be performed with replacement on the individual
  ##            hierarchy level. In this case we perform resampling
  ##            with replacement on all levels c("TRUE", "TRUE",
  ##            "TRUE").
  ##
  ## Returns:
  ##   A data.frame with resampled factors. The observations remain
  ##   unchanged.
  ##
  ## This code has been adapted from the code published at
  ## http://biostat.mc.vanderbilt.edu/wiki/Main/HowToBootstrapCorrelatedData
...
}
```

R Code B.10: Function to obtain the 1-α CI using parametric bootstrap using the SN-LMM.

```
getParmBoot.ucty <- function(cmip.ccs){
  ## Obtains the 1-alpha CI using parametric bootstrap using the
  ## SN-LMM. As for the non-parametric case, this function uses
  ## parallelization with the "parallel" package to use multiple
  ## cores. This function first fits the original data.frame to obtain
  ## the MLEs. Using those MLEs, the function "parm.boot.resample"
  ## is called 1000 times to simulate new observational data. Then the
  ## new data.frames are re-fit to obtain a sample of MLEs. The 1-alpha
  ## CI is obtained in the same manner as for the non-parametric case by
  ## taking the q0.025 and q0.975 quantiles of the MLE sample.
  ##
  ## Input:
  ##   cmip.ccs: data.frame of climate change signals of the CMIP5
  ##             multi-model ensembles obtained from the function
  ##             "aggregate.ccs".
  ##
  ## Returns:
  ##   Data.frame of 1-alpha confidence intervals obtained by
  ##   parametric bootstrapping.
...
}
```

R Code B.11: Function for parametric block-bootstrap resampling.

```
parm.boot.resample <- function(parm.mle, cmip.ccs){
  ## Hierarchically simulates new observational values for a given
  ## data.frame using pre-specified distributional assumptions from a
  ## statistical model based on MLEs. In this case, the average
  ## climate change signal for the top hierarchy "modsim" is simulated
  ## using a normal distribution. Then for each mid-hierarchy level of
  ## "gcm" skew-normal data is created and added to the top
  ## hierarchy. At last, for each lowest-level "run" data is created
  ## using a normal distribution and added to the hierarchy structures
  ## above.
  ##
  ## Input:
  ## parm.mle: Vector of MLEs obtained by the SN-LMM.
  ## cmip.ccs: Original data.frame which has been used to get the
  ##           MLEs. This data.frame will be refilled with the
  ##           simulated values.
  ## Returns:
  ##   Data.frame filled with simulated data from a hierarchy of
  ##   parametric distributions.
...
}
```

Bibliography

Abramowitz, G. and C. H. Bishop (2015). 'Climate model dependence and the ensemble dependence transformation of CMIP projections'. In: *Journal of Climate* 28.6, pp. 2332–2348. DOI: 10.1175/JCLI-D-14-00364.1 (cit. on pp. 19, 21, 22, 29).

Abramowitz, G. (2010). 'Model independence in multi-model ensemble prediction'. In: *Australian Meteorological and Oceanographic Journal* 59, pp. 3–6 (cit. on p. 21).

Abramowitz, G. and H. Gupta (2008). 'Toward a model space and model independence metric'. In: *Geophysical Research Letters* 35. DOI: 10.1029/2007GL032834 (cit. on pp. 29, 69).

Annan, J. D. and J. C. Hargreaves (2010). 'Reliability of the CMIP3 ensemble'. In: *Geophysical Research Letters* 37.2. DOI: 10.1029/2009GL041994 (cit. on p. 20).

Arellano-Valle, R., H. Bolfarine and V. Lachos (2005). 'Skew-normal linear mixed models'. In: *Journal of Data Science* 3.4, pp. 415–438 (cit. on pp. 93, 97, 100, 140).

Arellano-Valle, R. and A. Azzalini (2008). 'The centred parametrization for the multivariate skew-normal distribution'. In: *Journal of Multivariate Analysis* 99.7, pp. 1362–1382 (cit. on pp. 93, 95).

Azzalini, A. (1985). 'A class of distributions which includes the normal ones'. In: *Scandinavian Journal of Statistics* 12.2, pp. 171–178 (cit. on pp. 93, 94).

Azzalini, A. and A. Dalla Valle (1996). 'The multivariate skew-normal distribution'. In: *Biometrika* 83.4, pp. 715–726. DOI: 10.1093/biomet/83.4.715 (cit. on pp. 93, 100).

Bates, D., M. Maechler, B. M. Bolker and S. Walker (2014). *lme4: Linear mixed-effects models using Eigen and S4*. ArXiv e-print; submitted to *Journal of Statistical Software* (cit. on p. 49).

Bishop, C. H. and G. Abramowitz (2012). 'Climate model dependence and the replicate Earth paradigm'. In: *Climate Dynamics*, pp. 1–16 (cit. on pp. 20–22, 28–30, 69).

Buser, C. M., H. R. Künsch and A. Weber (2010). 'Biases and uncertainty in climate projections'. en. In: *Scandinavian Journal of Statistics* 37.2, pp. 179–199. DOI: 10.1111/j.1467-9469.2009.00686.x (cit. on p. 21).

Byrd, R. H., P. Lu, J. Nocedal and C. Zhu (1995). 'A limited memory algorithm for bound constrained optimization'. In: *SIAM Journal on Scientific Computing* 16.5, pp. 1190–1208 (cit. on p. 132).

Cannon, A. J. (2015). 'Selecting GCM scenarios that span the range of changes in a multimodel ensemble: Application to CMIP5 climate extremes indices'. In: *Journal of Climate* 28.3, pp. 1260–1267. DOI: 10.1175/JCLI-D-14-00636.1 (cit. on pp. 17, 27, 70).

Casella, G. and R. Berger (2002). *Statistical Inference*. Duxbury advanced series in statistics and decision sciences. Thomson Learning (cit. on p. 107).

Chandler, R. E. (2013). 'Exploiting strength, discounting weakness: combining information from multiple climate simulators'. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 371.1991. DOI: 10.1098/rsta.2012.0388 (cit. on p. 20).

Christensen, J. H. and O. B. Christensen (2007). 'A summary of the PRUDENCE model projections of changes in European climate by the end of this century'. English. In: *Climatic Change* 81.1, pp. 7–30. DOI: 10.1007/s10584-006-9210-7 (cit. on pp. 41, 63, 111).

Collins, M., R. Knutti, J. Arblaster, J.-L. Dufresne, T. Fichefet, P. Friedlingstein, X. Gao, W. Gutowski, T. Johns, G. Krinner, M. Shongwe, C. Tebaldi, A. Weaver and M. Wehner (2013). 'Long-term Climate Change: Projections, Commitments and Irreversibility'. In: *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Ed. by T. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P. Midgley. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press. Chap. 12, pp. 1029–1136. DOI: 10.1017/CBO9781107415324.024 (cit. on pp. 11, 12, 114).

Collins, M., B. S. F. Tett and C. Cooper (2001). 'The internal climate variability of HadCM3, a version of the Hadley Centre coupled model without flux adjustments'. In: *Climate Dynamics* 17.1, pp. 61–81. DOI: 10.1007/s003820000094 (cit. on pp. 15, 30).

Collins, M., B. B. B. Booth, G. R. Harris, J. M. Murphy, D. M. H. Sexton and M. J. Webb (2006). 'Towards quantifying uncertainty in transient climate change'. In: *Climate Dynamics* 27.2-3, pp. 127–147. DOI: 10.1007/s00382-006-0121-0 (cit. on p. 17).

Déqué, M., D. P. Rowell, D. Lüthi, F. Giorgi, J. H. Christensen, B. Rockel, D. Jacob, E. Kjellström, M. Castro and B. Hurk (2007). 'An intercomparison of regional climate simulations for Europe: assessing uncertainties in model projections'. In: *Clim. Change* 81.S1, pp. 53–70 (cit. on pp. 18, 47, 53).

Déqué, M., S. Somot, E. Sanchez-Gomez, C. M. Goodess, D. Jacob, G. Lenderink and O. B. Christensen (2011). 'The spread amongst ENSEMBLES regional scenarios: regional climate models, driving general circulation models and interannual variability'. In: *Clim. Dyn.* (Cit. on pp. 18, 30, 47).

Dillane, D. (2005). 'Deletion Diagnostics for the Linear Mixed Model'. PhD thesis. Trinity College, Dublin (cit. on p. 123).

Efron, B. and R. J. Tibshirani (1994). *An introduction to the bootstrap*. CRC press (cit. on p. 109).

Evans, J. P., F. Ji, C. Lee, P. Smith, D. Argüeso and L. Fita (2013). 'A regional climate modelling projection ensemble experiment - NARCliM'. In: *Geoscientific Model Development Discussions* 6.3, pp. 5117–5139. DOI: 10.5194/gmdd-6-5117-2013 (cit. on pp. 18, 28).

Evans, J. P., F. Ji, G. Abramowitz and M. Ekström (2013). 'Optimally choosing small ensemble members to produce robust climate simulations'. In: *Environmental Research Letters* 8.4, p. 044050 (cit. on p. 22).

Finger, D., G. Heinrich, A. Gobiet and A. Bauder (2012). 'Projections of future water resources and their uncertainty in a glacierized catchment in the Swiss Alps and the subsequent effects on hydropower production during the 21st century'. In: *Water Resources Research* 48.2, n/a–n/a. DOI: 10.1029/2011WR010733 (cit. on p. 28).

Fischer, A. M., A. P. Weigel, C. M. Buser, R. Knutti, H. R. Künsch, M. A. Liniger, C. Schür and C. Appenzeller (2012). 'Climate change projections for Switzerland based on a Bayesian multi-model approach'. In: *International Journal of Climatology* 32.15, pp. 2348–2371. DOI: 10.1002/joc.3396 (cit. on p. 21).

Hartmann, D., A. Klein Tank, M. Rusticucci, L. Alexander, S. Brönnimann, Y. Charabi, F. Dentener, E. Dlugokencky, D. Easterling, A. Kaplan, B. Soden, P. Thorne, M. Wild and P. Zhai (2013). 'Observations: Atmosphere and Surface'. In: *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Ed. by T. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P. Midgley. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press. Chap. 2, pp. 159–254. DOI: 10.1017/CBO9781107415324.008 (cit. on p. 8).

Hawkins, E. and R. Sutton (2009). 'The potential to narrow uncertainty in regional climate predictions'. In: *Bulletin of the American Meteorological Society* 90.8, pp. 1095–1107. DOI: 10.1175/2009BAMS2607.1 (cit. on pp. 15, 16, 30).

Heinrich, G., A. Gobiet and T. Mendlik (2014). 'Extended regional climate model projections for Europe until the mid-twentyfirst century: combining ENSEMBLES and CMIP3'. In: *Climate Dynamics* 42.1-2, pp. 521–535 (cit. on pp. 17, 19, 30, 47, 53, 69).

Henderson, C. R. (1982). 'Analysis of covariance in the mixed model: Higher-level, nonhomogeneous, and random regressions'. In: *Biometrics* 38.3, pp. 623–640 (cit. on p. 78).

Hewitt, C. D. and D. J. Griggs (2004). 'Ensembles-based predictions of climate changes and their impacts (ENSEMBLES)'. In: *Eos* 85.52, p. 566 (cit. on pp. 13, 18).

Huth, R., C. Beck, A. Philipp, M. Demuzere, Z. Ustrnul, M. Cahynová, J. Kyselý and O. E. Tveito (2008). 'Classifications of atmospheric circulation patterns: recent advances and applications.' In: *Annals Of The New York Academy Of Sciences* 1146.1, pp. 105–152 (cit. on p. 57).

IPCC (2013). *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press, p. 1535. DOI: 10.1017/CBO9781107415324 (cit. on pp. 7, 11, 13).

IPCC-TGICA, A. (2007). *General Guidelines on the Use of Scenario Data for Climate Impact and Adaptation Assessment. Version 2.* Tech. rep., p. 66 (cit. on pp. 27, 28, 69).

Jacob, D., J. Petersen, B. Eggert, A. Alias, O. Christensen, L. Bouwer, A. Braun, A. Colette, M. Déqué, G. Georgievski, E. Georgopoulou, A. Gobiet, L. Menut, G. Nikulin, A. Haensler, N. Hempelmann, C. Jones, K. Keuler, S. Kovats, N. Kröner, S. Kotlarski, A. Kriegsmann, E. Martin, E. Meijgaard, C. Moseley, S. Pfeifer, S. Preuschmann, C. Radermacher, K. Radtke, D. Rechid, M. Rounsevell, P. Samuelsson, S. Somot, J.-F. Soussana, C. Teichmann, R. Valentini, R. Vautard, B. Weber and P. Yiou (2013). 'EURO-CORDEX: new high-resolution climate change projections for European impact research'. English. In: *Regional Environmental Change*, pp. 1–16. DOI: 10.1007/s10113-013-0499-2 (cit. on pp. 13, 18, 47).

Jolliffe, I. T. (2002). *Principal Component Analysis*. Second. Springer (cit. on p. 57).

Jury, M. W., A. F. Prein, H. Truhetz and A. Gobiet (2015). 'Evaluation of CMIP5 models in the context of dynamical downscaling over Europe'. In: *Journal of Climate* 28.14, pp. 5575–5582. DOI: 10.1175/JCLI-D-14-00430.1 (cit. on p. 28).

Knutti, R. (2010). 'The end of model democracy?' In: *Climatic Change* 102 (3-4), pp. 395–404. DOI: 10.1007/s10584-010-9800-2 (cit. on pp. 20, 21).

Knutti, R., G. Abramowitz, M. Collins, V. Eyring, P. J. Gleckler, B. Hewitson and L. Mearns (2010). 'Good practice guidance paper on assessing and combining multi model climate projections'. In: *IPCC Expert Meeting on Assessing and Combining Multi Model Climate Projections*. Ed. by T. Stocker, D. Qin, G.-K. Plattner, M. Tignor and P. Midgley, p. 1 (cit. on pp. 20, 21, 28, 69).

Knutti, R., R. Furrer, C. Tebaldi, J. Cermak and G. A. Meehl (2010). 'Challenges in combining projections from multiple climate models'. In: *Journal of Climate* 23 (10). DOI: 10.1175/2009JCLI3361.1 (cit. on pp. 21, 22, 24, 25, 28, 53).

Knutti, R., D. Masson and A. Gettelman (2013). 'Climate model genealogy: Generation CMIP5 and how we got there'. In: *Geophys. Res. Lett.* DOI: 10.1002/grl.50256 (cit. on pp. 22–24, 29, 30, 73, 111, 114, 142).

Lin, T. I. and J. C. Lee (2008). 'Estimation and prediction in linear mixed models with skew-normal random effects for longitudinal data'. In: *Statistics in medicine* 27.9, pp. 1490–1507 (cit. on pp. 97, 100, 140).

Linden, P. van der and J. F. B. Mitchell (2009). *ENSEMBLES: Climate Change and its Impacts: Summary of research and results from the ENSEMBLES project*. Met Office Hadley Centre (cit. on pp. 45, 47).

Loy, A. and H. Hofmann (2014). 'HLMdiag: A suite of diagnostics for hierarchical linear models in R'. In: *Journal of Statistical Software* 56.1, pp. 1–28. DOI: 10.18637/jss.v056.i05 (cit. on pp. 118, 119).

Maraun, D., F. Wetterhall, A. M. Ireson, R. E. Chandler, E. J. Kendon, M. Widmann, S. Brienen, H. W. Rust, T. Sauter, M. Themeßl, V. K. C. Venema, K. P. Chun, C. M. Goodess, R. G. Jones, C. Onof, M. Vrac and I. Thiele-Eich (2010). 'Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user'. In: *Rev. Geophys.* 48.3, pp. 1–34. DOI: 10.1029/2009RG000314 (cit. on p. 53).

Masson, D. and R. Knutti (2011). 'Climate model genealogy'. In: *Geophysical Research Letters* 38.8. DOI: 10.1029/2011GL046864 (cit. on pp. 20, 22, 29, 73).

Mays, N. and C. Pope (1995). 'Rigour and qualitative research.' In: *BMJ : British Medical Journal* 311.6997, pp. 109–112 (cit. on p. 28).

McSweeney, C. F., R. G. Jones and B. B. B. Booth (2012). 'Selecting ensemble members to provide regional climate change information'. In: *J. Climate* 25.20, pp. 7100–7121. DOI: 10.1175/jcli-d-11-00526.1 (cit. on p. 28).

Mearns, L. O., F. Giorgi, P. Whetton, D. Pabon, M. Hulme and M. Lal (2003). *Guidelines for Use of Climate Scenarios Developed from Regional Climate Model Experiments*. Tech. rep. IPCC (cit. on pp. 13, 14).

Mearns, L. O. (2010). 'The drama of uncertainty'. In: *Climatic Change* 100 (1), pp. 77–85. DOI: 10.1007/s10584-010-9841-6 (cit. on p. 21).

Mearns, L. O., W. Gutowski, R. Jones, R. Leung, S. McGinnis, A. Nunes and Y. Qian (2009). "A regional climate change assessment program for North America". In: *Eos, Transactions American Geophysical Union* 90.36, pp. 311–311. DOI: 10.1029/2009EO360002 (cit. on p. 18).

Mendlik, T. and A. Gobiet (2016). 'Selecting climate simulations for impact studies based on multivariate patterns of climate change'. In: *Climatic Change* 135.3, pp. 381–393. DOI: 10.1007/s10584-015-1582-0 (cit. on pp. 3, 4, 17, 22, 29, 47, 58, 59, 65–67, 147).

Mendlik, T., G. Heinrich, A. Gobiet and A. Leuprecht (2016). 'From climate simulations to statistics - Introducing the wux package'. In: *Austrian Journal of Statistics* 45, pp. 81–96. DOI: 10.17713/ajs.v45i1.98 (cit. on pp. 3, 4, 25, 34, 35, 43, 48, 51, 64, 147).

Mendlik, T., G. Heinrich and A. Leuprecht (2015). *wux: Wegener Center Climate Uncertainty Explorer.* In collab. with A. Gobiet. R package version 2.1-0 (cit. on pp. 3, 147).

Murdock, T. and D. Spittlehouse (2011). *Selecting and Using Climate Change Scenarios for British Columbia.* Tech. rep. Pacific Climate Impacts Consortium, University of Victoria, Victoria BC, p. 39 (cit. on p. 27).

Murphy, J. M., B. B. Booth, M. Collins, G. R. Harris, D. M. Sexton and M. J. Webb (2007). 'A methodology for probabilistic predictions of regional climate change from perturbed physics ensembles'. In: *Royal Society of London Philosophical Transactions Series A* 365, pp. 1993–2028. DOI: 10.1098/rsta.2007.2077 (cit. on p. 17).

Nakicenovic, N., J. Alcamo and G. Davis (2000). *IPCC Special Report on Emissions Scenarios.* Cambridge University Press, Cambridge, United Kingdom and New York (cit. on p. 63).

Oldenborgh, G. J. van, S. Drijfhout, A. van Ulden, R. Haarsma, A. Sterl, C. Severijns, W. Hazeleger and H. Dijkstra (2009). 'Western Europe is warming much faster than expected'. In: *Clim. Past* 5.1, pp. 1–12. DOI: 10.5194/cp-5-1-2009 (cit. on p. 26).

Pennell, C. and T. Reichler (2010). 'On the effective number of climate models'. In: *J. Climate* 24.9, pp. 2358–2367. DOI: 10.1175/2010JCLI3814.1 (cit. on pp. 21, 29, 69).

Pirtle, Z., R. Meyer and A. Hamilton (2010). 'What does it mean when climate models agree? A case for assessing independence among general circulation models'. In: *Environmental Science* 13.5, pp. 351–361. DOI: 10.1016/j.envsci.2010.04.004 (cit. on pp. 21, 28).

Prein, A. F., A. Gobiet and H. Truhetz (2011). ‘Analysis of uncertainty in large scale climate change projections over Europe’. In: *Meteorol. Z.* 20.4, pp. 383–395. DOI: 10.1127/0941-2948/2011/0286 (cit. on p. 47).

Ravazzani, G., M. Ghilardi, T. Mendlik, A. Gobiet, C. Corbari and M. Mancini (2014). “Investigation of climate change impact on water resources for an Alpine basin in northern Italy: Implications for evapotranspiration modeling complexity”. In: *PLoS ONE* 9.10. Ed. by J. M. Dias, e109053. DOI: 10.1371/journal.pone.0109053 (cit. on p. 13).

Rougier, J., M. Goldstein and L. House (2013). ‘Second-order exchangeability analysis for multimodel ensembles’. In: *Journal of the American Statistical Association* 108.503, pp. 852–863. DOI: 10.1080/01621459.2013.802963 (cit. on p. 20).

Saha, K. (2008). *The Earth’s Atmosphere: Its Physics and Dynamics.* Google-Books-ID: Jlb5PtwpkI8C. Springer Science & Business Media. 374 pp. (cit. on p. 9).

Salathe, E. P., P. W. Mote and M. W. Wiley (2007). ‘Review of scenario selection and downscaling methods for the assessment of climate change impacts on hydrology in the United States pacific northwest’. In: *International Journal of Climatology* 27.12, pp. 1611–1621. DOI: 10.1002/joc.1540 (cit. on p. 27).

Sarkar, D. (2008). *Lattice: Multivariate Data Visualization with R.* ISBN 978-0-387-75968-5. New York: Springer (cit. on p. 49).

Smith, J. B. and M. Hulme (1998). ‘Climate change scenarios’. In: *UNEP handbook on methods for climate change impact assessment and adaptation studies.* Ed. by J. Feenstra, I. Burton, J. Smith and R. Tol. United Nations Environment Programme, Nairobi, Kenya and Institute for Environmental Studies, Amsterdam, pages (cit. on p. 27).

Solomon, S. (2007). *Climate change 2007-the physical science basis: Working group I contribution to the fourth assessment report of the IPCC.* Vol. 4. Cambridge University Press (cit. on pp. 9, 16).

Steinschneider, S., R. McCrary, L. O. Mearns and C. Brown (2015). ‘The effects of climate model similarity on probabilistic climate projections and the implications for local, risk-based adaptation planning’. en. In: *Geophysical Research Letters* 42.12, 2015GL064529. DOI: 10.1002/2015GL064529 (cit. on pp. 22, 29, 30).

Storch, H. and F. Zwiers (2013). ‘Testing ensembles of climate change scenarios for “statistical significance”’. In: *Climatic Change* 117 (1-2), pp. 1–9. DOI: 10.1007/s10584-012-0551-0 (cit. on p. 21).

Taylor, K. E., R. J. Stouffer and G. A. Meehl (2012). ‘An overview of CMIP5 and the experiment design’. In: *Bulletin of the American Meteorological Society* 93.4, pp. 485–498. DOI: 10.1175/BAMS-D-11-00094.1 (cit. on pp. 13, 39).

Tebaldi, C. and R. Knutti (2007). ‘The use of the multi-model ensemble in probabilistic climate projections’. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 365.1857, pp. 2053–2075. DOI: 10.1098/rsta.2007.2076 (cit. on pp. 3, 16, 17, 21, 24).

Tebaldi, C., R. L. Smith, D. Nychka and L. O. Mearns (2005). ‘Quantifying uncertainty in projections of regional climate change: A Bayesian approach to the analysis of multi-model ensembles’. In: *Journal of Climate* 18, pp. 1524–1540 (cit. on pp. 18, 21, 29).

Themeßl, M., A. Gobiet and A. Leuprecht (2011). ‘Empirical-statistical downscaling and error correction of daily precipitation from regional climate models’. In: *International Journal of Climatology* 31.10, pp. 1530–1544. DOI: 10.1002/joc.2168 (cit. on p. 53).

Vautard, R., A. Gobiet, S. Sobolowski, E. Kjellström, A. Stegehuis, P. Watkiss, T. Mendlik, O. Landgren, G. Nikulin, C. Teichmann and D. Jacob (2014). ‘The European climate under a 2 °C global warming’. In: *Environmental Research Letters* 9.3, p. 034006 (cit. on p. 63).

Whetton, P., K. Hennessy, J. Clarke, K. McInnes and D. Kent (2012). ‘Use of representative climate futures in impact and adaptation assessment’. In: *Climatic Change* 115 (3). 10.1007/s10584-012-0471-z, pp. 433–442 (cit. on pp. 17, 27).

Zender, C. S. (2008). ‘Analysis of self-describing gridded geoscience data with NetCDF operators (NCO)’. In: *Environ. Modell. Softw.* 23.10, pp. 1338–1342. DOI: 10.1016/j.envsoft.2008.03.004 (cit. on p. 26).

Zubler, E. M., A. M. Fischer, F. Fröb and M. A. Liniger (2015). ‘Climate change signals of CMIP5 general circulation models over the Alps – impact of model selection’. en. In: *International Journal of Climatology*. DOI: 10.1002/joc.4538 (cit. on pp. 17, 22, 29, 73).