

Dissertation zur Erlangung des  
naturwissenschaftlichen Doktorgrades der  
Bayerischen Julius-Maximilians-Universität  
Würzburg

**Entwicklung und Vergleich von  
Gewichtungsmetriken zur Analyse  
probabilistischer Klimaprojektionen  
aktueller Modellensembles**

vorgelegt von  
Christoph Ring  
aus Dortmund

Würzburg 2017



Eingereicht am:

5. September 2017

1. Gutachter:

Prof. Dr. Heiko Paeth

2. Gutachter:

Prof. Dr. Jucundus Jacobeit

der Dissertation

1. Prüfer:

Prof. Dr. Heiko Paeth

2. Prüfer:

Prof. Dr. Jürgen Rauh

der mündlichen Prüfung

Tag der mündlichen Prüfung:

24. Januar 2018

Doktorurkunde ausgehändigt am:



*'Try not!  
Do or do not.  
There is no try.'*

- Master Yoda  
The Empire Strikes Back  
(Lucasfilm Ltd)



# Zusammenfassung

Der anthropogene Klimawandel ist eine der größten Herausforderungen des 21. Jahrhunderts. Eine Hauptschwierigkeit liegt dabei in der Unsicherheit bezüglich der regionalen Änderung von Niederschlag und Temperatur. Hierdurch wird die Entwicklung geeigneter Anpassungsstrategien deutlich erschwert.

In der vorliegenden Arbeit werden vier Evaluationsansätze mit insgesamt 13 Metriken für aktuelle globale (zwei Generationen) und regionale Klimamodelle entwickelt und verglichen, um anschließend eine Analyse der Projektionsunsicherheit vorzunehmen. Basierend auf den erstellten Modellbewertungen werden durch Gewichtung Aussagen über den Unsicherheitsbereich des zukünftigen Klimas getroffen. Die Evaluation der Modelle wird im Mittelmeerraum sowie in acht Unterregionen durchgeführt. Dabei wird der saisonale Trend von Temperatur und Niederschlag im Evaluationszeitraum 1960–2009 ausgewertet. Zusätzlich wird für bestimmte Metriken jeweils das klimatologische Mittel oder die harmonischen Zeitreiheneigenschaften evaluiert. Abschließend werden zum Test der Übertragbarkeit der Ergebnisse neben den Hauptuntersuchungsgebieten sechs global verteilte Regionen untersucht. Außerdem wird die zeitliche Konsistenz durch Analyse eines zweiten, leicht versetzten Evaluationszeitraums behandelt, sowie die Abhängigkeit der Modellbewertungen von verschiedenen Referenzdaten mit Hilfe von insgesamt drei Referenzdatensätzen untersucht.

Die Ergebnisse legen nahe, dass nahezu alle Metriken zur Modellevaluierung geeignet sind. Die Auswertung unterschiedlicher Variablen und Regionen erzeugt Modellbewertungen, die sich in den Kontext aktueller Forschungsergebnisse einfügen. So wurde die Leistung der globalen Klimamodelle der neusten Generation (2013) im Vergleich zur Vorgängergeneration (2007) im Schnitt ähnlich hoch bzw. in vielen Situationen auch stärker eingeordnet. Ein durchweg bestes Modell konnte nicht festgestellt werden. Der Großteil der entwickelten Metriken zeigt für ähnliche Situationen übereinstimmende Modellbewertungen. Bei der Gewichtung hat sich der Niederschlag als besonders geeignet herausgestellt. Grund hierfür sind die im Schnitt deutlichen Unterschiede der Modelleleistungen in Zusammenhang mit einer geringeren Simulationsgüte. Umgekehrt zeigen die Metriken für die Modelle der Temperatur allgemein überwiegend hohe Evaluationsergebnisse, wodurch nur wenig Informationsgewinn durch Gewichtung erreicht werden kann. Während die Metriken gut für unterschiedliche Regionen und Skalenniveaus verwendet wer-

den können, ist eine Übertragbarkeit der Modellbewertungen für verschiedene Evaluationszeiträume nicht grundsätzlich gegeben. Zusätzlich zeigen die Modellranglisten unterschiedlicher Regionen und Jahreszeiten häufig nur geringe Korrelationen. Dies gilt besonders für den Niederschlag. Bei der Temperatur sind hingegen leichte Übereinstimmungen auszumachen. Beim Vergleich der mittleren Ranglisten über alle Modellbewertungen und Situationen der Hauptregionen des Mittelmeerraums mit den Globalregionen besteht eine signifikante Korrelation von 0,39 für Temperatur, während sie für Niederschlag um null liegt. Dieses Ergebnis ist für alle drei verwendeten Referenzdatensätze im Mittelmeerraum gültig. So schwankt die Korrelation der Modellbewertungen des Niederschlags für unterschiedliche Referenzdatensätze immer um Null und die der Temperaturranglisten zwischen 0,36 und 0,44.

Generell werden die Metriken als geeignete Evaluationswerkzeuge für Klimamodelle eingestuft. Daher können sie einen Beitrag zur Änderung des Unsicherheitsbereichs und damit zur Stärkung des Vertrauens in Klimaprojektionen leisten. Die Abhängigkeit der Modellbewertungen von Region und Untersuchungszeitraum muss dabei jedoch berücksichtigt werden. So besitzt die Analyse der Konsistenz von Modellbewertungen sowie der Stärken und Schwächen der Klimamodelle großes Potential für folgende Studien, um das Vertrauen in Modellprojektionen weiter zu steigern.

# Abstract

Climate change is one of the major tasks of the 21<sup>st</sup> century. The uncertainty of precipitation and temperature change is considered as a main challenge in this context. Thus, the development of appropriate adaptation strategies is very difficult.

In this study, four climate model evaluation approaches with 13 metrics in total are developed and compared. Current global (two generations) and regional climate models are evaluated to assess projection uncertainty. Based on model performances, weighting is applied to future climate projections to estimate simulation uncertainty. The evaluations are performed in the Mediterranean and eight sub-regions. Seasonal trend of temperature and precipitation are evaluated for the period 1960–2009. For some metrics, the climatological mean and the spectra of the time series are evaluated as well. In addition, six globally distributed study areas are evaluated to test the metrics' transferability. Further, temporal consistency is assessed by the evaluation of a second slightly shifted timeframe. Finally, three reference datasets are considered in order to analyse the dependence of the evaluation results between each other.

Results indicate that most metrics are suitable to evaluate climate models. Their application to different variables and regions generates reasonable model assessments which fit in the context of current publications in this field of research. In many situations, the results of the current model generation (2013) are similar or better compared to those of the last generation (2007). One single model with superior performance for all variables and situations cannot be found. Most metrics show similar estimations of performances for the same situations. Precipitation turned out to be more suitable for model weighting. Here, the differences between model weights are larger because of overall higher spread and lower model performances. Against this, there are mostly high performances on an equal level for simulations of temperature which lead to a minor added value of weighting. While metrics can easily be transferred and applied to different regions and scales, some evaluation results depend on the evaluated timeframe. Further, the model rankings for different regions and seasons show only minor correlations for most situations. This is particularly true for precipitation. However, for temperature there are some significant positive correlations. Comparing the mean ranking over all evaluation results of the main study areas of the Mediterranean

with that of the globally distributed regions, there is a significant correlation of 0.39 for temperature and a correlation around zero for precipitation. The choice of reference dataset for the Mediterranean areas is subordinated in this context. For different reference datasets, the overall rankings show correlations around zero for precipitation while those for temperature are between 0.36 and 0.44.

Overall, the metrics are suitable for the evaluation of climate models. Thus, they offer promising contributions to improve the range of uncertainty and therefore to enhance the general confidence in climate projections. However, dependence of model assessments on the analysed region and evaluation timeframe has to be considered. Consequently, the analyses of consistency of model evaluations and of climate model strengths and weaknesses have great potential for future studies, to further enhance confidence in climate projections.

# Inhaltsverzeichnis

Zusammenfassung	i
Abstract	iii
Tabellenverzeichnis	viii
Abbildungsverzeichnis	ix
Abkürzungsverzeichnis	xi
<b>1 Einführung und Hintergründe</b>	<b>1</b>
1.1 Klimawandel und Modellprojektionen . . . . .	1
1.2 Bewertung von Modellgüte . . . . .	3
1.3 Die Unsicherheit der Klimaprojektionen . . . . .	5
1.4 Untersuchungsgebiete . . . . .	7
1.5 Ziele und Aufbau der Arbeit . . . . .	11
<b>2 Datengrundlage</b>	<b>13</b>
2.1 Referenzdatensätze . . . . .	13
2.1.1 E-OBS . . . . .	14
2.1.2 Climate Research Unit . . . . .	15
2.1.3 ERA-20C Reanalyse . . . . .	16
2.2 Analysierte Klimamodelldaten . . . . .	16
2.2.1 Coupled Model Intercomparison Projekt Phase 3 . . . . .	17
2.2.2 Coupled Model Intercomparison Projekt Phase 5 . . . . .	19
2.2.3 Coordinated Regional Downscaling Experiment . . . . .	20
<b>3 Statistische Methoden</b>	<b>26</b>
3.1 Arithmetisches und gewichtetes Mittel . . . . .	26
3.2 Varianz und Standardabweichung . . . . .	28
3.3 Wahrscheinlichkeitsdichtefunktion . . . . .	29
3.3.1 Normalverteilungsansatz . . . . .	29
3.3.2 Kerndichteschätzer . . . . .	30
3.4 Der Kolmogorov-Smirnov-Test . . . . .	31

3.5	Kovarianz und Korrelation . . . . .	32
3.6	Regressionsanalyse . . . . .	34
3.7	Signifikanztests . . . . .	35
3.7.1	Test des Spearman-Korrelationskoeffizienten . . . . .	36
3.7.2	Test des Erwartungswertes . . . . .	36
3.7.3	Test zweier Varianzen . . . . .	37
3.7.4	Test zweier Regressionskoeffizienten . . . . .	38
3.7.5	$u$ -Anpassungstest . . . . .	39
<b>4</b>	<b>Evaluationsmetriken</b>	<b>41</b>
4.1	4-Felder-Metriken (4FM) . . . . .	41
4.2	Metrik der quadratische Abweichung (RM) . . . . .	45
4.3	Zeitreihenmetrik (ZM) . . . . .	47
4.4	Fingerprinting-Ansätze . . . . .	51
4.4.1	Optimaler und suboptimaler Fingerprintansatz (FPA) . . . . .	52
4.4.2	Zusätzliche Fingerprintansätze (FPA <sub>z</sub> ) . . . . .	57
<b>5</b>	<b>Modellevaluation im Mittelmeerraum</b>	<b>59</b>
5.1	Regionale Variabilität der Bewertung . . . . .	59
5.1.1	Unterschiede der Modellbewertungen . . . . .	60
	4FM . . . . .	60
	RM . . . . .	63
	ZM . . . . .	65
	FPA . . . . .	66
5.1.2	Regionale Korrelation der Bewertungen . . . . .	70
5.2	Saisonaler Einfluss auf die Modellbewertung . . . . .	73
5.2.1	Vergleich der saisonalen Simulationsgüte . . . . .	73
5.2.2	Saisonale Muster der Modellbewertung . . . . .	74
5.3	Analyse der Modellgenerationen . . . . .	77
5.3.1	Unterschiede der Simulationsgüte . . . . .	78
5.3.2	Einfluss dynamischen Forcings . . . . .	82
5.4	Korrelation der Evaluationsansätze . . . . .	84
5.5	Hintergründe der Modellbewertung . . . . .	87
5.5.1	Vergleich der Evaluationsgrundlagen . . . . .	87
5.5.2	Bedeutung der relativen Gewichte . . . . .	90
<b>6</b>	<b>Gewichtungen der Zukunftsprojektionen</b>	<b>94</b>
6.1	Auswertung der Zeitreihengewichtung . . . . .	95
6.1.1	Zeitreihengewichtung . . . . .	95
6.1.2	Gewichtungseffekt auf den Trend . . . . .	98

---

6.2	Änderung des Unsicherheitsbereichs . . . . .	100
6.2.1	Analyse einzelner Emissionsszenarien . . . . .	102
6.2.2	Multi-Szenario-Analyse . . . . .	107
<b>7</b>	<b>Analyse auf globaler Ebene</b>	<b>113</b>
7.1	Evaluation der Modelle . . . . .	113
7.1.1	Analyse der globalen Gewichte . . . . .	114
	Regionale und saisonale Unterschiede . . . . .	114
	Unterschiede zwischen den Jahreszeiten . . . . .	120
	Analyse der Modellgenerationen . . . . .	122
7.1.2	Korrelation der Evaluationsansätze . . . . .	125
7.2	Gewichtung der Zukunftsprojektionen . . . . .	127
7.2.1	Auswertung der Zeitreihengewichtung . . . . .	128
7.2.2	Änderung des Unsicherheitsbereichs . . . . .	132
<b>8</b>	<b>Integrität der Evaluationsmetriken</b>	<b>138</b>
8.1	Konsistenz der Modellbewertungen . . . . .	138
8.2	Zeitliche Übertragbarkeit der Modellbewertungen . . . . .	143
8.3	Ansatz zur Metrikvalidierung . . . . .	147
8.4	Abhängigkeit von Referenzdaten . . . . .	151
<b>9</b>	<b>Zusammenfassung und Diskussion</b>	<b>155</b>
<b>10</b>	<b>Ausblick</b>	<b>163</b>
	<b>Literaturverzeichnis</b>	<b>165</b>
	<b>Danksagungen</b>	<b>193</b>

# Tabellenverzeichnis

2.1	Übersicht über die verwendeten Referenzdatensätze . . . . .	13
2.2	Analysierte globale Klimamodelle von CMIP3 . . . . .	18
2.3	Analysierte globale Klimamodelle von CMIP5 . . . . .	24
2.4	Analysierte regionale Klimamodelle von CORDEX . . . . .	25
4.1	4-Felder-Tabelle . . . . .	42
4.2	Formeln zur Berechnung der sechs 4-Felder-Metriken . . . . .	43
4.3	Summe der erklärten Varianzen der EOF . . . . .	57
8.1	Gesamtranglisten der mittleren Modellrangplätze . . . . .	140

# Abbildungsverzeichnis

1.1	Klimatologie des Mittelmeerraums . . . . .	9
1.2	Überblick über die Untersuchungsgebiete . . . . .	10
2.1	Trend für DJF aller Referenzdatensätze . . . . .	14
2.2	Simulierte Temperatur- und Niederschlagstrends des Mittelmeerraums . . . . .	22
4.1	Veranschaulichung der Zeitreihenmetrik . . . . .	50
4.2	1. EOF mit Landungen für Niederschlag und Temperatur . . . . .	56
5.1	Annuelle PHI-Rohgewichte der Modellevaluation . . . . .	61
5.2	Evaluationsergebnis der FPA mit übereinstimmenden Zeitfenstern	67
5.3	Beispiel des Evaluationsergebnisses der FPA mit dynamischen Zeitfenstern . . . . .	68
5.4	Grundlage der zusätzlichen Fingerprint-Ansätze . . . . .	69
5.5	Korrelation der PHI-Ranglisten aller Situationen im Mittelmeerraum	70
5.6	Korrelation der RMSE-Ranglisten aller Situationen im Mittelmeerraum . . . . .	71
5.7	Jahreszeitlich unterteilte PHI-Rohgewichte . . . . .	74
5.8	Mittlere saisonale Rohgewichte der vier Metrikgruppen im Mittelmeerraum . . . . .	75
5.9	Vergleich der CMIP3-, CMIP5- und CORDEX-Evaluationen . . . . .	79
5.10	Einfluss dynamischen Forcings auf die Simulationsgüte . . . . .	83
5.11	Korrelation aller Evaluationsansätze für den Gesamtmittelmeerraum	84
5.12	Beispiele für das Trendmuster des Niederschlags . . . . .	88
5.13	Beispiele für das Trendmuster der Temperatur . . . . .	89
5.14	Relative 4FM-Gewichte der Hauptuntersuchungsregionen . . . . .	91
6.1	Gewichtungsergebnisse der annualen Zeitreihen . . . . .	96
6.2	Effekte der metrikgewichteten Zeitreihen auf den Trend . . . . .	99
6.3	Niederschlags- und Temperaturänderung zum Ende des 21. Jahrhunderts . . . . .	101
6.4	WDF der Klimaänderung aller MMEs für die Ägäis . . . . .	103
6.5	Übersicht über die Gewichtungseffekte auf die WDF . . . . .	106

---

6.6	Multi-Szenario-KDF der Klimaänderung für Ägäis . . . . .	108
6.7	MSK-Konfidenzintervalle aller Regionen . . . . .	110
7.1	Annuelle PHI-Rohgewichte für Arktis und Global . . . . .	115
7.2	Relative 4FM-Gewichte der Globalregionen . . . . .	118
7.3	Korrelation der SUB-8-Ranglisten aller Situationen der Globalregionen . . . . .	119
7.4	Mittlere saisonale Rohgewichte der vier Metrikgruppen für Global	121
7.5	Vergleich der MME-Evaluationsergebnisse für die Globalregionen .	123
7.6	Einfluss dynamischen Forcings auf die glob. Simulationsgüte . . .	125
7.7	Korrelation aller Evaluationsansätze für alle Globalregionen . . .	126
7.8	Gewichtungsergebnisse der annualen globalen Zeitreihen . . . . .	128
7.9	Effekte der metrikgewichteten Zeitreihen auf den Trend der Globalregionen . . . . .	131
7.10	Übersicht der Gewichtungseffekte auf Global-WDF . . . . .	133
7.11	Multi-Szenario-KDF der Klimaänderung der Globalregionen . . .	135
7.12	MSK-Konfidenzintervalle aller Globalregionen . . . . .	136
8.1	Korrelationen der Modellbewertungen für unterschiedliche Referenzzeiträume (MMR). . . . .	144
8.2	Korrelationen der Modellbewertungen für unterschiedliche Referenzzeiträume (Global) . . . . .	146
8.3	Ansatz zur Metrikvalidierung für die Mittelmeerregionen . . . . .	149
8.4	Ansatz zur Metrikvalidierung für die Globalregionen . . . . .	150
8.5	Korrelationen der Modellranglisten für unterschiedliche Referenzdatensätze . . . . .	152

# Abkürzungsverzeichnis

## Hintergründe

<b>COMEPRO</b>	Projekt: Vergleich von Metriken für probabilistische Klimaprojektionen des Niederschlages im Mittelmeerraum
<b>GKM</b>	Globale Klimamodelle
<b>RKM</b>	Regionale Klimamodelle
<b>IPCC</b>	Weltklimarat (Intergovernmental Panel on Climate Change)
<b>AR4</b>	4. Sachstandsbericht (Assessment Report) des IPCC
<b>AR5</b>	5. Sachstandsbericht (Assessment Report) des IPCC
<b>WCRP</b>	World Climate Research Programme
<b>PCMDI</b>	Program for Climate Model Diagnosis and Intercomparison
<b>MME</b>	Multi-Modell Ensemble
<b>MMR</b>	Gesamtmittelmeerregion
<b>NAO</b>	Nordatlantische Oszillation
<b>ENSO</b>	El Niño-Southern Oszillation

## Daten und Methodik

<b>CRU</b>	Climate Research Unit
<b>CMIP3</b>	Coupled Model Intercomparison Project Phase 3
<b>CMIP5</b>	Coupled Model Intercomparison Project Phase 5
<b>CORDEX</b>	Coordinated Regional Downscaling Experiment
<b>SRES</b>	Special Report on Emission Scenario (CMIP3)
<b>A1B</b>	CMIP3 Emissionszenario I: SRES A1B
<b>A2</b>	CMIP3 Emissionszenario II: SRES A2
<b>RCP</b>	Responsive climate pathway (CMIP5)
<b>RCP4.5</b>	CMIP5/CORDEX Emissionszenario I: RCP4.5
<b>RCP8.5</b>	CMIP5/CORDEX Emissionszenario II: RCP8.5
<b>SN</b>	Simulationsnummer
<b>RMSE</b>	Wurzel der mittleren quadratischen Abweichung (Root Mean Square Error)
<b>WDF</b>	Wahrscheinlichkeitsdichtefunktion
<b>KDF</b>	Kerndichteschätzfunktion
<b>MSK</b>	Multi-Szenario-Kerndichteschätzfunktion
<b>EOF</b>	Empirische Orthogonalfunktion

**IDW** Invertierte Distanzgewichtung (Interpolation)

**Evaluationsmetriken**

**4FM** 4-Felder-Metriken  
**PHI** Phi-Metrik (4FM)  
**HEI** Heidke-Metrik (4FM)  
**CHI** Chi<sup>2</sup>-Metrik (4FM)  
**GSS** Gilbert-Metrik (4FM)  
**PIE** Pierce-Metrik (4FM)  
**LOR** Log Odds Ratio-Metrik (4FM)  
**ZM** Zeitreihen-Metrik  
**RM** RMSE-Metrik  
**FPA** Fingerprinting-Ansätze  
**FPA<sub>z</sub>** Zusätzliche FPA-Metriken  
**SUB-Px** Suboptimale FPA-Metrik basierend auf Originaldaten  
**SUB-\*** Suboptimale FPA-Metrik basierend auf \* EOF  
**OPT-\*** Optimale FPA-Metrik basierend auf \* EOF  
**SUB<sub>z</sub>-\*** Suboptimale FPA<sub>z</sub>-Metrik basierend auf \* EOF  
**OPT<sub>z</sub>-\*** Optimale FPA<sub>z</sub>-Metrik basierend auf \* EOF





# 1 Einführung und Hintergründe

In diesem einleitenden Kapitel werden die Motivation sowie die Hintergründe dieser Arbeit erläutert. Dazu wird zunächst die Bedeutung des Klimawandels und der Unsicherheit bezüglich der Klimamodellsimulationen kurz herausgestellt. Im Anschluss werden die Herausforderungen der Modellevaluierung erläutert und dazu ein Einblick in den bisherigen Stand der Klimaforschung gegeben. Zusätzlich werden bisherige Ansätze zur Steigerung des Vertrauens in die Modellprojektionen betrachtet. Dabei wird auch das Projekt, in dessen Rahmen diese Studie entstand, kurz beschrieben. Anschließend werden die Untersuchungsregionen vorgestellt. Abgeschlossen wird dieses Kapitel mit einer Zusammenfassung der Ziele sowie einem kurzen Überblick über den Aufbau dieser Arbeit.

## 1.1 Klimawandel und Modellprojektionen

Den Menschen wird ihre Abhängigkeit vom Klima meist nur dann vor Augen geführt, wenn ein Klimaereignis eintritt, das außergewöhnlich ist. Solche Extremereignisse sind meist negativer Natur, wie beispielsweise Hitzeperioden, die im Zusammenspiel mit Dürren zu Waldbränden und Trinkwassermangel führen können, oder von Starkniederschlägen ausgelöste Überschwemmungen und Hangrutschungen. So ist der Mensch stärker vom Klima abhängig als er sich dessen vielleicht bewusst ist, wenn alles in seinen bekannten Bahnen verläuft (IPCC, 2014a). Dabei war das Klima auf der Erde nie ein statischer Zustand. Veränderungen sind ein Teil der Entwicklung des Planeten und damit auch der Menschen. Das Außergewöhnliche am anthropogenen Klimawandel ist daher nicht die Tatsache, dass sich das Klima ändert, sondern vielmehr die Ursache und Geschwindigkeit (LOARIE ET AL., 2009). Als Hauptgrund muss hier der sich weiter intensivierende anthropogene Ausstoß von Treibhausgasen seit Mitte des 19. Jahrhunderts genannt werden (IPCC, 2013). So legen Eisbohrkerne nahe, dass bereits heute die Konzentration bestimmter Treibhausgase wie  $\text{CO}_2$ ,  $\text{CH}_4$  und  $\text{N}_2\text{O}$  deutlich über den Maxima der letzten 800.000 Jahre liegt. Dazu kommt, dass eine vergleichbar rasche Zunahme der atmosphärischen Konzentrationen, wie sie über die letzten 100 Jahre beobachtet wurde, in den letzten 22.000 Jahren nicht existierte (IPCC, 2013). Der Klimawandel zeigt sich nicht nur in der Atmosphäre, sondern beispielsweise auch

auf der Landoberfläche in Form von Gletscherschwund oder im Ozean durch Meeresspiegelanstieg oder Korallensterben (HOEGH-GULDBERG ET AL., 2007; FLATO ET AL., 2013). Kein direkter klimatologischer Nachweis, dennoch auffällige Indizien für die globale Erwärmung sind die gehäuften Extremereignisse der letzten Jahrzehnte, wie das 2015 gemessene und zu diesem Zeitpunkt deutsche absolute Temperaturmaximum in Kitzingen mit  $40,3^{\circ}\text{C}$  (BECKER ET AL., 2015). Veränderungen beispielsweise des Niederschlags auf regionaler Ebene, der Intensität und Häufigkeit von Extremereignissen oder der Anstieg des Meeresspiegels sind nur ein kleiner Teil möglicher Konsequenzen des anthropogenen Klimawandels (IPCC, 2007c, 2013). Ein Problem in der Wahrnehmung der globalen Erwärmung ist darüber hinaus, dass sich, anders als bei den aufgezählten Wetterphänomenen, viele Veränderungen des Klimawandels langsam entwickeln. Die Folgen, die beispielsweise das Abschmelzen der Eisschilde und Gebirgsgletscher oder der Anstieg des Meeresspiegels haben, sind heute nicht mit absoluter Sicherheit abzuschätzen. Wenn es allerdings durch den konstanten Ausstoß von Treibhausgasen zu einer nachhaltigen Veränderung gekommen ist, muss davon ausgegangen werden, dass der neue Zustand für lange Zeit unumkehrbar bleiben wird (SOLOMON ET AL., 2009; HOEGH-GULDBERG & BRUNO, 2010). Da neben dem Meeresspiegelanstieg auch zunehmende Wasserknappheit in ariden Regionen der Erde (bspw. dem Mittelmeerraum) wahrscheinliche Szenarien sind (ALLEN ET AL., 2010; GIORGI & LIONELLO, 2008; IGLESIAS ET AL., 2007), muss mit schwerwiegenden sozioökonomischen Folgen für die Menschen gerechnet werden (PERNETTA, 1992; KATES ET AL., 2012). Da Treibhausgase bereits seit der Industrialisierung im 19. Jahrhundert in die Atmosphäre emittiert werden (IPCC, 2013), sind auch mit einem völligen Stopp weiterer Emissionen bestimmte Veränderungen für die nächsten Jahrzehnte unseres Klimasystems irreversibel. Das bedeutet jedoch nicht, dass nachhaltige Politik und die Reduzierung der Emission von klimawirksamen Gasen wie  $\text{CO}_2$  oder  $\text{CH}_4$  keine Auswirkungen hätten (DIMITROV, 2016; KNUTTI ET AL., 2015). So zeigen aktuelle Klimamodellprojektionen anhand von verschiedenen Emissionszenarien, dass Geschwindigkeit und Intensität des Klimawandels eingedämmt bzw. abgeschwächt werden können (IPCC, 2013).

Da der anthropogene Klimawandel beispiellos ist und Vergleiche mit der Vergangenheit sich daher als schwierig gestalten, bieten Klimamodelle die besten Informationsquellen für das zukünftige Klima (COLLINS, 2007). Auf Grundlage der Modellprojektionen können Probleme frühzeitig antizipiert und Lösungsstrategien erarbeitet werden (SMITH, 1997). Allerdings ist kein Modell in der Lage, alle Elemente des Klimasystems perfekt abzubilden (IPCC, 2001c), weshalb es zu Verunsicherung der Menschen kommt, welche die geeigneten Anpassungsstrategien sind. Um die beste Informationsgrundlage zu bieten, werden daher Gruppen von Modellen zusammengefasst und als Multi-Modell Ensemble (MME) ausge-

wertet (MEEHL ET AL., 2007). In ihrer heutigen Form erfassen sie einen großen Teil der physikalischen Prozesse, die zur Entstehung des Klimasystems beitragen. Obwohl eine Mehrheit der Klimamodelle auf globaler Ebene hohe Übereinstimmungen beispielsweise bei der jährlichen Mitteltemperatur besitzen, zeigt sich insbesondere auf regionaler Ebene ein hohes Maß an Unsicherheit zwischen den verschiedenen Modellen (GIORGI ET AL., 2009). Da es aufgrund der hohen Komplexität der atmosphärischen Prozesse noch nicht möglich ist, das Klimasystem vollständig in Form eines Modells zu rekonstruieren, sind Vereinfachungen und Parametrisierungen erforderlich (REICHLER & KIM, 2008; WANG ET AL., 2014). Für die Entwicklung geeigneter Anpassungsstrategien ist eine möglichst hohe Verlässlichkeit der Zukunftsprojektionen jedoch entscheidend. Allerdings zeigen die Projektionen unterschiedlicher Modelle eine beträchtliche Schwankungsbreite. Neben den Herausforderungen der Modellentwicklung ist auch die Zukunft selbst eine Quelle von großer Unsicherheit. So sind das Emissionsverhalten oder die Landnutzungsformen im 21. Jahrhundert von vielen Faktoren abhängig und daher nicht eindeutig voraussagbar. Aus diesem Grund werden für die Zukunftsprojektionen mehrere Emissionsszenarien simuliert, um verschiedene Eventualitäten abzudecken (MOSS ET AL., 2010; NAKICENOVIC ET AL., 2000). Neben den Zukunftsläufen für das 21. Jahrhundert besitzen alle Klimamodelle auch Simulationen, die im 20. Jahrhundert oder noch früher beginnen. Im Gegensatz zu den historischen Zeiträumen, die alle Modelle statistisch möglichst realistisch reproduzieren sollten, ist die Unsicherheit bzgl. des zukünftigen Klimas basierend auf den Emissionsszenarien folglich wichtig. So können Abschätzungen eines Unsicherheitsbereichs der Klimaänderung ermöglicht werden (ALLEN ET AL., 2009). Welche Modellsimulation der tatsächlichen klimatischen Entwicklung in der Zukunft am nächsten kommt, lässt sich allerdings nicht mit Sicherheit sagen. Jedoch können Aussagen über die Simulation des aktuellen bzw. historischen Klimazustands getroffen werden. Es ist auffällig, dass die Modelle in den historischen Simulationen ebenso wie in den Zukunftsprojektionen teilweise beträchtliche Unterschiede aufweisen (REICHLER & KIM, 2008; FORSTER ET AL., 2013). Allerdings hat sich gezeigt, dass viele Klimamodelle Simulationen des globalen und teilweise auch regionalen Klimas mit hoher Genauigkeit (Modellgüte) bieten können (FLATO ET AL., 2013; RANDALL ET AL., 2007).

## 1.2 Bewertung von Modellgüte

In diesem Abschnitt soll ein Eindruck von den Evaluationsmetriken bisheriger Studien gegeben werden. Dabei kann jedoch kein Anspruch auf Vollständigkeit gestellt werden. Für weitergehende Informationen zur Einschätzung der Modell-

leistungen der unterschiedlichen Ansätze wird daher auf die Modellevaluationskapitel der letzten beiden Sachstandsberichte des Weltklimarates verwiesen (RANDALL ET AL., 2007; FLATO ET AL., 2013).

Generell handelt es sich bei der Wahl eines Evaluationskriteriums für Klimamodelle um eine komplexe Fragestellung. Da man Modelle bezüglich ihrer Zukunftsprojektionen aufgrund fehlender Vergleichswerte nicht objektiv bewerten kann, bedient man sich ihrer Simulationen des 20. und frühen 21. Jahrhunderts (TEBALDI & KNUTTI, 2007; NIKULIN ET AL., 2012). Hintergrund ist die Annahme, dass ein Modell, das die Klimaentwicklung in der Vergangenheit realitätsgetreu simulieren kann, mehr Vertrauen bei der Betrachtung der Projektion von zukünftigen Klimaverhältnissen erhalten sollte (REICHLER & KIM, 2008). Eine weitere Herausforderung liegt in den unterschiedlichen Herangehensweisen bei der Evaluation von Modellen. Während es eine Vielzahl an Studien gibt, die sich mit der Modellevaluation beschäftigen, wird gleichzeitig auch Zweifel an der Interpretation der Ergebnisse deutlich. Die von CHRISTENSEN ET AL. (2010) und WEIGEL ET AL. (2010) formulierte Kritik, dass durch Modellevaluation eine neue Unsicherheit erzeugt werde, da es keine universell gültige Metrik gibt, kann nicht von der Hand gewiesen werden. Neben der Wahl der Metrik sind beispielsweise auch die untersuchten Variablen und Regionen entscheidend dafür, welche Modelle eine hohe Bewertung erhalten (PERKINS ET AL., 2007). Allerdings gilt trotz dieser berechtigten Kritik, dass die Modelle über unterschiedliche Leistungsfähigkeiten verfügen (LEDUC ET AL., 2016). So können die bisher angewandten Evaluationsmetriken deutliche Unterschiede zwischen den Modellen aufzeigen, auch wenn die Ergebnisse verschiedener Ansätze nicht in jedem Punkt übereinstimmen.

Das Fehlen einer allgemeingültigen Metrik (GLECKLER ET AL., 2008) führt dazu, dass die bisherigen Studien ein umfangreiches Bild unterschiedlicher Evaluationsansätze abbilden. Diese unterscheiden sich deutlich hinsichtlich der Komplexität, der evaluierten Klimavariablen, der Region oder anderer Charakteristika. Ein möglicher Ansatz wurde im Vorfeld dieser Arbeit von RING ET AL. (2016) durchgeführt. Hier wurde die Abweichung der aktuellen Klimamodelle von verschiedenen Beobachtungsdaten (Bias) für den mittleren Niederschlag des 20. Jahrhunderts in mehreren Untersuchungsgebieten untersucht. KUMAR ET AL. (2014) nutzen verschiedene Herangehensweisen, darunter ebenfalls einen räumlichen Bias und aggregierte Zeitreihen, um eine Modellbewertung auf Basis des 20. Jahrhunderts zu generieren. Die überwiegende Zahl der Ansätze ist jedoch deutlich komplexer (u. a. KUMAR ET AL., 2013; KOUTROULIS ET AL., 2016; BACHNER ET AL., 2008; CHEN & FRAUENFELD, 2014; MARKOVIC ET AL., 2013; SANDERSON ET AL., 2015b,a). Die vielen aktuellen Studien zum Thema der Modellevaluation zeigen, dass die Suche nach einer gültigen Metrik hochrelevant ist. Einen wichtigen Beitrag dazu haben GIORGI & MEARN (2002) mit der Entwicklung

der REA-Methode (reliability ensemble averaging) geleistet. Auch hier wird der historische Zeitraum der Modellläufe evaluiert, um ein Kriterium für die Simulationsgüte zu schaffen. Zusätzlich wird die Übereinstimmung der Klimaänderung im 21. Jahrhundert zwischen den Simulationen betrachtet und zur Bewertung der Modelle berücksichtigt. Die Studie von GIORGI & MEARNs (2002) wurde für 22 kleinskalige Regionen für Niederschlag und Temperatur durchgeführt und bietet daher gute Vergleichsmöglichkeiten für weitere Studien (GIORGI & MEARNs, 2003; TEBALDI ET AL., 2004; GIORGI, 2005). Als eine sehr komplexe Evaluationsmetrik wurde von TEBALDI ET AL. (2004) die Erweiterung der REA-Methode basierend auf dem Bayes-Theorem vorgestellt, die auf bedingte Wahrscheinlichkeiten setzt.

Trotz vieler weiterer aktueller Studien hat sich jedoch auch der Bayes-Ansatz nicht als allgemeingültige Metrik herauskristallisiert. Viele Studien setzten weiterhin auf Evaluation des historischen Zeitraums und unterschiedliche Modelleigenschaften (u.a. DANANDEH MEHR & KAHYA, 2016; KATRAGKOU ET AL., 2015; SHASHIKANTH ET AL., 2014; GROSE ET AL., 2014; LI ET AL., 2016). Diese vielfältigen Entwicklungen zeigen, dass die Frage nach einer geeigneten und möglicherweise allgemeingültigen Modellevaluation entscheidend ist. Um den vielen Fragen der Modellbewertung gerecht werden zu können, werden in dieser Arbeit vier unterschiedliche Ansätze, die sich teilweise deutlich in ihrer Komplexität unterscheiden, entwickelt, angewandt und verglichen. Indem verschiedene Klimavariablen in denselben Regionen ausgewertet werden, soll ein differenziertes Bild der Modelleistung erzeugt und neue Aspekte beleuchtet werden. Zusätzlich kann so auch die Frage nach der allgemeinen Gültigkeit der Metriken diskutiert werden.

Zuletzt sind für eine sinnvolle Anwendung von Evaluationsmetriken auch die Referenzdaten entscheidend (FEKETE ET AL., 2004; RING ET AL., 2016). Insbesondere bei Niederschlagsdatensätzen fanden YIN ET AL. (2015) und PRAKASH ET AL. (2015) teilweise deutliche Unterschiede. Die richtige Wahl der Referenzdaten gerade auch hinsichtlich regionaler Stärken und Schwächen ist daher von großer Bedeutung (THORNE ET AL., 2005; KIM ET AL., 2015; GEHNE ET AL., 2016). Um dieser Bedeutung Rechnung zu tragen, werden in dieser Arbeit drei unterschiedliche Typen von Referenzdaten (Beobachtungen und eine Reanalyse) verwendet. Damit wird neben der Aussage zur Modellgüte auch ein Anteil der Unsicherheit basierend auf den Referenzdaten diskutiert. Ziel der meisten Studien ist es Modelle zu finden, die besser für bestimmte Fragestellungen geeignet sind, um dadurch Aussagen über die Unsicherheit der Zukunftsprojektionen treffen zu können. In Abschnitt 1.3 wird daher ein Überblick über die Modellentwicklung seit 1990 und die bisherigen Versuche gegeben, die Unsicherheit der Klimasimulationen zu reduzieren.

### 1.3 Die Unsicherheit der Klimaprojektionen

Zuverlässige Projektionen sind mit die wichtigsten Ziele der Klimaforschung (RÄISÄNEN, 2007). Nur so können fundierte Zukunftsstrategien entwickelt werden, um die Anpassung an den anthropogenen Klimawandel nachhaltig zu ermöglichen. Hierzu wurde 1990 im Rahmen des ersten Sachstandberichts (eng. Assessment Report, AR) des Weltklimarats (eng. Intergovernmental Panel on Climate Change, IPCC) der Grundstein gelegt. Dieser sollte eine unabhängige und wertungsfreie Instanz bieten, um die Menschen und die Politik über die Veränderungen des Klimas zu informieren (IPCC, 1990a,b,c). Die fortschreitende Bedeutung und Entwicklung zeigt sich in den vier folgenden Sachstandberichten des IPCC, veröffentlicht 1995 (IPCC, 1996a,b,c), 2001 (IPCC, 2001a,b,c), 2007 (IPCC, 2007a,b,c) und 2013 (IPCC, 2013, 2014a,b). Auch die Klimamodellierung hat sich in dieser Zeit deutlich weiterentwickelt. Dies lässt sich anhand immer höherer räumlicher und vertikaler Auflösungen der neueren Modellgenerationen leicht nachvollziehen (RANDALL ET AL., 2007). Infolgedessen wurde neben globalen Klimamodellen auch die Entwicklung sehr hoch aufgelöster regionaler Klimamodelle vorangetrieben, die zuverlässigere Projektionen auf regionaler Ebene bieten sollen (CHRISTENSEN ET AL., 2007, 2010; KATRAGKOU ET AL., 2015; GIORGI ET AL., 2009). Dazu kommen auch neue Variablen und Verknüpfungen, die seit 1990 in die Modellen implementiert wurden, wie die Kopplung von Ozean und Atmosphäre oder die Berücksichtigung von Änderungen der Biosphäre (IPCC, 2007c, 2013). Neben dem technischen Fortschritt hat die seit 1990 vergangene Zeit selbst zur Steigerung des Vertrauens der Modelle beigetragen, da die ersten 25 Jahre, die damals noch Teil von Zukunftsprojektionen waren, heute bereits messbar sind. Gerade im Bezug auf die globale Mitteltemperatur sind hier hohe Übereinstimmungen der Modelle mit den gemessenen Beobachtungen festzustellen (FLATO ET AL., 2013). Frühere Studien bescheinigen darüber hinaus dem neueren MME (2013) für verschiedene Regionen insgesamt eine bessere Leistung gegenüber den Modellen der Vorgängerversion (CHEN & FRAUENFELD, 2014; GROSE ET AL., 2014; SHASHIKANTH ET AL., 2014; RING ET AL., 2016). Eine ähnliche Entwicklung wurde auch für die aktuelle Generation der regionalen Modelle festgestellt (PREIN ET AL., 2016; NIKIEMA ET AL., 2017).

Allerdings bleibt trotz aller Verbesserungen und Indizien Unsicherheit (STEVENS & BONY, 2013). Hintergrund sind u.a. Parametrisierungen, die in der Modellierung erforderlich sind, um beispielsweise die Entstehung von Niederschlag zu berechnen. Manche Wechselwirkungen sind noch nicht vorhersehbar oder existieren auf zu kleinem Maßstab für die Modellierung (BACHNER ET AL., 2008; EMORI ET AL., 2005; ANSTEY ET AL., 2013; PAETH ET AL., 2006). Dabei ist gerade die Frage nach dem regionalen und lokalen Klimawandel von höchstem

Interesse, insbesondere in von Desertifikation beeinflussten Übergangsregionen oder mit Blick auf die Entwicklung des Niederschlagsverhaltens für die Landwirtschaft (ASSENG ET AL., 2013; GOBIET ET AL., 2014). In diesem Zusammenhang stellt sich die Frage, ob es ein Modell gibt, das zuverlässig für alle Regionen oder Variablen verwendet werden kann. Dazu zeigen jedoch beispielsweise NIKULIN ET AL. (2012), dass ein MME meist einem einzelnen überlegen ist. Mit Hilfe von MME können so Konfidenzintervalle bestimmt werden, um natürliche Variabilität zu simulieren und statische Aussagen treffen zu können (TEBALDI & KNUTTI, 2007; GILLETT, 2015). Diese Intervalle schwanken, abhängig von der untersuchten Klimagröße oder Region, jedoch beträchtlich. In ariden Gebieten kann aber schon eine leichte Zu- oder Abnahme von Niederschlag entscheidend sein. Das Ziel der Verringerung von Unsicherheit ist entsprechend von hoher Relevanz (STAINFORTH ET AL., 2005; HAWKINS & SUTTON, 2009; PALMER ET AL., 2005).

Trotz der vielen Verbesserungen durch die Entwicklung neuer Klimamodelle und die Verwendung eines MME zur Abschätzung der Unsicherheit bleibt die Aufgabe, das Vertrauen in die Zukunftsprojektionen weiter zu steigern, um den Menschen so zeitnah angemessene Anpassungsstrategien zu ermöglichen (KNUTTI & SEDLÁČEK, 2012; HAWKINS & SUTTON, 2009). Um die Unsicherheit zu reduzieren, werden die Klimamodelle daher evaluiert (siehe Abschnitt 1.2). In diesem Zusammenhang haben HAUGHTON ET AL. (2015) und RÄISÄNEN & YLHÄISI (2012) festgestellt, dass teilweise eine deutliche Verringerung der Unsicherheit erreicht werden kann. Auch HAWKINS & SUTTON (2009) weisen darauf hin, dass dies einen besonders wichtigen Aspekt der zukünftigen Klimaforschung darstellt. Da bisher keine allgemeingültige Metrik anerkannt wurde (GLECKLER ET AL., 2008), wird meist die gleichgewichtete Betrachtung aller verfügbaren Modelle genutzt (RANDALL ET AL., 2007; FLATO ET AL., 2013). Dennoch liegt die These nahe, dass eine leistungsorientierte Gewichtung von Modellen zu einer zuverlässigen Veränderung des Unsicherheitsbereichs führen sollte (RÄISÄNEN ET AL., 2010; DELSOLE ET AL., 2013).

So wird die „Modell-Demokratie“ auch von LEDUC ET AL. (2016) und REICHLER & KIM (2008) in Frage gestellt. Beide Studien kommen trotz achtjährigem Veröffentlichungsunterschied zu dem Schluss, dass globale Klimamodelle zu abhängig voneinander sind, um allen die gleiche Bedeutung zuzuteilen. Zudem zeigen u.a. RÄISÄNEN & YLHÄISI (2012) und GILLETT (2015), dass eine sinnvolle Gewichtung helfen kann, zuverlässigere Aussagen über Klimaprojektionen zu treffen. Dies deckt sich mit HIDALGO & ALFARO (2015), die nahe legen, dass Klimamodelle evaluiert und entsprechend ihrer Leistung eingesetzt werden sollten. YLHÄISI ET AL. (2015) führen darüber hinaus an, dass nur durch Modellbewertung und gewichteter Anwendung der Ergebnisse eine angemessene Nutzung

neuer Modelle und ihrer verbesserten Fähigkeiten erreicht werden kann. Auch KLOCKE ET AL. (2011) und GILLET (2015) empfehlen, auf eine fundierte Gewichtung einzelner Modelle innerhalb des MME zurückzugreifen. BOÉ & TERRAY (2015) stimmen mit WEIGEL ET AL. (2010) überein, dass die korrekte Wahl der Evaluationsmetrik dabei von entscheidender Bedeutung ist. Daher liegt ein Fokus dieser Studie auf der Entwicklung und dem Vergleich mehrerer Evaluationsansätze.

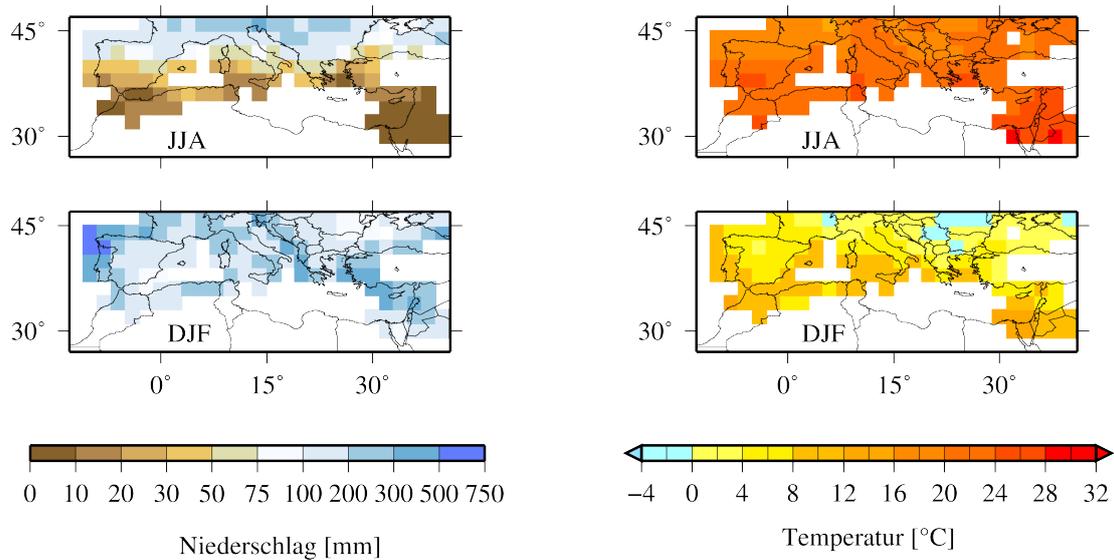
Die vorliegende Arbeit entstand im Rahmen des Projekts „Comparison of Metrics for Probabilistic Climate Change Projections of Mediterranean Precipitation“ (COMEPRO). Dieses wurde durch die Deutsche Forschungsgemeinschaft (DFG) gefördert und in Kooperation der Universitäten Würzburg und Augsburg durchgeführt. COMEPRO hat das Ziel, neue Metriken zur Einschätzung der Simulationsleistung aktueller Klimamodelle zu erstellen und zu vergleichen. Dazu liegt der Schwerpunkt auf der Region des Mittelmeerraums. Als Indikator für die Leistung der Modelle wurde für COMEPRO der Niederschlag gewählt. Basierend auf den individuellen Modellgewichten der Evaluationsmetriken sollen im Anschluss die Zukunftsprojektionen gewichtet und die Änderung des Unsicherheitsbereichs analysiert werden. Die vorliegende Arbeit ist damit als eine Auskopplung des Gesamtprojekts zu verstehen, die neben dem Niederschlag in der Hauptregion des Mittelmeerraums (siehe Abschnitt 1.4) auch die Temperatur als Bewertungskriterium der Modelle nutzt. Zusätzlich wird die Studie auf global verteilte Regionen ausgeweitet, um die Übertragbarkeit der Metriken zu prüfen. Außerdem wird so der Vergleich der Modellbewertungen für unterschiedliche Skalen und Regionen ermöglicht.

## 1.4 Untersuchungsgebiete

In diesem Abschnitt werden die Untersuchungsregionen vorgestellt, die zur Evaluation der Klimamodelle verwendet werden. Da die meisten Studien zeigen, dass die Leistungen der Modelle in verschiedenen Regionen variieren, wird hier nicht nur eine einzelne Region evaluiert. Der Schwerpunkt dieser Studie liegt auf der Analyse des nördlichen Teils des Mittelmeerraums. Dabei wird eine Region von 40° West bis 14° Ost und 20° Nord bis 29° Nord abgedeckt. Beim Mittelmeerraum handelt es sich um einen Brennpunkt des Klimawandels (GIORGI, 2006; PAETH ET AL., 2016). So hat die Region eine starke Sensibilität bzw. Vulnerabilität gegenüber den Folgen des anthropogenen Klimawandels. Beispielsweise sind im Mittelmeerraum die Temperatur und die Anzahl der Hitzetage stärker angestiegen als in vergleichbaren Regionen (GIORGI, 2006; BALDI ET AL., 2006). Zusätzlich führt das komplexe Niederschlagsverhalten mit hoher Variabilität sowie starken

Unterschieden hinsichtlich des Niederschlagserspartrags zwischen den Jahreszeiten (HERTIG & TRAMBLAY, 2016) zu einem hohen Anspruch an die Klimamodelle. Eine Vielzahl von Studien hat sich daher mit der Evaluation der Unsicherheit sowie der Analyse der Zukunftsprojektionen von Klimatologie und Extremereignissen dieser Regionen befasst (u.a. PAETH & HENSE, 2005; GIORGI & LIONELLO, 2008; HERTIG & JACOBET, 2008; KELLEY ET AL., 2012a). Damit bietet sich der Mittelmeerraum als Gradmesser für die Modellgüte an. Um Fehlwerte für die Evaluation auszuschließen, wurde zunächst eine Datenkontrolle der verwendeten E-OBS-Daten (siehe Kapitel 2.1) durchgeführt. Die gesamte Auswahl der Hauptuntersuchungsgebiete erfolgte im Verbund des COMEPRO-Projekts. So wurden für den Gesamtmittelmeerraum nur solche Werte verwendet, die über weniger als 12 Fehlmonate verfügen, wobei ein Fehlmonat durch mehr als zwei aufeinanderfolgende Tage mit Fehlwerten charakterisiert wurde (HERTIG & JACOBET, 2014). In den anderen Fällen wurde eine lineare Interpolation durchgeführt. Die Auswahl und Anzahl der Subregionen erfolgte anschließend durch Auswertung des sogenannten „Dominanzkriteriums“ (JACOBET, 1993; PHILIPP ET AL., 2007). Basierend auf s-modalen Empirischen Orthogonalfunktions-Analysen (Hintergründe dazu in Kapitel 4.4) wurden auf Grundlage der annualen Niederschlagssummen acht Variationszentren mit mindestens einer Gitterbox gefunden. Die Anzahl der nötigen Hauptkomponenten wurde mit Hilfe des Dominanzkriteriums bestimmt. Die Variationszentren werden im Folgenden als acht Subregionen neben dem Gesamtmittelmeerraum betrachtet. Für die Bezeichnung der Jahreszeiten werden die Anfangsbuchstaben der drei betreffenden Monate genannt („DJF“ entspricht: Dezember, Januar und Februar). Dies ist übliche Praxis in der Klimaforschung und vereinfacht das Verständnis, da sich ein erklärender Zusatz bezüglich der zutreffenden Hemisphäre erübrigt.

Abbildung 1.1 stellt die mittlere Temperatur und Niederschlagssumme für DJF und JJA über den Zeitraum von 1960–2009 dar. Aufgrund der Datengrundlage des Hauptreferenzdatensatz E-OBS (siehe Kapitel 2.1.1) wird die Analyse nur für Landflächen vorgenommen. Der Mittelmeerraum ist klimatisch durch eine hohe mittlere Jahrestemperatur sowie einem Niederschlagsmaximum in DJF gekennzeichnet. JJA ist dagegen größtenteils arid mit subtropischen Temperaturen (siehe Abbildung 1.1). Da Niederschlag in MAM und JJA nur in geringem Umfang auftritt, sind insbesondere Änderungen von DJF und SON sowie des Jahresniederschlags entscheidend in der Region. Den Untersuchungsgebieten wurde zur leichteren Verortung jeweils ein Name zugeteilt. Dieser orientiert sich teilweise grob am politischen Grenzverlauf, aber meist auch an der Geographie bzw. den Einflussfaktoren der jeweiligen Klimatologie. Im Ergebnis wurden die folgenden Namen gewählt: Ägäis, Nordatlantik (N.Atlantik), Italien, Naher Osten, Spanien, Schwarzes Meer, Nordafrika (N.Afrika) und Balkan. Zur Vereinfachung wird im

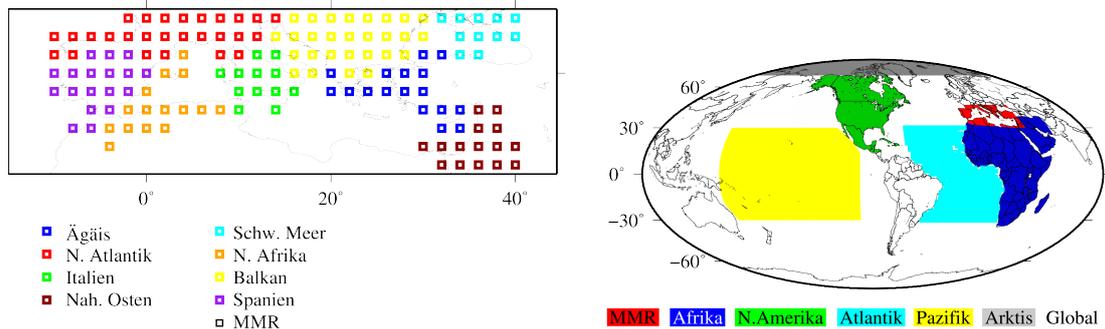


**Abbildung 1.1** Klimatologie des Mittelmeerraums. Niederschlagssumme und Temperaturmittel für DJF und JJA des Zeitraums 1960–2009.

Folgenden die Gesamtmittelmeerregion als MMR bezeichnet (siehe 1.2).

Zum Test der Übertragbarkeit der entwickelten Metriken wurden darüber hinaus sechs global verteilte Regionen von kontinentalem Maßstab untersucht. Bei deren Auswahl wurde Wert darauf gelegt sowohl Wasser- als auch Landflächen zu berücksichtigen. Daher sind die ersten beiden Erweiterungsregionen der tropische Atlantik und Pazifik. Beide erstrecken sich über die gesamte Fläche zwischen den Landmassen von Afrika und Südamerika bzw. Südamerika und Australien zwischen  $30^\circ$  Nord und  $30^\circ$  Süd. Gerade hier ist die Simulationsgüte aktueller Klimamodelle von sehr hoher Bedeutung, da diese Regionen eine entscheidende Rolle bei der Entstehung vieler Klimaphänomene und Extremereignisse einnehmen. Als Beispiel seien hier nur El Niño-Southern Oszillation (ENSO) für die Pazifikregion und die Entwicklung von Wirbelstürmen für den Atlantik genannt (SABA ET AL., 2016; CHEN ET AL., 2017).

Die Analyse der überwiegend maritimen Regionen wird durch die Arktis vervollständigt. Hier wird die gesamte Fläche nördlich des  $75.$  nördlichen Breitengrads untersucht. Neben den Wasserflächen werden hier auch Teile Nordgrönlands und Sibiriens betrachtet. Hierbei handelt es sich um eine von zwei Erweiterungsregionen in denen sowohl Land- als auch Seemassen enthalten sind. Die Arktis ist ein wichtiger Faktor bei der Erforschung des Klimawandels. So wird häufig die Meereisdicke am Nordpol als Beleg für den anthropogenen Klimawandel herangezogen oder der Neuschneeertrag bzw. die Abnahme der Gletscher von Grönland untersucht (KHAN ET AL., 2014; FELIKSON ET AL., 2017; KEEGAN ET AL.,



**Abbildung 1.2** Überblick über die Untersuchungsgebiete. Hauptuntersuchungsgebiete des Mittelmeerraums (9) und globale Erweiterungsregionen (6).

2014). Die zweite Region mit Land- und Wassermassen ist der gesamte Globus (Global). Dieser ist aus Gründen der Übersichtlichkeit in Abbildung 1.2 nicht explizit vermerkt, da sowohl alle Wasser- als auch Landflächen in die Evaluation eingehen. Die Entscheidung für dieses größtmögliche Gebiet beruht auf der allgemeinen Bedeutung der globalen Mitteltemperatur. Diese gilt in vielen Studien als Hauptindikator für die globale Änderung des Klimas (TANAKA ET AL., 2017; WARTENBURGER ET AL., 2017). Die drei weiteren Untersuchungsgebiete liegen ausschließlich über Landflächen. Die Region Afrika umfasst den afrikanischen Kontinent sowie die Arabische Halbinsel und südliche Teile des Iraks und Irans. Hiermit wird eine Region betrachtet, die größtenteils innerhalb der Subtropen liegt und daher einerseits über ein sehr hohes Niederschlagsaufkommen auf Höhe des Äquators, aber auch große aride Bereiche insbesondere im Norden und Süden verfügt. Die Temperaturen schwanken, über das Jahr gesehen, weniger stark als in anderen Teilen der Erde und liegen meist auf relativ hohem Niveau. Die letzte Erweiterungsregion (Nordamerika) umfasst den Nordamerikanischen Kontinent sowie Teile Zentralamerikas, beginnend ab 14° Nord. Diese Region wurde exemplarisch für die Außertropen gewählt. Hier ist ein kleiner Teil der Landfläche Nordkanadas deckungsgleich mit Teilen der Arktis-Region. Dies dient dem Zweck, in diesem Untersuchungsgebiet einen hohen Temperaturgradienten von Nord nach Süd mit dem Übergang von möglichst unterschiedlichen Klimazonen zu erhalten. Durch das ausgeprägte Relief und die langen heterogenen Küstenlinien wird mit Nordamerika ein sehr komplexes Untersuchungsgebiet für die Evaluation gewählt.

## 1.5 Ziele und Aufbau der Arbeit

Diese Arbeit widmet sich der Bewertung aktueller globaler und regionaler Klimamodelle zur Analyse der Unsicherheit probabilistischer Zukunftsprojektionen.

Die Ziele sind dabei in drei Schwerpunkte untergliedert. Die erste Zielsetzung stellt die Entwicklung neuer Metriken zur Evaluation von Klimamodellen dar. So werden basierend auf der Übereinstimmung historischer Modellläufe mit dem beobachteten Klima des 20. Jahrhunderts aktuelle Simulationen für Niederschlag und Temperatur evaluiert. Damit soll es ermöglicht werden, eine Aussage über die Zuverlässigkeit der verfügbaren Modellsimulationen (Läufe) aktueller globaler und regionaler Klimamodelle zu treffen. Ein besonderer Fokus wird dabei auf die Untersuchung der regionalen sowie saisonalen Konsistenz der Modellbewertungen gerichtet. Der zweite Schwerpunkt liegt auf dem Vergleich der Evaluationsergebnisse untereinander sowie auf der Analyse von Vor- und Nachteilen der Bewertungskriterien. Als dritter Schwerpunkt wird anschließend der Einfluss der metrikgewichteten Modellgruppen auf die Unsicherheit probabilistischer Klimaprojektionen untersucht. Modelle, die eine hohe Zuverlässigkeit gezeigt haben, werden bei der Bestimmung eines Unsicherheitsbereichs entsprechend stärker berücksichtigt als andere. Die aktuell standardmäßig verwendete gleichgewichtete Betrachtungsweise (KNUTTI, 2010) dient hier als Vergleichswert. Neben diesen drei Schwerpunkten werden weitere Fragestellungen wie beispielsweise die Übereinstimmung der Modellbewertung verschiedener Regionen, Jahreszeiten sowie Metriken beleuchtet oder der Einfluss unterschiedlicher Evaluationsgrößen sowie Variablen untersucht.

Die weitere Arbeit gliedert sich wie folgt: Kapitel 2 widmet sich der Darstellung der verwendeten Modell- und Referenzdaten. Anschließend werden in Kapitel 3 die statistischen Methoden und in Kapitel 4 die entwickelten Gewichtungsmetriken vorgestellt. Kapitel 5 stellt die Ergebnisse der vier Metrikblöcke hinsichtlich der verschiedenen Variablen und Hauptuntersuchungsregionen vor. Im Anschluss daran wird in Kapitel 6 der Einfluss der Metrikgewichtung auf die Unsicherheit der Zukunftsprojektionen untersucht. Kapitel 7 behandelt die Übertragbarkeit der Evaluationsmetriken auf die global verteilten Untersuchungsgebiete und stellt den Vergleich zu den Ergebnissen des Mittelmeerraums an. Schließlich werden in Kapitel 8 verschiedene Aspekte behandelt, die vor dem Hintergrund der zuvor veranschaulichten Ergebnisse die Integrität der Metriken untersuchen sollen. Dabei wird auch die Frage nach der Validität der Metriken sowie der Abhängigkeit der Modellbewertung von unterschiedlichen Referenzdaten erörtert. Danach folgt eine Diskussion und eine Zusammenfassung aller zuvor erläuterten Ergebnisse (Kapitel 9). Abgeschlossen wird die Arbeit mit einem kurzen Fazit sowie einem Ausblick (Kapitel 10) auf offene Fragestellungen, die sich im Rahmen dieser Untersuchung ergeben haben.

## 2 Datengrundlage

In diesem Kapitel werden die in dieser Arbeit verwendeten Datensätze vorgestellt. Sie werden in die Gruppe der Referenz- und Modelldatensätze unterteilt. Zu den Referenzdaten zählen zwei Beobachtungsdatsätze und eine Reanalyse.

### 2.1 Referenzdatensätze

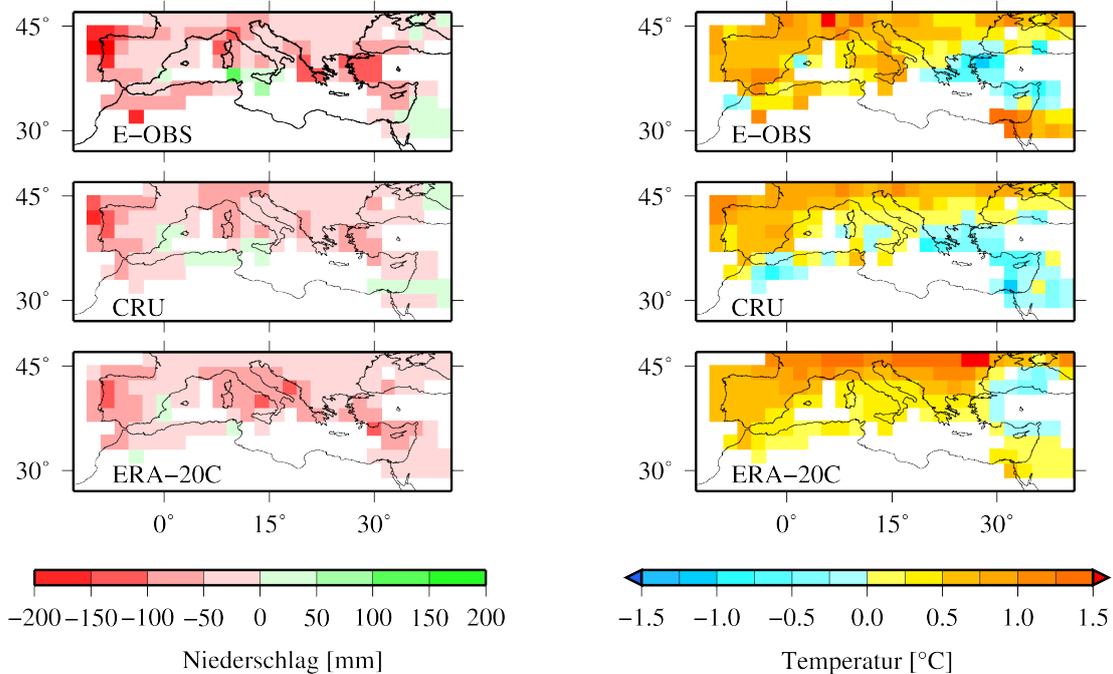
E-OBS wurde als Hauptreferenz gewählt. Die anderen beiden Referenzdatensätze, der Climate Research Unit (CRU) und die Reanalyse ERA-20C, dienen dem Vergleich der Ergebnisse sowie zur Übertragung der Metriken auf weitere Regionen. Von allen Datensätzen wurde die zum Zeitpunkt der Erstellung dieser Arbeit aktuelle Version verwendet (Tabelle 2.1).

**Tabelle 2.1** Übersicht über die verwendeten Referenzdatensätze.

Name	Abdeckung	Zeit	Referenz
E-OBS (V12)	Europa (Land)	1950–2015	HAYLOCK ET AL. (2008)
ERA-20C	Global	1900–2010	POLI ET AL. (2013)
CRU (TS3.23)	Global (Land)	1901–2014	MITCHELL & JONES (2005)

Die horizontale Auflösung wurde für alle drei Datensätze mit Hilfe von REMAPCON CDO (SCHULZWEIDA ET AL., 2009) auf ein gemeinsames  $2^\circ \times 2^\circ$  Gitter interpoliert. Damit wird die beste Vergleichsmöglichkeit untereinander sowie mit den Modelldaten ermöglicht. Abbildung 2.1 zeigt die Niederschlags- und Temperaturtrends für DJF der drei Referenzdatensätze im Untersuchungszeitraum 1960–2009. Auch wenn hier gewisse Unterschiede zutage treten, zeigt sich zumindest für die Gesamtregion, dass es beim Niederschlag größtenteils zu einer Abnahme kommt. Dies wird besonders deutlich im Westen. Im Osten zeigen sich abhängig von der Region auch leicht gegensätzliche Trends. Bei der Temperatur kommt es dagegen überwiegend zu einer Zunahme. Wobei für den Osten alle Datensätze übereinstimmend eine leichte Temperaturabnahme zeigen. Dieser Eindruck zeigt bereits, dass zwischen Referenzdaten gewisse Unsicherheiten bestehen (GEHNE ET AL., 2016). Diese Feststellung ist entscheidend für die Interpretation

der Ergebnisse der Modellevaluation. Um dennoch möglichst zuverlässige Einschätzungen zu bieten, wird die Stabilität bzw. Vulnerabilität der Ergebnisse bezogen auf die verschiedenen Referenzdaten ebenfalls analysiert (siehe hierfür Kapitel 8.4).



**Abbildung 2.1** Temperatur- [°C] und Niederschlagstrend [mm] aller Referenzdatensätze für DJF im Zeitraum 1960 – 2009.

### 2.1.1 E-OBS

Die Hauptreferenz E-OBS deckt die Landfläche Europas mit dem äußersten Norden Afrikas bei 21° südlicher bis 75° nördlicher Breite und 49° westlicher bis 68° östlicher Länge ab. E-OBS basiert auf über 2300 Messstationen. Dabei ist die Stationsdichte in Irland, den Niederlanden sowie der Schweiz am höchsten. Durch drei Verarbeitungsschritte und Fehlerkorrekturen werden die Werte auf ein 0,5° x 0,5° Gitter in der Fläche interpoliert. Damit können sie ähnlich wie die Daten von Modellen und Reanalysen verwendet werden. Aufgrund der geringeren Anzahl an Messstationen muss jedoch eine erhöhte Unsicherheit im Süden des Mittelmeerraums angenommen werden (HOFSTRA ET AL., 2009). Da die E-OBS-Daten dennoch zu den zuverlässigsten stationsbasierten Datensätzen für den Mittelmeerraum zählen, werden sie in aktuellen Studien aufgrund der ansonsten guten Datenlage und hohen Auflösung häufig verwendet (u.a. KOTLARSKI ET AL.,

2014; KATRAGKOU ET AL., 2015; HERTIG ET AL., 2013; HERTIG & JACOBET, 2014; PAXIAN ET AL., 2013, 2014). Die Daten sind für eine Vielzahl verschiedener Klimavariablen verfügbar. In dieser Arbeit werden die bodennahe monatliche Temperatur und der Niederschlag verwendet. E-OBS deckt aktuell ein Zeitintervall von 1950-2015 ab, wobei die Daten immer wieder aktualisiert werden, sodass sich der Stand nach der Veröffentlichung dieser Arbeit bereits verändert haben dürfte. Neben dem Zeitraum ändert sich auch die Verfügbarkeit der Stationen. Nach HOFSTRA ET AL. (2009) liegt die höchste Dichte an verfügbaren Stationen im Zeitraum 1961–1990. Da PAXIAN ET AL. (2013) nahe legen, dass eine Evaluation eine möglichst große Periode von deutlich mehr als 30 Jahre abdecken sollte, wurde als Untersuchungszeitraum ein 50-jähriges Intervall von 1960–2009 festgelegt. Zusätzlich wäre es von großem Interesse, zwei unabhängige Zeitintervalle zu analysieren, um die zeitliche Konsistenz der Modellbewertungen einschätzen zu können. Da die Datenverfügbarkeit dies nicht unter den genannten Bedingungen (zwei 50-jährige Intervalle) zulässt, wurde ein zweites Zeitfenster gewählt, das allerdings nur um 10 Jahre verschoben ist und den Zeitraum 1950–1999 abdeckt. Auch hier ist der Abschnitt mit der höchsten Stationsdichte eingeschlossen, der die beste Datenqualität insbesondere für die Regionen im Süden des Mittelmeerraums bietet. Der Start des zweiten 50-jährigen Zeitintervalls wurde um 10 Jahre nach vorne verlegt, da dieser Zeitraum auch von den beiden ergänzenden Referenzdaten CRU und insbesondere der Reanalyse ERA-20C (siehe folgende Abschnitte) vollständig abgedeckt wird. Allerdings sind diese beiden Referenzzeiträume mit einer Schnittmenge von 80 % nicht unabhängig. In der Untersuchung der zeitlichen Konsistenz der Modellevaluation (Kapitel 8.2) werden sie folglich eher als leicht abgewandelte Untersuchungszeiträume behandelt. Für die Anwendung der Metriken und Analysen hat dies zwar keine Relevanz, allerdings muss bei der Interpretation die Schlussfolgerung der Ergebnisse entsprechend zurückhaltend durchgeführt werden.

### 2.1.2 Climate Research Unit

Ergänzend für den Mittelmeerraum wurde zu den E-OBS-Daten als weitere Referenz der Datensatz der Climate Research Unit (CRU) verwendet. Er gehört aufgrund der großen Anzahl an Informationsquellen und seiner globalen Abdeckung zu einem der meistgenutzten Beobachtungsdatensätze in der Klimaforschung, wie die hohe Zahl der verwendenden Studien nahe legt (u.a. ASADI ZARCH ET AL., 2015; WU ET AL., 2015; RING ET AL., 2016). Er deckt, bis auf die Antarktis, die gesamten Landflächen des Globus mit einer Auflösung von  $0,5^\circ \times 0,5^\circ$  ab (NEW ET AL., 1999, 2000). Wie bei den E-OBS-Daten stehen zusätzlich zu den in dieser Arbeit verwendeten monatlichen Niederschlags- und Temperaturinfor-

mationen noch weitere Klimavariablen zur Verfügung (HARRIS ET AL., 2014). CRU bietet in der verwendeten Version monatliche Daten für den Zeitraum 1901 bis 2010 (vgl. Tabelle 2.1). Aufgrund der Einschränkung auf Landflächen wird CRU hier nur ergänzend zu E-OBS zur Evaluierung der Modelldaten und zum Abgleich der Metriken in der Region des Mittelmeerraums sowie deren Unterregionen verwendet.

### 2.1.3 ERA-20C Reanalyse

ERA-20C ist ein Reanalyse-Datensatz, der basierend auf dem „Integrated Forecast System cy38r1 (ECMWF 2013)“ erstellt wurde (POLI ET AL., 2013). Dabei wurde auf eine Vielzahl von Beobachtungsdaten bzw. Stationsmessungen zurückgegriffen, um das Reanalyse-Modell zu verbessern und zu korrigieren. Das Ergebnis ist eine lückenlose zeitliche und räumliche Abdeckung für eine Vielzahl atmosphärischer Parameter. ERA-20C bietet atmosphärische Daten für den gesamten Globus über den Zeitraum 1900–2010 mit einer räumlichen Auflösung von ca. 125km (POLI ET AL., 2016). Im Gegensatz zu den beiden zuvor genannten Beobachtungsdatensätzen steht hinter ERA-20C ein Modell und besitzt damit auch eine vertikale Auflösung von 91 Schichten. Allgemein gilt ERA-20C als sehr zuverlässig für die Verwendung zur Evaluationen im 20. Jahrhundert. DONAT ET AL. (2016) und DITTUS ET AL. (2016) bescheinigen ERA-20C insbesondere zuverlässige Ergebnisse für Extremereignisse ab dem Jahr 1950. Der große Vorteil, den die Reanalyse bietet, ist, dass sie lückenlos den gesamten Globus für das 20. Jahrhundert abdeckt. Zur Analyse der Übertragbarkeit der Metriken auf andere Regionen, insbesondere Ozeane, ist dies unverzichtbar.

## 2.2 Analyisierte Klimamodelldaten

Die in dieser Arbeit verwendeten Klimamodelle lassen sich ebenfalls in zwei Kategorien unterteilen: Globale und regionale Klimamodelle. Dabei werden auch die globalen Modelle nochmals in zwei Generationen untergliedert. Die Modelle des vierten Sachstandberichts von 2007 (AR4) und die des fünften von 2013 (AR5). Sie sind Teil des „Coupled Model Intercomparison Projekts“ Phase 3 (CMIP3, 2007) und Phase 5 (CMIP5, 2013). Dieses wurde vom IPCC ins Leben gerufen, um die Erkenntnisse unterschiedlicher Forschergruppen und Modellierer durch eine gemeinsame Auswertung zu verbessern. Neben den globalen Klimamodellen werden auch aktuelle regionale Klimamodelle des „Coordinated Regional Downscaling Experiment“ (CORDEX) analysiert. Da diese jedoch jeweils nur einen bestimmten Abschnitt der Erdoberfläche abdecken, werden sie hier ausschließlich

für die Mittelmeerregion verwendet. Alle globalen und regionalen Modellsimulationen wurden mit Hilfe einer invertierten Distanzgewichtung (IDW) auf ein gemeinsames  $2^\circ \times 2^\circ$  Gitter interpoliert (BABAK & DEUTSCH, 2009). Dies ist insbesondere für die Modelldaten von großer Bedeutung, da die Unterschiede zwischen den einzelnen Modellen zwischen  $5^\circ$  (INM-CM3.0, CMIP3) und  $0,11^\circ$  (z.B. SMHI-RCA4, CORDEX) liegen. Durch die Interpolation auf  $2^\circ \times 2^\circ$  kann es zu einem gewissen Informationsverlust von höher aufgelösten Modellen kommen. Dieser Nachteil wird dadurch relativiert, dass kein Modell die jeweilige Originalauflösung in diese Untersuchung einbringen kann und daher bei allen eine leichte Anpassung stattgefunden hat. Für die Modelle hat sich die IDW als praktikabler erwiesen als die REMAPCON-Interpolation der Referenzdaten. Zur Überprüfung wurden die Referenzdaten testweise mit beiden Interpolationen betrachtet und verschiedene Evaluationen wiederholt. Das Ergebnis zeigte jedoch keine oder nur minimale Unterschiede, die sich nicht auf die Einschätzungen der Simulationsgüte auswirkten.

Insgesamt werden 62 globale Klimamodelle aus CMIP3 (24) bzw. CMIP5 (38) und 18 regionalen CORDEX Simulationen analysiert. Wichtig für die Evaluierung ist es, durchgehende Simulationen (sog. transiente Läufe) zu verwenden, die sowohl für das 20. als auch für das 21. Jahrhundert verfügbar waren. Da jedoch nicht alle Simulationen für jeweils beide Emissionsszenarien zur Verfügung stehen, schwankt die Zahl der analysierten Modelle bzw. Simulationen zwischen den Variablen und betrachteten Szenarien. Generell gilt, dass verschiedene Simulationen bzw. Läufe eines Modells als Ensemble bezeichnet werden.

### 2.2.1 Coupled Model Intercomparison Projekt Phase 3

Da generell versucht wurde, eine möglichst große Anzahl aller zur Verfügung stehenden globalen Klimamodelle (GKM) zu verwenden, wurde auch auf die Modelle von 2007 zurückgegriffen. Dass neue Modellgenerationen nicht zwingend für eine bessere Simulationsgüte stehen, wurde bereits mehrfach bestätigt (KNUTTI & SEDLÁČEK, 2012; RAMESH & GOSWAMI, 2015). Dies rechtfertigt auch die weitere Berücksichtigung der Modelle von 2007. Dabei nutzen alle Modelle gekoppelte Systeme von Ozean und Atmosphäre, um das Klima bestmöglich zu simulieren (IPCC, 2007c). Dies ist gerade für globale Phänomene wie beispielsweise ENSO eine wichtige Voraussetzung. Alle hier verwendeten Modelle von CMIP3 (Tabelle 2.2) verfügen über historische Simulationen, die das Klima des 20. Jahrhunderts und Projektionen für das 21. Jahrhundert berechnen. Um der Unsicherheit über die zukünftigen Entwicklungen des Klimas Rechnung zu tragen, werden für alle Modelle unterschiedliche Emissionsszenarien analysiert. Es muss angenommen werden, dass einige Aspekte des Klimasystems auch im nächsten Jahrhundert

relativ konstant bleiben werden (bspw. die Sonnenaktivität). Andere Informationen, die für das Emissionsverhalten von Treibhausgasen relevant sind, wie sozioökonomische und demographische Entwicklungen sowie technologische Fortschritte, werden für CMIP3 anhand von sechs Emissionsszenarien, den „Special Report on Emission Scenarios“ (SRES), antizipiert: A1B, A1FI, A1T, A2, B1 und B2 (NAKICENOVIC ET AL., 2000). In dieser Arbeit werden das SRES A1B- (A1B) und SRES A2- (A2) Zukunftsszenario für das 21. Jahrhundert verwendet. A1B geht von einer gleichmäßigen Nutzung aller verfügbaren Energiequellen aus. Bei der Weltbevölkerung wird bis Mitte des 21. Jahrhunderts ein Wachstum erwartet, das anschließend stagniert oder leicht abnimmt. Zusätzlich können technologische Fortschritte zu einer Verbesserung der Effizienz beispielsweise in der Industrie führen. Dieses Szenario wird in dieser Arbeit als das optimistischere der beiden CMIP3-Emissionsszenarien verstanden. A2 geht von einer heterogenen Welt aus mit einer über die Mitte des 21. Jahrhunderts weiter steigenden Weltbevölkerung sowie langsameren technologischen Entwicklungen bei gleichzeitig regionalem Schwerpunkt des Wirtschaftswachstums (NAKICENOVIC ET AL., 2000). Dieses Szenario wird im Folgenden als das realistischere der beiden Szenarien verstanden.

**Tabelle 2.2** Analyisierte globale Klimamodelle von CMIP3 (RANDALL ET AL., 2007). Modelle mit dynamischer Simulation der drei ausgewählten Forcings sind mit \* markiert.

Modelle	Niederschlag		Temperatur		Referenz
	A1b	A2	A1b	A2	
BCCR BCM2.0	1	1	1	1	DÉQUÉ ET AL. (1994)
CGCM3.1 (T47)*	2-6	2-6	2-6	2-6	McFARLANE ET AL. (1992)
CGCM3.1 (T63)*	7	-	7	-	McFARLANE ET AL. (1992)
CNRM-CM3	8	7	8	7	DÉQUÉ ET AL. (1994)
GFDL-CM2.0*	9	8	9	8	FREIDENREICH ET AL. (2004)
GFDL-CM2.1*	10	9	10	9	FREIDENREICH ET AL. (2004)
GISS-AOM	11-12	-	11-12	-	RUSSELL ET AL. (1995)
GISS-EH*	13-15	10	13-15	10	SCHMIDT ET AL. (2006)
GISS-ER*	16-17	-	16-20	-	SCHMIDT ET AL. (2006)
FGOALS-g1.0	18-20	-	21-23	-	WANG ET AL. (2004)
INM-CM3.0	21	11	24	-	ALEKSEEV ET AL. (1998)
IPSL-CM4(LMDZ)	22	12	25	12	HOUDIN ET AL. (2006)
INGV-SXG*	23	13	26	13	SCOCIMARRO ET AL. (2007)
MIROC3.2(hr)*	24	-	27	-	ABE-OUCHI ET AL. (2004)
MIROC3.2(mr)*	25-27	14-16	28-30	14-16	ABE-OUCHI ET AL. (2004)
MRI-CGCM2.3.2	28-32	17-21	31-35	17-21	SHIBATA ET AL. (1999)
ECHO-G	33-35	22-24	36-38	22-24	ROECKNER ET AL. (1996)
CSIRO-Mk3.0	36	25	39	25	GORDON ET AL. (2002)
CSIRO-Mk3.5	37	26	40	26	GORDON ET AL. (2002)
ECHAM5/MPI-OM	38-41	27-29	41-44	27-29	ROECKNER ET AL. (2003)
NCAR-CCSM3*	42-48	30-33	45-51	30-33	COLLINS ET AL. (2006)
NCAR-PCM1	49-52	34-37	52-55	34-37	KIEHL ET AL. (1998)
UKMO-HadCM3	53	38	56	3	POPE ET AL. (2000)
UKMO-HadGEM1*	54	39	57	39	MARTIN ET AL. (2004)

In Tabelle 2.2 sind alle verwendeten Modelle und zugehörigen Simulationen aufgelistet. Dabei gilt, dass viele Modelle über mehrere Läufe mit unterschiedlichen Eingangsparametern verfügen. Sowohl für Temperatur als auch Niederschlag haben alle Modelle Zukunftsprojektionen für das A1B-, nicht aber für das A2-Szenario. Um für alle Simulationen eine Evaluationsgrundlage zu schaffen, werden für die letzten Jahre des Evaluationsintervalls 1960–2009, die nicht mehr im historischen Zeitraum aller Simulationen enthalten sind, die Läufe von A1B verwendet. Insgesamt stehen 54 (57) Simulationen für Niederschlag (Temperatur) von A1B und 39 (39) von A2 zur Verfügung.

Die CMIP3-Modelle unterscheiden sich teilweise stark in ihrer physischen Beschaffenheit, wie zum Beispiel ihrer vertikalen und horizontalen Auflösung oder unterschiedlichen Parametrisierungen (RANDALL ET AL., 2007). Eine noch größere Rolle bei der Zuverlässigkeit der Simulationen könnten die verschiedenen Strahlungsantriebe (eng. Radiative Forcings) spielen, die in den untersuchten Modelle zum Einsatz gekommen sind. Bei Forcings handelt es sich um ein Maß, das das Verhältnis von einfallender zu ausgehender Strahlung in der Atmosphäre angibt (IPCC, 2007c). Während manche Modelle über statische Bedingungen beispielsweise für das Forcing der Solarvariabilität verfügen, simulieren andere Modelle dies dynamisch. Eine genaue Auflistung, welche Modelle über welche dynamischen Forcings verfügen, ist im Sachstandsbericht von 2007 (IPCC, 2007c) auf S. 756 zu finden. Um einen Teil dieser Unterschiede zu berücksichtigen, wird in der Auswertung neben der Gesamtanalyse der CMIP3-Modelle eine Aufteilung in zwei Gruppen durchgeführt. Während die eine Gruppe statische Bedingungen für Meersalz- und Vulkan-Aerosole sowie solare Variabilität simuliert, wird dies in der anderen Gruppe dynamisch umgesetzt. Damit soll geprüft werden, ob die zusätzlichen dynamischen Forcings der Modelle eine insgesamt bessere Bewertung der Metriken zur Folge haben. Diese Unterscheidung wird nur für CMIP3 vorgenommen. Die Modelle der neueren Generation (siehe 2.2.2) verfügen in diesem Zusammenhang alle über dynamische Forcings.

### 2.2.2 Coupled Model Intercomparison Projekt Phase 5

Im aktuellen Sachstandsbericht von 2013 wurden die Modelle von CMIP5 verwendet. Dabei handelt es sich in vielen Fällen um Weiterentwicklungen der Modelle von CMIP3. Zusätzlich sind auch neue Modelle aufgenommen worden, sodass hier 38 Modelle (vgl. Tabelle 2.3) im Gegensatz zu 24 von CMIP3 zur Verfügung stehen. Eine wichtige Verbesserung des aktuellen MME ist insgesamt eine höhere regionale Auflösung (FLATO ET AL., 2013; SCHIEMANN ET AL., 2014). Hierdurch können Vorgänge auf kleinen Skalen, die gerade für die Simulation von Niederschlag eine Rolle spielen, besser erfasst werden.

**Tabelle 2.3** Analysierte globale Klimamodelle von CMIP5 (FLATO ET AL., 2013).

Modelle	Niederschlag		Temperatur		Referenz
	RCP4.5	RCP8.5	RCP4.5	RCP8.5	
ACCESS1-0	1	1	1	1	DIX ET AL. (2013)
ACCESS1-3	2	2	2	2	DIX ET AL. (2013)
BCC-CSM1.1	3	3	3	3	WU ET AL. (2008, 2010)
BCC-CSM1.1(m)	4	4	4	4	WU ET AL. (2008, 2010)
CanESM2	5-9	5-9	5-9	5-9	SCINOCCA ET AL. (2008)
CCSM4	10-15	10-15	10-15	10-15	GENT ET AL. (2011)
CESM1-BGC	16	16	16	16	HURRELL ET AL. (2013)
CESM1-CAM5	17-19	17-19	17-19	17-19	HURRELL ET AL. (2013)
CMCC-CM	20	21	20	21	FOGLI ET AL. (2009)
CMCC-CMS	21	22	21	22	FOGLI ET AL. (2009)
CMCC-CM5	22	23-27	22	23	VOLDOIRE ET AL. (2013)
CSIRO-Mk3-6-0	23-32	28-37	23-32	24-33	GORDON ET AL. (2002)
CSIRO-Mk3L-1-2	33-35	-	33-35	-	GORDON ET AL. (2002)
EC-EARTH	36-39	38-42	-	-	HAZELEGER ET AL. (2012)
FGOALS-g2	40	43	36	-	LI ET AL. (2013)
FIO-ESM	-	-	37	-	QIAO ET AL. (2013)
GFDL-CM3	41	45	38	34	DONNER ET AL. (2011)
GFDL-ESM2G	42	46	39	35	DUNNE ET AL. (2012)
GFDL-ESM2M	43	47	40	36	DUNNE ET AL. (2012)
GISS-E2-H-CC	44	48	-	-	SCHMIDT ET AL. (2006, 2014)
GISS-E2-H	45-60	49-53	-	-	SCHMIDT ET AL. (2006, 2014)
GISS-E2-R-CC	61	54	-	-	SCHMIDT ET AL. (2006, 2014)
GISS-E2-R	62-78	55-59	-	-	SCHMIDT ET AL. (2006, 2014)
HadGEM2-AO	79	60	41	37	COLLINS ET AL. (2011)
HadGEM2-CC	80	61	-	-	COLLINS ET AL. (2011)
HadGEM2-ES	81-84	62-65	-	-	COLLINS ET AL. (2011)
INMCM4	85	66	42	38	VOLODIN ET AL. (2010)
IPSL-CM5A-LR	86-89	67-70	43-46	39-42	DUFRESNE ET AL. (2013)
IPSL-CM5A-MR	90	71	47	43	DUFRESNE ET AL. (2013)
IPSL-CM5B-LR	91	72	48	44	DUFRESNE ET AL. (2013)
MIROC5	92-94	73-75	49-51	45-47	WATANABE ET AL. (2010)
MIROC-ESM-CHEM	95	76	52	48	WATANABE ET AL. (2011)
MIROC-ESM	96	77	53	49	WATANABE ET AL. (2011)
MPI-ESM-LR	97-99	78-80	54-56	50-52	STEVENS ET AL. (2013)
MPI-ESM-MR	100-102	81	57-59	53	STEVENS ET AL. (2013)
MRI-CGCM3	103	82	60	54	YUKIMOTO ET AL. (2012)
NorESM1-ME	104	83	61	55	BENTSEN ET AL. (2013)
NorESM1-M	105	84	62	56	BENTSEN ET AL. (2013)

Auch wenn es teilweise zu widersprüchlichen Ergebnissen kommt, ist die Mehrheit der Forschergruppen der Ansicht, dass CMIP5 dem Vorgänger CMIP3 in vieler Hinsicht überlegen ist (vgl. Abschnitt 1.3). Angesichts der unterschiedlichen Resultate wäre es jedoch voreilig, diese Untersuchungen ausschließlich auf CMIP5 zu konzentrieren, da hierdurch wichtige Vergleichsmöglichkeiten verloren gehen würde. Darüber hinaus unterscheiden sich die Zukunftsprojektionen der CMIP5-Modelle deutlich von den CMIP3 Szenarien. Der Schwerpunkt im aktuellen Sachstandsbericht wird auf der Entwicklung der Konzentration von klimawirksamen

Treibhausgasen, den „Representative Concentration Pathways“ (RCP) (MOSS ET AL., 2010), gelegt. So werden zum Beispiel neben CO<sub>2</sub> auch andere klimarelevante Gase wie Ozon, Methan etc. in die Projektionen integriert. Auf Grundlage der Veränderung dieser Gase wird der zusätzliche Strahlungsantrieb im Jahr 2100 im Vergleich zum vorindustriellen Zeitraum, der Mitte des 19. Jahrhunderts, in W/m<sub>2</sub> berechnet. Da dieser Wert direkt in den Namen einfließt (z.B. 4.5 W/m<sub>2</sub>), lässt sich daran die jeweilige Kategorie der Szenarien ablesen. Insgesamt wurden für CMIP5 vier Szenarien entwickelt: RCP2.6, RCP4.5, RCP6.0, RCP8.5. In dieser Arbeit wurden die Zukunftsprojektionen basierend auf den RCP4.5- und RCP8.5-Szenarien untersucht. Dabei entspricht RCP8.5 in etwa dem antizipierten Emissionsverhalten von A2 (CMIP3) mit einem langfristig hohen Ausstoß von Treibhausgasen bis Ende des 21. Jahrhunderts. RCP4.5 wird zusätzlich als optimistischeres Szenario untersucht. Dabei ist RCP4.5 noch optimistischer als das CMIP3-Äquivalent A1B. Wie für CMIP3 gilt auch für CMIP5, dass nicht alle Modelle über alle Emissionszenarien verfügen. Tabelle 2.3 zeigt die Übersicht über alle verwendeten Modelle und Simulationen von CMIP5. Dabei stehen für Niederschlag 105 (84) Simulationen für RCP4.5 (RCP8.5) und 84 (56) für Temperatur zur Verfügung. Analog zu CMIP3 gilt für CMIP5 und CORDEX, dass für die Evaluation im Zeitraum 1960–2009 die Läufe von RCP4.5 verwendet werden, um die fehlenden letzten Jahre der Simulationen des 20. Jahrhunderts zu ergänzen.

### **2.2.3 Coordinated Regional Downscaling Experiment**

Der Trend zu räumlich höher aufgelösten Modellen (FLATO ET AL., 2013; RANDALL ET AL., 2007) hat zur Gründung des Coordinated Regional Downscaling Experiment (CORDEX) geführt (GIORGI ET AL., 2009). Dabei handelt es sich um das Folgeprojekt des älteren PRUDANCE (CHRISTENSEN ET AL., 2007) und ENSEMBLES (VAN DER LINDEN & MITCHELL, 2009). CORDEX wurde durch das Weltklima-Forschungsprogramm (WCRP) 2009 initiiert, um ein koordiniertes Netzwerk für hochaufgelöste regionale Klimamodelle (RKM) zu schaffen. RKM nutzen die Daten von GKM und rechnen sie auf ein sehr hoch aufgelöstes Raster von aktuell bis zu 0,11° herunter. Hierdurch werden neue Möglichkeiten beispielsweise bei der Differenzierung von Orographie oder Strömungsverhalten möglich, die bei der im Vergleich dazu relativ niedrigen Auflösung der GKM nicht wahrnehmbar sind. Verschiedene Studien haben dazu Ergebnisse erbracht, dass der Informationsgewinn regionaler Klimamodelle beispielsweise durch die höhere Auflösung die Simulationsgüte in vielen Fällen verbessert (PAETH & MANNIG, 2013; FESER, 2006; FESER ET AL., 2011; PRÖMMEL ET AL., 2010; PREIN ET AL., 2016).

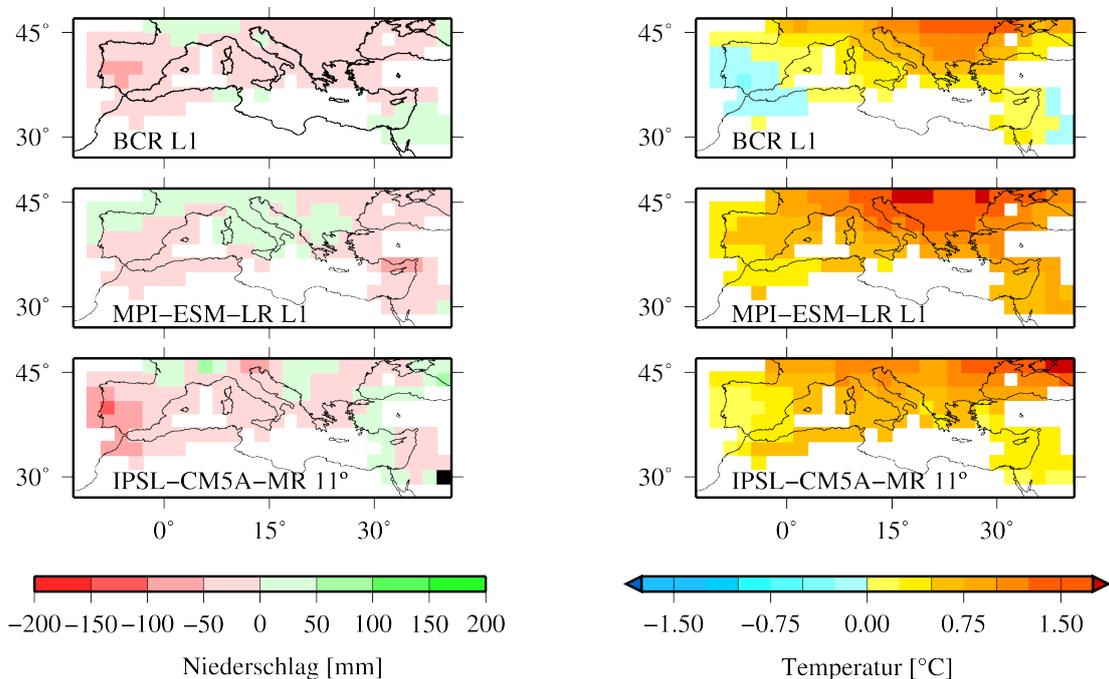
Die räumlichen Auflösungen der RKM in dieser Arbeit liegen bei  $0,11^\circ$  und  $0,44^\circ$ . In dieser Arbeit werden die Simulationen der CORDEX-Europaregion verwendet (JACOB ET AL., 2014), um die regionalen Klimamodelle zu evaluieren und mit den deutlich niedriger aufgelösten Modellen von CMIP3 und CMIP5 zu vergleichen. Die Zukunftsszenarien sind dabei äquivalent zu CMIP5 und haben entsprechend auch die gleiche Bezeichnung. Insgesamt werden 18 Simulationen untersucht, die sich teilweise nur im Antrieb oder in der räumlichen Auflösung unterscheiden. Globale und regionale Modelle nutzen häufig gleiche Komponenten zur Simulation des Klimas. So ist generell von keiner Unabhängigkeit der Modelle auszugehen. Diese Abhängigkeit ist für diese Studie jedoch nicht weiter von Bedeutung, da die Modelle im Anschluss abhängig von ihrer Leistung innerhalb des jeweiligen MME gewichtet werden. Im Gegensatz zu CMIP3 und CMIP5 sind alle 18 Modelle sowohl für Niederschlag als auch Temperatur für beide Emissionsszenarien verfügbar (siehe Tabelle 2.4).

**Tabelle 2.4** Analyisierte regionale Klimamodelle von CORDEX: RCA (SAMUELSSON ET AL., 2011); HIRHAM (CHRISTENSEN ET AL., 1998); KNMI (VAN MEIJGAARD ET AL., 2008).

GKM	RKM	Auflösung	NS/T
CNRM-CERFACS-CNRM-CM5	SMHI-RCA4	$0,11^\circ$	1
ICHEC-EC-EARTH	SMHI-RCA4	$0,11^\circ$	2
ICHEC-EC-EARTH	DMI-HIRHAM5	$0,11^\circ$	3
IPSL-IPSL-CM5A-MR	SMHI-RCA4	$0,11^\circ$	4
MOHC-HadGEM2-ES	SMHI-RCA4	$0,11^\circ$	5
MPI-M-MPI-ESM-LR	SMHI-RCA4	$0,11^\circ$	6
CCCma-CanESM2	SMHI-RCA4	$0,11^\circ$	7
CNRM-CERFACS-CNRM-CM5	SMHI-RCA4	$0,11^\circ$	8
CSIRO-QCCCE-CSIRO-Mk3-6-0	SMHI-RCA4	$0,11^\circ$	9
ICHEC-EC-EARTH	SMHI-RCA4	$0,44^\circ$	10
ICHEC-EC-EARTH	KNMI-RACMO22E	$0,44^\circ$	11
ICHEC-EC-EARTH	DMI-HIRHAM5	$0,44^\circ$	12
IPSL-IPSL-CM5A-MR	SMHI-RCA4	$0,44^\circ$	13
MIROC-MIROC5	SMHI-RCA4	$0,44^\circ$	14
MOHC-HadGEM2-ES	SMHI-RCA4	$0,44^\circ$	15
MPI-M-MPI-ESM-LR	SMHI-RCA4	$0,44^\circ$	16
NCC-NorESM1-M	SMHI-RCA4	$0,44^\circ$	17
NOAA-GFDL-GFDL-ESM2M	SMHI-RCA4	$0,44^\circ$	18

Der Hauptuntersuchungszeitraum in dieser Arbeit wurde, mit Blick auf die Datenlage der Referenzdaten sowie der klimatischen Charakteristik des Mittelmeerraums, auf das Intervall 1960–2009 gelegt. Dieser Zeitraum wird von den GKM vollständig abgedeckt. Die CORDEX-Modelle decken allerdings nur einen historischen Zeitraum seit 1970 ab. Dies hat zur Folge, dass nur ein Zeitraum von 40 Jahren für die regionalen Klimamodelle untersucht werden kann. Um deswegen nicht vollständig auf die Evaluation der CORDEX-Modelle und damit dem Vergleich zu CMIP3 und CMIP5 verzichten zu müssen, wurde die Evalua-

tion in dem 40-jährigen Zeitraum von 1970–2009 vorgenommen. Daher müssen die Evaluationsergebnisse der RKM im Vergleich zu den GKM mit Vorbehalt betrachtet werden. Aufgrund der regionalen Abdeckung gilt, dass die CORDEX-Simulationen über einen Datenpunkt weniger im äußersten Südosten verfügen. Der räumliche Abdeckung der verwendeten CORDEX-Simulationen deckt dieses Feld nicht ab. Für den MMR bedeutet dies, dass hier nur 142 (anstatt 143) und im Nahen Osten 14 (statt 15) Datenpunkte in die Untersuchung eingehen. Wie für CMIP3 und CMIP5 werden auch alle Simulationen von CORDEX auf das deutlich niedriger aufgelöste Gitter von  $2^\circ \times 2^\circ$  interpoliert, um die beste Vergleichbarkeit zu gewährleisten.



**Abbildung 2.2** Temperatur- (rechts) und Niederschlagstrend (links) einzelner Modelle jedes der drei MME. Exemplarisch dargestellt ist DJF über den Zeitraum 1960–2009 (CORDEX 1970–2009).

Abbildung 2.2 gibt einen Eindruck von den Unterschieden zwischen den Simulationen. Da der Schwerpunkt in dieser Arbeit auf der Evaluation einzelner Modelle liegt, sind hier zur Verdeutlichung der Unterschiede nicht der MME-Wert, sondern exemplarisch einzelne Simulationen von CMIP3, CMIP5 und CORDEX dargestellt. Alle gezeigten Simulationen sind auf das  $2^\circ \times 2^\circ$  Gitter interpoliert. Dargestellt ist jeweils der lineare Trend (siehe Kapitel 3.6) über den Zeitraum 1960–2009 (1970–2009 für IPSL-CM5A-MR). Die Darstellung des Trends zeigt sowohl für Niederschlag als auch für Temperatur ein relativ ähnliches Bild zwi-

schen den Simulationen. Beim Niederschlag sind über einen Großteil des Mittelmeerraums von allen Simulationen Abnahmen dargestellt, diese unterscheiden sich jedoch teilweise erheblich im Betrag. Darüber hinaus zeigen MPI-ESM-LR und BCCR-BCM2.0 im Norden leichte Niederschlagszunahmen. Die stärkste Niederschlagsabnahme (120mm, IPSL-CM5A-MR) im Untersuchungszeitraum wird im Westen über Spanien übereinstimmend von allen Simulationen dargestellt. Auch beim Temperaturtrend zeigt sich ein homogenes Bild. Alle drei Simulationen zeigen hier überwiegend eine Zunahme. Das Maximum wird jeweils für die Balkan-Region bzw. den nördlichen Teil der Region Schwarzes Meer angezeigt mit bis zu  $2,8^{\circ}\text{C}$  (IPSL-CM5A-MR). Dieser Wert ist jedoch deutlich extremer als im übrigen Teil der MMR. Hier werden größtenteils Werte zwischen  $0,1^{\circ}\text{C}$  und  $1,5^{\circ}\text{C}$  angezeigt. Einzig BCCR-BCM2.0 zeigt in zwei Regionen auch leichte Temperaturabnahmen mit maximal  $-0,3^{\circ}\text{C}$  an der Südküste Spaniens. Anders als beim Niederschlagstrend, dessen Ursache nicht eindeutig dem anthropogenen Klimawandel zugeschrieben werden kann, steht die positive Temperaturentwicklung zu hoher Wahrscheinlichkeit in Zusammenhang mit der globalen Klimaerwärmung (DIFFENBAUGH ET AL., 2007; DIFFENBAUGH & GIORGI, 2012). Beim Niederschlag wirkt sich hingegen beispielsweise auch die in der zweiten Hälfte des 20. Jahrhunderts verstärkte Nordatlantische Oszillation (NAO) auf den Trend aus (JACOBET ET AL., 2001; SOUSA ET AL., 2011; KELLEY ET AL., 2012a,b). Für die Evaluation der Simulationen spielen die Ursachen des beobachteten Klimas bzw. der Klimaänderung jedoch nur eine untergeordnete Rolle.

Im folgenden Kapitel werden zunächst die statistischen Grundlagen erklärt, die zur Durchführung dieser Untersuchung notwendig sind. Im Anschluss daran wird die Entwicklung und Funktion der Evaluationsmetriken veranschaulicht bevor die Auswertung der Modellergebnisse ab Kapitel 5 vorgenommen wird.

## 3 Statistische Methoden

In diesem Kapitel werden die in dieser Arbeit eingesetzten statistischen Methoden erläutert. Die deskriptiven Methoden wie Berechnung der Mittelwerte, Regression und Varianzen dienen der Auswertung von Niederschlag und Temperatur der Modell- und Evaluationsdaten. Mit Hilfe der Spearman- und Pearson-Korrelationen werden Zusammenhänge und Unterschiede zwischen den Ergebnissen und Gewichtungsmetriken quantifiziert. Allgemein dienen die Methoden der deskriptiven Statistik dem besseren Verständnis und Veranschaulichung von Ergebnissen. Dabei ist es unerheblich, ob es sich um die Ergebnisse einer individuellen Modellsimulation oder die weitere Auswertung aller Simulationen zusammen handelt. Neben wichtigen Grundmethoden wie der Berechnung von Mittelwerten, werden auch weniger triviale Kenngrößen wie bspw. die Regressions- und Korrelationsanalyse dazu gezählt. Die Hauptreferenzen in dieser Arbeit sind WILKS (2006) und VON STORCH & ZWIERS (1999). Ergänzend wurden auch die Werke von BAHRENBERG ET AL. (2003), BAHRENBERG ET AL. (2010), SCHÖNWIESE (2013), PRUSCHA (2006), SACHS (1997) und CLAUSS ET AL. (2004) verwendet.

### 3.1 Arithmetisches und gewichtetes Mittel

Die Auswertung statistischer Zusammenhänge erfordert meist die Bestimmung von Anomalien (z.B. die Extremwerte) in Bezug zum Verhalten der Gesamtheit aller Werte. Letztere wird häufig in Form eines Mittelwertes bestimmt. Dabei stehen verschiedene Formen der Berechnung zur Auswahl. In dieser Arbeit kommen das arithmetische Mittel, der Median sowie gewichtete Mittelwerte zum Einsatz. Das arithmetische Mittel  $\bar{x}$  ist der am häufigsten verwendete Mittelwert. Es wird bestimmt, in dem man alle verfügbaren Variablenwerte  $x_i$  addiert und die Summe anschließend durch die Anzahl  $n$  der Werte dividiert.

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} \quad (3.1)$$

Beim arithmetischen Mittel erhalten alle Variablenwerte dasselbe Gewicht. In der Auswertung der Evaluationsergebnisse bezieht sich daher die gleichgewichtete Be-

trachtung auf das arithmetische Mittel über alle verfügbaren Modelle. Auch wenn die Gleichgewichtung aller  $x_i$  häufig gewünscht ist, bietet es auch eine erhöhte Störanfälligkeit für Ausreißer. Einzelne Extremwerte können so starken Einfluss auf den Mittelwert ausüben. Wenn es darum geht, die Effekte solcher Anomalien zu reduzieren und eine Aussage über die Lage des Mittelwertes der Mehrheit der Werte zu treffen, ist es daher sinnvoll den Median zu verwenden.

Beim Median werden alle Variablenwerte  $x_i$  der Größe nach sortiert und jedem Wert ein Rangplatz von 1 bis  $n$  vergeben. Der Median beschreibt nun den Wert, der bei einer ungeraden Anzahl  $n$ , auf der mittleren Rangposition, bzw. das arithmetische Mittel der beiden mittleren Rangpositionen bei einem geraden  $n$ .

Bei der Beschreibung des Medians schließt sich die Erklärung von Quantilen, bzw. Perzentilen an: Das  $q$  %-Perzentil beschreibt den Wert der nach Größe sortierten Variablenwerte unterhalb dem  $q$  % der Werte und oberhalb dessen  $100 - q$  % der Werte liegen. Unter der Voraussetzung, dass die Variablenwerte  $x_i$  der Größe nach sortiert sind  $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ , mit  $x_{(1)}$  als dem kleinsten und  $x_{(n)}$  als dem größten  $x_i$ , werden die Perzentile nach den folgenden Formeln berechnet. Wenn bei  $n \cdot q/100 = k$ ,  $k$  einer ganzen Zahl entspricht ergibt sich für das  $q$  %-Perzentil:

$$M_q = \frac{(100 - q)x_k + qx_{k+1}}{100} \quad (3.2)$$

Andernfalls, wenn  $n \cdot q/100 = k$  keine ganze Zahl ist, gilt nach SACHS (1997, S. 79):

$$M_q = x_{\lceil n \cdot q \rceil} \quad (3.3)$$

Durch  $\lceil \cdot \rceil$  wird die kleinste ganze Zahl, welche nicht kleiner als das Ergebnis aus  $n \cdot q$  ist, gewählt. Dem Median entspricht das 50 %-Perzentil, da genau 50 % der Werte oberhalb und 50 % der Werte unterhalb diesem liegen. Diese Form der Einteilung von Daten wird häufig zur Veranschaulichung von Datenstreuung mit Hilfe eines Boxplots verwendet.

Die dritte Form des Mittelwertes, der in dieser Arbeit berechnet wird, ist ein gewichtetes Mittel. Dieses ist entscheidend, wenn es darum geht, bestimmten Variablenwerten eine größere Bedeutung zuzuteilen als anderen. Hierbei ist jedoch weiteres Wissen über die Datenlage erforderlich. In dieser Arbeit werden die Ergebnisse verschiedener Metriken herangezogen um abzuschätzen, welchen Modellen ein höheres Gewicht zugeteilt werden sollte. Allgemein ist die Berechnung des arithmetischen Mittels  $\bar{x}$  eine Sonderform des gewichteten arithmetischen Mittels  $\bar{x}_g$  mit  $\omega_i = \text{const.} = 1$ .

$$\bar{x}_g = \frac{\sum_{i=1}^n \omega_i x_i}{\sum_{i=1}^n \omega_i} \quad (3.4)$$

Der entscheidende Unterschied zum arithmetischen Mittel ist, dass die Gewichte  $\omega_i$  jeweils nicht gleich groß sind. Stattdessen bekommt jede Variable  $x_i$  ein individuell durch verschiedene Metriken bestimmtes Gewicht  $\omega_i$ , das jeden positiven Wert annehmen kann.

## 3.2 Varianz und Standardabweichung

Neben den Mittelwerten und Extremwerten ist bei der statistischen Auswertung von Klimamodellensembles die Datenstreuung eine wichtige Informationsquelle. Diese wird häufig über die Varianz und Standardabweichung ausgedrückt. Darüber hinaus dienen diese Größen auch als Grundlage für eine Reihe weiterer statistischer Indikatoren. Die einfachste Information zur Datenstreuung ist die Variationsbreite. Diese drückt den Bereich aus der zwischen Minimum  $x_{min}$  und Maximum  $x_{max}$  der Werte liegt. Der Umfang dieses Intervalls ist genau wie der Maximum- und Minimum-Wert von Bedeutung bei der Beschreibung metrischer Daten (z.B. Zeitreihen). Da hierfür jedoch nur zwei Werte Beachtung finden, bei denen es sich meist nur um Ausreißer handelt, ist der Informationsgewinn über die allgemeine Streuung der Daten begrenzt. Eine bevorzugtes Maß zum Ausdruck der Variation ist daher die Standardabweichung  $s$ .

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \quad (3.5)$$

Die Stichprobenumfang ist hier  $n$ . Das Quadrat der Standardabweichung  $s^2$  wird als Varianz bezeichnet. Da die Standardabweichung  $s$  empfindlich gegenüber Ausreißern ist wird sie weder als robust noch als widerstandsfähig angesehen (WILKS, 2006, S.26). Stattdessen verbirgt sich hinter ihr die Wurzel über die mittlere Differenz zum Quadrat zwischen den Zufallsvariablen  $x_i$  und dem arithmetischen Mittel  $\bar{x}$ . Die Standardabweichung gibt ebenso wie die Varianz Aufschluss über die Streuung einer Datenreihe, ist aber nicht mit dieser gleich zu setzen. In dieser Arbeit werden die Varianz und Standardabweichung zur Bestimmung und Interpretation der Wahrscheinlichkeitsdichtefunktion verwendet.

### 3.3 Wahrscheinlichkeitsdichtefunktion

Die Wahrscheinlichkeitsdichtefunktionen (WDF) können als kontinuierliches und theoretisches Äquivalent eines Histogramms und des nichtparametrischen Kerndichteschätzers verstanden werden (WILKS, 2006, S. 85).

In dieser Arbeit werden Wahrscheinlichkeitsdichte- und Kerndichteschätzfunktion verwendet, um Auskunft über Verteilungen der Zukunftsprojektionen zu geben. Dabei werden sowohl metrik- sowie gleichgewichtete Unsicherheitsbereiche bestimmt und analysiert.

#### 3.3.1 Normalverteilungsansatz

Diskrete Zufallsvariablen lassen sich mit Hilfe einer Verteilungsfunktion charakterisieren. Sei  $Z$  eine beliebige diskrete Zufallsvariable, die über den Ergebnisraum  $R = z_1, z_2, \dots, z_n$  verfügt. Dann gilt die Funktion, die dem jeweiligen Ereignis  $Z = z_i$  eine Wahrscheinlichkeit  $P$  zuordnet und ansonsten 0 annimmt als Wahrscheinlichkeitsfunktion. Das Integral über das Intervall der untersuchten Ereignisse beschreibt die Wahrscheinlichkeit. In dieser Studie werden verschiedene WDF über die Klimaänderung der unterschiedlichen Zukunftsprojektionen berechnet. Damit wird untersucht, wie groß der Unsicherheitsbereich der verschiedenen Modelle für das 21. Jahrhundert ist bzw. ob die Unsicherheit reduziert werden kann, indem unterschiedliche Gewichte den Modellen zugeteilt werden. Die zu Grunde liegenden Klimaänderungen vom Ende des 20. zum Ende des 21. Jahrhunderts werden als normalverteilt angenommen. Dies wurde durch den Kolmogorov-Smirnov-Test zuvor überprüft (siehe Abschnitt 3.4). Zur Berechnung der WDF über die Klimaänderung der gleichgewichteten Modelle wurde das arithmetische Mittel und die Standardabweichung verwendet.

$$f(z) = \frac{1}{s \cdot \sqrt{2\pi}} \exp\left[-0,5\left(\frac{z - \bar{z}}{s}\right)^2\right] \quad (3.6)$$

$\bar{z}$  ist arithmetischer Mittelwert bzw. Erwartungswert der Datenpunkte und  $s$  die Standardabweichung. Das Ergebnis ist eine normalverteilte Glockenkurve, die abhängig von Standardabweichung und Erwartungswert ist. Dabei gibt die Abszisse die Klimaänderung (hier in mm oder °C) und die Ordinate die Wahrscheinlichkeitsdichte an. Letztere ist dabei nicht mit der Wahrscheinlichkeit zu verwechseln. Zur Bestimmung der Wahrscheinlichkeit muss über die Fläche unterhalb der Kurve integriert werden. Die Kurve reicht grundsätzlich von  $-\infty$  bis  $\infty$ . Bei der Interpretation werden häufig die Über- bzw. Unterschreitungswahrscheinlichkeit eines bestimmten Ereignisses wie beispielsweise dem Erwartungswert betrachtet. Bezogen auf die Fragestellung dieser Arbeit ist eine große Standardabweichung mit

einer hohen Unsicherheit über die zu erwartenden Klimaänderungen zu verstehen. Hintergrund ist, dass durch die hohe Standardabweichung eine große Streuung der Modelle ausgedrückt wird, die in diesem Fall nur wenig Übereinstimmung zeigen. Umgekehrt steht eine schmale WDF für eine hohe Übereinstimmung der Modelle und damit geringe Unsicherheit. Die Lage des Erwartungswerts gibt darüber Auskunft, in welchem Bereich die mittlere Klimaänderung über Modelle projiziert wird. Der Erwartungswert stimmt durch die normalverteilte WDF immer mit dem Wert der höchsten Wahrscheinlichkeitsdichte überein.

Die Berechnung der ungleich- oder leistungs-basiert gewichteten WDF erfolgt analog zu Formel 5.7 mit der Änderung, dass nun das gewichtete  $\bar{z}_g$  anstelle dem arithmetischen Mittel, sowie die gewichtete Standardabweichung  $s_g$  eingesetzt werden.

$$f_g(z) = \frac{1}{s_g \cdot \sqrt{2\pi}} \exp\left[-0,5\left(\frac{z - \bar{z}_g}{s_g}\right)^2\right] \quad (3.7)$$

Diese Änderung hat in der Folge sowohl Einfluss auf die Lage der WDF als auch auf das Unsicherheitsverhalten. Abseits von der optischen Interpretation wird im Folgenden mit Hilfe eines Tests geprüft, ob sich die beiden Arten von WDF signifikant unterscheiden.

### 3.3.2 Kerndichteschätzer

Als eine weitere Form der Darstellung und Analyse der Modelldaten wird der Gauß-Kerndichteschätzer verwendet. Hierbei handelt es sich im Prinzip um eine erweiterte Darstellungsform eines Histogramms. Es wird in diesem Zusammenhang auch von Kerndichte-Glättung gesprochen (WILKS, 2006, S. 35f.). Die Anwendung der Kerndichte-Glättung auf die Häufigkeitsverteilung eines Datensatzes ergibt die Kerndichteschätzfunktion (KDF), bei dem es sich um eine nicht-parametrische Form der WDF handelt. Dabei wird die Anpassung über die verschiedenen Werte des Datensatzes basierend auf einem Glättungs-Kern vorgenommen. Jeder Wert  $z_i$  wird so zu einem Zentrum für eine Kern-Funktion. Das Intervall bzw. die Bandbreite  $h$ , über das der Kern berechnet wird, ist dabei variabel. In dieser Arbeit wurden situationsabhängig verschiedene Bandbreiten getestet, um eine glatte Kurve zu erreichen, die noch genügend Flexibilität besitzt, um multimodale Funktionen zu darzustellen. Die insgesamt besten Ergebnisse zeigten sich basierend auf dem Ansatz von SILVERMAN (1986) beispielsweise nach LAHR (2014) zur Berechnung der idealen Bandbreite. Durch manuelles Nachjustieren ist situationsabhängig eine stärkere Betonung multimodaler Funktionen möglich. Die finale KDF wird durch die Addition der Höhen aller Kern-Funktionen berechnet. Der Wert an der jeweiligen Stelle  $z_i$  trägt so zu der geglätteten KDF

bei.

$$\hat{f}(z_0)_g = \frac{h}{\sqrt{2\pi} \sum_{i=1}^n \omega_i} \sum_{i=1}^n \exp[-0,5(\frac{z_0 - z_i}{h})^2] \cdot \omega_i \quad (3.8)$$

Hier entspricht  $n$  dem Umfang des Datensatzes und  $z_0$  dem jeweiligen Zentrum der Kern-Funktionen. Die KDF werden analog zu den WDF im Rahmen der Auswertung von Zukunftsprojektionen angewandt. Da es hier um den Vergleich von ungleich- zu gleichgewichteten Datensätzen geht wurde der Faktor  $\omega_i$  eingefügt, der dem jeweiligen Gewicht entspricht. Im gleichgewichteten Fall gilt stets  $\omega_i = 1/n$ . Dadurch hat die Fläche unter der gesamten KDF den Wert Eins, wodurch die Funktion eine Vergleichsmöglichkeit zur WDF bietet.

### 3.4 Der Kolmogorov-Smirnov-Test

Die Analyse der zuvor erläuterten WDF über die Klimaänderung wird in dieser Studie zusammen mit den KDF zur Abschätzung des Unsicherheitsbereichs verwendet. Für die Klimaänderungen verschiedener Modelle, die zur Erstellung der WDF genutzt werden wird eine Gaußverteilung angenommen. Um dies statistisch zu belegen wird für jede Situation der Kolmogorov-Smirnoff-Test (KS-Test) angewandt. Dieser dient dem Vergleich der Verteilung zweier Zufallsvariablen. Hier wird die empirische kumulative Wahrscheinlichkeit  $F_n(x)$  der Klimaänderungen unterschiedlicher Modelle eines MME, abgeschätzt über  $F_n(x_{(i)}) = i/n$ , mit einer theoretischen kumulativen Normalverteilungsfunktion  $F(x)$  verglichen, um zu prüfen, ob beide derselben Verteilung angehören. Die Teststatistik  $D_s$  berechnet sich über:

$$D_s = \max|F_n(x) - F(x)| \quad (3.9)$$

$D_s$  sucht nach der betragsmäßig größten Differenz zwischen der empirischen und theoretischen kumulativen Verteilungsfunktion. Wenn die Differenz hinreichend groß wird, kann die Nullhypothese abgelehnt werden. Zur Überprüfung wurde die Irrtumswahrscheinlichkeit  $\alpha = 0,05$  (bzw. das Signifikanzniveau  $S_i = 0,95$ ) gewählt. Die kritische Wert  $KS_{krit}$  zur Prüfung der Signifikanz kann für  $n > 35$  berechnet werden (SACHS, 1997, S.462).

$$KS_{krit} = \frac{1,358}{\sqrt{n}}, \text{ für } n > 35 \quad (3.10)$$

Für kleinere Anzahlen, wie sie beispielsweise bei CMIP3 (24) oder CORDEX (18), wird der Wert direkt aus einem Tafelwerk zum KS-Test abgelesen (z.B. MILLER

(1956)). Die abschließende Überprüfung erfolgt durch den Vergleich  $KS_{krit}$  und  $D_s$ . Wenn gilt,

$$D \geq KS_{krit}$$

wird die Nullhypothese auf dem Signifikanzniveau von 95 % verworfen. Für die Betrachtung und Analyse der WDF wurden nur Verteilungen berücksichtigt, die sich nach dem KS-Test als normalverteilt herausgestellt haben.

### 3.5 Kovarianz und Korrelation

Die Kovarianz und Korrelation sind wichtige Bestandteile der deskriptiven Statistik, um die Stärke von Zusammenhängen von Zufallsvariablen zu beschreiben. Um die Art des Zusammenhangs zu bestimmen wird dagegen die Regressionsanalyse verwendet (siehe Abschnitt 3.6). Die Unterscheidung zwischen einer abhängigen und einer unabhängigen Variablen ist nur für die Regressionsanalyse notwendig. In dieser Arbeit werden zwei Kategorien von Korrelationskoeffizienten verwendet. Pearson- und Spearman-Korrelationen dienen im Verlauf dieser Arbeit als Indikatoren über die Stärke des Zusammenhangs bzw. der Ähnlichkeit beispielsweise zwischen den Modellranglisten basierend auf unterschiedlichen Metriken. Der andere große Block von Korrelationskoeffizienten wird als Metriken eingesetzt und daher in Kapitel 4 erläutert. Diese werden im Folgenden zur Evaluation von Klimamodellen genutzt. Für weitere Information sei an dieser Stelle auf Kapitel 4 verwiesen. Zunächst wird die Kovarianz anhand der folgenden Formel erläutert.

$$Cov(x, y) = \frac{1}{n-1} \sum_{i=1}^n [(x_i - \bar{x})(y_i - \bar{y})] \quad (3.11)$$

Die Kovarianz wird über die Summe der Produkte der jeweiligen Anomalien, Abweichungen vom Stichprobenmittel z.B.  $x_i - \bar{x}$ , beider Zufallsgrößen  $x_i, y_i$  berechnet und anschließend durch den, um 1 reduzierten Stichprobenumfang  $n$ , dividiert. Da die Kovarianz nicht normiert ist, kann sie theoretisch beliebige Werte annehmen, was die Vergleichbarkeit jedoch erschwert. In dieser Arbeit wird die Kovarianz daher nicht als Indikator des Zusammenhangs zweier Variablen direkt verwendet. Stattdessen kommt sie bei der Berechnung der natürlichen Variabilität zur Erstellung der Fingerprintmetrik (siehe 4.4) in Form einer Kovarianzmatrix zum Einsatz. Drüber hinaus ist sie für die Bestimmung der Hauptkomponenten zur Berechnung der Fingerprintmetrik (siehe 4.4) erforderlich.

Basierend auf der Kovarianz ist darüber hinaus eine einfache Erläuterung der Pearson-Korrelation möglich. Die Pearson-Korrelation wird für metrische Daten eingesetzt, während die Spearman-Korrelation, die auch als Rang-Korrelation

bezeichnet wird, auch für ordinal skalierte Zufallsvariablen genutzt wird. Die Pearson-Korrelation wird häufig als ein Standardverfahren zu Bestimmung der Stärke des Zusammenhangs zwischen zwei Zufallsvariablen verwendet. Für eine explizite Herleitung der Berechnung sei auf WILKS (2006, S.50ff). oder BAHRENBURG ET AL. (2010, S.197ff.) verwiesen. Die Herleitung erfolgt in den meisten Fällen basierend auf der Regressionsanalyse (3.6). Die folgende Formel beschreibt die Berechnung des Korrelationskoeffizienten  $r_{xy}$ . Die Pearson-Korrelation kann auf dem Verhältnis zweier bereits zuvor erläuterten statistischen Größen betrachtet werden: Der normierten Kovarianz  $Cov(x, y)$  und dem Produkt der Standardabweichungen  $s_x$  und  $s_y$ .

$$r_{xy} = \frac{Cov(x, y)}{s_x s_y} \quad (3.12)$$

Durch Einsetzen der zuvor erklärten Formeln folgt:

$$r_{xy} = \frac{\frac{1}{n-1} \sum_{i=1}^n [(x_i - \bar{x})(y_i - \bar{y})]}{\sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}} \quad (3.13)$$

Die Pearson-Korrelation kann Werte im Intervall  $[-1; 1]$  annehmen. Dabei entspricht ein Wert von 1 einer idealen positiven bzw. ein Wert von -1 einer negativen linearen Beziehung zwischen den beiden Zufallsvariablen. Um zu prüfen, ob auch Korrelationen dazwischen, die nicht  $\pm 1$  annehmen, einen relevanten Zusammenhang ausdrücken, bedient man sich Signifikanztests. In dieser Arbeit wird hierzu der t-Test verwendet (vgl. Abschnitt 3.7). Korrelationen, die um 0 liegen, geben an, dass keinerlei Beziehung zwischen zwei Variablen besteht.

Generell drücken Korrelationskoeffizienten die Güte des Zusammenhang zweier Zufallsvariablen aus. Die Besonderheit der Pearson-Korrelation ist, dass die Linearität der metrischen Daten vorausgesetzt wird. Dies ist bezogen auf den Informationsgewinn ein Vorteil, allerdings können Ausreißer hier sehr stark ins Gewicht fallen. Soll durch eine Korrelation überprüft werden, ob die Reihenfolge zweier Zufallsvariablen einen Zusammenhang aufweist, ist dies jedoch von Nachteil. Hierzu wird die Spearman- oder Rang-Korrelation  $r_s$  verwendet. Diese gilt als sehr robust und ist unempfindlich gegenüber Ausreißern. Die Berechnung ist dabei identisch mit der Pearson-Korrelation, allerdings basierend auf den Rängen der Zufallsvariablen. Zunächst müssen dazu die Werte  $x_i, y_i$  der Größe nach sortiert und mit Rangpositionen versehen werden. Somit werden aus den metrischen ordinal skalierte Variablen, denen die Werte  $x_{(1)} = 1, x_{(2)} = 2, \dots, x_{(n)} = n$  bzw.  $y_{(1)} = 1, y_{(2)} = 2, \dots, y_{(n)} = n$  zugeteilt werden. Anschließend werden die Da-

tensätze wieder in ihre Ausgangsreihenfolge gebracht und über die beiden Rangfolgen  $x_{(i),r}$  und  $y_{(i),r}$  die Formel der Pearson-Korrelation angewandt. Im Fall, dass zwei Zufallsvariablen dieselbe Rangposition einnehmen, wird die mittlere Rangposition für alle betroffenen Werte berechnet und verwendet. Dies ist in dieser Arbeit jedoch nicht erforderlich, da alle Rangplätze aufgrund der Werte und Nachkommastellen immer eindeutig vergeben werden können. Die Pearson- und Spearman-Korrelation sind zwei sehr wichtige statistische Indikatoren für Zusammenhänge in dieser Arbeit. Zusätzlich bieten die verschiedenen Ansätze der 4-Felder-Korrelation noch weitere Möglichkeiten zur Berechnung der Ähnlichkeit bzw. der Zusammenhänge zweier Zufallsvariablen. Diese werden in Kapitel 4 behandelt, da sie als eine der experimentellen Evaluationsmetriken von Klimamodellen behandelt werden.

## 3.6 Regressionsanalyse

Neben der Stärke, ausgedrückt über die Korrelation, ist auch die Art des Zusammenhangs von Zufallsvariablen von großer Bedeutung. Eine der wichtigsten statistischen Größen bei der Auswertung von Klimamodellen ist der Regressionskoeffizient, der häufig auch als linearer Trend bezeichnet wird. Dies kann im einfachsten Fall durch eine Gerade dargestellt werden. Im Falle dieser Arbeit ist immer die Zeit als unabhängige Größe festgelegt, während sich die Variablenwerte in Abhängigkeit von der Zeit verändern (z.B. BAHRENBERG ET AL., 2010, S. 184ff.).

Für die Berechnung der Trendgerade wird die allgemeine Formel

$$\hat{y}_i = a + b \cdot x_i \quad (3.14)$$

aufgestellt. Hier beschreibt  $x_i$  einen spezifischen Zeitpunkt, der über die beiden Konstanten  $a$  und  $b$  den Funktionswert  $\hat{y}_i$  ergibt. Um optimal positioniert zu sein, muss die Summe der vertikalen Entfernungen der Punkte von der Gerade minimal sein. Hierzu wird die vertikale Entfernung, die parallel zur Y-Achse verläuft, verwendet. Man bezeichnet die Abweichungen von der Trendgeraden auch als Residuen, die die Differenz zwischen dem tatsächlichen y-Variablenwert und dem zugehörigen Wert auf der Geraden beschreibt. Zur Aufstellung der Geraden müssen die beiden Unbekannten, die Regressionskonstante  $a$  und der Regressionskoeffizient  $b$  bestimmt werden (z.B. WILKS, 2006, S. 182).

$$b = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2} \quad (3.15)$$

$$a = \bar{y} - b\bar{x} \quad (3.16)$$

Durch Umformung ist es möglich  $b$  auch als den Quotienten aus Kovarianz und Varianz darzustellen. Damit beschreibt der Regressionskoeffizient die Kovarianz von  $X$  und  $Y$ , normiert durch die Varianz von  $X$  (z.B. BAHRENBERG ET AL., 2010, S. 190). Die Interpretation von  $b$  erfolgt abhängig vom Vorzeichen. Ein positiver Regressionskoeffizient deutet auf eine Zunahme und ein negatives Vorzeichen auf eine Abnahme innerhalb des Zeitraumes hin. Alle natürlichen metrischen Daten besitzen einen Trend. Möchte man die Verteilung der Werte ohne den langfristigen Trend untersuchen, ist es möglich diesen Trend herauszurechnen, dabei spricht man von Enttrends. Hierzu wird ebenfalls zunächst die Trendgerade mit Hilfe von Regressionskonstante und Regressionskoeffizienten berechnet. Anschließend werden die  $y$  Werte der Trendgerade vom Originalwert in Abhängigkeit vom jeweiligen  $x$ -Wert abgezogen. Durch dieses Vorgehen wird der langfristige Trend entfernt. Dieses Vorgehen wird bei der Berechnung der Kovarianzmatrix der natürlichen Variabilität des Fingerprintansatzes verwendet (Kapitel 4.4). Für die Berechnung des Temperaturtrends der Referenz- und Modelldaten wurde eine nichtlineare Form der Trendberechnung verwendet. Hierzu wurde das arithmetische Mittel der jeweils ersten und letzten 20 Jahre des 50-jährigen Intervalls bestimmt und die Differenz berechnet. Der „50-jährige“ Trend von Niederschlag oder Temperatur in dieser Arbeit bezieht sich daher auf das Ergebnis der zuvor erläuterten nichtlinearen Differenzuntersuchung, während die Trendanalyse von Zeitreihen (in Kapitel 6.1 und Kapitel 7.2.1) auf der linearen Regression beruhen.

### 3.7 Signifikanztests

Als abschließende Indikatoren bei der Untersuchung von zwei Zufallsvariablen sind Signifikanztests zu nennen. Während die Regression die Art und die Korrelation die Stärke eines Zusammenhangs untersuchten, gibt ein Signifikanztest Auskunft, ob überhaupt ein relevanter Zusammenhang zwischen den Zufallsvariablen besteht. Sowohl die Berechnung der Korrelation als auch der Regression werden immer ein Ergebnis haben. Der Signifikanztest muss im Anschluss klären, ob die Aussagekraft dieser Werte von Bedeutung ist. Damit wird die Fragestellung beantwortet, ob zwei Stichproben derselben Grundgesamtheit angehören oder sich signifikant unterscheiden. In dieser Arbeit werden ausschließlich zweiseitige Tests

durchgeführt. Für die Untersuchung der Mittelwerte bzw. der Erwartungswerte werden die Werte als T-verteilt, für die Varianz und die Regression als F-verteilt angenommen.

### 3.7.1 Test des Spearman-Korrelationskoeffizienten

Zur Prüfung der statistischen Signifikanz der Spearman-Korrelation  $r_s$  (siehe Abschnitt 3.5) wird nach SACHS (1997, S. 745f.) die Prüfgröße  $P_r$  über:

$$P_r = \frac{r_s}{2} \left[ \sqrt{(n-1)} + \sqrt{\frac{n-2}{1-r_s^2}} \right] \quad (3.17)$$

berechnet. Diese Berechnung (3.17) dient als Approximation für Stichprobenumfänge  $n > 30$ . Für kleinere Umfänge kann die Signifikanz beispielsweise aus Tafelwerken von SACHS (1997) abgelesen werden. Die Prüfgröße  $P_r$  wird anschließend mit dem Vergleichswert  $V_p$

$$V_p = [z_{\alpha/2} + t_{n-2, \alpha/2}] / 2 \quad (3.18)$$

verglichen. Da es sich um einen zweiseitigen Test handelt, wird  $\alpha/2$  für die Irrtumswahrscheinlichkeit  $\alpha = 0,05$  verwendet. Dazu wird aus der Standardnormalverteilung  $z_{0,975} = 1,960$  entnommen. Analog wird  $t$  abhängig vom jeweiligen  $n$  aus der t-Verteilung abgelesen. Gilt  $|P_r| \geq V_p$  so ist der Zusammenhang unter den genannten Voraussetzungen statistisch signifikant.

### 3.7.2 Test des Erwartungswertes

Beim Signifikantest des Erwartungswertes handelt es sich um einen Vergleichstest zweier Mittelwerte (SCHÖNWIESE, 2013, S. 125f.). Es werden die Erwartungswerte der WDF einmal gleich- und einmal metrikgewichtet für die Zukunftsprojektionen berechnet. Dabei wird untersucht, ob es einen signifikanten Unterschied zwischen den beiden Erwartungswerten  $\bar{x}_0, \bar{x}_g$  gibt. In dieser Arbeit wird über diesen Test geprüft, ob der Effekt der Gewichtung signifikant bzgl. der WDF-Verschiebung ist. Dazu werden zunächst die Nullhypothese  $H_0$  sowie die Alternativhypothese  $H_1$  aufgestellt.

$$H_0 : \bar{x}_0 = \bar{x}_g$$

$$H_1 : \bar{x}_0 \neq \bar{x}_g$$

Die hier getroffene Annahme der Nullhypothese besagt, dass der Mittelwert der

ersten Stichprobe derselben Grundgesamtheit angehört wie der der zweiten Stichprobe. Dies wird bei einer Irrtumswahrscheinlichkeit von  $\alpha = 0,05$  getestet. Die einfachste Herangehensweise ist, die Differenz zwischen den jeweiligen Mittel beider Stichproben zu analysieren. Im Fall der Gleichgewichtung handelt es sich um das arithmetische Mittel während es bei der Metrikgewichtung das gewichtete Mittel ist.

$$\hat{t} = \frac{|\bar{x}_0 - \bar{x}_g|}{\sqrt{\frac{s_0^2 + s_g^2}{n}}} \quad (3.19)$$

Hier beschreiben  $s_0^2$  die Varianz der gleichgewichteten Modelle und  $s_g^2$  der metrikgewichteten Modelle während  $n$  für die Modellanzahl steht. Die Freiheitsgrade FG werden über

$$FG = n - 1 + \frac{2n - 2}{\frac{s_0^2}{s_g^2} + \frac{s_g^2}{s_0^2}} \quad (3.20)$$

berechnet. Anschließend kann der kritische Wert  $t_{krit}$  aus der t-Verteilung überprüft werden. Gilt  $\hat{t} > t_{krit}$  wird die Nullhypothese  $H_0$  bei einer Irrtumswahrscheinlichkeit von  $\alpha = 0,05$  verworfen. In dieser Arbeit würde das bedeuten, dass der Einfluss der Metrikgewichtung eine signifikante Änderung des Erwartungswerts der Klimaänderung für das 21. Jahrhundert im Vergleich zu Gleichgewichtung hat.

### 3.7.3 Test zweier Varianzen

Neben dem zuvor beschriebenen Test zum Vergleich der Erwartungswerte der metrik- und gleichgewichteten Klimaänderungs-WDF ist auch die Ähnlichkeit der Varianzen bzw. der Standardabweichungen von großem Interesse. Diese beschreiben den Unsicherheitsanteil der WDF. Neben der Frage nach der Stärke der Verschiebung der Klimaänderung ist auch die Untersuchung der Veränderung des Unsicherheitsbereichs von ebenso großem Interesse. Dazu wird hier mit Hilfe eines Signifikanztests geprüft, ob sich die Varianzen gleich- ( $s_0^2$ ) und metrikgewichtet ( $s_g^2$ ) signifikant ( $\alpha = 0,05$ ) unterscheiden. Für die Anwendung des Tests sollten die beiden getesteten Stichproben unabhängig voneinander und normalverteilt sein (VON STORCH & ZWIERS, 1999, S. 119). Die Bedingung der Normalverteilung wird mit Hilfe des KS-Tests (vgl. Abschnitt 3.4) geprüft, sodass nur WDF untersucht werden, für die normalverteilte Klimaänderungen der Modelle nachgewiesen werden.

$$H_0 : s_0^2 = s_g^2$$

$$H_1 : s_0^2 \neq s_g^2$$

Zur Prüfung der Signifikanz zweier Varianzen wird der F-Test verwendet. Hierzu wird die Prüfgröße  $\hat{F}$  über

$$\hat{F} = \frac{s_0^2}{s_g^2} \quad (3.21)$$

berechnet. Hier gilt, dass die größere der beiden Varianzen im Zähler stehen soll (SCHÖNWIESE, 2013, S. 168). Für die F-Verteilung gilt analog zum Test des Erwartungswertes  $n - 1$  Freiheitsgrade wobei gilt  $n = n_0 = n_g$ . Gilt  $\hat{F} > F_{n;n;0,025}$  wird die Nullhypothese  $H_0$  verworfen und die Alternativhypothese  $H_1$  angenommen.

### 3.7.4 Test zweier Regressionskoeffizienten

Der Regressionkoeffizient wird zur Berechnung des linearen Trends verwendet. Dies ist in dieser Arbeit insbesondere hinsichtlich der Projektionen für das 21. Jahrhundert von großer Relevanz. Die Evaluation der Modelle basierend auf den Ergebnissen der Metriken führt zu einer unterschiedlichen Verteilung innerhalb des MME. Berechnet man den linearen Trend bis zum Ende des 21. Jahrhunderts können so Abweichungen zwischen dem Trend bzw. den Regressionskoeffizienten der gleich- ( $r_0$ ) und metrikgewichteten ( $r_g$ ) MME entstehen. Um allgemein gültige Aussagen zu treffen wird getestet, ob es durch die Gewichtung zu einer signifikanten Abweichung der Regressionskoeffizienten kommt. In dieser Arbeit kommt der Ansatz von SACHS (1997, S. 755) zum Einsatz. Zunächst werden die Null- und Alternativhypothese aufgestellt.

$$H_0 : r_0 = r_g$$

$$H_1 : r_0 \neq r_g$$

Hier gilt, dass  $r_0, r_g$  den Regressionkoeffizienten entsprechen. Die Prüfgröße  $\hat{t}$  berechnet sich im Folgenden über:

$$\hat{t} = \frac{|r_0 - r_g|}{\sqrt{\frac{s_0^2(n-2) + s_g^2(n-2)}{(2n-4)} \left( \frac{1}{Q_{X_0}} + \frac{1}{Q_{X_1}} \right)}} \quad (3.22)$$

Dabei sind  $s_0^2, s_g^2$  die Varianzen der Verteilungen denen die Regressionen zu Grunde liegen.  $Q_{X_0}$  entspricht der Summe der Abweichungsquadrate  $Q_{X_0} = \sum_{i=1}^n (x_i - \bar{x})^2$ . Da diese Werte für beide Zeitreihen (metrik- und gleichgewichtet) identisch

sind, gilt  $Q_{X_0} = Q_{X_1}$ . Der kritische Wert  $t_{krit}$  wird anschließend bei einer Irrtumswahrscheinlichkeit von  $\alpha = 0,05$  aus der t-Verteilung mit

$$FG = 2 \cdot n - 4$$

Freiheitsgraden (FG) abgelesen. Ist die Prüfgröße  $\hat{t}$  größer als  $t_{krit}$  kann  $H_0$  verworfen werden und die beiden Regressionskoeffizienten unterscheiden sich signifikant voneinander.

### 3.7.5 *u*-Anpassungstest

Als abschließender Signifikanztest wird ein *u*-Test in diesem Abschnitt eingeführt. Für den Fall, dass nur zwei Merkmalsausprägungen  $a, b$  zur Wahl stehen, wird mit dem *u*-Test die Wahrscheinlichkeit einer dieser Ausprägungen bestimmt. Es handelt sich hierbei um eine Untergruppe der Anpassungstests (CLAUSS ET AL., 2004, S.186). Dieser Test ist auf einen Stichprobenumfang  $n \geq 60$  ausgelegt. Zusätzlich darf die Wahrscheinlichkeit  $p_{i,0}$  nicht zu stark von 0,5 abweichen und das Minimum der Erwartungshäufigkeit  $n \cdot p_{min,0} \geq 5$  sein. Es wird getestet, ob die eine der beiden Ausprägungen häufiger vorkommt und sich das Ergebnis signifikant von 50 % Wahrscheinlichkeit und damit zufälligen Ergebnissen unterscheidet (CLAUSS ET AL., 2004, S. 186). In dieser Studie kommt dieser Test zum Einsatz, um die Gewichte basierend auf den angewandten Metriken zu validieren (Kapitel 8). In dieser Arbeit gilt  $p_{i,0} = 0,5$  daher ist  $p_{min,0}$  immer gleich  $p_{i,0}$ . Bei  $p_{min,0} = 0,5$  und  $n > 60$  ist die Bedingung der Erwartungshäufigkeit immer erfüllt. Anschließend wird zur Durchführung die relative Häufigkeit  $h_{a,b}$  der untersuchten Zufallsereignisse über

$$h_i = \frac{i}{n}$$

bestimmt. Hier entspricht  $i$  der absoluten Häufigkeit, die überprüft werden soll. Die Hypothese, die hierzu aufgestellt wird, ist:

$$H_0 : p_i = p_{i,0} = 0,5$$

$$H_1 : p_i \neq p_{i,0}$$

In dieser Arbeit wird als Signifikanzniveau  $\alpha = 0,05$  gewählt. Die Prüfgröße  $\hat{t}$  wird anschließend über die folgende Formel berechnet:

$$\hat{u} = \frac{h_i - p_{i,0}}{\sqrt{p_{i,0}(1 - p_{i,0})}} \cdot \sqrt{n} \quad (3.23)$$

Da  $p_{1,0}$  in dieser Untersuchung immer 0,5 entspricht, ist die Bestimmung der Signifikanz nur vom Stichprobenumfang und der relativen Häufigkeit  $h_i$  abhängig. Der kritische Bereich wird durch das Quantil  $u_{\alpha,zweis} = z_1 - \alpha/2$  aus der Normalverteilung abgelesen (CLAUSS ET AL., 2004, S. 186). Gilt,

$$|\hat{u}| \geq u_{\alpha,zweis}$$

wird  $H_0$  abgelehnt und die Alternativhypothese angenommen.

## 4 Evaluationsmetriken

Ein Schwerpunkt dieser Studie liegt auf der Erstellung und Anwendungen von Metriken zur Bewertung und Gewichtung von Klimamodellsimulationen. Im Rahmen des COMEPRO-Projekts wurden verschiedene Metriken aus vorangegangenen Studien aufgegriffen und erweitert oder neu für die Aufgabe der Klimamodell-evaluation entwickelt (STEPHENSON, 2000; PAETH ET AL., 2006; HASSELMANN, 1979). Alle Metriken stellen die Frage nach der Übereinstimmung von globalen oder regionalen Klimamodellen mit den Referenzdaten in verschiedenen Untersuchungsgebieten und unterschiedliche Variablen. Die Ergebnisse der Evaluationen auf Grundlage der Metriken bzw. der darauf basierenden Gewichtungen werden in Kapitel 5 bis 9 dargestellt. Allgemein gilt, dass alle einzelnen Simulationen ausgewertet werden und ein Rohgewicht zugeteilt bekommen. Da die Modelle über unterschiedlich viele Simulationen verfügen, wird anschließend das mittlere Rohgewicht für jedes Modell über alle zugehörigen Simulationen berechnet. Daher gilt zu beachten, dass während die Auswertung der Rohgewichte alle Simulationen beschreibt, die Auswertung der Gewichtung selbst auf dem mittleren Gewicht der Modelle basiert, um eine Überrepräsentation einiger weniger Modelle mit hoher Anzahl an Läufen (z.B. GISS-EH, CMIP5) zu vermeiden. Im Folgenden wird die Funktionsweise und Bedeutung der verwendeten Metriken erläutert.

### 4.1 4-Felder-Metriken (4FM)

Die erste Evaluationsmetrik basiert auf einem 4-Felder-Korrelationskoeffizienten. Diese Art der Datenauswertung existiert bereits seit dem späten 19. Jahrhundert. So haben bereits GILBERT (1884) und PIERCE (1884) diese Statistik zur Bewertung des Erfolgs der Tornadovorhersagen von FINLEY (1884) verwendet. Anfang des neuen Jahrtausends hat STEPHENSON (2000) einen Vergleich unterschiedlicher 4-Felder-Metriken angestellt und ihre Vor- und Nachteile wiederum am Beispiel der Tornado Vorhersage veranschaulicht. Heute kommen ähnliche Ansätze in der Regel in der Hydrologie oder numerischen Wettervorhersage zum Einsatz (DONE ET AL., 2004; GHELLI & PRIMO, 2009; WILKINSON, 2017; GILL & BUCHANAN, 2014). Für den Einsatz in klimatologischen Fragestellungen wurde er bisher selten in Erwägung gezogen (WOODCOCK, 1976; PAETH ET AL., 2006).

Beim 4-Felder-Test oder auch der 2 x 2-Tabelle wird ursprünglich der statistische Zusammenhang zweier nominal skalierten Variablen ausgedrückt. Am Beispiel der Tornado Vorhersage bedeutet dies, in insgesamt  $N$  Ereignissen tritt ein Tornado auf (1) oder er tritt nicht auf (0). Im Anschluss daran wurde die Übereinstimmung mit den Beobachtungen der tatsächlichen Tornadosichtungen abgeglichen und der statistische Zusammenhang berechnet. In dieser Arbeit wird dieser Test zur Evaluation von globalen und regionalen Klimamodellen basierend auf Referenzdaten in verschiedenen Regionen verwendet werden. In der Klimaforschung werden in der Regel metrische Daten, wie Temperatur- oder Niederschlagsunterschiede analysiert. Daher muss die Datengrundlage zunächst in nominal skalierte Daten umgeformt werden. Hierzu wird ein dynamischer Schwellenwert festgelegt. Die Wahl dieses Schwellenwerts hat dabei großen Einfluss auf die untersuchte Zielgröße. In dieser Arbeit wurde das regionale Mittel  $\bar{r}$  gewählt. Dabei handelt es sich um das arithmetische Mittel über alle Datenpunkte bzw. Gridboxen innerhalb einer Region. Dies wurde sowohl für das Modell  $\bar{r}_m$  als auch die Evaluationsdaten  $\bar{r}_e$  bestimmt. Die Werte aller untersuchten Datenpunkte  $N$  werden für das Modell ( $m_i$ ) und die Referenzdaten ( $e_i$ ) in jeweils zwei Kategorien unterteilt, je nachdem ob sie den Schwellenwert überschreiten oder darunter liegen. Mit diesen Werten wird die 4-Felder-Tabelle 6.1 gefüllt.

**Tabelle 4.1** 4-Felder-Tabelle.

	$e_i \geq \bar{r}_e$	$e_i < \bar{r}_e$	$\Sigma$
$m_i \geq \bar{r}_m$	$a$	$b$	$a + b$
$m_i < \bar{r}_m$	$c$	$d$	$c + d$
$\Sigma$	$a + c$	$b + d$	$N$

Die Felder  $a, b, c$  und  $d$  geben das Maß der Übereinstimmung bzw. der Abweichung wieder. Dabei zeigen hohe  $a$ - und  $d$ -Werte eine Übereinstimmung und umgekehrt vermehrt Werte für  $b$  und  $c$  größere Unterschiede. Bei der Betrachtung der 4-Felder-Metriken werden häufig die statistischen Größen wie Treffer-, oder Fehlerrate untersucht, die sich auch über  $a, b, c$  und  $d$  bestimmen lassen (STEPHENSON, 2000). Da es bei der Anwendung zum Vergleich von Klimamodellen mit Referenzdaten nicht um eine Vorhersage sondern statistische Evaluation geht, wird dies hier nicht weiter behandelt. Für weitere Informationen hierzu sei auf FINLEY (1884) und STEPHENSON (2000) verwiesen. In dieser Studie werden vier verschiedene Klimagrößen evaluiert: Das klimatologische Mittel und der Trend von jeweils saisonalem Niederschlag und Temperatur basierend auf dem Zeitraum 1960–2009.

Ein häufiges Problem der Klimamodellevaluation ist der sogenannte Bias. Mo-

delle können im statistischen Vergleich mit Beobachtungsdaten, insbesondere beim Trend, hohe Ähnlichkeiten zeigen, dennoch ist oft eine gewisse absolute Abweichung zu registrieren. Beim Jahresniederschlag kann das um mehrere 100mm der Fall sein. Durch die separate Betrachtung des regionalen Mittels von Beobachtungs- sowie Modelldaten wird dieser Bias eliminiert und die Analyse auf den Vergleich der regionalen Muster beschränkt. Dies ist ein Vorteil, den der hier entwickelte Ansatz zur Modellevaluation bietet.

**Tabelle 4.2** Formeln zur Berechnung der sechs 4-Felder-Metriken.

Bezeichnung	Formel	W	Referenz
Phi (PHI)	$\frac{ad-bc}{\sqrt{(a+b)(c+d)+(a+c)(b+d)}}$	$[-1;1]$	z.B. BORTZ ET AL. (2008)
Chi <sup>2</sup> (CHI)	$\frac{(ad-bc)^2}{(a+b)(c+d)+(a+c)(b+d)}$	$[0;1]$	DOOLITTLE (1885)
Heidke (HEI)	$\frac{2(ad-bc)}{(a+b)(c+d)+(a+c)(b+d)}$	$[-1;1]$	HEIDKE (1926)
Gilbert (GSS)	$\frac{a}{a+b+c}$	$[-1;1]$	GILBERT (1884)
Pierce (PIE)	$\frac{ad-bc}{(a+c)+(b+d)}$	$[0;1]$	PIERCE (1884)
Log Odds Ratio (LOR)	$\ln a + \ln d - \ln b - \ln c$	$]-\infty;\infty[$	z.B. STEPHENSON (2000)

In dieser Arbeit werden nicht nur eine 4-Felder-Metrik berechnet und untersucht, sondern insgesamt sechs. Diese unterscheiden sich ausschließlich hinsichtlich ihrer Formel und des Wertebereichs. Andere Parameter wie Eingangsdaten oder Funktionsweise sind identisch. Die Auswahl in Tabelle 6.2 orientiert sich dabei am Vergleich von STEPHENSON (2000) ergänzt um die PHI-Korrelation. Die sechs Metriken zeigen hinsichtlich ihrer Berechnungsweise teilweise hohe Ähnlichkeiten, die Wertebereiche schwanken dabei allerdings. So sind die PHI, Heidke (HEI) und Gilbert (GSS)-Korrelationen auf den Wertebereich  $[-1;1]$  definiert. Am Beispiel PHI bedeutet dies, auf die Evaluation der Modelle übertragen, dass Werte nahe an  $+1$  eine hohe Übereinstimmung zwischen Beobachtungsdaten und Modell ausdrücken, während Werte nahe an  $-1$  theoretisch eine gegensätzliche Beziehung darstellen. Werte um  $0$  zeigen an, dass kein relevanter Zusammenhang besteht. Den kleinsten Wertebereich mit  $[0;1]$  besitzen die PIE und CHI Metriken. Dagegen ist der Wertebereich von LOR theoretisch unendlich groß. Da negative Werte nicht zur Berechnung von Gewichten verwendet werden können, werden die Ergebnisse der 4-Felder-Metriken im Folgenden Rohgewichte  $g$  genannt. Durch Anwendung der Exponentialfunktion werden sie anschließend in positive Werte umgewandelt und basierend darauf die finalen Gewichte ( $\omega_i$ ) berechnet.

$$\omega_i = \frac{1}{\sum_{i=1}^n e^{g_i}} \cdot e^{g_i} \quad (4.1)$$

Hier beschreibt  $n$  die Anzahl aller evaluierten Modelle je MME. Die Unterschiede des Wertebereichs der 4-Felder-Metriken haben einen Einfluss auf die Anwendbarkeit ihrer Ergebnisse hinsichtlich einer Gewichtung. Neben der Bewertung und dem Vergleich der Modelle wird eine Gewichtung basierend auf den Ergebnissen der Metriken vorgenommen. Hier sind durch die stärkere Differenzierung von LOR die deutlichsten Auswirkungen zu erwarten. Allerdings hat LOR die Einschränkung, dass der Logarithmus nur für Werte  $a, b, c, d > 0$  definiert ist.

$$LOR = \ln a + \ln d - \ln b - \ln c \quad (4.2)$$

für  $a, b, c, d > 0$

Diese Einschränkung ist bei der Bewertung von Modellen aber nicht zielführend, da so gerade Modelle mit absoluter Übereinstimmung oder Widerspruch nicht betrachtet werden können, bzw. ein theoretisch unendlich positives oder negatives Gewicht erhalten würden. Aus diesem Grund behilft man sich in solchen Fällen damit, einen kleinen konstanten Wert 0,1 zu allen Feldern zu addieren, siehe KATERI (2014, S. 26), PARZEN ET AL. (2002) und HALDANE (1956). Dies übt zwar eine leichte Beeinflussung der Metrik aus, ist aber eindeutig nachzuvollziehen, da es sich nur um das jeweilige Extrem handelt. Die Addition von 0,1 ermöglicht für alle vier Felder aber gerade für kleinere Regionen, dass Modelle mit sehr hoher Trefferrate nicht aus der Betrachtung fallen. Gleichzeitig hat diese Addition keine Auswirkungen auf die Rangfolge der übrigen Simulationen.

$$LOR = \ln(a + 0,1) + \ln(d + 0,1) - \ln(b + 0,1) - \ln(c + 0,1) \quad (4.3)$$

für  $a, b, c, d \leq 0$

Im Gegensatz zum LOR benötigen die übrigen Metriken diese Anpassung nicht, da sie für  $a, b, c, d \geq 0$  definiert sind. Allerdings zeigt sich, dass die, im Vergleich zu LOR, kleineren Wertebereiche der fünf übrigen 4-Felder Ansätze in einer geringeren Differenzierung der Modelle bzw. der Rohgewichte untereinander resultieren. Als Ergänzung wurde hier daher für jede der 4-Felder-Metriken (4FM), die auch einen negativen Wertebereich besitzen, eine intensiviertere Version ( $\text{PHI}_i$ ,  $\text{HEI}_i$ ,  $\text{GSS}_i$ ,  $\text{LOR}_i$ ) berechnet, die stärkere Effekte, bezogen auf die finale Gewichtung, zur Folge haben. So wird ein Sub-Ensemble erstellt, d.h. Modelle, die eine negative Korrelation zeigen, erhalten automatisch ein Nullgewicht. Entsprechend gilt für die  $4\text{FM}_i$  bezogen auf Formel 4.1 für  $g_i \leq 0$  entsprechend  $\omega_i = 0$ . Diese

Ergänzung hat sich als hilfreich erwiesen, Effekte der Gewichtung auf das MME zu intensivieren und gleichzeitig die Art der Veränderung stabil zu behalten.

Teilergebnisse der Evaluation durch die 4-Felder-Metriken, hauptsächlich basierend auf dem Referenzdatensatz ERA-20C, wurden im Verlauf der Entstehung dieser Arbeit in RING ET AL. (2017) veröffentlicht. Die in dieser Arbeit dargestellten Resultate, mit Schwerpunkt auf die Evaluation der Mittelmeerregion auf Grundlage der E-OBS Beobachtungsdaten, stellen die vollständige Analyse dar, die sowohl alle Ergebnisse von RING ET AL. (2017) abdeckt als auch darüber hinausgehende Untersuchungen und die Verwendung weiterer Metriken.

## 4.2 Metrik der quadratische Abweichung (RM)

Die zweite Metrik, die als Klimamodellevaluation in dieser Arbeit Verwendung findet, basiert auf der mittleren quadratischen Abweichung  $MSE$ . Sie ist ein weit verbreitetes statistisches Maß und wird im geographischen Kontext zumeist eingesetzt, um die Genauigkeit von Vorhersagen zu bewerten (WILKS, 2006, S. 307). Es handelt sich hier um ein sehr intuitives Maß, da ohne besondere Vorverarbeitung für die Felder (Gridboxen) jeder individuellen Region ein Vergleich von Vorhersage  $x_i$  mit Beobachtung  $y_i$  angestellt wird. Dabei wird jeweils die quadrierte Differenz jedes Feldes aufsummiert und durch die Anzahl aller betrachteten Felder  $n$  dividiert.

$$MSE = \frac{1}{n} \sum_{i=1}^n (x_i - y_i)^2, \quad \mathbb{W} = [0; +\infty[ \quad (4.4)$$

Der Wertebereich  $\mathbb{W}$  liegt zwischen 0, bei absoluter Übereinstimmung von Simulation und Referenzdatensatz, und theoretisch unendlich. Zur leichteren Interpretation wird häufig die Wurzel der mittleren quadratischen Abweichung (eng. root mean square error, RMSE) verwendet.

$$RMSE = \sqrt{MSE} \quad (4.5)$$

Der Vorteil liegt hier darin, dass die Ausgangsgröße der untersuchten Variablen erhalten bleibt (z.B. mm oder °C). In dieser Arbeit wird der RMSE als Metrik (RM) zur Evaluation von Klimamodellen verwendet. Dabei orientiert sich das Verfahren an dem Vorgehen von RING ET AL. (2016). Dort wurde bereits der RMSE über die Abweichung von Modell- und Beobachtungsdaten als Vergleichsgröße berechnet. Allerdings wurde als  $n$  die Anzahl der zwölf untersuchten Monate gewählt und der RMSE anschließend über den Vergleich des mittleren Niederschlags über ein 30-jähriges Zeitfenster je Monat berechnet. Im Gegensatz

dazu wird in dieser Arbeit  $RM$  nicht über die Monate sondern über die Datenpunkte in jeder untersuchten Region  $n_r$  berechnet. Als untersuchte Größen  $(x_i, y_i)$  dienen hier, ebenso wie bei den 4FM (siehe Abschnitt 4.1), das 50-jährige klimatologische Mittel, der Trend des saisonalen und annuellen Niederschlags sowie der Temperatur.

$$RM_k = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_{k_i} - y_{k_i})^2} \quad (4.6)$$

Hier ist  $n$  abhängig von den untersuchten Regionen (bspw.  $n_{MMR} = 143$  für den Gesamtmitteleerraum) und  $k$  entspricht dem jeweiligen evaluierten Modell. Die  $RM_k$  Ergebnisse werden im Folgenden als die Rohgewichte der RMSE-Metrik bezeichnet. Dazu erfordert die Interpretation der Rohgewichte etwas Vorwissen. Im Gegensatz zu den 4FM steigen die  $RM$ -Rohgewichte mit zunehmender Abweichung an und das (theoretisch) bestmögliche Ergebnis ist ein  $RM_k$  von 0. Die Evaluation verschiedener Modelle basierend auf einer untersuchten Variable ist mit diesem Wissen sehr einfach. Beim Vergleich verschiedener Variablen muss jedoch die allgemeine Größe der Ausgangsdaten berücksichtigt werden. So kann es am Beispiel des Niederschlags leicht vorkommen, dass ein Modell einen  $> 100\text{mm}$  hohes Rohgewicht erhält und gemessen am Niederschlag der Region dennoch eine relativ hohe Übereinstimmung mit den Beobachtungsdaten hat. Entsprechend müsste ein hohes Gewicht zugeteilt werden. Gleichzeitig kann ein wenige Millimeter hohes Rohgewicht bezogen auf den Niederschlagstrend ein niedriges Gewicht zugeteilt bekommen, wenn der Trend der Beobachtungsdaten insgesamt sehr gering ist. Diese Komplikation ist bei den 4FM nicht zu beobachten, da hier die meisten Wertebereiche fest vorgegeben sind und sich die verschiedenen Größen daher leicht vergleichen lassen. Um die Resultate der  $RM$  ebenfalls zu standardisieren wird das jeweils kleinste Rohgewicht ( $\min(RM)$ ) jeder Situation als bestmögliches Ergebnis definiert. Das Modell, dass dieses Ergebnis erreicht hat, bekommt so das Gewicht  $\omega_{RM_{\min}} = 1$  zugeteilt.

$$\omega_{RM_k} = \frac{\min(RM)}{RM_k} \quad (4.7)$$

Diese Umrechnung ermöglicht zum einen den leichteren Vergleich der  $RM$ -Ergebnisse, zum anderen ist er notwendig, um die Berechnung der Gewichte vorzunehmen, da Modellen ( $k = 1, \dots, m$ ) mit höherer Abweichung so ein geringes Gewicht zugeteilt werden können. Mit jeder Region bzw. Variablen ändert sich das minimale Rohgewicht. Die Berechnung der finalen Gewichte ist auf Grundlage dieses, zwischen 1 und für sehr große Abweichungen gegen 0 tendierenden, definierten Wertebereichs analog zu 4FM. Zur Verbreitung des Spektrums wird

auch hier die Exponentialfunktion von  $\omega_{RM_k}$  verwendet. Dies dient der besseren Vergleichbarkeit der Metriken untereinander und ermöglicht zusätzlich ein etwas breiteres Spektrum der Gewichte. Eine intensiviertere Form nach dem Muster der  $4FM_i$  wurde nicht erstellt. Aufgrund des Wertebereichs  $\mathbb{W}_{\omega_{RM}} = ]0; 1]$  wäre dies jedoch über die Bestimmung eines Schwellenwertes der Rohgewichte oder eines anderen Wertes als 1 für das bestplatzierte Modell zwar möglich, aufgrund der Variabilität der unterschiedlichen Regionen aber nur mit vielen weiteren Hintergrundinformationen sinnvoll umsetzbar. Da ein Vorteil dieser Metrik die einfache und unkomplizierte Berechnung ist, wurde auf weitere Anpassungen in dieser Arbeit verzichtet. Die Berechnung der finalen Gewichte  $\omega_i$  erfolgt basierend auf den  $\omega_{RM_k}$  analog zu Formel 4.1.

### 4.3 Zeitreihenmetrik (ZM)

Die zuvor veranschaulichten Metriken (4FM, RM) basieren auf der statistischen Auswertung der Werte der einzelnen Gridboxen in jeder Region. Bei der hier vorgestellten Zeitreihenmetrik (ZM) wird die Evaluation der Klimamodelle basierend auf der Zeitreihe jeder Region für alle Variablen durchgeführt. Dazu werden aus den Werten zunächst jährliche Zeitreihen über den untersuchten Zeitabschnitt 1960–2009 erstellt.

$$y_r(t) = \sum_{i=1}^n x_i g_i \quad (4.8)$$

Hier entspricht  $y_r(t)$  dem Wert, der zum Zeitpunkt  $t$  angenommen wird.  $n$  beschreibt die Anzahl aller Felder der jeweiligen Region  $r$  und  $g_i$  das Gewicht, das abhängig von der geographischen Breite  $lat$  zwischen  $\pm 90^\circ$  auf die Werte  $x_i$  angewandt werden muss.

$$g_i = \frac{\cos\left(\frac{\pi \cdot lat_i}{180}\right)}{\sum_{j=1}^n \cos\left(\frac{\pi \cdot lat_j}{180}\right)} \quad (4.9)$$

Hintergrund ist, dass die Gitterpunkte der Modelle in jeder geographischen Breite gleich groß sind. Die Fläche wird jedoch aufgrund der Erdkrümmung kleiner mit steigendem Abstand zum Äquator. Die Gewichtung ist daher erforderlich, um keine Überrepräsentation der Außertropen in die Zeitreihen einfließen zu lassen. Eine Zeitreihenanalyse kann in der Regel auf zwei Ebenen durchgeführt werden: Der Zeit- und der Frequenzebene. Analysen auf Zeitebene betrachten und zerlegen die Zeitreihe in ihrem Originalzustand. Bezogen auf die Evaluation einer Niederschlagszeitreihe wäre hier die Anwendung einer mittleren quadratischen

Abweichung im Vergleich zu den Referenzdaten zu jedem Zeitpunkt eine einfache Herangehensweise. Diese Überlegung birgt jedoch einen elementaren Fehler: Da die Modelle nur eine statistische Simulation des Klimas liefern ist die Annahme, dass der Wert  $y_0(t)$  eines Modells mit dem Wert  $y_r(t)$  der Referenzdaten übereinstimmen könnte, falsch. Eine korrekte statistische Annahme wäre dagegen, dass der lineare Trend, der sich hinter der Zeitreihe von Modell- und Beobachtungsdaten verbirgt, übereinstimmen könnte, wenn das Zeitfenster groß genug gewählt ist. Da eine ähnliche Form der Auswertung bereits durch 4FM und RM durchgeführt wird, untersucht die Spektral- oder Zeitreihenmetrik (ZM) die Frequenzebene der Zeitreihen, als Indikator der Simulationsgüte. Hierbei geht es darum, dass Werte der Zeitreihe nicht in ihrer ursprünglichen Form betrachtet werden sondern eine Charakterisierung durch verschiedene Zeitskalen oder Frequenzen erfolgt. Die Formeln und Erläuterungen des Berechnungswegs bis zur Erstellung des Periodogramms beruhen auf der Harmonischen Analyse bzw. der spektralen Zerlegung von Zeitreihen nach WILKS (2006, S. 371-385). Tiefer gehende Hintergrundinformationen können aus KAY & MARPLE (1981) und GHIL ET AL. (2002) entnommen werden.

Es gilt, dass jede Zeitserie durch die Addition vieler verschiedener Sinus- und Kosinusfunktionen dargestellt werden kann. Da es sich bei Sinus und Kosinus um unendliche Kurven handelt, die zwischen  $[-1,1]$  pendeln, werden diese als harmonische Kurven bezeichnet. Daher wird die Zerlegung auch als harmonische Analyse bezeichnet. Die in dieser Arbeit angewandte Metrik der spektralen Zerlegung vergleicht die Bestandteile der Modell- und Referenzzeitreihen und erstellt daraus ein Ähnlichkeitsmaß, das in der Folge zur Modellbewertung und Gewichtung dient. Hierzu ist eine diskrete Fourier-Transformation erforderlich. Um die Erklärung nachvollziehbar zu gestalten, wird zunächst die Umsetzung jeder beliebigen Zeitreihe als Kombination von Sinus- und Kosinusfunktionen nach WILKS (2006, S. 372) erläutert. Dies sei zunächst am einfachsten Beispiel, der Darstellung einer Zeitreihe durch nur eine Funktion, gezeigt. Da sich Sinus und Kosinus alle  $360^\circ$  wiederholen gilt  $\cos(2\pi k + \alpha) = \cos(\alpha)$  und analog  $\sin(2\pi k + \alpha) = \sin(\alpha)$ . Hier entspricht  $\alpha$  einem beliebigen Winkel und  $k$  einer ganzen beliebigen Zahl. Der Versatz um  $90^\circ$  oder  $\frac{\pi}{2}$  von Sinus und Kosinus erlaubt die eine Funktion mit Hilfe der anderen auszudrücken.

$$\cos\left(\alpha - \frac{\pi}{2}\right) = \sin(\alpha) \quad (4.10)$$

$$\sin\left(\alpha + \frac{\pi}{2}\right) = \cos(\alpha) \quad (4.11)$$

Um jede beliebige Zeitreihe durch die Addition von einer endlichen Zahl von Sinus

und Kosinusfunktionen darzustellen, ist die Anpassung von Frequenz, Amplitude und Mittelwert der Kurve notwendig. Darüber hinaus muss jedoch die Funktion in Abhängigkeit der Zeit  $t$  der Originalzeitserie als eine Funktion in Abhängigkeit eines Winkels  $\alpha$  dargestellt werden, um die harmonischen Funktionen anwenden zu können. Hier gilt, dass ein vollständiger Durchlauf der Funktionen ( $2\pi$ ) mit der vollständigen Anzahl der Zeitschritte  $n$  (für die GKM  $n = 50$ ) gleichgesetzt wird. Über die Formel

$$\alpha = 2\pi \frac{t}{n} \quad (4.12)$$

kann so durch den Winkel  $\alpha$  der einzelne Zeitschritt  $t$  dargestellt werden. Für die  $t = 1$  gilt  $\omega_1 = 2\pi/n$ , was als die „fundamentale Frequenz“ (WILKS, 2006, S. 373) bezeichnet wird, die den Abschnitt von  $2\pi$  beschreibt, der bei einem einzelnen Zeitschritt vollzogen wird. Anschließend kann über die Addition des Mittelwertes der Ausgangsdaten  $\bar{y}$  die Ebene, um die die Funktionen pendeln, bestimmt und mit Hilfe einer Konstanten  $C_1$  die Frequenz festgelegt werden. Durch die Subtraktion des Winkels  $\Phi_1$ , der Phase, wird die Verschiebung entlang der Abszisse beeinflusst.

$$y(t) = \bar{y} + C_1 \cos\left(\frac{2\pi t}{n} - \Phi_1\right) \quad (4.13)$$

Hier gilt, dass der Kosinus sein Maximum annimmt wenn  $\Phi_1 = 2\pi t/n$ . Das Ziel zur perfekten Darstellung der Zeitreihe ist es, die beiden unbekanntes  $C_1$  und  $\Phi_1$  zu berechnen. Aufgrund der trigonometrischen Identität kann die Formel

$$C_1 \cos\left(\frac{2\pi t}{n} - \Phi_1\right) = A_1 \cos\left(\frac{2\pi t}{n}\right) + B_1 \sin\left(\frac{2\pi t}{n}\right) \quad (4.14)$$

aufgestellt werden. Zur Herleitung sei auf WILKS (2006, S. 375) verwiesen. Die Gleichungen drückt mathematisch aus, dass eine harmonische Welle entweder als eine Kosinusfunktion mit der Amplitude  $C_1$  und Phase  $\Phi_1$  oder als die Summe einer nicht verschobenen Kosinus- und Sinuswelle mit den Amplituden  $A_1$  und  $B_1$  dargestellt werden kann. Durch die Berechnung der beiden Amplituden lässt sich  $C_1$  bestimmen. Formel 4.14 beschreibt die Bestimmung exakt einer Kosinusfunktion. In den wenigsten Fällen ist eine Welle ausreichend, um eine Zeitreihe  $y(t)$  perfekt darzustellen. Hierzu ist die Addition von  $k = n/2$  harmonischen Funktionen bei geradem, bzw.  $k = (n - 1)/2$  bei ungeradem  $n$  notwendig (WILKS, 2006, S. 379).

$$y(t) = \bar{y} + \sum_{k=1}^{n/2} A_k \cos\left(\frac{2\pi t k}{n}\right) + B_k \sin\left(\frac{2\pi t k}{n}\right) \quad (4.15)$$

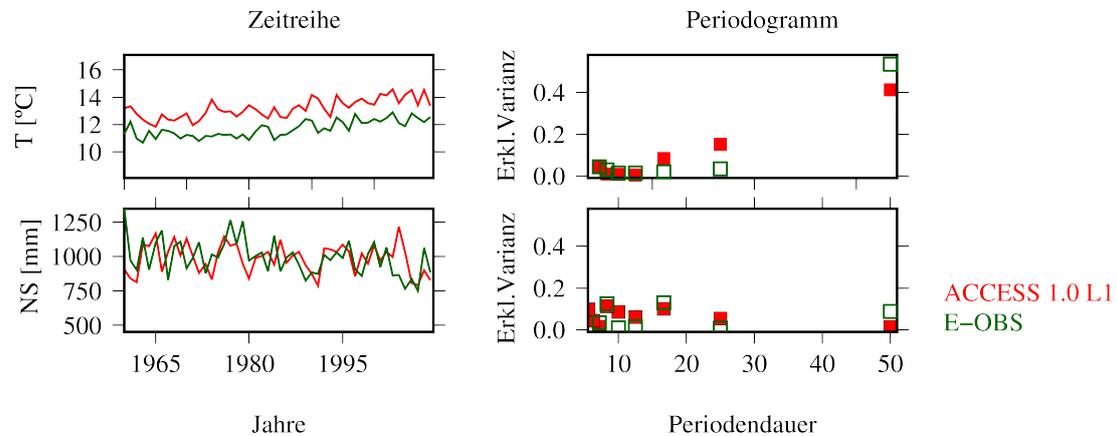
$A_k$  und  $B_k$  können über

$$A_k = \frac{2}{n} \sum_{t=1}^n y(t) \cos\left(\frac{2\pi tk}{n}\right) \quad B_k = \frac{2}{n} \sum_{t=1}^n y(t) \sin\left(\frac{2\pi tk}{n}\right) \quad (4.16)$$

berechnet werden. Diese Umformung wird als diskrete Fourier Transformation bezeichnet. Hierdurch wird die Niederschlags- und Temperaturzeitreihe jedes Modells durch  $n/2$  (in dieser Arbeit  $k = 1, \dots, 25$ ) Amplituden  $A_k$  und  $B_k$  bzw. die Amplitude  $C_k$  mit der Phase  $\Phi_k$  ausgedrückt. Da jede harmonische Funktion aufgrund der Beziehung von Sinus und Kosinus unabhängig von den anderen ist, kann über die Amplitude  $C_k$  und die Varianz  $s^2$  der Ausgangsdaten der Zeitreihe jeweils ein fester Anteil der erklärten Varianz  $R_k^2$  errechnet werden.

$$R_k^2 = \frac{(n/2)C_k^2}{(n-1)s^2} \quad (4.17)$$

Bezogen auf die Zeitreihen drücken sie den Anteil der erklärten Varianz der jeweiligen harmonischen Welle  $k$  aus. Entsprechend ist  $\sum_{k=1}^{n/2} R_k^2 = 1$ . Die Erstellung der Metrik, aus der sich die Modellbewertung ergibt, sei am Beispiel der Zeitreihen und zugehörigen Periodogrammen in Abbildung 4.1 erläutert.



**Abbildung 4.1** Veranschaulichung der Zeitreihenmetrik. Die eingefärbten Quadrate des Periodogramms (rechts) bzw. Zeitreihen (links) basieren auf der Auswertung der DJF Temperatur (oben) und Niederschlag (unten) der E-OBS Referenzdaten (grün) und ACCESS1.0 L1 (CMIP5) (rot) für den Gesamtmitteleerraum. Im Periodogramm ist auf der Abszisse die Periodendauer der jeweiligen harmonischen Welle und auf der Ordinate die dadurch erklärte Varianz  $R^2$  von Beobachtungs- und Modelldaten angetragen.

Für die Temperaturzeitreihen zeigt sich in diesem Fall ein leichter Bias mit einer etwas höheren Ausgangstemperatur in den Modelldaten aber einem relativ gleichmäßigen Anstieg. Die höchste erklärte Varianz wird hier durch die längste

Periodendauer von 50 Jahren erklärt und kann als der langfristige Temperaturtrend, sprich dem anthropogenen Klimawandel interpretiert werden.

Für eine hohe Bewertung eines Modells wird die Anforderung gestellt, dass die erklärten Varianzen der harmonischen Schwingungen eines Modells  $m$  möglichst nahe an denen der Beobachtungsdaten liegen sollten. Hierbei werden jedoch nur Periodenlängen über sieben Jahren betrachtet ( $p = 7, \dots, 50$  Jahre). Dies entspricht der durchschnittlichen Periode des ENSO-Ereignisses. Perioden darunter können physikalisch nicht mehr zugeordnet werden und gehen damit als Rauschen nicht in die Evaluation ein. Für den Niederschlag zeigt sich ein anderes Bild. Hier ist in der Zeitreihe links zwar ebenfalls ein Bias auszumachen, allerdings tritt kein deutlicher Trend zu Tage. In der Folge ist auch im Periodogramm die erklärte Varianz deutlich niedriger auf der 50 jährigen Periodendauer. Der Ähnlichkeitsindex, bzw das Rohgewicht,  $ZM_m$  zwischen den jeweilig erklärten Varianzen von Modell  $R_{p_m}^2$  und Referenz  $R_{p_0}^2$  wird über einen RMSE berechnet:

$$ZM_m = \sqrt{\frac{1}{7} \sum_{p=1}^7 (R_{p_m}^2 - R_{p_0}^2)^2} \quad (4.18)$$

Die Ergebnisse von  $ZM_m$  werden im weiteren Verlauf dieser Arbeit als Rohgewichte bezeichnet. Da hier wie beim  $RM$  kleine Werte von  $ZM_k$  einer hohen Simulationsgüte entsprechen, werden für die Berechnung der tatsächlichen Gewichte analog zur  $RM$  die Rohgewichte zunächst auf Grundlage der geringsten Abweichung  $\min(ZM)$  über Formel

$$\omega_{ZM_m} = \frac{\min(ZM)}{ZM_m} \quad (4.19)$$

normiert und abschließend die tatsächlichen Gewichte über die Summe aller Modelle je MME analog zu Formel 4.1 berechnet. Da bei der Zeitreihenmetrik die Modellevaluation, im Gegensatz zu allen anderen Metriken, nicht basierend auf dem Raummuster erfolgt und die erklärten Varianzen der Zeitreihenfunktionen nicht mit klimatologischen Mitteln oder Trends gleichgesetzt werden können, wird sie gesondert betrachtet.

## 4.4 Fingerprinting-Ansätze

Als vierte Metrik wird in diesem Kapitel das Fingerprinting zur Evaluation von Klimamodellen vorgestellt. Diese Methodik wurde erstmals von HASSELMANN (1979, 1993) vorgestellt. Beim Fingerprint handelt es sich generell um einen optimalen linearen Filter. Dies soll die Detektion des Klimawandelsignals ermöglichen.

Dieser Ansatz ist als zuverlässig anerkannt und wird häufig in der Klimaforschung für unterschiedliche Fragestellungen eingesetzt (HEGERL & NORTH, 1997; PAETH & HENSE, 2001; RIBES ET AL., 2010, 2013; RIBES & TERRAY, 2013; ZWIERS & ZHANG, 2003). Die in dieser Arbeit verwendeten Fingerprinting-Ansätze (FPA) basieren auf HASSELMANN (1993) und HEGERL ET AL. (1996). Der Aufbau zur Verwendung als Evaluationsmetrik und die Beschreibung der Formeln basiert auf den Studien von PAETH & HENSE (2001) und PAETH & MANNIG (2013).

Die Kenngröße  $d$  zur Beschreibung der Ähnlichkeit wird über das Skalarprodukt eines Referenzmusters (den Referenzdaten) und dem simulierten Muster (den Klimamodellen) berechnet. Diese werden als Vektoren dargestellt. Es gilt die Annahme, dass sowohl im Muster der Referenz- als auch der Modelldaten  $\vec{y}$  das Signal des Klimawandels  $\vec{y}_s$  kontaminiert durch Hintergrundrauschen  $\vec{y}_r$  ist.

$$\vec{y} = \vec{y}_s + \vec{y}_r \quad (4.20)$$

Es werden generell zwei FPA unterschieden. In dieser Arbeit wird die Nomenklatur von PAETH & HENSE (2001) übernommen, d.h. die beiden Fingerprintansätze werden als optimaler und suboptimaler Fingerprintansatz bezeichnet. In beiden Versionen wird eine Kenngröße  $d$  berechnet, die als Ähnlichkeitsmaß dient. Vereinfacht beschrieben ist das Ziel des optimalen Ansatzes (OPT), den Einfluss der natürlichen Variabilität dabei so stark wie möglich zu reduzieren, während der suboptimale Ansatz (SUB) dies vernachlässigt und Auskunft über die allgemeine Übereinstimmung gibt. Im Folgenden werden beide Ansätze vorgestellt. Zusätzlich wird ein erweiterter FPA (FPA<sub>z</sub>) erläutert. Zunächst wird die Idee des optimalen Fingerprints erläutert. Anschließend werden die Unterschiede von SUB und OPT sowie FPA und FPA<sub>z</sub> beschrieben.

#### 4.4.1 Optimaler und suboptimaler Fingerprintansatz (FPA)

In dieser Arbeit werden die Muster bzw. Vektoren  $\vec{y}$  des 50-jährigen Trends von saisonalem Niederschlag und Temperatur für 1960–2009 untersucht. Dabei besitzt  $\vec{y}$  die Dimension  $m$  abhängig von der untersuchten Größe der Regionen. In dieser Arbeit wird analog zu PAETH & MANNIG (2013) angenommen, dass kein linearer Zusammenhang zwischen  $\vec{y}_s$  und  $\vec{y}_r$  existiert. Die Kernaufgabe, die der Fingerprint zu lösen hat ist, die in Formel 4.20 formulierte Trennung von Signal und Klimarauschen, die zusammen dem beobachtbaren Klima  $\vec{y}$  entsprechen. Dazu müssen zunächst einige Annahmen festgelegt werden. So gilt, dass das Klimawandelsignal  $\vec{y}_s$  am besten gefiltert werden kann je größer der Anteil des Signals in Bezug zum Rauschen  $\vec{y}_r$  ist. Zusätzlich wird angenommen, dass das ungefähre Muster des

Signals aus theoretischem Wissen abgeschätzt werden kann. Hierzu empfehlen sich Klimamodellsimulationen, die unter realistischen Startbedingungen, durch beispielsweise erhöhte Treibhausgasemissionen eine Klimaänderung mit den zugehörigen Wechselwirkungen bezogen auf die Landoberfläche und Ozeane simulieren. Dieses als bekannt geltende Muster des Signals wird als Annahmevektor  $\vec{x}$  bezeichnet, der dieselben räumlichen Dimensionen  $m$  wie der beobachtete Klimavektor  $\vec{y}$  besitzt. Zwischen  $\vec{y}$  und  $\vec{x}$  besteht ein unbekannter Zusammenhang, der über einen konstanten Faktor  $A$  formuliert wird. Damit entspricht der FPA im Grundsatz einer allgemeinen linearen Regression (vgl. Abschnitt 3.6) (ALLEN & TETT, 1999).

Der Faktor  $A$ , der die Amplitude symbolisiert, kann für die Bestimmung der Rohgewichte  $d$  vernachlässigt werden, da hier nur die Muster relevant sind (HASSELMANN, 1997). Zur Bestimmung des Zusammenhangs wird das Skalarprodukt zwischen dem beobachteten und dem simulierten Trendmuster berechnet (HEGERL ET AL., 1996):

$$d = \vec{x} \cdot \vec{y} \quad (4.21)$$

Es gilt, dass das Skalarprodukt den Wert 0 annimmt, wenn beide Vektoren orthogonal (d.h. unabhängig) zueinander werden. Die Maximalwerte werden entsprechend erreicht, wenn die Vektoren parallel verlaufen. Nach PAETH & HENSE (2001) wird das Rohgewicht der FPA  $d(t_{sim}, t_{obs})$  für die 50-jährigen Trends der Referenz-  $\vec{y}(t_{obs})$  und der Modelldaten  $\vec{x}(t_{sim})$  über die Formel

$$d(t_{sim}, t_{obs}) = \frac{\sum_{i=1}^m x_i(t_{sim})y_i(t_{obs})}{\sqrt{\sum_{i=1}^m [x_i(t_{sim})]^2 \sum_{i=1}^m [y_i(t_{obs})]^2}} \quad (4.22)$$

berechnet und auf den Wertebereich  $[-1;1]$  normalisiert. Hiermit kann die Interpretation analog zum räumlichen Korrelationskoeffizienten oder auch teilweise den Ergebnissen der 4-Felder-Metriken erfolgen. Es ist nicht zwingend erforderlich dieselben Zeiträume  $t_{sim}, t_{obs}$  der 50-jährigen Trends miteinander zu vergleichen. Da das Ziel dieser Studie die Evaluation der Modelle für die zweite Hälfte des 20. Jahrhunderts ist, stimmen die Zeitabschnitte hier für den FPA jedoch überein. Für den im Anschluss vorgestellten zusätzlichen FPA wird von dieser Vorgabe abgewichen und es werden auch andere Zeitfenster berücksichtigt. Eine Einschränkung gilt generell für Modelle von CORDEX, die nur einen kürzeren Zeitabschnitt (seit 1970) abdecken. Hier wurde das Zeitfenster der Referenzdaten auf den 40-jährigen Zeitraum 1970–2009 verkürzt. Aus Formel 4.21 folgt, dass der Annahmevektor als Filter für den Klimavektor dient. Ersterer wird aus Modellen

berechnet.

Die Verwendung von  $\vec{x}$  als Filter bzw. Fingerprint lässt bei diesem Ansatz außer acht, dass sowohl  $\vec{x}$  als auch  $\vec{y}$  durch Klimarauschen kontaminiert sind (HEGERL ET AL., 1996). Dieses Rauschen kann sowohl zu einer Verstärkung als auch einer Abschwächung des Rohgewichts führen. Aus diesem Grund wird dieser Ansatz zur Berechnung der Rohgewichte als suboptimaler Fingerprint SUB bezeichnet (PAETH & HENSE, 2001). Für die Bestimmung des unverfälschten Rohgewichts muss  $d$  ebenfalls in einen Signal- und Rauschanteil unterteilt werden:

$$d = d_s + d_n \quad (4.23)$$

Aus statistischer Sicht gilt, dass der optimale Filter das quadrierte Verhältnis von Signal-zu-Rauschen maximiert.

$$R^2 = \frac{d_s^2}{d_n^2} = \max \quad (4.24)$$

Zu diesem Zweck wird ein optimaler Filter  $\vec{x}_{opt}(t)$  benötigt. HASSELMANN (1979 & 1993) zeigt, dass die Anforderung aus Formel 4.21 durch Verwendung der inversen Kovarianzmatrix  $C^{-1}$  der natürlichen Variabilität erreicht wird. Damit ist die Berechnung des optimalen Rohgewichts  $d_{opt}$  über

$$d_{opt}(t_{sim}, t_{obs}) = \vec{x}_i(t_{sim}) \cdot \vec{y}_j(t_{obs}) \quad (4.25)$$

mit dem Filter

$$\vec{y}_{opt}(t_{obs}) = C^{-1} \cdot \vec{y}_{orig}(t_{obs}) \quad (4.26)$$

möglich. Die inverse Kovarianzmatrix  $C^{-1}$  besitzt  $m \times m$  Dimensionen. Durch die Anwendung des optimalen Filters wird der Annahmevektor rotiert, sodass er nicht mehr in die Richtung des maximalen Rauschens zeigt und mit dem Klimavektor korreliert. In der Folge berechnet sich das optimale Rohgewicht  $d_{opt}$  über

$$d_{opt}(t_{sim}, t_{obs}) = \frac{\sum_{i=1}^m \sum_{j=1}^m x_i(t_{sim}) y_j(t_{obs}) C_{ij}^{-1}}{\sqrt{\left[ \sum_{i=1}^m \sum_{j=1}^m x_i(t_{sim}) C_{ij}^{-1} x_j(t_{sim}) \right] \left[ \sum_{i=1}^m \sum_{j=1}^m y_i(t_{obs}) C_{ij}^{-1} y_j(t_{obs}) \right]}} \quad (4.27)$$

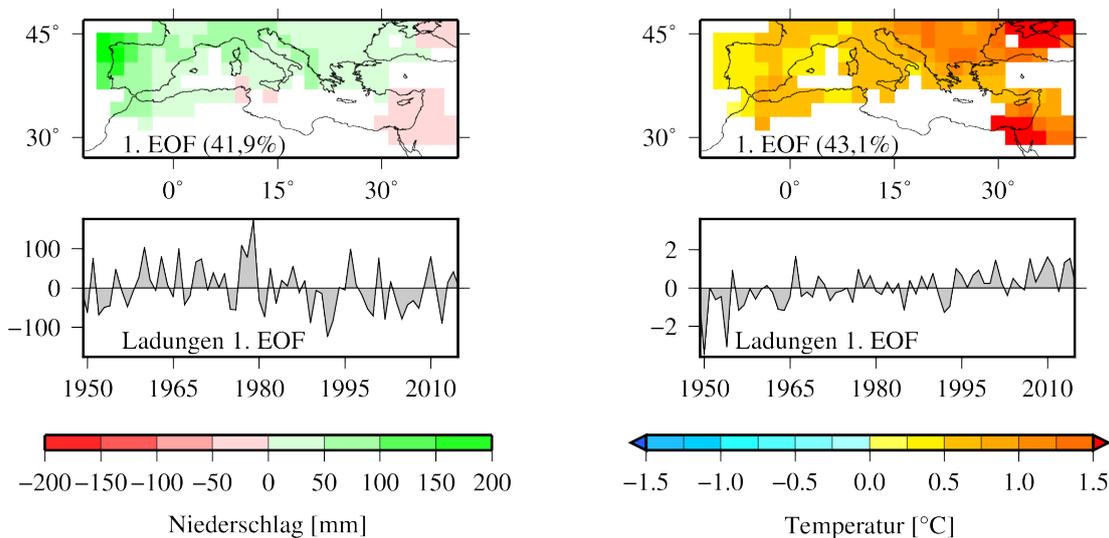
Dieser Ansatz wird zur Berechnung der OPT-Rohgewichte verwendet. Beide Ansätze sind für die Fragestellung dieser Arbeit relevant und werden gleichgestellt analysiert. Dabei gibt SUB analog zu den 4-Felder-Metriken eine Auswertung der Ähnlichkeit des Gesamttrends von Modell- und Beobachtungsdaten aus.

OPT zeigt hingegen eine komplexere Analyse, welche die Übereinstimmung der Modelle mit den Referenzdaten, bezogen auf das Klimawandelsignal, untersucht. Wie zuvor erläutert wird der Anteil des Rauschens, der im SUB möglicherweise einen positiven Effekt auf das Rohgewicht hat, über die Verwendung des optimalen Filters entfernt.

Die Berechnung des Filters  $C^{-1}$  ist hierbei die größte Herausforderung. Die natürliche Variabilität des Klimas ist im Prinzip unbekannt. Die ideale Quelle für Informationen hierzu wären Zeitreihen aus Beobachtungsdaten in einem vom Menschen unbeeinflussten Zeitraum. Da die Referenzdaten nur den Zeitraum seit 1950 abdecken, wurde in dieser Arbeit auf Klimamodellsimulationen zurückgegriffen. Es gilt, dass der Einfluss der Treibhausgase auf das Klima zu Beginn der historischen Modellsimulationen im 19. Jahrhundert sehr gering ist. Daher bietet dieser Zeitraum Informationen zur Abschätzung einer simulierten natürlichen Variabilität. Mit diesem Wissen wurden alle verfügbaren 50-jährigen Trends aktueller historischer GKM-Simulationen bis einschließlich des Zeitintervalls 1900–1949 verwendet.

Die Abschätzung der natürlichen Variabilität, bzw. die Berechnung der invertierten Kovarianzmatrix  $C^{-1}$ , macht aufgrund der regionalen Dimensionen eine Datenreduktion erforderlich. Zu diesem Zweck wurde eine Hauptkomponentenanalyse durchgeführt. Ihr Ziel ist es einen möglichst hohen Anteil der erklärten Varianz einer großen Datengrundlage durch neue künstliche Variablen wiederzugeben und so die Daten auf wenige unkorrelierte Variablen zu reduzieren. Diese Variablen werden als Hauptkomponenten oder Empirische Orthogonalfunktionen (EOF) bezeichnet. Eine detaillierte Herleitung der EOF-Analyse kann von VON STORCH & ZWIERS (1999, S. 293ff.) nachvollzogen werden. In dieser Arbeit wurde auf die EISPACK-Routine RS (SMITH ET AL., 1974; GARBOW, 1974) zurückgegriffen, um die EOFs sowie die entsprechenden Ladungen zu berechnen. Nach HASSELMANN (1993) wurde für die Berechnung des Fingerprintingansatzes dazu eine rotierte Kovarianzmatrix verwendet. Zunächst wurde eine EOF-Analyse über die enttrenteten (vgl. Kapitel 3.6) Referenzdaten im verfügbaren Zeitraum durchgeführt. Da jede Hauptkomponente einen unterschiedlichen Betrag der Varianz erklärt, wurden sie in absteigender Reihenfolge geordnet, sodass die ersten EOF die größte erklärte Varianz besitzen. Anschließend wurden die historischen Läufe bis 1949 der Modelldaten auf die EOF projiziert. Dabei wurde mit einem Moving-Window-Verfahren das Start- und Endjahr der Trends um ein Jahr nach vorne verschoben, um dadurch eine deutlich größere Anzahl an Trends (>2000 Trends) zu ermöglichen. Über diese Trends wurde anschließend die Kovarianzmatrix berechnet. Obwohl hier keine Beobachtungsdaten zur Bestimmung der tatsächlichen natürlichen Variabilität verwendet wurden, zeigen die Ergebnisse aufgrund der großen Datenmenge eine zufriedenstellende Abschätzung des Kli-

marauschens. Das Signal des Klimawandels wird deutlich wahrnehmbar gefiltert. Zwar werden nicht alle Effekte des Signals dargestellt aber zumindest optimiert die Anwendung des Filters das Verhältnis von Signal und Rauschen im Vergleich zum suboptimalen Ansatz. Abbildung 4.2 zeigt exemplarisch die ersten EOFs sowie deren Ladungen für Niederschlag und Temperatur.



**Abbildung 4.2** Die ersten EOF (oben) sowie zugehörige Ladungen (unten) des Gesamtmitteleerraums für DJF Temperatur (rechts) und Niederschlag (links), basierend auf dem Gesamtzeitraum der E-OBS Daten von 1950–2015.

Bei der Temperatur wird für die erste EOF ein sehr homogenes Muster angezeigt, wobei die Intensität im Osten erhöht ist. Auch die Ladungen zeigen seit Mitte des 20. Jahrhunderts einen stetigen positiven Trend. Daher liegt die Vermutung nahe, dass hier die anthropogene Erwärmung mit einer erklärten Varianz von 43,1 % abgebildet wird. Beim Niederschlag zeigt die erste EOF ein zweigeteiltes Bild. So ist im Norden und Westen der Region eine Zunahme und im Osten und Süden eine leichte Abnahme zu beobachten. Auch die Ladungen zeigen hier ein deutlich heterogenes Bild, wobei ca. ab 1980 negative Ladungen zunehmen. Dieses Muster kann mit den Einflüssen der NAO in Verbindung gebracht, die sich in der zweiten Hälfte des 20. Jahrhunderts besonders stark auf die Klimatologie des Niederschlags ausgewirkt hat (DÜNKELOH & JACOBET, 2003; KELLEY ET AL., 2012b; PAXIAN ET AL., 2013). Mit einer erklärten Varianz von 41,9 % zeigt diese, in der ersten Niederschlags-EOF, nahezu den gleichen Anteil, den der anthropogene Klimawandel in der ersten Temperatur-EOF dargestellt. Auch die weiteren EOF zeigen kein Muster, das eindeutig letzterem zugeordnet werden kann. Daher ist davon auszugehen, dass die Filterung des Klimawandelsignals für den Niederschlag eine deutlich größere Herausforderung als für die Temperatur

ist.

Die Berechnung der FPA wurden letztendlich für 8 und 15 EOFs durchgeführt. Die analysierte erklärte Varianz sollte nicht einen zu großen Teil abbilden, da hierdurch die Vorteile, die durch die EOF-Analyse gewonnene Datenreduktion, verloren gehen. Mit 72,7 % bis 84,0 % lassen die Werte des Niederschlags noch mindestens 15 % erklärter Varianz offen. Da bei der Temperatur die Werte bereits mit deutlich weniger EOF über 90 % liegen, wurde hier 8 EOF als Kompromiss gewählt. Um die Effekte einer nahezu vollständigen erklärten Varianz zu untersuchen wurden die FPA außerdem basierend auf den ersten 15 EOF berechnet (vgl. Tabelle 4.3).

**Tabelle 4.3** Summe der erklärten Varianzen der verwendeten Anzahlen von empirischen Orthogonalfunktionen des Gesamtmitteleerraums. Angetragen sind alle saisonalen Werte für Niederschlag (rechts) und Temperatur (links).

	Temperatur		Niederschlag	
	8 EOF	15 EOF	8 EOF	15 EOF
MAM	95,0	98,0	72,7	86,4
JJA	94,1	97,5	77,6	88,0
SON	94,0	97,6	75,1	86,0
DJF	95,3	98,0	84,2	92,4

Da für den SUB keine Kovarianzmatrix erforderlich ist, wurde hier die Berechnung der Rohgewichte zusätzlich basierend auf den Originaldaten ohne Datenreduktion durchgeführt, um eventuelle Unterschiede ansprechen zu können (SUB-Px). Zur Vereinfachung werden bei der späteren Auswertung die FPA Ansätze zusammen mit der verwendeten EOF Zahl bezeichnet: So entspricht bspw. SUB-8 dem suboptimalen Ansatz basierend auf 8 EOF. Da das analysierte Trendmuster bei kleineren Untersuchungsgebieten einen großen Teil der Aussagekraft verliert, wurden die FPA als einzige Metrik nur auf die Gesamtregion des Mittelmeerraums und die global verteilten Erweiterungsregionen angewandt. Auch hier erfolgt die Berechnung der finalen Gewichte analog zu 4FM, RM und ZM über Formel 4.1.

#### 4.4.2 Zusätzliche Fingerprintansätze (FPA<sub>z</sub>)

Neben den zuvor erläuterten FPA gibt es noch zusätzliche Fingerprintingansätze (FPA<sub>z</sub>) in dieser Arbeit. Dieser wurde zum einen entwickelt, um eine intensivierte Form der Gewichtung nutzen zu können bei gleichzeitiger Nutzung eines größeren Hintergrundwissen bei der Bewertung der Simulationsgüte. Generell gilt, dass die FPA durch Normierung, ebenso wie beispielsweise PHI, auf den Wertebereich  $[-1,1]$  festgelegt sind. Dies hat zum einen den Vorteil, dass es eine sehr gute

Vergleichsmöglichkeit zwischen den Rohgewichten der Simulationen und Metriken gibt. Andererseits ist die Möglichkeit der Abstufung zwischen den Modellgewichten, analog zu beispw. PHI, begrenzt. Um allein dies zu ändern, wäre es möglich, dieselbe Intensivierungstechnik wie bei den 4FM zu nutzen. Dazu würden Modelle, die ein negatives Rohgewicht zugeteilt bekommen haben, nicht weiter berücksichtigt werden. Da dies jedoch bereits für die 4FM verwendet und analysiert wird, wurde hier für die  $FPA_z$  ein anderer Ansatz gewählt.

Die Annahme der  $FPA_z$  lautet, dass Modelle trotz des 50-jährigen Intervalls nicht zwingend in derselben Phase wie die Referenzdaten sein müssen. Daher wird nicht nur ein Rohgewicht pro Simulation berechnet sondern sämtliche, die durch alle drei Referenzdatensätze (1950–1999, ..., 1961–2009) möglich sind. Die gleichen 12 Zeitabschnitte werden auch für die Modelldaten genutzt, wobei die Zeitfenster von Modell und Referenzdatensatz nicht identisch sein müssen. So können insgesamt 144 Rohgewichte  $d$  pro Simulation berechnet werden. Zur Bestimmung des finalen Rohgewichts  $d_z$  werden die Einzelergebnisse anschließend der Größe nach geordnet und der Median bestimmt. Da dieser in jedem Fall im Wertebereich  $[-1,1]$  liegt, wird abschließend 0 als Schwellenwert für das erste Quartil festgelegt. Die Wahl dieses Schwellenwertes erfolgte mit Blick auf die zu Grunde liegenden Daten. Beliebige andere Grenzwerte wären ebenfalls möglich. Der Vergleich unterschiedlicher Versionen brachte für das erste Quartil die subjektiv beste Filterung von Modellmedianen. Damit werden nur solche Simulationen zur Gewichtung genutzt, von denen 75 % der Werte (108 von 144) ein positives Rohgewichte  $d$  besitzen. Somit können Simulationen, die nur zufällig eine hohe Übereinstimmung im  $FPA$  Zeitraum 1960–2009 besitzen, aussortiert werden und andere, die konstant hohe Werte zeigen, intensiviert werden. Da die  $FPA_z$  generell keine zeitliche Übereinstimmung nutzen, bieten die  $FPA_z$  auch für CORDEX die Möglichkeit einen besseren Eindruck der Simulationsgüte zu erhalten.

Grundsätzlich ist diese Methode der Intensivierung für alle Metriken möglich. Da dies den Rahmen dieser Studie jedoch sprengen würde, wurde sie explizit nur für die Metrik des Fingerprintings angewandt. Die Begrenzung auf generell positive Rohgewichte zur Erzeugung eines Sub-Ensembles wurde hingegen ausschließlich nur für die 4FM ( $4FM_i$ ) verwendet. Das Fingerprinting bietet den Vorteil, dass es hier bereits zu einer zweistufigen Betrachtung des Klimas (SUB; OPT) kommt und somit die Phasenverschiebung sowohl für das beobachtete Klima (SUB) als auch reduziert auf das anthropogene Signal (OPT) untersucht wird. Zusätzlich ist durch die ebenfalls phasensynchrone Evaluation ( $FPA$ ) auch eine Vergleichsmöglichkeit gegeben, wodurch die Unterschiede der Modellbewertung aufgezeigt werden. Um dennoch auch für die übrigen Metriken weitere Informations- und Vergleichsmöglichkeiten zu bieten, wurde die Evaluation der Klimamodelle, neben dem Zeitraum 1960–2009, auch noch für das Zeitintervall

1950–1999 durchgeführt. Damit wird abschließend die Stabilität der Modellgewichte sowie die zeitliche Übertragbarkeit beleuchtet (siehe dazu Kapitel 8.2). Nachdem die Evaluationsmetriken vorgestellt wurden, erfolgt in den anschließenden Kapiteln die Auswertung der Ergebnisse mit Schwerpunkt auf der Region des Mittelmeerraums.

## 5 Modellevaluation im Mittelmeerraum

In diesem Kapitel werden die Ergebnisse der Modellevaluationen basierend auf allen Metriken für den Gesamtmittelmeerraum sowie seiner Unterregionen vorgestellt. Lediglich die Fingerprinting-Metriken werden ausschließlich für den Gesamtmittelmeerraum angewandt. Die dargestellten Ergebnisse beziehen sich auf die Evaluationsperiode 1960–2009. Der zweite Zeitraum 1950–1999 wird in Kapitel 8 zur Untersuchung der zeitlichen Konsistenz herangezogen.

Dieses Kapitel ist so aufgebaut, dass zunächst die Ergebnisse der Modellevaluationen basierend auf den Rohgewichten aller Metriken (vgl. Kapitel 4) regional analysiert werden. Anschließend werden saisonale Unterschiede herausgearbeitet. Danach folgt eine allgemeine Analyse der Entwicklung der MMEs. Abschließend wird die Übereinstimmung der Ergebnisse verschiedener Metriken untersucht und die Eignung der Modellbewertungen für Gewichtungen eingeschätzt. Damit soll Aufschluss über die Bedeutung der unterschiedlichen Metriken für die in Kapitel 6 durchgeführten Gewichtungen geben werden.

Im Folgenden sind bei Darstellungen der Simulationsrohgewichte auf der Abszisse grundsätzlich Simulationsnummern (SN) angetragen. Diese ermöglichen durch Abgleich mit den Tabellen 2.2, 2.3 und 2.4 die Zuordnung des Ergebnisses jeder einzelnen Simulation. Die Reihenfolge ist dabei immer CMIP3-CMIP5-CORDEX. Die SN sind abhängig von der jeweiligen Variablen (Temperatur, Niederschlag).

### 5.1 Regionale Variabilität der Bewertung

In diesem Abschnitt werden die individuellen Evaluationsergebnisse der Simulationen untersucht. Zunächst werden die Unterschiede zwischen den Simulationen innerhalb einer Region sowie die Korrelation der Ranglisten zwischen verschiedenen Regionen betrachtet. Neben der annualen Evaluation wurde dies für jede Jahreszeit sowohl für Niederschlag als auch Temperatur durchgeführt. Die Kombination aus einer bestimmten Region, Jahreszeit und Variablen wird als „Situation“ bezeichnet. Die dargestellten annualen Ergebnisse sind dabei exemplarisch

für das Erscheinungsbild der vier Jahreszeiten, auch wenn die einzelnen Werte variieren können. Ein detaillierter Vergleich der Simulationsgüte für unterschiedliche Jahreszeiten wird im Anschluss in Kapitel 5.2 vorgenommen. Im Folgenden wird der Schwerpunkt auf die regionalen Unterschiede der Modellbewertung gelegt.

### 5.1.1 Unterschiede der Modellbewertungen

Zur Evaluation der Modelle werden insgesamt 13 Grund-Metriken, eingeteilt in vier Metrikblöcke, angewandt. Diese teilen sich in sechs 4FM, fünf FPA und je eine RM und ZM auf (siehe Kapitel 4). Ergänzend wird an bestimmten Stellen auf die Sub-Ensemble-Ansätze ( $4FM_i$ ) für vier der 4FM sowie auf die zusätzlichen Ansätze des Fingerprintings ( $FPA_z$ ) hingewiesen. Für die 4FM und RM wurde neben dem 50-jährigen Trend auch das klimatologische Mittel zur Evaluation herangezogen. Da es nicht möglich ist, alle Ergebnisse für jede untersuchte Situation zu zeigen, werden exemplarisch Beispiele dargestellt und auf Besonderheiten der verschiedenen Metriken und Situationen hingewiesen.

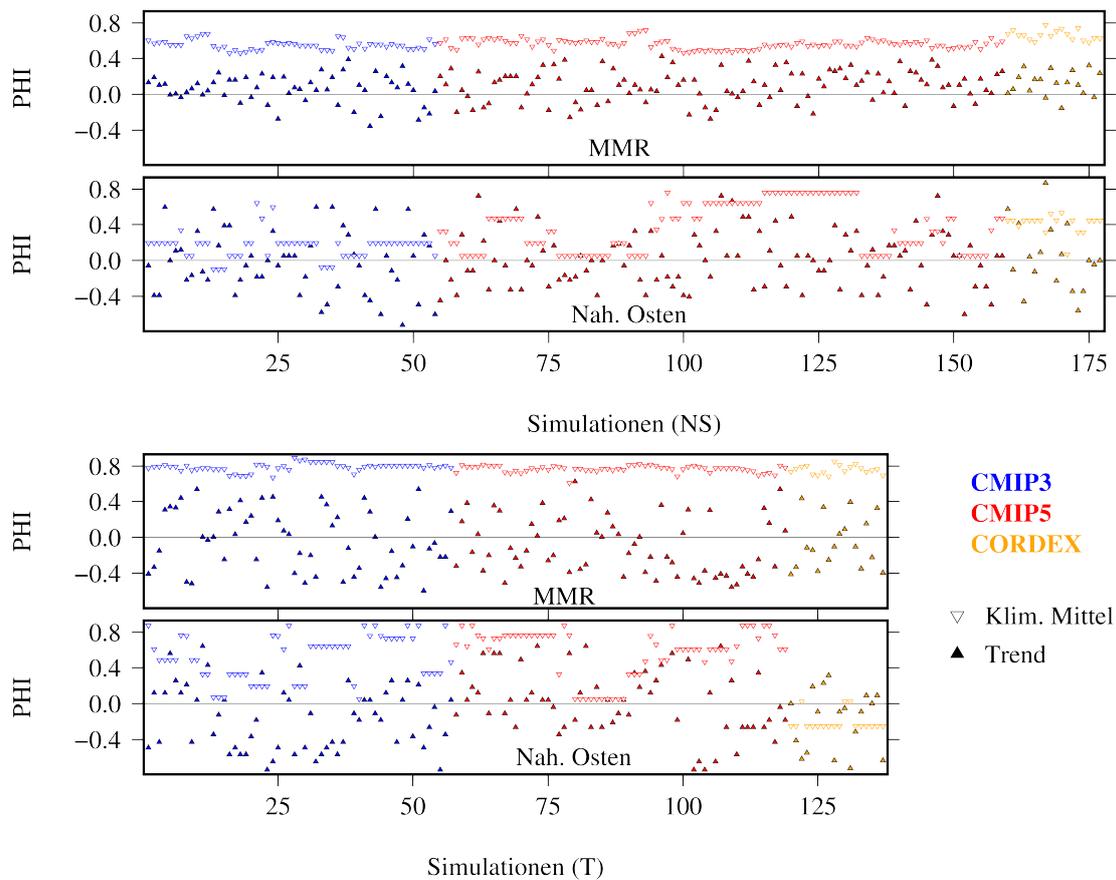
#### 4FM

Abbildung 5.1 zeigt die Evaluationsergebnisse basierend auf PHI für den annualen Niederschlag und die mittlere Jahrestemperatur. Für beide Variablen ist das PHI-Rohgewicht des MMR und für die Region Naher Osten exemplarisch angebracht. Neben den Trendrohgewichten sind auch die des klimatologischen Mittels für jede Simulation eingezeichnet.

Beim Vergleich der Niederschlagsrohgewichte fällt auf, dass es bei MMR eine relativ klare Trennung der Ergebnisse von Trend und Mittel gibt. Während die Ergebnisse des klimatologischen Mittels homogen zwischen 0,4 und 0,8 schwanken, ist die Streuung für den Trend zwischen  $\pm 0,4$  deutlich stärker ausgeprägt. Dies lässt darauf schließen, dass der Trend komplexer zu simulieren ist bzw. das klimatologische Mittel der Modelle überwiegend besser mit dem der Referenzdaten übereinstimmt.

Für den Trend der Unterregionen ist ein noch heterogeneres Bild als beim MMR zu beobachten. Insbesondere für den Nahen Osten und Italien, zwei recht niederschlagsarme Untersuchungsgebiete, wird die Skala sowohl positiv als auch negativ nahezu vollständig ausgeschöpft. Bei den meisten anderen Unterregionen ist das Bild zwar ähnlich, der Wertebereich liegt jedoch innerhalb von  $\pm 0,8$ . Mit dem Nahen Osten und MMR sind in Abbildung 5.1 die beiden Regionen mit der jeweils größten bzw. geringsten Streuung der Trendwerte angetragen. Die übrigen Unterregionen liegen zwischen diesen Extremen.

Während die Regionen Nordatlantik und Balkan (nicht gezeigt) beispielsweise



**Abbildung 5.1** Annuelle PHI-Rohgewichte der Modellevaluation für Niederschlag (oben) und Temperatur (unten). Angetragen sind die Ergebnisse des klimatologischen Mittels und des Trends im Zeitraum 1960–2009 für den Gesamtmittelmeerraum und den Nahen Osten. Über die Ordnungszahl (SN) auf der Abszisse ist die Zuordnung der Ergebnisse für jede Simulation (vgl. Kapitel 2.2) möglich.

in Wertebereich und Streuung eher MMR ähneln, besitzen Nordafrika und besonders Italien eine vergleichbare Streuung wie der Nahe Osten. Die Gründe hierfür können in der Größe der Region liegen. Mit 143 Datenpunkten ist MMR mit Abstand die größte Region, gefolgt vom Balkan (29), während die wenigsten Datenpunkte für Italien (12) betrachtet werden. Daneben scheint auch die jährliche Niederschlagsmenge eine relevante Rolle zu spielen. Insbesondere humide Regionen haben eine geringere Streuung mit insgesamt nur durchschnittlichen PHI-Korrelationen. Umgekehrt zeigen aride Regionen extremere PHI-Werte. Dieses Bild des Streuungsverhaltens lässt sich auch auf die Ergebnisse des 50-jährigen klimatologischen Mittels übertragen. Allerdings liegen hier, anders als beim Trend, bis auf vereinzelte Simulationen beispielsweise in Italien die Rohgewichte für al-

le Regionen überwiegend deutlich im positiven Bereich. Auch hier liefert MMR das einheitlichste Bild mit relativ homogenen Werten. Die teilweise sehr hohen Ergebnisse für das Niederschlagsmittel müssen zum einen auf die hohe Qualität der Simulationen zurückgeführt werden, zum anderen sind analog zum Trend Faktoren wie beispielsweise die jährliche Niederschlagsmenge vermutlich mitentscheidend. Insgesamt fällt beim klimatologischen Mittel teilweise eine Häufung verschiedener Simulationen mit sehr ähnlichen Rohgewichten in den Teilregionen auf. Hintergrund sind hier unterschiedliche Realisationen (Läufe) desselben Modells (Ensemble). Trotz unterschiedlicher Startbedingungen zeigen sie häufig sehr ähnliche Resultate für das klimatologische Mittel in den untersuchten Regionen. Da dies jedoch nicht in allen Fällen zutrifft, scheint die Größe der Region hier ebenfalls maßgeblich zu sein. So kann es bei der Berechnung der 4-Felder-Korrelationen bei einer geringen Anzahl von Datenpunkten sehr leicht zu hohen Übereinstimmungen kommen. Allerdings gilt hier, dass diese Art der Gruppierung in keiner Region für die Trendergebnisse auftritt. Die verschiedenen Realisationen scheinen sich daher beim Trend stärkerer auszuwirken als beim Mittel. Bezogen auf die verschiedenen MMEs ist auf dieser Betrachtungsebene kein entscheidender Unterschied auszumachen.

Die Temperaturergebnisse zeigen im Vergleich zum Niederschlag insgesamt deutlich höhere Rohgewichte für alle Simulationen und über alle Regionen. Für den Trend sind hier Italien, das Schwarze Meer und insbesondere die Ägäis zu nennen. Hier liegen die Rohgewichte zu großen Teilen oberhalb von 0,5 (nicht gezeigt). Die übrigen Untersuchungsgebiete zeigen eine relativ gleichmäßige Streuung zwischen 0–1, wobei auch vereinzelt Werte im negativen Bereich liegen. Auffällig ist hier ebenfalls MMR, bei dem vergleichbar zum Niederschlag größtenteils eine homogene Streuung um  $\pm 0,5$  zu sehen ist. Zusammen mit dem Nahen Osten und dem Balkan ist hier die Simulationsgüte am geringsten bzw. der Unterschied zwischen den Rohgewichten am größten.

Die Auswertung der mittleren Jahrestemperatur liefert ein anderes Bild. MMR zeigt zusammen mit dem Balkan und Spanien insgesamt die besten Resultate mit einem Großteil der Werte oberhalb von 0,7. Analog zum Niederschlag ist die Streuung allgemein niedriger als noch beim Trend. In den meisten Regionen liegen die Werte im positiven, teilweise sehr hohen Bereich. Nur für die Ägäis und das Schwarze Meer trifft diese Beobachtung nicht zu. Während für Erstere die meisten Ergebnisse zwischen  $\pm 0,4$  schwanken, liegen sie bei Letzterem nahezu vollständig im teilweise stark negativen Bereich. Da für die übrigen Regionen insgesamt hohe bis sehr hohe PHI-Korrelationen berechnet werden, scheint das Temperaturmuster dieser beiden angrenzenden Region nicht von den Modellen wiedergegeben zu werden.

Insgesamt zeigt sich, dass es deutliche Unterschiede zwischen den Simulatio-

nen gibt. Diese sind sowohl bei Niederschlag als auch bei Temperatur für das Mittel und den Trend zu beobachten. Dies muss nur für MMR relativiert werden, da hier vermutlich die Größe der Region dafür sorgt, dass die Ergebnisse des Mittels relativ homogen über alle Regionen ausfallen. Darüber hinaus gilt, dass die Rohgewichte der Temperatur im Schnitt etwas höher als die des Niederschlag ausfallen. Diese Eindrücke decken sich grundsätzlich mit anderen Studien zur Modellevaluation (z.B. KUMAR ET AL., 2013).

Die Ergebnisse der fünf übrigen 4FM zeigen überwiegend identische Ergebnisse. Dabei gilt zu beachten, dass der Wertebereich der Metriken schwankt. So fallen auch die Abstufungen der Rohgewichte teilweise anders aus, was sich auf die Streuung auswirkt. Hier sind besonders PIE und LOR zu nennen. Erstere liefern ein sehr ähnliches Bild wie PHI für alle Regionen, allerdings mit deutlich minimierter Streuung. Das Gegenteil ist bei LOR zu beobachten, die bei angepasster Skala ein identisches Bild wie PHI, jedoch mit Werten zwischen  $\pm 8$ , zeigt. Diese Unterschiede sind wichtig für die in Kapitel 6 vorgenommene Gewichtung. Dabei dürften sich stärker differenzierende Metriken als effizienter herausstellen, um bei der Gewichtung signifikante Unterschiede zum gleichgewichteten MME zu erzeugen. Auffällig sind zudem die CHI-Ergebnisse. Diese Metrik besitzt als einzige deutlich andere Ergebnisse als die fünf übrigen 4FM. Grund hierfür ist die Formel (vgl. Kapitel 4.1), über die der Betrag der PHI-Korrelation berechnet wird. Der Wertebereich wird wie bei PIE dadurch auf  $[0-1]$  festgelegt. Anders als bei PIE bekommen sehr niedrige Übereinstimmungen nicht einen Wert um 0 und hohe Übereinstimmungen den Wert 1 zugeteilt. Stattdessen erhalten sowohl Simulationen mit sehr niedriger als auch solche mit sehr hoher Übereinstimmung Werte um 1. Diese Form der Klassifizierung ist vor dem Hintergrund der leistungsorientierten Erzeugung von Gewichten nicht geeignet. An dieser Stelle kann CHI bereits als ungeeignete Metrik zur Modellevaluation eingestuft werden. Im Weiteren wird sie daher nur noch zu Vergleichszwecken erwähnt. Die fünf übrigen 4FM zeigen dagegen soweit alle logische Ergebnisse. Auf die Ansätze zur Intensivierung ( $4FM_i$ ) wird an dieser Stelle nicht explizit eingegangen, da sie lediglich allen negativen Korrelationen ein Nullgewicht zuteilen und damit die Bedeutung der positiven Rohgewichte erhöhen. Dies wird erst im späteren Verlauf eine wichtige Rolle einnehmen.

## RM

In diesem und den beiden folgenden Abschnitten werden die Besonderheiten und Unterschiede der Evaluationsergebnisse der übrigen (nicht gezeigten) Metriken im Vergleich zu denen der 4FM bzw. Abbildung 5.1 im Speziellen herausgestellt. Dabei gilt zu berücksichtigen, dass zur Erstellung der Rohgewichte von RM und

ZM ein RMSE berechnet wurde. Hohe Werte sind damit beispielsweise als schwaches Rohgewicht bzw. niedrige Simulationsgüte anzusehen. Dies kann gerade im direkten Vergleich zu den 4FM- und FPA-Rohgewichten zunächst zu Verwirrungen führen. Zur Berechnung der finalen Gewichte werden die Rohgewichte nach Formel 4.7 (vgl. Kapitel 4) auf den Wertebereich  $]0;1]$  festgelegt, wobei die beste Simulation in jeder Situation den Wert 1 erhält. Bei der folgenden Analyse der RM- und ZM-Rohgewichte entsprechen niedrige Rohgewichte einer hohen Simulationsgüte bzw. einer geringen Abweichung von den Referenzdaten.

Wie bei den 4FM wurde für RM sowohl das klimatologische Mittel wie auch der Trend der Simulationen evaluiert. Bei den RM-Ergebnissen fällt, analog zu MMR bei PHI, eine deutliche Separierung zwischen den Werten des Trends und des Mittels auf (vgl. Abbildung 5.1). Dies gilt sowohl für die Temperatur als auch für den Niederschlag. Anders als bei den 4FM, wo dies in den Unterregionen schwächer ausgeprägt ist, zieht sich dieses Muster für RM durch alle Regionen. Dieses Phänomen ist allerdings naheliegend, da der Trend sowohl für Temperatur als auch für Niederschlag im Schnitt deutlich niedriger als das klimatologische Mittel ist. Da der RMSE die Abweichung detektiert, ist diese Trennung folgerichtig. Anders als bei den 4FM ist die Streuung für alle Situationen deutlich reduziert. Die Trendergebnisse sind hier sogar noch etwas homogener über die Simulationen verteilt als die des Mittels. Bei Letzterem fällt ähnlich zu den PHI-Rohgewichten des Mittels eine häufig auftretende Verkettung der Werte einzelner Simulationen auf. Dies gilt nicht für den Trend, wo alle Simulationen relativ gleichmäßig auf ähnlichem Niveau streuen. Insgesamt zeigt der Temperaturtrend sehr hohe Übereinstimmungen mit den Referenzdaten und Abweichungen, die im Schnitt unter  $1^{\circ}\text{C}$  für nahezu alle Regionen liegen. Beim Niederschlag sind die Ergebnisse des Trends stark von der Gesamtniederschlagsmenge der jeweiligen Region abhängig. So liegen die Werte in den humiden Regionen Nordatlantik und Balkan zwischen 100–200mm und für den Nahen Osten unterhalb von 50mm. Diese Zuordnung ist für das klimatologische Mittel nur eingeschränkt gültig. So schwanken die Rohgewichte des Nahen Ostens und Balkans zwischen 100–300mm. Auch bei der mittleren Temperatur ist keine direkte Abhängigkeit von der Region zu beobachten. Hier liegen alle Werte zwischen ca.  $1^{\circ}\text{C}$  und knapp unter  $5^{\circ}\text{C}$ .

Insgesamt ist im Vergleich zu den 4FM festzustellen, dass sich Trend und Mittel mit Blick auf die Differenzierbarkeit der Simulationsergebnisse gegensätzlich verhalten. Dies liegt an der offenen RM-Skala. So ermöglicht sie bei größeren Schwankungen bzw. Abweichungen mehr Spielraum als bei ähnlichem Verhalten. Die Bewertung bezüglich der allgemeinen Simulationsgüte ist für den Niederschlag nur relativ je Region vorzunehmen, da die jährliche Niederschlagsmenge entscheidend ist. Bei der Temperatur sind dagegen allgemeine Aussagen möglich, da die Abweichung des Trends mit meist  $<1^{\circ}\text{C}$  in allen Regionen gleichmäßig

gering ausfällt. Insgesamt sind auch die RM-Ergebnisse als Einschätzung der Modellqualität vielversprechend. Wie schon bei den 4FM wird die Simulationsgüte der Temperatur für Trend und Mittel größtenteils als hoch eingestuft. Die Simulationsgüte des Niederschlags muss dagegen regionsabhängig klassifiziert werden.

## ZM

Die ZM untersucht, anders als die übrigen Metriken, weder den Trend noch das Mittel des 50-jährigen Untersuchungsintervalls. Allerdings gilt, dass das Muster der ZM-Rohgewichte eher dem des Trends in Abbildung 5.1 ähnelt als dem Mittel. Es muss jedoch beachtet werden, dass die Rohgewichte der ZM auf dem RMSE über erklärte Varianzen beruhen und dies daher lediglich als optische Einordnung zu verstehen ist. Alle Ergebnisse liegen im Intervall  $]0; 0,24]$ . Da die Rohgewichte der Simulationen hier nicht weiter interpretiert werden können, werden nur die Unterschiede zwischen den Simulationen analysiert. Auch hier gilt wie bei RM, dass niedrige Werte einer geringen Abweichung und damit besseren Simulationsgüte entsprechen.

Generell besteht in allen Regionen eine relativ gleichmäßige Streuung der Rohgewichte. Beim Niederschlag liegen die meisten Werte zwischen 0,04–0,10. Nur für die Ägäis ist die Mehrheit der Werte oberhalb von 0,1. Bis auf je eine Ausnahme in Italien und dem Schwarzen Meer übersteigen die Rohgewichte 0,2 nicht. Hierbei handelt es sich in beiden Fällen um eine Simulation von GISS-E2-CC (SN: 115). Diese zeigt in den meisten Regionen einen auffällig hohen Wert und damit ein schwächeres Ergebnis im Vergleich zu den restlichen Simulationen. Insgesamt sind die Unterschiede zwischen minimalem und maximalem Rohgewicht innerhalb der MME eher gering. Zwischen den MME fallen die Rohgewichte von CORDEX insbesondere in Italien und dem Nahen Osten auf. Hier ist die Streuung deutlich stärker und hat höhere Abweichungen als die der anderen MME. Ansonsten liegen die Ergebnisse trotz des kürzeren Zeitraums auf vergleichbarem Niveau wie die von CMIP3 und CMIP5.

Bei der Temperatur ist die Streuung insgesamt für die GKM viel größer, während die RKM hier sehr ähnliche Rohgewichte aufzeigen. Nur für die Region Schwarzes Meer liegen die Ergebnisse der GKM unter 0,12. In den übrigen Regionen variieren die Werte stark zwischen minimal 0,02 für Nordafrika und maximal 0,22 für Italien. Auf Grundlage dieser ersten Eindrücke lassen die Rohgewichte der Temperatur stärkere Gewichtungseffekte erwarten als die des Niederschlags. Ein Rückschluss auf die allgemeine Simulationsgüte ist nur schwer möglich. Die erklärte Varianz, veranschaulicht im Periodogramm (vgl. Kapitel 4.3), nimmt die entscheidende Rolle ein. Da bei der Temperatur ein hoher Anteil der erklärten Varianz auf den langfristigen Trend des Klimawandels fällt, sind hier die größten

Abweichungen wahrscheinlich. Beim Niederschlag drückt keine der analysierten Perioden einen mit dem langfristigen Temperaturtrend vergleichbar hohen Anteil der erklärten Varianz aus. Die Unterschiede zu den Referenzdaten fallen daher beim Niederschlag deutlich geringer aus.

Die Einordnung der ZM-Simulationsbewertung im Vergleich zu den übrigen Metriken wird in Kapitel 5.4 dargestellt. Im nächsten Abschnitt werden abschließend die Rohgewichte der FPA-Metriken vorgestellt. Da diese nur für den Gesamtmittelmeerraum bestimmt wurden, wird ergänzend auch die Vorstellung der  $FPA_z$  angefügt.

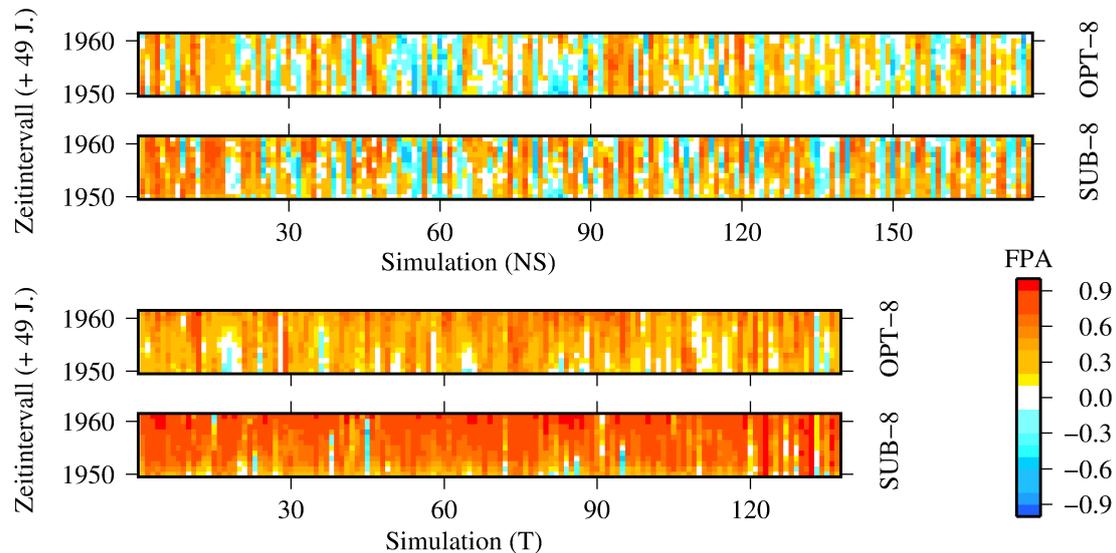
## FPA

Die Fingerprinting-Ansätze wurden aufgrund der Ausmaße der Unterregionen nur für das Trendmuster des Gesamtmittelmeerraums bestimmt. Der Hintergrund hierzu wurde in Abschnitt 4.4 erläutert. Die Rohgewichte des Niederschlags zeigen für den Untersuchungszeitraum von 1960–2009 ein ähnliches Erscheinungsbild wie die der PHI-Korrelation (vgl. Abbildung 5.1). Da beide den Wertebereich  $[-1;1]$  besitzen, ist die Interpretation vergleichbar: Hohe Übereinstimmungen werden durch Werte am oberen Ende des Intervalls angegeben. Über alle Simulationen ist eine hohe Streuung sowohl für SUB-8 als auch SUB-15 mit Werten zwischen  $\pm 0,8$  zu beobachten. Das Muster von OPT-8 und OPT-15 ist dem der PHI-Rohgewichte noch ähnlicher. Hier ist eine vergleichbar starke Streuung im Bereich  $\pm 0,6$  zu beobachten. Anders als bei PHI zeigen die FPA-Rohgewichte der Temperatur eine starke Verschiebung in den positiven Bereich. Für SUB-8 und SUB-15 liegen nahezu alle Werte oberhalb von 0,5. In der Folge sind die meisten Ergebnisse auf einem ähnlichem Niveau, was für eine vergleichbar hohe Simulationsgüte spricht. Bei den optimalen FPA sinkt diese Schwelle bei zugleich leicht ansteigender Streuung der einzelnen Simulationen etwas ab. Generell gilt, dass Unterschiede zwischen Ergebnissen der FPA mit verschiedenen EOF-Anzahlen auf dieser Betrachtungsebene nicht ersichtlich werden.

Für die FPA gilt, dass die Unterschiede zwischen den suboptimalen Rohgewichten höher ausfallen. Das Muster bleibt im Vergleich zu den optimalen Rohgewichten konstant. Wird die Analyse auf das Klimawandelsignal der optimalen FPA reduziert, so fallen die Ergebnisse der Simulationen mit den besten suboptimalen Rohgewichten etwas niedriger aus. Umgekehrt schneiden Modelle, die beim Gesamttrendmuster eine schwächere Leistung gezeigt haben, etwas besser ab. Um diese Beobachtungen noch weiter verfolgen zu können, werden die Ergebnisse der zusätzlichen Fingerprint-Metriken ( $FPA_z$ ) für den Mittelmeerraum betrachtet.

Anders als die  $4FM_i$  werden diese nicht allein durch die Bildung eines Subensembles basierend auf dem Rohgewicht erzeugt. Stattdessen werden die Eva-

luationsergebnisse aller möglichen Zeiträume aus Beobachtungs- und Simulationsdaten zur Erstellung der Rohgewichte herangezogen. Zur Veranschaulichung zeigt Abbildung 5.2 zunächst die SUB-8 und OPT-8 Rohgewichte über alle deckungsgleichen Zeiträume.

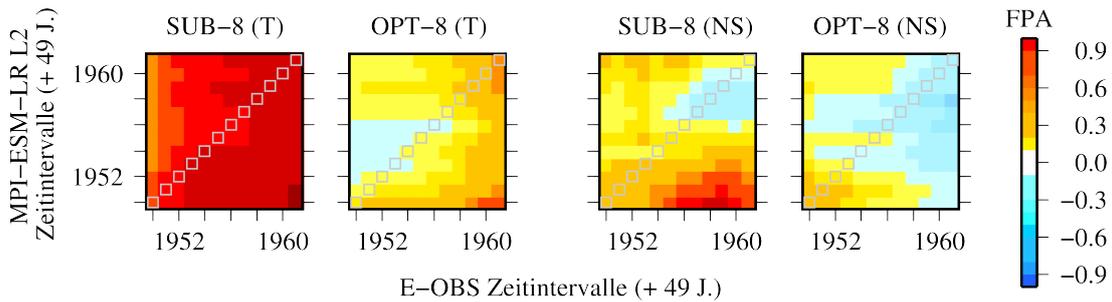


**Abbildung 5.2** Annuelle Evaluationsergebnisse der suboptimalen und optimalen FPA basierend auf 8 EOF. Angetragen sind die Rohgewichte aller Simulationen für Niederschlag (oben) und Temperatur (unten) im MMR mit übereinstimmenden 50-jährigen Zeitfenstern (Startjahre angetragen) von Referenz- und Modelldaten.

Für jede Simulation sind die Evaluationsergebnisse des Niederschlags und der Temperatur für beide Ansätze farblich dargestellt. Die zuvor diskutierten Rohgewichte des Hauptuntersuchungszeitraums 1960–2009 sind folglich entlang einer horizontalen Reihe auf Höhe des Zeitintervalls „1960“ dargestellt. Beim Niederschlag zeigt sich ein heterogenes Muster sowohl in vertikaler (zeitlicher) als auch horizontaler (SN) Richtung. Die Unterschiede bei der Temperatur sind hier insbesondere zwischen den Simulationen gering. Dieses Bild wurde bereits zuvor für den Zeitraum 1960–2009 beschrieben, setzt sich allerdings auch für die übrigen Zeitintervalle fort. Die Intensität der Muster (vertikal) der optimalen Temperaturrohgewichte ändert sich allerdings über den kurzen Zeitraum von 12 Jahren teilweise deutlich. An dieser Stelle ist jedoch nicht abzulesen, ob dieser Beobachtung eine Änderung in den Trendmustern der Simulationen oder der Beobachtungsdaten zu Grunde liegt. Beim Niederschlag tritt dieses Phänomen nicht auf. Bei der Temperatur zeigen die meisten Simulationen eine hohe Übereinstimmung über einen Großteil der Trends mit einem Maximum zum Ende der Zeitintervalle. Auch hier ist eine vergleichbare Tendenz wie beim Niederschlag nicht festzustellen. Aller-

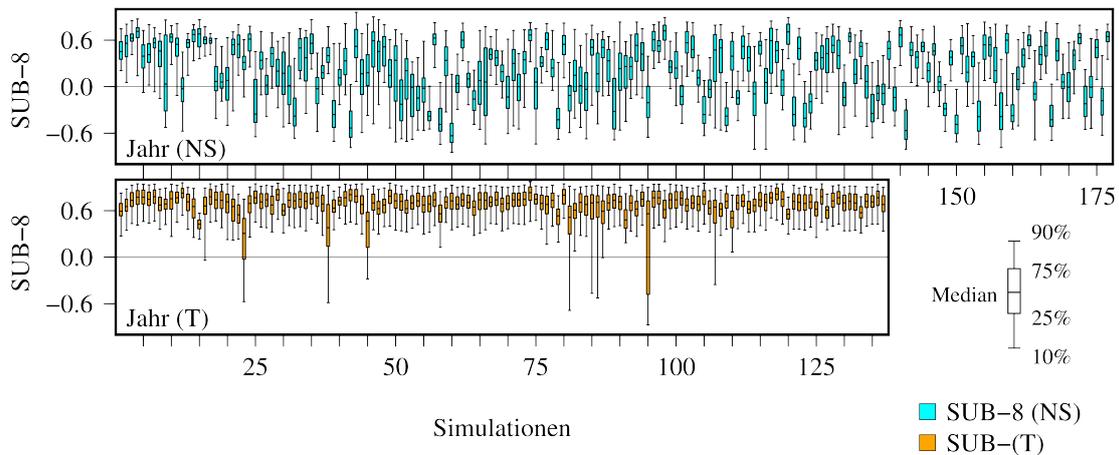
dings fällt auf, dass sich die Rohgewichte der meisten Niederschlagssimulationen nur wenig über die 12 Intervalle ändern. Dennoch legen die Unterschiede zwischen den Zeitabschnitten nahe, dass es bei den  $FPA_z$  zu einer deutlichen Verschiebung der Modellbewertung insbesondere beim Niederschlag kommen kann.

Zur Erstellung der  $FPA_z$ -Gewichte werden neben den 12 übereinstimmenden Zeitintervallen auch die 132 Intervalle mit variablen Zeiträumen berücksichtigt. Diese 144 Rohgewichte wird in Abbildung 5.3 an dem Beispiel MPI-ESM-LR Lauf 2 veranschaulicht.



**Abbildung 5.3** Evaluationsergebnis der suboptimalen (unten) und optimalen Fingerprint-Ansätze (oben) für MPI-ESM-LR L2, SN: 152 (NS); 112 (T). Dargestellt sind die annuellen Rohgewichte der Temperatur im MMR mit variablen Zeitfenstern von Referenz- und Modelldaten.

Abbildung 5.3 zeigt die Grundlage des  $FPA_z$  am Beispiel MPI-ESM-LR L2 (CMIP5) für Temperatur. Die Diagramme selbst sind in  $12 \times 12$  Felder aufgeteilt. Die Ordinate zeigt die Startjahre der 50-jährigen Trendmuster von MPI-ESM-LR L2 und die Abszisse das E-OBS-Äquivalent. Folglich entsprechen die Werte auf der Diagonale der annuellen Auswertung in Abbildung 5.2 mit übereinstimmenden zeitlichen Intervallen. Auffällig ist hier, dass das generelle Muster, das sich aus der Diagonale ablesen lässt, gut mit dem der restlichen Felder übereinstimmt. So kommt es eher zu einer stufenweisen Änderung, sobald eine Veränderung innerhalb der Referenzdaten auftritt. Die Modelltrends sind im Vergleich zu den Beobachtungsdaten stabiler, wie aus der geringeren Variabilität entlang der Ordinate deutlich wird. Für MPI-ESM-LR L2 ist erst ab dem Zeitintervall 1955–2004 ein deutlicher Einschnitt, der auf eine Trendänderung in den Modelldaten zurückgeführt werden kann, für Niederschlag und Temperatur zu erkennen. Auch wenn diese Tendenz in den meisten Fällen relativ konsistent ist, lässt sich das Erscheinungsbild aller 144 FPA Ergebnisse der übrigen Simulationen nicht verallgemeinern. Zur Veranschaulichung hierzu zeigt Abbildung 5.4 die Verteilungen der 144 annuellen FPA-Rohgewichte für SUB-8. Diese werden anschließend verwendet, um über den Median der jeweils 144 verfügbaren Werte einer Simulation das  $FPA_z$ -Rohgewicht, im weiteren  $SUB_z$  bzw.  $OPT_z$ , zu bestimmen.



**Abbildung 5.4** Grundlage der zusätzlichen Fingerprint-Ansätze. Boxplots über die zeitlich dynamischen Rohgewichte der Temperatur im MMR. Angetragen sind die Verteilungen der annualen SUB-8 Rohgewichte für Niederschlag (oben) und Temperatur (unten) zur Darstellung des Medians ( $SUB_z-8$ , bzw.  $OPT_z-8$ ).

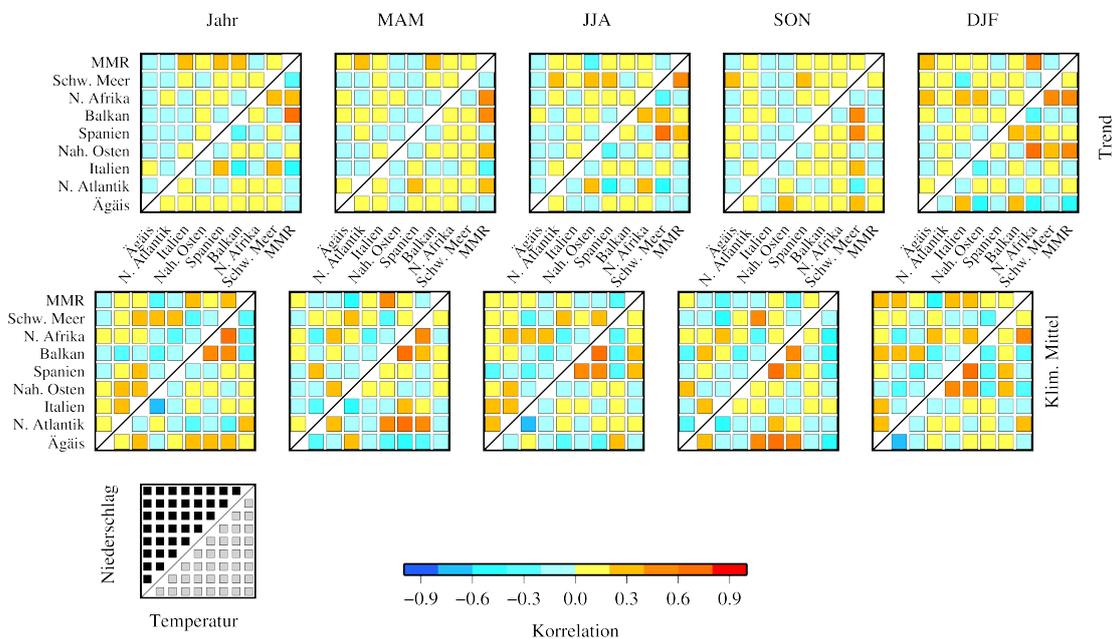
Ein Großteil der SUB-8-Verteilungen der Temperatur in Abbildung 5.4 liegt im Bereich zwischen 0,6–1,0, wobei vereinzelt Ausreißer auch stärker streuen. Für OPT-8 sind die Werte insgesamt etwas niedriger. Die Ergebnisse für den Niederschlag sind dagegen deutlich heterogener und decken einen Großteil des positiven und negativen Wertebereichs ab. Auch die Unterschiede zwischen  $SUB_z-8$  und  $OPT_z-8$  fallen beim Niederschlag weniger ins Gewicht. Allerdings lassen hier die Verteilungen und daraus resultierenden Mediane Unterschiede bezüglich der finalen Gewichte im Vergleich zu den FPA-Ergebnissen für 1960–2009 (vgl. Abbildung 5.2) erahnen. Der Vergleich von FPA-Rohgewichten und  $FPA_z$ -Medianen zeigt für die unterschiedlichen Situationen und MME im Schnitt eine Pearson- und auch Spearman-Korrelation von ca. 0,7. Dieser Wert gilt sowohl für Niederschlag als auch für Temperatur. Dabei ist die Korrelation der Niederschlagsergebnisse im Schnitt bei 0,67–0,80 und die der Temperatur im Bereich 0,59–0,84. Der überwiegende Teil der Ranglisten bleibt stabil. Das deutet darauf hin, dass auch das Intervall 1960–2009 als zulässiges Beispiel zur Bestimmung der Simulationsgüte für die evaluierten Situationen gewählt werden kann.

Neben der Verwendung aller möglichen Trends wurde für die  $FPA_z$  auch ein Schwellenwert eingefügt, um basierend auf der Verteilung aller Rohgewichte eine Vorauswahl bzw. ein Sub-Ensemble zu erstellen. Für die Temperatur zeigen die finalen Ergebnisse der  $FPA_z$  hier ein sehr ähnliches Bild im Vergleich zu den FPA. Hier kommt zum Tragen, dass die Simulationen alle relativ hohe Rohgewichte zugeteilt bekommen und dies auch über verschiedene Zeiträume konstant

bleibt. Anders sieht es beim Niederschlag aus. Durch die hohe Streuung und die vielen Rohgewichte im negativen Bereich bekommen viele Simulationen eine Nullgewicht zugeteilt. Die übrigen Mediane streuen dennoch relativ stark im Bereich zwischen 0,2–0,7. Aufgrund der geringeren Anzahl an Simulationen, die in den verschiedenen MME erhalten bleiben, kann davon ausgegangen werden, dass die einzelnen Gewichte so eine deutlich größere Bedeutung erhalten als bei den FPA.

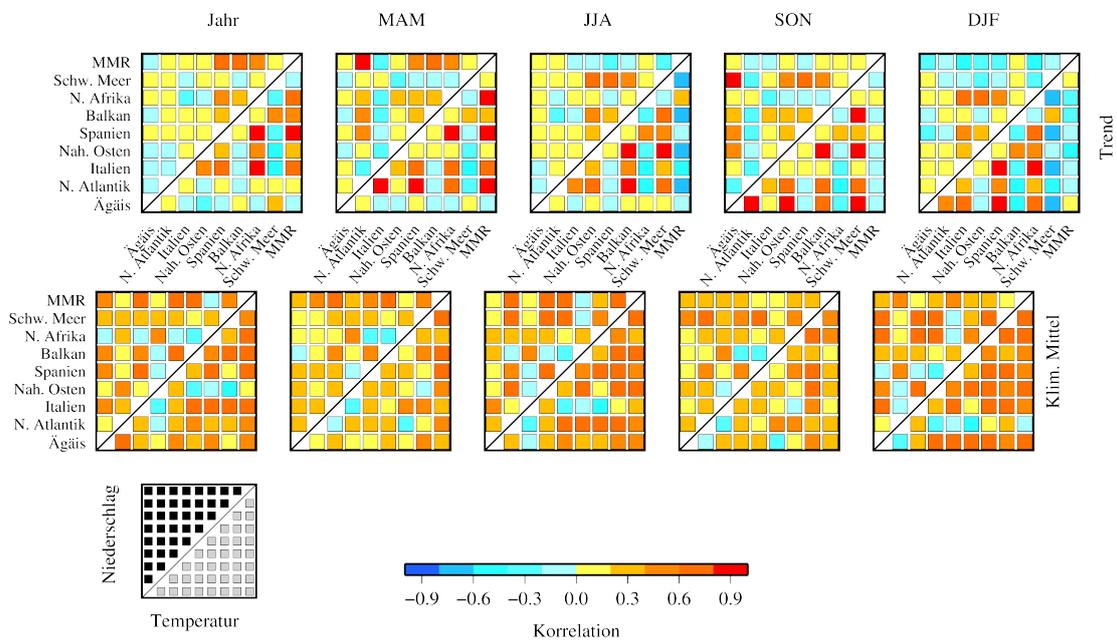
### 5.1.2 Regionale Korrelation der Bewertungen

Der Vergleich der Evaluationen hat gezeigt, dass die Simulation des 50-jährigen Trends die größere Herausforderung für die Klimamodelle ist als die des Mittels. Obwohl es sich mit MMR um eine relativ kleine Region bzw. Unterregionen handelt, lassen die Trendergebnisse meist eine deutliche Abstufung zwischen einzelnen Simulationen zu. Darüber hinaus gilt, dass die Simulationen Unterschiede hinsichtlich der Rohgewichte in den einzelnen Regionen zeigen. So konnte ein klares Muster mit Modellen, die überall überzeugen, nicht ausgemacht werden. Um hierzu genauere Aussagen treffen zu können, wird im Folgenden der Zusammenhang zwischen den metrikbasierten Ranglisten der verschiedenen Regionen veranschaulicht.



**Abbildung 5.5** Spearman-Korrelation der PHI-Ranglisten aller Situationen im Mittelmeerraum. Angetragen sind die saisonalen Ergebnisse des Trends (oben) und Mittels (unten) für Niederschlag und Temperatur.

Abbildung 5.5 zeigt die Spearman-Korrelationen zwischen den PHI-Modellranglisten jedes Untersuchungsgebietes für Temperatur und Niederschlag. Es zeigt sich für alle Situationen ein heterogenes Bild. Während ein Großteil der regionalen Trendkorrelationen im Bereich  $\pm 0,4$  liegt, nehmen die Ergebnisse des 50-jährigen Mittels vereinzelt Werte außerhalb dieses Intervalls an. Ein einheitliches Muster ist jedoch weder innerhalb einer Jahreszeit noch für bestimmte Regionen zu erkennen. Regionen, die direkt nebeneinander liegen, tendieren vereinzelt dazu, schwache Zusammenhänge zu zeigen. Dies lässt sich allerdings nicht verallgemeinern und ist stark von der jeweiligen Situation abhängig. Eine generelle Aussage auf einer Region oder Jahreszeit basierend ist daher nicht zulässig. Auch die Unterscheidung von Niederschlag und Temperatur liefert hier keine anderen Ergebnisse. Lediglich die positiven Extreme der Temperaturkorrelationen fallen etwas intensiver aus als beim Niederschlag. Ansonsten sind die Korrelationen stark situationsabhängig und erlauben keine Verallgemeinerungen. Auch die Korrelationen zwischen den Ranglisten verschiedener Regionen basierend auf den übrigen 4FM verhalten sich analog zu PHI. So gilt, dass für die 4FM kein entscheidender Zusammenhang zwischen den Rohgewichtsranfolgen der Simulationen unterschiedlicher Regionen besteht. Für eine zuverlässige Bewertung der Simulationen muss daher jede Situation für sich untersucht und entsprechend verwendet werden, um das richtige Modell stark bzw. schwach zu gewichten.



**Abbildung 5.6** Spearman-Korrelation der RM-Ranglisten aller Situationen im Mittelmeerraum. Angetragen sind die saisonalen Ergebnisse des Trends (oben) und Mittels (unten) für Niederschlag und Temperatur.

Analog zu Abbildung 5.5 zeigt Abbildung 5.6 die Spearman-Korrelation zwischen den RM-Rohgewichten jedes Untersuchungsgebiets. Hier werden Unterschiede zu den Ergebnissen der 4-Felder-Metriken deutlich. Insbesondere für das Mittel werden Korrelationen zwischen 0,4–0,8 sowohl für Niederschlag als auch für Temperatur berechnet. Negative Korrelationen sind dagegen nur vereinzelt und mit schwachen Beträgen festzustellen. Für das klimatologische Mittel kann gefolgert werden, dass die Simulationen, die in einer Region besonders hohe bzw. niedrige Rohgewichte zugeteilt bekommen haben, auch in anderen Regionen entsprechend bessere bzw. schlechtere Resultate gezeigt haben. Damit ist basierend auf RM für das Mittel eine regionale Abhängigkeit der Simulationsqualität in den Regionen des Mittelmeerraums festzustellen. Etwas anders verhält sich die regionale Korrelation der Ranglisten beim 50-jährigen Trend. Hier ist das Muster für Niederschlag dem von PHI ähnlich. Es können keine eindeutigen Tendenzen ausgemacht werden und die berechneten Zusammenhänge sind größtenteils eher schwach. Punktuell sind allerdings auch sehr hohe Werte von über 0,9 zu erkennen wie beispielsweise in MAM zwischen MMR und Nordatlantik oder zwischen Schwarzes Meer und Ägäis für SON. Auch hier ist auffällig, dass die positiven Korrelationen eher höhere Beträge annehmen als die negativen Korrelationen. Noch deutlicher fällt dies bei der Temperatur auf. Auch hier sind die Korrelationen positiv und teilweise oberhalb von 0,4. Allerdings sind insbesondere für JJA und DJF die Korrelationen zwischen MMR und den meisten übrigen Regionen negativ. Gerade JJA ist in Bezug auf MMR auffällig. Hier sind beinahe ausschließlich deutlich negative Korrelationen sowohl für den Temperatur- als auch den Niederschlagstrend zu beobachten.

Für die ZM (nicht gezeigt) kann das Ergebnis nicht in Trend und Mittel unterschieden werden, da die Evaluationsbasis eine andere ist. Trotzdem ähneln die Korrelationen sehr stark denen der Trendkorrelationen von RM (siehe Abbildung 5.6). Die regionalen Korrelationen der Niederschlagsrangfolgen sind deutlich schwächer als die der Temperaturrengfolgen. Letztere liegen in allen Jahreszeiten teilweise oberhalb von 0,9. Auch insgesamt überwiegen positive Korrelationen zwischen 0,2–0,8. Die niedrigsten Werte liegen dagegen beispielsweise im Bereich von  $-0,3$  bis maximal  $-0,44$  für Italien und Schwarzes Meer in SON. Insbesondere für DJF und JJA sind in den meisten Regionen mittlere bis hohe positive Korrelationen zu erkennen. Auch die regionalen Übereinstimmungen für die Ranglisten der Niederschlagsbewertung sind insgesamt eher positiv, allerdings auf einem sehr niedrigen Niveau. Nur vier Werte liegen außerhalb des Intervalls  $[-0,2; 0,4]$ . Die Jahreszeit scheint hier Einfluss auf die Stärke und Art des Zusammenhangs zu haben. Das gilt beispielsweise für MMR. Hier zeigen die Rangfolgen der Temperatur in 7 von 8 Kombinationen einen negativen Zusammenhang in DJF und damit wiederum eine Ähnlichkeit zu RM. Für das Gesamtjahr, JJA und SON dagegen

ist die Korrelation in 20 von 24 Fällen positiv und teilweise über 0,9. Damit zeigt nur ZM einen jahreszeitlichen Einfluss auf die Ergebnisse. 4FM und RM weisen jeweils für sich ein stimmiges Ergebnis auf, das sich über die Jahreszeiten zwar ändert, jedoch nicht nach einem eindeutigen Schema. Daran anschließend werden im folgenden Abschnitt die Rohgewichte der unterschiedlichen Jahreszeiten untersucht.

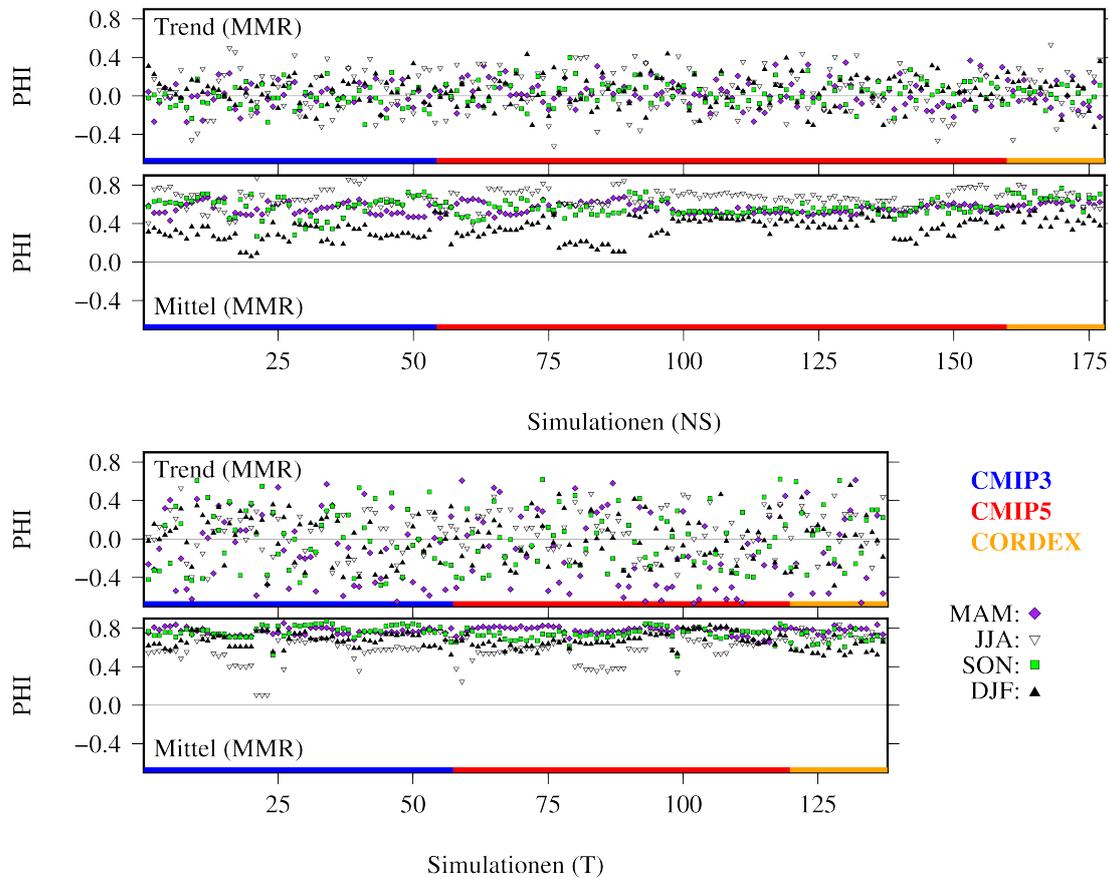
## 5.2 Saisonaler Einfluss auf die Modellbewertung

Das vorangegangene Kapitel hat gezeigt, dass die Simulationen in der Lage sind, sehr hohe Übereinstimmung mit den Beobachtungsdaten, abhängig von Region und Jahreszeit, zu erreichen. Die Korrelation der Ranglisten zwischen den Regionen schwankt zwischen den Metriken und Variablen. In diesem Abschnitt wird untersucht, wie sich die unterschiedlichen Jahreszeiten auf die Bewertung der Simulationsgüte auswirken. Anschließend werden die saisonalen Evaluierungsmuster verglichen.

### 5.2.1 Vergleich der saisonalen Simulationsgüte

Abbildung 5.7 zeigt die PHI-Rohgewichte für alle Simulationen exemplarisch für MMR. Bei den Rohgewichten des 50-jährigen Trends fällt auf, dass keine Simulation in allen Jahreszeiten die höchsten und keine die niedrigsten Koeffizienten vereint. Stattdessen gibt es für jede Simulation eine gewisse Streuung, die teilweise stärker oder schwächer ausfällt. Gemittelt über alle Jahreszeiten wäre eine Rangfolge der Simulationen zwar möglich, allerdings kommt es in den wenigsten Fällen dazu, dass alle vier Jahreszeiten auf einem gleichmäßig hohen Niveau liegen. Dies gilt sowohl für Niederschlag als auch für Temperatur. Des Weiteren fällt für beide Variablen auf, dass sich die Rohgewichte des klimatologischen Mittels deutlich homogener verhalten und im oberen Drittel der Skala befinden. Die des Trends streuen dagegen relativ gleichmäßig um 0. Diese Tendenz setzt sich aus den annualen Ergebnissen (vgl. Abbildung 5.1) für die einzelnen Jahreszeiten fort.

Beim klimatologischen Mittel ist im Gegensatz zum Trend eine deutliche Abstufung der Jahreszeiten für jede Simulation, abhängig von der untersuchten Variablen, zu erkennen. Im ariden JJA sind die PHI-Korrelationen über alle Simulationen am höchsten, während die Werte in der Hauptregenzeit DJF überwiegend niedrig ausfallen. Gegensätzlich verhält es sich bei der Temperatur. Die kleinsten Rohgewichte werden überwiegend in JJA berechnet, während hier die Übergangsjahreszeiten die höchsten Werte zeigen. Dieses Bild des MMR setzt sich für die



**Abbildung 5.7** Saisonale Rohgewichte der Modellevaluation am Beispiel PHI (4FM) für Niederschlag und Temperatur. Angetragen sind die Ergebnisse des klimatologischen Mittels und des Trends des Zeitraum 1960–2009 für den Gesamtmittelmeerraum.

übrigen Untersuchungsgebiete fort. Dabei ändert sich das Korrelationsniveau analog zu den in Abschnitt 5.1.1 diskutierten regionalen Unterschieden der Simulationsgüte. Mit diesem Hinweis sind die in Abbildung 5.7 gezeigten Rohgewichte für die übrigen 4FM exemplarisch. Auch die nicht dargestellten Ergebnisse von ZM, RM und FPA besitzen keine Simulationen, die über alle Jahreszeiten die höchsten Rohgewichte besitzen. Allerdings sind Muster zu erkennen, die im Folgenden herausgearbeitet werden.

### 5.2.2 Saisonale Muster der Modellbewertung

In diesem Abschnitt werden die saisonalen Muster analysiert, die durch die Modellevaluation der verschiedenen Metriken zu beobachten sind. Die in Abbildung 5.8 dargestellten Boxplots zeigen für PHI (4FM), RM, ZM und SUB-/OPT-8



als Zentrum, wobei die Stärke der Streuung zwischen den Jahreszeiten etwas variiert. Dagegen gilt beim Mittel, dass sich MAM, JJA und SON beim Niederschlag von DJF abheben. Bei der Temperatur werden dagegen für JJA leicht niedrigere Werte berechnet. Bei beiden Variablen handelt es sich jeweils um die Jahreszeit, in der betragsmäßig die höchsten absoluten Werte (Niederschlagsmaximum in DJF und Temperaturmaximum in JJA) in der Region angenommen werden. Es kann nicht ausgeschlossen werden, dass dieses Ergebnis auf die Funktionsweise der Evaluationsmetrik zurückzuführen ist. Allerdings könnten die Simulationen hier auch eine Unsicherheit zeigen, die von den 4FM registriert wird.

Beim Trend bleiben die Mediane zwar um 0 zentriert, der Schwankungsbereich ist jedoch beim Niederschlag in JJA, der niederschlagsärmsten Jahreszeit, und bei der Temperatur in den Übergangsjahreszeiten MAM und SON am größten. MAM zeigt darüber hinaus die größte Streuung und den niedrigsten Median aller Trends. Diese Ergebnisse können jedoch nicht verallgemeinert werden, da sich das Muster in anderen Regionen teilweise ändert. Während die Unterscheidung von Trend und Mittel sowie die saisonalen Besonderheiten bei Letzterem in allen Regionen stabil sind, ist der Schwankungsbereich des Trends stark vom gewählten Untersuchungsgebiet abhängig. Die meisten Regionen zeigen zwar grundsätzlich ein ähnliches Bild wie MMR, der Betrag der Streuung und die Jahreszeit mit höchster Varianz sind aber nicht einheitlich.

Die Ergebnisse der FPA zeigen sich sehr stabil über alle Jahreszeiten. Vergleichbar zu PHI sind die Trendwerte des Niederschlags relativ stark um 0 gestreut. Die Streuung der optimalen FPA sind allerdings in jeder Situation geringer als die der suboptimalen FPA. Die größte Varianz zeigt sich in DJF. Für die Temperatur unterscheidet sich das Ergebnis von SUB-8 und OPT-8 etwas. So liegen die SUB-8-Werte deutlich abgegrenzt über den OPT-8-Ergebnissen. Darüber hinaus ist eine ansteigende Ordnung zu erkennen mit dem Minimum in DJF und Maximum in JJA. Ein vergleichbares Bild gilt für alle FPA. Im Gegensatz zu den 4FM kann hier gefolgert werden, dass die FPA auf dem Trendmuster basierend eine Differenzierung der Simulationsgüte unterschiedlicher Jahreszeiten messen. Während beim Niederschlag analog zu PHI alle Verteilungen um 0 liegen, zeigen die FPA für Temperatur im Schnitt höhere Rohgewichte für MAM und JJA als für SON und DJF.

Für RM sind die saisonalen Evaluationsmuster ebenfalls auffällig. Wie schon in Kapitel 5.1.1 angesprochen, ist das Niveau der Gewichte bei dieser Metrik direkt von der ausgewerteten Größe bzw. dem vorhandenen Betrag abhängig. Allerdings zeigt sich bei der Varianz beispielsweise des JJA-Niederschlagsmittels, dass hier der größte Schwankungsbereich besteht. Dagegen liegt das Varianzmaximum des Trends erwartungsgemäß in der Jahreszeit des Niederschlagsmaximums DJF. Ein ähnliches Bild zeigt sich bei der Temperatur. Hier ist die größte Varianz des

Mittels in DJF, der Saison mit dem Temperaturminimum, zu beobachten, wohingegen der Trend die maximale Streuung in JJA anzeigt. Diese Muster sind über alle Regionen des Mittelmeerraums gültig, wenngleich der Wertebereich regional variiert.

Bei ZM ähnelt das Ergebnis wie beim Vergleich der regionalen Korrelationen (vgl. Abschnitt 5.1.2) dem der RM-Trendergebnisse. Beim Niederschlag ist die Varianz für alle Jahreszeiten ungefähr gleich stark ausgeprägt. Die DJF-Werte liegen jedoch in allen Regionen etwas höher. Dies entspricht hier einer geringeren Simulationsgüte. MAM, JJA und SON sind dagegen immer auf einem Level, der jedoch regional leicht variiert. Bei der Temperatur sind in JJA höhere Werte als in den übrigen Jahreszeiten zu beobachten. Anders als beim Niederschlag ist hier die Varianz in allen Regionen größer.

Die saisonalen Muster der vier Metrikblöcke zeigen vereinzelt Übereinstimmungen. Allerdings fällt auf, dass sich die Metriken in keinem Fall für beide Variablen und gegebenenfalls untersuchten Größen einig sind. Saisonale Muster scheinen vorhanden zu sein, allerdings unterscheiden sie sich abhängig von der Metrik bzw. Situation. Dies könnte darauf hindeuten, dass sich entweder nicht alle Metriken gleich gut zur Evaluation eignen oder sie zu unterschiedliche Aspekte untersuchen, um ein einheitliches Bild zu liefern. Diese Fragestellung wird in Kapitel 8 aufgegriffen, indem ein Ansatz zur Metrikevaluierung vorgestellt wird. Im folgenden Abschnitt wird die Simulationsebene verlassen und die Analyse bezüglich der drei unterschiedlichen MMEs CMIP3, CMIP5 und CORDEX angestellt.

## 5.3 Analyse der Modellgenerationen

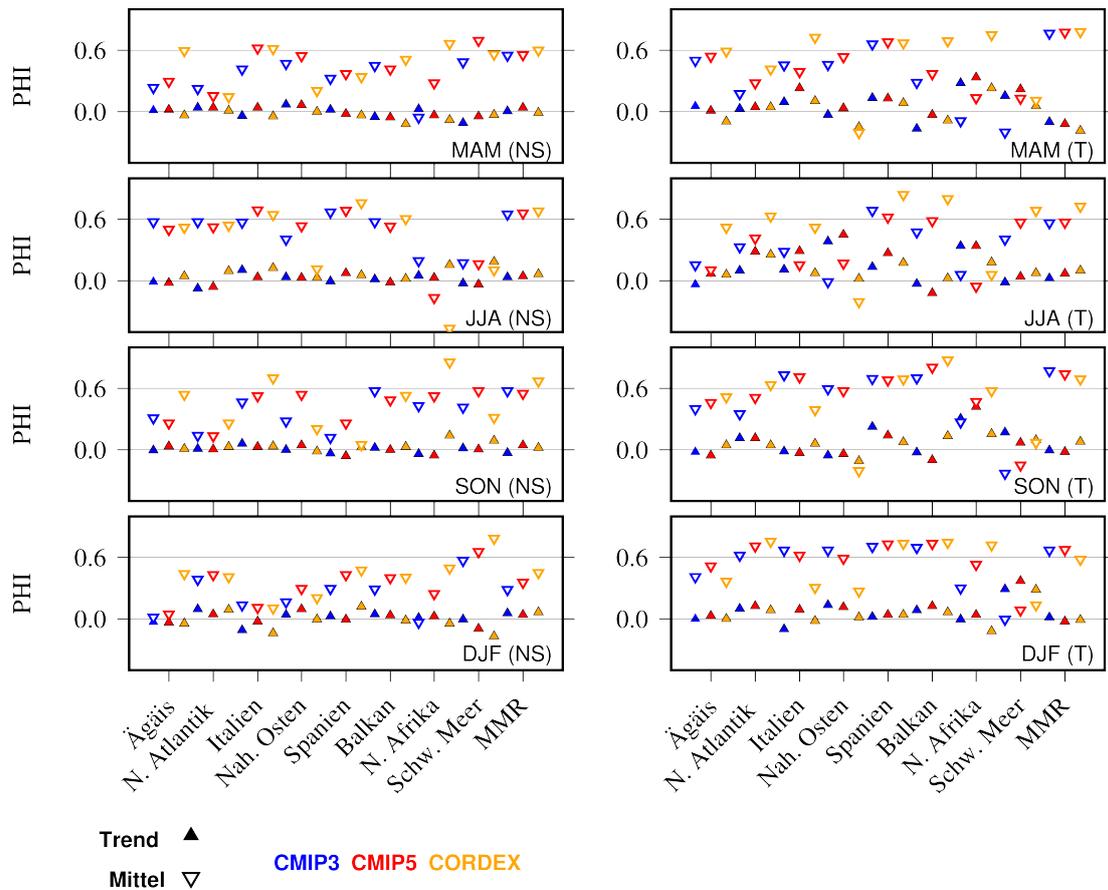
Die vorangegangenen Kapitel haben gezeigt, dass es basierend auf den entwickelten Metriken Unterschiede in der Simulationsgüte gibt. Diese haben jedoch weder regional noch saisonal ein klares Muster ergeben. Dieses Kapitel widmet sich der Auswertung der MME-Aggregationsebene. Hierzu wird untersucht, ob es qualitative Unterschiede zwischen CMIP3, CMIP5 und CORDEX gibt und wenn ja, in welcher Form sich diese äußern. Wie bereits in Kapitel 1.3 angesprochen, gibt es teilweise widersprüchliche Ansichten hinsichtlich der Entwicklung der Simulationsqualität zwischen den verschiedenen Modellgenerationen. Der letzte Sachstandsbericht geht jedoch von einer Verbesserung neuerer Modelle aus (FLATO ET AL., 2013). Dies deckt sich auch mit dem Ergebnis der Mehrheit aktueller Studien und dient daher als Vergleichsmaß (u.a. CHEN & FRAUENFELD, 2014; GROSE ET AL., 2014; SHASHIKANTH ET AL., 2014).

### 5.3.1 Unterschiede der Simulationsgüte

Zum Vergleich der Unterschiede zwischen den MME wurde jeweils das mittlere Rohgewicht jeder Metrik bestimmt. Bei 4FM und FPA kann über den vordefinierten Wertebereich neben dem MME-Vergleich auch eine Aussage zur generellen Simulationsgüte getroffen werden. Bei RM und ZM sind die Wertebereiche zwar offen, der Vergleich zwischen den MME ist jedoch dennoch möglich, da hier die relative Auswertung der Rohgewichte auf unterschiedliche Modellgüte schließen lässt. Abbildung 5.9 zeigt das mittlere Rohgewicht von Niederschlag und Temperatur für CMIP3, CMIP5 und CORDEX für MMR sowie alle acht Unterregionen des Mittelmeerraums. Exemplarisch sind die mittleren PHI-Rohgewichte dargestellt. Es gilt zu beachten, dass die Simulationen, die dem CORDEX-Ergebnis zu Grunde liegen, über einen 40-jährigen Zeitraum untersucht wurden. Dies muss berücksichtigt und die Ergebnisse daher mit Vorsicht interpretiert werden.

Der in den vorherigen Abschnitten gewonnene Eindruck setzt sich fort: Die Simulationsgüte des klimatologischen Mittels ist, auch aggregiert für das jeweilige MME, in fast allen Situationen höher als die des Trends. Allerdings fallen zwischen CMIP3 und CMIP5 leichte Abstufungen auf. Bei Niederschlag und Temperatur ist das mittlere Rohgewicht von CMIP5 in 66,7 % bzw. 69,4 % der Fälle höher als das von CMIP3. Damit können die PHI-Ergebnisse, repräsentativ für alle 4FM, bereits auf dieser Basis als vergleichbar mit denen anderer Studien ansehen werden, die dem neueren MME eine Leistungssteigerung attestieren (vgl. Abschnitt 1.3). Zusätzlich zeigt CORDEX bezogen auf das klimatologische Mittel in 18 der 36 Situationen des Niederschlags einen höheren Wert als CMIP3 und CMIP5. Auffällig ist darüber hinaus, dass CORDEX bei der Temperatur im Nahen Osten häufig den niedrigsten Wert aufweist. In allen anderen Regionen liegt der Wert zumeist auf Höhe von CMIP5 bzw. ist in 63,4 % der Fälle am höchsten. Aufgrund der Abdeckung von CORDEX geht in die Auswertung des Nahen Ostens ein Datenpunkt weniger in die Untersuchung ein. In weiteren Analysen hat sich dieser Umstand jedoch nicht als Ursache für die schlechtere Bewertung herausgestellt, sondern Ungenauigkeiten und Abweichungen in den übrigen 15 Datenpunkten.

Insgesamt niedriger liegen die Werte des 50-jährigen Trends. Beim Niederschlag werden sowohl Maximum (0,19, JJA) als auch Minimum (-0,17, DJF) für das Schwarze Meer durch CORDEX berechnet. Die Unterschiede zwischen CMIP3 und CMIP5 innerhalb dieses Intervalls sind darüber hinaus, verglichen mit dem klimatologischen Mittel, weniger stark ausgeprägt. Ein eindeutiges Muster von Verbesserungen oder Verschlechterungen kann nicht ausgemacht werden, da die Unterschiede eher zufällig erscheinen. Anders verhält es sich beim Temperaturtrend. Hier liegen CMIP3 und CMIP5 größtenteils im positiven Bereich und er-



**Abbildung 5.9** Vergleich der mittleren PHI-Rohgewichte von CMIP3, CMIP5 und CORDEX für alle Jahreszeiten und Untersuchungsregionen. Angetragen sind die Werte des Trends und des klimatologischen Mittels von Niederschlag (links) und Temperatur (rechts).

reichen Werte von maximal 0,45 für Naher Osten (CMIP5, JJA) und minimal  $-0,18$  im MMR (CORDEX, MAM). Insgesamt kommt es bei der Temperatur in 61,1 % der Fälle zu einer Verbesserung des Ergebnisses von CMIP5 im Vergleich zu CMIP3. Die von verschiedenen Studien nachgewiesene Verbesserung der aktuellen Modellgeneration von 2013 bzw. von aktuellen Regionalmodellen (PAETH & MANNIG, 2013; PAXIAN ET AL., 2016; NIKULIN ET AL., 2012) wird auch durch die 4FM sowohl für Niederschlag als auch Temperatur bestätigt. Dies gilt sowohl für das klimatologische Mittel beider Variablen als auch für den Temperaturtrend. Die Unstimmigkeiten bezogen auf den Niederschlagstrend lassen sich auch mit den vorherigen Studien in Einklang bringen (MIAO ET AL., 2012; KUMAR ET AL., 2013; KNUTTI & SEDLÁČEK, 2012; DAI, 2006). Auf dieser Grundlage können die 4FM bezogen auf die MME-Simulationsgüte als geeignete Evaluationswerkzeuge

angesehen werden.

Bei RM (nicht gezeigt) fallen die Unterschiede zwischen den einzelnen Regionen für Niederschlag relativ gering aus. Sowohl Trend als auch Mittel liegen, bis auf wenige Ausnahmen, unterhalb von 100mm. Die Ausnahmen werden insbesondere für die humide Region Nordatlantik bzw. generell für DJF berechnet. Die gemittelten Rohgewichte des Niederschlags basierend auf dem klimatologischen Mittel zeigen für CORDEX in ca. der Hälfte der Fälle das beste Ergebnis unter den drei MME. Die Ergebnisse in der anderen Hälfte sind teilweise auch schwächer als von CMIP3 und CMIP5. Zwischen diesen beiden MME ist die Tendenz allerdings relativ eindeutig. Basierend auf dem klimatologischen Mittel hat CMIP3 in 22 von 36 Fällen ein besseres Ergebnis als CMIP5. Hier besteht ein deutlicher Unterschied zu PHI. Ein Grund hierfür liegt bei den Simulationen des GISS-Modells, deren Mittelwert über das gesamte MME deutlich höhere Abweichungen besitzt als alle CMIP3-Simulationen. Dasselbe Phänomen ist beim Vergleich der Trendergebnisse zu erkennen. Auch hier ist CMIP3 im Mittel in nahezu allen Situationen besser als CMIP5. Allerdings sind hier die Unterschiede teilweise erst in der zweiten Nachkommastelle zu finden.

Bei der Temperatur gilt bezogen auf das klimatologische Mittel ein ähnliches Bild. Nur in 50 % der Situationen besitzt CMIP5 das stärkere Ergebnis. Dabei liegt der Unterschied der Abweichung zwischen den beiden MME teilweise bei bis zu 1°C. Für CORDEX gilt hier, dass es in 29 von 36 Situationen zu einer Verbesserung der Ergebnisse in Bezug zu beiden MMEs kommt. Insbesondere in JJA fallen deutliche Verbesserungen sowohl von CMIP5 gegenüber CMIP3 als auch von CORDEX im Vergleich zu CMIP5 auf. Die Auswertung des 50-jährigen Temperaturtrends ergibt ein weniger eindeutiges Ergebnis. CMIP3 und CMIP5 liegen in den meisten Situationen gleichauf im Bereich zwischen 0,3–0,9°C. Im Detail ist das CMIP5-Ergebnis allerdings in 61,1 % der Fälle stärker als das von CMIP3. Die CORDEX-Trends sind hier wiederum in fast allen Situationen auf dem Niveau von CMIP3 oder schlechter. Die RM-Werte haben für Niederschlag und Temperatur insgesamt ein relativ ähnliches Niveau von CMIP3 und CMIP5. CMIP5 ist hier nur bezogen auf den Trend stärker bewertet worden. Die CORDEX-Simulationen können für das Mittel in den meisten Situationen überzeugen. Trotz der Unterschiede insbesondere bei der Bewertung des Niederschlags im Vergleich zu 4FM scheinen die RM-Ergebnisse ebenfalls als gültige Einschätzungen der Simulationsqualität interpretiert werden zu können.

Bei der Analyse der ZM-Rohgewichte (nicht gezeigt), gemittelt je MME, fällt, ähnlich zu den anderen Metriken, ein Unterschied zwischen Niederschlag und Temperatur auf. Beim Niederschlag liegen die Werte für die meisten Regionen auf einem ähnlichen Niveau und es zeigen sich nur geringe Unterschiede. Das schwächste Ergebnis, d. h. die höchste Abweichung, wird in jeder Jahreszeit für

alle drei MME in Nordafrika berechnet. In den übrigen Regionen fallen die Abweichungen im Schnitt etwas geringer aus. Für fast alle Jahreszeiten und Regionen des Niederschlags sind die relativen Unterschiede so gering, dass von einer vergleichbaren Simulationsqualität basierend auf ZM ausgegangen werden kann. Ähnlich wie bei RM, ist dieses Ergebnis nicht zwingend als fragwürdig anzusehen. Verschiedene Studien mit unterschiedlichen Metriken konnten ebenfalls keine eindeutige Tendenz ausmachen. Einzige Ausnahme ist hier der Nahe Osten, wo CORDEX eine ca. doppelt so hohe Abweichung wie CMIP3 und CMIP5 besitzt.

Bei der Temperatur variieren die Werte zwischen den Regionen stärker. Allerdings ist auch hier keine klare Verbesserung von CMIP5 gegenüber CMIP3 bzw. CORDEX festzustellen. CMIP5 zeigt so in 50 % der Fälle ein besseres Ergebnis als CMIP3. Dasselbe gilt für CORDEX gegenüber CMIP5. Dabei gilt, dass in nahezu allen Situationen, in denen CMIP5 eine Verbesserung zeigt, diese von CORDEX nochmals gesteigert wird. In den übrigen 50 % werden die besten Ergebnisse von CMIP3 simuliert. 9 der 18 Verbesserungen von CMIP5 sind in JJA zu finden. Davon werden 8 nochmals von CORDEX verbessert. Die saisonale Untersuchung (Kapitel 5.2) hat JJA, aufgrund der stärksten Variabilität der Ergebnisse, als wichtiges ZM-Kriterium für die Bewertung im Mittelmeerraum ergeben. Den MME-Ergebnissen zu Folge haben die neuen Modellgenerationen hier die deutlichsten Fortschritte vollzogen, da die übrigen Jahreszeiten eher Stagnation oder leichte Verschlechterungen zeigen. Dennoch kann aufgrund mangelnder Vergleichsmöglichkeiten nicht abschließend geklärt werden, ob ZM ein geeignetes Bewertungskriterium für die Simulationsqualität von Klimamodellen im Untersuchungsraum ist. Die verbesserten Ergebnisse für die JJA-Temperatur der neueren MME lassen aber eine sinnvolle Einschätzung der Simulationen vermuten. Dazu passt auch die uneinheitliche Bewertung der Niederschlagsgüte, die von den zuvor analysierten Metriken geteilt wird.

Für die FPA (nicht gezeigt) gilt im MMR ein ähnliches Muster wie bei der Trendanalyse von PHI. Beim Niederschlag streuen die Werte um  $\pm 0,1$ , wohingegen sie bei der Temperatur deutlich im positiven Bereich verortet sind. Zu Ersterem fällt auf, dass hier weder große Unterschiede zwischen den Jahreszeiten noch zwischen SUB und OPT zu erkennen sind. Angesichts der minimalen Unterschiede sind Schlussfolgerungen bzgl. der Modellentwicklung nicht möglich. Sowohl CMIP3 als auch CMIP5 und CORDEX liegen mal an erster, mal an letzter Stelle mit nur marginalen Unterschieden. Hier besteht für CMIP3 und CMIP5 Einklang mit den Ergebnissen der ZM. Anders ist die Situation bei der Temperatur. Hier liegen die mittleren Rohgewichte über alle vier Jahreszeiten über 0,3 und reichen teilweise über 0,9. In allen MME zeigt JJA die für alle FPA höchsten Ergebnisse. Dabei liegen die SUB-8 Werte zwischen 0,6–0,92, während die OPT-8 Werte zwischen 0,35–0,45 liegen. In allen Situationen gilt dabei, dass CMIP5 die höchsten

Ergebnisse zeigt. Im Gegensatz dazu liegen die mittleren CORDEX-Ergebnisse überwiegend an dritter Position. Dies stimmt mit den trendbasierten Ergebnissen der anderen Metriken bezogen auf den Gesamtmittelmeerraum überein.

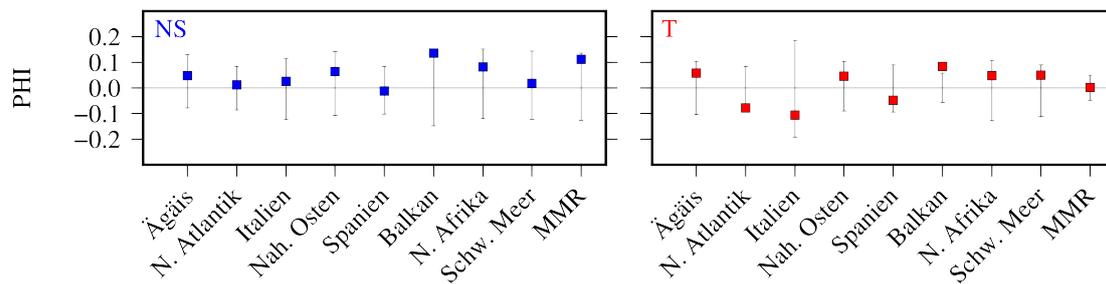
### 5.3.2 Einfluss dynamischen Forcings

Der Vergleich zwischen den verschiedenen MME hat gezeigt, dass die CMIP5-Simulationsgüte durch die Evaluationsmetriken meist ähnlich oder höher als die von CMIP3 eingeschätzt wird. Dieses Resultat bescheinigt der ständigen Weiterentwicklung der Modelle zumindest für die Temperatur ein positives Zeugnis. Allerdings bestehen auch innerhalb gleicher Modellgenerationen teilweise relevante Unterschiede mit Blick auf die technischen Modellfähigkeiten. Neben offensichtlichen Änderungen, wie horizontaler und vertikaler Auflösung, sind auch unterschiedliche Forcings und Prozesse, die in einem Modell nur statisch und in einem anderen dafür dynamisch simuliert werden, relevant. Zum Einfluss unterschiedlicher Auflösungen auf die Simulationsgüte wurden bereits verschiedene Analysen durchgeführt, die meist einen positiven Effekt ergaben (SCHIEMANN ET AL., 2008, 2014, u.a.). Vor dem Hintergrund der verschiedenen Evaluationsansätze wird daher in diesem Abschnitt ein Blick auf den Einfluss der Forcing-Unterschiede der CMIP3-Modelle geworfen.

Unter den CMIP3-Modellen existieren zwei ca. gleichgroße Gruppen von Modellen (11; 13), die vulkanische und Meersalzaerosole sowie die Solarvariabilität dynamisch simulieren (11 Modelle). Bei der Gruppe der anderen Modelle wird mindestens bei einer dieser Variablen ein statisches Forcing angewandt (13 Modelle). Im neueren MME ist diese Unterscheidung nicht möglich, da hier alle Modelle über die dynamische Berechnung verfügen. Diese Erkenntnis lässt den Schluss zu, dass die dynamische Simulation hochwertigere Ergebnisse erzeugen sollte. Mit dieser Grundannahme wurden zwei Sub-Ensembles von CMIP3 erstellt und die jeweils mittleren Rohgewichte der entwickelten Metriken analysiert.

Neben dem Vergleich der beiden Sub-Ensembles wurden basierend auf zufällig gezogenen Rohgewichten Mittelwerte gebildet. Dies wurde 10.000-fach wiederholt und die zweifache Standardabweichung über alle Werte berechnet. Da die Werte normalverteilt sind, wird so eine Art Konfidenzintervall abgeschätzt, in dem über 95 % der mittleren Rohgewichte liegen (z.B. CLAUSS ET AL., 2004, S. 149). In Abbildung 5.10 ist das mittlere Rohgewicht des Sub-Ensembles der Modelle mit dynamischen Forcing für PHI exemplarisch für den Trend des annualen Niederschlags (blau) sowie der Jahresmitteltemperatur (rot) dargestellt. Zusätzlich zeigen die Balken die zweifache Standardabweichung des Mittels der zufällig gezogenen Gewichte.

Beim Niederschlag liegen die Werte für alle Regionen außer Spanien leicht im



**Abbildung 5.10** Einfluss des dynamischen Forcings auf die Simulationsgüte von CMIP3 im Mittelmeerraum. Dargestellt sind die mittleren annualen PHI-Trendrohgewichte des Sub-Ensembles mit dynamischem Forcing für Niederschlag (blau) und Temperatur (rot). Die Balken zeigen die zweifache Standardabweichung über das Mittel 10.000-fach zufällig gezogener Rohgewichte für jede Region.

positiven Bereich. Nur für den Balkan und MMR erreicht die Differenz die Grenze der zweifachen Standardabweichung. Ansonsten liegen die Werte deutlich innerhalb der zufälligen Varianz bzw. in den meisten Fällen sogar um 0 gestreut. Ein sehr ähnliches Bild zeigt sich auch für die Temperatur. Hier liegen die Werte in sechs Regionen im positiven Bereich. Auch hier wird die Grenze der zweifachen Standardabweichung in zwei Regionen erreicht bzw. für die Balkan-Region leicht überschritten. Da der überwiegende Teil der Ergebnisse deutlich innerhalb der Standardabweichungen liegt, kann das Ergebnis für Balkan als Ausreißer bewertet werden. Dieses Fazit deckt sich mit den übrigen Jahreszeiten. Während die Stärke und Art der Unterscheidung etwas variieren, liegen die Werte in nahezu allen Situationen im Bereich zufälliger Ziehungen. Dieses Ergebnis deckt sich auch mit den nicht gezeigten Ergebnissen des klimatologischen Mittels der RM und 4FM. Da sich in keiner Region besonders hohe Gewichte zeigen, sind bezogen auf den regionalen Zusammenhang keine Rückschlüsse möglich. Analog zu diesen Ergebnissen basierend auf PHI verhalten sich auch die Werte sämtlicher anderen Metriken. Während LOR eine deutlich stärkere Unterscheidung generiert, sind hier allerdings auch die Bereiche der zweifachen Standardabweichung analog viel größer. In der Konsequenz zeigen alle Metriken für die absolute Mehrheit der Situationen (>95 %) keinen relevanten statistischen Zusammenhang zwischen den dynamisch simulierenden Sub-Ensembles und einer verbesserten Simulationsgüte in den untersuchten Regionen.

Da grundsätzlich davon auszugehen ist, dass die dynamische Modellierung qualitativ hochwertiger ist, liegt der Schluss nahe, dass hier andere Modelleigenschaften bei den Evaluationen eine größere Rolle einnehmen. Ein weiterer Grund könnte sein, dass die betrachteten Prozesse für die untersuchte Region des Mittelmeerraums nur eine untergeordnete Rolle spielen und in anderen Teilen der

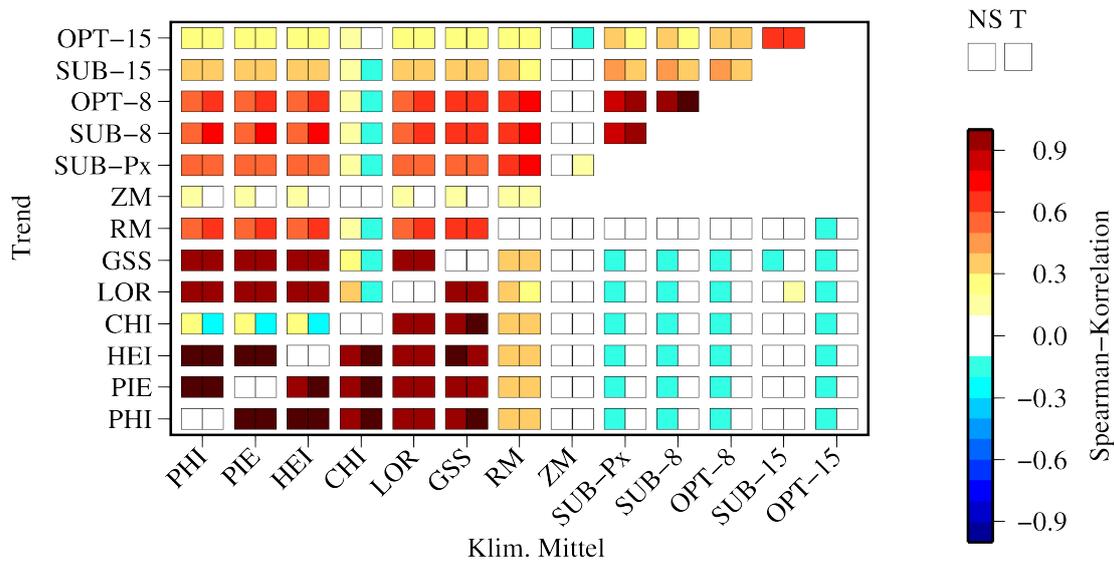
Erde relevanter sind. Diese These wird in Kapitel 7 bei der Untersuchung sechs global verteilter Regionen überprüft. Neben der regionalen Bedeutung könnten jedoch auch Modelleigenschaften wie die bereits angesprochene Auflösung, aber auch andere Parametrisierungen, entscheidend für eine hohe Übereinstimmung mit Beobachtungen bzw. zuverlässigen Projektionen sein. Desweiteren könnten auch die untersuchten Temperatur- und Niederschlagswerte weniger stark davon betroffen sein als andere Variablen. Daher wird an dieser Stelle der generelle Zusammenhang zwischen höherer Simulationsgüte und Modellen mit dynamischem Forcing gegenüber Modellen mit einem statischen Ansatz nicht ausgeschlossen. Die Ergebnisse der Metriken für den Mittelmeerraum legen unter diesen Experimentbedingungen einen Zusammenhang jedoch nicht nahe. Im folgenden Kapitel wird die Korrelation der individuellen Gewichte aller Metriken untersucht.

## 5.4 Korrelation der Evaluationsansätze

Abschließend zur Auswertung der Rohgewichte des Mittelmeerraums werden in diesem Abschnitt die Ergebnisse aller Metriken in Relation gesetzt. Zum Vergleich der Übereinstimmung zwischen den Metriken wurden alle berechneten Gewichte für jede Metrik in eine Rangliste geschrieben und anschließend die Spearman-Korrelation zwischen den Metriken berechnet. Da für die FPA nur MMR evaluiert wurde, zeigt Abbildung 5.11 die Spearman-Korrelation zwischen allen Metriken für diese Region. Die Ergebnisse in den übrigen Regionen bzw. insgesamt über alle Regionen werden im Anschluss für die restlichen Metriken zusammengefasst.

Abbildung 5.11 ist in ihrer Diagonale zweigeteilt zu lesen. Im linken oberen Bereich sind die Korrelationen basierend auf den Ergebnissen des 50-jährigen Trends angetragen. Unterhalb der Diagonale sind die Korrelationen für die äquivalenten Ergebnisse des klimatologischen Mittels angetragen. Auf der Diagonale selbst ist die Korrelation zwischen den Rohgewichten von Trend und Mittel derselben Metrik (wo vorhanden) abzulesen. Da für ZM und FPA hier nur eine Größe untersucht wurde, ist dieser Bereich nur für die 4FM und RM gefüllt. Für jede Gegenüberstellung wurde die Korrelation für jeweils Temperatur und Niederschlag separat untersucht.

Zwischen den 4FM sind die Ergebnisse des 50-jährigen Mittels durchweg über 0,8 mit einem Maximum von 0,99 zwischen PHI und PIE sowie HEI. Die niedrigsten Übereinstimmungen sind für LOR zu beobachten, während CHI und GSS ebenfalls Werte über 0,9 angeben. Markante Unterschiede zwischen Temperatur und Niederschlag sind hier nicht festzustellen. Auffälliger ist das Bild des 50-jährigen Trends. Während die ersten drei Metriken wieder sehr hohe Übereinstimmungen, über 0,99 sowohl für Niederschlag als auch Temperatur, zeigen,



**Abbildung 5.11** Spearman-Korrelation aller Evaluationsansätze für den Gesamtmittelmeerraum. Angetragen sind jeweils die Ergebnisse des Niederschlags (linkes Feld) und der Temperatur (rechtes Feld) für den Trend (oben-links) sowie, gegebenenfalls das klimatologische Mittel (rechts-unten). Die Diagonale zeigt die Korrelation zwischen Trend und Mittel für die 4FM und RM.

ist die Korrelation für CHI hier deutlich geringer. Insbesondere für die Temperatur liegen die Werte im negativen Bereich. Für den Niederschlag sind die Korrelationen mit  $\geq 0,2$  zwar etwas höher, zeigen aber trotzdem die niedrigsten Übereinstimmungen zwischen allen 4-Felder-Metriken. LOR und PIE verhalten sich mit Werten von 0,8 bzw. 0,9 vergleichbar zu den Ergebnissen des Mittels. Die auffällige Abnahme der Spearman-Korrelation für CHI kann auf die Quadrierung in ihrer Formel zurückgeführt werden. Hierdurch werden Simulationen, die eine besonders schlechte Übereinstimmung mit den Beobachtungsdaten aufzeigen, ebenso behandelt wie Simulationen mit nahezu identischen Ergebnissen mit den Beobachtungsdaten.

Da der Unterschied besonders deutlich beim 50-jährigen Trend des Niederschlags zu beobachten ist, muss es sich hier um eine sehr komplexe Größe für die Simulationen handeln. Entsprechend haben viele Simulationen niedrige Ergebnisse in den anderen Metriken zugeteilt bekommen. Der abschließende Vergleich der Übereinstimmung zwischen 50-jährigem Trend und Mittel, veranschaulicht auf der Diagonale, zeigt keinerlei Zusammenhang. Damit geben alle 4FM an, dass gute Ergebnisse bei der Darstellung des Trends nichts darüber aussagen, ob das langjährige Mittel auch gute Bewertungen bezüglich der Simulationsqualität erhält. Dies gilt ebenso für CHI. Für die übrigen Regionen des Mittelmeerraums ist das Bild konsistent. Die Korrelation zwischen den fünf geeigneten 4FM liegt

sogar noch näher an 1. Auch hier gilt, dass ein Zusammenhang zwischen den Ranglisten des Trends und des Mittels nicht nachvollzogen werden kann. Ansonsten scheinen die 4FM sowohl für Niederschlag als auch Temperatur zumindest für die Erstellung von Ranglisten zur Modellbewertung untereinander austauschbar zu sein.

Für die FPA sind die Korrelationen zwar sowohl für Temperatur als auch für Niederschlag durchweg positiv, das Niveau schwankt jedoch relativ deutlich. Die höchste Übereinstimmung liegt zwischen SUB-Px und SUB-8 für Temperatur. Auch generell ist die Spearman-Korrelation für Temperatur in den meisten Fällen höher als die für Niederschlag. Die beiden Ausnahmen sind hier im Vergleich von SUB-Px, SUB-8 und OPT-15 zu finden. So zeigt sich eine gewisse Gruppenbildung von Metriken, die besonders stark korrelieren. Des Weiteren ist eine klare Trennung zwischen den Metriken mit unterschiedlicher EOF-Anzahl zu erkennen. Die Korrelationen der Rohgewichte mit gleicher EOF-Zahl sind dabei sowohl für 15 als auch 8 EOFs deutlich höher als bei gemischten Rangfolgen.

Da für ZM und die FPA hier nur eine Variable genutzt wurde, ist diese nur auf der Trendseite angetragen. Es fällt eine deutliche Zweiteilung hinsichtlich der Übereinstimmungen zwischen den verschiedenen Evaluationsansätzen auf. Werte oberhalb der Diagonale zeigen meist positive Korrelationen, während die Übereinstimmungen unterhalb der Diagonale im Vergleich zwischen den FPA und den 4FM um 0 liegen. RM hat hier mit Werten zwischen 0,3–0,4 leicht positive Korrelationen mit den 4FM und FPA, während die ZM keine Korrelation mit den anderen Metriken zeigt. Analog sind auch keine Übereinstimmungen zwischen den Ergebnissen des 50-jährigen Trends und dem klimatologischen Mittel festzustellen.

Beim Vergleich der Trendkorrelationen zwischen den verschiedenen Metrikgruppen zeigen sich zwei auffällige Muster. Zum einen besitzen die Ergebnisse von CHI und ZM mehrheitlich keine Korrelation mit den weiteren Metriken. Zum anderen gilt für die übrigen Ergebnisse, dass leichte bis mittlere positive Zusammenhänge existieren. Besonders hoch fallen diese in den Kombinationen von den 4FM, RM sowie SUB-Px und FPA-8 Ergebnissen ( $> 0,5$ ) aus. Diese Korrelationen zeigen, dass die Metriken für den Mittelmeerraum sowohl für Niederschlag als auch Temperatur überwiegend denselben Modellen hohe Simulationsgüte bescheinigen. Dies ist insofern auffällig, als dass die Metriken sich stark in ihrer Komplexität und Durchführung unterscheiden. Insbesondere die Rohgewichtsauswertungen der FPA haben deutlich höhere Bewertungen für Temperatur ergeben als die der 4FM. Dennoch wird für den Mittelmeerraum hier eine hohe Korrelation in der Rangfolge der Modelle deutlich. Diese Übereinstimmung ist gerade vor dem Hintergrund, dass die 4FM bisher noch nicht als Kriterium zur Evaluation von GKM oder RKM verwendet wurden, sehr interessant. Auch RM, das die mit

Abstand einfachste Modellevaluation nutzt, zeigt sowohl für die Temperatur als auch den Niederschlag tendenziell konsistente Ergebnisse mit den beiden anderen Metrikblöcken. Generell gilt auch für den Trend, dass sich die 4FM-, RM- und ZM-Ergebnisse des Mittelmeerraums auf die Unterregionen übertragen lassen. Die Korrelationen von 4FM und RM nehmen dabei noch leicht zu. ZM stellt sich hier dagegen als eine unabhängige Methode der Evaluierung heraus. So sind keine Zusammenhänge mit den Ergebnissen anderer Metriken zu beobachten. Allerdings gilt, dass die Grundlage der Bewertung nur für ZM nicht auf einem Trendmuster beruht. In der Folge muss sie also nicht als eine unzuverlässige Metrik, sondern vielmehr als eine alternative Herangehensweise mit anderen Kriterien betrachtet werden.

Insbesondere die Ergebnisse der 4FM zeigen sich sehr stabil untereinander und weisen eine hohe Korrelationen zu den Ergebnissen von RM und Teilen der FPA auf. Des Weiteren zeigen die FPA viele Überschneidungsflächen und korrelieren untereinander durchweg positiv. Die Ergebnisse der Korrelationen mit anderen Metriken machen die FPA-15 im weiteren Verlauf redundant, da hier durch erhöhtes Rauschen (vgl. Abbildung 5.2) eher ein Verlust des Signals als eine gesteigerte Qualität der Evaluation zu beobachten ist.

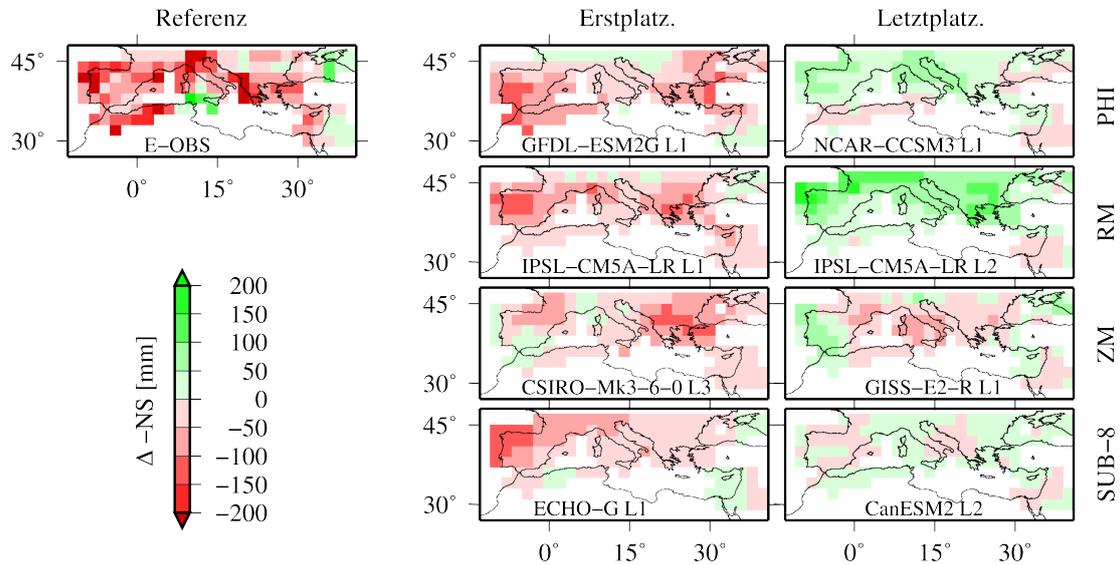
## 5.5 Hintergründe der Modellbewertung

In diesem Abschnitt werden abschließend die Hintergründe der Evaluationsrohgewichte veranschaulicht. Damit wird gezeigt, welches Modell ein hohes und welches ein niedriges Gewicht, abhängig von der jeweiligen Metrik, zugeteilt bekommen hat. Anschließend werden die relativen Gewichte der Simulationen verglichen, um einen Überblick über die bisherigen Ergebnisse zu liefern und ihre Bedeutung für die in Kapitel 6 durchgeführte Gewichtung einschätzen zu können.

### 5.5.1 Vergleich der Evaluationsgrundlagen

Obwohl die angewandten Metriken alle das Niederschlags- und Temperaturverhalten im Zeitraum 1960–2009 analysieren, untersuchen sie verschiedene Aspekte des Klimas. Die Folge ist, dass, wie zuvor gezeigt, nicht dieselben Modelle das gleiche Gewicht bzw. dieselben Positionen in einer Rangliste für die jeweilige Situation erhalten. Im Folgenden werden daher die jeweils erst- und letztplatzierten Simulationen exemplarisch von vier Metriken für MMR abgebildet. Dabei gilt zu beachten, dass die Unterschiede zwischen den Ranglisten teilweise sehr gering sind. Zusätzlich basiert die Wahl der jeweiligen Simulation lediglich auf der ersten bzw. letzten Rangposition der betroffenen Situation und ist nicht zur

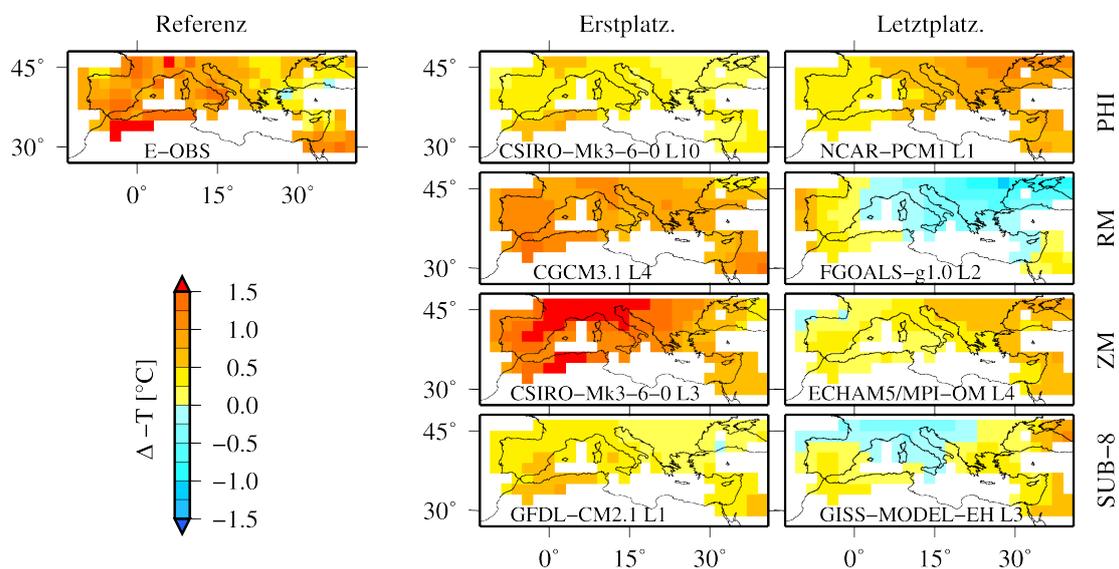
Verallgemeinerung gedacht. Abbildung 5.12 zeigt die entsprechenden Simulationen des annuellen Niederschlagstrends im MMR für PHI, RM, ZM und SUB-8 sowie das Referenzmuster von E-OBS.



**Abbildung 5.12** Vergleich der annuellen Trendmuster des Niederschlags [mm] für den Gesamtmittelmeerraum. Angetragen sind jeweils die Trendmuster im Zeitraum 1960–2009 der erst- und letztplatzierten Simulationen von PHI, RM, ZM und SUB-8 sowie der Referenzdatensatz.

Vergleicht man das Muster von E-OBS mit den dargestellten Simulationen fällt auf, dass ein ähnliches Bild bei den erstplatzierten Simulationen zu sehen ist. Die Muster der letztplatzierten Simulationen weichen dagegen teilweise deutlich ab. Dies gilt insbesondere für PHI und RM. Hier sind bei den erstplatzierten Simulationen, vergleichbar zu E-OBS, in großen Teilen der Region deutliche Abnahmen mit einem Maximum im Westen über der Iberischen Halbinsel sowie in der Mitte für Griechenland bzw. der Ägäis zu verzeichnen. Darüber hinaus zeigen die beiden Simulationen auch eine leichte Niederschlagszunahme im Osten, wobei für PHI dies im Nahen Osten und für RM im Bereich des Schwarzen Meeres verzeichnet wird. Entgegen den Referenzdaten zeigen beide erstplatzierten Simulationen (GFDL-ESM2G Lauf 1, IPSL-CM5A-LR L1) auch eine leichte Abnahme an den südlichen Alpen. Die beiden letztplatzierten Läufe dieser beiden Metriken (NCAR-CCSM3.0 L1, IPSL-CM5A-LR L2) zeigen dagegen fast ein gegensätzliches Bild mit einer deutlichen Niederschlagszunahme im MMR und nur leichten Abnahmen im Osten bzw. Süden. Auffällig ist hier, dass RM zwei Läufe von IPSL-CM5A-LR auf die extremen Rangpositionen platziert. Die Darstellung liefert allerdings die Erklärung hierfür.

Dieses klar getrennte Muster ist bei ZM und SUB-8 nicht zu beobachten. Dabei muss allerdings beachtet werden, dass PHI und RM im Gegensatz zu ZM und SUB-8 direkt auf die Trendmuster der Simulationen zur Evaluation zurückgreifen. Die erstplatzierte Simulation von SUB-8 (ECHO-G L1) zeigt dennoch ein recht ähnliches Bild mit überwiegender Abnahme im MMR. Hier ist ein Maximum über dem nordöstlichen Teil der Iberischen Halbinsel und nur leichte Zunahme im Osten zu beobachten. CanESM2 L2 zeigt dagegen überwiegend leichte Abnahmen in großen Teilen der Region. Die beiden ZM-Simulationen, CSIRO-MK-6-0 L3 und GISS-E2-R L1, zeigen eine deutlich höhere Ähnlichkeit untereinander als dies bei den übrigen Metriken zu beobachten war. Aus Gründen der Vergleichbarkeit zu den anderen Metriken wurde hier als Darstellungsform ebenfalls das regionale Niederschlagsmuster gewählt. Da ZM die erklärten Varianzen untersucht, wäre die Darstellung des Periodogramms naheliegender (vgl. Kapitel 4.3). Hier liegen die Unterschiede zwischen den Referenzdaten und CSIRO-MK-6-0 L3 deutlich niedriger als für GISS-E2-R L1. Dennoch kann davon ausgegangen werden, dass die für CSIRO-MK-6-0 L3 dargestellte überwiegende Niederschlagsabnahme auch für die Zeitreihenuntersuchung relevant ist. Die regionalen Unterschiede der Niederschlagsveränderung in Abbildung 5.12 sind vor diesem Hintergrund nicht überzubewerten.



**Abbildung 5.13** Vergleich der annualen Trendmuster der Temperatur [°C] für den Gesamtmitelmeerraum. Angetragen sind jeweils die Trendmuster im Zeitraum 1960–2009 der erst- und letztplatzierten Simulationen von PHI, RM, ZM und SUB-8 sowie der Referenzdatensatz.

Analog zum zuvor diskutierten Niederschlagstrend zeigt Abbildung 5.13 die jeweils erst- und letztplatzierten Simulationen für den Temperaturtrend des MMR für PHI, RM, ZM und SUB-8. Das Muster der E-OBS-Referenzdaten zeigt eine durchgehende Erwärmung im Bereich zwischen  $0,2\text{--}1,8^\circ\text{C}$ . Das Maximum liegt hier im Norden Afrikas und das Minimum in der Ägäis, wo einzelne Felder minimale Abnahmen zeigen. Das Muster der erstplatzierten Simulationen ist hier für alle Metriken sehr homogen. Alle Simulationen zeigen für das gesamte Untersuchungsgebiet eine mittlere bis starke Erwärmung. Die maximale Erwärmung von über  $1,5^\circ\text{C}$  im gesamten westlichen Teil von MMR wird von CSIRO-MK-6-0 L3 (ZM) abgebildet, während CSIRO-MK-6-0 L10 (PHI) insgesamt die niedrigste Erwärmungsrate zeigt.

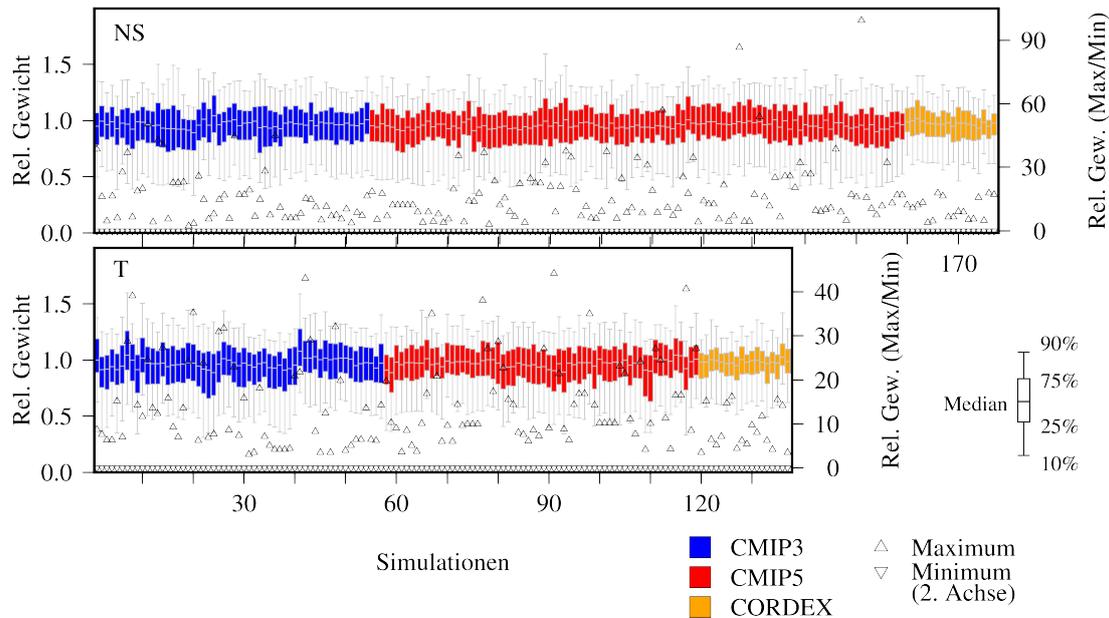
Hier wird deutlich, dass PHI den Bias zwischen Referenz- und Modelldaten ignoriert. Stattdessen ist die räumliche Verteilung mit einem Maximum für Nordafrika bzw. einer stärkeren Erwärmung im Westen von MMR entscheidend. Dies zeigt auch den Unterschied zu NCAR-PCM1 L1. Während für RM und SUB-8 hier Simulationen mit einem hohen Anteil an Abkühlung ausgewählt wurden, ist das Raummuster für PHI (exemplarisch für alle 4FM) entscheidend. So ist die Verteilung der Datenpunkte von NCAR-PCM1 L1 im Vergleich zu den Referenzdaten vertauscht, sodass die stärkste Erwärmung im Nordosten simuliert wird. Dies führt zu einer schlechten Bewertung. Ähnlich sieht es bei ZM aus, wo ECHAM5/MPI-OM L4 dieselbe Verteilung zeigt. Zur Bestätigung liegt dieses Modell in der Rangliste von PHI auf der drittletzten Position. Die Interpretation der ZM-Bewertung ist dagegen basierend auf dem regionalen Muster analog zum Niederschlag nicht zielführend.

Diese Veranschaulichung der Hintergründe der Modellevaluationen zeigt, dass alle Metriken sinnvolle Ergebnisse basierend auf den von ihnen untersuchten Parametern bieten. Dies gilt auch für die hier nicht gezeigten übrigen Situationen sowie für die Metriken aus der 4FM- und FPA-Gruppe. Auch wenn die den Abbildungen 5.12 und 5.13 zu Grunde liegenden Ranglisten unterschiedliche Simulationen für die gezeigten Metriken ergaben, gilt dies nicht für alle Situationen. Wie bereits kurz für den Vergleich von PHI und ZM bei der Temperaturänderung angedeutet, liegen viele Simulationen in ähnlichen Rangpositionen von unterschiedlichen Metriken und sind abhängig von der Situation, teilweise sind sie auch identisch.

### 5.5.2 Bedeutung der relativen Gewichte

Die in Kapitel 5.1 durchgeführte Analyse der Rohgewichte für die verschiedenen Regionen hat gezeigt, dass es teilweise deutliche Diskrepanzen zwischen den Simulationen gibt. Hierfür sind jedoch nicht nur die verschiedenen Jahreszeiten und Regionen verantwortlich. Die unterschiedlichen Metriken beurteilen die Eva-

luationsergebnisse auch verschieden stark. Da auf Grundlage dieser Bewertungen eine Gewichtung vorgenommen werden soll, ist es von Interesse, sich neben dem Rohgewicht auch mit dem relativen Gewicht, das jede Simulation erhält, auseinanderzusetzen. Dies wird im folgenden Abschnitt eingeschoben, bevor im nächsten Kapitel die Gewichtung durchgeführt wird.



**Abbildung 5.14** Relative Gewichte der 4FM über sämtliche Situationen von Niederschlag (oben) und Temperatur (unten) für alle Regionen des Mittelmeerraums. Dargestellt sind neben den 10 %-, 25 %-, 75 %-, 90 %-Perzentilen und dem Median (linke Y-Achse) das jeweilig maximale und minimale Gewicht (rechte Y-Achse).

Abbildung 5.14 zeigt die Verteilung sämtlicher relativen Gewichte der 4FM je Simulation in Form von Boxplots. Grundlage sind die Rohgewichte basierend auf sämtlichen Situationen des klimatologischen Mittels sowie des Trends (insgesamt 540 Werte). Die Ordinate auf der linken Seite zeigt die relativen Gewichte der Verteilungen. Dieser Bereich bildet hier 80 % der Werte ab, die durch die Boxplots dargestellt sind. Für die Maxima und Minima ist die rechte Ordinate mit einer Skala von 0 bis 105 (NS) bzw. 49 (T) angetragen. Alle Gewichte wurden mit der Anzahl der im MME enthaltenen Simulationen multipliziert, wodurch die Gewichte nicht zwischen 0–1 liegen. Im Fall von Gleichgewichtung würden entsprechend die Gewichte aller Simulationen auf einer horizontalen Linie bei 1 liegen.

Zunächst fallen die Extremwerte auf. Jede Simulation besitzt in mindestens einer Situation ein Gewicht knapp über. Daher liegen alle Minima nahe der Grundli-

nie der zweiten Ordinate. Die Maxima streuen stark über alle Simulationen. Beim Niederschlag werden stärkere Unterschiede deutlich. So wird das bis zu 99,4-fache Gewicht (97, CMIP5) angenommen. Die Mehrheit der Maxima liegt allerdings im Bereich zwischen dem fünf- und zwanzigfachen Gewicht. Die Metrik, die diese Gewichte erzeugt, ist ausnahmslos LOR. Auch die  $4FM_i$ -Ansätze erzeugen keine Gewichte, die ähnlich stark differenzieren, wie bei den Maxima von LOR zu beobachten ist. Trotzdem ist der Effekt durch die stärkeren Abstufungen im Vergleich zu den normalen 4FM ebenfalls deutlich. Bei der Temperatur liegt das absolute Maximum beim 44,2-fachen Gewicht (34, CMIP5). Die übrigen Maxima haben größtenteils ebenfalls Werte zwischen dem fünf- und fünfzehnfachen Gewicht.

Die Boxplots zeigen sowohl für Niederschlag als auch für Temperatur ein sehr homogenes Muster über alle Simulationen. So liegen ca. 50 % der Werte zwischen 0,75–1,25 mit einem Median, der für alle Simulationen um 1 liegt. Hier gibt es einige Ausreißer bei der Temperatur, die an diese Grenzen heranreichen, während die Mehrheit über alle MME eher um 1 liegt. Erst bei Betrachtung der 10 %- und 90 %-Intervalle werden deutlichere Unterschiede erkennbar. Beim Niederschlag sind die größten Abweichungen der Gewichte im negativen Bereich mit einem Minimum von ca. 0,33 (20, CMIP3) zu beobachten. Das positive Extremum liegt dagegen bei 1,6 (35, CMIP5). Bei der Temperatur ist die Verschiebung nach unten etwas schwächer ausgeprägt mit einem Minimum von 0,31 (22, CMIP3) und einem Maximum von 1,59 (41, CMIP3). Insgesamt sind die Simulationen innerhalb der MME relativ homogen, sodass kein Ausreißer in die eine oder andere Richtung auffällt. Die Streuung der CORDEX-Simulationen ist generell etwas niedriger als bei CMIP3 und CMIP5. Dies könnte an der relativ hohen Abhängigkeit der 18 CORDEX-Simulationen liegen. So unterscheiden sie sich teilweise nur durch verschiedene räumliche Auflösungen.

Für die nicht dargestellten Ansätze fällt auf, dass keine der andern Metriken ähnlich hohe relative Gewichte erzeugt wie LOR. Am nächsten kommen hier die Gewichte der  $FPA_z$ , die jedoch ein Maximum bei 9,4 (39, CMIP3) für Niederschlag und 18 (8, CORDEX) für Temperatur haben. Allerdings muss hier beachtet werden, dass die FPA und  $FPA_z$  nur für den MMR berechnet wurden und daher eine deutlich geringere Basis haben als alle anderen Metriken. Auch bei LOR wird kein relatives Maximum für den Gesamtmitelmeerraum berechnet. Ansonsten streuen die relativen Gewichte des Fingerprintings deutlich stärker als bei den anderen Metriken. Auch wenn der Median bei Niederschlag und Temperatur um 1 liegt, reichen die Werte des 90 %-Perzentils bis 3,1 (8, CMIP5) beim Niederschlag. Da durch die  $FPA_z$  beim Niederschlag viele Simulationen auch Nullgewichte zugeteilt bekommen, reicht das 25 %-Perzentil teilweise ebenfalls bis 0. Die Temperaturergebnisse sind dagegen noch deutlich homogener als bei den 4FM mit Werten zwischen 0–1,6 (26, CMIP3).

RM und ZM besitzen Maxima zwischen 2,2 (32, CMIP5) bzw. 3,1 (41, CMIP5). Ein Grund für diese geringen Unterschiede liegt darin, dass beide Metriken im Gegensatz zu  $FPA_z$  und  $4FM_i$  keine Nullgewichte zuteilen. Die Minima liegen so relativ gleichmäßig über alle Simulationen für RM und ZM um 0,6 bzw. 0,3. Bei ZM sind insgesamt die geringsten Schwankungen zu beobachten. Hier liegen knapp 80 % der Werte zwischen 0,8–1,2. Etwas stärker ist die Varianz bei RM mit einem Intervall von 0,6–1,6 für Temperatur und 0,1–1,5 für Niederschlag. Für beide Metriken gilt, dass kaum Unterschiede zwischen Temperatur und Niederschlag zu beobachten sind.

Insgesamt sind die Unterschiede der relativen Gewichte zweigeteilt. Der absolute Großteil der Werte liegt bei allen Metriken um 1 gestreut. Die Bedeutung ist hier aber von Metrik zu Metrik verschieden. RM und ZM sind aufgrund ihrer Berechnungsweise kaum fähig, stärkere Gewichte zu verteilen. Die überwiegende Zahl der Metriken zeigt sich aufgrund hoher Übereinstimmung zwischen den verschiedenen Metriken eher als Schätzer für die Simulationsgüte und damit zur Erstellung von Bewertungsranklisten. Da meist jedoch nur geringe Abstufungen zwischen den Modellen berechnet werden, ist das Potential anschließender Gewichtungen eingeschränkt. Vor dem Hintergrund der Modellevaluation sind jedoch auch diese Metriken wichtig zur Einordnung der Ergebnisse. Darüber hinaus sind mit LOR (bzw. den  $4FM_i$ ) und den  $FPA_z$  auch Metriken in der Lage, teilweise starke Gewichte zu verteilen und so möglicherweise signifikante Änderungen des MME herbeizuführen. Im folgenden Kapitel wird die Gewichtung der Zukunftsprojektionen basierend auf den Gewichten der Evaluationsmetriken durchgeführt.

## 6 Gewichtungen der Zukunftsprojektionen

In diesem Kapitel werden die zuvor analysierten Gewichte auf die Zukunftsprojektionen der aktuellen globalen und regionalen MMEs angewandt. Für alle Gewichtungen gilt, dass zunächst das mittlere Gewicht jedes Modells über alle verfügbaren Endgewichte der zugehörigen Simulationen berechnet wurde. Dieses Verfahren dient dem Ausgleich von Modellen, die eine größere Anzahl an Simulationen in die Untersuchung einbringen, wie beispielsweise GISS-E2-H (CMIP5). Häufig wird für jedes Modell nur eine Simulation verwendet, um den gleichen Effekt zu erzielen. Da die Auswertung der Rohgewichte jedoch gezeigt hat, dass es insbesondere für den Niederschlag erhebliche Unterschiede der Simulationsqualität innerhalb des Ensembles (alle Simulationen) eines Modells gibt, hätte dieser Ansatz hier zu einem Informationsverlust geführt. Für alle gemittelten Modellgewichte wurde abschließend eine Normierung für jedes MME vorgenommen. Da die Erstellung der Gewichte auf dem Zeitraum 1960–2009 basiert, sind diese für die unterschiedlichen Szenarien identisch und nur durch die Verfügbarkeit der Simulationen unterschieden (vgl. Tabellen 2.2, 2.3, 2.4). Zur Gewährleistung der besten Vergleichbarkeit wird im Folgenden für 4FM und RM der Schwerpunkt auf die Trendgewichte gelegt. In den folgenden Untersuchungen liegt das Hauptaugenmerk auf der Analyse der Unterschiede zwischen gleich- und metrikgewichteten Zukunftsprojektionen. Im ersten Schritt werden dabei die Effekte auf die gewichteten Zeitreihen untersucht. Anschließend wird der Einfluss der Gewichtung auf die Unsicherheit der Klimaänderung vom Ende des 20. zum Ende 21. Jahrhunderts beleuchtet. Für jedes der drei MME werden Niederschlag und Temperatur sowie jeweils zwei Emissionsszenarien von CMIP3 (A1B; A2) und CMIP5 sowie CORDEX (RCP4.5; RCP8.5) untersucht. Dabei wird zunächst eine separate Analyse der einzelnen Szenarien und anschließend ein Multi-Szenario-Ansatz verwendet.

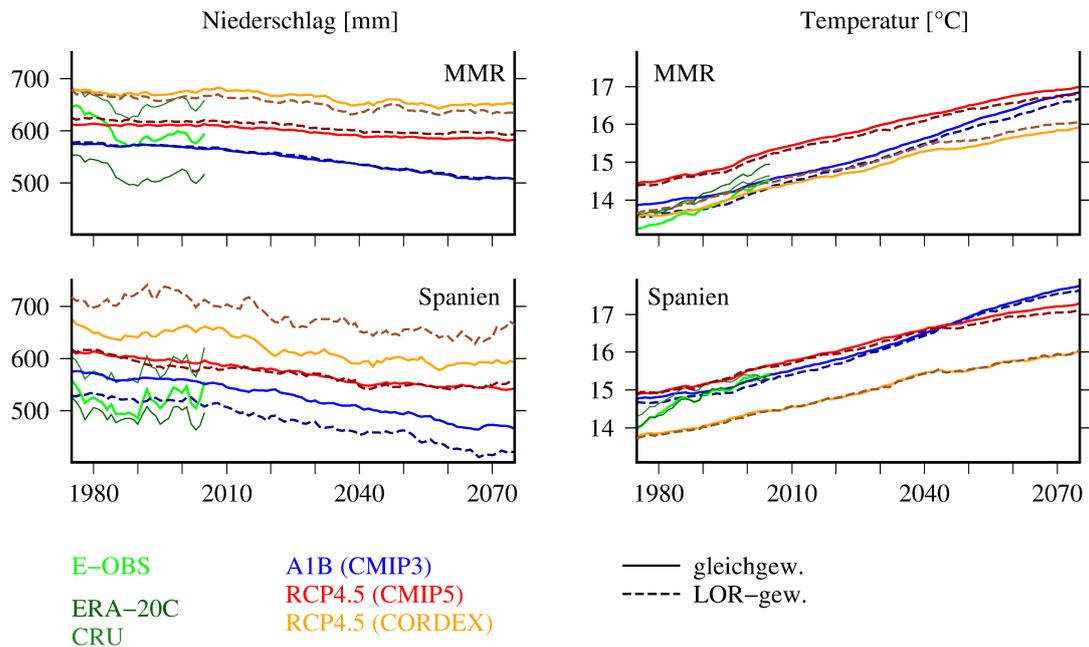
## 6.1 Auswertung der Zeitreihengewichtung

Im Folgenden werden Ergebnisse der metrikgewichteten Zeitreihen für CMIP3, CMIP5 und CORDEX diskutiert und die Unterschiede zu den gleichgewichteten Zeitreihen analysiert. Anschließend wird die Bedeutung für den Trend der Zukunftsprojektionen veranschaulicht. Dies umfasst die jährlichen Werte von Niederschlag und Temperatur des Zeitraums 1970–2098. Zur Berechnung der Zeitreihen wurde eine Kosinusgewichtung angewandt, um die Abhängigkeit der Datenpunkte vom Breitengrad zu berücksichtigen (vgl. Kapitel 4.3).

### 6.1.1 Zeitreihengewichtung

Abbildung 6.1 zeigt die annualen Zeitreihen von MMR und Spanien exemplarisch für die neun Untersuchungsregionen. Angetragen sind jeweils das A1B- (CMIP3) bzw. RCP4.5- (CMIP5/CORDEX) Szenario für Niederschlag und die mittlere Jahrestemperatur. Die Darstellung der gewichteten Zeitreihen basiert auf den Trendgewichten von LOR. Hier sind die höchsten Unterschiede zwischen gleich- und metrikgewichteten Zeitreihen zu beobachten. Dies hilft bei der Verdeutlichung der Gewichtungseffekte bei gleichzeitig hoher Korrelation der Modellbewertungen zu anderen Metriken. Über die Zeitreihen wurde zudem ein gleitender 11-jähriger Filter gelegt, um die Variabilität insbesondere beim Niederschlag zu reduzieren (vgl. Kapitel 3.1). Dies ist notwendig, da die drei MME über unterschiedlich viele Modelle verfügen und andernfalls der eigentliche Effekt der Gewichtung durch jährliche Schwankungen überlagert werden würde. Infolgedessen ist in Abbildung 6.1 der Zeitraum 1975–2093 dargestellt. Neben den MMEs sind für die Jahre bis 2009 die Zeitreihen der drei in dieser Studie verwendeten Referenzdatensätze E-OBS, CRU und ERA-20C angetragen. Dies dient der Verortung des Bias und dem generellen Trend und darf nicht als direkter Vergleich in diesem Zeitabschnitt fehlinterpretiert werden.

In Abbildung 6.1 tritt ein deutliches Muster zu Tage. Alle MMEs zeigen für Niederschlag eine Abnahme, während umgekehrt für Temperatur eine übereinstimmende Zunahme simuliert wird. MMR und Spanien sind vor diesem Hintergrund exemplarisch für die meisten Untersuchungsregionen. Insbesondere bei der Temperatur ist eine sehr homogen und stetig ansteigende Kurve für alle MMEs in beiden Regionen zu beobachten. Zwischen den MMEs ist nur eine geringe Differenz von unter  $1^{\circ}\text{C}$  zu beobachten. Alle zeigen ungefähr die gleiche Zunahme, wengleich CMIP5 die höchste Ausgangstemperatur besitzt und mit über  $18^{\circ}\text{C}$  die maximale Temperatur für das Ende des 21. Jahrhundert simuliert. CORDEX liegt dagegen für MMR auf ähnlichem Niveau wie CMIP3 bzw. am niedrigsten für Spanien. Der Vergleich mit den LOR-gewichteten Zeitreihen zeigt hier generell



**Abbildung 6.1** Vergleich der LOR-Gewichtungsergebnisse der anuellen Zeitreihen der drei MMEs (A1B; RCP4.5) sowie der Referenzdatensätze. Dargestellt sind die Werte des MMR (oben) sowie Spanien (unten) von Niederschlag in mm (links) und Temperatur in °C (rechts).

nur geringe Unterschiede. Lediglich für CMIP5 ist eine deutliche Änderung des Trends zu beobachten, die zum Ende des Zeitraums am größten in Form einer leichten Abnahme ist. Bei CORDEX und CMIP3 wirken sich die Unterschiede zwischen LOR- und gleichgewichteter Zeitreihe nicht stark aus. Beim Niederschlag zeigen sich insbesondere für Spanien deutlichere Unterschiede, sowohl was den Bias zwischen den MMEs als auch die Trendänderung angeht. Insgesamt zeigt hier CORDEX die höchsten Niederschlagswerte bei gleichzeitig der stärksten Abnahme bis 2075. CMIP3 und CMIP5 liegen dabei etwas niedriger, eher auf dem Niveau der Beobachtungsdaten, nehmen aber ebenfalls deutlich ab. Das Bild lässt sich auf MMR übertragen, wo der Unterschied zwischen den MMEs ebenso wie die Niederschlagsänderung etwas niedriger ausfällt. Die LOR-Gewichtung wirkt sich am stärksten in Spanien für CORDEX aus. Hier wird das Niederschlagsniveau deutlich angehoben und auch der Abnahmetrend verringert. Bei CMIP3 bleibt das Niederschlagsniveau dagegen konstant und der Trend nimmt zu, was zu einer stärkeren Niederschlagsabnahme führt. Bei CMIP5 sind dagegen nur minimale Unterschiede zur gleichgewichteten Zeitreihe zu beobachten. Im MMR sind die Unterschiede zur Gewichtung deutlich geringer und wirken sich nur in einer leicht verstärkten Abnahme für CORDEX bzw. etwas höherem Nieder-

schlagsaufkommen für CMIP5 aus. Der in Abbildung 6.1 gewonnene Eindruck von A1B/RCP4.5 bezogen auf die LOR-Gewichtungseffekte lässt sich nahtlos auch auf A2/RCP8.5 übertragen, da die Gewichte im historischen Zeitraum erstellt wurden und sich entsprechend ähnlich auswirken. Insgesamt ist dort eine etwas stärkere Niederschlagsab- und Temperaturzunahme zu beobachten. Die übrigen Charakteristika sind trotz leicht abgewandelter Modellverfügbarkeit stabil.

Für die übrigen (nicht gezeigten) Regionen fällt beim Niederschlag allgemein eine hohe Diversität der Niederschlagsmenge zwischen den Regionen im Mittelmeerraum auf. So zeigt Nordatlantik mit ca. 750–1150mm den mit Abstand höchsten Niederschlag über alle MMEs. Das Minimum ist im Nahen Osten mit Werten < 200mm zu finden. Die übrigen Regionen zeigen jährliche Niederschläge zwischen 300–750mm. Die unterschiedlichen Emissionsszenarien fallen dagegen über weite Strecken des betrachteten Zeitintervalls nur schwach ins Gewicht. Es fällt auf, dass es zum Ende des 21. Jahrhunderts ab ca. 2060 zu einer etwas stärkeren Differenzierung der Szenarien kommt. Im direkten Vergleich zwischen Start- und Endjahr verstärkt sich dieser Effekt. Daraus folgt, dass sich die veränderten Emissionen im 21. Jahrhundert im Mittelmeerraum erst mit einiger Verzögerung auf die Entwicklung des Niederschlags auswirken. In nahezu allen Regionen ist dieser Effekt soweit ausgeprägt, dass zum Ende des 21. Jahrhunderts A2/RCP8.5 etwas niedriger als A1B/RCP4.5 liegen. Des Weiteren werden Unterschiede auch bei der Betrachtung der Trends über das Zeitintervall deutlich. So zeigen ca. 56 % der A1B/RCP4.5-Trends eine Abnahme des Niederschlags. Der gleiche Anteil ist auch bei A2/RCP8.5 zu beobachten. Damit gibt es zwischen den Jahreszeiten und Regionen zwar eine relativ gleichmäßige Aufteilung von Zu- bzw. Abnahmen, die Emissionsszenarien haben aber auf das Vorzeichen keinen Einfluss, sondern verstärken den Trend bzw. schwächen ihn ab. Im Gegensatz zu den unterschiedlichen Emissionsszenarien wirkt sich die LOR-Gewichtung in den meisten Situationen über einen Großteil des Zeitintervalls aus. Die Änderung äußert sich in vielen Situationen in Form einer Verschiebung des Betrags. Ob es durch die Gewichtung zu einer Verstärkung oder Abschwächung des jährlichen Niederschlags kommt, ist auch hier von dem jeweiligen Szenario, MME, der Region und Jahreszeit abhängig. Neben der Verschiebung zeigen sich darüber hinaus auch teilweise deutliche Effekte für den Trend. Art und Stärke sind jedoch wie schon für die Verschiebung situationsabhängig.

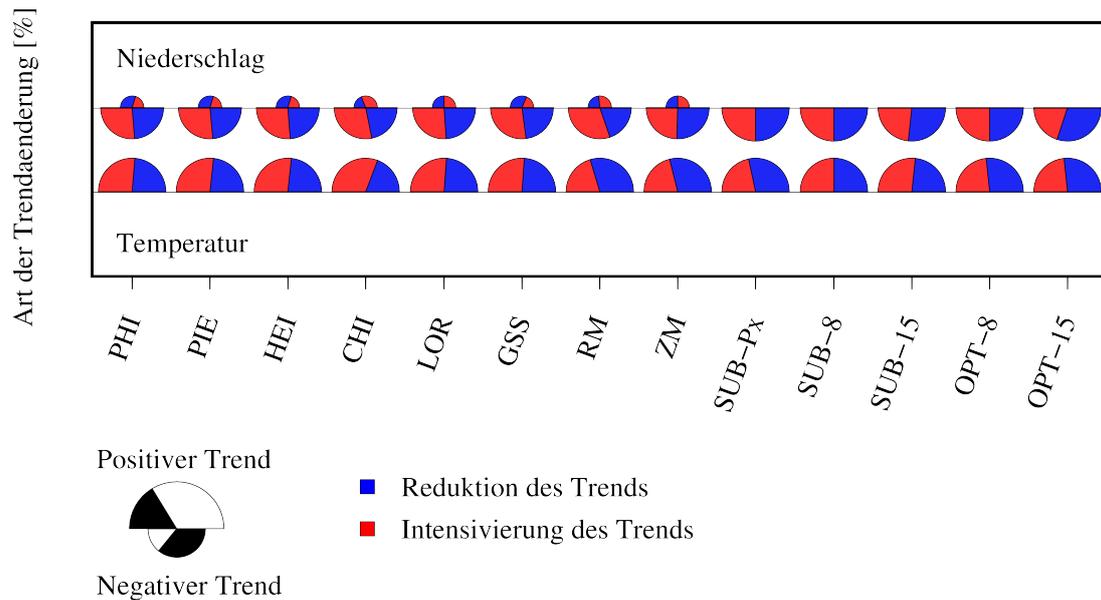
Die Temperaturzeitreihen der übrigen Regionen zeigen die deutlichsten Unterschiede beim Vergleich der verschiedenen Emissionsszenarien. So steigt die Temperatur zum Ende des 21. Jahrhunderts um bis zu 6°C unter dem A2/RCP8.5-Szenario und um 2–3°C für A1B/RCP4.5. Die Temperatur ist auch hier stark abhängig von der betrachteten Situation. So werden gerade im ariden Nahen Osten oder auch in den Regionen Ägäis und Nordafrika Werte von über 21°C bis zum

Ende des 21. Jahrhunderts im Mittel projiziert. Wie in Abbildung 6.1 dargestellt, ähneln sich die MMEs bei der Temperatur deutlich stärker als es beim Niederschlag der Fall ist. So liegen die Werte zum Ende des 20. Jahrhunderts in den meisten Regionen innerhalb von ca. 1–2°C. Auch die generelle Entwicklung der Szenarien wird in den meisten Situationen übereinstimmend ausgeführt. Dabei ist allerdings nochmals darauf hinzuweisen, dass gerade CORDEX und CMIP5 nicht als völlig unabhängige Informationsquellen gewertet werden können. CMIP3 zeigt im Vergleich dazu die größten Unterschiede hinsichtlich der Projektionen. Die übrigen Jahreszeiten fügen sich meist in das zuvor beschriebene Muster ein, wobei sich hier insbesondere die Temperaturniveaus ändern. Einzig für JJA ist neben der generell höheren Ausgangstemperatur für einen Großteil der Regionen eine nochmals verstärkte Erwärmung zu beobachten. Für die Trends aller Regionen und Jahreszeiten gilt, dass es über den untersuchten Zeitraum zu einer Erwärmung kommt. Der Vergleich der LOR-gewichteten zu den gleichgewichteten Zeitreihen zeigt für die Temperatur im Gegensatz zum Niederschlag eher geringe Unterschiede. Deutliche Abweichungen oder Änderungen des Trends sind für alle Jahreszeiten nicht zu beobachten. Dieses Bild deckt sich auch mit den Ergebnissen der übrigen Gewichtungsmetriken. Gerade bei Betrachtung des Trends sind kaum größere Divergenzen festzustellen. Dieses Ergebnis ist jedoch nicht unerwartet. So hat die Analyse der Modellbewertungen (vgl. Kapitel 5) gerade für die Temperatur überwiegend hohe Bewertungen über die Mehrheit der Modelle ergeben. Entsprechend gilt für die Trends der LOR-gewichteten Zeitreihen ebenso wie für die gleichgewichteten Ergebnisse, dass es in allen Regionen und Jahreszeiten zu einer relativ stabilen Erwärmung kommt.

### 6.1.2 Gewichtungseffekt auf den Trend

Nachdem zuvor der Gewichtungseffekt für LOR auf die Zeitreihen von Niederschlag und Temperatur analysiert wurde, wird nun die Trendänderungstendenz aller Metriken veranschaulicht. Zuvor wurde LOR aufgrund seiner intensiven Gewichte gewählt, um Unterschiede möglichst deutlich hervorzuheben. In Abbildung 6.2 sind die Trendänderungen aller Metriken in Form eines Tortendiagramms dargestellt. Hier werden die Änderungen aller Metriken unabhängig von ihrer Intensität veranschaulicht. Bei der Trendänderung wird der Zeitraum 1970–2098 untersucht. Die Fläche aus dem oberen und unteren Halbkreis ergibt 100 % aller Trends. Dies sind für alle Situationen 540 Trends. Für die Metriken der FPA ist die Anzahl auf 60 reduziert, da hier nur MMR untersucht wurde. Für jeden Halbkreis wurde analysiert, ob es zu einer Intensivierung (rot) oder Reduktion (blau) des bestehenden Trends gekommen ist.

Zunächst fällt in Abbildung 6.2 ein sehr deutliches Muster für Temperatur und



**Abbildung 6.2** Effekte der metrikgewichteten Zeitreihen auf den Trend der Zukunftsprojektionen. Dargestellt für die jeweilige Metrik ist der Anteil der Trendänderung für Niederschlag (oben) und Temperatur (unten). Die Fläche jedes Diagramms ist auf 100 % normiert und teilt sich auf den relativen Anteil an positiven (oberer Halbkreis) und negativen Trends (unterer Halbkreis) auf.

Niederschlag auf. Bei der Temperatur sind ausnahmslos alle Trends im positiven Bereich, d.h. für jede der 540 (bzw. 60) Situationen wird eine Temperaturzunahme beobachtet. Umgekehrt sind für den Niederschlag deutlich überwiegend Abnahmen festzustellen. Dies deckt sich mit den Eindrücken aus Abschnitt 6.1.1. Bei den Effekten der Gewichtungen auf die Trends zeigen sich einige unterschiedliche Muster. Für den kleinen Anteil der positiven Niederschlagstrends ist das Bild relativ homogen. Die meisten Metriken zeigen hier ein ausgeglichenes Ergebnis, wobei die 4FM und ZM eher eine Reduktion und RM (und CHI) eher eine schwache Intensivierung ausgeben. Für den MMR ist in keiner Situation ein positiver Trend zu beobachten, daher wird hier auch kein Ergebnis für die FPA angezeigt. Zweigeteilt ist das Bild auch für die Änderung der negativen Trends. So zeigen die 4FM und RM überwiegend eine Intensivierung, während ZM und die FPA eher schwache Reduktionen darstellen. Auch hier sind die Unterschiede allerdings eher gering. Bei den 4FM und RM kommt es in 53,0 % bis 60,1 % zu einer Intensivierung. Für die ZM und FPA ist das Ergebnis umgekehrt. So zeigen maximal 60,1 % (OPT-15) der Situationen eine Reduktion des negativen Trends. Die Mehrheit liegt hier jedoch um 50 %. Die Temperaturergebnisse sind ebenfalls zweigeteilt. Allerdings zeigen hier 4FM und einige FPA die gleiche Tendenz. So kommt es in bis zu 61,5 % der Fälle zu einer Intensivierung. Umgekehrt finden

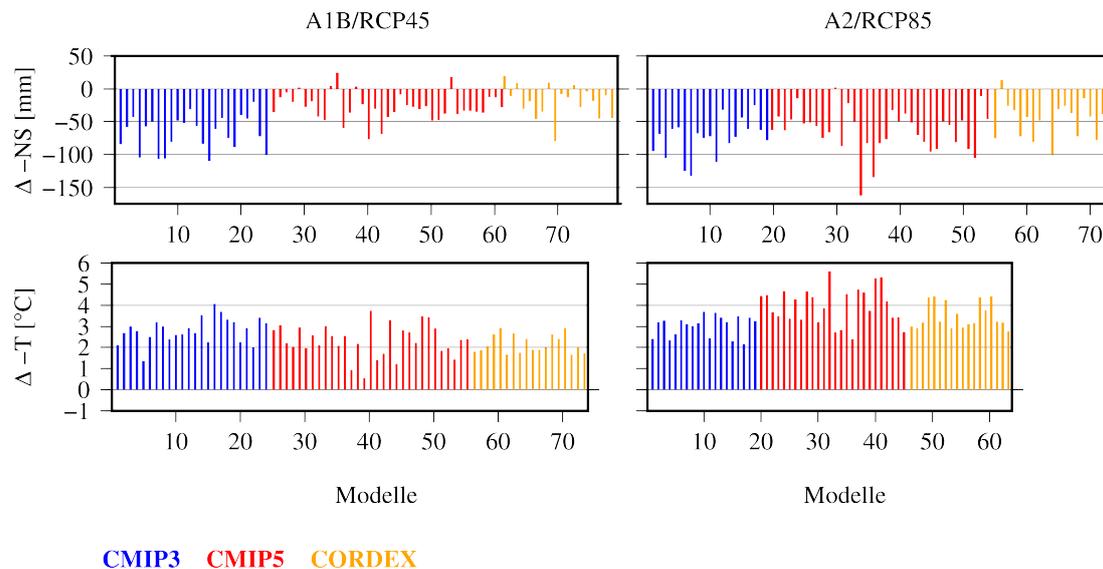
sich bei RM, ZM, SUB-Px sowie OPT-8 und OPT-15 Reduktionen der Trends von bis zu 59,0 %.

Sowohl für Niederschlag als auch Temperatur ist überwiegend eine etwas häufigere Verstärkung der Trends zu beobachten. Dieses Ergebnis darf jedoch nicht überinterpretiert werden, da die Stärke der Änderung nicht berücksichtigt wurde. Bei allen untersuchten Metriken kam es lediglich für LOR in 3,0 % der Fälle zu signifikanten Trendänderungen. Anders ist es bei Betrachtung der verstärkten Metriken ( $4FM_i$ ,  $FPA_z$ ). Hier steigt der Anteil der signifikanten Trendänderungen leicht bei gleichbleibender Art des Gewichtungseffekts. Die These einer generellen Intensivierung der Trends kann jedoch aufgrund der Unterschiede zwischen den einzelnen Metriken und untersuchten Situationen nur unter Vorbehalt aufgestellt werden. Zur Prüfung der Validität muss die Untersuchung hierzu auf weitere Regionen ausgeweitet werden, um beispielsweise auch für die FPA eine größere Datengrundlage zu schaffen. Dazu sei hier ein Verweis auf Kapitel 7 eingebracht. Dort werden alle zuvor dargestellten Untersuchungen auf die sechs großskaligen, global verteilten Regionen übertragen und vorgestellt. Im folgenden Abschnitt werden die Auswirkungen der verschiedenen Gewichtungen auf den Unsicherheitsbereich der Zukunftsprojektionen herausgearbeitet.

## 6.2 Änderung des Unsicherheitsbereichs

Dieses Kapitel widmet sich der Veränderung des Unsicherheitsbereichs der Zukunftsprojektionen. Dazu wird in jeder Region die Änderung des saisonalen und annuellen Klimas vom Ende des 20. Jahrhunderts im Vergleich zum Ende des 21. Jahrhunderts (Delta:  $\Delta$ ) untersucht. Zur Bestimmung dieses Deltas wurde für jede Simulation die Differenz zwischen dem arithmetische Mittel über die Jahre 1970–1990 und 2068–2098 berechnet. Anschließend wurde analog zur Zeitreihenanalyse (Kapitel 6.1) für jedes Modell mit mehr als einer Simulation der Mittelwert über diese berechnet. Abbildung 6.3 zeigt die daraus resultierenden Niederschlags- und Temperaturänderungen für MMR für die A1B/RCP4.5- und A2/RCP8.5-Szenarien.

Für beide Variablen bestätigt sich der aus den Abbildungen 6.1 und 6.2 gewonnene Eindruck: Nahezu alle Modelle projizieren eine Abnahme des Niederschlags und eine Zunahme der Temperatur. Dies gilt unabhängig vom Emissionsszenario. Beim Niederschlag zeigt sich unter dem A1B-Szenario eine stärkere Abnahme bis über 100mm bei den Modellen von CMIP3. Die CMIP5- und CORDEX-Modelle zeigen unter dem gemäßigeren RCP4.5-Szenario größtenteils Werte zwischen 0–50mm sowie vereinzelte Zunahmen. Unter A2 bzw. RCP8.5 bleibt die Änderung von CMIP3 relativ konstant, während CMIP5 und CORDEX verstärkte Ab-



**Abbildung 6.3** Annuelle Niederschlags- [oben] und Temperaturänderung [unten] ( $\Delta$ ) zum Ende des 21. Jahrhunderts für den Gesamtmittelmeerraum. Ange tragen sind die Werte jedes Modells für jeweils beide analysierten Emissionsszenarien A1B/A2 (CMIP3) und RCP4.5/RCP8.5 (CMIP5, CORDEX).

nahmen zeigen. Die maximale Abnahme liegt hier bei Werten von über 140mm (CMIP5), während nur noch ein Modell eine Zunahme von 10mm (CORDEX) ausgibt. Unterteilt in die vier Jahreszeiten ist das Bild relativ homogen. In MAM, JJA und SON werden über nahezu alle Simulationen Abnahmen im Bereich zwischen 0–30mm angegeben. Für DJF variiert die Niederschlagsänderung bei allen Modellen zwischen  $\pm 20$ mm.

Bei der Temperatur zeigen alle Modelle eine deutliche Erwärmung. Für A1B bzw. RCP4.5 liegt der Anstieg zwischen minimal  $0,4^{\circ}\text{C}$  und  $4,1^{\circ}\text{C}$  für MMR. Neben diesen Extremen liegt ein Großteil der Werte jedoch zwischen  $1,8^{\circ}\text{C}$  und  $3^{\circ}\text{C}$ . Die MMEs unterscheiden sich hier nur wenig. Für A2/RCP8.5 ist die Situation dagegen anders. So zeigen die CMIP3-Modelle insgesamt eine Zunahme mit Werten zwischen  $2^{\circ}\text{C}$  und  $4^{\circ}\text{C}$ . Bei CMIP5 und CORDEX fällt die Zunahme deutlich stärker ins Gewicht. Ein Großteil der Modelle zeigt eine Änderung von über  $3^{\circ}\text{C}$  bei einem Maximum von  $5,2^{\circ}\text{C}$ . Die annualen Ergebnisse decken sich mit den saisonalen Veränderungen, wobei hier insbesondere in JJA maximale Erwärmungsraten von bis zu  $7,6^{\circ}\text{C}$  projiziert werden. In den übrigen Jahreszeiten liegen die Werte dagegen für beide Szenarien ungefähr auf gleichem Niveau wie in Abbildung 6.3 exemplarisch dargestellt.

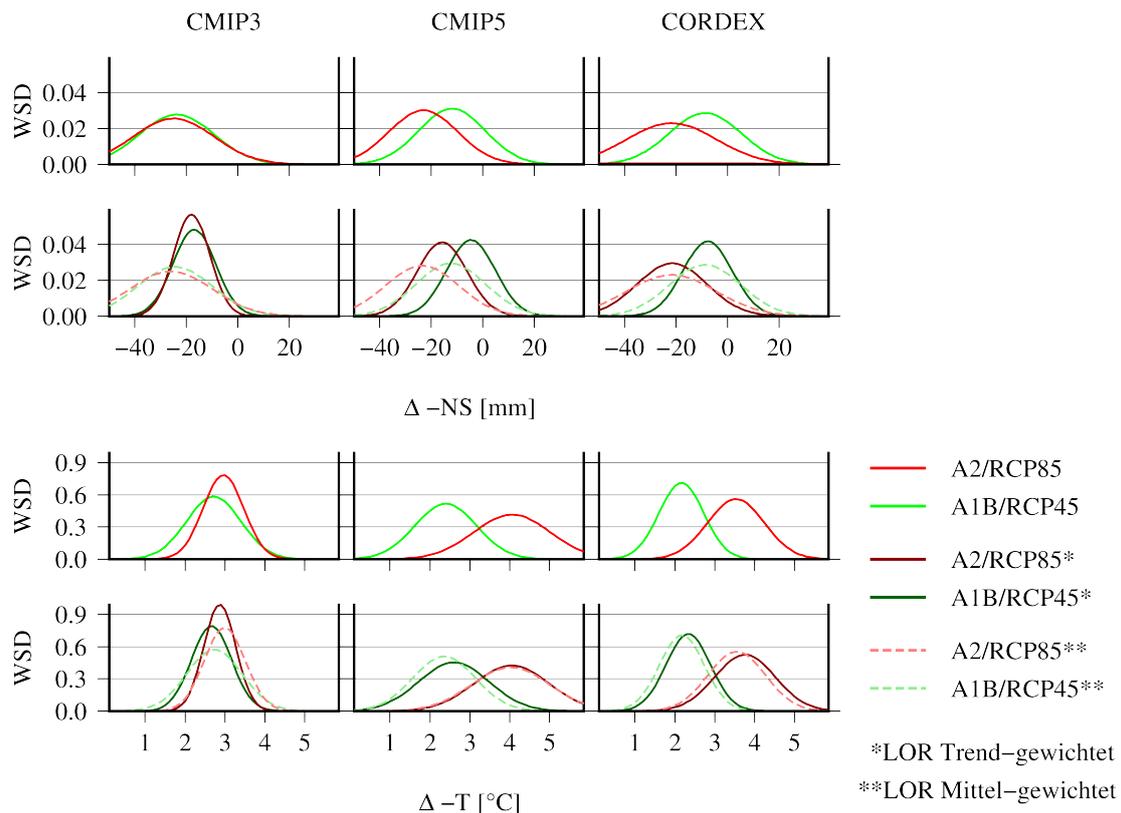
Die Unterregionen (nicht gezeigt) bilden ein überwiegend ähnliches Bild ab. So werden in den meisten Regionen für den Niederschlag Abnahmen zwischen

0–100mm angezeigt und für die Temperatur je nach Szenario gleichmäßig eine Erwärmung zwischen 2°C und 3°C bzw. 3°C und 4,5°C. Auffällige Werte bezüglich des Niederschlags sind für Spanien zu beobachten, wo mit Abstand die stärksten Abnahmen mit bis zu 200mm berechnet werden. Für A1B/RCP4.5 schwankt der Betrag etwas zwischen CMIP3 und CMIP5 bzw. CORDEX. Dagegen sind die Werte für A2/RCP8.5 gleichmäßig auf sehr hohem Niveau. Lediglich für die Regionen Balkan und das Schwarzes Meer werden deutliche positive Niederschlagsänderungen für einige Modelle mit Zunahmen von bis zu 100mm angezeigt. Hier sind allerdings Unterschiede zwischen den MMEs zu erkennen. Während die Modelle des älteren MME eher eine Mischung von leichten Zu- und Abnahmen zeigen, liegt die Mehrheit der Modelle von CMIP5 und CORDEX im positiven Bereich. Darüber hinaus zeigt sich auch hier, wie schon bei den übrigen Regionen, dass RCP8.5 zu einer Verstärkung des Signals von RCP4.5 führt. Die Unterschiede zwischen A1B und A2 sind dagegen geringer ausgeprägt. Für die Temperatur zeigen die Unterregionen des Mittelmeerraums keine nennenswerten Abweichungen vom Bild der Gesamtregion. In allen Regionen und Szenarien sind für nahezu alle Modelle deutliche Temperaturzunahmen zu beobachten, wobei sich A2/RCP8.5 immer verstärkend auswirkt. Insbesondere die südlichen Regionen wie Nordafrika oder der Nahe Osten zeigen in diesem Zusammenhang besonders hohe Werte. Darüber hinaus gilt für alle Situationen, dass für JJA die mit Abstand höchsten absoluten Erwärmungsraten projiziert werden.

### 6.2.1 Analyse einzelner Emissionsszenarien

Die absoluten Niederschlags- und Temperaturänderungen ( $\Delta$ ) zwischen dem Ende des 21. Jahrhunderts im Vergleich zum Ende des 20. Jahrhundert zeigen für viele Situationen ähnliche Ergebnisse für die meisten Modelle. Die zuvor angesprochenen Unterschiede zwischen 1,8°C und 3,0°C können jedoch große Bedeutung für die Menschen und ihre sie umgebende Natur haben. Darüber hinaus gilt, dass kleine Verschiebungen der Unsicherheit bezüglich Niederschlagsabnahme insbesondere in ariden Regionen wie Nordafrika oder den Nahen Osten umfassende Bedeutung für die Gesundheit und Nahrungsmittelversorgung haben können (IPCC, 2014a). Abbildung 6.3 hat exemplarisch für MMR gezeigt, dass, obwohl die Trends in den meisten Regionen dasselbe Vorzeichen haben (vgl. Abbildung 6.1 und 6.2), beträchtliche Unterschiede hinsichtlich des Betrags der einzelnen Modelle zu beobachten sind. Diese Unterschiede symbolisieren die MME-Unsicherheit. Um diese zu quantifizieren, wird für jedes MME eine WDF bestimmt (siehe Abschnitt 3.3). Die Unsicherheit wird anhand der Größe ihrer Standardabweichung gefolgert. Diese wird im Folgenden in metrik- und gleichgewichteter Form analysiert. Dabei wurde zunächst der KS-Test angewandt, um zu prüfen, ob die

$\Delta$ -Werte der Modelle normalverteilt sind. Dies ist in fast allen Situationen der Fall. Exemplarisch für die 45 untersuchten Situationen im Mittelmeerraum zeigt Abbildung 6.4 die WDF für die Region Ägäis. Dabei gilt es, die unterschiedlichen Skalen der Wahrscheinlichkeitsdichte auf der Ordinate von Temperatur und Niederschlag zu beachten. Bei Temperatur wird generell eine deutlich höhere Wahrscheinlichkeitsdichte und damit niedrigere Unsicherheit im Vergleich zu den WDF des Niederschlags angegeben.



**Abbildung 6.4** Wahrscheinlichkeitsdichtefunktionen der Klimaänderung aller MMEs für die Ägäis. Dargestellt sind für Niederschlag (oben) und Temperatur (unten) die jeweils gleich- (1. Zeile) und LOR-gewichteten (2. Zeile) WDF der einzelnen Emissionsszenarien in SON. Die LOR-WDF sind sowohl basierend auf den Gewichten des klimatologischen Mittels als auch auf denen des Trends angetragen.

Bei der Betrachtung der gleichgewichteten WDF des Niederschlags fällt für alle MMEs und Szenarien eine Verschiebung in den negativen Bereich auf. Dies deckt sich mit den Ergebnissen in Abbildung 6.3 für das gesamte Jahr. Dazu passend bestätigt sich insgesamt, dass CORDEX und CMIP5 stärker zwischen RCP4.5 und RCP8.5 differenzieren. Während hier RCP8.5 etwas stärker im negativen Bereich liegt, zeigen die A1B und A2 WDF zwei nahezu deckungsgleiche Kurven.

Die Unsicherheit, ausgedrückt durch die Standardabweichung der WDF, ist für alle MMEs ungefähr gleich mit einem leichten Maximum für RCP8.5 (CORDEX). Der Vergleich mit den LOR-gewichteten WDF (2. Zeile) zeigt für alle MMEs eine Verringerung der Unsicherheit. Am deutlichsten tritt dies bei CMIP3 zu Tage. Für CMIP5 und CORDEX verschiebt sich zudem das Maximum der WDF (der Erwartungswert) leicht in Richtung einer geringeren Abnahme. Diese Beobachtung der Veränderung muss allerdings als stark situationsabhängig behandelt und kann nicht verallgemeinert werden. Allerdings ist die Art der Änderung hier exemplarisch für die meisten Metriken. Dabei muss jedoch zwischen dem Trend und dem klimatologischen Mittel unterschieden werden. Der Vergleich der Evaluationsergebnisse in Kapitel 5 hat gezeigt, dass kein Zusammenhang zwischen beiden Bewertungskriterien besteht. Dies setzt sich hier auch bei der WDF-Gewichtung fort. In Abbildung 6.4 sind die Unterschiede zwischen den LOR- (klim. Mittel) und gleichgewichteten WDF nur minimal. Weder Änderungen der Unsicherheit noch des Erwartungswertes werden ersichtlich. Allerdings gilt hier wie für die trendbasierten Gewichtungen, dass Art und Stärke der Änderungen in den anderen Jahreszeiten sehr variabel sind. Insbesondere in MAM sind deutliche Abnahmen der Unsicherheit zu beobachten, wohingegen die Unsicherheit in JJA und DJF gesteigert wird.

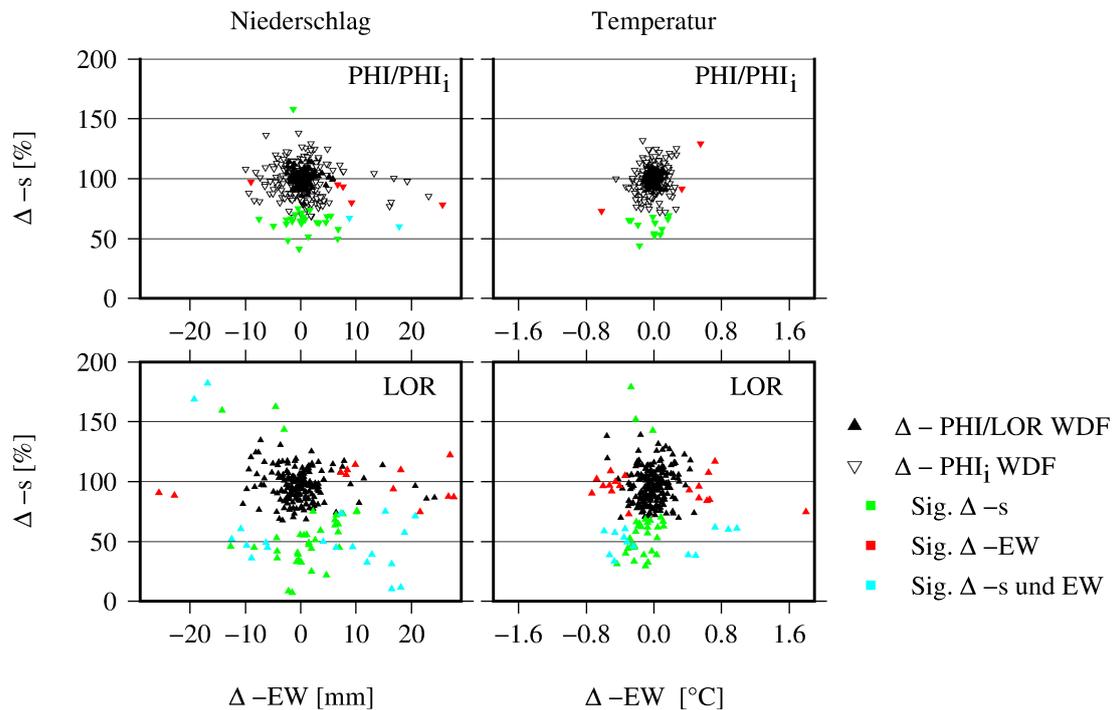
Bezogen auf die gleichgewichteten WDF ist das generelle Erscheinungsbild der Temperatur vergleichbar mit dem des Niederschlags. So sind die CMIP3-Szenarien fast deckungsgleich, während RCP4.5 und RCP8.5 eine Differenzierung bei CORDEX und CMIP5 zeigen. Anders als beim Niederschlag liegen die WDF deutlich im positiven Bereich, d.h. im Bereich einer wahrscheinlichen Temperaturzunahme. Während RCP4.5 eine Erwärmung zwischen  $1^{\circ}\text{C}$  und  $4^{\circ}\text{C}$  zeigt, liegt die Spanne von RCP8.5 zwischen  $2^{\circ}\text{C}$  und  $5,5^{\circ}\text{C}$ . A1B und A2 zeigen dagegen das Intervall von  $1,5^{\circ}\text{C}$  bis  $4^{\circ}\text{C}$ . Für alle MMEs und Szenarien wird folglich eine Erwärmung von mindestens  $1,5^{\circ}\text{C}$  projiziert. Die unterschiedlichen Szenarien und Modellgenerationen variieren lediglich im Erwartungswert und in der Standardabweichung. Der Erwartungswert schwankt im Bereich  $2,5\text{--}3^{\circ}\text{C}$  für A1B, A2, RCP4.5 (CMIP5 und CORDEX) und  $3,5\text{--}4^{\circ}\text{C}$  RCP8.5 (CMIP5 und CORDEX). Ein entscheidender Unterschied zu den WDF des Niederschlags ist die deutlich höhere Wahrscheinlichkeitsdichte bei der Temperatur. So sind die Ergebnisse hier um ein zehnfaches höher und verdeutlichen die hohe Unsicherheit der Niederschlagsprojektionen im Vergleich zur hohen Übereinstimmung der Modelle bezüglich der Temperatur. Allerdings sind die Schwankungsbereiche von teilweise bis zu  $3^{\circ}\text{C}$  nicht zu vernachlässigen. Anders als beim Niederschlag gilt, dass die LOR-gewichteten WDF in den meisten Fällen sehr dicht an der gleichgewichteten WDF liegen. Auch die Sub-Ensemble-Ansätze ( $4\text{FM}_i$ ,  $\text{FPA}_z$ ) zeigen nur geringe Effekte. Hintergrund ist die allgemein hohe Simulationsqualität, die alle

Metriken den Modellen für Temperatur attestieren (siehe Kapitel 5.3). Infolgedessen bieten die Sub-Ensemble-Ansätze hier keinen Vorteil, da kaum Modelle den Schwellenwert zur Berücksichtigung nicht erreichen. Die deutlichsten Effekte sind bei CMIP3 zu beobachten, wo beide Szenarien eine leicht verringerte Unsicherheit für die LOR-gewichte WDF (Trend) zeigen. Auch die Gewichtung basierend auf der Auswertung des klimatologischen Mittels weist keine großen Veränderungen auf. Insgesamt zeigt sich für die Temperatur bezogen auf die Standardabweichung über alle Jahreszeiten ein ähnlich stabiles Bild wie für SON. Die deutlichsten Unterschiede sind für DJF zu beobachten, wo insbesondere für beide CORDEX-Szenarien eine deutliche Reduktion der Unsicherheit dargestellt wird. Gleichzeitig bleiben hier die Änderungen von CMIP3 und CMIP5 minimal.

Generell gilt, dass die Effekte auf die WDF beim Niederschlag insbesondere von LOR und den  $FPA_z$  deutlich stärker in Erscheinung treten als bei der Temperatur. Hintergrund sind die bei der Temperatur insgesamt gleichmäßig höheren Rohgewichte bei geringen Unterschieden zwischen den Modellen. Auch wenn die überwiegende Zahl der Metriken nur geringe Anpassungen der Unsicherheitsbereiche und Erwartungswerte produzieren, zeigen zumindest die  $FPA_z$  und LOR teilweise deutliche Anpassungen der WDF. Analog zu den  $FPA_z$  bieten die Sub-Ensemble-Ansätze ( $4FM_i$ ) die größten Unterschiede beim Niederschlag. Dabei gilt, dass die  $4FM_i$  die gleichen Tendenzen wie die  $4FM$ -Gewichtungen zeigen, allerdings in intensivierter Form. Aufgrund der geringen Abstufung der RM- und ZM-Gewichte sind hier praktisch keine Unterschiede zu den gleichgewichteten WDF zu erkennen.

Die in Abbildung 6.4 dargestellte Situation (SON, Ägäis) ist exemplarisch zur Verdeutlichung der Heterogenität zwischen den unterschiedlichen MMEs sowie verschiedenen Gewichtungseffekten. Das Erscheinungsbild der WDF sowie der Gewichtungseffekte lässt sich abseits von einer Zunahme der Temperatur bzw. Abnahme des Niederschlags nicht nach einem homogenen Muster zusammenfassen. Um dennoch eine allgemeine Aussage über die Auswirkungen der Gewichtungen treffen zu können, zeigt Abbildung 6.5 die Veränderungen der WDF aller Situationen und MMEs. Eine prozentuale Verkleinerung der Standardabweichung ( $\Delta$ -s) ist dabei mit einer Reduzierung der Unsicherheit gleichzusetzen. Zur besseren Einordnung ist die 100 %-Stufe, die keine Veränderung anzeigt, durch eine horizontale Linie in der Mitte gekennzeichnet. Neben den Ergebnissen für alle Situationen basierend auf LOR ist das entsprechende Ergebnis für PHI und  $PHI_i$  angegeben.

Für PHI zeigt sich sowohl für Niederschlag als auch Temperatur ein sehr ähnliches Bild. Die Veränderungen streuen relativ gleichmäßig um das Zentrum (0 [mm/°C]; 100 %). Der Vergleich zu  $PHI_i$  lässt hier ein etwas differenzierteres Bild erkennen. Während der Großteil der Werte weiterhin um das Zentrum streuen,



**Abbildung 6.5** Übersicht über die Gewichtungseffekte auf die WDF. Dargestellt sind die Gewichtsergebnisse von PHI bzw.  $\text{PHI}_i$  (oben) und LOR (unten) für Standardabweichung und Erwartungswert der Niederschlags- (links) und Temperatur-WDF (rechts). Auf der Abszisse sind jeweils die Änderungen ( $\Delta$ -EW) des Erwartungswertes in mm (NS) bzw.  $^{\circ}\text{C}$  (T) angetragen. Die Ordinate zeigt die Änderung ( $\Delta$ -s) der Standardabweichung in Prozent an.

zeigt ein deutlicher Anteil für beide Variablen  $\Delta$ -s zwischen 50 % und 150 %. Bei der Änderung des Erwartungswertes ( $\Delta$ -EW) wirken die Unterschiede des Niederschlags mit ca. 20mm im Vergleich zu  $0,5^{\circ}\text{C}$  etwas größer. Dies ist jedoch lediglich der Achsen-Darstellung geschuldet. Diese vorliegende Darstellung wurde gewählt, da so eine relativ gleichmäßige Verteilung dargestellt und damit beste Übersicht geboten werden kann. Prozentual betrachtet sind die Unterschiede des Erwartungswertes von Temperatur und Niederschlag vergleichbar gering. Es fällt auf, dass alle signifikanten Änderungen der WDF durch  $\text{PHI}_i$  erzeugt werden. Hier ist ein Übergewicht an signifikanten  $\Delta$ -s zu erkennen, die zumeist eine Verringerung anzeigen. Signifikante  $\Delta$ -EW oder die Kombination von beidem fallen nur vereinzelt auf. Die übrigen 4FM sowie RM, ZM und FPA (nur MMR) haben keine signifikanten Ergebnisse und streuen analog zu PHI um das Zentrum (nicht gezeigt).

Einzigste Ausnahme ist LOR. Hier streuen die Ergebnisse zwar ebenfalls um das Zentrum, weisen dabei jedoch einen deutlich höheren Betrag auf. Dies gilt sowohl

für Niederschlag als auch Temperatur. Der überwiegende Teil der signifikanten  $\Delta$ -s zeigt eine Verringerung der Unsicherheit. Dieses Bild passt zu den Ergebnissen der  $4FM_i$ . Zusätzlich sind auch deutlich mehr signifikante  $\Delta$ -EW sowie die Kombination von beidem zu erkennen. Die deutlichen Übereinstimmungen bestätigen, dass die entscheidenden Unterschiede zwischen den  $4FM$  nur in der Abstufung der Gewichte zu finden sind. Alle Metriken ändern die WDF in derselben Weise, nur mit unterschiedlicher Intensität. Durch stärker ausgeprägte Gewichte ( $4FM_i$ ) kann dieser Effekt jedoch künstlich herbeigeführt werden.

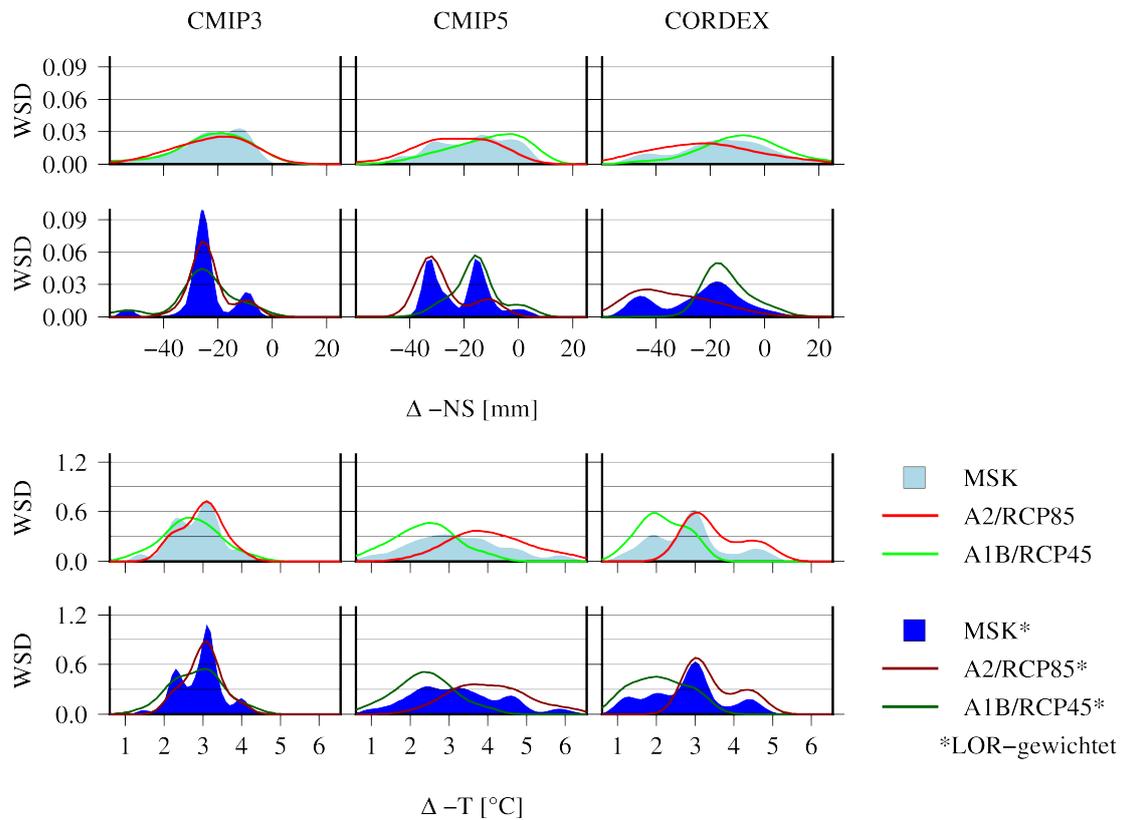
Insgesamt gilt für alle Metriken und Situationen, dass keine Verallgemeinerung über die metrikbasierte Änderung der WDF möglich ist. Die Entscheidung, ob die Unsicherheit verkleinert oder vergrößert wird, ist abhängig von Szenario, Jahreszeit, Region und Variablen. Die Stärke des Effekts auf die WDF ist dabei besonders von den Abstufungen der Gewichte abhängig, da die meisten Metriken hohe Übereinstimmungen bezogen auf die Änderungstendenz zeigen. Der Vergleich in Abbildung 6.5 zeigt, dass ein Übergewicht der Werte unterhalb von 100 % liegt, d.h. die Unsicherheit der WDF wird mehrheitlich kleiner. Diese Aussage ist jedoch in Anbetracht der zuvor beobachteten Abhängigkeit von der jeweiligen Situation nur mit äußerster Vorsicht zu interpretieren.

## 6.2.2 Multi-Szenario-Analyse

Nach der Betrachtung der WDF einzelner Szenarien wird in diesem Abschnitt der Gewichtungseffekt auf die Kombination aus jeweils beiden Emissionsszenarien der MMEs untersucht. Die verschiedenen Szenarien dienen ebenso wie die Modelle eines MME als Unsicherheitsfaktor, da das Emissionsverhalten der Menschen nicht mit Sicherheit antizipiert werden kann. Um dies zu berücksichtigen, wird nun anders als im vorangegangenen Kapitel nicht eine WDF, sondern eine Multi-Szenario-Kerndichteschätzfunktion (MSK) über beide Emissionsszenarien berechnet.

Abbildung 6.6 zeigt analog zu Abbildung 6.4 die Multi-Szenario-Kerndichteschätzfunktionen (MSK) für Ägäis und SON. Die LOR-Gewichte basieren auf der Auswertung des Trends. Zusätzlich zu den MSK sind die entsprechenden KDF der einzelnen Szenarien A1B/RCP4.5 und A2/RCP8.5 angetragen. Auf der Abszisse ist das Niederschlags- $\Delta$  in mm bzw. Temperatur- $\Delta$  in  $^{\circ}\text{C}$  angegeben. Da die Fläche unter den Kernen ebenso wie die MSK und KDF auf 1 normiert ist und die Fläche darunter der Wahrscheinlichkeit entspricht, ist ein Vergleich mit Abbildung 6.4 möglich. Analog zu den WDF ist auch hier die insgesamt höhere Wahrscheinlichkeitsdichte auf der Ordinate der Temperatur im Vergleich zu der des Niederschlags zu beachten.

Zunächst werden erneut die gleichgewichteten Ergebnisse der Niederschlagsän-



**Abbildung 6.6** Multi-Szenario-Kerndichteschätzfunktionen (MSK) der MME-Klimaänderung für SON in der Ägäis. Dargestellt sind für die Niederschlags- (oben) und Temperaturänderung (unten) die jeweils gleich- (1. Zeile) und LOR-gewichteten (2. Zeile) MSK. Zusätzlich sind die KDF der separaten Szenarien dargestellt.

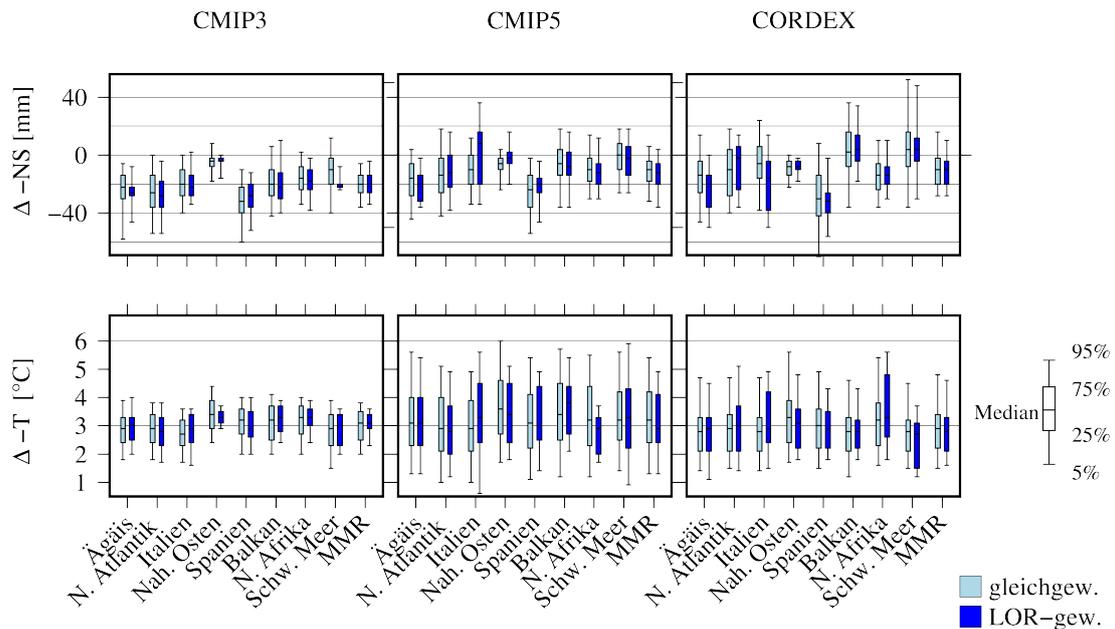
derung betrachtet. Die Berechnung des MSK hat insbesondere beim Niederschlag zu einer Verdichtung geführt. Die einzelnen Szenarien liegen eng zusammen. Die Ausprägung multipler lokaler Maxima ist bei den gleichgewichteten MMEs nur sehr schwach zu beobachten. Stärker sind die Effekte bei den LOR-gewichteten MSK von CMIP3, CMIP5 und CORDEX. Bei CMIP3 kommt es zu einer starken Verdichtung im Bereich einer 40–20mm hohen Niederschlagsabnahme. Die Werte liegen im Bereich des lokalen Maximums um das Dreifache erhöht im Vergleich zur gleichgewichteten MSK. Hintergrund ist, dass die Maxima beider Szenarien hier deutlich ausgeprägter zur Geltung kommen. Daneben fallen zwei kleine Bereiche um  $-50\text{mm}$  sowie bei  $-10\text{mm}$  fallen auf. Im Vergleich zur normalverteilten WDF (vgl. Abschnitt 6.2.1) zeigt sich, dass der Effekt durchaus konsistent ist. Allerdings ist die Stärke der Anpassung hier deutlich intensiver. Gleiches gilt für CMIP5. Auch hier wird das Muster der einzelnen WDF eingehalten. Die MSK

zeigt aber anders als bei CMIP3 zwei deutliche lokale Maxima im Bereich zwischen  $-20-0\text{mm}$  sowie zwischen  $-20$  bis  $-40\text{mm}$ . Die MSK von CORDEX ist der von CMIP5 sehr ähnlich. Allerdings sind die Maxima weniger ausgeprägt und insgesamt leicht in Richtung stärkerer Abnahmen verschoben. Für alle MMEs gilt, dass der Großteil der Modelle eine Niederschlagsabnahme mit Schwerpunkt im Intervall von  $-40-0\text{mm}$  zeigt.

Bei der Temperatur zeigt sich analog zu Abbildung 6.4 für alle MMEs und Szenarien eine deutliche Erwärmung. Die Unterschiede zwischen den Szenarien sind bei CMIP5 und CORDEX stärker ausgeprägt. Die Unterschiede bei CMIP3 sind dagegen sehr gering. Insgesamt hat die LOR-Gewichtung bei der Temperatur nur schwache Veränderungen zur Folge. Dies gilt sowohl für die einzelnen Szenarien als auch die MSK. Lediglich für CMIP3 bilden sich zwei lokale Maxima zwischen  $2-3,5^\circ\text{C}$  mit etwas höherer WSD aus. Damit fällt auch der Unterschied zu den WDF generell geringer aus als beim Niederschlag. Trotz der größeren Unterschiede der Szenarien zeigt die MSK der Temperatur eine relativ gleichmäßige Funktion. Allerdings gelten diese Ergebnisse nicht für alle Situationen. Zumeist sorgt die Trennung der Szenarien dafür, dass mindestens zwei lokale Maxima für die MSK gebildet werden. Beim Niederschlag gilt, dass sich die Szenarien häufig ähnlich verhalten. Dadurch führen auch intensive Gewichte, anders als in Abbildung 6.6 für CMIP3 und CMIP5 dargestellt, meist eher zu einer Verdichtung bzw. nur einem Maximum. Bei der Temperatur liegen die Szenarien insbesondere bei CMIP5 und CORDEX meist streng getrennt, wobei RCP8.5 eine höhere Erwärmung anzeigt. Dies hat zur Folge, dass, sobald es zu einer Gewichtung mit relativ deutlichen Unterschieden kommt, dies häufig zu klar getrennten multiplen Extrema führt.

Für Abbildung 6.6 wurde SON exemplarisch für alle Jahreszeiten dargestellt. Allerdings setzt sich der Eindruck der zuvor diskutierten Kapitel fort. So ist der Gewichtungseffekt neben den Regionen und MMEs auch von den Jahreszeiten abhängig. Hier bilden DJF und JJA die Extreme. Während in DJF für alle Regionen die stärksten Niederschlagsabnahmen angezeigt werden, zeigen die KDF für JJA maximale Temperaturzunahmen an. Die Art und Weise der KDF-Anpassung durch Gewichtung ist dabei analog zu den in Abbildung 6.5 gezeigten Ergebnissen stark situationsabhängig und kann nicht verallgemeinert werden. Auf eine Darstellung analog zu Abbildung 6.5 wird an dieser Stelle verzichtet, da die gewichteten Standardabweichungen zur Berechnung der KDF identisch mit denen in Abbildung 6.5 sind. Zur Veranschaulichung der unterschiedlichen Erwartungswerte und Gewichtungseffekte auf die generelle Streuung bietet Abbildung 6.7 einen Überblick über das 90 %-Konfidenzintervall der MSK.

Analog zu Abbildung 6.6 zeigen die Boxplots für CMIP3, CMIP5 und CORDEX die gleich- sowie LOR-gewichteten MSK für SON. Bei der Interpretation der



**Abbildung 6.7** Überblick der MSK-Konfidenzintervalle aller Regionen und MMEs. Angetragen sind die gleich- und LOR-gewichteten Konfidenzintervalle der Niederschlagsänderung  $\Delta$ -NS (oben) und Temperatur  $\Delta$ -T (unten) für SON. Neben dem Median sind die 5 %-, 25 %-, 75 %- und 95 %-Perzentile angegeben.

Darstellung gilt hier, dass ein schmalere Wertebereich der Boxplots eine Verringerung der Unsicherheit zu bedeuten hat. Beim Niederschlag zeigen nahezu alle Regionen einen überwiegenden Anteil des Konfidenzintervalls im negativen Bereich zwischen 0–60mm. Anders als bei CMIP3 zeigen CMIP5 und CORDEX einzeln gleichgewichtete Konfidenzintervalle, die über das 75 %-Perzentil hinaus positiv sind. Hier sind insbesondere Balkan und das Schwarze Meer zu nennen. Die stärksten Niederschlagsabnahmen sind für Spanien zu beobachten. Der Nahe Osten zeigt mit dem geringsten jährlichen Niederschlagsaufkommen auch die geringste Streuung und eine Abnahme im Bereich von 0–50mm. Für die übrigen Regionen verhält es sich hier wie für die Ägäis, mit einer leichten Abnahme. Dieses Muster bleibt auch für die übrigen (nicht gezeigten) Jahreszeiten konstant. Auffällig ist hier nur, dass für DJF neben dem Schwarzen Meer auch Nordatlantik und Nordafrika eine leichte Niederschlagszunahme zeigen. Im JJA ist dagegen für alle Regionen ein deutlich kleinerer Unsicherheits- sowie Wertebereich der gleichgewichteten Intervalle zu beobachten. Im ariden Nahen Osten werden beispielsweise für alle MMEs Werte um 0 angezeigt. Der Grund dürfte in der allgemein niedrigen Niederschlagsmenge der Region liegen. Nur Nordatlantik und Nordafrika zeigen wiederum deutliche Abnahmen, die über 60mm hinausreichen.

Der Unterschied zwischen gleich- zu LOR-gewichteten Konfidenzintervallen ist, wie zuvor angedeutet, sehr variabel. So ist über alle Regionen und Jahreszeiten keine eindeutige Tendenz bezüglich der Unsicherheit der Niederschlagsänderung auszumachen. Neben der Ägäis zeigen noch vier weitere Regionen bei CMIP3 für SON eine leichte Verringerung. Bei CMIP5 und CORDEX ist das Bild jedoch teilweise gegensätzlich. Auffällig ist zudem, dass sich die Wertebereiche zwischen den MMEs durch die Gewichtung nicht in allen Fällen annähern. Die fehlende Übereinstimmung zwischen den MMEs könnte mit der Region bzw. der Größe der Untersuchungsgebiete zusammenhängen. Hier könnten sich Unterschiede wie niedrige Auflösungen negativ auswirken. Für MMR zeigen alle MMEs eine ähnliche Anpassung des LOR-gewichteten MSK. So bleiben CMIP3 und CMIP5 relativ stabil, während sich das Ergebnis von CORDEX dem Bereich der anderen MMEs annähert. Auf die Bedeutung der Regionen wird in Kapitel 8.1 noch detailliert eingegangen. Für die Regionen des Mittelmeerraums gilt, dass das Intervall der MSK in den meisten Situationen durch LOR-Gewichtung teilweise deutliche Anpassungen erfährt. Neben einer leichten Verschiebung des Erwartungswertes sind die Effekte der LOR-Gewichtung auf den Unsicherheitsbereich vielversprechend. Dabei ist allerdings die Art der Verschiebung wie auch die Änderung des Unsicherheitsbereichs situationsabhängig.

Im Gegensatz zu der Niederschlagsänderung zeigt sich bei der Temperatur insgesamt ein homogeneres Bild. Für sämtliche Regionen zeigen die Konfidenzintervalle in Abbildung 6.7 einen Bereich zwischen 2–4°C. Dabei liegt der Schwerpunkt zwischen 2,5–3,4°C. Auffällig ist, dass die Ergebnisse von CMIP5 und CORDEX im Gegensatz zu CMIP3 im äußeren Bereich des Intervalls deutlich höhere Werte von bis zu 5,5°C für Nordafrika annehmen. Auch insgesamt ist die Streuung bei CMIP3 deutlich geringer. Hintergrund sind die geringeren Unterschiede zwischen A1B und A2 im Vergleich zu RCP4.5 und RCP8.5 (vgl. Abbildung 6.6). Insbesondere RCP8.5 zeigt hier deutlich intensivere Erwärmungsraten in allen Situationen. Dieses Bild ist sehr stabil über alle Jahreszeiten. Das 10 %-Perzentil liegt so in keiner Situation unterhalb von 1,8°C. Auch das 95 %-Perzentil ist abhängig vom MME meist stabil. Einzig für JJA liegt es insbesondere bei CMIP5 und CORDEX in den meisten Regionen über 5°C. Bei CMIP3 steigt es ebenfalls auf Werte zwischen 3–4°C. Mit Blick auf die LOR-gewichteten MSK ist ebenfalls ein deutlicher Unterschied zum Niederschlag auszumachen. So zeigen sich hier nahezu ausnahmslos deutliche Reduktionen der Streuung bzw. Unsicherheit. Zusätzlich ist die im Gegensatz zum Niederschlag hohe Ähnlichkeit der Mediane auffällig. Dies gilt für alle Situationen. In den meisten Regionen ähneln sich diese über die Multi-Modellgenerationen hinweg, wobei saisonale Unterschiede der Lage des Konfidenzintervalls bestehen. Allerdings ist auch hier analog zum Niederschlag festzuhalten, dass es zwischen den Jahreszeiten gewisse Unterschiede gibt und

nicht von einer generellen Reduktion der Unsicherheit ausgegangen werden kann.

Dennoch sind vor dem Hintergrund der hier aufgeführten Temperaturergebnisse die Gewichtungsergebnisse als sehr positiv zu bewerten. Die gewichteten MSK zeigen für die unterschiedlichen MMEs durchweg homogene Ergebnisse, die meist zu einer Reduktion der Unsicherheit führen. Vor dem Hintergrund der Verknüpfung verschiedener Szenarien, die in dieser Analyse mit einbezogen wurden, könnte dieses Ergebnis bei der Planung von Anpassungsstrategien hilfreich sein. So wird die Unsicherheit bezüglich des Emissionsverhaltens der Menschen berücksichtigt und dennoch eine Reduktion basierend auf der Analyse von Simulationsleistung erreicht. Dass dieser Ansatz beim Niederschlag nur punktuell zu Verbesserungen geführt hat, ist nicht als Nachteil zu werten. So ist auch die in vielen Fällen gesteigerte Unsicherheit eine wichtige Information. Insgesamt ist das Bild, das durch die LOR- bzw. gleichgewichteten MSK erzeugt wird, insofern stimmig, als dass alle eine hohe Variabilität und Unsicherheit nahe legen. Die Auswirkungen auf das Unsicherheitsverhalten der Klimaänderung machen deutlich, dass die Metrikgewichtung einen beträchtlichen Effekt insbesondere in den Unterregionen des Mittelmeerraums haben kann. Die 4FM stimmen zusammen mit RM hinsichtlich dieser Tendenz überein. Dabei hat sich LOR als die effektivste Gewichtungsmetrik herauskristallisiert. ZM besitzt im direkten Vergleich ähnlich schwache Unterscheidungen zwischen den Modellen wie beispielsweise PHI oder RM. Im Gegensatz zu den übrigen Metriken zeigt ZM jedoch nur selten Übereinstimmungen bezogen auf die Bewertungen von Modellen.

Einerseits zur weiteren Unterstützung der gewonnenen Ergebnisse, andererseits zum Test der Übertragbarkeit wurden alle Metriken und Analysen auf sechs weitere Untersuchungsräume außerhalb des Mittelmeerraums übertragen bzw. durchgeführt. Das folgende Kapitel widmet sich der Auswertung dieser sechs Untersuchungsgebiete.

## 7 Analyse auf globaler Ebene

Die Ergebnisse der Evaluation der sechs globalen Vergleichsregionen sowie der Änderung des Unsicherheitsbereichs der Zukunftsprojektionen werden in diesem Kapitel erläutert. Der Aufbau orientiert sich dabei an den Kapiteln 5 und 6. Da es sich um eine Ergänzung zur Studie des Mittelmeerraums handelt, liegt der Schwerpunkt auf der Herausstellung von Besonderheiten sowie Unterschieden zu den bisher im MMR gewonnenen Erkenntnissen.

### 7.1 Evaluation der Modelle

Die sechs global verteilten Untersuchungsgebiete (siehe Kapitel 1.4) bieten an erster Stelle die Möglichkeit, die Übertragbarkeit aller entwickelten Metriken vom Mittelmeerraum auf kontinentale Skalen zu analysieren. Zusätzlich sind nun auch weitergehende Vergleiche für die FPA möglich. Eine Einschränkung, die sich durch die neuen Regionen wie Pazifik oder auch den gesamten Globus ergibt, gilt hinsichtlich der nutzbaren Referenzdaten. E-OBS ist durch seine Beschränkung auf den Mittelmeerraum bzw. Europa ebenso wenig geeignet wie die CRU-Daten, die nur Abdeckung für Landflächen bieten. Infolgedessen wurde die Reanalyse ERA-20C als Evaluationsdatensatz gewählt. Welchen Einfluss der Austausch des Referenzdatensatzes auf die Metrikergebnisse hat, wird im Anschluss in Kapitel 8 am Beispiel der Hauptuntersuchungsregionen analysiert. Eine weitere Einschränkung, die durch die Ausweitung der Untersuchungsgebiete entsteht, ist der nötige Verzicht auf die CORDEX-Simulationen. Diese bilden in ihrer verwendeten Version ausschließlich den Mittelmeerraum ab und können daher nicht auf die weiteren Regionen übertragen werden.

Die Analyse der globalen Untersuchungsgebiete in den folgenden Kapiteln ist in drei Abschnitte unterteilt. Zunächst werden die regionalen und saisonalen Evaluationsergebnisse der Metriken herausgearbeitet. Zusätzlich wird auch hier die Bedeutung der MME-Generationen beleuchtet. Im Anschluss werden die Ranglisten der unterschiedlichen Metriken verglichen, um die Konsistenz der Evaluationsansätze auf globaler Skala zu untersuchen. Abschließend werden die Gewichte auf die Zukunftsprojektionen angewandt und deren Effekte auf das Unsicherheitsverhalten bzw. die Trends analysiert.

### 7.1.1 Analyse der globalen Gewichte

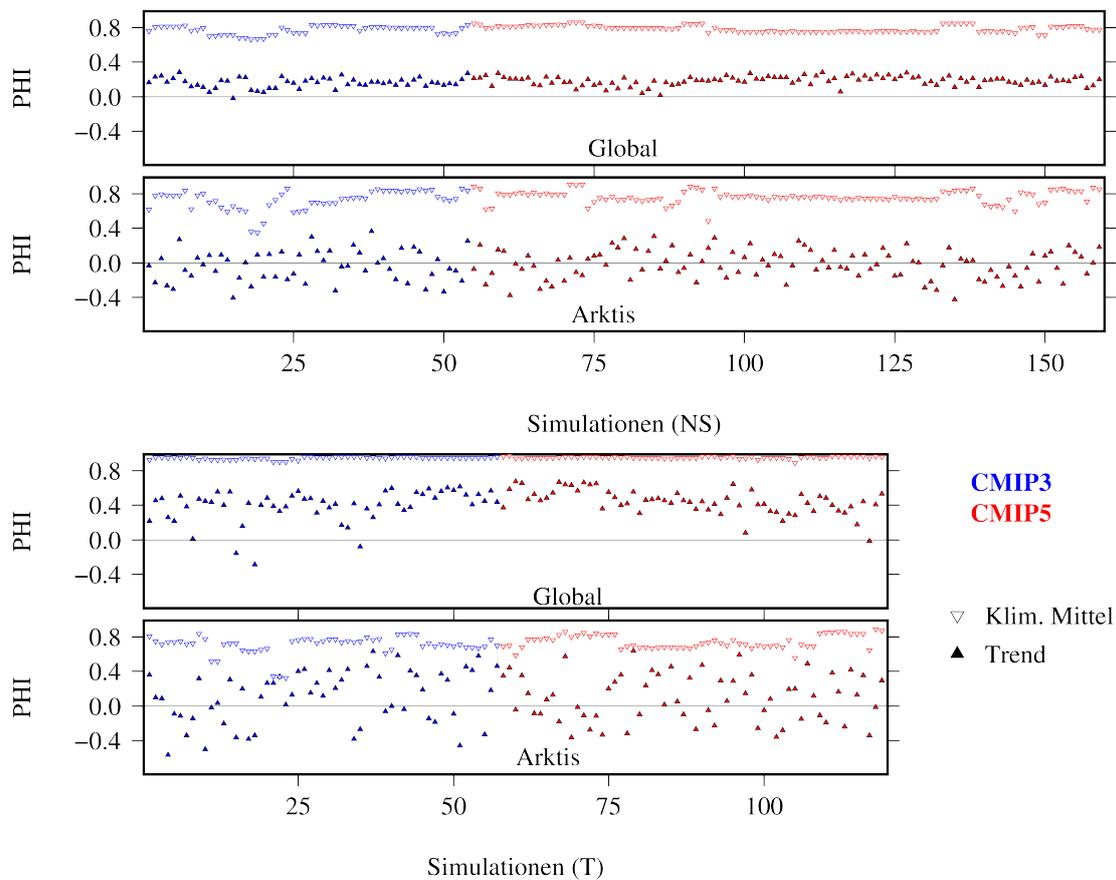
Dieses Kapitel beleuchtet die Evaluationsergebnisse aller Metriken für die globalen Untersuchungsgebiete. Analog zum Mittelmeerraum werden der Niederschlag und die Temperatur des Evaluationszeitraums 1960–2009 verwendet und basierend auf dem Trend (FPA, 4FM, RM), dem klimatologischen Mittel (4FM, RM) sowie den Eigenschaften der spektral zerlegten Zeitreihe (ZM) evaluiert. Zur besseren Vergleichsmöglichkeit wird in den folgenden Kapiteln der MMR ebenfalls aufgeführt. Die Ergebnisse hier basieren allerdings alle, wie die der globalen Regionen, auf der Verwendung von ERA-20C als Referenzdatensatz. Dies hat zur Folge, dass teilweise leichte Abweichungen bzgl. der absoluten Werte zu den Ergebnissen aus Kapitel 5 auftreten können. Für die zusammenfassende Analyse beispielsweise der Korrelation aller Ranglisten wurden jedoch ausschließlich die sechs globalen Regionen verwendet.

#### Regionale und saisonale Unterschiede

Analog zur Analyse des Mittelmeerraums wurden zur Veranschaulichung der Simulationsgüte der Globalregionen die PHI-Rohgewichte verwendet. Da nur die Simulationen von CMIP3 und CMIP5 untersucht werden, ist die Anzahl entsprechend um die 18 CORDEX-Simulationen reduziert. Für beide Variablen ist jeweils das Ergebnis des gesamten Globus und der Arktis angetragen (siehe Abbildung 7.1).

Die Muster der PHI-Rohgewichte in Abbildung 7.1 stimmen auf den ersten Blick mit denen des Mittelmeerraums überein. Dies gilt für Niederschlag und Temperatur. Es zeigt sich, dass die Ergebnisse der Trendauswertung auch hier insgesamt niedriger als für das klimatologische Mittel liegen. Dabei unterscheiden sich die Werte der sechs neuen Untersuchungsgebiete deutlich weniger als dies bei den Unterregionen des Mittelmeerraums der Fall ist. Insbesondere beim Niederschlag liegen die Werte basierend auf dem Mittel relativ gleichmäßig zwischen 0,7–0,9 gestreut. Lediglich bei den beiden ozeanischen Untersuchungsregionen sowie der Arktis (siehe Abbildung 7.1) ist das Niveau mit im Schnitt 0,65 etwas niedriger bei gleichzeitig etwas höherer Streuung zwischen den Simulationen. Die geringsten Rohgewichtsunterschiede sind für den gesamten Globus zu beobachten, wo nahezu alle Werte denselben hohen Wert besitzen.

Analog sind die Ergebnisse für Temperatur. Innerhalb der jeweiligen Regionen liegen die meisten Simulationen im Schnitt auf demselben Niveau zwischen 0,75 (Atlantik) und 0,98 (Globus). Dabei ist die Streuung zwischen den Simulationen innerhalb der jeweiligen Regionen nochmals geringer als beim Niederschlag. Besonders deutlich wird dies für Nordamerika und Global (siehe Abbildung 7.1).



**Abbildung 7.1** Annuelle PHI-Rohgewichte der Modellevaluation für Niederschlag (oben) und Temperatur (unten). Angetragen sind die Ergebnisse des klimatologischen Mittels und des Trends im Zeitraum 1960–2009 Global und für die Arktis. Über die Ordnungszahl auf der Abszisse ist die Zuordnung der Ergebnisse für jede Simulation möglich (vgl. Kapitel 2.2).

Dieser Eindruck deckt sich auch mit den nicht gezeigten 4FM. Wie bei den Ergebnissen des Mittelmeerraums wird CHI dabei vernachlässigt, da diese Metrik sich nicht zur Modellevaluation eignet. Abhängig vom jeweiligen Wertebereich liegen die Rohgewichte bei allen Metriken auf einem ähnlichen Level. Auch saisonal sind hier kaum Unterschiede zum Gesamtjahreszeitraum zu beobachten. Einzig in JJA (nicht gezeigt) werden für den Niederschlag über alle Simulationen etwas geringere Rohgewichte berechnet. Die Unterschiede zwischen den Simulationen sind hier allerdings vergleichbar. Die geringen Unterschiede bzw. die gleich hohe Simulationsgüte der Modelle lassen bereits den Schluss zu, dass das klimatologische Mittel für die globalen Regionen nicht zur Differenzierung der Modelle geeignet ist. Eine Gewichtung auf Basis der hier dargestellten Rohgewichte kommt einer Gleichgewichtung sehr nahe. Dieses Zwischenergebnis stimmt mit den Ergebnissen der

Mittelmeerregionen überein.

Auch bei der Betrachtung des 50-jährigen Trends fällt auf, dass sich die Rohgewichte der sechs Erweiterungsregionen deutlich homogener verhalten als die der Unterregionen des Mittelmeerraums. Beim Niederschlag sind die Rohgewichte der meisten Regionen relativ gleichmäßig zwischen  $\pm 0,4$  gestreut. Lediglich die globalen Rohgewichte zeigen hier eine geringere Streuung mit Werten zwischen  $0-0,4$ . Im Vergleich zu den Hauptuntersuchungsregionen wird deutlich, dass die Größe keinen besonderen Einfluss auf die generelle Simulationsqualität des Niederschlagstrends zu haben scheint. Das generelle Verhalten der Rohgewichte mit einer gleichmäßigen Streuung um 0 bleibt über alle Jahreszeiten konstant. Die einzigen Unterschiede liegen hier in den Wertebereichen der Rohgewichte. Die Analyse der Temperatur zeigt am Beispiel PHI ebenfalls ein deutlich homogeneres Verhalten der Rohgewichte als bei den Unterregionen. Auch hier sind sich die Ergebnisse der Simulationen deutlich ähnlicher und zeigen darüber hinaus für die meisten Regionen positive Werte. Lediglich in der Arktis und für den Atlantik liegt ein relevanter Anteil der Werte im negativen Bereich. Beim Pazifik hingegen wird die insgesamt geringste Streuung ausgegeben mit Werten, die gleichmäßig zwischen  $\pm 0,2$  streuen. Damit ist der Pazifik die einzige Region, in der die Untersuchung des Temperaturtrends keine besseren Ergebnisse im Vergleich zu den Niederschlagstrends ergibt. Die insgesamt überzeugendste Leistung wird dagegen für Global mit maximalen Werten um  $0,6$  und einer geringen Streuung um  $0,4$  gezeigt. Auch hier sind zwar vereinzelt negative Ausreißer zu beobachten, diese stellen mit unter 4 % der Simulationen allerdings nur eine Minderheit dar. Im Atlantik, Pazifik und der Arktis ist die Verteilung der Rohgewichte vergleichbar zu der des Gesamtmittelmeerraums. Ein ähnliches Bild wie für Global zeigt sich für Nordamerika und Afrika, allerdings auf etwas geringerem Niveau. Hier sind die absoluten Maxima  $0,7$  sowie Minima  $-0,6$  zu finden. In beiden Regionen überwiegen allerdings die positiven Werte. Generell gilt, dass das Bild der Temperaturendergebnisse über alle Jahreszeiten basierend auf dem Trend konstant ist. Dabei gibt es zwar deutliche Unterschiede zwischen den jeweiligen Simulationswerten, das Muster über alle Rohgewichte bleibt jedoch gleich.

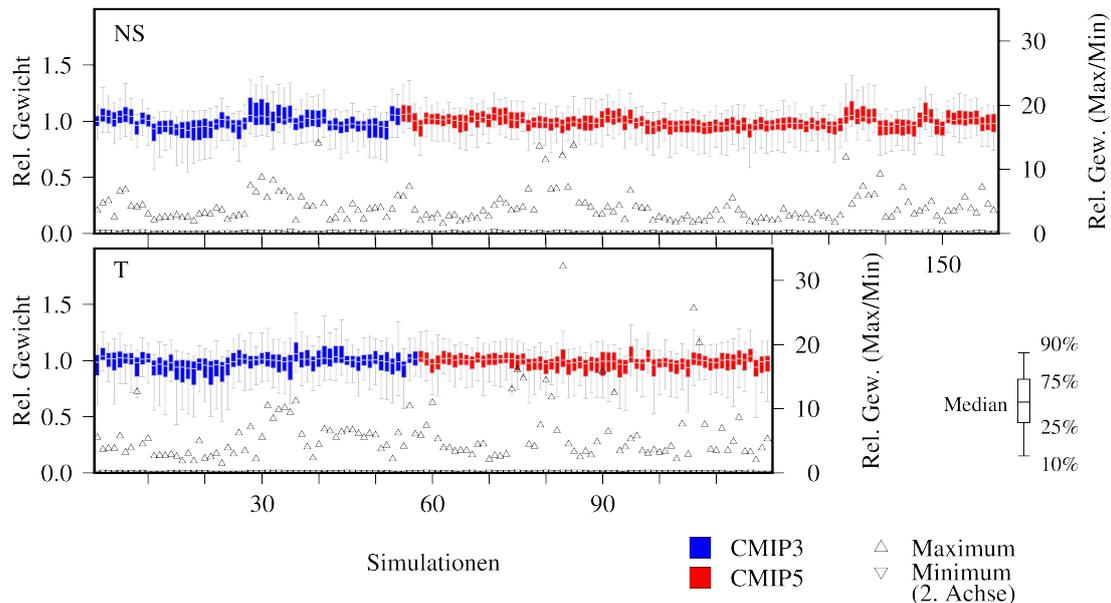
Insgesamt ähnelt das Erscheinungsbild der Evaluationsergebnisse eher dem des Gesamtmittelmeerraums als dem der Unterregionen. Die starke Streuung der Rohgewichte von Trend und Mittel in den Unterregionen ist in keiner der Erweiterungsregionen zu beobachten. Stattdessen ist eine klare Trennung erkennbar, bei der die Simulationsgüte für das Mittel durchweg über dem des Trends liegt. Es zeigt sich, dass die GKM insgesamt besser mit den Referenzdaten für größere Untersuchungsregionen übereinstimmen. Dies hat sich bereits bei der Analyse der Hauptuntersuchungsgebiete für den Gesamtmittelmeerraum angedeutet. Die Art der Region selbst kann dafür nicht als Grund angeführt werden, da dieses

Muster für alle Erweiterungsregionen gültig ist. Auch die Ergebnisse der übrigen 4FM verhalten sich analog zu denen von PHI. Dabei sind allerdings Unterschiede insbesondere durch den Wertebereich der verschiedenen Metriken zu beobachten. LOR zeigt hier wiederum die höchsten Abstufungen zwischen den einzelnen Simulationen.

RM und ZM (nicht gezeigt) besitzen ebenfalls eine hohe Übereinstimmung zu ihren Ergebnissen der Gesamtmittelmeerregion. Bei RM gilt, dass sowohl für Niederschlag als auch Temperatur die Rohgewichte basierend auf dem Trend wiederum über alle Regionen und Jahreszeiten deutlich geringer sind und damit eine kleinere Abweichung als die des klimatologischen Mittels zeigen. Da die Skalen hier allerdings auch unterschiedlich sind, gibt dies noch keine Auskunft über die generelle Simulationsgüte. Beim Vergleich der globalen Regionen mit dem Mittelmeerraum zeigt sich, dass die annuellen Trendrohgewichte des Niederschlags in den meisten Regionen mit ca. 100mm auf relativ ähnlichem Niveau liegen. Ausnahmen sind hier nur die Arktis mit ca. 20mm und der Pazifik mit 200mm. Bei den Evaluationsergebnissen der FPA fällt insgesamt auf, dass sie sich ebenfalls kohärent zu denen des Gesamtmittelmeerraums über alle Regionen verhalten. Die hohe Streuung der Werte des Niederschlags setzt sich in den meisten Regionen fort. Als Ausnahme ist hier nur die Arktis zu nennen, wo die FPA-Ergebnisse relativ gleichmäßig zwischen 0,4–0,9 variieren und damit mehr den PHI-Temperaturrohgewichten (z.B. Global) ähneln. Bei der Temperatur liegen die Werte für die Erweiterungsregionen sogar höher als für den Gesamtmittelmeerraum.

Es lässt sich zunächst folgern, dass alle Evaluationsmetriken positiv auf die größeren Untersuchungsgebiete reagieren. Von 4FM und FPA werden in den meisten Situationen höhere Werte für die globalen Modelle ausgegeben. Da die Modelle eher für die globale Klimasimulation ausgelegt sind und dies auch mit der  $2^\circ \times 2^\circ$  Auflösung besser übereinstimmt, ist diese Erkenntnis nicht überraschend. Allerdings ist auch hier, analog zum MMR, keine Simulation zu beobachten, die in allen Regionen Bestwerte erreicht. Die Variabilität der Modelle zwischen den verschiedenen Regionen bestätigt sich auch auf kontinentalem Maßstab. Allerdings führt die insgesamt höhere Simulationsgüte auch dazu, dass die Unterschiede zwischen den Rohgewichten kleiner ausfallen. Abbildung 7.2 veranschaulicht die Bedeutung der relativen 4FM-Gewichte für alle Globalregionen (insgesamt 360 Werte).

Analog zu den relativen Gewichten des Mittelmeerraums liegt der überwiegende Teil der Werte für Niederschlag und Temperatur zwischen dem 0,8–1,2-fachen Gewicht. Dagegen zeigen die Maxima ein anderes Bild im Vergleich zu den Hauptuntersuchungsregionen. Das Maximum für Niederschlag liegt bei 14,0 (40, CMIP3) und bei 25,7 (49, CMIP5) für Temperatur. Die Mehrheit der maximalen relati-



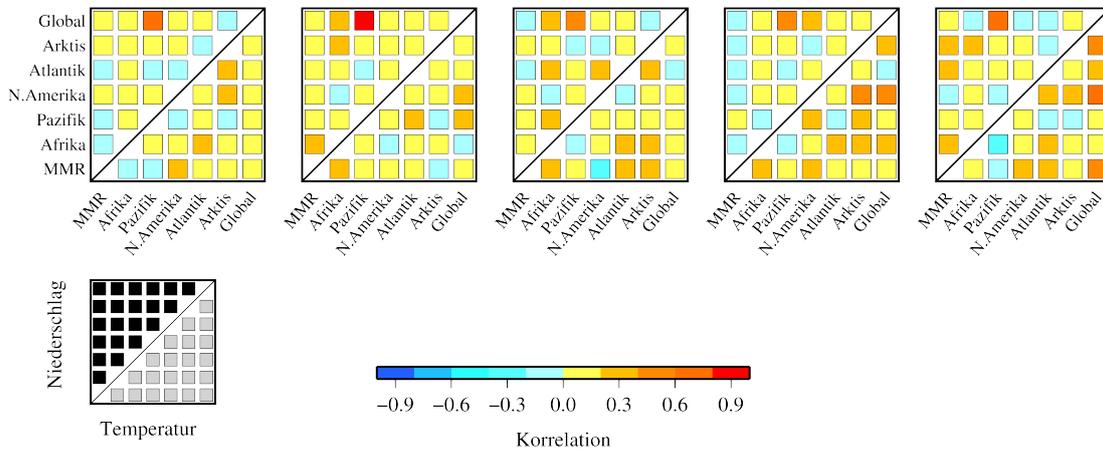
**Abbildung 7.2** Relative Gewichte der 4FM über sämtliche Situationen von Nierenschlag (oben) und Temperatur (unten) für alle sechs Globalregionen. Dargestellt sind neben den 10 %-, 25 %-, 75 %-, 90 %-Perzentilen und dem Median (linke Y-Achse) das jeweilig maximale und minimale Gewicht (rechte Y-Achse).

ven Simulationengewichte liegt im Bereich zwischen dem 1,5–15-fachen Gewicht. Wie im MMR stammen diese Gewichte von LOR. Die geringen Unterschiede der Simulationsgüte führen allerdings dazu, dass die Effekte deutlich schwächer zum Tragen kommen als dies im MMR der Fall war.

Die  $FPA_z$  (nicht gezeigt) haben für Niederschlag ähnliche Unterschiede mit nur sehr geringen relativen Gewichten. Dagegen besitzen die Verteilungen der Boxplots eine höhere Streuung der relativen Gewichte mit Werten bis zu 2,9. Im Gegensatz dazu werden für Temperatur nur geringe Unterschiede zwischen den Simulationen festgestellt, die zu einem 1,5- (34, CMIP3) bis 4,2-fachen (10, CMIP3) relativen Gewicht im Maximum und ansonsten meist zwischen 1,2–0,7 führen. Bei RM ist das Bild vergleichbar zum MMR. Hier liegt der Großteil der Werte zwischen 0,7–1,4 mit Maxima von bis zu 1,7 (22, CMIP5) und Minima bei 0,7 (z.B. 99, CMIP5). Entsprechend können auch zwischen den Simulationen nur geringe Unterschiede ausgemacht werden. Gleiches gilt auch für ZM, wo die meisten relativen Rohgewichte zwischen dem 0,6–1,5-fachen Gewicht liegen und Extremwerte zwischen 0,1 (2, CMIP5) und 3,0 (2, CMIP3) angenommen werden.

Abschließend zur regionalen Analyse der Rohgewichte wurde analog zu den Hauptuntersuchungsregionen die Spearman-Korrelation der Ranglisten zwischen

den unterschiedlichen Erweiterungsregionen berechnet (Abbildung 7.3). Es zeigt sich, dass für nahezu alle Metriken die Korrelationen merklich ansteigen. Dies gilt sowohl für Niederschlag als auch für Temperatur.



**Abbildung 7.3** Spearman-Korrelation der SUB-8-Ranglisten aller Situationen der Globalregionen und des Gesamtmitteleerraums. Angetragen sind die saisonalen Ergebnisse des Trendmusters für Niederschlag (oben-rechts) und Temperatur (unten-links).

Die SUB-8-Rohgewichte, die zuvor dahingehend noch nicht untersucht werden konnten, fügen sich nahtlos in den Kontext der anderen Metriken ein und zeigen ähnliche Werte wie die Trendresultate von 4FM und RM (nicht dargestellt). Dabei gilt beim Trend allerdings, dass die Mehrheit der Kombinationen weder für Temperatur noch Niederschlag extrem hohe Korrelationen annimmt. Die meisten Werte liegen vielmehr zwischen 0,2–0,6, wobei auch weiterhin einige negative Korrelationen (bis  $-0,2$ ) zu beobachten sind. Eine Ausnahme ist hier die Korrelation zwischen den Ranglisten des Niederschlags für Pazifik und der Globalregion. Hier werden in allen Jahreszeiten Werte zwischen 0,5–0,8 angenommen. Insbesondere bei der Temperatur überwiegen positive Korrelationen. Hier werden besonders für DJF hohe Zusammenhänge berechnet. Dieses Bild ist konsistent mit den 4FM-Trendergebnissen, wobei die einzelnen Korrelationen teilweise etwas anders ausfallen. Allerdings gilt auch für die 4FM, dass insbesondere die Ranglisten der Globalregion für Niederschlag hohe Korrelationen mit denen der übrigen Regionen zu allen Jahreszeiten zeigen. Das gilt auch für RM und ZM. Dabei sind die regionalen Ranglistenkorrelationen des Niederschlagstrends bei RM und der Temperatur bei ZM mit meist  $> 0,6$  durchweg am höchsten.

Noch höher fallen die Korrelationen bei den klimatologischen Mitteln von 4FM und RM (nicht gezeigt) aus. Während Letzteres schon für die Hauptregionen relativ hohe Werte zeigte, werden diese für die Globalregionen nochmal verstärkt, wo-

durch einige Kombinationen (beispielsweise Temperatur: Global-Arktis/Afrika; oder Niederschlag: Atlantik-Pazifik) über 0,95 liegen. Diese sehr hohen Korrelationen werden für die 4FM zwar nicht erreicht, allerdings liegen die Werte überwiegend zwischen 0,4–0,9. Das bedeutet für das klimatologische Mittel, dass sich die Spearman-Korrelationen der Ranglisten zwischen den Globalregionen für Niederschlag und Temperatur kaum unterscheiden und 4FM und RM auf ähnlich hohem Niveau liegen.

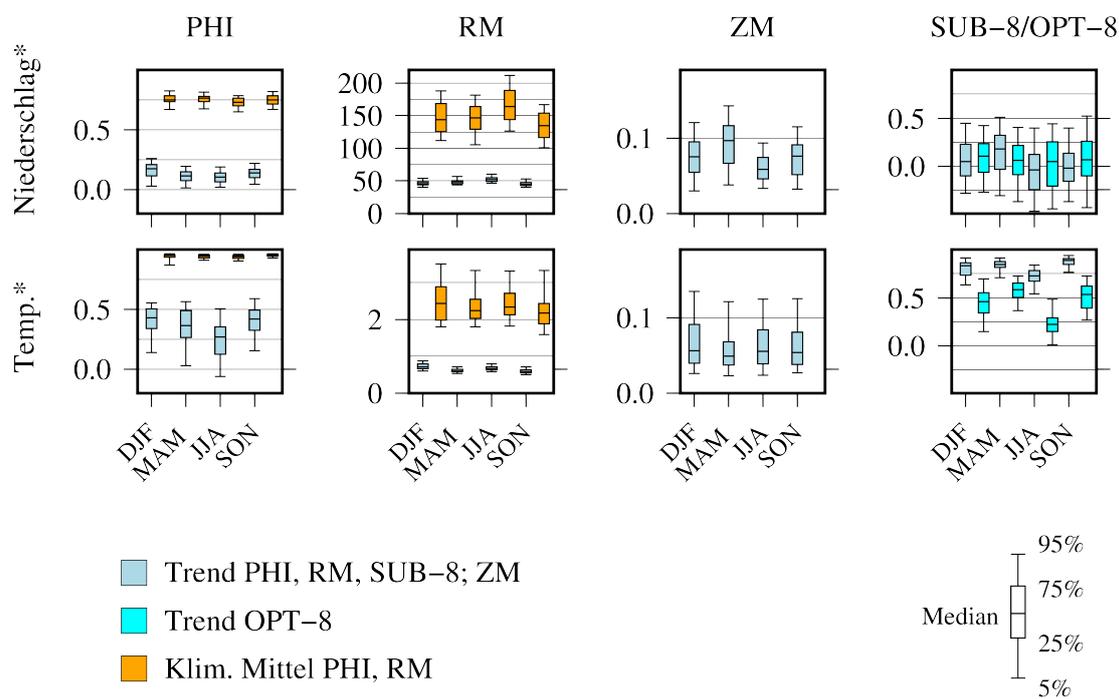
Über alle Situationen und Metriken bleibt festzuhalten, dass die Übereinstimmung der Ranglisten auf großskaliger Ebene höher als auf der regionalen Ebene des Mittelmeerraums ist. Besonders stark äußert sich dies für das klimatologische Mittel, während die Korrelationen beim Trend eher vereinzelt höher ausfallen. Auch die Ergebnisse der FPA fügen sich in die Zusammenhänge der übrigen Metriken nahtlos ein.

### Unterschiede zwischen den Jahreszeiten

Da sich die einzelnen Simulationsergebnisse der verschiedenen Jahreszeiten ähnlich verhalten wie beim MMR bzw. im Erscheinungsbild stark dem annuellen Ergebnis gleichen (vgl. Abbildung 7.1), wird im Folgenden nur die Analyse der Muster vorgenommen. Abbildung 7.4 zeigt analog zu Abbildung 5.8 die Boxplots zur Verteilung der saisonalen Rohgewichte für Global.

Zunächst zeigt sich für PHI und RM eine ähnliche Trennung zwischen den Verteilungen des klimatologischen Mittels und des Trends. Bei PHI werden sehr hohe Werte erreicht, während der RM hier eine höhere Abweichung als beim Trend misst. Diese hat sich bereits im vorangegangenen Abschnitt abgezeichnet und bestätigt sich auch bei saisonaler Betrachtung sowohl für Niederschlag als auch für Temperatur. Darüber hinaus fällt auf, dass für die Temperatur von allen Metriken bessere Ergebnisse bescheinigt werden als für den Niederschlag. Lediglich bei ZM zeigen beide Variablen sehr ähnliche Abweichungen und Streuungsmuster. Ansonsten lässt sich anders als beim MMR kein Muster zwischen den unterschiedlichen Jahreszeiten ablesen. Übertragen auf die Rohgewichte bedeutet dies, dass alle Jahreszeiten eine ähnliche Simulationsgüte zeigen. So unterscheiden sich die Verteilungen von denen der Mittelmeerregionen. Der Grund hierfür muss zum einen auf die geographische Lage der Globalregionen und zum anderen auf die allgemein höhere Simulationsgüte zurückgeführt werden. So sind auch schwächere Ergebnisse überwiegend noch auf relativ hohem Niveau in den meisten Gebieten.

In vier der sechs globalen Regionen sind daher analog zu Abbildung 7.4 keine auffälligen Unterschiede zwischen den vier Jahreszeiten zu beobachten. Anders verhält es sich nur für Nordamerika und die Arktis (nicht dargestellt). Bei Letzterer zeigt PHI für JJA beim klimatologischen Mittel des Niederschlag und der



**Abbildung 7.4** Mittlere saisonale Rohgewichte der vier Metrikgruppen für Niederschlag (oben) und Temperatur (unten) für Global. Angetragen sind die Werte von Trend (blau) und Mittel (orange) für PHI und RM sowie die Ergebnisse von ZM und den FPA-8. Die Boxplot-Verteilungen zeigen neben dem Median die 5 %-, 25 %-, 75 %- und 95 %-Perzentile.

Temperatur etwas niedrigere Ergebnisse als in den anderen Jahreszeiten. Zusätzlich steigen die PHI-Rohgewichte der Temperatur für den Trend und für RM die Streuung des klimatologischen Mittels in DJF stark an. Für Nordamerika sind Unterschiede in allen Metriken zu beobachten, die sich allerdings auch in allen Jahreszeiten unterschiedlich auf die Variablen auswirken. In den meisten Fällen sinken die Werte insgesamt und die Streuung nimmt zu. Besonders deutlich ist sie bei ZM, wo für Niederschlag die höchsten Abweichungen für SON und die niedrigsten für DJF zu beobachten sind. Gleichzeitig ist das Bild für Temperatur hier genau gegensätzlich. Der entscheidende Grund für das Muster der Jahreszeiten dürfte in der Lage und der Komplexität des vorherrschenden saisonalen Klimas zu finden sein. Die Arktis und Nordamerika liegen im Gegensatz zu den anderen globalen Regionen ausschließlich auf der Nordhalbkugel. Damit ist nur hier ein explizierter Jahresgang zu beobachten. Die Unterschiede zwischen den Jahreszeiten der anderen Regionen sind so gering, dass sie sich nicht auf die Simulationsgüte aller Modelle auswirken. Dass sich die Effekte für den MMR und Nordamerika stärker auf die saisonalen Unterschiede als für die Arktis auswirken,

liegt an den speziellen Herausforderungen der Jahreszeiten an die Modelle. So sind bei der Arktis scheinbar nur DJF und JJA als Gegenpole auszumachen, während bei Nordamerika auch die Übergangsjahreszeiten komplexe Herausforderungen an die Modelle stellen.

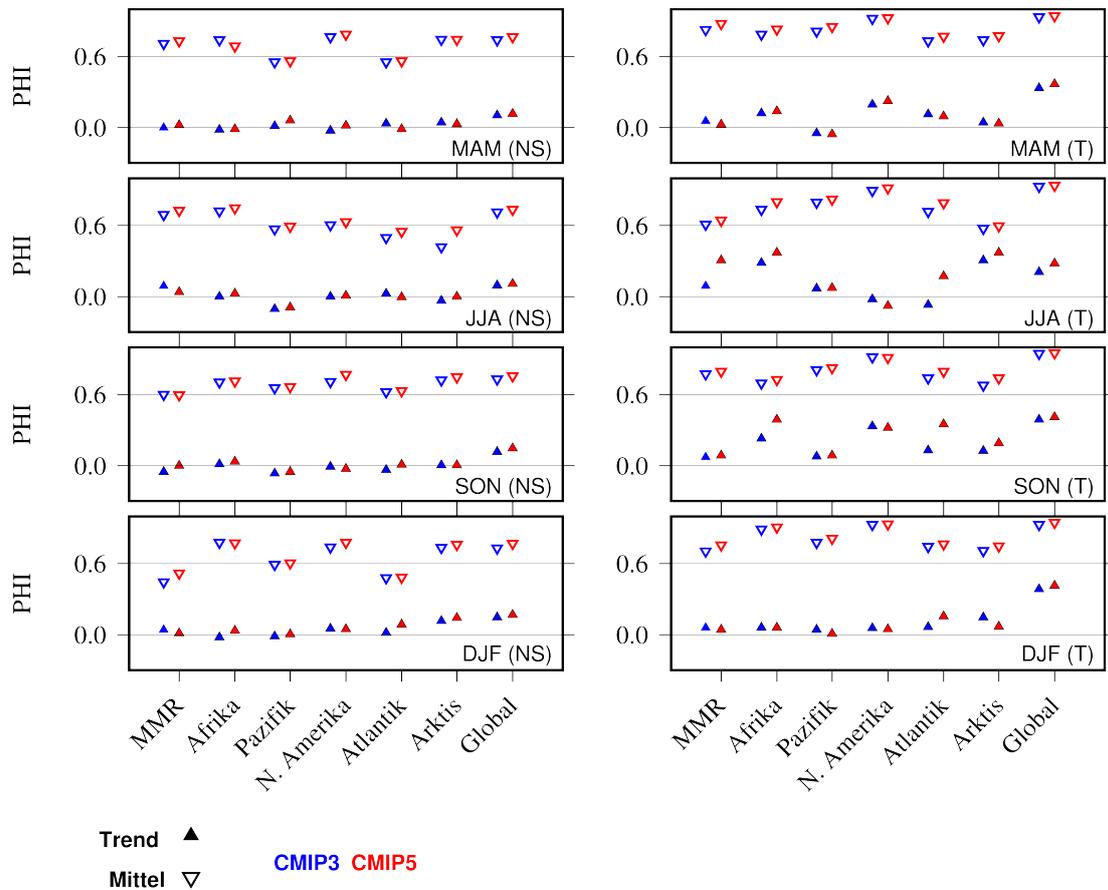
Generell gilt, dass die Jahreszeiten auch auf globaler Ebene eine Rolle bei der Bewertung der Simulationsgüte spielen. Dabei sind jedoch die Intensitäten der Jahreszeiten relevant. Regionen, die gleichmäßig um den Äquator liegen, zeigen entsprechend konstante Ergebnisse. Zusätzlich gilt, wie bei den Hauptuntersuchungsgebieten, dass keine Simulation in allen Situationen Bestwerte einer oder gar aller Metriken erreicht.

### Analyse der Modellgenerationen

In diesem Abschnitt werden die Unterschiede zwischen den MME-Evaluationsergebnissen für die Globalregionen veranschaulicht. Dabei handelt es sich, analog zur Auswertung der Mittelmeerregionen, um das arithmetische Mittel aller Rohgewichte des jeweiligen MME.

In Abbildung 7.5 sind für alle globalen Regionen und den MMR basierend auf ERA-20C-Evaluation die mittleren PHI-Rohgewichte aller Jahreszeiten angetragen. Insgesamt ist das Bild ähnlich dem des MMR. Für Niederschlag liegt das mittlere PHI-Gewicht des Trends bei beiden MMEs ca. um 0 und für das klimatologische Mittel zwischen 0,5–0,8. Ähnliche ist es bei der Temperatur, wo die Trendwerte allerdings eher im positiven Bereich zu finden sind und nur für DJF um 0 streuen. Die Werte des Mittels sind hier mit 0,6–0,95 noch etwas höher. Die insgesamt besten Ergebnisse über alle Situationen werden meist für Global berechnet. Auch die Unterschiede zwischen den beiden MMEs stimmen mit den zuvor im Mittelmeerraum gewonnenen Erkenntnissen überein. So gilt beim Niederschlag, dass für das Mittel in 91,7 % und für den Trend in 79,2 % aller Situationen CMIP5 ein höheres mittleres PHI-Gewicht zeigt als CMIP3. Bei der Temperatur sind es 66,6 % beim Trend und 95,8 % beim Mittel. Damit kann hier insgesamt eine überwiegende Verbesserung des neueren MME basierend auf PHI attestiert werden. Dieses Ergebnis deckt sich mit den übrigen 4FM.

Die FPA-Ergebnisse der Globalregionen (nicht gezeigt) sind ebenfalls mit denen des Gesamtmittelmeerraums konsistent. Die Werte des Niederschlags liegen für SUB-8 um 0 gestreut und die der Temperatur zwischen 0,6–0,98. Die OPT-8-Ergebnisse sind beim Niederschlag sehr ähnlich, zeigen für Temperatur jedoch insgesamt etwas niedrigere Werte. Durch die Verwendung der  $FPA_z$  wird das Ergebnis insbesondere für Niederschlag allerdings deutlich verbessert. Da hier nur positive Werte eingehen, liegen die Niederschlagsergebnisse zwischen 0,3–0,6, während die Temperaturwerte unverändert bleiben. Insgesamt sind CMIP3 und



**Abbildung 7.5** Vergleich der MME-Evaluationsergebnisse für die Globalregionen. Angetragen sind die gemittelten PHI-Werte aller Jahreszeiten und Regionen für Trend und klimatologisches Mittel von Niederschlag (links) und Temperatur (rechts).

CMIP5 in den meisten Fällen auf einem Level. Die geringen Unterschiede weisen jedoch beim Niederschlag beispielsweise für SUB-8 in 54,2 % der Situationen eine Verbesserung von CMIP5 aus. Bei der Temperatur sind es hier 58,3 % für SUB-8 und 54,2 % für OPT-8. Die Unterschiede fallen allerdings häufig minimal aus. Am deutlichsten ist das Bild für JJA und SON, wo nahezu alle Situationen ein besseres Ergebnis für CMIP5 zeigen.

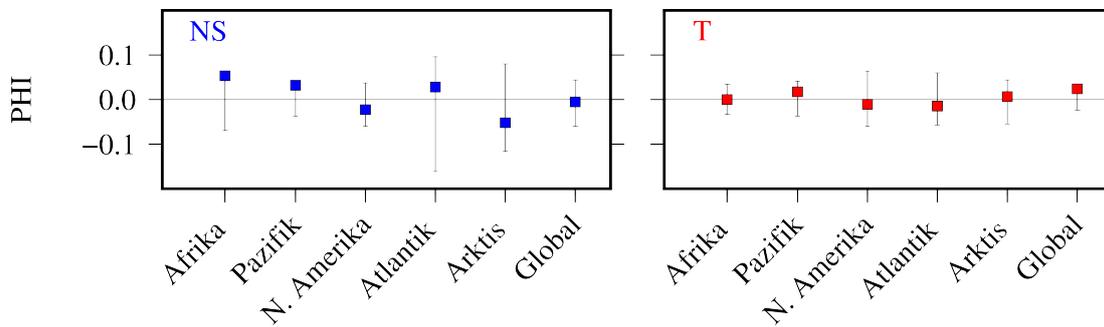
Auch für RM (nicht gezeigt) gilt, dass das Muster der mittleren MME-Rohgewichte dem des MMR entspricht. Abhängig vom Niederschlagsaufkommen sind für RM teilweise sehr hohe Werte des klimatologischen Mittels für CMIP3 und CMIP5 zu beobachten. Besonders die maritimen Regionen Atlantik und Pazifik zeigen für alle Jahreszeiten Werte zwischen 150–200mm, die auf einem ähnlichen Level liegen. Am geringsten fallen die mittleren RM-Rohgewichte für die Ark-

tis aus. Allerdings zeigt sich, dass die Werte von CMIP3 in vielen Situationen minimal besser sind als die von CMIP5. Da es sich hier jedoch meist um Unterschiede von  $< 1\%$  handelt, werden die Werte als vergleichbar gewertet. Auch beim Niederschlagstrend ist keine klare Tendenz zwischen CMIP3 und CMIP5 festzustellen bzw. die Unterschiede erscheinen zu gering, um daraus den Schluss zu ziehen, dass sich die Simulationsgüte von CMIP3 zu CMIP5 verbessert oder verschlechtert hat. Bei der Temperatur ist ein zweigeteiltes Bild zu erkennen. Während beim Trend ebenfalls kaum Unterschiede auszumachen sind und alle Werte zwischen  $0,1\text{--}1,6^\circ\text{C}$  liegen, ist bei der mittleren Temperatur eine deutliche Tendenz pro CMIP5 zu erkennen. Dabei liegen die Unterschiede zwischen den MMEs bei bis zu  $20\%$ . So zeigen die RM-Ergebnisse beispielsweise für Afrika (DJF) bei CMIP3 eine Abweichung von  $2,45^\circ\text{C}$  und bei CMIP5  $2,07^\circ\text{C}$  an. Insgesamt variieren die Unterschiede abhängig von Saison und Region. Insbesondere die Arktis ist sowohl für CMIP3 als auch CMIP5 eine hohe Herausforderung und zeigt hohe Abweichungen zu den Referenzdaten. Allerdings wird in  $87,5\%$  der Situationen ein meist deutlich besseres Ergebnis für CMIP5 ausgegeben. In den übrigen  $12,5\%$  liegen die Werte nahezu gleichauf.

Bei den ZM-Ergebnissen (nicht gezeigt) fällt auf, dass die Unterschiede zwischen CMIP3 und CMIP5 nochmals geringer ausfallen als für MMR. So weichen sie meist erst auf der dritten Nachkommastelle ab. Dies gilt besonders für den Niederschlag, trifft aber teilweise auch auf die Temperatur zu. Bei Letzterer sind die Unterschiede wieder in JJA am größten. Allerdings hat CMIP5 hier in keiner Situation einen besseren Wert als CMIP3. Insgesamt muss für die Globalregionen zusammengefasst werden, dass eine Unterscheidung der Simulationsgüte von CMIP3 und CMIP5 nur schwer möglich ist. Mit Blick auf die dazu passenden Ergebnisse in Kapitel 5.3.1 kann gefolgert werden, dass basierend auf ZM keine eindeutige Verbesserung von CMIP3 zu CMIP5 zu beobachten ist. Die schwachen Unterschiede, die sich zuvor angedeutet haben, wirken zusammen mit denen der globalen Regionen eher wie zufällige Variationen als fundierte Leistungssteigerungen. Da die übrigen Metriken ein anderes Bild zeigen, ist eine Weiterentwicklung trotz der ZM-Ergebnisse zu bestätigen. Diese wirkt sich aber offensichtlich eher auf das räumliche Muster aus als auf die Untersuchungsgrößen, die mit RM und ZM analysiert werden.

Analog zum MMR wurde ergänzend zum Vergleich der MME-Unterschiede auch die Analyse der Einflüsse des dynamischen Forcings von vulkanischen Aerosolen und Meersalzaerosolen sowie Solarvariabilität auf die CMIP3-Rohgewichte durchgeführt. Dabei hat sich das Ergebnis aus der Analyse der Hauptuntersuchungsregionen bestätigt. Auch für die globalen Regionen gilt, dass die mittleren Gewichte des Sub-Ensembles mit dynamischem Forcing nahezu immer innerhalb des Schwankungsbereichs der zufällig zusammengestellten Modellgruppen liegen

(vgl. Abbildung 7.6).

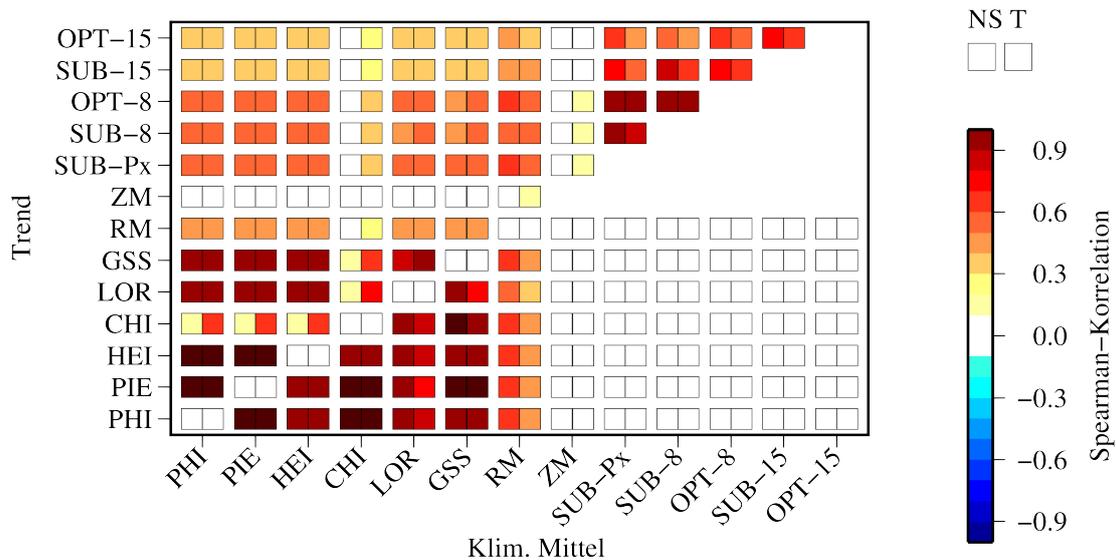


**Abbildung 7.6** Einfluss des dynamischen Forcings auf die Simulationsgüte von CMIP3 in den Globalregionen. Dargestellt sind die mittleren annualen PHI-Trendrohgewichte des Sub-Ensembles mit dynamischem Forcing für Niederschlag (blau) und Temperatur (rot). Die Balken zeigen die zweifache Standardabweichung über das Mittel 10.000-fach zufällig gezogener Rohgewichte für jede Region.

Die dargestellten annualen Ergebnisse decken sich mit denen der Jahreszeiten. Auch für die beiden rein maritimen Untersuchungsgebiete ist keine auffällige Abweichung zu erkennen. Gleiches gilt auch für die nicht gezeigten Ergebnisse des klimatologischen Mittels der RM und 4FM. Basierend auf den hier untersuchten Stichproben lassen sich keine allgemeingültigen Aussagen treffen. Allerdings geben die übereinstimmenden Ergebnisse aller Untersuchungsgebiete Anlass, das dynamische Forcing nicht als entscheidenden Grund für die Simulationsgüte anzunehmen. Dass die dynamische Berechnung verschiedener Parameter hohe Bedeutung für die Simulation des Klimas haben kann, ist naheliegend. Wie bereits in Kapitel 5.3.2 angesprochen, wird in den aktuellen MMEs (CMIP5, CORDEX) versucht, einen möglichst hohen Anteil an Parametern dynamisch zu erzeugen. Dass sich die drei untersuchten Faktoren auf keine der entwickelten Metriken positiv oder negativ auswirken, muss an dieser Stelle jedoch ebenfalls festgehalten werden.

### 7.1.2 Korrelation der Evaluationsansätze

Im diesem Abschnitt wird abschließend zur Modellevaluation der globalen Untersuchungsgebiete die Korrelation der Gewichte aller Metriken untersucht, um die Aussagen bezüglich der Übereinstimmungen und Unterschiede der Evaluationsansätze zu quantifizieren. Abbildung 7.7 zeigt die Spearman-Korrelation über alle Simulationsbewertungen und Situationen der Globalregionen für den Zeitraum 1960–2009. Das entspricht je Metrik insgesamt 1830 Werten für Niederschlag und 1650 für Temperatur.



**Abbildung 7.7** Spearman-Korrelation aller Evaluationsansätze für alle Globalregionen. Angetragen sind jeweils die Ergebnisse des Niederschlags (linkes Feld) und der Temperatur (rechtes Feld) für den Trend (oben-links) sowie, gegebenenfalls das klimatologische Mittel (rechts-unten). Die Diagonale zeigt die Korrelation zwischen Trend und Mittel für die 4FM und RM.

Das Muster in Abbildung 7.7 verdeutlicht eine sehr hohe Ähnlichkeit zu dem der Mittelmeerregion. Die hohe Korrelation zwischen den 4FM fällt sowohl für den Trend als auch das klimatologische Mittel auf. Einzige Ausnahme ist hier wieder CHI, das beim Niederschlag basierend auf dem Trend mit ca. 0,1–0,2 die kleinste Korrelation zu den übrigen 4FM zeigt. Auch bei der Temperatur sind mit 0,5–0,6 die niedrigsten Werte innerhalb dieser Metrikgruppe bei CHI zu beobachten. Dass hier Temperatur weniger starke Unterschiede aufweist, ist der allgemein sehr hohen Simulationsgüte in allen Situationen geschuldet, in der kaum negative Rohgewichte der übrigen 4FM erzeugt werden. Für die übrigen Kombinationen liegen die Korrelationen hier zwischen 0,90–0,99. Neben CHI fällt auch ZM auf. Hier werden für alle Kombinationen unabhängig von der untersuchten Größe keinerlei Übereinstimmungen angezeigt. Die RM zeigt dagegen im Vergleich zu den anderen Metriken Werte zwischen 0,22–0,64 für Trend und Mittel. Ähnlich positive Übereinstimmungen der Trendergebnisse von 4FM und RM bestehen auch mit den FPA. Diese sind erneut für die FPA-8 mit 0,51–0,58 am höchsten. Die Korrelationen zu den Ranglisten der FPA-15 sind mit 0,34–0,53 etwas niedriger. Auch die Übereinstimmungen zwischen den FPA selbst sind vergleichbar zu den Ergebnissen des Mittelmeerraums. Dabei muss jedoch beachtet werden, dass in Kapitel 5.4 nur eine Region zur Analyse verwendet wurde. Die FPA-8 und SUB-Px zeigen hier auch über sechs Regionen die höchsten Übereinstimmungen ( $\geq 0,9$ ).

Die Korrelationen der FPA-15 sind mit 0,5–0,7 etwas niedriger. Die Unterschiede zwischen den beiden Variablen sind dagegen für alle Metriken minimal.

Abschließend gilt sowohl für RM als auch für 4FM, dass keine Korrelationen zwischen den Modellbewertungen des klimatologischen Mittels und des Trends zu beobachten sind. Dieses Ergebnis ist die Bestätigung der Eindrücke im Mittelmeerraum, jedoch zeigt sich hier, dass es unabhängig von der untersuchten Region oder Größe des Gebietes gültig ist. Auch die Untersuchung der einzelnen Regionen hat sehr ähnliche Ergebnisse gezeigt. Gerade Global, Afrika und Nordamerika zeigen ein nahezu identisches Bild im Vergleich zur Analyse über alle Regionen (vgl. Abbildung 7.7). Einzig für die Arktis sind leichte Abweichungen zu beobachten. So sind die Spearman-Korrelationen hier zwischen den 4FM auf ähnlichem Niveau, zwischen den FPA und den übrigen Metriken sind die Werte jedoch nur für FPA-8 und SUB-Px  $\geq 0,7$ . Die übrigen Korrelationen liegen mit 0,1–0,25 deutlich niedriger.

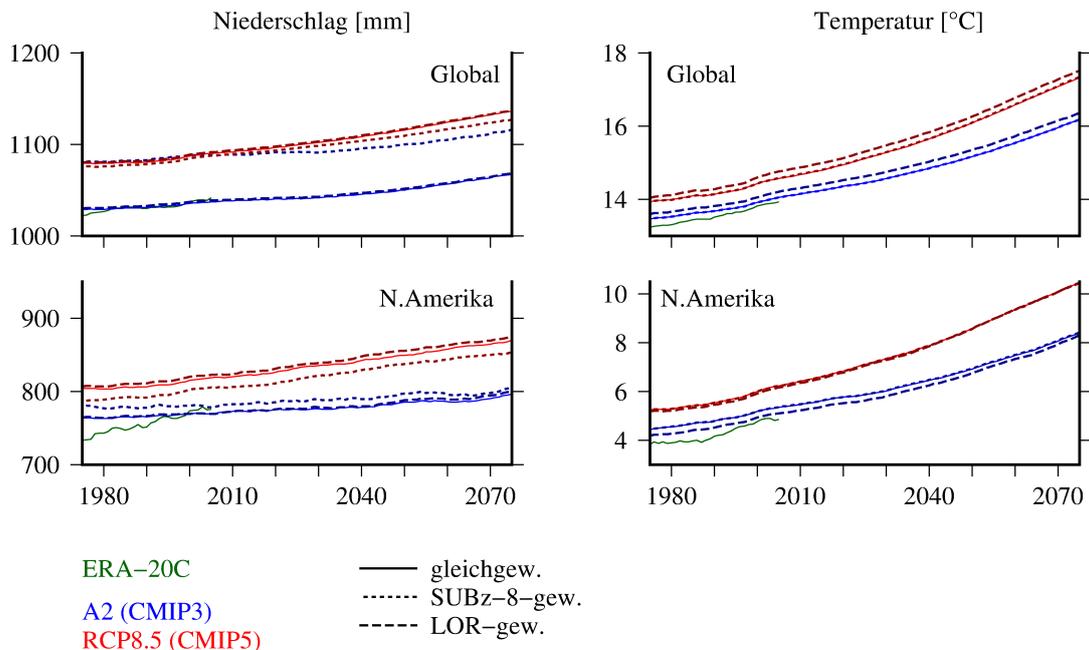
Insgesamt kann gefolgert werden, dass die Muster, die für den Mittelmeerraum berechnet wurden, sich auch für die Erweiterungsregionen bestätigen. Basierend auf dem 50-jährigen Trend zeigen die unterschiedlichen Metriken der 4FM, RM und FPA hohe Übereinstimmungen bezüglich der Bewertung von Modellen. Gleiches gilt auch für 4FM und RM bezogen auf das klimatologische Mittel. Die ZM zeigt sich ebenfalls konsistent, da ihre Rangfolgen auch für die globalen Regionen faktisch keine Korrelationen zu den übrigen Metriken zeigen. Nachdem die Rohgewichte der verschiedenen Metriken analysiert und in einen Kontext mit den Ergebnissen des MMR gestellt wurden, wird im Anschluss die Gewichtung für die globalen Regionen durchgeführt.

## 7.2 Gewichtung der Zukunftsprojektionen

Dieser Abschnitt widmet sich der Anwendung der Modellgewichte auf die vier verschiedenen Zukunftsprojektionen für die globalen Regionen. Dabei wird analog zu Kapitel 6 zunächst der Gewichtungseffekt auf die Entwicklung der Zeitreihen betrachtet. Im Anschluss daran wird die Veränderung des Unsicherheitsbereichs anhand der Wahrscheinlichkeitsdichte- bzw. Kerndichteschätzfunktionen analysiert. Da die Modellbewertungen teilweise hohe Korrelationen zeigen (siehe Abschnitt 7.1.2), werden zur besseren Übersicht in den folgenden Abbildungen jeweils nur die Ergebnisse einer Auswahl von Metriken dargestellt. Analog zum Mittelmeerraum werden die 4FM- und RM-Modellgewichte basierend auf dem 50-jährigen Trend verwendet.

### 7.2.1 Auswertung der Zeitreihengewichtung

Abbildung 7.8 zeigt die jährlichen Niederschlags- und Temperaturzeitreihen von A2 (CMIP3) und RCP8.5 (CMIP5) exemplarisch für den gesamten Globus sowie Nordamerika. Die aus Gründen der Übersichtlichkeit nicht gezeigten A1B- sowie RCP4.5-Zeitreihen werden im Folgenden allerdings ebenfalls beschrieben. Wie im vorangegangenen Kapitel gezeigt, haben insbesondere FPA-8 eine hohe Korrelation mit den 4FM und mit RM. Die Korrelationen zwischen den Rangfolgen von FPA-8 und  $FPA_z$ -8 liegen zumeist mit 0,7–0,8 auf einem vergleichbaren Niveau wie im Mittelmeerraum. Nur für Atlantik und Pazifik gilt, dass die Korrelationen der Ranglisten von CMIP3 mit  $< 0,3$  niedriger ausfallen. Da die FPA-Gewichte, ähnlich wie beispielsweise PHI oder RM, nur schwer erkennbare Auswirkungen auf die MME haben, zeigt Abbildung 7.8 die SUB $_z$ -8- sowie LOR-gewichteten Zeitreihen. Für die Zeitreihen wurde ein 11-jähriges gleitendes Mittel berechnet.



**Abbildung 7.8** Vergleich der LOR- und SUB-8-gewichteten annualen Zeitreihen von A2 und RCP8.5. Dargestellt sind die Werte für Global (oben) sowie Nordamerika (unten) von Niederschlag in mm (links) und Temperatur in °C (rechts). Zusätzlich ist die Zeitreihe des Referenzdatensatzes ERA-20C ange tragen.

Der geringste Niederschlag wird für die Arktis mit ca. 310mm simuliert. Die Beträge von Afrika, Nordamerika und Atlantik liegen alle zwischen 650–800mm. Die höchsten Niederschläge mit über 1050mm werden Global und bei Pazifik simuliert. Beide Regionen in Abbildung 7.8 zeigen eine deutliche Temperatur-

und Niederschlagszunahme. Dieses Bild ist exemplarisch für die globalen Regionen und steht im Gegensatz zu den überwiegenden Niederschlagsabnahmen des Mittelmeerraums. Ausnahmen stellen Afrika und der Atlantik dar, die eine Stagnation bzw. schwache Abnahme der Niederschlagsmenge zeigen. Während sich A1B und A2 in den meisten Situationen nahezu identisch verhalten, geht die Schere zwischen RCP4.5 und RCP8.5 ab der zweiten Hälfte des 21. Jahrhunderts deutlich auseinander. Dabei ist die Niederschlagszunahme für RCP8.5 intensiver als für RCP4.5. Dieses Bild ist auffällig, da im Mittelmeerraum RCP8.5 zu einer verstärkten Abnahme tendiert. Als Konsequenz daraus kann gefolgert werden, dass die Niederschlagsentwicklung im RCP8.5-Szenario für alle untersuchten Regionen generell zu einer Verstärkung des bestehenden Trends von RCP4.5 führt. Dies setzt sich auch in den Jahreszeiten fort. Beispielsweise liegen die Unterschiede zwischen den CMIP5-Szenarien für DJF in Nordamerika oder der Arktis bei bis zu 40 % bzw. 60 % für das Ende des 21. Jahrhunderts. Auch die übrigen Jahreszeiten zeigen ein ähnliches Bild. Des Weiteren gilt, dass es in fast allen Regionen einen deutlichen Unterschied zwischen CMIP5 und CMIP3 bezüglich der generellen Niederschlagsmenge gibt. Diese Differenz bleibt auch für die gesamte Zeitreihe des 21. Jahrhunderts relativ konstant. Einzig in Afrika zeigen CMIP3 und CMIP5 nahezu identische Niederschlagsbeträge.

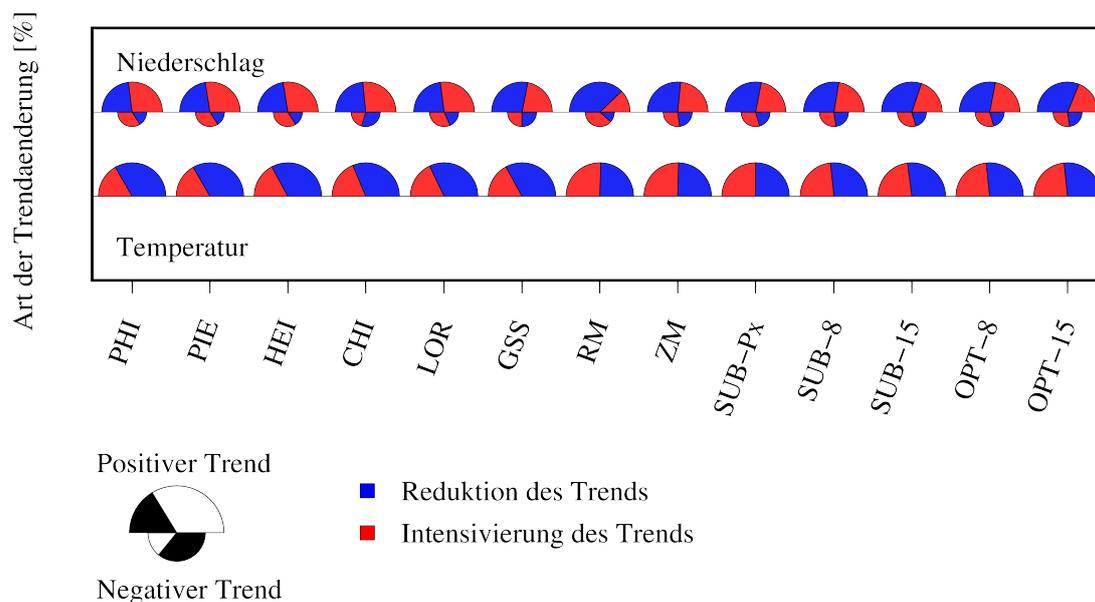
Die Unterschiede zwischen den gleich- und den metrikgewichteten Niederschlagszeitreihen wirken in Abbildung 7.8 etwas kleiner als dies im Mittelmeerraum der Fall war. Allerdings gilt zu beachten, dass das allgemeine Niederschlagsniveau hier deutlich höher ist und so auch prozentual geringe Unterschiede höhere absolute Werte zur Folge haben. Die in Abbildung 7.8 gezeigten Situationen veranschaulichen die deutlichsten Formen der Anpassung. Insbesondere der A2-Niederschlag (CMIP3) nimmt durch die  $SUB_z$ -8-Gewichtung deutlich zu und liegt auf einem Level mit RCP8.5 (CMIP5). RCP8.5 ändert sich durch die Gewichtung dagegen kaum. Für Nordamerika wird eine leichte Abnahme von RCP8.5 und eine Zunahme von A2 erzeugt. In beiden Fällen nähern sich die Kurven an. Dieses Bild kann jedoch nicht verallgemeinert werden. Beispielsweise ist für Afrika ein gegensätzliches Bild zu beobachten. Hier zeigen sich vereinzelt auch deutliche Unterschiede zwischen  $SUB_z$ -8 ( $FPA_z$ ) und LOR (4FM). Die LOR-gewichteten Zeitreihen des Niederschlags zeigen hier kaum Unterschiede zur Gleichgewichtung. Dies gilt insbesondere für CMIP3. Allgemein haben die Gewichtungen den stärksten Effekt auf das Niederschlagsniveau. Signifikante Veränderungen des Trends sind wie im Mittelmeerraum selten zu beobachten. Im Anschluss an die Analyse der Temperaturzeitreihen werden die Trendänderungen ausführlich dargestellt.

Gleichmäßiger als die Niederschläge verhalten sich die Temperaturzeitreihen und zeigen eine deutliche Erwärmung für alle Situationen und Szenarien. Die in Abbildung 7.8 gezeigten Szenarien A2 und RCP8.5 simulieren für Global und

Nordamerika eine Erwärmung um bis zu  $4^{\circ}\text{C}$  zum Ende des 21. Jahrhunderts. Dabei sind die Werte von A2 immer etwas niedriger als von RCP8.5. Dieser Effekt ist in allen Regionen zu beobachten, allerdings unterschiedlich stark ausgeprägt. Die wärmsten Regionen mit ca.  $22\text{--}25^{\circ}\text{C}$  sind Atlantik, Pazifik und Afrika. Die globale Mitteltemperatur liegt mit  $13\text{--}15^{\circ}\text{C}$  im Wertebereich des Gesamtmitteleerraums. Die mit Abstand kälteste Region ist die Arktis mit unter  $-15^{\circ}\text{C}$ . Auch die Stärke der Temperaturzunahmen ist abhängig von der jeweiligen Region. Das Maximum wird für die Arktis mit bis zu  $9^{\circ}\text{C}$  Erwärmung zum Ende des 21. Jahrhunderts simuliert. Analog zum Niederschlag zeigt sich auch für die Temperatur ein Unterschied zwischen jeweiligen Szenarien von CMIP3 und CMIP5. A1B und A2 sind sich deutlich ähnlicher als RCP4.5 und RCP8.5. Bei Letzterem tritt ab Mitte des 21. Jahrhundert eine deutliche Schere der Zeitreihen auf. Das gemäßigte RCP4.5-Szenario stabilisiert sich anschließend, während RCP8.5 weiter monoton ansteigt. Auch hier gilt, dass das Bild für alle Jahreszeiten im Kern identisch ist. Lediglich die Ausgangstemperaturen und der Betrag der Erwärmung variieren. Alles in allem sind die gezeigten Zeitreihen sowohl hinsichtlich der Unterschiede der MME-Szenarien als auch der allgemeinen positiven Temperaturentwicklung bis zum Ende des 21. Jahrhunderts mit den Subregionen des Mittelmeerraums konsistent.

Auch die Unterschiede der gleich- und metrikgewichteten Zeitreihen ähneln grundsätzlich denen des MMR. Über alle Regionen gibt es leichte Abweichungen im Betrag, die sich situationsabhängig in positiver oder negativer Richtung äußern. Diese Variabilität ist auch in den vier Jahreszeiten zu beobachten. Da die Ergebnisse der Simulationsgüte für die Temperatur sehr gute Resultate gezeigt haben (vgl. Abschnitt 7.1), waren umfassende Änderungen der Zeitreihen nicht zu erwarten. Die SUB<sub>z</sub>-8 Zeitreihen (Abbildung 7.8) sind nahezu deckungsgleich mit den Ausgangskurven. Generell sind bei der Temperatur nur durch LOR-Gewichtung sichtbare Änderungen der Zeitreihen in den Globalregionen zu erzeugen. Allerdings gilt auch hier, dass die Art und Stärke von der jeweiligen Situation abhängig ist.

Um die Effekte der Gewichtungen zu analysieren, wird analog zum MMR die Trendänderungstendenz aller Metriken veranschaulicht. In Abbildung 7.9 sind die Trendänderungen in Form von Kreisdiagrammen dargestellt. Analog zum Mittelmeerraum gilt, dass die Änderungen aller Metriken unabhängig von der Stärke der Effekte betrachtet werden. Oberhalb der horizontalen Linie sind die Veränderungen der positiven und unterhalb der negativen Ausgangstrends angetragen. Die Gesamtkreisfläche entspricht damit 100 % der Trends (hier 120). Wie für MMR sind alle Ausgangstemperaturtrends positiv. Dagegen fällt auf, dass alle 4FM mit bis zu 64,2 % (LOR) eine überwiegende Reduktion des positiven Trends zeigen. Auch die FPA zeigen keine Intensivierung, sondern einen leicht höheren Anteil



**Abbildung 7.9** Effekte der metrikgewichteten Zeitreihen auf den Trend der Zukunftsprojektionen in den Globalregionen. Dargestellt für die jeweilige Metrik ist der Anteil der Trendänderung für Niederschlag (oben) und Temperatur (unten). Die Fläche jedes Diagramms ist auf 100 % normiert und teilt sich auf den relativen Anteil an positiven (oberer Halbkreis) und negativen Trends (unterer Halbkreis) auf.

an Reduktionen. Bei RM- und ZM-Gewichtungen halten sich beide Entwicklungen die Waage. Ein zweigeteiltes Bild ist beim Niederschlag zu beobachten. Der überwiegende Teil zeigt positive Trends. Niederschlagsabnahmen wurden nur für den Atlantik und Afrika simuliert. Dabei zeigen bis auf GSS und CHI alle Metriken eine überwiegende Intensivierung. Bei den Niederschlagszunahmen gibt es hingegen zwei größere Gruppen. Die 4FM zeigen hier bis auf GSS alle eine leicht überwiegende Intensivierung, die jedoch maximal 54,1 % (LOR) annimmt. Die übrigen Metriken erzeugen dagegen häufiger verringerte Niederschlagsabnahmen. Am deutlichsten trifft dies auf RM mit 75,5 % zu.

Die Auswirkungen der Gewichtungen auf die Zeitreihen sind in höchstem Maße heterogen. Dass sogar innerhalb der 4FM neben CHI mit GSS eine weitere Metrik eine andere Tendenz als die übrigen Ansätze zeigt, legt nahe, dass aus diesen Ergebnissen keine Interpretationen zulässig sind. Wie schon für den Mittelmeerraum muss die Stärke der Trendänderung hier ebenfalls berücksichtigt werden. Insgesamt wurden nur maximal 12 signifikante Trendänderungen (LOR) erzeugt. Dies entspricht einem Anteil von 5,0 % aller gewichteten Zeitreihen. Die Ausgangstrends selbst zeigen eine sehr hohe Übereinstimmung zwischen den MMEs. Die Tendenzen sind sehr stabil und können daher über alle untersuchten Regio-

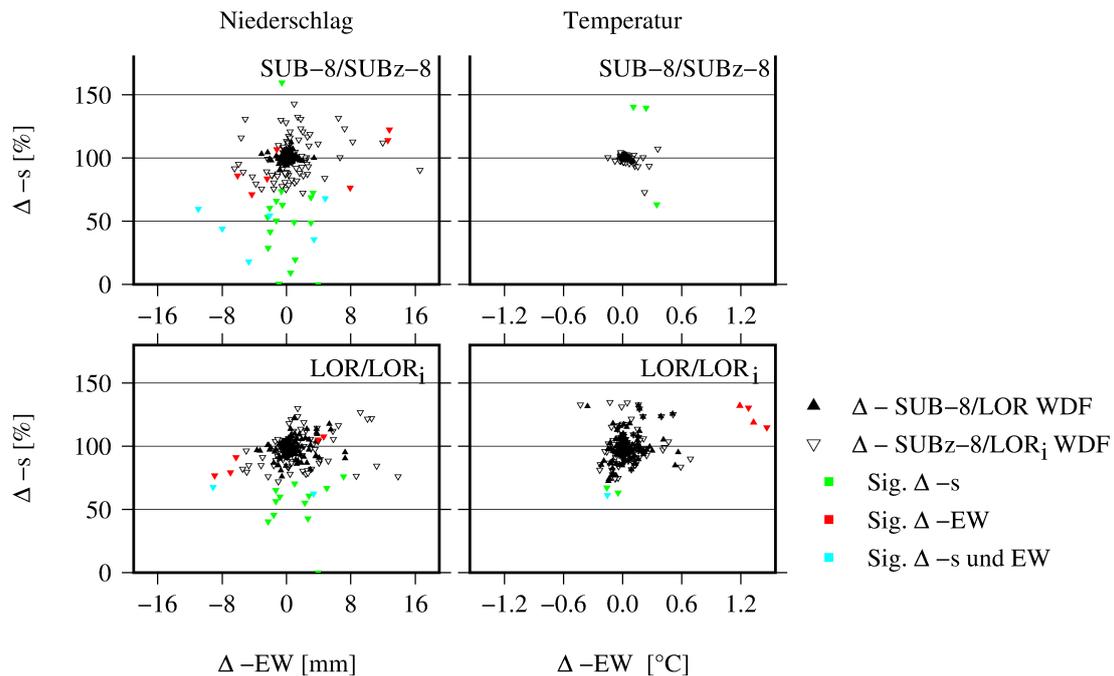
nen und Jahreszeiten als zuverlässig angenommen werden. Die Gewichtung der Zeitreihen kann hierzu keinen weiteren Beitrag leisten. Die beobachteten Effekte zeigen für die Globalregionen wie für den Mittelmeerraum ein heterogenes Bild. Daher sind allgemeine Aussagen bezüglich Verstärkung oder Reduktion bestehender Trends nicht möglich. Stattdessen muss für jede spezielle Situation eine Auswertung vorgenommen werden, um fundierte Informationen liefern zu können.

### 7.2.2 Änderung des Unsicherheitsbereichs

Dieser Abschnitt widmet sich abschließend der Auswertung der Änderung des Unsicherheitsbereichs für die Globalregionen. Dabei wird analog zu Kapitel 6.2 die Veränderung des Klimawandels basierend auf den metrikgewichteten Zukunftsprojektionen von CMIP3 und CMIP5 analysiert. Hierzu wird für jedes Modell in allen Erweiterungsregionen die Änderung von Niederschlag und Temperatur vom Ende des 20. zum Ende 21. Jahrhunderts berechnet und die Differenz im Folgenden als Klimaänderung bezeichnet. Die Auswertung der Ergebnisse wurde analog zu Kapitel 6.2 für die globalen Regionen durchgeführt, um die beste Vergleichbarkeit zu gewährleisten. Wie für den Mittelmeerraum wurde mit dem KS-Test zunächst die Normalverteilung der Werte geprüft. Dies trifft auf alle Situationen zu.

Auf eine exemplarische Darstellung einzelner WDF wird in diesem Abschnitt verzichtet, da die Art der Klimaänderung bereits durch die Analyse der Zeitreihen deutlich wurde. So ist für jedes Untersuchungsgebiet eine intensive Temperatur- und bis auf Afrika und Atlantik auch eine Niederschlagszunahme zu beobachten. Die Unterschiede zwischen den beiden Emissionsszenarien fallen in den Globalregionen bei CMIP3 ebenfalls moderat aus. A1B und A2 zeigen meist eine hohe Schnittmenge, die sich durch überlagerte WDF darstellt. Bei RCP4.5 und RCP8.5 ist dagegen analog zu MMR eine deutlichere Trennung erkennbar, wobei Letzteres immer zu einer Intensivierung des Trends von RCP4.5 tendiert. Generell ist die Übereinstimmung zwischen den Modellen auch bei der Temperatur höher als beim Niederschlag. Dies führt zu höheren Wahrscheinlichkeitsdichtewerten bei der Temperatur. Insgesamt zeigen die WDF in den sechs Regionen für alle Situationen ein einheitlicheres Bild als im Mittelmeerraum. Die Effekte der Metrikgewichtungen fallen hingegen deutlich geringer aus. Abbildung 7.10 fasst alle Gewichtungsergebnisse der WDF zusammen. Dazu sind analog zu Abbildung 6.5 die prozentualen Veränderungen des Erwartungswerts  $\Delta$ -EW und der Standardabweichung  $\Delta$ -s für jede einzelne WDF dargestellt.

Die WDF-Änderungen zeigen im Vergleich zur Auswertung des Mittelmeerraums für die Erweiterungsregionen ein sehr ähnliches Bild. Weder für Temperatur noch für Niederschlag haben die metrikbasierten Gewichtungen auf den



**Abbildung 7.10** Übersicht der Gewichtungseffekte auf die WDF der Globalregionen. Dargestellt sind die Gewichtsergebnisse von SUB-8 bzw. SUB<sub>z</sub>-8 (oben) und LOR (unten) auf Standardabweichung und Erwartungswert der Niederschlags- (links) und Temperatur-WDF (rechts). Auf der Abszisse sind jeweils die Änderungen ( $\Delta$ -EW) des Erwartungswertes in mm (NS) bzw. °C (T) angetragen. Die Ordinate zeigt die Änderung ( $\Delta$ -s) der Standardabweichung in Prozent an.

Erwartungswert einen eindeutigen Effekt. Stattdessen ist die jeweilige Situation dafür entscheidend, welche Art von Änderungen durch die Gewichtung hervorgerufen wird. Beim Niederschlag werden die meisten signifikanten Änderungen der Standardabweichung basierend auf SUB<sub>z</sub>-8-Gewichtung erreicht. Es fällt auf, dass LOR schwächere Effekte erzeugt. Signifikante Änderungen sind hier im Gegensatz zu SUB<sub>z</sub>-8 nicht zu beobachten. Durch Verwendung des LOR<sub>i</sub> Sub-Ensemble-Ansatzes können die Gewichte verstärkt werden, um vergleichbare Ergebnisse zu generieren. Ein ähnliches Bild ergibt sich auch für die übrigen 4FM<sub>i</sub> sowie FPA<sub>z</sub>. Das stärkste  $\Delta$ -s wird zweimal für Pazifik (A2) erzeugt, wo die Standardabweichung auf 0 % reduziert wird. Hier hat lediglich ein Modell ein gültiges Gewicht zugeteilt bekommen. In diesem speziellen Fall kann nicht von einer sinnvollen Reduzierung der Unsicherheit ausgegangen werden. Vielmehr gilt, dass hier eine allgemein niedrige Simulationsgüte der Modelle bezüglich der beiden Situationen besteht. Im Schnitt liegen für alle Metriken etwas mehr Ergebnisse unter 100 % als darüber. Auch wenn die Verteilung nicht eindeutig ist, scheint die überwie-

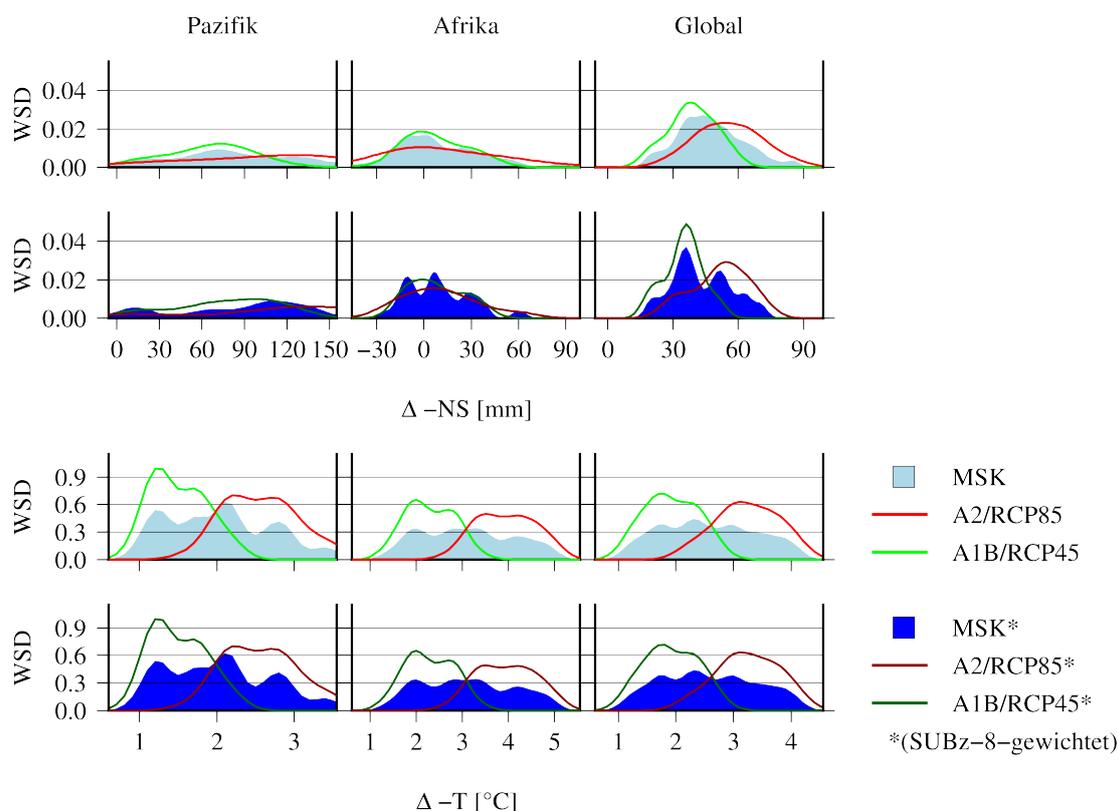
gende Zahl der Metrikgewichtungen zu einer Verringerung der Unsicherheit zu führen.

Ein vergleichbares Bild zeigt sich bei der Temperatur. Die maximale Änderung der Standardabweichung auf 140,4 % wird durch  $SUB_{z-8}$  erzeugt. Allgemein fallen bei der Temperatur die Effekte der Metrikgewichtungen geringer aus, was sich besonders stark bei den FPA zeigt. Hier sind die meisten Werte direkt um das Zentrum gestreut, d.h. es treten nur minimale Veränderungen auf. Die LOR-Ergebnisse zeigen hingegen eine ähnliche Streuung wie beim Niederschlag. Allerdings sind auch hier die Werte relativ gleichmäßig ober- und unterhalb von 100 % verteilt. Deutliche Unterschiede zwischen den Metriken sind dabei nicht zu beobachten. So werden insgesamt für LOR bzw.  $LOR_i$  ( $SUB_{z-8}$ ) maximal zwei (drei) signifikante Änderungen der Standardabweichung und zwei (keine) des Erwartungswertes erzeugt.

Alle analysierten Metriken haben hinsichtlich des Einflusses auf die WDF einen ähnlichen Effekt gezeigt. Zudem haben sich die Ergebnisse der Regionen des Mittelmeerraums auf die globalen Erweiterungsregionen übertragen lassen. Die Tatsache, dass eine generelle Verringerung des Unsicherheitsbereichs nicht möglich ist, bestätigt die zuvor gewonnenen Zwischenergebnisse. Dass auch ein relevanter Anteil an Situationen durch leistungsorientierte Gewichtung eine erhöhte Unsicherheit erhält, sollte bei der Interpretation für Anpassungsstrategien berücksichtigt werden. Um die Unsicherheit der beiden verwendeten Zukunftsszenarien mit in die Untersuchung einzubeziehen, wurden zusätzlich die MSK analog zu Kapitel 6.2.2 analysiert. Abbildung 7.11 zeigt exemplarisch für CMIP5 die Ergebnisse des annualen Niederschlags und der Jahresmitteltemperatur von Pazifik, Afrika und Global. Diese vermitteln einen allgemeinen Eindruck und lassen sich mit Vermerk auf die Unterschiede der Szenarien von CMIP3 zu CMIP5 gut übertragen.

Zunächst gilt zu beachten, dass die Abszisse in allen Regionen für Niederschlag und Temperatur leicht unterschiedliche Wertebereiche abbildet. Dies ist notwendig, um die MSK der verschiedenen Regionen vollständig darstellen zu können. Der Großteil der gleichgewichteten MSK des Niederschlags liegt im positiven Bereich, was einer projizierten Niederschlagszunahme entspricht. Die Streuung für Pazifik und Afrika ist dabei jedoch sehr hoch. Die KDF der getrennten Emissionsszenarien RCP4.5 und RCP8.5 zeigen kaum Unterschiede und entsprechen auch größtenteils dem Bild der MSK. Global ist die Streuung nicht so hoch.

Mit Blick auf die  $SUB_{z-8}$ -gewichteten MSK werden Veränderungen insbesondere hinsichtlich Global und Afrika deutlich. Bei beiden prägen sich basierend auf den Schwerpunkten der Szenarien multimodale Kurven aus. Global führen diese zu einer Zentrierung bei einer Niederschlagszunahme von ca. 35–50mm, wobei sich die Kurve auch insgesamt im Bereich zwischen 30–60mm verdichtet. Das Ergebnis der Gewichtung entspricht hier daher der Reduktion des Unsicher-

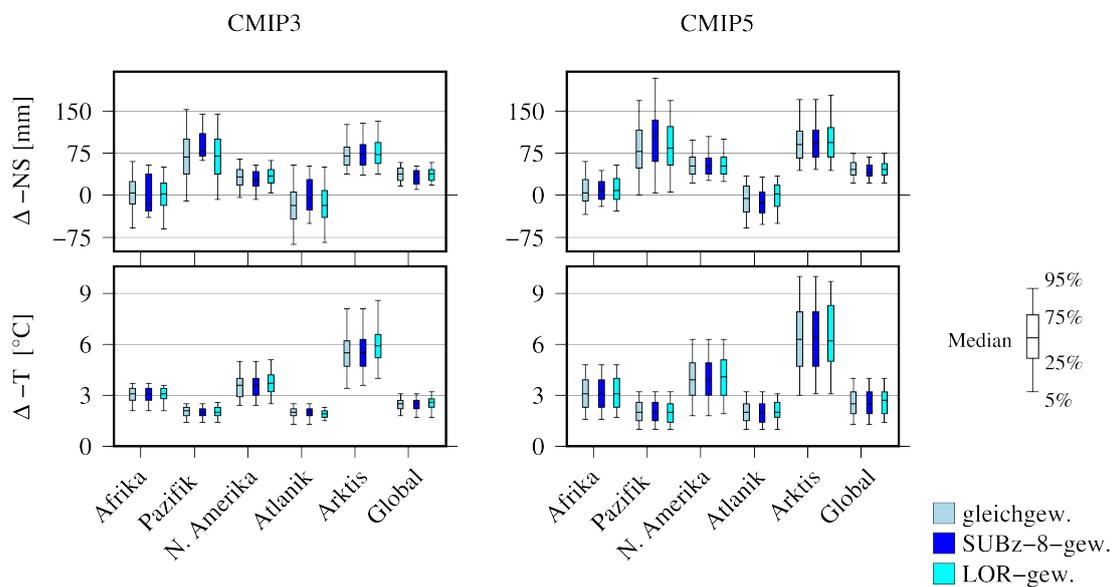


**Abbildung 7.11** Multi-Szenario-Kerndichteschätzfunktionen (MSK) der annualen MME-Klimaänderung in Afrika, Pazifik und Global. Dargestellt sind für die Niederschlags- (oben) und Temperaturänderung (unten) die jeweils gleiche (1. Zeile) und SUB<sub>z</sub>-8-gewichteten (2. Zeile) CMIP5-MSK. Zusätzlich sind die KDF der separaten Szenarien dargestellt.

heitsbereichs. In Afrika bilden sich noch mehr lokale Extrema aus, gleichzeitig verschiebt sich die MSK in Richtung einer leichten Abnahme, obwohl weiterhin der überwiegende Teil der MSK eine schwache Niederschlagszunahme zeigt. Beim Pazifik ist der Gewichtungseffekt nur schwach in Richtung einer etwas stärkeren Niederschlagszunahme ausgeprägt.

Bei der Temperatur sind dagegen kaum Unterschiede zwischen den gleich- und SUB<sub>z</sub>-8-gewichteten MSK auszumachen. Dieses Bild musste auch nach den Erkenntnissen aus Abbildung 7.10 erwartet werden. Lediglich für LOR-gewichtete MSK sind vereinzelt leichte Unterschiede zu erkennen. Da die Mehrheit der Situationen und aller Metriken eher mit dem in Abbildung 7.11 gezeigten Effekt übereinstimmt, wurde dies hier exemplarisch dargestellt. Auch wenn sich keine deutliche Ausbildung neuer lokaler Extrema zeigt, ist für alle drei Regionen eine minimale Verdichtung und damit Abnahme der Unsicherheit zu verzeichnen. Dabei gilt zu beachten, dass die Unsicherheit der Temperaturänderung erneut ins-

gesamt deutlich niedriger als die der Niederschlagsänderung ist. Stärker als beim Niederschlag zeigt sich bei der Temperatur erneut die Trennung zwischen den beiden Emissionsszenarien. Der Wertebereich von RCP8.5 ist dabei um ca.  $2^{\circ}\text{C}$  höher. Die exakte Lage ist allerdings situationsbedingt. Die MSK zeigt hier im Gegensatz zum Niederschlag entsprechend ein deutlich anderes Erscheinungsbild als die einzelnen KDF und bildet einen relativ gleichmäßigen Unsicherheitsbereich zwischen  $1\text{--}3,5^{\circ}\text{C}$  für Pazifik und  $1,5\text{--}5^{\circ}\text{C}$  für Afrika ab. Dies veranschaulicht die Brisanz der Frage nach dem Konfidenzintervall der Klimaänderung. Die Auswirkungen auf die Natur und den Menschen bei einer Temperaturzunahme von bis zu  $3,5^{\circ}\text{C}$  müssen so anders bewertet werden als bei einer Zunahme um  $1\text{--}2^{\circ}\text{C}$ .



**Abbildung 7.12** Überblick der MSK-Konfidenzintervalle aller Globalregionen und MMEs. Angetragen sind neben den gleichgewichteten Konfidenzintervallen auch die basierend auf Gewichtung mit SUB<sub>z</sub>-8 und LOR der annuellen Niederschlags-  $\Delta\text{-NS}$  (oben) und Temperaturänderungen  $\Delta\text{-T}$  (unten). Neben dem Median sind die 5 %-, 25 %-, 75 %- und 95 %-Perzentile angegeben

Beim Niederschlag fällt auf, dass die 90 %-Konfidenzintervalle in allen Regionen bis auf Pazifik und Atlantik trotz verschiedener Gewichtungen relativ stabil bleiben. Unterschiede sind eher durch die verschiedenen Regionen zu erkennen. Der regionale Vergleich zeigt hier für beide MMEs ein ähnliches Bild. So liegt der deutlich überwiegende Anteil der Konfidenzintervalle von Pazifik, Nordamerika, Arktis und Global im Bereich einer Niederschlagszunahme. Lediglich für Afrika und den Atlantik ist hier eine ausgewogene Verteilung von Zu- und Abnahme zu erkennen. Die Unterschiede zwischen CMIP3 und CMIP5 fallen beim Niederschlag eher gering aus. CMIP5 zeigt ein etwas stärker gestreutes Konfidenzintervall für Arktis

und Pazifik. Für Afrika und Atlantik ist dagegen umgekehrt eine höhere Streuung von CMIP3 zu beobachten. Die Ergebnisse der Jahreszeiten zeigen sich insgesamt konsistent mit den in Abbildung 7.12 dargestellten annualen Ergebnissen. Dabei sind hauptsächlich Schwankungen im Betrag des Niederschlags zu beobachten. Die Tendenz der Regionen bleibt hingegen stabil. Der Vergleich mit den metrikgewichteten Ergebnissen bestätigt, dass die Wirkung der Gewichtung insgesamt schwach ausfällt. Auch  $SUB_z-8$  und LOR, die Metriken mit den am stärksten differenzierenden Gewichten, zeigen beim Niederschlag überwiegend ähnliche Resultate. Für Afrika, Pazifik und Atlantik ist ein reduzierter Unsicherheitsbereich durch  $SUB_z-8$ -Gewichtung gegenüber LOR und dem gleichgewichteten MSK zu beobachten. Die Unterschiede sind ansonsten eher gering. Auch Verschiebungen des Intervalls bezogen auf den Wertebereich sind eher schwach ausgeprägt.

Bei der Temperatur ist das Bild der unterschiedlich gewichteten Metriken dagegen absolut homogen. Die Metrikgewichtung führt in nahezu keiner Situation zu einer markanten Änderung des Konfidenzintervalls. Lediglich beim Atlantik ist basierend auf der LOR-Gewichtung eine leichte Reduktion des Unsicherheitsbereichs auszumachen. Es fällt auf, dass CMIP3 insgesamt etwas kleinere Konfidenzintervalle zeigt als CMIP5. Dies ist anders als beim Niederschlag und ist auf die größeren Unterschiede der Emissionsszenarien von CMIP5 zurückzuführen. Die größte Unsicherheit besteht für beide MMEs übereinstimmend für Nordatlantik und Arktis. Die übrigen Regionen zeigen ein Intervall zwischen  $1-4^{\circ}\text{C}$ . Auch für die übrigen Jahreszeiten unterscheidet sich das Ergebnis nicht maßgeblich. Das Intervall der Erwärmung ist in DJF und SON höher als in MAM und JJA. Die Ergebnisse der Metrikgewichtungen zeigen auch hier keine nennenswerten Unterschiede.

Für die Gewichtung in den Globalregionen haben sich LOR (bzw. die  $4FM_i$ ) und die  $FPA_z$  als am besten geeignet erwiesen, wobei die meisten Metriken eine hohe Korrelation in der Einschätzung der Modelle für die jeweilige Situation gezeigt haben. Allerdings kann eine Gewichtung nur dort signifikante Effekte zeigen, wo die Unterschiede zwischen den Modellen dies zulassen. Hierzu ist auf globaler Ebene nur der Niederschlag zu nennen, da die Modellgewichte der Metriken für Temperatur hier auf einem hohen Niveau zu ähnlich sind.

Die Metriken haben sich als leicht übertragbar zwischen beiden Skalenniveaus herausgestellt, wobei gleichzeitig unterschiedliche Ergebnisse auch in Hinsicht auf die resultierenden Gewichtungen zu erkennen sind. Als abschließendes Kapitel wird im Folgenden die Integrität der in dieser Studie entwickelten und angewandten Evaluationsmetriken untersucht. Dazu werden alle zuvor erstellten Ergebnisse der unterschiedlichen Regionen berücksichtigt, um eine abschließende Einschätzung vornehmen zu können.

## 8 Integrität der Evaluationsmetriken

Im Überblick zum aktuellen Stand der Forschung (Kapitel 1.3) wurde angedeutet, dass eine ideale Metrik bisher noch nicht entwickelt wurde bzw. keine Metrik den Status der Allgemeingültigkeit erreicht hat. Die in dieser Studie verwendeten Metriken zeigen zwar hohe Korrelationen, untersuchen jedoch meist nur einen kleinen Aspekt der Klimatologie. Vor dem Hintergrund der unterschiedlichen Eigenschaften der Modelle und Anforderungen, die jede klimatologische Fragestellung mit sich bringt, kann daher nicht angenommen werden, dass eine Metrik Allgemeingültigkeit erreicht. Die Übertragbarkeit und Vielseitigkeit zeichnen dennoch ein vielversprechendes Bild.

In diesem Kapitel werden vier Fragestellungen zur Analyse der Integrität und Gültigkeit der entwickelten Metriken in dieser Studie untersucht. Zunächst wird die Konsistenz der Metriken hinsichtlich der Modellbewertung beleuchtet. Dabei wird geprüft, wie sich die Modellgewichte beim Vergleich unterschiedlicher regionaler Skalenniveaus ändern und der Versuch einer allgemeinen Rangliste basierend auf allen Ergebnissen dieser Arbeit vorgenommen. Danach kommt es zur Analyse der zeitlichen Übertragbarkeit der Rohgewichte. Hier wird die Konsistenz der Modellbewertungen der beiden leicht versetzten Evaluationszeiträume analysiert. Anschließend wird ein Ansatz zur Validierung der Metriken vorgestellt. Letztlich wird untersucht, welchen Einfluss die Wahl von verschiedenen Beobachtungsdaten auf die Ergebnisse hat.

### 8.1 Konsistenz der Modellbewertungen

Die Konsistenz der Modellbewertungen wird in diesem Abschnitt hinsichtlich der regionalen Skala vorgenommen. Dabei wird untersucht, ob Modelle auf unterschiedlichen Skalenniveaus konsistente Ergebnisse erreichen. Diese Fragestellung kann jedoch nicht als eine Anforderung an die Metriken oder Modelle verstanden werden. Vielmehr geht es dabei um das Abstecken des Rahmens, der bei der Interpretation der Modellbewertungen berücksichtigt werden sollte. Die Spearman-Korrelation der Rohgewichte zwischen verschiedenen Regionen innerhalb des Mit-

telmeerraums ist für die meisten Metriken gering (vgl. Abschnitt 5.1.2). Unter den globalen Regionen liegen die Korrelationen der Modellbewertung für alle Metriken dagegen meist etwas höher. Allerdings muss festgehalten werden, dass kein Modell in allen Regionen bzw. Situationen hohe Ergebnisse zeigt. Trotzdem wird an dieser Stelle eine Rangliste über alle untersuchten Metriken und Situationen vorgestellt. Um ein Ungleichgewicht zu vermeiden, wird für jeden der vier Metrikblöcke eine eigene Wertung über alle Situationen erstellt. Aufgrund unterschiedlicher Wertebereiche und Abstufungen werden nicht die Rohgewichte, sondern die jeweiligen Ranglistenpositionen verwendet. Dadurch wird eine Überrepräsentation beispielsweise von LOR, das sehr hohe Gewichte erstellen kann, vermieden. Für die 4FM und RM wurden, um ein allgemeines Bild liefern zu können, sowohl die Gewichte des Trends als auch die des klimatologischen Mittels berücksichtigt.

Über die resultierenden vier gemittelten Metrikranglisten wurde anschließend das arithmetische Mittel berechnet, um eine Gesamtwertung zu erzeugen. Da die Globalregionen nur mit ERA-20C evaluiert wurden, wurde zur Erstellung der Gesamtwertung auch für MMR auf die Rohgewichte basierend auf ERA-20C als Referenzdatensatz zurückgegriffen. Da allerdings alle Ergebnisse der Hauptuntersuchungsregionen auch für E-OBS und CRU erzeugt wurden, werden Unterschiede, die sich durch die Wahl der drei Datensätze ergeben, anschließend zusätzlich erläutert. Eine weiterführende Diskussion der Abhängigkeit von Referenzdaten wird in Abschnitt 8.4 vorgenommen. Um einen Vergleich zwischen der globalen und regionalen Skala zu ermöglichen, ist die Gesamtbewertung jeweils zweigeteilt für Niederschlag und Temperatur in Tabelle 8.1 dargestellt. Da die Modellverfügbarkeit zwischen Niederschlag und Temperatur nicht identisch ist, sind in Tabelle 8.1 die Gesamtwertungen für beide Variablen dargestellt. Zur Vereinfachung und leichteren Zuordnung sind die Modelle von CMIP3 farblich unterlegt.

Beim Niederschlag erreichen die besten Modelle MPI-ESM-MR und GFDL-ESM2M insgesamt einen mittleren Rangplatz von 25,2 bzw. 25,5. Beide sind Teil von CMIP5. Allerdings zeigt sich mit Blick auf die Verteilung der übrigen Platzierungen, dass hier keine Trennung zwischen den Modellgenerationen zu beobachten ist. Die Modelle sind gleichmäßig gemischt und die letzten beiden Positionen belegen mit NorESM1-M und HadGEM2-ES ebenfalls Modelle von CMIP5. Mit einer gemittelten Rangposition von 36,1 und 36,2 liegen die Modelle im Schnitt 10 Rangplätze niedriger als die besten Modelle. Dazwischen ist ein relativ gleichmäßiger Übergang zu beobachten. Angesichts des Umfangs von 61 Modellen für Niederschlag entspricht das einem Gesamtunterschied von 17,9 % zwischen den am besten und am schlechtesten bewerteten Modellen. Dieser Wert ist dabei jedoch nicht als Gütesiegel, sondern als statistisches Resümee zu verstehen.

Gerade die Auswertung der Niederschlagsergebnisse hat eine extrem hohe Va-

**Tabelle 8.1** Gemittelte Mittelmeer- (Mi.-R.) bzw. Globalregionen (Gl.-R.) sowie Gesamttrangelisten (Ges.) der durchschnittlichen Modellplatzierungen aller Metrikgruppen von Niederschlag (links) und Temperatur (rechts). Zur Erstellung aller Rohgewichte wurde ERA20-C als Referenzdatensatz zur Evaluation im Zeitraum 1960–2009 verwendet. Dargestellt sind die mittleren Modellrangplätze für die globalen Klimamodelle von CMIP3 (unterlegt) und CMIP5.

Modell	Niederschlag			Temperatur			Modell
	Ges.	Mi.-R.	Gl.-R.	Mi.-R.	Gl.-R.	Ges.	
MPI-ESM-MR	25,2	21,9	28,6	15,2	22,2	18,7	CNRM-CM3
GFDL-ESM2M	25,5	25,2	25,8	15,7	21,7	18,7	CGCM3.1 (T63)
IPSL-CM5A-MR	26,9	22,0	31,8	23,3	20,6	22,0	CMCC-CMS
BCCR-BCM2.0	27,1	27,3	26,8	21,1	24,0	22,5	CMCC-CM
CNRM-CM3	27,4	24,7	30,1	21,3	24,0	22,6	ACCESS1-0
GISS-EH	27,6	23,6	31,6	18,6	26,9	22,8	CGCM3.1 (T47)
CGCM3.1 (T63)	28,2	26,3	30,0	23,3	22,8	23,0	INGV-SXG
CGCM3.1 (T47)	28,4	27,6	29,2	21,4	24,9	23,1	UKMO-HadGEM1
HadGEM2-AO	28,6	27,1	30,1	20,5	25,9	23,2	CSIRO-Mk3.0
CSIRO-Mk3.0	28,6	27,7	29,6	21,4	26,1	23,7	HadGEM2-AO
CMCC-CM	28,9	26,1	31,7	24,5	23,3	23,9	CCSM4
IPSL-CM5B-LR	28,9	26,0	31,8	25,5	22,9	24,2	IPSL-CM5A-MR
CCSM4	28,9	29,4	28,4	26,0	24,9	25,4	CanESM2
GFDL-ESM2G	29,2	32,3	26,1	23,4	27,8	25,6	UKMO-HadCM3
MIROC3.2(hr)	29,6	28,8	30,4	24,3	26,8	25,6	IPSL-CM4(LMDZ)
CanESM2	29,7	32,1	27,3	20,7	30,8	25,8	BCCR-BCM2.0
GISS-E2-H-CC	29,7	30,2	29,2	28,1	23,9	26,0	CNRM-CM5
INGV-SXG	29,9	30,8	29,0	23,1	29,3	26,2	ECHO-G
ECHAM5/MPI-OM	30,1	32,2	27,9	25,7	27,0	26,3	NCAR-PCM1
MRI-CGCM3	30,1	29,2	31,0	26,7	26,2	26,5	MPI-ESM-MR
GISS-E2-H	30,2	31,4	29,0	25,7	28,4	27,0	ECHAM5/MPI-OM
NCAR-PCM1	30,2	34,0	26,5	26,6	27,7	27,1	MIROC3.2(hr)
GISS-ER	30,4	28,5	32,2	29,7	25,1	27,4	GFDL-ESM2M
UKMO-HadCM3	30,4	34,6	26,3	25,7	29,3	27,5	MPI-ESM-LR
MPI-ESM-LR	30,5	29,2	31,9	30,0	25,2	27,6	BCC-CSM1.1(m)
FGOALS-g1.0	30,6	29,2	31,9	26,8	28,8	27,8	NorESM1-ME
ACCESS1-3	30,6	32,7	28,5	24,6	31,6	28,1	MIROC5
MRI-CGCM2.3.2	30,7	31,6	29,9	28,0	28,2	28,1	CSIRO-Mk3.5
BCC-CSM1.1(m)	30,8	33,7	27,8	33,4	22,9	28,1	GFDL-ESM2G
CESM1-BGC	30,8	30,5	31,1	29,2	27,5	28,4	MIROC-ESM-CHEM
NCAR-CCSM3	30,9	31,9	29,8	34,0	23,0	28,5	GFDL-CM3
IPSL-CM4(LMDZ)	31,1	27,8	34,5	28,9	28,6	28,8	INMCM4
BCC-CSM1.1	31,2	33,6	28,7	25,5	32,1	28,8	CESM1-BGC
GISS-E2-R	31,2	30,6	31,7	31,6	26,0	28,8	BCC-CSM1.1
GFDL-CM3	31,3	36,5	26,1	28,0	29,9	29,0	CSIRO-Mk3-6-0
CSIRO-Mk3.5	31,4	31,1	31,6	28,3	29,7	29,0	NCAR-CCSM3
CESM1-CAM5	31,6	32,8	30,4	34,2	24,2	29,2	FIO-ESM
FGOALS-g2	31,7	32,2	31,2	30,1	29,1	29,6	ACCESS1-3
ACCESS1-0	31,7	32,7	30,7	28,0	31,6	29,8	MIROC3.2(mr)
NorESM1-ME	31,8	32,7	30,9	31,8	28,0	29,9	GISS-AOM
CSIRO-Mk3-6-0	32,0	29,2	34,7	31,7	28,2	29,9	IPSL-CM5A-LR
CMCC-CMS	32,0	32,2	31,7	30,4	29,6	30,0	CESM1-CAM5
ECHO-G	32,0	31,2	32,8	31,9	28,9	30,4	GISS-ER
CSIRO-Mk3L-1-2	32,2	29,4	35,0	35,6	26,2	30,9	IPSL-CM5B-LR
MIROC3.2(mr)	32,2	29,2	35,3	35,1	27,2	31,1	GFDL-CM2.0
MIROC-ESM	32,3	33,0	31,6	30,2	32,2	31,2	MRI-CGCM2.3.2
GISS-AOM	32,5	32,4	32,6	34,2	29,5	31,8	GFDL-CM2.1
MIROC-ESM-CHEM	32,5	35,2	29,9	37,5	26,8	32,2	MIROC-ESM
EC-EARTH	32,6	30,7	34,4	31,7	33,7	32,7	NorESM1-M
INM-CM3	33,2	30,2	36,2	31,8	33,9	32,9	MRI-CGCM3
GISS-E2-R-CC	33,4	33,0	33,8	36,3	33,5	34,9	FGOALS-g2
UKMO-HadGEM1	33,6	35,9	31,3	34,5	35,6	35,0	GISS-EH
CNRM-CM5	33,7	34,6	32,8	32,6	38,3	35,5	CSIRO-Mk3L-1-2
IPSL-CM5A-LR	33,7	29,4	38,0	33,7	38,2	36,0	INM-CM3.0
HadGEM2-CC	33,8	36,0	31,6	39,7	38,3	39,0	FGOALS-g1.0
MIROC5	33,8	33,7	33,9				
GFDL-CM2.0	34,5	37,9	31,0				
GFDL-CM2.1	34,6	37,7	31,5				
INMCM4	35,7	33,5	38,0				
NorESM1-M	36,1	38,0	34,1				
HadGEM2-ES	36,2	41,0	31,4				

riabilität in der Modellbewertung gezeigt. Auch mit Blick auf die Unterschiede zwischen den mittleren Ranglisten der Mittelmeer- und der Globalregionen wird dies deutlich. Besonders Erstere tendieren hier zu einer stärkeren Trennung der Modelle. So besteht ein Unterschied von über 31,3 % zwischen dem ersten und letzten Modell, das identisch mit dem der Gesamtrangfolge ist. Damit ist der Einfluss der Mittelmeerregionen auf die Gesamtrangfolge auch höher als der der Globalregionen, da hier nur ein Unterschied von 20 % zwischen den Extremen besteht. Der Vergleich der beiden gemittelten Regionenranglisten zeigt ein bunt durchmischtes Bild. Dies verdeutlicht auch eine Spearman-Korrelation von  $-0,06$ . Ein Zusammenhang der über alle Metriken gemittelten Modellbewertungen des Mittelmeerraums und der Globalregionen muss damit abgelehnt werden. Mit Blick auf die mittleren Ranglisten der einzelnen Metrikblöcke ist das Bild etwas stabiler. Einen signifikanten Zusammenhang zwischen der gemittelten Rangfolge der Global- und der Rangfolge der Mittelmeerregionen zeigt allerdings nur RM mit einer Spearman-Korrelation von  $0,37$ . Die Korrelationen der 4FM-Ranglisten sind mit  $0,25$  gerade nicht mehr signifikant, wohingegen das Ergebnis für ZM ( $0,14$ ) und FPA ( $0,07$ ) jeweils noch niedriger ausfällt. Hier muss allerdings beachtet werden, dass die FPA nur für den MMR und nicht seine Unterregionen angewandt wurden. Daher drückt der Wert hier nur die Korrelation der mittleren Ränge des MMR mit dem Durchschnitt der Globalrangfolgen aus. Allerdings sind auch die übrigen Korrelationen so niedrig, dass nicht von einer Übertragbarkeit der Ranglisten gesprochen werden kann. Dieses Bild hat sich allerdings, insbesondere für den Niederschlag, auch bei den Evaluationen der einzelnen Simulationen angedeutet.

Ein anderes Bild zeigt sich bei der Temperatur. Zunächst fällt hier auf, dass die beiden bestplatzierten Modelle CNCRM-CM3 und CGCM3.1 (T63) (beide  $18,7$ ) ebenso wie die letztplatzierten INMCM3.0 ( $36,0$ ) und FGOALS-g1.0 ( $39,0$ ) Teil von CMIP3 sind. Dieses Ergebnis ist überraschend, da die Einzelgewichte der Temperatur in einer Vielzahl von Situationen eine bessere Simulationsgüte für CMIP5 gezeigt haben. Dies scheint jedoch nur auf das Mittel zuzutreffen und nicht auf die einzelnen Modelle, die auch bei CMIP3 teilweise sehr niedrige Rangpositionen annehmen. Des Weiteren fällt auf, dass die Unterschiede zwischen den Extremen bei der Temperatur deutlich ausgeprägter sind als beim Niederschlag. So besitzt CNCRM-CM3 insgesamt einen um  $36,9$  % besseren mittleren Gesamtrang als FGOALS-1.0g. Auch die Ähnlichkeit zwischen den einzelnen regionalen Rangfolgen ist deutlich höher bei der Temperatur. So liegen die Werte insbesondere der ersten und letzten 15 Modelle häufig gleichauf. Dies zeigt sich in einer signifikanten Spearman-Korrelation von  $0,4$ . Auch zwischen allen gemittelten Metrikrangfolgen der Global- und Mittelmeerregionen sind die Werte mit  $0,49$  (4FM),  $0,34$  (RM),  $0,35$  (ZM) und  $0,45$  (FPA) signifikant. Dabei gilt

wiederum, dass die FPA nur über alle Ranglisten des Gesamtmittelmeerraums berechnet wurden. Auch wenn die Werte hier ebenfalls eher in einem niedrigen Korrelationsbereich liegen, kann doch von einem positiven Zusammenhang gesprochen werden. Es ist bemerkenswert, dass damit aus den deutlich kleineren Unterregionen des Mittelmeerraums teilweise Tendenzen über die Simulationsgüte auf globalem Maßstab abgeleitet werden können. Für FPA gilt entsprechend, dass hier der Zusammenhang des gesamten MMR mit den großskaligen Gebieten bestätigt wird.

Die dargelegten Ergebnisse lassen den Schluss zu, dass die Metriken eine Korrelation der Modellbewertungen zwischen unterschiedlichen Skalen detektieren, falls sie bestehen. In dieser Arbeit wurde dies nur für die Temperatur und nicht für den Niederschlag erkannt. Als Grund hierfür können auch Unsicherheiten bezüglich der Referenzdaten nicht ausgeschlossen werden. Gerade der Trend des Niederschlags war in den meisten Regionen des Mittelmeerraums nicht eindeutig. Die zuvor erläuterten Ergebnisse beruhen auf den Gewichten, die basierend auf ERA-20C für die Hauptuntersuchungsregionen verwendet wurden. Tauscht man diese durch die Gewichte der Hauptreferenz E-OBS oder CRU aus, so treten teilweise deutliche Veränderungen der Niederschlagsrangfolge auf. Die Modelle der Temperatur bleiben dagegen größtenteils konstant. Auch die Korrelation der Rangfolgen bleibt auf vergleichbarem Niveau mit 0,44 (E-OBS) bzw. 0,36 (CRU). Analog verhält es sich mit der Korrelation der einzelnen Metriken. Beim Niederschlag ändert sich die allgemeine Korrelation nicht und liegt zwischen  $-0,01$  (E-OBS) und  $0,02$  (CRU). Für die Korrelation der Gesamtranglisten scheint die Wahl der Beobachtungsdaten daher nur eine untergeordnete Rolle zu spielen. YIN ET AL. (2015) und PRAKASH ET AL. (2015) detektierten allerdings für verschiedene Referenzdaten insbesondere beim Niederschlag teilweise deutliche Abweichungen. Diese Erkenntnisse decken sich auch mit den Niederschlagsranglisten der einzelnen Metriken, wo es durch unterschiedliche Referenzdaten zu Verschiebungen der Reihenfolge kommt. Die richtige Wahl der Beobachtungsdaten ist gerade auch hinsichtlich regionaler Stärken und Schwächen entsprechend von großer Bedeutung (THORNE ET AL., 2005; KIM ET AL., 2015). Eine weiterführende Analyse bezüglich der Abhängigkeit der Metriken bzw. der einzelnen Simulationsbewertungen von der Wahl der Referenzdaten wird in Abschnitt 8.4 durchgeführt.

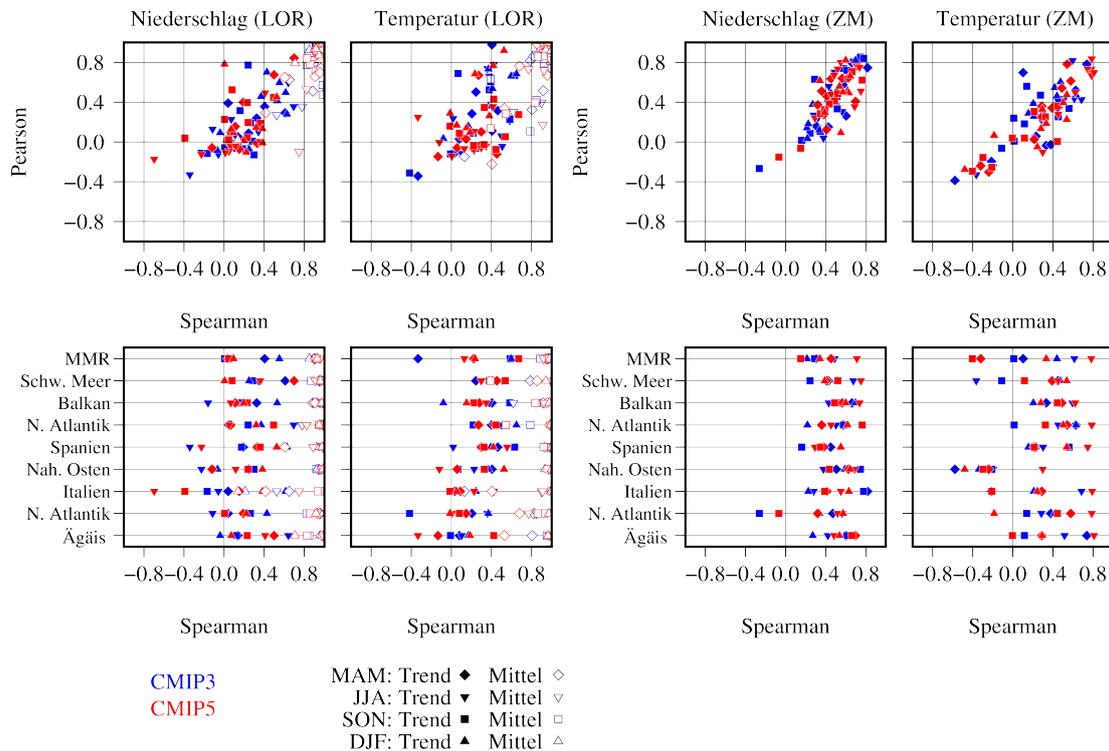
In den vorangegangenen Kapitel wurde verdeutlicht, dass die Metriken selbst sehr gut auf unterschiedliche Regionen und Skalen übertragbar sind. Diese Konsistenz der Ergebnisse gilt jedoch nur eingeschränkt. Daher bleibt festzuhalten, dass gerade für den Niederschlag so instabile Modellbewertungen selbst innerhalb einer Region erzeugt werden, dass keine Verallgemeinerung möglich ist. Da die Modellbewertungen der Metriken jedoch plausibel ausfallen, ist der Grund hierfür in den Modellen selbst zu suchen. Gerade mit Blick auf die Temperatur zeigt

sich, dass durchaus Konsistenz besteht. Dies gilt trotz teilweise sehr verschiedener Ansätze der Metrikgruppen. Der Niederschlag ist dagegen eine deutlich komplexere Größe, die von vielen Modellen nur für einzelne Regionen bzw. Situationen überzeugend bewältigt wird. Somit sind hier anders als bei der Temperatur kaum Schlüsse von einer auf die andere Skala oder Region möglich.

## 8.2 Zeitliche Übertragbarkeit der Modellbewertungen

In diesem Abschnitt wird die zeitliche Konsistenz der Modellbewertungen untersucht. Die analysierten Ergebnisse der Modellevaluation beruhen alle auf dem Zeitintervall 1960–2009, da CORDEX nur einen deutlich kürzeren Zeitraum abdeckt, der für das zweite Untersuchungsintervall 1950–1999 lediglich 30 statt 40 Jahre im Vergleich zu 50 Jahren von CMIP3 und CMIP5 zur Verfügung gehabt hätte. Dies hätte sich insbesondere auf den Trend deutlich ausgewirkt. Zur Analyse der zeitlichen Übertragbarkeit wird an dieser Stelle daher die Übereinstimmung der Rohgewichte von CMIP3 und CMIP5 basierend auf beiden Zeiträumen verglichen. Damit wird untersucht, ob Modelle, die eine hohe Bewertung für ein bestimmtes Zeitfenster erhalten haben, auch in anderen Zeiträumen bessere Ergebnisse liefern. Da nur um 10 Jahre verschobene Zeiträume zur Verfügung stehen, müssen die Ergebnisse dieses Vergleichs als ein erster Schritt bzw. erste Hinweise verstanden werden. Die Analyse wird sowohl für die Hauptuntersuchungsgebiete (Abbildung 8.1) als auch die Globalregionen (Abbildung 8.2) durchgeführt.

Abbildung 8.1 ist zweigeteilt. Die Diagramme im oberen Abschnitt zeigen die Gegenüberstellung von Pearson- und Spearman-Korrelation der Ranglisten von Rohgewichten aller zuvor analysierten Situationen des Mittelmeerraums zwischen den Zeiträumen 1950–1999 und 1960–2009. Bei den Korrelationen der ZM-Ranglisten sind die Unterschiede zwischen Pearson- und Spearman-Korrelation relativ gering. Die Werte streuen über alle Regionen zwischen  $-0,4$ – $0,8$ . Der Großteil der Korrelationen liegt allerdings zwischen  $0$ – $0,8$  beim Niederschlag und  $0$ – $0,6$  bei der Temperatur. Negative Korrelationen sind dagegen nur vereinzelt zu beobachten. Hier fällt besonders der Nahe Osten bei der Temperatur mit Korrelationen auf, die bis auf eine Ausnahme alle negativ sind. Beim Niederschlag zeigt der Nahe Osten dagegen mit Werten zwischen  $0,35$ – $0,8$  die besten Ergebnisse. Hier ist der Nordatlantik mit zwei Situationen im negativen Bereich zu nennen. Die Werte der übrigen Regionen liegen im leicht bis mittleren positiven Bereich. Da die Zeitfenster, die hier verglichen werden, nur um 10 Jahre versetzt sind, müssen die Korrelationen hier insgesamt eher als gering bewertet werden,



**Abbildung 8.1** Korrelation der Modellbewertung von CMIP3 und CMIP5 basierend auf E-OBS zwischen den Referenzzeiträumen 1960–2009 und 1950–1999. Angetragen sind die Spearman- und Pearson-Korrelation der Rangfolgen (oben) basierend auf LOR (links) und ZM (rechts) für Niederschlag (1. Spalte) und Temperatur (2. Spalte) sowie die regionale Unterteilung für die Mittelmeerregionen (unten). Für LOR sind die Korrelationen basierend auf dem Trend und dem Mittel angetragen.

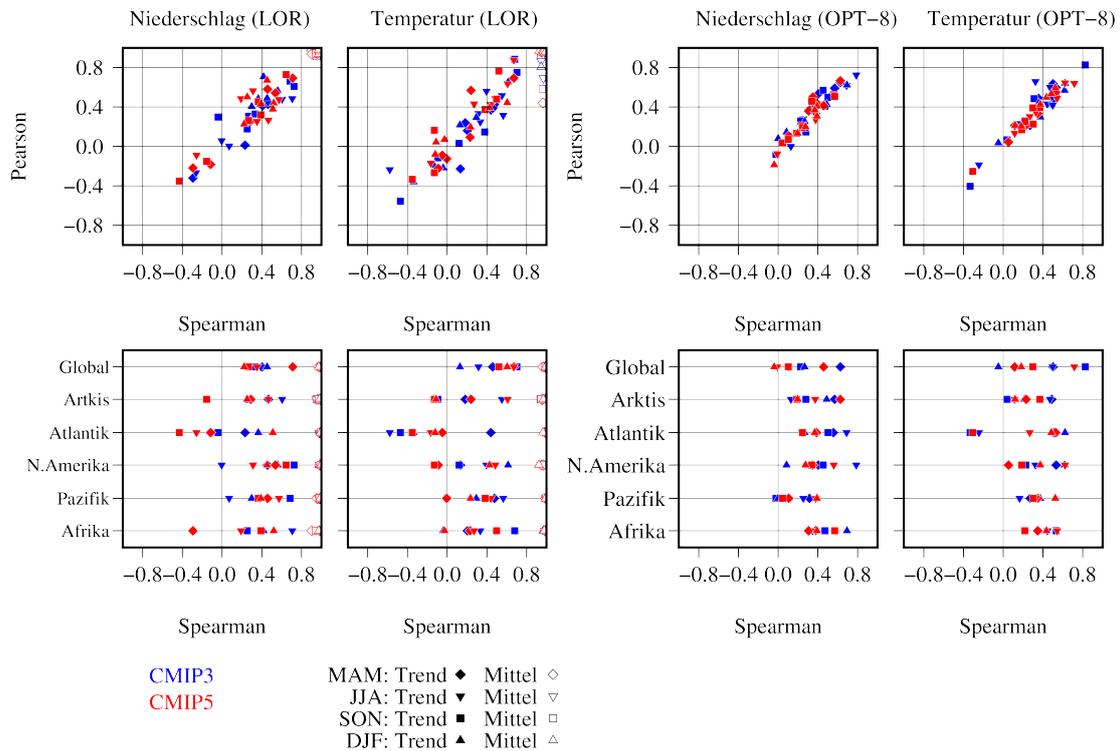
auch wenn die Mehrheit der Situationen bei 0,4 oder höher liegt. Generell zeigen die ZM-Korrelationen einen relativ linearen monotonen Zusammenhang. Dieser ist auch bei den Trendergebnissen von fünf der sechs 4FM und RM zu beobachten. Auch die Korrelationen der FPA-Ranglisten zeigen für die acht Situationen des Gesamtmittelmeerraums Werte, die eine identische Verteilung erahnen lassen. Einzige Ausnahme ist das Muster der LOR-Korrelationen.

Bei LOR fällt zunächst eine Zweiteilung auf. Das klimatologische Mittel zeigt sowohl für Temperatur als auch Niederschlag für nahezu alle Situationen Korrelationen, die höher sind als die des klimatologischen Mittels. In den meisten Fällen liegen die Werte des Mittels zwischen 0,8–1. Vereinzelt sind auch Ausreißer mit niedrigeren Korrelationen insbesondere für Italien zu beobachten. Hier fallen die Trendkorrelationen dann allerdings noch geringer aus. Besonders Atlantik zeigt hier sehr niedrige Korrelationen für Niederschlag und Temperatur. Dieses Bild

ist auch für die übrigen 4FM gültig. Die Modellbewertungen basierend auf dem klimatologischen Mittel für 1960–2009 sind entsprechend auch für Evaluationen für 1950–1999 gültig. Am deutlichsten ist dies bei RM (nicht gezeigt), bei der alle Werte zwischen 0,95–1 liegen. Für die Trendkorrelationen ergibt sich ein anderes Bild. Während die Spearman-Korrelationen in der regionalen Betrachtung noch vergleichbare Ergebnisse wie ZM oder die anderen Metriken anzeigen, ist die Monotonie der Werte hier deutlich schwächer ausgeprägt. Dies verdeutlicht, dass die Reihenfolge der Modellbewertungen mit Blick auf den Trend für LOR eine ähnliche Stabilität wie für ZM oder die übrigen Metriken zeigt. Die Abstufungen der individuellen Gewichte innerhalb der Ranglisten variieren dabei allerdings teilweise sehr stark. Die Unterschiede in Bezug auf die Monotonie gelten bei LOR in abgeschwächter Form auch für das klimatologische Mittel. Allerdings zeigt sich hier eine höhere Spearman-Korrelation als beim Trend. Dadurch wird der optische Effekt abgeschwächt. Die Person-Korrelationen liegen dennoch auch beim Mittel im negativen Bereich bei einer Spearman-Korrelation von über 0,7 beim Niederschlag und 0,4 bei der Temperatur.

Im Mittelmeerraum zeigen sich generell ähnliche Ergebnisse. Der überwiegende Teil der Korrelationen liegt im teilweise stark positiven Bereich. Dies gilt für Niederschlag und Temperatur insbesondere für das klimatologische Mittel der 4FM und RM. Beim Trend ist die Übereinstimmung zwischen den beiden Zeiträumen deutlich schwächer, wenngleich zumeist noch positiv. Ein Großteil der Werte liegt hier zwischen 0–0,8. Dieses Ergebnis ist allerdings dennoch auffällig gering, da der Überschneidungsbereich zwischen beiden Zeiträumen bei 80 % liegt. Vor diesem Hintergrund kann eine zeitliche Übertragbarkeit der Trendmodellbewertungen auf zwei unabhängige Zeiträume nicht angenommen werden. Zum Abgleich der Ergebnisse zeigt Abbildung 8.2 die Korrelationen der Ranglisten der Globalregionen basierend auf LOR und OPT-8.

Das Bild ist hierbei nahezu identisch mit dem des Mittelmeerraums. Die insgesamt schwächsten Korrelationen werden bei LOR und OPT-8 für den Atlantik gezeigt. Ein saisonales Muster ist nicht auszumachen. Stattdessen kann jede Jahreszeit abhängig von der Variablen und der Region hohe oder eher niedrige Korrelationen zeigen. Analog zu Abbildung 8.1 sind die LOR-Trendkorrelationen überwiegend positiv. Dabei besteht für die Globalregionen für LOR analog zu allen anderen Metriken auch eine hohe Übereinstimmung zwischen Pearson- und Spearman-Korrelationen. Die Ergebnisse von OPT-8 fügen sich hier in die der übrigen Trendkorrelationen ein. Allerdings fällt auf, dass hier bis auf für Atlantik kaum negative Werte berechnet werden. Außerdem liegen auch die übrigen Korrelationen für Niederschlag und Temperatur überwiegend zwischen 0,2–0,7 und zeigen damit die stabilsten positiven Werte. Die Korrelationen basierend auf SUB-8 ähneln hingegen eher denen von LOR und den übrigen Metriken. Die Reduktion



**Abbildung 8.2** Korrelationen der Modellbewertungen von CMIP3 und CMIP5 basierend auf ERA-20C zwischen den Referenzzeiträumen 1960–2009 und 1950–1999. Angetragen sind die Spearman- und Pearson-Korrelation der Rangfolgen (oben) basierend auf LOR (links) und OPT-8 (rechts) für Niederschlag (1. Spalte) und Temperatur (2. Spalte) sowie die regionale Unterteilung für die Globalregionen (unten). Für LOR sind die Korrelationen basierend auf dem Trend und dem Mittel angetragen.

der Rohgewichte auf das Klimawandelssignal scheint daher einen positiven Effekt auf die Stabilität der Modellbewertungen auszuüben. Wobei die genauen Übereinstimmungen in den einzelnen Situationen keine Verallgemeinerung zulassen. Die Werte des klimatologischen Mittels fallen mit Werten  $\geq 0,9$  durchweg höher aus als für den Mittelmeerraum. Darauf reduziert scheint die Veränderung der untersuchten Regionen bzw. der ihrer Größe einen positiven Effekt auf die zeitliche Konsistenz zu haben. Der Vergleich von zwei Zeiträumen, die sich zu 80 % überlappen, hat nicht die gleiche Aussagekraft wie zwei völlig unabhängige Zeiträume. Daher müssen die hier vorgestellten Ergebnisse zurückhaltend interpretiert werden und allgemeine Schlüsse sind nicht zulässig.

Die Metriken stimmen darin überein, dass die Ranglisten des klimatologischen Mittels überwiegend sehr hohe Korrelationen zwischen den untersuchten Zeiträumen zeigen. Dagegen ist der Zusammenhang der Modellbewertungen zwischen

den beiden Zeiträumen basierend auf dem 50-jährigen Trend in großen Teilen nur noch gering. Der Grund hierfür kann nicht bei speziellen Charakteristiken der Regionen gesucht werden, da sich diese grundlegend unterscheiden und auch bei der Evaluation unterschiedliche Ergebnisse gezeigt haben. Zusätzlich gilt, dass die zuvor analysierten Ergebnisse konstant sowohl für CMIP3 und CMIP5 als auch für Niederschlag und Temperatur sind. Es muss vermutet werden, dass der Grund in der Länge des Untersuchungszeitraums und des Trends selbst liegt.

Wie die Zeitreihenanalysen (vgl. Kapitel 6.1 und 7.2.1) gezeigt haben, simulieren die Klimamodelle für das 20. Jahrhundert einen langfristigen und stetigen Trend, der nicht zu plötzlichen Abweichungen oder interner Variabilität neigt. Bei den Referenzdaten gibt es dagegen auch kurzfristige Variationen, die den Trend teilweise stark beeinflussen können. Da PAXIAN ET AL. (2013) bereits mit 30-jährigen Zeitintervallen auf den starken Einfluss solcher Variationen im Mittelmeerraum verwiesen haben, wurden in dieser Studie 50-jährige Zeitintervalle, mit Blick auf die maximale Länge der Referenzdaten, gewählt. Zur Evaluation der Modelle in den gewählten Zeiträumen hat sich dieser als geeignet herausgestellt. Um allerdings Gewichte zu erstellen, die über unterschiedliche Zeiträume konsistent sind, könnten noch längere Zeiträume von beispielsweise 100 Jahren erforderlich sein. Auch wenn die erstellten Ranglisten der Trends keine universelle Gültigkeit besitzen, bieten sie doch eine wichtige Information zur Simulationsgüte. Für das klimatologische Mittel gilt, dass die zeitliche Konsistenz, zumindest auf der Basis einer Verschiebung um 10 Jahre, gegeben ist.

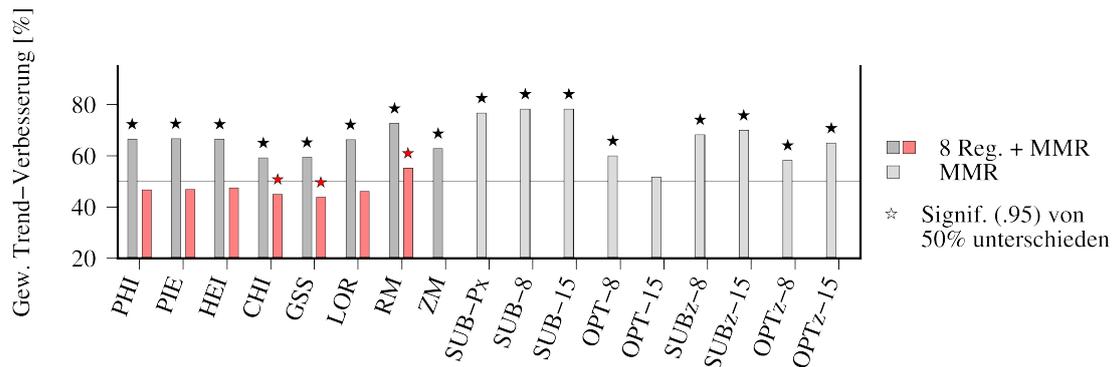
### 8.3 Ansatz zur Metrikvalidierung

Nach der Prüfung der räumlichen und zeitlichen Integrität der Evaluationsergebnisse wird an dieser Stelle der Ansatz einer Validierung vorgenommen. Da ein Schwerpunkt dieser Arbeit auf den experimentellen Anwendungen neuer Metriken lag, wird in diesem Kapitel der Erfolg der Gewichtungszuteilung durch die Metriken bewertet. Damit wird der Versuch unternommen, die paradoxe Frage zu beantworten, ob die angewandten Evaluationsmetriken tatsächlich die Modelle besser bewerten, deren untersuchte Variablen die größte Ähnlichkeit mit den Referenzdaten aufweisen.

Zur Beantwortung dieser Fragestellung wird die folgende These aufgestellt: Eine erfolgreiche Gewichtung basierend auf dem 50-jährigen Trend sollte einen positiven Effekt auf den Trend des MME im Vergleich zu dem der Referenzdaten haben. Hierbei gilt zu beachten, dass die Evaluation auf dem Raummuster des Trends für 4FM, RM und FPA sowie auf der Zeitreihenbeschaffenheit für ZM durchgeführt wurde. Die Anwendung der Gewichte auf die Zeitreihe zur Berechnung des

resultierenden Trends muss daher nicht zwingend zu demselben Ergebnis führen. Um den Metriken eine sinnvolle Gewichtungszuteilung zu attestieren, sollte jedoch der metrikgewichtete Trend der MMEs näher an dem der Beobachtungsdaten liegen als der gleichgewichtete Trend. Ändert sich die untersuchte Evaluationsgröße von Trend zu klimatologischem Mittel, so sollte dies zu einem anderen Resultat führen. Abbildung 8.3 zeigt die Auswertung dieser These. Insgesamt wurden für alle 4FM sowie für die RM und ZM 540 Situationen untersucht. Diese setzen sich aus den drei MMEs, allen Jahreszeiten, zwei Zeiträumen, den beiden untersuchten Variablen und neun Regionen zusammen. Für die FPA und FPA<sub>z</sub> wurden entsprechend 60 Situationen nur für den Gesamtmittelmeerraum untersucht. Für die 4FM und RM wurde die Auswertung sowohl basierend auf den Trend- (grau) als auch den Mittelevaluationsergebnissen (rot) durchgeführt. Für jede Situation wurde anschließend geprüft, wie sich der Trend des MMEs über den untersuchten Zeitraum in Bezug auf die Referenzdaten verhält. Zur Berechnung des Validierungstrends wurde für jedes Jahr des 50-jährigen Intervalls ein metrikgewichteter Wert erstellt und für die resultierende Zeitreihe anschließend der Regressionskoeffizient berechnet. Wenn der metrikgewichtete Regressionskoeffizient des MMEs im Vergleich zum gleichgewichteten dem Regressionskoeffizienten der Referenzdaten ähnlicher wurde, wurde dies für die jeweilige Situation als Erfolg der Metrik gewertet. Um die Signifikanz der Ergebnisse zu prüfen, wurde hier der *u*-Test angewandt. Dieser untersucht, ob sich die Anzahl der Änderungen im Bereich des Zufalls bewegt. Es muss angenommen werden, dass ein zufälliges Ergebnis gleichmäßig zwischen Verbesserungen und Verschlechterungen streuen würde. Mit Hilfe des *u*-Tests wird daher untersucht, ob sich die Ergebnisse mit einer Irrtumswahrscheinlichkeit  $\alpha = 0,05$  von einem zufälligen Ergebnis (50 %) unterscheiden (Abbildung 8.3).

Die Auswertung der Validierung basierend auf Trend oder Zeitreihenanalyse (für ZM) zeigt für alle der 17 dargestellten Metrikansätze einen Verbesserungsanteil von über 50 %. Im Schnitt nähert sich der gewichtete Trend in 62,2 % der Fälle dem der Referenzdaten an. Dabei zeigt sich auch ein leichtes Ungleichgewicht zwischen den einzelnen Metriken. Insbesondere OPT-15 besitzt hier mit 51,1 % den niedrigsten Anteil, der sich auch nicht signifikant vom Zufall unterscheidet. Dagegen sind bei SUB-Px, SUB-8 und SUB-15 mit über 76,7 % die meisten Verbesserungen zu beobachten. Dabei gilt jedoch, dass hier auch nur der Gesamtmittelmeerraum mit 60 untersuchten Situationen betrachtet werden konnte. Dies ist auch wichtig für die Prüfung der Signifikanz, da diese stark von der Anzahl der untersuchten Situationen abhängig ist. Bei einer großen Stichprobe genügen schon Werte, die sich nur geringfügig von 50 % unterscheiden, um als signifikant klassifiziert zu werden (vgl. Abschnitt 3.7.5). Knapp hinter den SUB-Metriken liegen mit 72,8 % die RM-Ergebnisse. Die übrigen 4FM-Ergebnisse sind

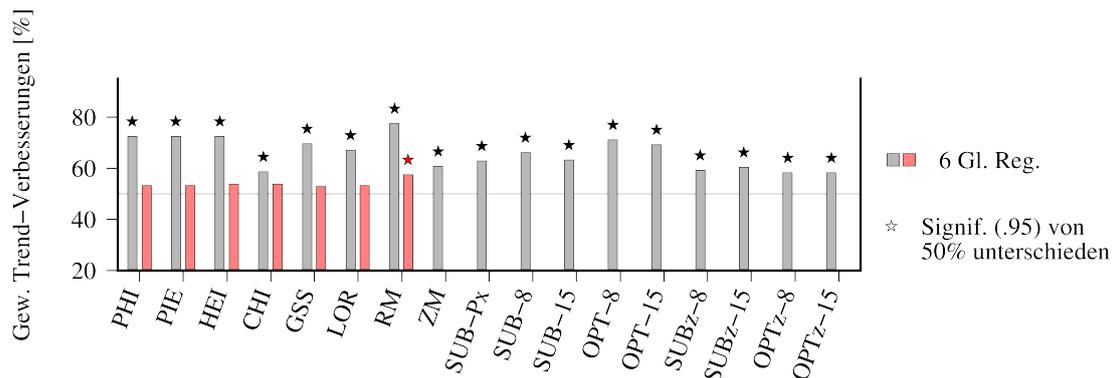


**Abbildung 8.3** Ansatz zur Metrikvalidierung für die Hauptuntersuchungsgebiete des Mittelmeerraums. Angetragen sind die durch Metrikgewichtung erreichten Annäherungen des Regressionskoeffizienten an den der Referenzdaten in beiden Untersuchungszeiträumen 1960–2009 und 1950–1999. Für die 4FM und RM ist dies sowohl basierend auf den Trend- (grau) als auch den klimatologischen Mittelgewichten (rot) angetragen.

mit 59,4–66,7 % ungefähr auf dem Level der  $FPA_z$ . Das Gesamtbild zeigt für alle Metriken ein positives Verhältnis der Verbesserungen des Trends durch die Gewichtung der MMEs. Alle Metriken außer OPT-15 zeigen einen Verbesserungsanteil, der sich signifikant von 50 % unterscheidet. Daraus lässt sich folgern, dass die Modelle, die basierend auf ihrem räumlichen Muster ein hohes Gewicht erhalten haben, überwiegend auch zu einer Verbesserung des Trends bezogen auf die Referenzdaten führen. Da dies nur ein Validierungsansatz ist, müssen diese Ergebnisse vorsichtig interpretiert werden und können lediglich als Hinweis verstanden werden, dass die Metriken eine sinnvolle Evaluation durchführen. Dies wird dadurch untermauert, dass sich die Metriken teilweise stark unterscheiden und hier trotzdem dieselbe Schlussfolgerung zulassen.

Um dem Anspruch einer Validierung gerecht zu werden, sind zur Gegenprüfung in Abbildung 8.3 die Ergebnisse basierend auf den Gewichten des klimatologischen Mittels (rot) angetragen. Diese wurden von 4FM und RM bestimmt und analog zu den Trendgewichten angewandt, um als Gegenprüfung zu dienen. Es zeigt sich, dass sechs von sieben Metriken unter 50 % liegen. Damit kommt es hier in der Mehrheit der Fälle zu einer vergrößerten Abweichung vom Trend der Referenzdaten. Nur bei RM ist mit 55,9 % eine signifikante positive Bilanz zu beobachten. CHI und GSS zeigen basierend auf der Evaluation des klimatologischen Mittels eine signifikante negative Bilanz. Die übrigen Ergebnisse der 4FM sind nicht signifikant von 50 % unterschieden und liegen damit im Bereich zufälliger Gewichtungen. Dieses Bild deckt sich mit der Auswertung der Korrelation der verschiedenen Metrikansätze (vgl. Kapitel 5.4). Die hier analysierten Trendän-

derungen gehen allerdings darüber hinaus und verdeutlichen, dass sich die Verwendung der Mittelgewichte im Gegensatz zu den Trendgewichten eher negativ auf den Trend der Zeitreihen auswirkt. Dass OPT-8 keinen signifikant positiven Anteil an Verbesserungen aufweist, kann möglicherweise auf den begrenzten Umfang mit nur 60 Situationen zurückgeführt werden. Um dies zu testen bzw. um die Gültigkeit aller Ergebnisse zu überprüfen, wurde die gesamte Validierung auf die Globalregionen übertragen (vgl. Abbildung 8.4).



**Abbildung 8.4** Ansatz zur Metrikvalidierung für die Globalregionen. Angetragen sind die durch Metrikgewichtung erreichten Annäherungen des Regressionskoeffizienten an den der Referenzdaten in beiden Untersuchungszeiträumen 1960–2009 und 1950–1999. Für die 4FM und RM ist dies sowohl basierend auf den Trend- (grau) als auch den klimatologischen Mittelgewichten (rot) angetragen.

Da hier nur sechs Regionen untersucht wurden, ist der Stichprobenumfang auf 240 reduziert. Allerdings ist dieser hier für alle Metriken gegeben. Das Ergebnis der trendbasierten Gewichtung zeigt sich konsistent mit dem des Mittelmeerraums. Alle Metriken zeigen einen signifikant positiven Anteil an Trendverbesserungen. Umgekehrt zeigt bei den mittelbasierten Gewichtungen erneut nur RM einen signifikanten Anteil an Verbesserungen, der jedoch deutlich unter dem der Trendgewichtung liegt. Auch die 4FM zeigen zwar Werte, die knapp über 50 % liegen, da die Ergebnisse hier jedoch nicht signifikant sind, ist dies im Rahmen einer zufälligen Streuung. Auch der Test der intensivierten Ansätze ( $4FM_i$ ) zeigt für alle Metriken unabhängig vom Untersuchungsraum konsistente Ergebnisse.

Bei der Interpretation der Resultate dieses Ansatzes gilt es zu berücksichtigen, dass die Validierung von Evaluationsmetriken eine komplexe wenn nicht paradoxe Fragestellung ist. Die dargestellten Ergebnisse spiegeln entsprechend nur den Versuch einer Validierung wider. Dass diese Herangehensweise nicht frei von Unsicherheiten ist, liegt auf der Hand. Aufgrund der hohen Übereinstimmung und Konsistenz zwischen globaler und regionaler Skala wird das Ergebnis des Vali-

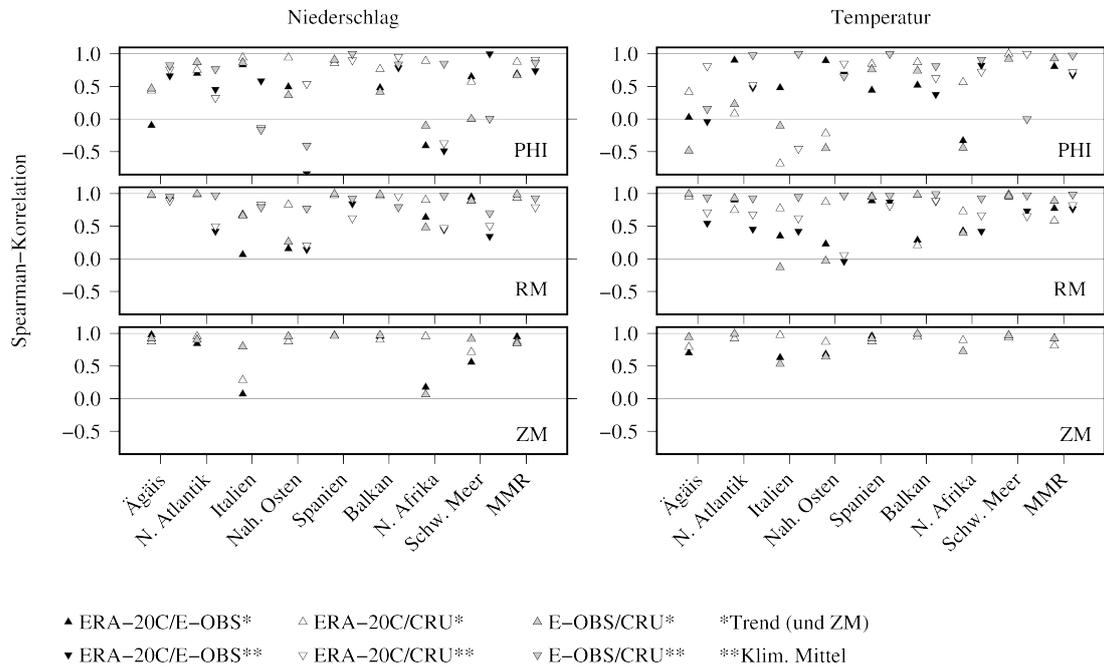
dierungsansatzes als positiv eingestuft. Allerdings sollte beachtet werden, dass bei der Auswertung lediglich die Tendenz der Änderung und nicht der Betrag berücksichtigt wurde. Daher muss darauf hingewiesen werden, dass die Trendänderungen in den meisten Fällen wie auch bei den Zukunftszeitreihen (vgl. Kapitel 6.1) nicht signifikant waren.

## 8.4 Abhängigkeit von Referenzdaten

Die Modellevaluationen in dieser Arbeit basierten je nach Region für den Mittelmeerraum auf E-OBS und für die Globalregionen auf ERA-20C. Da die Erweiterungsregionen nur vollständig von ERA-20C abgedeckt werden, wurde in Abschnitt 8.1 bereits ein kurzer Eindruck der Unterschiede bezüglich der allgemeinen Ranglisten des Mittelmeerraums basierend auf ERA-20C vorgestellt. E-OBS wird als sehr zuverlässiger Datensatz für Mitteleuropa häufig als Referenz verwendet (u.a. KOTLARSKI ET AL., 2014; HERTIG ET AL., 2013; HERTIG & JACOBET, 2014; PAXIAN ET AL., 2013), ist aber regional auch nur auf diesen begrenzt. Um eine Abschätzung der Unterschiede zwischen den Datensätzen und damit auch zur Einordnung der Ergebnisse der Globalregionen zu bieten, wurde die Evaluation der Hauptuntersuchungsregionen des Mittelmeerraums vollständig auch auf Grundlage von CRU und der ERA-20C-Reanalyse durchgeführt. Damit wird veranschaulicht, wie sich der Austausch der Referenzdaten auf die Bewertung der Modelle auswirkt. Abbildung 8.5 zeigt dazu exemplarisch die Spearman-Korrelation der DJF Modellranglisten jeder Region basierend auf jeweils zwei unterschiedlichen Referenzdatensätzen.

Für alle Metriken zeigt sich, dass der Großteil der Korrelationen positiv und meist im Bereich zwischen 0,5–1 liegt. Zwischen den einzelnen Werten existieren hier abhängig von der Situation teilweise deutliche Unterschiede. Am stabilsten sind die ZM-Ranglisten. Insbesondere für die Temperatur zeigen die Ranglisten für Regionen und Jahreszeiten Werte, die oberhalb von 0,5, meist sogar 0,8 liegen. Beim Niederschlag ist das Bild ähnlich, allerdings sind hier auch vereinzelt Werte um 0 z.B. für Italien und Nordafrika zu beobachten. Auffällig ist hier, dass auch bei der Temperatur die niedrigste Korrelation für Nordafrika berechnet wird. Insgesamt muss jedoch festgehalten werden, dass die Ranglisten über die absolute Mehrheit der Regionen sehr stabil und damit unempfindlich für die Wahl der Beobachtungsdaten sind.

Bei RM ist das Bild vergleichbar, allerdings zeigen sich hier vermehrt Korrelationen, die zwischen 0–0,5 liegen. Dabei besteht insgesamt eine hohe Übereinstimmung zwischen den Ergebnissen des Niederschlags sowie der Temperatur. Wie für ZM werden auch für RM die niedrigsten Korrelationen für Italien berechnet. Ana-



**Abbildung 8.5** Spearman-Korrelationen der Modellbewertungen (CMIP3, CMIP5, CORDEX) für DJF basierend auf unterschiedlichen Referenzdatensätzen. Angetragen sind die annualen Ergebnisse von Niederschlag (links) und Temperatur (rechts) für PHI (1. Zeile), RM (2.) und ZM (3.) in allen Regionen des Mittelmeerraums. Für PHI und RM sind die Ergebnisse des Trends (markiert wie ZM) sowie des klimatologischen Mittels angetragen.

log zu ZM ist auch für RM die Übereinstimmung zwischen ERA-20C und E-OBS am niedrigsten. Allerdings ist die Korrelation von E-OBS und CRU für Italien und den Nahen Osten ebenfalls niedrig. Davon abgesehen liegt der Großteil der Werte im Bereich zwischen 0,5–0,95. Die Maxima werden darüber hinaus zwischen E-OBS und CRU angegeben. Das absolute Minimum wird hier für den Trend berechnet. Ansonsten sind die Unterschiede zwischen Trend und klimatologischem Mittel eher untergeordnet.

PHI zeigt exemplarisch für alle 4FM die höchste Variabilität der Ergebnisse. Dabei gilt zunächst auch hier, dass der überwiegende Teil der Korrelationen positiv und im Bereich zwischen 0,3–1 liegt. Ansonsten zeigt sich beim Niederschlag ein ähnliches Bild wie bei RM. Die niedrigsten Korrelationen werden für Italien und den Nahen Osten sowie Nordafrika berechnet. Das absolute Minimum stellt sich erneut zwischen ERA-20C und E-OBS für Italien dar. Hier zeigen alle Metriken eine gewisse Übereinstimmung hinsichtlich der Einordnung der Referenzdaten. Bei der Temperatur ist die Verteilung vergleichbar. Hier kommt zu Italien, dem Nahen Osten und Nordafrika auch noch die Ägäis mit einer deut-

lich negativen Korrelation hinzu. Analog zu RM sind auch hier die Werte für die Kombination zwischen E-OBS und CRU auffällig niedrig. Anders als bei RM sind die Korrelationen des klimatologischen Mittels für die 4FM insgesamt meist höher als die des Trends und damit unempfindlicher für unterschiedliche Referenzdaten. Abgesehen von den Ausreißern bei der Temperatur sind die höchsten Korrelationen hier häufig für CRU und E-OBS zu finden. Hier handelt es sich in beiden Fällen um tatsächliche Stationsdaten, die entsprechend eine ähnliche Datengrundlage nutzen.

Auch bei den übrigen Metriken fällt auf, dass CRU und E-OBS meist die höchsten Korrelationen zeigen. Allerdings ist die Streuung bei RM und insbesondere PHI stärker ausgeprägt. Für DJA fällt in jeder Metrik auf, dass die Korrelationen des Gesamtmittelmeerraum auch bei den 4FM und den FPA in nahezu jeder Jahreszeit meist positiv mit über 0,6 sind. Dies könnte darauf hindeuten, dass größere Regionen noch unempfindlicher auf verschiedene Referenzdatentypen reagieren. Das in Abbildung 8.5 dargestellte Bild ist exemplarisch für die übrigen Jahreszeiten, wobei die exakten Werte variieren können. Für JJA werden für alle Metriken die insgesamt höchsten Korrelationen berechnet. Anders bei MAM und SON, wo situationsabhängig auch teilweise niedrigere Werte angenommen werden. Die Mehrheit der Werte bleibt jedoch im positiven Bereich bzw. überwiegend zwischen 0,5–1 stabil.

Insgesamt bleibt festzustellen, dass die Auswahl der Referenzdaten regional Einfluss auf die Bewertung der Klimamodelle hat. Dieses Ergebnis muss aufgrund unterschiedlicher Datenquellen und Ausgangsaufösungen als naheliegend erachtet werden. Allerdings hat sich auch gezeigt, dass ein überwiegender Anteil der Ranglisten unempfindlich gegenüber der Wahl der Referenzdatensätze ist. Dies ist besonders bei ZM und RM auffällig. Aber auch die 4FM zeigen in den meisten Situationen positive bis stark positive Korrelationen. Das Bild, das sich bei der Bewertung der Modelle im Verlauf dieser Studie herauskristallisiert hat, trifft offensichtlich auch auf die Beobachtungsdatensätze zu: Entscheidend für die Übereinstimmung der Referenzdaten sind Region, Jahreszeit und Variable. Die meisten dieser Situationen der drei untersuchten Datensätzen zeigen eine hohe Übereinstimmung. Eine Verallgemeinerung ist jedoch zumindest für die 4FM nicht möglich.

Als abschließende Untersuchung der Abhängigkeit der Metriken von den Beobachtungsdaten wurde analog zu Kapitel 5.4 die Spearman-Korrelation zwischen den Metrikranglisten aller untersuchten Situationen basierend auf den beiden anderen Referenzdatensätzen berechnet. Hier zeigte sich ein konsistentes Bild zu Abbildung 5.11 mit Korrelationen von über 95 % zwischen den 4FM (außer CHI) und hohen Korrelationen zu RM. Auch die ZM zeigt unabhängig vom zu Grunde liegenden Datensatz konstant keine Übereinstimmung zu den übrigen Metriken.

Dieses Resultat legt in Zusammenhang mit den in Abbildung 8.5 dargestellten Ergebnissen zwei mögliche Interpretationen nahe. Zum einen ähneln sich möglicherweise die drei verwendeten Referenzdatensätze im Mittelmeerraum so stark, dass hier keine großen Unterschiede in der Modellbewertung ersichtlich werden. Da diese Unterschiede jedoch bei den in Abbildung 8.5 exemplarisch gezeigten Korrelationen der Metrikranglisten unterschiedlich stark ausgeprägt sind, muss zum anderen die Wahl der Metrik ebenfalls eine Rolle spielen. Insbesondere ZM und RM zeigen so deutlich geringere Abweichungen als PHI, wobei gerade RM mit dem Trend dieselbe Größe analysiert wie PHI. Diese Unterschiede verdeutlichen, dass neben einer relativ hohen Übereinstimmung der Referenzdaten auch die Wahl der Metrik einen zumindest bedeutenden Anteil an der Abhängigkeit von Referenzdaten hat. So zeigt sich, dass die Wahl der Referenzdaten mitentscheidend für die Bewertung der Modelle ist. Für die Regionen des Mittelmeerraums attestieren die Evaluationsergebnisse der analysierten Metriken den untersuchten Datensätzen eine hohe Korrelation und damit relativ stabile Modellbewertungen.

## 9 Zusammenfassung und Diskussion

Das Ziel dieser Arbeit war die Entwicklung und der Vergleich neuer Metriken zur Evaluation von Klimamodellen. Zusätzlich wurden die so gewonnenen Informationen genutzt, um die Zukunftsprojektionen metrikgestützt zu gewichten und so eine Veränderung des Unsicherheitsbereichs der Klimaprojektionen zu erreichen. Obwohl die Evaluationsbedingungen zu keiner anderen Studie identisch sind, können viele Ergebnisse in den Kontext bisheriger Arbeiten eingeordnet werden. Dabei zeigen sich neben Übereinstimmungen auch Unterschiede und Herausforderungen der Modellbewertung.

Generell erfüllen die Metriken ihre Aufgabe, eine unabhängige und vergleichbare Bewertung von Klimamodellen zu ermöglichen. Dies gilt sowohl für die regionalen CORDEX- als auch die globalen CMIP3- und CMIP5-Klimamodelle. Die Ergebnisse der Metrikevaluationen legen nahe, dass der Niederschlag die deutlich größere Herausforderung für die untersuchten Klimamodelle ist. Dies gilt unabhängig von der Modellgeneration. Sowohl in CMIP3 als auch CMIP5 und CORDEX haben sich jedoch Modelle herauskristallisiert, die auch für Niederschlag sehr hohe Gewichte durch die unterschiedlichen Metriken zugeteilt bekommen haben. Die allgemein niedrigere Simulationsgüte für Niederschlag hat sich als großes Potential für Gewichtung herausgestellt. Während man von den geringen Unterschieden der Modelle bei Temperatur von einer bereits bestehenden hohen Sicherheit der Szenarien ausgehen kann, ist dies beim Niederschlag durch Gewichtung eher zu verbessern.

Die Auswertung verschiedener Regionen bestätigt GIORGI (2006), dass es sich beim Mittelmeerraum tatsächlich um eine durch den Klimawandel besonders bedrohte Region handelt. Die hohen Unsicherheiten zwischen den Modellen verdeutlichen dies durch große Konfidenzintervalle. Die Unterschiede der Gewichtungen sind vor diesem Hintergrund in dieser Region umso relevanter, da sich hier deutliche Leistungsabstufungen zwischen den Modellen herausgestellt haben. Dies gilt für die globalen Regionen nur beim Niederschlag und dort auch lediglich in abgeschwächter Form. Allgemein wurde den Modellen für die größeren Regionen eine bessere Leistung bei gleichzeitig schwächeren Unterschieden zwischen den Simulationen attestiert.

Die Evaluation der regionalen und globalen Klimamodelle in verschiedenen Regionen hat gezeigt, dass jede Metrik für sich in der Lage ist, nachvollziehbare Bewertungen der Simulationsgüte zu erstellen. Dabei waren insbesondere bezogen auf 4FM, RM und FPA viele Übereinstimmungen zu erkennen. Die ZM hat dagegen aufgrund einer abgewandelten Datengrundlage gezeigt, dass auch völlig andere Gewichte für die Modelle möglich sind. Die grundsätzliche Übertragbarkeit der Metriken auf unterschiedliche Variablen, Regionen und Skalenniveaus verbindet jedoch alle Metriken und empfiehlt sie so für weitere Fragestellungen und Variablen.

Die angewandten Metriken haben eine ähnliche oder verbesserte Simulationsgüte von CMIP5 gegenüber der von CMIP3 ergeben. Letztere zeigte sich zumeist bei der Simulation der saisonalen Niederschlagssumme sowie der Temperaturmittel. Bezogen auf die Trends war überwiegend keine eindeutige Tendenz zu beobachten. Ähnliche Ergebnisse finden sich in den meisten aktuellen Studien wieder (u.a. GROSE ET AL., 2014; ZHANG ET AL., 2013; KUMAR ET AL., 2014). KOUTROULIS ET AL. (2016) zeigen beispielsweise für die Evaluation der CMIP5- und CMIP3-Modelle für Niederschlag und Temperatur über 26 unterschiedliche Regionen insgesamt eine deutliche Steigerung der Simulationsleistung von CMIP5. Auch wenn es große Schnittmengen gibt, sehen auch einige Studien keine eindeutigen oder andere Ergebnisse. RAMESH & GOSWAMI (2015) beschreiben für den Indischen Monsun basierend auf Evaluation historischer Trends keine Steigerung des Vertrauens für die neuere Modellgeneration. Stattdessen werden Unterschiede in der Simulationsleistung einzelner Modelle deutlich. Daher schlagen RAMESH & GOSWAMI (2015) die Wahl von Modellen mit hoher Simulationsgüte für spezifische Fragestellungen vor. Auch KNUTTI & SEDLÁČEK (2012) sehen insgesamt eine vergleichbare Entwicklung der Unsicherheit bezogen auf die globale Mitteltemperatur und die Niederschlagsmuster von CMIP3 und CMIP5. Sie stellen heraus, dass insbesondere die Effekte, die auf eine Reduzierung der Unsicherheit abzielen, zu gering ausfallen, um erhöhtes Vertrauen in die Simulationen haben zu können. Dennoch legen auch KNUTTI & SEDLÁČEK (2012) nahe, dass dies aufgrund des generell hohen Vertrauens in die Simulationen kein Hindernis für das Festlegen politischer Entscheidungen sein sollte.

Die in dieser Arbeit entwickelten Ergebnisse fügen sich in den Kontext aktueller Studien ein. Keine Metrik zeigte für jede Situation Verbesserungen und in vielen Fällen wurden nur geringe Unterschiede zwischen den Modellgenerationen oder den separaten Simulationsrohgewichten beobachtet. Der Grund für die Unstimmigkeiten zu manchen Evaluationsansätzen scheint in der Art der Metriken bzw. der untersuchten Variablen zu liegen. Die Frage nach der Validität einer Evaluationsmetrik ist vor diesem Hintergrund hochsensibel. WEIGEL ET AL. (2010) bemerken, dass fehlerhafte Annahmen bei der Gewichtung zu einer Ver-

schlechterung im Vergleich zur Gleichgewichtung führen können. Da kein universell gültiger Ansatz oder anerkannte Metriken zur objektiven Evaluation von Klimamodellen existieren (KNUTTI ET AL., 2010; GLECKLER ET AL., 2008), wurden unterschiedliche Herangehensweisen an diese Problematik angewandt. Neben einem experimentellen Ansatz einer Metrikvalidierung (vgl. Kapitel 8.3) und dem Vergleich der Gewichte von unterschiedlichen Evaluationszeiträumen (vgl. Kapitel 8.2) diente der Vergleich der Simulationsgüte von CMIP3 und CMIP5 in dieser Arbeit dazu, ein möglichst umfassendes Bild der Metrikvalidität darstellen zu können. Während der Validierungsansatz sowie die Analyse der Modellgenerationen vielversprechende Ergebnisse lieferten, ist eine zeitliche Konsistenz der Modellbewertung nicht generell gegeben. So zeigte der Vergleich zweier leicht versetzter Evaluationszeiträume für das klimatologische Mittel zwar stabile Ergebnisse, für die Analyse der Zeitreihen und des Trends konnte dies jedoch für alle Metriken in vielen Situationen nicht bestätigt werden. Allerdings muss generell hinterfragt werden, ob Modellsimulationen zu zwei Zeiträumen dieselbe Übereinstimmung mit Referenzdaten haben können. Die Evaluation der einzelnen Läufe hat eine sehr hohe Abhängigkeit von der Region, der Variable und der Jahreszeit ergeben. Daher kann auch ein leicht versetzter Evaluationszeitraum gerade bei einer zeitabhängigen Untersuchungsgröße wie dem Trend zu einer Abänderung der jeweiligen Simulationsgüte führen. Da dieses Ergebnis von allen Metriken für den Trend übereinstimmend für alle analysierten Regionen abgebildet wird, muss dies daher als wichtige Erkenntnis der Auswertung von Klimamodellen angesehen werden. Dieses Resultat ist jedoch insbesondere Angesichts der hohen Überschneidung der beiden untersuchten 50-jährigen Zeiträume von 80 % auffällig. Die Variabilität innerhalb der Referenzdaten scheint so stark ausgeprägt zu sein, dass bereits leichte Verschiebungen zu deutlichen Änderungen des Trends führen können und sich damit auf die Modellbewertung auswirken. Von zusätzlichem Interesse wäre es daher übereinstimmend mit PAXIAN ET AL. (2013) noch deutlich längere Zeitintervalle, die beispielsweise doppelt so lang sind wie die in dieser Arbeit, zu evaluieren, um die Erkenntnis der fehlenden Übereinstimmung zu bestätigen. Zwei so lange unabhängige Zeitintervalle sind aufgrund der Referenzdatensätze jedoch nicht möglich. Auch wenn CRU und ERA-20C einen relativ langen Zeitraum abdecken, ist auf absehbare Zeit kein zuverlässiges zweites Zeitintervall verfügbar, um die Konsistenzprüfung vornehmen zu können. Um dennoch eine allgemeingültige Aussage über Simulationsgüte treffen zu können, wäre es ein Lösungsansatz, eine umfassende Datenbank von Modellbewertungen unterschiedlicher Zeiträume und Regionen zu erstellen. Welches historische Zeitintervall am geeignetsten ist, um für das zukünftige Klima die beste Evaluationsbasis zu bieten, kann nicht mit Sicherheit gesagt werden. Da eher mit langfristigen Veränderungen wie dem Temperaturanstieg zu rechnen ist, bieten sich allerdings möglichst aktuelle Zeiträume

an, die bereits stark vom anthropogenen Einfluss betroffen sind und geringe natürliche Variabilität zeigen. Dazu passt der Eindruck, dass die Rohgewichte der optimalen Fingerprints (siehe Abbildung 8.2) insgesamt deutlich weniger negative Korrelationen und etwas höhere Trendwerte annehmen. Zum Ziel einer umfassenden Datengrundlage liefert diese Arbeit einen ersten Schritt, da die letzten 50-Jahre bis zum Jahr 2009 für 15 Regionen und verschiedene Variablen betrachtet wurden. Anschließende Studien könnten diesen Weg weiterverfolgen und zum Vergleich nutzen.

Die Analyse der regionalen Klimamodelle von CORDEX zeigte ein vergleichbares Bild zu dem der globalen Klimamodelle von CMIP3 und CMIP5. Auch wenn die RKM aufgrund ihrer sehr hohen räumlichen Auflösung ( $0,11^\circ$ ) großes Potential für die Klimaforschung haben (DI LUCA ET AL., 2013; FESER ET AL., 2011; RUMMUKAINEN, 2010; CHRISTENSEN ET AL., 2007), waren auch hier heterogene Ergebnisse zu beobachten. Dennoch zeigten sich trotz Abweichungen in der Länge des Untersuchungszeitraums der CORDEX-Simulationen vielversprechende Ergebnisse. Die schwächeren Bewertungen des Trends können mit dem unterschiedlichen Zeitintervall erklärt werden. Bei Tests mit einem verkürzten Zeitraum aller Modelle lagen die Ergebnisse meist auf einem Level mit den GKM oder darüber. Allgemein zeigen die Metriken überwiegend eine bessere Leistung bei der Simulation des klimatologischen Mittels von Niederschlag und Temperatur. Hier wurde die Temperatur zu jeder Zeit nochmals stärker eingeschätzt als der Niederschlag. LI ET AL. (2016) kommen für den Niederschlag für die zweite Hälfte des 20. Jahrhunderts zu einem ähnlichen Ergebnis mit einer deutlich schwächeren Übereinstimmung des Trends mit dem der Referenzdaten. Das wird darauf zurückgeführt, dass letztere eher durch interne Variationen des Klimasystems beeinflusst sind als durch anthropogene oder externe Forcings. Dies deckt sich mit PAXIAN ET AL. (2013) und dient als eine Erklärung dafür, warum ähnliche starke Diskrepanzen zwischen Temperaturmittel und dem stark anthropogen beeinflussten Temperaturtrend nicht zu beobachten sind.

Generell zeigten die 4FM und FPA überwiegend sehr hohe Modellbewertungen für die Temperatur und eine etwas niedrigere Leistung bezüglich des Niederschlags. Analog zeigen SU ET AL. (2013) für das Tibetplateau, dass alle 24 untersuchten GKM Schwierigkeiten hatten, das Niederschlagsmuster zuverlässig zu reproduzieren. Diese Ergebnisse decken sich mit den Ergebnissen von PERKINS ET AL. (2007), KUMAR ET AL. (2013) und MIAO ET AL. (2012). Auch SU ET AL. (2013) empfehlen, dass zur Steigerung der Zuverlässigkeit der Klimamodelle ein Schwerpunkt auf die Verbesserung der Niederschlagssimulation gelegt werden sollte. Diese Empfehlung kann auch auf Grundlage der Evaluationsergebnisse in dieser Arbeit gegeben werden. Aufgrund der generellen Übereinstimmung mit den Ergebnissen anderer Studien sowie des positiven Ergebnisses des Vali-

dierungsansatzes werden die in dieser Arbeit entwickelten Metriken daher als geeignet zur Bewertung von Klimamodellen eingestuft. Da für RM und ZM kein vordefinierter Wertebereich wie bei den anderen Metrikgruppen existiert, mussten diese gesondert untersucht werden. Die hohe positive Spearman-Korrelation der RM-Bewertungen sowohl mit den FPA- (für den Trend) als auch den 4FM-Ergebnissen (für Trend und Mittel) legen den Schluss nahe, dass auch RM ein geeigneter Schätzer für die Simulationsgüte ist. Lediglich ZM zeigte weder bezogen auf den Trend noch auf das Mittel eine positive oder negative Korrelation mit den anderen Metriken. Hier muss in Betracht gezogen werden, dass sie sich nicht als geeigneter Schätzer der Simulationsgüte eignet. Allerdings handelt es sich bei der ZM um die einzige Metrik, die weder Trend noch Mittel als Grundlage ihrer Evaluation verwendet. In der Folge könnte die fehlende Übereinstimmung weniger an einer ungeeigneten Metrik als vielmehr einer anderen untersuchten Klimagröße der Modelle liegen. Dies würde mit der zuvor beschriebenen Dissonanz der Modelleinschätzungen basierend auf Temperatur und Niederschlag oder klimatologischen Mittel und Trend übereinstimmen. Die 4FM und RM zeigen, dass es keinerlei Zusammenhang zwischen den Rangfolgen basierend auf dem Trend mit denen des klimatologischen Mittels gibt. Eine hohe Bewertung der Modelle für eine der untersuchten Variablen lässt keine Rückschlüsse auf die Simulationsgüte einer anderen Variablen zu. Dies gilt sowohl im Vergleich von Niederschlag und Temperatur als auch klimatologischem Mittel und Trend und deckt sich mit den Ergebnissen von LI ET AL. (2016). Auch sie kommen zu dem Schluss, dass es deutliche Unterschiede zwischen der Simulationsgüte des saisonalen Niederschlagsmittels und dem Trend gibt. In diesem Zusammenhang fügt sich auch die Zeitreihenmetrik ein. Infolgedessen werden ihre Ergebnisse und Rangfolgen nicht als ungeeignet eingestuft, sondern die Zeitreihenevaluation wird als ein weiterer Untersuchungsaspekt betrachtet, der von den Klimamodellen unterschiedlich stark simuliert wird. Da auch hier bei der Temperatur insbesondere für JJA eine Verbesserung des neueren MME zu beobachten war, scheint auch diese Metrik Potential zur Modellevaluation zu besitzen. Da sich ihre Ergebnisse jedoch nicht so leicht in den Kontext anderer Studien einordnen lassen wie die der anderen Metriken, sind hier weitergehende Untersuchungen erforderlich, um eine abschließende Bewertung der Metrik vornehmen zu können.

Bei der Evaluation der Hauptuntersuchungsgebiete des Mittelmeerraums und der globalen Erweiterungsregionen haben sich keine Modelle herauskristallisiert, die überall überzeugen konnten. Vielmehr zeigten sich in verschiedenen Untersuchungen verschiedene Modelle als besonders geeignet (HUANG ET AL., 2013; LIU ET AL., 2014; RAMESH & GOSWAMI, 2015). Da in dieser Arbeit vier große Metrikblöcke für zwei Variablen auf unterschiedlichen Untersuchungsebenen vereint wurden, wurde dennoch jeweils eine Gesamtrangfolge erstellt, um die Übertrag-

barkeit zu testen und eine mittlere Modellbewertung abschätzen zu können. Der Vergleich der Rangfolgen des Mittelmeerraums und der Globalregionen zeigte für den Niederschlag keinerlei Zusammenhang. Für die Temperatur wurde mit 0,39 jedoch eine signifikante Korrelation festgestellt. Dieses Ergebnis spiegelt nur einen kleinen Teilausschnitt der Modellqualitäten wider. Allerdings gilt, dass zumindest unter den in dieser Arbeit untersuchten Bedingungen eine relativ konsistente Gruppe von Modellen beobachtet werden konnte, die die Temperatur in unterschiedlichen Regionen sowie auf unterschiedlichen Skalen ähnlich gut simulieren.

Neben der Metrikerstellung und dem -vergleich war die Anpassung des Unsicherheitsbereichs der dritte Schwerpunkt dieser Arbeit. Die Frage nach der Modellgewichtung bei der Erstellung eines MME zur sinnvollen Beeinflussung des Unsicherheitsbereichs von Zukunftsprojektionen ist ein aktuelles Thema in der Klimaforschung. In den letzten Sachstandsberichten des IPCC (IPCC, 2007c, 2013) werden die Modelle demokratisch gleich gewichtet. Aktuelle Studien zeigen vermehrt Tendenzen, die zu einer leistungsorientierten Bewertung von Modellen und entsprechend ungleicher Gewichtung oder Sub-Ensembles führen (LEDUC ET AL., 2016; SANDERSON ET AL., 2015b,a). So schlagen beispielsweise WALSH ET AL. (2008) auch für die Arktis und Grönland vor, durch sinnvolle Vorauswahl nur bestimmte Modelle zu verwenden, um Unsicherheit zu verringern. Dies wird auch für eine Vielzahl weiterer Regionen von RAMESH & GOSWAMI (2015) vorgeschlagen. Vor diesem Hintergrund stellt sich die Frage wie im kommenden Sachstandsbericht (MEEHL ET AL., 2014; O'NEILL ET AL., 2016) mit der Thematik verfahren werden sollte. Die Auswertung der Klimamodellergebnisse bleibt unabhängig davon, ob es zukünftig zur metrikbasierten Untersuchung des Klimawandels kommt, ein elementarer Teil der Klimaforschung, wie die umfassende Anzahl von Studien, die sich dieser Aufgabe widmen, belegt (u.a. SILLMANN ET AL., 2013; BACHNER ET AL., 2008; DANANDEH MEHR & KAHYA, 2016; CHEN & FRAUENFELD, 2014; EMANUEL & ŽIVKOVIĆ-ROTHMAN, 1999). Dass sowohl die gleichgewichtete Verwendung von allen Modellen als auch eine metrikgewichtete Betrachtung der Unsicherheit Vor- wie auch Nachteile mit sich bringen, kann jedoch nicht bestritten werden. So zeigen CHRISTENSEN ET AL. (2010) bei der Gewichtung von regionalen Klimamodellen, dass der Erfolg von Gewichtung unter anderem stark von der Wahl der Evaluationsmetrik abhängig ist. Danach muss metrikbasierte Gewichtung nicht zwingend einen positiven Effekt gegenüber der Gleichgewichtung haben. Vielmehr entstehe durch die Gewichtung eine eigene Unsicherheit, die es weiter zu untersuchen gelte, bevor die Frage, was zu bevorzugen sei, beantwortet werden könne. Dagegen verweisen beispielsweise HAUGHTON ET AL. (2015) und GILLET (2015) eher auf die Vorteile der Metrikgewichtung der MME. Insgesamt hat sich gezeigt, dass bereits der gleichgewichtete Unsi-

cherheitsbereich der Modelle bezogen auf die Klimaänderung des Niederschlags deutlich größer als der Unsicherheitsbereich der Temperatur ist. Zum gleichen Resultat kommen auch HAWKINS & SUTTON (2009) und KHARIN ET AL. (2007, 2013).

Die Anwendung der Gewichte auf die Zukunftsprojektionen führte in dieser Arbeit zu einem heterogenen Bild. Das Ziel, durch leistungsorientierte Gewichtung der Modelle die Unsicherheit der Zukunftsprojektionen zu verringern, konnte pauschal nicht erfüllt werden. Stattdessen kam es sowohl zur Verringerung als auch zur Vergrößerung der Unsicherheitsbereiche. Entgegen WEIGEL ET AL. (2010) stimmte die Mehrheit der Metriken hier in der Art der Verschiebung überein. Stattdessen war insbesondere bei der Intensität der Änderung des Unsicherheitsbereichs die Wahl der Metrik entscheidend. So wurden die meisten signifikanten Verschiebungen sowohl des Erwartungswertes als auch der Standardabweichung der WDF durch die Verwendung der LOR-Gewichte erzeugt. Diese haben gegenüber den anderen Metriken den Vorteil, dass der Wertebereich nicht auf ein vordefiniertes Intervall wie beispielsweise bei den 4FM (außer LOR) oder den FPA beschränkt ist. Zum Test wurde für die meisten Metriken (außer RM und ZM) zusätzlich ein intensivierter Ansatz angewandt, der ein Sub-Ensemble und damit stärkere Gewichte erzeugt. Die Art der Verschiebung blieb durch diese Anpassung meist stabil. Im Ergebnis konnten so jedoch auch mit Hilfe der anderen Metriken deutlich mehr signifikante Änderungen des Unsicherheitsbereichs realisiert werden. Hier sind für die Globalregionen insbesondere die  $FPA_z$  zu nennen, die neben der Intensivierung der Gewichte zusätzlich eine größere Informationsgrundlage nutzten. Insgesamt hat sich der Niederschlag aufgrund ausgeprägterer Unterschiede zwischen den Modellergebnissen als sensitiver für die Gewichtung herausgestellt. Dadurch waren deutlichere Abstufungen möglich und in der Folge stärkere Gewichte. Insgesamt führte der überwiegende Teil der signifikanten Ergebnisse zu einer Reduktion des Unsicherheitsbereichs. Allerdings können auch die Vergrößerungen des Unsicherheitsbereichs für die Entwicklung von Anpassungsstrategien hohe Relevanz haben. Die Bedeutung metrikbasierter Gewichtung sollte daher nicht ignoriert werden. Dabei gilt jedoch, dass die Art und Weise der Änderung stark von der jeweiligen Region, Saison oder Variablen abhängig ist. Die meisten der entwickelten Metriken sind zudem besser dazu geeignet, eine Rangliste der Simulationsleistung zu erstellen, als basierend auf ihren Rohgewichten zur Gewichtung verwendet zu werden. Angesichts der insgesamt relativ hohen Simulationsgüte ist die Stärke des Einflusses der Gewichtung auf die MMEs durch den von den Metriken vorgegebenen Wertebereich beschränkt. Durch weitere Evaluation und Intensivierung der Gewichte könnte dies zwar verbessert werden, allerdings ist diese Art der Veränderung der Rohgewichte ohne weitere Analysen eher als experimentell anzusehen. Dies bietet ein großes Potential für weitere

Studien. So könnten die umfangreichen Ergebnisse der Modellbewertungen weiter evaluiert werden, um basierend auf einem allgemeinen Gewicht eine fundierte Aussage über die Abstufungen zwischen den Modellqualitäten treffen zu können. Die Gesamtrangliste in Abschnitt 8.1 bietet hierzu möglicherweise einen ersten Schritt.

Neben der Abhängigkeit von Regionen und Jahreszeiten wurde in dieser Arbeit auch die Bedeutung der Referenzdaten betrachtet. Zur Evaluation auf Basis des Referenzdatensatzes E-OBS wurde die gesamte Untersuchung für den Mittelmeerraum auch basierend auf CRU und der ERA-20C-Reanalyse durchgeführt. Vorangegangene Studien zeigen teilweise große Abweichungen unterschiedlicher Beobachtungsdatentypen (FEKETE ET AL., 2004; RING ET AL., 2016). Neben Unsicherheiten bei der Erzeugung der Datenprodukte können auch Messfehler an Stationen die Anwendung beeinflussen (LIU ET AL., 2015). Außerdem ist die Stationsdichte ein mitentscheidender Faktor für die Unsicherheit von Beobachtungsdaten. Dies wirkt sich insbesondere auf die Datengrundlage der in dieser Hinsicht schlechter erschlossenen Gebiete wie die Sahara oder Zentral Asien negativ aus (SCHNEIDER ET AL., 2008). Die Abhängigkeit der Modellbewertungen hat sich jedoch als geringer herausgestellt als zunächst erwartet. Insbesondere die Zeitreiheneigenschaften der verschiedenen Datensätze scheinen eine hohe Ähnlichkeit zu besitzen, da ZM nahezu in allen Situationen sehr hohe Korrelationen der Modellbewertungen für unterschiedliche Referenzdatensätze zeigt. Auch die Trendmuster und klimatologischen Mittel zeigten überwiegend hohe Korrelationen, wobei dies stärker für RM als für die 4FM gilt. Neben regionalen Unterschieden wurden auch Muster zwischen den Variablen deutlich. So zeigte insbesondere die Auswertung der Ranglisten basierend auf unterschiedlichen Referenzdaten für Niederschlag hohe negative Korrelationen. Die Ergebnisse der Temperatur waren hingegen zumeist positiv. Dies deutet darauf hin, dass auch bei den Beobachtungsdaten der Niederschlag mit erhöhter Unsicherheit versehen ist. Dies bestätigen auch MIAO ET AL. (2012) und KUMAR ET AL. (2013) gerade für den Niederschlag der niederen Breiten. Letztendlich zeigen die Ergebnisse, dass die Wahl der Referenzdaten eine mitentscheidende Rolle bei der Modellevaluation spielt. Der Einfluss der Wahl auf das generell heterogene Bild der Simulationsgüte unterschiedlicher Modelle ist nach den Ergebnissen dieser Arbeit jedoch untergeordnet.

## 10 Ausblick

Die Ergebnisse in dieser Studie zeigen trotz des großen Umfangs von 13 untersuchten Einzelmetriken in insgesamt 15 Untersuchungsgebieten in vieler Hinsicht eine hohe Konsistenz. Darüber hinaus haben sich die Ergebnisse zumeist gut in den Kontext aktueller Studien einordnen lassen und deren Ergebnisse bestätigt oder ergänzt. Die einfache Übertragbarkeit aller Metriken für unterschiedliche Situationen bei gleichzeitig hoher Übereinstimmung untereinander haben zusätzlich vielversprechende Perspektiven für weitere Anwendungen aufgezeigt. Dabei besteht noch weiteres Potential, die Metriken auf andere Fragestellungen wie beispielsweise Extremereignisse oder weitere Variablen anzuwenden, um die Gültigkeit der beobachteten Ergebnisse besser einordnen zu können.

Ob die metrikbasierte Gewichtung von Multi-Modell Ensembles die beste Lösung ist, zuverlässigere Aussagen über den Unsicherheitsbereich der Klimaänderung zu erhalten, kann nicht zweifelsfrei beantwortet werden. Zwar zeigten sich viele Situationen insbesondere beim Niederschlag, in denen die Gewichtung zu einer signifikanten Verringerung der Unsicherheit bzw. zu einer Verschiebung des Erwartungswertes geführt hat, allerdings war die überwiegende Mehrheit der Situationen nicht statistisch signifikant.

Für die meisten Situationen gab es Modelle, die hohe Ähnlichkeiten mit den Referenzdaten besaßen. Die Übereinstimmung der Modelle hinsichtlich des Temperaturanstiegs zum Ende des 21. Jahrhunderts ist darüber hinaus so hoch, dass die Unsicherheit hier hauptsächlich auf dem menschlichen Verhalten verdeutlicht durch die vier Emissionsszenarien beruht. Allerdings bleibt festzustellen, dass es in den unterschiedlichen Regionen stark unterschiedliche Leistungen der Modelle gibt. Die Metriken zeigen, dass die Simulationen mancher Modelle deutlich näher an denen der Referenzdaten liegen als an den Simulationen anderer Modelle. So muss die Frage beantwortet werden, wie mit der Erkenntnis umgegangen werden soll, dass manche Modelle bestimmte Situationen besser simulieren als andere.

Die Ergebnisse dieser Studie legen nahe, dass Gewichtungsmetriken als Möglichkeit genutzt werden sollten, den Unsicherheitsbereich anzupassen. Dass es in einem überwiegenden Teil der Situationen keine signifikante Änderung der Unsicherheit gegeben hat, ist der Struktur der Metriken geschuldet. LOR und die zusätzlichen Fingerprint-Ansätze haben sich insgesamt als stärkste Gewichtungswerkzeuge herausgestellt. Neben der eigentlichen Änderung des Unsicher-

heitsbereichs kann die fortgeführte Evaluation der Metriken basierend auf deren Leistung einen positiven Effekt auf die Entwicklung neuer Klimamodelle bzw. die Verbesserung aktueller Modelle haben. Gerade durch die vielfältige Analyse der Simulationsleistung für verschiedene Regionen, Variablen etc. können Stärken und Schwächen von Modellen aufgedeckt und übertragen oder verbessert werden. Aus diesem Grund sollten die Ergebnisse von Klimamodellen intensiv evaluiert und anschließend gewichtet werden. Dazu bieten die hier entwickelten Metriken eine Vergleichsmöglichkeit für folgende Studien mit anderen Ansätzen oder neuen Variablen und Regionen.

Diese Arbeit könnte so als ein Mosaikstein im Gesamtbild der Modelleinschätzung dazu beitragen, Konsistenz in die Simulationsbewertung zu bringen, Schwächen oder Stärken aufzuzeigen und so die Unsicherheit der Zukunftsprojektionen situationsabhängig sinnvoll zu verändern. Das könnte ein Weg sein, unvoreingenommen über den aktuellen Stand der Klimaforschung und den anthropogenen Klimawandel zu informieren und transparent die Leistungsfähigkeit aktueller globaler und regionaler Modelle aufzuzeigen, um damit mehr Vertrauen in die Zukunftsprojektionen zu erzeugen. So besteht die Chance, dass die Klimaprojektionen das richtige Echo und Bewusstsein in der Politik und Bevölkerung finden und sinnvolle, reflektierte Anpassungsstrategien zeitnah entwickelt und umgesetzt werden.

## Literaturverzeichnis

- ABE-OUCHI, A., EMORI, S., HASEGAWA, A., HASUMI, H., INQUE, T., KIMOTO, M., MATSUMURA, S., NAGASHIMA, T., NAKANO, H., NISHIMURA, T., NOZAWA, T., OHGAITO, R., OKA, A., OKADA, N., OGOCHI, K., OGURA, T., SAITO, F., SAITO, K., SAKAMATO, T., SEGAWA, T., SOGA, T., SUDO, K., SUMI, A., SUZUKI, T., SUZUKI, T., TAKAHASHI, H., TAKATA, K., TAKEMURA, T., TAKIGAWA, M., TSUSHIMA, Y., WATANABE, M., WATANABE, S. & YOKOHATA, T. (2004): *K-1 Coupled GCM (MIROC) Description*. Techn. Ber., 1, Center for Climate Research, University of Tokyo.
- ALEKSEEV, V. A., VOLODIN, E. M., GALIN, V. Y., DYMNIKOV, V. P. & LYKOSOV, V. N. (1998): *Modelling of the present-day climate by the INM RAS atmospheric model DNM GCM*. Institute of numerical Mathematics, Moscow, Russia.
- ALLEN, C. D., MACALADY, A. K., CHENCHOUNI, H., BACHELET, D., MCDOWELL, N., VENNETIER, M., KITZBERGER, T., RIGLING, A., BRESHEARS, D. D., HOGG, E. H. T., GONZALEZ, P., FENSHAM, R., ZHANG, Z., CASTRO, J., DEMIDOVA, N., LIM, J. H., ALLARD, G., RUNNING, S. W., SEMERCI, A. & COBB, N. (2010): *A global overview of drought and heat-induced tree mortality reveals emerging climate change risks for forests*. In: *For. Ecol. Manage.*, **259**, 4: 660–684.
- ALLEN, M. R., FRAME, D. J., HUNTINGFORD, C., JONES, C. D., LOWE, J. A., MEINSHAUSEN, M. & MEINSHAUSEN, N. (2009): *Warming caused by cumulative carbon emissions towards the trillionth tonne*. In: *Nature*, **458**, 7242: 1163–1166.
- ALLEN, M. R. & TETT, S. F. B. (1999): *Checking for model consistency in optimal fingerprinting*. In: *Clim. Dyn.*, **15**, 6: 419–434.
- ANSTEY, J. A., DAVINI, P., GRAY, L. J., WOOLLINGS, T. J., BUTCHART, N., CAGNAZZO, C., CHRISTIANSEN, B., HARDIMAN, S. C., OSPREY, S. M. & YANG, S. (2013): *Multi-model analysis of Northern Hemisphere winter blocking: Model biases and the role of resolution*. In: *J. Geophys. Res. Atmos.*, **118**, 10: 3956–3971.

- ASADI ZARCH, M. A., SIVAKUMAR, B. & SHARMA, A. (2015): *Droughts in a warming climate: A global assessment of Standardized precipitation index (SPI) and Reconnaissance drought index (RDI)*. In: *J. Hydrol.*, **526**: 183–195.
- ASSENG, S., EWERT, F., ROSENZWEIG, C., JONES, J. W., HATFIELD, J. L., RUANE, A. C., BOOTE, K. J., THORBURN, P. J., RÖTTER, R. P., CAMMARANO, D., BRISSON, N., BASSO, B., MARTRE, P., AGGARWAL, P. K., ANGULO, C., BERTUZZI, P., BIERNATH, C., CHALLINOR, A. J., DOLTRA, J., GAYLER, S., GOLDBERG, R., GRANT, R., HENG, L., HOOKER, J., HUNT, L. A., INGWERSEN, J., IZAURRALDE, R. C., KERSEBAUM, K. C., MÜLLER, C., NARESH KUMAR, S., NENDEL, C., O’LEARY, G., OLESEN, J. E., OSBORNE, T. M., PALOSUO, T., PRIESACK, E., RIPOCHE, D., SEMENOV, M. A., SHCHERBAK, I., STEDUTO, P., STÖCKLE, C., STRATONOVITCH, P., STRECK, T., SUPIT, I., TAO, F., TRAVASSO, M., WAHA, K., WALLACH, D., WHITE, J. W., WILLIAMS, J. R. & WOLF, J. (2013): *Uncertainty in simulating wheat yields under climate change*. In: *Nat. Clim. Chang.*, **3**, 9: 827–832.
- BABAK, O. & DEUTSCH, C. V. (2009): *Statistical approach to inverse distance interpolation*. In: *Stoch. Environ. Res. Risk Assess.*, **23**, 5: 543–553.
- BACHNER, S., KAPALA, A. & SIMMER, C. (2008): *Evaluation of daily precipitation characteristics in the CLM and their sensitivity to parameterizations*. In: *Meteorol. Zeitschrift*, **17**, 4: 407–419.
- BAHRENBERG, G., GIESE, E., MEVENKAMP, N. & NIPPER, J. (2003): *Statistische Methoden in der Geographie. Band 2: Multivariate Statistik*. Borntraeger, Berlin.
- BAHRENBERG, G., GIESE, E., MEVENKAMP, N. & NIPPER, J. (2010): *Statistische Methoden in der Geographie. Band 1: Univariate und bivariate Statistik*. 5. Aufl. Borntraeger, Stuttgart.
- BALDI, M., DALU, G., MARACCHI, G., PASQUI, M. & CESARONE, F. (2006): *Heat waves in the Mediterranean: A local feature or a larger-scale effect?* In: *Int. J. Climatol.*, **26**, 11: 1477–1487.
- BECKER, P., IMBERY, F., FRIEDRICH, K., RAUTHE, M., MATZARAKIS, A., GRÄTZ, A. & JANSSEN, W. (2015): *Klimatologische Einschätzung des Sommer 2015*. Techn. Ber., Deutscher Wetterdienst (DWD). URL [https://www.dwd.de/DE/leistungen/besondereereignisse/temperatur/20151013\\_bericht\\_sommer\\_2015.pdf?\\_\\_blob=publicationFile&v=5\(16.08.2017\)](https://www.dwd.de/DE/leistungen/besondereereignisse/temperatur/20151013_bericht_sommer_2015.pdf?__blob=publicationFile&v=5(16.08.2017)).

- BENTSEN, M., BETHKE, I., DEBERNARD, J. B., IVERSEN, T., KIRKEVÅG, A., SELAND, Ø., DRANGE, H., ROELANDT, C., SEIERSTAD, I. A., HOOSE, C. & KRISTJÁNSSON, J. E. (2013): *The Norwegian Earth System Model, NorESM1-M-Part 1: Description and basic evaluation of the physical climate*. In: Geosci. Model Dev., **6**, 3: 687–720.
- BOÉ, J. & TERRAY, L. (2015): *Can metric-based approaches really improve multi-model climate projections? The case of summer temperature change in France*. In: Clim. Dyn., **45**, 7-8: 1913–1928.
- BORTZ, J., LIENERT, G. A. & BOEHNKE, K. (2008): *Verteilungsfreie Methoden in der Biostatistik*. Springer-Lehrbuch, 3 Aufl. Springer Berlin Heidelberg, Berlin, Heidelberg.
- CHEN, C., CANE, M. A., WITTENBERG, A. T. & CHEN, D. (2017): *ENSO in the CMIP5 Simulations: Life Cycles, Diversity, and Responses to Climate Change*. In: J. Clim., **30**, 2: 775–801.
- CHEN, L. & FRAUENFELD, O. W. (2014): *A comprehensive evaluation of precipitation simulations over China based on CMIP5 multimodel ensemble projections*. In: J. Geophys. Res. Atmos., **119**, 10: 5767–5786.
- CHRISTENSEN, J., KJELLSTRÖM, E., GIORGI, F., LENDERINK, G. & RUMMUKAINEN, M. (2010): *Weight assignment in regional climate models*. In: Clim. Res., **44**, 2-3: 179–194.
- CHRISTENSEN, J. H., CARTER, T. R., RUMMUKAINEN, M. & AMANATIDIS, G. (2007): *Evaluating the performance and utility of regional climate models: the PRUDENCE project*. In: Clim. Change, **81**, S1: 1–6.
- CHRISTENSEN, O. B., CHRISTENSEN, J. H., MACHENHAUER, B. & BOTZET, M. (1998): *Very High-Resolution Regional Climate Simulations over Scandinavia Present Climate*. In: J. Clim., **11**, 12: 3204–3229.
- CLAUSS, G., FINZE, F.-R. & PARTZSCH, L. (2004): *Statistik Grundlagen: Für Soziologen, Pädagogen, Psychologen und Mediziner*. 5. Aufl. Harri Deutsch, Dresden.
- COLLINS, M. (2007): *Ensembles and probabilities: a new era in the prediction of climate change*. In: Philos. Trans. R. Soc. A Math. Phys. Eng. Sci., **365**, 1857: 1957–1970.

- COLLINS, W. D., BITZ, C. M., BLACKMON, M. L., BONAN, G. B., BREHER-  
TON, C. S., CARTON, J. A., CHANG, P., DONEY, S. C., HACK, J. J., HEN-  
DERSON, T. B., KIEHL, J. T., LARGE, W. G., MCKENNA, D. S., SANTER,  
B. D. & SMITH, R. D. (2006): *The Community Climate System Model Version  
3 (CCSM3)*. In: J. Clim., **19**, 11: 2122–2143.
- COLLINS, W. J., BELLOUIN, N., DOUTRIAUX-BOUCHER, M., GEDNEY, N.,  
HALLORAN, P., HINTON, T., HUGHES, J., JONES, C. D., JOSHI, M., LID-  
DICOAT, S., MARTIN, G., O'CONNOR, F., RAE, J., SENIOR, C., SITCH, S.,  
TOTTERDELL, I., WILTSHIRE, A. & WOODWARD, S. (2011): *Development  
and evaluation of an Earth-System model-HadGEM2*. In: Geosci. Model Dev.,  
**4**, 4: 1051–1075.
- DAI, A. (2006): *Precipitation Characteristics in Eighteen Coupled Climate Mo-  
dels*. In: J. Clim., **19**, 18: 4605–4630.
- DANANDEH MEHR, A. & KAHYA, E. (2016): *Grid-based performance evalua-  
tion of GCM-RCM combinations for rainfall reproduction*. In: Theor. Appl.  
Climatol., **129**, 1-2: 47–57.
- DELSOLE, T., YANG, X. & TIPPETT, M. K. (2013): *Is unequal weighting si-  
gnificantly better than equal weighting for multi-model forecasting?* In: Q. J. R.  
Meteorol. Soc., **139**, 670: 176–183.
- DÉQUÉ, M., DREVETON, C., BRAUN, A. & CARIOLLE, D. (1994): *The ARPE-  
GE/IFS atmosphere model: a contribution to the French community climate  
modelling*. In: Clim. Dyn., **10**, 4-5: 249–266.
- DI LUCA, A., DE ELÍA, R. & LAPRISE, R. (2013): *Potential for small scale  
added value of RCM's downscaled climate change signal*. In: Clim. Dyn., **40**,  
3-4: 601–618.
- DIFFENBAUGH, N. S. & GIORGI, F. (2012): *Climate change hotspots in the  
CMIP5 global climate model ensemble*. In: Clim. Change, **114**, 3-4: 813–822.
- DIFFENBAUGH, N. S., PAL, J. S., GIORGI, F. & GAO, X. (2007): *Heat stress  
intensification in the Mediterranean climate change hotspot*. In: Geophys. Res.  
Lett., **34**, 11: L11706.
- DIMITROV, R. S. (2016): *The Paris Agreement on Climate Change: Behind Clo-  
sed Doors*. In: Glob. Environ. Polit., **16**, 3: 1–11.

- DITTUS, A. J., KAROLY, D. J., LEWIS, S. C., ALEXANDER, L. V. & DONAT, M. G. (2016): *A Multiregion Model Evaluation and Attribution Study of Historical Changes in the Area Affected by Temperature and Precipitation Extremes*. In: *J. Clim.*, **29**, 23: 8285–8299.
- DIX, M., VOHRALIK, P., BI, D., RASHID, H., MARSLAND, S., O'FARRELL, S., UOTILA, P., HIRST, T., KOWALCZYK, E., SULLIVAN, A., YAN, H., FRANKLIN, C., SUN, Z., WATTERSON, I., COLLIER, M., NOONAN, J., ROTSTAYN, L., STEVENS, L., UHE, P. & PURI, K. (2013): *The ACCESS couple model: documentation of core CMIP5 simulations and initial results, 3-99, 2013*. In: *Aust. Met. Ocean. J.*, **63**: 83–99.
- DONAT, M. G., ALEXANDER, L. V., HEROLD, N. & DITTUS, A. J. (2016): *Temperature and precipitation extremes in century-long gridded observations, reanalyses, and atmospheric model simulations*. In: *J. Geophys. Res. Atmos.*, **121**, 19: 11,174–11,189.
- DONE, J., DAVIS, C. A. & WEISMAN, M. (2004): *The next generation of NWP: explicit forecasts of convection using the weather research and forecasting (WRF) model*. In: *Atmos. Sci. Lett.*, **5**, 6: 110–117.
- DONNER, L. J., WYMAN, B. L., HEMLER, R. S., HOROWITZ, L. W., MING, Y., ZHAO, M., GOLAZ, J.-C., GINOUX, P., LIN, S.-J., SCHWARZKOPF, M. D., AUSTIN, J., ALAKA, G., COOKE, W. F., DELWORTH, T. L., FREIDENREICH, S. M., GORDON, C. T., GRIFFIES, S. M., HELD, I. M., HURLIN, W. J., KLEIN, S. A., KNUTSON, T. R., LANGENHORST, A. R., LEE, H.-C., LIN, Y., MAGI, B. I., MALYSHEV, S. L., MILLY, P. C. D., NAIK, V., NATH, M. J., PINCUS, R., PLOSHAY, J. J., RAMASWAMY, V., SEMAN, C. J., SHEVLIAKOVA, E., SIRUTIS, J. J., STERN, W. F., STOUFFER, R. J., WILSON, R. J., WINTON, M., WITTENBERG, A. T. & ZENG, F. (2011): *The Dynamical Core, Physical Parameterizations, and Basic Simulation Characteristics of the Atmospheric Component AM3 of the GFDL Global Coupled Model CM3*. In: *J. Clim.*, **24**, 13: 3484–3519.
- DOOLITTLE, M. H. (1885): *The verification of predictions*. In: *Amer. Meteor. J.*, **2**: 327–329.
- DUFRESNE, J. L., FOUJOLS, M. A., DENVIL, S., CAUBEL, A., MARTI, O., AUMONT, O., BALKANSKI, Y., BEKKI, S., BELLENGER, H., BENSHILA, R., BONY, S., BOPP, L., BRACONNOT, P., BROCKMANN, P., CADULE, P., CHERUY, F., CODRON, F., COZIC, A., CUGNET, D., DE NOBLET, N., DUVEL,

- J. P., ETHÉ, C., FAIRHEAD, L., FICHEFET, T., FLAVONI, S., FRIEDLINGSTEIN, P., GRANDPEIX, J. Y., GUEZ, L., GUILYARDI, E., HAUGLUSTAINE, D., HOURDIN, F., IDELKADI, A., GHATTAS, J., JOUSSAUME, S., KAGEYAMA, M., KRINNER, G., LABETOULLE, S., LAHELLEC, A., LEFEBVRE, M. P., LEFEBVRE, F., LEVY, C., LI, Z. X., LLOYD, J., LOTT, F., MADEC, G., MANCIP, M., MARCHAND, M., MASSON, S., MEURDESOF, Y., MIGNOT, J., MUSAT, I., PAROUTY, S., POLCHER, J., RIO, C., SCHULZ, M., SWINGEDOUW, D., SZOPA, S., TALANDIER, C., TERRAY, P., VIOVY, N. & VUICHARD, N. (2013): *Climate change projections using the IPSL-CM5 Earth System Model: From CMIP3 to CMIP5*. In: *Clim. Dyn.*, **40**, 9-10: 2123–2165.
- DÜNKELOH, A. & JACOBET, J. (2003): *Circulation dynamics of Mediterranean precipitation variability 1948-98*. In: *Int. J. Climatol.*, **23**, 15: 1843–1866.
- DUNNE, J. P., JOHN, J. G., ADCROFT, A. J., GRIFFIES, S. M., HALLBERG, R. W., SHEVLIAKOVA, E., STOUFFER, R. J., COOKE, W., DUNNE, K. A., HARRISON, M. J., KRASTING, J. P., MALYSHEV, S. L., MILLY, P. C. D., PHILLIPPS, P. J., SENTMAN, L. T., SAMUELS, B. L., SPELMAN, M. J., WINTON, M., WITTENBERG, A. T. & ZADEH, N. (2012): *GFDL's ESM2 Global Coupled Climate-Carbon Earth System Models. Part I: Physical Formulation and Baseline Simulation Characteristics*. In: *J. Clim.*, **25**, 19: 6646–6665.
- EMANUEL, K. A. & ŽIVKOVIĆ-ROTHMAN, M. (1999): *Development and Evaluation of a Convection Scheme for Use in Climate Models*. In: *J. Atmos. Sci.*, **56**, 11: 1766–1782.
- EMORI, S., HASEGAWA, A., SUZUKI, T. & DAIRAKU, K. (2005): *Validation, parameterization dependence, and future projection of daily precipitation simulated with a high-resolution atmospheric GCM*. In: *Geophys. Res. Lett.*, **32**, 6: L06708.
- FEKETE, B. M., VÖRÖSMARTY, C. J., ROADS, J. O. & WILLMOTT, C. J. (2004): *Uncertainties in Precipitation and Their Impacts on Runoff Estimates*. In: *J. Clim.*, **17**, 2: 294–304.
- FELIKSON, D., BARTHOLOMAUS, T. C., CATANIA, G. A., KORSGAARD, N. J., KJÆR, K. H., MORLIGHEM, M., NOËL, B., VAN DEN BROEKE, M., STEARNS, L. A., SHROYER, E. L., SUTHERLAND, D. A. & NASH, J. D. (2017): *Inland thinning on the Greenland ice sheet controlled by outlet glacier geometry*. In: *Nat. Geosci.*, **10**, 5: 366–369.

- FESER, F. (2006): *Enhanced Detectability of Added Value in Limited-Area Model Results Separated into Different Spatial Scales*. In: *Mon. Weather Rev.*, **134**, 8: 2180–2190.
- FESER, F., ROCKEL, B., VON STORCH, H., WINTERFELDT, J. & ZAHN, M. (2011): *Regional Climate Models Add Value to Global Model Data: A Review and Selected Examples*. In: *Bull. Am. Meteorol. Soc.*, **92**, 9: 1181–1192.
- FINLEY, J. P. (1884): *Tornado predictions*. In: *Amer. Meteor. J.*, **1**: 85–88.
- FLATO, G., MAROTZKE, J., ABIODUN, B., BRACONNOT, P., CHOU, S. C., COLLINS, W. J., COX, P., DRIQUECH, F., EMORI, S., EYRING, V., FOREST, C., GLECKLER, P., GUILYARDI, E., JAKOB, C., KATTSOV, V., REASON, C. & RUMMUKAINEN, M. (2013): *Evaluation of Climate Models*. In: *Clim. Chang. 2013 Phys. Sci. Basis. Contrib. Work. Gr. I to Fifth Assess. Rep. Intergov. Panel Clim. Chang.*, Kap. 9, 866. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.
- FOGLI, P. G., MANZINI, E., VICHI, M., ALESSANDRI, A., PATARA, L., GUALDI, S., SCOCCIMARRO, E., MASINA, S. & NAVARRA, A. (2009): *INGV-CMCC Carbon (ICC): A Carbon Cycle Earth System Model*. In: *CMCC Research Paper*, **61**: 31.
- FORSTER, P. M., ANDREWS, T., GOOD, P., GREGORY, J. M., JACKSON, L. S. & ZELINKA, M. (2013): *Evaluating adjusted forcing and model spread for historical and future scenarios in the CMIP5 generation of climate models*. In: *J. Geophys. Res. Atmos.*, **118**, 3: 1139–1150.
- FREIDENREICH, S. M., GARNER, S., GUDGEL, R. G., GORDON, C. & HELD, I. (2004): *The New GFDL Global Atmosphere and Land Model AM2-LM2. Evaluation with Prescribed SST Simulations*. In: *J. Clim.*, **17**, 24: 4641–4673.
- GARROW, B. S. (1974): *EISPACK-A package of matrix eigensystem routines*. In: *Comput. Phys. Commun.*, **7**, 4: 179–184.
- GEHNE, M., HAMILL, T. M., KILADIS, G. N. & TRENBERTH, K. E. (2016): *Comparison of Global Precipitation Estimates across a Range of Temporal and Spatial Scales*. In: *J. Clim.*, **29**, 21: 7773–7795.
- GENT, P. R., DANABASOGLU, G., DONNER, L. J., HOLLAND, M. M., HUNKE, E. C., JAYNE, S. R., LAWRENCE, D. M., NEALE, R. B., RASCH, P. J., VERTENSTEIN, M., WORLEY, P. H., YANG, Z.-L. & ZHANG, M. (2011): *The Community Climate System Model Version 4*. In: *J. Clim.*, **24**, 19: 4973–4991.

- GHELLI, A. & PRIMO, C. (2009): *On the use of the extreme dependency score to investigate the performance of an NWP model for rare events*. In: Meteorol. Appl., **16**, 4: 537–544.
- GHIL, M., ALLEN, M. R., DETTINGER, M. D., IDE, K., KONDRASHOV, D., MANN, M. E., ROBERTSON, A. W., SAUNDERS, A., TIAN, Y., VARADI, F. & YIOU, P. (2002): *Advanced Spectral Methods for Climatic Time Series*. In: Rev. Geophys., **40**, 1: 1–41.
- GILBERT, G. F. (1884): *Finley's tornado predictions*. In: Amer. Meteor. J., **1**: 166–172.
- GILL, P. G. & BUCHANAN, P. (2014): *An ensemble based turbulence forecasting system*. In: Meteorol. Appl., **21**, 1: 12–19.
- GILLET, N. P. (2015): *Weighting climate model projections using observational constraints*. In: Philos. Trans. R. Soc. A Math. Phys. Eng. Sci., **373**, 2054: 20140425.
- GIORGI, F. (2005): *Climate Change Prediction*. In: Clim. Change, **73**, 3: 239–265.
- GIORGI, F. (2006): *Climate change hot-spots*. In: Geophys. Res. Lett., **33**, 8: L08707.
- GIORGI, F., JONES, C. & ASRAR, G. (2009): *Addressing climate information needs at the regional level: The CORDEX framework*. In: WMO Bull., **58**, 3: 175–183.
- GIORGI, F. & LIONELLO, P. (2008): *Climate change projections for the Mediterranean region*. In: Glob. Planet. Change, **63**, 2-3: 90–104.
- GIORGI, F. & MEARN, L. O. (2002): *Calculation of Average, Uncertainty Range, and Reliability of Regional Climate Changes from AOGCM Simulations via the Reliability Ensemble Averaging (REA) Method*. In: J. Clim., **15**, 10: 1141–1158.
- GIORGI, F. & MEARN, L. O. (2003): *Probability of regional climate change based on the Reliability Ensemble Averaging (REA) method*. In: Geophys. Res. Lett., **30**, 12.
- GLECKLER, P. J., TAYLOR, K. E. & DOUTRIAUX, C. (2008): *Performance metrics for climate models*. In: J. Geophys. Res., **113**, D6: D06104.

- Gobiet, A., Kotlarski, S., Beniston, M., Heinrich, G., Rajczak, J. & Stoffel, M. (2014): *21st century climate change in the European Alps? A review*. In: *Sci. Total Environ.*, **493**: 1138–1151.
- Gordon, H. B., Rotstayn, L. D., McGregor, J. L., Dix, M. R., Kowalczyk, E. A., O’Farrell, S. P., Waterman, L. J., Hirst, A. C., Wilson, S. G., Collier, M. A., Watterson, I. G. & Elliott, T. I. (2002): *The CSIRO Mk3 Climate System Model*. Techn. Ber., CSIRO Atmospheric Research Tech. Pap. No. 60.
- Grose, M. R., Brown, J. N., Narsey, S., Brown, J. R., Murphy, B. F., Langlais, C., Gupta, A. S., Moise, A. F. & Irving, D. B. (2014): *Assessment of the CMIP5 global climate model simulations of the western tropical Pacific climate system and comparison to CMIP3*. In: *Int. J. Climatol.*, **34**, 12: 3382–3399.
- Haldane, J. B. S. (1956): *The estimation and significance of the logarithm of a ratio of frequencies*. In: *Ann. Hum. Genet.*, **20**, 4: 309–311.
- Harris, I., Jones, P., Osborn, T. & Lister, D. (2014): *Updated high-resolution grids of monthly climatic observations - the CRU TS3.10 Dataset*. In: *Int. J. Climatol.*, **34**, 3: 623–642.
- Hasselmann, K. (1979): *On the signal-to-noise problem in atmospheric response studies*. In: *Meteorol. Trop. Ocean.*, 251–259.
- Hasselmann, K. (1993): *Optimal Fingerprints for the Detection of Time-dependent Climate Change*. In: *J. Clim.*, **6**, 10: 1957–1971.
- Hasselmann, K. (1997): *Multi-pattern fingerprint method for detection and attribution of climate change*. In: *Clim. Dyn.*, **13**, 9: 601–611.
- Haughton, N., Abramowitz, G., Pitman, A. & Phipps, S. J. (2015): *Weighting climate model ensembles for mean and variance estimates*. In: *Clim. Dyn.*, **45**, 11-12: 3169–3181.
- Hawkins, E. & Sutton, R. (2009): *The Potential to Narrow Uncertainty in Regional Climate Predictions*. In: *Bull. Am. Meteorol. Soc.*, **90**, 8: 1095–1107.
- Haylock, M. R., Hofstra, N., Klein Tank, A. M. G., Klok, E. J., Jones, P. D. & New, M. (2008): *A European daily high-resolution gridded data set of surface temperature and precipitation for 1950 to 2006*. In: *J. Geophys. Res.*, **113**, D20: D20119.

- HAZELEGER, W., WANG, X., SEVERIJNS, C., STEFANESCU, S., BINTANJA, R., STERL, A., WYSER, K., SEMMLER, T., YANG, S., VAN DEN HURK, B., VAN NOIJE, T., VAN DER LINDEN, E. & VAN DER WIEL, K. (2012): *EC-Earth V2.2: description and validation of a new seamless earth system prediction model*. In: *Clim. Dyn.*, **39**, 11: 2611–2629.
- HEGERL, G. C. & NORTH, G. R. (1997): *Comparison of Statistically Optimal Approaches to Detecting Anthropogenic Climate Change*. In: *J. Clim.*, **10**, 5: 1125–1133.
- HEGERL, G. C., VON STORCH, H., HASSELMANN, K., SANTER, B. D., CUBASCH, U. & JONES, P. D. (1996): *Detecting Greenhouse-Gas-Induced Climate Change with an Optimal Fingerprint Method*. In: *J. Clim.*, **9**, 10: 2281–2306.
- HEIDKE, P. (1926): *Berechnung des Erfolges und der Güte der Windstärkevorhersagen im Sturmwarnungsdienst*. In: *Geogr. Ann. Stockholm*, **8**: 301.
- HERTIG, E. & JACOBET, J. (2008): *Assessments of Mediterranean precipitation changes for the 21st century using statistical downscaling techniques*. In: *Int. J. Climatol.*, **28**, 8: 1025–1045.
- HERTIG, E. & JACOBET, J. (2014): *Considering observed and future nonstationarities in statistical downscaling of Mediterranean precipitation*. In: *Theor. Appl. Climatol.*, **122**, 3-4: 667–683.
- HERTIG, E., SEUBERT, S., PAXIAN, A., VOGT, G., PAETH, H. & JACOBET, J. (2013): *Changes of total versus extreme precipitation and dry periods until the end of the twenty-first century: Statistical assessments for the Mediterranean area*. In: *Theor. Appl. Climatol.*, **111**, 1-2: 1–20.
- HERTIG, E. & TRAMBLAY, Y. (2016): *Regional downscaling of Mediterranean droughts under past and future climatic conditions*. In: *Glob. Planet. Change*, **151**: 36–48.
- HIDALGO, H. G. & ALFARO, E. J. (2015): *Skill of CMIP5 climate models in reproducing 20th century basic climate features in Central America*. In: *Int. J. Climatol.*, **35**, 12: 3397–3421.
- HOEGH-GULDBERG, O. & BRUNO, J. F. (2010): *The Impact of Climate Change on the World's Marine Ecosystems*. In: *Science*, **328**, 5985: 1523–1528.
- HOEGH-GULDBERG, O., MUMBY, P. J., HOOTEN, A. J., STENECK, R. S., GREENFIELD, P., GOMEZ, E., HARVELL, C. D., SALE, P. F., EDWARDS,

- A. J., CALDEIRA, K., KNOWLTON, N., EAKIN, C. M., IGLESIAS-PRIETO, R., MUTHIGA, N., BRADBURY, R. H., DUBI, A. & HATZIOLOS, M. E. (2007): *Coral Reefs Under Rapid Climate Change and Ocean Acidification*. In: *Science*, **318**, 5857: 1737–1742.
- HOFSTRA, N., HAYLOCK, M., NEW, M. & JONES, P. D. (2009): *Testing E-OBS European high-resolution gridded data set of daily precipitation and surface temperature*. In: *J. Geophys. Res.*, **114**, D21: D21101.
- HOUDIN, F., MUSAT, I., BONY, S., BRACONNOT, P., CODRON, F., DUFRESNE, J.-L., FAIRHEAD, L., FILIBERTI, M.-A., FRIEDLINGSTEIN, P., GRANDPEIX, J.-Y., KRINNER, G., LEVAN, P., LI, Z.-X. & LOTT, F. (2006): *The LMDZ4 general circulation model: climate performance and sensitivity to parametrized physics with emphasis on tropical convection*. In: *Clim. Dyn.*, **27**, 7-8: 787–813.
- HUANG, D.-Q., ZHU, J., ZHANG, Y.-C. & HUANG, A.-N. (2013): *Uncertainties on the simulated summer precipitation over Eastern China from the CMIP5 models*. In: *J. Geophys. Res. Atmos.*, **118**, 16: 9035–9047.
- HURRELL, J. W., HOLLAND, M. M., GENT, P. R., GHAN, S., KAY, J. E., KUSHNER, P. J., LAMARQUE, J.-F., LARGE, W. G., LAWRENCE, D., LINDSAY, K., LIPSCOMB, W. H., LONG, M. C., MAHOWALD, N., MARSH, D. R., NEALE, R. B., RASCH, P., VAVRUS, S., VERTENSTEIN, M., BADER, D., COLLINS, W. D., HACK, J. J., KIEHL, J. & MARSHALL, S. (2013): *The Community Earth System Model: A Framework for Collaborative Research*. In: *Bull. Am. Meteorol. Soc.*, **94**, 9: 1339–1360.
- IGLESIAS, A., GARROTE, L., FLORES, F. & MONEO, M. (2007): *Challenges to manage the risk of water scarcity and climate change in the Mediterranean*. In: *Water Resour. Manag.*, **21**, 5: 775–788.
- IPCC (1990a): *Climate change: The IPCC Impacts Assessment. Report prepared for the Intergovernmental Panel on Climate Change by Working Group II*. Australian Government Publishing Service, Canberra, Australia.
- IPCC (1990b): *Climate Change: The IPCC Response Strategies. Report prepared for the Intergovernmental Panel on Climate Change by Working Group III*. Cambridge University Press, Cambridge and New York.
- IPCC (1990c): *Climate Change: The IPCC Scientific Assessment. Report prepared for the Intergovernmental Panel on Climate Change by Working Group I*. Cambridge Univ. Press, Cambridge [u.a.].

- IPCC (1996a): *Climate Change 1995. Economic and Social Dimensions of Climate Change: Contribution of Working Group III to the Second Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge and New York.
- IPCC (1996b): *Climate Change 1995. Impacts, Adaptations and Mitigation of Climate Change: Scientific-Technical Analyses. Contribution of Working Group II to the Second Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, New York.
- IPCC (1996c): *Climate Change 1995. The Scientific Basis: Contribution of the Working Group I to the Second Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge and New York.
- IPCC (2001a): *Climate Change 2001: Impacts, Adaptation, and Vulnerability. A Report of Working Group II of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge and New York.
- IPCC (2001b): *Climate Change 2001: Mitigation. A Report of Working Group III of the Intergovernmental Panel on Climate Change*. Cambridge Univ. Press, Cambridge, New York.
- IPCC (2001c): *Climate Change 2001. The Scientific Basis: Contribution of the Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge and New York.
- IPCC (2007a): *Climate Change 2007: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom.
- IPCC (2007b): *Climate Change 2007: Mitigation of Climate Change. Contribution of Working Group III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge and New York.
- IPCC (2007c): *Climate Change 2007. The Physical Basis: Contribution of the Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge and New York.

- IPCC (2013): *Climate Change 2013. The Physical Basis: Contribution of the Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge and New York.
- IPCC (2014a): *Climate Change 2014 Impacts, Adaptation, and Vulnerability Part A: Global and Sectoral Aspects Working Group II Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom.
- IPCC (2014b): *Climate Change 2014 Mitigation of Climate Change Working Group III Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom.
- JACOB, D., PETERSEN, J., EGGERT, B., ALIAS, A., CHRISTENSEN, O. B., BOUWER, L. M., BRAUN, A., COLETTE, A., DÉQUÉ, M., GEORGIEVSKI, G., GEORGOPOULOU, E., GOBIET, A., MENUT, L., NIKULIN, G., HAENS-  
LER, A., HEMPELMANN, N., JONES, C., KEULER, K., KOVATS, S., KRÖ-  
NER, N., KOTLARSKI, S., KRIEGSMANN, A., MARTIN, E., VAN MEIJGAARD,  
E., MOSELEY, C., PFEIFER, S., PREUSCHMANN, S., RADERMACHER, C.,  
RADTKE, K., RECHID, D., ROUNSEVELL, M., SAMUELSSON, P., SOMOT,  
S., SOUSSANA, J.-F., TEICHMANN, C., VALENTINI, R., VAUTARD, R., WE-  
BER, B. & YIOU, P. (2014): *EURO-CORDEX: new high-resolution climate  
change projections for European impact research*. In: Reg. Environ. Chang.,  
**14**, 2: 563–578.
- JACOBET, J. (1993): *Regionale Unterschiede im atmosphärischen Zirkulations-  
geschehen bei globalen Klimaveränderungen (Regional differences of the atmo-  
spheric circulation under conditions of global climate change)*. In: Die Erde,  
**124**: 63–77.
- JACOBET, J., JÖNSSON, P., BÄRRING, L., BECK, C. & EKSTRÖM, M. (2001):  
*Zonal indices for Europe 1780 - 1995 and running correlations with temperature*.  
In: Clim. Chang., **48**: 219–241.
- KATERI, M. (2014): *Contingency Table Analysis*. Springer New York, New York,  
NY.
- KATES, R. W., TRAVIS, W. R. & WILBANKS, T. J. (2012): *Transformational  
adaptation when incremental adaptations to climate change are insufficient*. In:  
Proc. Natl. Acad. Sci. U. S. A., **109**, 19: 7156–7161.

- KATRAGKOU, E., GARCÍA-DÍEZ, M., VAUTARD, R., SOBOLOWSKI, S., ZANIS, P., ALEXANDRI, G., CARDOSO, R. M., COLETTE, A., FERNANDEZ, J., GOBIET, A., GOERGEN, K., KARACOSTAS, T., KNIST, S., MAYER, S., SOARES, P. M. M., PYTHAROULIS, I., TEGOULIAS, I., TSIKERDEKIS, A. & JACOB, D. (2015): *Regional climate hindcast simulations within EURO-CORDEX: evaluation of a WRF multi-physics ensemble*. In: Geosci. Model Dev., **8**, 3: 603–618.
- KAY, S. & MARPLE, S. (1981): *Spectrum analysis. A modern perspective*. In: Proc. IEEE, **69**, 11: 1380–1419.
- KEEGAN, K. M., ALBERT, M. R., MCCONNELL, J. R. & BAKER, I. (2014): *Climate change and forest fires synergistically drive widespread melt events of the Greenland Ice Sheet*. In: Proc. Natl. Acad. Sci., **111**, 22: 7964–7967.
- KELLEY, C., TING, M., SEAGER, R. & KUSHNIR, Y. (2012a): *Mediterranean precipitation climatology, seasonal cycle, and trend as simulated by CMIP5*. In: Geophys. Res. Lett., **39**, 21: L21703.
- KELLEY, C., TING, M., SEAGER, R. & KUSHNIR, Y. (2012b): *The relative contributions of radiative forcing and internal climate variability to the late 20th Century winter drying of the Mediterranean region*. In: Clim. Dyn., **38**, 9-10: 2001–2015.
- KHAN, S. A., KJAER, K. H., BEVIS, M., BAMBER, J. L., WAHR, J., KJELDSEN, K. K., BJORK, A. A., KORSGAARD, N. J., STEARNS, L. A., VAN DEN BROEKE, M. R., LIU, L., LARSEN, N. K. & MURESAN, I. S. (2014): *Sustained mass loss of the northeast Greenland ice sheet triggered by regional warming*. In: Nat. Clim. Chang., **4**, 4: 292–299.
- KHARIN, V. V., ZWIERS, F. W., ZHANG, X. & HEGERL, G. C. (2007): *Changes in Temperature and Precipitation Extremes in the IPCC Ensemble of Global Coupled Model Simulations*. In: J. Clim., **20**, 8: 1419–1444.
- KHARIN, V. V., ZWIERS, F. W., ZHANG, X. & WEHNER, M. (2013): *Changes in temperature and precipitation extremes in the CMIP5 ensemble*. In: Clim. Change, **119**, 2: 345–357.
- KIEHL, J. T., HACK, J. J., BONAN, G. B., BOVILLE, B. A., WILLIAMSON, D. L. & RASCH, P. J. (1998): *The National Center for Atmospheric Research Community Climate Model: CCM3\**. In: J. Clim., **11**, 6: 1131–1149.
- KIM, J., SANJAY, J., MATTMANN, C., BOUSTANI, M., RAMARAO, M. V. S., KRISHNAN, R. & WALISER, D. (2015): *Uncertainties in estimating spatial*

- and interannual variations in precipitation climatology in the India-Tibet region from multiple gridded precipitation datasets.* In: *Int. J. Climatol.*, **35**, 15: 4557–4573.
- KLOCKE, D., PINCUS, R. & QUAAS, J. (2011): *On Constraining Estimates of Climate Sensitivity with Present-Day Observations through Model Weighting.* In: *J. Clim.*, **24**, 23: 6092–6099.
- KNUTTI, R. (2010): *The end of model democracy?* In: *Clim. Change*, **102**, 3-4: 395–404.
- KNUTTI, R., FURRER, R., TEBALDI, C., CERMAK, J. & MEEHL, G. A. (2010): *Challenges in Combining Projections from Multiple Climate Models.* In: *J. Clim.*, **23**, 10: 2739–2758.
- KNUTTI, R., ROGELJ, J., SEDLÁČEK, J. & FISCHER, E. M. (2015): *A scientific critique of the two-degree climate change target.* In: *Nat. Geosci.*, **9**, 1: 13–18.
- KNUTTI, R. & SEDLÁČEK, J. (2012): *Robustness and uncertainties in the new CMIP5 climate model projections.* In: *Nat. Clim. Chang.*, **3**, 4: 369–373.
- KOTLARSKI, S., KEULER, K., CHRISTENSEN, O. B., COLETTE, A., DÉQUÉ, M., GOBIET, A., GOERGEN, K., JACOB, D., LÜTHI, D., VAN MEIJGAARD, E., NIKULIN, G., SCHÄR, C., TEICHMANN, C., VAUTARD, R., WARRACH-SAGI, K. & WULFMEYER, V. (2014): *Regional climate modeling on European scales: A joint standard evaluation of the EURO-CORDEX RCM ensemble.* In: *Geosci. Model Dev.*, **7**, 4: 1297–1333.
- KOUTROULIS, A. G., GRILLAKIS, M. G., TSANIS, I. K. & PAPADIMITRIOU, L. (2016): *Evaluation of precipitation and temperature simulation performance of the CMIP3 and CMIP5 historical experiments.* In: *Clim. Dyn.*, **47**, 5-6: 1881–1898.
- KUMAR, D., KODRA, E. & GANGULY, A. R. (2014): *Regional and seasonal intercomparison of CMIP3 and CMIP5 climate model ensembles for temperature and precipitation.* In: *Clim. Dyn.*, **43**, 9-10: 2491–2518.
- KUMAR, S., MERWADE, V., KINTER, J. L. & NIYOGI, D. (2013): *Evaluation of Temperature and Precipitation Trends and Long-Term Persistence in CMIP5 Twentieth-Century Climate Simulations.* In: *J. Clim.*, **26**, 12: 4168–4185.
- LAHR, H. (2014): *An Improved Test for Earnings Management Using Kernel Density Estimation.* In: *Eur. Account. Rev.*, **23**, 4: 559–591.

- LEDUC, M., LAPRISE, R., DE ELÍA, R. & ŠEPAROVIĆ, L. (2016): *Is Institutional Democracy a Good Proxy for Model Independence?* In: *J. Clim.*, **29**, 23: 8301–8316.
- LI, L., LIN, P., YU, Y., WANG, B., ZHOU, T., LIU, L., LIU, J., BAO, Q., XU, S., HUANG, W., XIA, K., PU, Y., DONG, L., SHEN, S., LIU, Y., HU, N., LIU, M., SUN, W., SHI, X., ZHENG, W., WU, B., SONG, M., LIU, H., ZHANG, X., WU, G., XUE, W., HUANG, X., YANG, G., SONG, Z. & QIAO, F. (2013): *The flexible global ocean-atmosphere-land system model, Grid-point Version 2: FGOALS-g2*. In: *Adv. Atmos. Sci.*, **30**, 3: 543–560.
- LI, X., HU, Z.-Z., JIANG, X., LI, Y., GAO, Z., YANG, S., ZHU, J. & JHA, B. (2016): *Trend and seasonality of land precipitation in observations and CMIP5 model simulations*. In: *Int. J. Climatol.*, **36**, 11: 3781–3793.
- LIU, Y., ZHUANG, Q., MIRALLES, D., PAN, Z., KICKLIGHTER, D., ZHU, Q., HE, Y., CHEN, J., TCHEBAKOVA, N., SIRIN, A., NIYOGI, D. & MELILLO, J. (2015): *Evapotranspiration in Northern Eurasia: Impact of forcing uncertainties on terrestrial ecosystem model estimates*. In: *J. Geophys. Res. Atmos.*, **120**, 7: 2647–2660.
- LIU, Z., MEHRAN, A., PHILLIPS, T. & AGHAKOUCHAK, A. (2014): *Seasonal and regional biases in CMIP5 precipitation simulations*. In: *Clim. Res.*, **60**, 1: 35–50.
- LOARIE, S. R., DUFFY, P. B., HAMILTON, H., ASNER, G. P., FIELD, C. B. & ACKERLY, D. D. (2009): *The velocity of climate change*. In: *Nature*, **462**, 7276: 1052–1055.
- MARKOVIC, M., DE ELÍA, R., FRIGON, A. & MATTHEWS, H. D. (2013): *A transition from CMIP3 to CMIP5 for climate information providers: the case of surface temperature over eastern North America*. In: *Clim. Change*, **120**, 1-2: 197–210.
- MARTIN, G. M., DEARDEN, C. & GREEVES C. (2004): *Hadley Centre Technical Note*. Techn. Ber., 54.
- McFARLANE, N. A., BOER, G. J., BLANCHET, J.-P. & LAZARE, M. (1992): *The Canadian Climate Centre Second-Generation General Circulation Model and Its Equilibrium Climate*. In: *J. Clim.*, **5**, 10: 1013–1044.
- MEEHL, G. A., COVEY, C., TAYLOR, K. E., DELWORTH, T., STOUFFER, R. J., LATIF, M., McAVANEY, B. & MITCHELL, J. F. B. (2007): *THE*

- WCRP CMIP3 Multimodel Dataset: A New Era in Climate Change Research.*  
In: Bull. Am. Meteorol. Soc., **88**, 9: 1383–1394.
- MEEHL, G. A., MOSS, R., TAYLOR, K. E., EYRING, V., STOUFFER, R. J., BONY, S. & STEVENS, B. (2014): *Climate model intercomparisons: Preparing for the next phase.* In: Eos Trans. AGU, **95**, 9: 77–78.
- MIAO, C., DUAN, Q., YANG, L. & BORTHWICK, A. G. L. (2012): *On the Applicability of Temperature and Precipitation Data from CMIP3 for China.* In: PLoS One, **7**, 9: e44659.
- MILLER, L. H. (1956): *Table of Percentage Points of Kolmogorov Statistics.* In: J. Am. Stat. Assoc., **51**, 273: 111–121.
- MITCHELL, T. D. & JONES, P. D. (2005): *An improved method of constructing a database of monthly climate observations and associated high-resolution grids.* In: Int. J. Climatol., **25**, 6: 693–712.
- MOSS, R. H., EDMONDS, J. A., HIBBARD, K. A., MANNING, M. R., ROSE, S. K., VAN VUUREN, D. P., CARTER, T. R., EMORI, S., KAINUMA, M., KRAM, T., MEEHL, G. A., MITCHELL, J. F. B., NAKICENOVIC, N., RIAHI, K., SMITH, S. J., STOUFFER, R. J., THOMSON, A. M., WEYANT, J. P. & WILBANKS, T. J. (2010): *The next generation of scenarios for climate change research and assessment.* In: Nature, **463**, 7282: 747–756.
- NAKICENOVIC, N., ALCAMO, J., DAVIS, G. ., DE VRIES, B., FENHANN, J., GAFFIN, S., GREGORY, K., GRUBLER, A., JUNG, T. Y., KRAM, T., LA ROVERE, E. L., MICHAELIS, L., MORI, S., MORITA, T. & ZHOU, D. (2000): *Special report on emissions scenarios : a special report of Working Group III of the Intergovernmental Panel on Climate Change.* Cambridge University Press, U.S.A.
- NEW, M., HULME, M. & JONES, P. (1999): *Representing Twentieth-Century Space Time Climate Variability . Part I : Development of a 1961 to 1990 Mean Monthly Terrestrial Climatology.* In: J. Clim., **12**, 3: 829–856.
- NEW, M., HULME, M. & JONES, P. (2000): *Representing twentieth-century space time climate variability. Part II: Development of 1901 to 1996 monthly grids of terrestrial surface climate.* In: J. Clim., **13**, 13: 2217–2238.
- NIKIEMA, P. M., SYLLA, M. B., OGUNJOBI, K., KEBE, I., GIBBA, P. & GIORGI, F. (2017): *Multi-model CMIP5 and CORDEX simulations of historical summer temperature and precipitation variabilities over West Africa.* In: Int. J. Climatol., **37**, 5: 2438–2450.

- NIKULIN, G., JONES, C., GIORGI, F., ASRAR, G., BÜCHNER, M., CERESO-MOTA, R., CHRISTENSEN, O. B., DÉQUÉ, M., FERNANDEZ, J., HÄNSLER, A., VAN MEIJGAARD, E., SAMUELSSON, P., SYLLA, M. B. & SUSHAMA, L. (2012): *Precipitation Climatology in an Ensemble of CORDEX-Africa Regional Climate Simulations*. In: *J. Clim.*, **25**, 18: 6057–6078.
- O'NEILL, B. C., TEBALDI, C., VAN VUUREN, D., EYRING, V., FRIEDLINGS-TEIN, P., HURTT, G., KNUTTI, R., KRIEGLER, E., LAMARQUE, J.-F., LOWE, J., MEEHL, J., MOSS, R., RIAHI, K. & SANDERSON, B. M. (2016): *The Scenario Model Intercomparison Project (ScenarioMIP) for CMIP6*. In: *Geosci. Model Dev.*, **9**, 9: 3461–3482.
- PAETH, H., GIRMES, R., MENZ, G. & HENSE, A. (2006): *Improving Seasonal Forecasting in the Low Latitudes*. In: *Mon. Weather Rev.*, **134**, 7: 1859–1879.
- PAETH, H. & HENSE, A. (2001): *Signal analysis of the atmospheric mean 500/1000 hPa temperature north of 55N between 1949 and 1994*. In: *Clim. Dyn.*, **18**, 3-4: 345–358.
- PAETH, H. & HENSE, A. (2005): *Mean versus extreme climate in the Mediterranean region and its sensitivity to future global warming conditions*. In: *Meteorol. Zeitschrift*, **14**, 3: 329–347.
- PAETH, H. & MANNIG, B. (2013): *On the added value of regional climate modeling in climate change assessment*. In: *Clim. Dyn.*, **41**, 3-4: 1057–1066.
- PAETH, H., VOGT, G., PAXIAN, A., HERTIG, E., SEUBERT, S. & JACOBET, J. (2016): *Quantifying the evidence of climate change in the light of uncertainty exemplified by the Mediterranean hot spot region*. In: *Glob. Planet. Change*, **151**: 144–151.
- PALMER, T., SHUTTS, G., HAGEDORN, R., DOBLAS-REYES, F., JUNG, T. & LEUTBECHER, M. (2005): *Representing model uncertainty in weather and climate prediction*. In: *Annu. Rev. Earth Planet. Sci.*, **33**, 1: 163–193.
- PARZEN, M., LIPSITZ, S., IBRAHIM, J. & KLAR, N. (2002): *An Estimate of the Odds Ratio That Always Exists*. In: *J. Comput. Graph. Stat.*, **11**, 2: 420–436.
- PAXIAN, A., HERTIG, E., SEUBERT, S., VOGT, G., JACOBET, J. & PAETH, H. (2014): *Present-day and future mediterranean precipitation extremes assessed by different statistical approaches*. In: *Clim. Dyn.*, **44**, 3-4: 845–860.

- PAXIAN, A., HERTIG, E., VOGT, G., SEUBERT, S., JACOBET, J. & PAETH, H. (2013): *Greenhouse gas-related predictability of regional climate model trends in the Mediterranean area*. In: *Int. J. Climatol.*, **34**, 7: 2293–2307.
- PAXIAN, A., SEIN, D., PANITZ, H.-J., WARSCHER, M., BREIL, M., ENGEL, T., TÖDTER, J., KRAUSE, A., CABOS NARVAEZ, W. D., FINK, A. H., AHRENS, B., KUNSTMANN, H., JACOB, D. & PAETH, H. (2016): *Bias reduction in decadal predictions of West African monsoon rainfall using regional climate models*. In: *J. Geophys. Res. Atmos.*, **121**, 4: 1715–1735.
- PERKINS, S. E., PITMAN, A. J., HOLBROOK, N. J. & MCANENEY, J. (2007): *Evaluation of the AR4 Climate Models' Simulated Daily Maximum Temperature, Minimum Temperature, and Precipitation over Australia Using Probability Density Functions*. In: *J. Clim.*, **20**, 17: 4356–4376.
- PERNETTA, J. C. (1992): *Impacts of climate change and sea-level rise on small island states. National and international responses*. In: *Glob. Environ. Chang.*, **2**, 1: 19–31.
- PHILIPP, A., DELLA-MARTA, P. M., JACOBET, J., FEREDAY, D. R., JONES, P. D., MOBERG, A. & WANNER, H. (2007): *Long-Term Variability of Daily North Atlantic European Pressure Patterns since 1850 Classified by Simulated Annealing Clustering*. In: *J. Clim.*, **20**, 16: 4065–4095.
- PIERCE, C. S. (1884): *The Numerical Measure of the Success of Predictions*. In: *Science*, **4**, 93: 453–454.
- POLI, P., HERBACH, H., DEE, D. P., BERRISFORD, P., SIMMONS, A. J., VITART, F., LALOYAUX, P., TAN, D. G. H., PEUBEY, C., THÉPAUT, J.-N., TRÉMOLET, Y., HÓLM, E. V., BONAVITA, M., ISAKSEN, L. & FISHER, M. (2016): *ERA-20C: An Atmospheric Reanalysis of the Twentieth Century*. In: *J. Clim.*, **29**, 11: 4083–4097.
- POLI, P., HERBACH, H., TAN, D., DEE, D., THÉPAUT, J.-N., SIMMONS, A., PEUBEY, C., LALOYAUX, P., KOMORI, T., BERRISFORD, P. & DRAGANI, R. (2013): *The data assimilation system and initial performance evaluation of the ECMWF pilot reanalysis of the 20th-century assimilating surface observations only (ERA-20C)*. In: *ERA Rep. Ser.*, **14**: 59.
- POPE, V. D., GALLANI, M. L., ROWNTREE, P. R. & STRATTON, R. A. (2000): *The impact of new physical parametrizations in the Hadley Centre climate model: HadAM3*. In: *Clim. Dyn.*, **16**, 2-3: 123–146.

- PRAKASH, S., GAIROLA, R. M. & MITRA, A. K. (2015): *Comparison of large-scale global land precipitation from multisatellite and reanalysis products with gauge-based GPCC data sets*. In: *Theor. Appl. Climatol.*, **121**, 1-2: 303–317.
- PREIN, A. F., GOBIET, A., TRUHETZ, H., KEULER, K., GOERGEN, K., TEICHMANN, C., FOX MAULE, C., VAN MEIJGAARD, E., DÉQUÉ, M., NIKULIN, G., VAUTARD, R., COLETTE, A., KJELLSTRÖM, E. & JACOB, D. (2016): *Precipitation in the EURO-CORDEX 0.11 and 0.44 degree simulations: high resolution, high benefits?* In: *Clim. Dyn.*, **46**, 1-2: 383–412.
- PRÖMMEL, K., GEYER, B., JONES, J. M. & WIDMANN, M. (2010): *Evaluation of the skill and added value of a reanalysis-driven regional simulation for Alpine temperature*. In: *Int. J. Climatol.*, **30**, 5: 760–773.
- PRUSCHA, H. (2006): *Statistisches Methodenbuch*. Springer Berlin Heidelberg, Berlin, Heidelberg.
- QIAO, F., SONG, Z., BAO, Y., SONG, Y., SHU, Q., HUANG, C. & ZHAO, W. (2013): *Development and evaluation of an Earth System Model with surface gravity waves*. In: *J. Geophys. Res. Ocean.*, **118**, 9: 4514–4524.
- RÄISÄNEN, J. (2007): *How reliable are climate models?* In: *Tellus A*, **59**, 1: 2–29.
- RÄISÄNEN, J., RUOKOLAINEN, L. & YLHÄISI, J. (2010): *Weighting of model results for improving best estimates of climate change*. In: *Clim. Dyn.*, **35**, 2-3: 407–422.
- RÄISÄNEN, J. & YLHÄISI, J. S. (2012): *Can model weighting improve probabilistic projections of climate change?* In: *Clim. Dyn.*, **39**, 7-8: 1981–1998.
- RAMESH, K. V. & GOSWAMI, P. (2015): *Assessing reliability of regional climate projections: the case of Indian monsoon*. In: *Sci. Rep.*, **4**, 1: 4071.
- RANDALL, D., BONY, R., S., W., COLMAN, R., FICHEFET, T., FYFE, J., KATTSOV, V., PITMAN, A., SHUKLA, J., SRINIVASAN, J., STOUFFER, R., SUMI, A. & TAYLOR, K. (2007): *Climate Models and Their Evaluation*. In: SOLOMON, S., QIN, D., MANNING, M., CHEN, Z., MARQUIS, M., AVERYT, K., M. TIGNOR & MILLER, H. [Hrsg.]: *Clim. Chang. 2007 Phys. Sci. Basis. Contrib. Work. Gr. I to Fourth Assess. Rep. Intergov. Panel Clim. Chang.*, Kap. 8, 662. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.
- REICHLER, T. & KIM, J. (2008): *How Well Do Coupled Models Simulate Today's Climate?* In: *Bull. Am. Meteorol. Soc.*, **89**, 3: 303–311.

- RIBES, A., AZAÏS, J.-M. & PLANTON, S. (2010): *A method for regional climate change detection using smooth temporal patterns*. In: *Clim. Dyn.*, **35**, 2-3: 391–406.
- RIBES, A., PLANTON, S. & TERRAY, L. (2013): *Application of regularised optimal fingerprinting to attribution. Part I: method, properties and idealised analysis*. In: *Clim. Dyn.*, **41**, 11-12: 2817–2836.
- RIBES, A. & TERRAY, L. (2013): *Application of regularised optimal fingerprinting to attribution. Part II: application to global near-surface temperature*. In: *Clim. Dyn.*, **41**, 11-12: 2837–2853.
- RING, C., MANNIG, B., POLLINGER, F. & PAETH, H. (2016): *Uncertainties in the simulation of precipitation in selected regions of humid and dry climate*. In: *Int. J. Climatol.*, **36**, 10: 3521–3538.
- RING, C., POLLINGER, F., KASPAR-OTT, I., HERTIG, E., JACOBET, J. & PAETH, H. (2017): *A comparison of metrics for assessing state-of-the-art climate models and implications for probabilistic projections of climate change*. In: *Clim. Dyn.*, 1–20.
- ROECKNER, E., ARPE, K., BENGTSSON, L., CHRISTOPH, M., CLAUSSEN, M., DÜMENIL, L., ESCH, M., GIORGETTA, M., SCHLESE, U. & SCHULZWEIDA, U. (1996): *The atmospheric general circulation model ECHAM-4: Model description and simulation of present-day climate*. Techn. Ber., 218, Max-Planck-Institut für Meteorologie Rep.
- ROECKNER, E., BÄUML, G., BONAVENTURA, L., BROKOPF, R., ESCH, M., GIORGETTA, M., HAGEMANN, S., KIRCHNER, I., KORNBLUEH, L., MANZINI, E., RHODUN, A., SCHLESE, U., SCHULZWEIDA, U. & TOMPHINS, A. (2003): *The Atmospheric General Circulation Model ECHAM5. Part I: Model Description*. Techn. Ber., 349, MPI Report.
- RUMMUKAINEN, M. (2010): *State-of-the-art with regional climate models*. In: *Wiley Interdiscip. Rev. Clim. Chang.*, **1**, 1: 82–96.
- RUSSELL, G. L., MILLER, J. R. & RIND, D. (1995): *A coupled atmosphere ocean model for transient climate change studies*. In: *Atmosphere-Ocean*, **33**, 4: 683–730.
- SABA, V. S., GRIFFIES, S. M., ANDERSON, W. G., WINTON, M., ALEXANDER, M. A., DELWORTH, T. L., HARE, J. A., HARRISON, M. J., ROSATI,

- A., VECCHI, G. A. & ZHANG, R. (2016): *Enhanced warming of the Northwest Atlantic Ocean under climate change*. In: J. Geophys. Res. Ocean., **121**, 1: 118–132.
- SACHS, L. (1997): *Angewandte Statistik : Anwendung statistischer Methoden*. 8 Aufl. Springer, Berlin [u.a.].
- SAMUELSSON, P., JONES, C. G., WILLÉN, U., ULLERSTIG, A., GOLLVIK, S., HANSSON, U., JANSSON, C., KJELLSTRÖM, E., NIKULIN, G. & WYSER, K. (2011): *The Rossby Centre Regional Climate model RCA3: Model description and performance*. In: Tellus, Ser. A Dyn. Meteorol. Oceanogr., **63**, 1: 4–23.
- SANDERSON, B. M., KNUTTI, R., CALDWELL, P., SANDERSON, B. M., KNUTTI, R. & CALDWELL, P. (2015a): *A representative democracy to reduce interdependency in a multimodel ensemble*. In: J. Clim., **28**, 13: 5171–5194.
- SANDERSON, B. M., KNUTTI, R., CALDWELL, P., SANDERSON, B. M., KNUTTI, R. & CALDWELL, P. (2015b): *Addressing interdependency in a multimodel ensemble by interpolation of model properties*. In: J. Clim., **28**, 13: 5150–5170.
- SCHIEMANN, R., DEMORY, M.-E., MIZIELINSKI, M. S., ROBERTS, M. J., SHAFFREY, L. C., STRACHAN, J. & VIDALE, P. L. (2014): *The sensitivity of the tropical circulation and Maritime Continent precipitation to climate model resolution*. In: Clim. Dyn., **42**, 9-10: 2455–2468.
- SCHIEMANN, R., LUETHI, D., VIDALE, P. L. & SCHAER, C. (2008): *The precipitation climate of Central Asia. Intercomparison of observational and numerical data sources in a remote semiarid region*. In: Int. J. Climatol., **28**, 3: 295–314.
- SCHMIDT, G. A., KELLEY, M., NAZARENKO, L., RUEDY, R., RUSSELL, G. L., ALEINOV, I., BAUER, M., BAUER, S. E., BHAT, M. K., BLECK, R., CANUTO, V., CHEN, Y.-H., CHENG, Y., CLUNE, T. L., DEL GENIO, A., DE FAINCHEIN, R., FALUVEGI, G., HANSEN, J. E., HEALY, R. J., KIANG, N. Y., KOCH, D., LACIS, A. A., LEGRANDE, A. N., LERNER, J., LO, K. K., MATTHEWS, E. E., MENON, S., MILLER, R. L., OINAS, V., OLOSO, A. O., PERLWITZ, J. P., PUMA, M. J., PUTMAN, W. M., RIND, D., ROMANOU, A., SATO, M., SHINDELL, D. T., SUN, S., SYED, R. A., TAUSNEV, N., TSGARIDIS, K., UNGER, N., VOULGARAKIS, A., YAO, M.-S. & ZHANG, J. (2014): *Configuration and assessment of the GISS ModelE2 contributions to the CMIP5 archive*. In: J. Adv. Model. Earth Syst., **6**, 1: 141–184.
- SCHMIDT, G. A., RUEDY, R., HANSEN, J. E., ALEINOV, I., BELL, N., BAUER, M., BAUER, S., CAIRNS, B., CANUTO, V., CHENG, Y., DEL GENIO, A.,

- FALUVEGI, G., FRIEND, A. D., HALL, T. M., HU, Y., KELLEY, M., KIANG, N. Y., KOCH, D., LACIS, A. A., LERNER, J., LO, K. K., MILLER, R. L., NAZARENKO, L., OINAS, V., PERLWITZ, J. J., PERLWITZ, J. J., RIND, D., ROMANOU, A., RUSSELL, G. L., SATO, M., SHINDELL, D. T., STONE, P. H., SUN, S., TAUSNEV, N., THRESHER, D. & YAO, M.-S. (2006): *Present-Day Atmospheric Simulations Using GISS ModelE: Comparison to In Situ, Satellite, and Reanalysis Data*. In: *J. Clim.*, **19**, 2: 153–192.
- SCHNEIDER, U., FUCHS, T., MEYER-CHRISTOFFER, A. & RUDOLF, B. (2008): *Global Precipitation Analysis Products of the GPCC*. Techn. Ber., Deutscher Wetterdienst, Offenbach. URL [http://www.mapcruzin.com/environmental-shapefile-maps/water/precipitation/GPCC\\_intro\\_products\\_2008.pdf](http://www.mapcruzin.com/environmental-shapefile-maps/water/precipitation/GPCC_intro_products_2008.pdf)(16.08.2017).
- SCHÖNWIESE, C.-D. (2013): *Praktische Statistik für Meteorologen und Geowissenschaftler*. 5. Aufl. Borntraeger, Stuttgart.
- SCHULZWEIDA, U., KORNBLUEH, L. & QUAST, R. (2009): *CDO User's Guide. Climate data operators. Version 1.4.1*. URL <https://www.rsmas.miami.edu/users/rajib/cdo.pdf>(16.08.2017).
- SCINocca, J. F., MCFARLANE, N. A., LAZARE, M., LI, J. & PLUMMER, D. (2008): *Technical Note: The CCCma third generation AGCM and its extension into the middle atmosphere*. In: *Atmos. Chem. Phys.*, **8**, 23: 7055–7074.
- SCOCCIMARRO, E., GUALDI, S., FOGLI, P. G., MANZINI, E., GREZIO, A. & NAVARRA, A. (2007): *INGV-SXG: A Coupled Atmosphere Ocean Sea-Ice General Circulation Climate Model*. In: CMCC Research Paper, **15**.
- SHASHIKANTH, K., SALVI, K., GHOSH, S. & RAJENDRAN, K. (2014): *Do CMIP5 simulations of Indian summer monsoon rainfall differ from those of CMIP3?* In: *Atmos. Sci. Lett.*, **15**, 2: 79–85.
- SHIBATA, K., YOSHIMURA, H., OHIZUMI, M., HOSAKA, M. & SUGI, M. (1999): *A Simulation of Troposphere, Stratosphere and Mesosphere With an MRI/JMA98 GCM*. In: *Pap. Meteorol. Geophys.*, **50**, 1: 15–53.
- SILLMANN, J., KHARIN, V. V., ZHANG, X., ZWIERS, F. W. & BRONAUGH, D. (2013): *Climate extremes indices in the CMIP5 multimodel ensemble: Part 1. Model evaluation in the present climate*. In: *J. Geophys. Res. Atmos.*, **118**, 4: 1716–1733.

- SILVERMAN, B. W. (1986): *Density Estimation for Statistics and Data Analysis*. 1 Aufl. CRC Press, London.
- SMITH, B. T., BOYLE, J. M., GARBOW, B. S., IKEBE, Y., KLEMA, V. C. & MOLER, C. B. (1974): *Matrix Eigensystem Routines - EISPACK Guide*. In: GOOS, G. & HARTMANIS, J. [Hrsg.]: *Lect. Notes Comput. Sci., Lecture Notes in Computer Science*, Bd. 6, 6 Aufl., 387. Springer Berlin Heidelberg, Berlin, Heidelberg.
- SMITH, J. B. (1997): *Setting priorities for adapting to climate change*. In: *Glob. Environ. Chang.*, **7**, 3: 251–264.
- SOLOMON, S., PLATTNER, G.-K., KNUTTI, R. & FRIEDLINGSTEIN, P. (2009): *Irreversible climate change due to carbon dioxide emissions*. In: *Proc. Natl. Acad. Sci.*, **106**, 6: 1704–1709.
- SOUSA, P. M., TRIGO, R. M., AIZPURUA, P., NIETO, R., GIMENO, L. & GARCIA-HERRERA, R. (2011): *Trends and extremes of drought indices throughout the 20th century in the Mediterranean*. In: *Nat. Hazards Earth Syst. Sci.*, **11**, 1: 33–51.
- STAINFORTH, D. A., AINA, T., CHRISTENSEN, C., COLLINS, M., FAULL, N., FRAME, D. J., KETTLEBOROUGH, J. A., KNIGHT, S., MARTIN, A., MURPHY, J. M., PIANI, C., SEXTON, D., SMITH, L. A., SPICER, R. A., THORPE, A. J. & ALLEN, M. R. (2005): *Uncertainty in predictions of the climate response to rising levels of greenhouse gases*. In: *Nature*, **433**, 7024: 403–406.
- STEPHENSON, D. B. (2000): *Use of the Odds Ratio for Diagnosing Forecast Skill*. In: *Weather Forecast.*, **15**, 2: 221–232.
- STEVENS, B. & BONY, S. (2013): *What Are Climate Models Missing?* In: *Science*, **340**, 6136: 1053–1054.
- STEVENS, B., GIORGETTA, M., ESCH, M., MAURITSEN, T., CRUEGER, T., RAST, S., SALZMANN, M., SCHMIDT, H., BADER, J., BLOCK, K., BROKOPF, R., FAST, I., KINNE, S., KORNBLUEH, L., LOHMANN, U., PINCUS, R., REICHLER, T. & ROECKNER, E. (2013): *Atmospheric component of the MPI-M Earth System Model: ECHAM6*. In: *J. Adv. Model. Earth Syst.*, **5**, 2: 146–172.
- SU, F., DUAN, X., CHEN, D., HAO, Z. & CUO, L. (2013): *Evaluation of the Global Climate Models in the CMIP5 over the Tibetan Plateau*. In: *J. Clim.*, **26**, 10: 3187–3208.

- TANAKA, A., TAKAHASHI, K., SHIOGAMA, H., HANASAKI, N., MASAKI, Y., ITO, A., NODA, H., HIJIOKA, Y. & EMORI, S. (2017): *On the scaling of climate impact indicators with global mean temperature increase: a case study of terrestrial ecosystems and water resources*. In: *Clim. Change*, **141**, 4: 775–782.
- TEBALDI, C. & KNUTTI, R. (2007): *The use of the multi-model ensemble in probabilistic climate projections*. In: *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.*, **365**, 1857: 2053–2075.
- TEBALDI, C., MEARN, L. O., NYCHKA, D. & SMITH, R. L. (2004): *Regional probabilities of precipitation change: A Bayesian analysis of multimodel simulations*. In: *Geophys. Res. Lett.*, **31**, 24: L24213.
- THORNE, P. W., PARKER, D. E., CHRISTY, J. R. & MEARS, C. A. (2005): *Uncertainties in Climate Trends: Lessons from Upper-Air Temperature Records*. In: *Bull. Am. Meteorol. Soc.*, **86**, 10: 1437–1442.
- VAN DER LINDEN, P. & MITCHELL, J. F. B. (2009): *ENSEMBLES: Climate Change and its Impacts: Summary of research and results from the ENSEMBLES project*. Techn. Ber., Met Office Hadley Centre, Exeter, U. K.
- VAN MEIJGAARD, E., VAN ULFT, L., VAN DE BERG, W., BOSVELD, F. C., VAN DEN HURK, B., LENDERINK, G. & SIEBESMA, A. (2008): *The KNMI regional atmospheric climate model RACMO version 2.1*. Techn. Ber., 302, Royal Netherlands Meteorological Institute. URL <http://www.sciamachy-validation.org/bibliotheek/knmipubTR/TR302.pdf>(16.08.2017).
- VOLDOIRE, A., SANCHEZ-GOMEZ, E., SALAS Y MÉLIA, D., DECHARME, B., CASSOU, C., SÉNÉSI, S., VALCKE, S., BEAU, I., ALIAS, A., CHEVALLIER, M., DÉQUÉ, M., DESHAYES, J., DOUVILLE, H., FERNANDEZ, E., MADEC, G., MAISONNAVE, E., MOINE, M.-P., PLANTON, S., SAINT-MARTIN, D., SZOPA, S., TYTECA, S., ALKAMA, R., BELAMARI, S., BRAUN, A., COQUART, L. & CHAUVIN, F. (2013): *The CNRM-CM5.1 global climate model: description and basic evaluation*. In: *Clim. Dyn.*, **40**, 9-10: 2091–2121.
- VOLODIN, E. M., DIANSKII, N. A. & GUSEV, A. V. (2010): *Simulating present-day climate with the INMCM4.0 coupled model of the atmospheric and oceanic general circulations*. In: *Izv. Atmos. Ocean. Phys.*, **46**, 4: 414–431.
- VON STORCH, H. & ZWIERS, F. W. (1999): *Statistical analysis in climate research*. 1 Aufl. Cambridge Univ. Press, Cambridge [u.a.].

- WALSH, J. E., CHAPMAN, W. L., ROMANOVSKY, V., CHRISTENSEN, J. H. & STENDEL, M. (2008): *Global Climate Model Performance over Alaska and Greenland*. In: J. Clim., **21**, 23: 6156–6174.
- WANG, B., WAN, H., JI, Z., ZHANG, X., YU, R., YU, Y. & LIU, H. (2004): *Design of a new dynamical core for global atmospheric models based on some efficient numerical methods*. In: Sci. China Ser. A, **47**, S1: 4–21.
- WANG, C., ZHANG, L., LEE, S.-K., WU, L. & MECHOSO, C. R. (2014): *A global perspective on CMIP5 climate model biases*. In: Nat. Clim. Chang., **4**, 3: 201–205.
- WARTENBURGER, R., HIRSCHI, M., DONAT, M. G., GREVE, P., PITMAN, A. J. & SENEVIRATNE, S. I. (2017): *Changes in regional climate extremes as a function of global mean temperature: an interactive plotting framework*. In: Geosci. Model Dev. Discuss., 1–30.
- WATANABE, M., SUZUKI, T., O’ISHI, R., KOMURO, Y., WATANABE, S., EMORI, S., TAKEMURA, T., CHIKIRA, M., OGURA, T., SEKIGUCHI, M., TAKATA, K., YAMAZAKI, D., YOKOHATA, T., NOZAWA, T., HASUMI, H., TATEBE, H. & KIMOTO, M. (2010): *Improved Climate Simulation by MIROC5: Mean States, Variability, and Climate Sensitivity*. In: J. Clim., **23**, 23: 6312–6335.
- WATANABE, S., HAJIMA, T., SUDO, K., NAGASHIMA, T., TAKEMURA, T., OKAJIMA, H., NOZAWA, T., KAWASE, H., ABE, M., YOKOHATA, T., ISE, T., SATO, H., KATO, E., TAKATA, K., EMORI, S. & KAWAMIYA, M. (2011): *MIROC-ESM 2010: model description and basic results of CMIP5-20c3m experiments*. In: Geosci. Model Dev., **4**, 4: 845–872.
- WEIGEL, A. P., KNUTTI, R., LINIGER, M. A. & APPENZELLER, C. (2010): *Risks of Model Weighting in Multimodel Climate Projections*. In: J. Clim., **23**, 15: 4175–4191.
- WILKINSON, J. M. (2017): *A Technique for Verification of Convection-Permitting NWP Model Deterministic Forecasts of Lightning Activity*. In: Weather Forecast., **32**, 1: 97–115.
- WILKS, D. S. (2006): *Statistical methods in the atmospheric sciences*. 2. Aufl. Elsevier, Amsterdam.
- WOODCOCK, F. (1976): *The Evaluation of Yes/No Forecasts for Scientific and Administrative Purposes*. In: Mon. Weather Rev., **104**, 10: 1209–1214.

- WU, D., ZHAO, X., LIANG, S., ZHOU, T., HUANG, K., TANG, B. & ZHAO, W. (2015): *Time-lag effects of global vegetation responses to climate change*. In: Glob. Chang. Biol., **21**, 9: 3520–3531.
- WU, T., YU, R. & ZHANG, F. (2008): *A Modified Dynamic Framework for the Atmospheric Spectral Model and Its Application*. In: J. Atmos. Sci., **65**, 7: 2235–2253.
- WU, T., YU, R., ZHANG, F., WANG, Z., DONG, M., WANG, L., JIN, X., CHEN, D. & LI, L. (2010): *The Beijing Climate Center atmospheric general circulation model: Description and its performance for the present-day climate*. In: Clim. Dyn., **34**, 1: 123–147.
- YIN, H., DONAT, M. G., ALEXANDER, L. V. & SUN, Y. (2015): *Multi-dataset comparison of gridded observed temperature and precipitation extremes over China*. In: Int. J. Climatol., **35**, 10: 2809–2827.
- YLHÄISI, J. S., RÄISÄNEN, J., MASSON, D., RÄTY, O. & JÄRVINEN, H. (2015): *How does model development affect climate projections?* In: Atmos. Sci. Lett., **16**, 3: 414–419.
- YUKIMOTO, S., ADACHI, Y., HOSAKA, M., SAKAMI, T., YOSHIMURA, H., HIRABARA, M., TANAKA, T. Y., SHINDO, E., TSUJINO, H., DEUSHI, M., MIZUTA, R., YABU, S., OBATA, A., NAKANO, H., KOSHIRO, T., OSE, T. & KITO, A. (2012): *A New Global Climate Model of the Meteorological Research Institute: MRI-CGCM3 -Model Description and Basic Performance*. In: J. Meteorol. Soc. Japan, **90A**: 23–64.
- ZHANG, J., LI, L., ZHOU, T. & XIN, X. (2013): *Evaluation of spring persistent rainfall over East Asia in CMIP3/CMIP5 AGCM simulations*. In: Adv. Atmos. Sci., **30**, 6: 1587–1600.
- ZWIERS, F. W. & ZHANG, X. (2003): *Toward Regional-Scale Climate Change Detection*. In: J. Clim., **16**, 5: 793–797.



# Danksagungen

Die vorliegende Studie entstand an der Julius-Maximilians-Universität Würzburg am Institut für Geographie und Geologie am Lehrstuhl für Physische Geographie. Ich möchte mich bei der Deutschen Forschungsgemeinschaft (DFG) für die Finanzierung, im Rahmen des COMEPRO-Projekts (Comparison of Metrics for Probabilistic Climate Change Projections of Mediterranean Precipitation), bedanken. Dank gilt außerdem dem „Program for Climate Model Diagnosis and Intercomparison“ (PCMDI) und „World Climate Research Programme“ (WCRP), welche die Modelldatensätze von CMIP3, CMIP5 und CORDEX zur Verfügung gestellt haben. Gleiches gilt für die Climate Research Unit (CRU), das „European Centre for Medium-Range Weather Forecasts“ (ECMWF) sowie dem „ECA&D project“ für die Bereitstellung der drei verwendeten Referenzdatensätze.

An erster Stelle gilt mein Dank Prof. Dr. Heiko Paeth für die Anleitung, Koordination und Inspiration die mir zuteil wurde. Für seine Verfügbarkeit bezüglich Fragen und Unterstützung durch Vertrauen und neue Denkanstöße mit fachlicher Expertise möchte ich mich insbesondere für die letzten Jahre, aber auch schon für die Zeit der Masterarbeit zuvor, herzlich bedanken. Des Weiteren bedanke ich mich bei Prof. Dr. Jucundus Jacobeit, der sich bereit erklärt hat, die Aufgabe des Zweitgutachters meiner Dissertation zu übernehmen. Die Projekttreffen mit ihm, Elke Hertig und Irena Kaspar-Ott, fanden in harmonischer Atmosphäre statt und waren immer hilfreich und informativ. Dies gilt auch für das Arbeitsklima der Klimaarbeitsgruppe in Würzburg. So konnte ich stets auf den Rat und die Hilfe aller Mitarbeiter, Dr. Felix Pollinger, Dr. Andreas Paxian, Dr. Hervé Awoyé, Doro Schill und Luzia Keupp vertrauen. Einen besonderen Dank möchte ich an dieser Stelle Dr. Felix Pollinger für seine unermüdliche Unterstützung und zahllosen Besprechungen ausdrücken.

Abseits der fachlichen Hilfestellung möchte ich mich auch für die moralische Unterstützung und teilweise auch willkommene Ablenkung durch meine Freunde bedanken. Da ich nicht jeden aufzählen kann, möchte ich mich hiermit bei allen herzlich bedanken.

Die wichtigsten Stützen für mich waren meine Schwester, meine Eltern und Bianca. Sie alle unterstützten mich zu jeder Phase, bis zuletzt beim Korrekturlesen dieser Dissertation. Das werde ich nie vergessen und ich bedanke mich von ganzen Herzen, für Alles.



# Erklärung

Hiermit versichere ich, Christoph Ring, geboren am 28.06.1987 in Dortmund, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe, dass alle Stellen der Arbeit, die wörtlich oder sinngemäß aus anderen Quellen übernommen wurden, als solche kenntlich gemacht sind und dass die Arbeit in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegt wurde.

Würzburg, Monat Jahr