


A comparison of metrics for assessing state-of-the-art climate models and implications for probabilistic projections of climate change

Christoph Ring¹  · Felix Pollinger¹ · Irena Kaspar-Ott² · Elke Hertig² · Jucundus Jacobeit² · Heiko Paeth¹

Abstract A major task of climate science are reliable projections of climate change for the future. To enable more solid statements and to decrease the range of uncertainty, global general circulation models and regional climate models are evaluated based on a 2×2 contingency table approach to generate model weights. These weights are compared among different methodologies and their impact on probabilistic projections of temperature and precipitation changes is investigated. Simulated seasonal precipitation and temperature for both 50-year trends and climatological means are assessed at two spatial scales: in seven study regions around the globe and in eight sub-regions of the Mediterranean area. Overall, 24 models of phase 3 and 38 models of phase 5 of the Coupled Model Intercomparison Project altogether 159 transient simulations of precipitation and 119 of temperature from four emissions scenarios are evaluated against the ERA-20C reanalysis over

the 20th century. The results show high conformity with previous model evaluation studies. The metrics reveal that mean of precipitation and both temperature mean and trend agree well with the reference dataset and indicate improvement for the more recent ensemble mean, especially for temperature. The method is highly transferrable to a variety of further applications in climate science. Overall, there are regional differences of simulation quality, however, these are less pronounced than those between the results for 50-year mean and trend. The trend results are suitable for assigning weighting factors to climate models. Yet, the implications for probabilistic climate projections is strictly dependent on the region and season.

Keywords Model-weighting · CMIP3 · CMIP5 · Mediterranean · Spatial scales · Probabilistic climate projections

✉ Christoph Ring
christoph.ring@uni-wuerzburg.de

Felix Pollinger
felix.pollinger@uni-wuerzburg.de

Irena Kaspar-Ott
irena.kaspar-ott@geo.uni-augsburg.de

Elke Hertig
elke.hertig@geo.uni-augsburg.de

Jucundus Jacobeit
jucundus.jacobeit@geo.uni-augsburg.de

Heiko Paeth
heiko.paeth@uni-wuerzburg.de

¹ Institute of Geography and Geology, University of Wuerzburg, Am Hubland, 97074 Wuerzburg, Germany

² Institute of Geography, University of Augsburg, Alter Postweg 118, 86159 Augsburg, Germany

1 Introduction

In spite of increasing knowledge of climatic processes and enhanced computer performances, the estimation of uncertainty about future regional climate change remains a particularly challenging task (e.g. Power et al. 2012; Knutti and Sedláček 2012). It is crucial to quantify uncertainty to allow sound adaptation strategies and sustainable political decisions (e.g. Hawkins et al. 2016; Clark et al. 2016).

The ideal way of evaluating climate models is still subject to discussion (e.g. Gleckler et al. 2008; Knutti 2010). Different metrics have been used in climate science to evaluate different model characteristics (e.g. Bishop and Abramowitz 2013; Sanderson et al. 2015; Ring et al. 2016). Further, the model performance does not only depend on the choice of metrics but of region and variable as well.

This study utilizes simple but powerful metrics, with universal applicability, based on a 2×2 contingency table, to address the assessment of uncertainty of temperature and precipitation changes towards the end of the twenty-first century. We consider 24 models from CMIP3 and 38 models from CMIP5 (Randall et al. 2007; Flato et al. 2013) which participated in the fourth (AR4) and fifth Assessment Report (AR5) from 2007 to 2013, respectively. In addition, 18 simulations from RCMs from the World Climate Research Program Coordinated Regional Downscaling Experiment (CORDEX) are considered (Giorgi et al. 2009). The 2×2 table metrics have frequently been used in hydrological and meteorological research (e.g. Stephenson 2000; Thornes and Stephenson 2001; Armistead 2013), however, they have barely been used in climate science (Woodcock 1976; Paeth et al. 2006). Most recent studies utilize contingency tables for numerical weather prediction (NWP) (e.g. Done et al. 2004; Ghelli and Primo 2009; Gill and Buchanan 2014; Wilkinson 2017).

Previous studies show distinct differences in simulation quality for different regions and variables (i.e. Power et al. 2012; Miao et al. 2014; Eum et al. 2014; Ring et al. 2016). As a consequence, Perkins et al. (2007), Gillett et al. (2015) and Haughton et al. (2015) showed that sophisticated weighting of climate models may affect probability density functions (PDF) of future climate change and, hence, could decrease the range of projected climate change. However, Knutti et al. (2010) suggested to be cautious when weighting multi model ensembles (MMEs). The projected changes of the CMIP3 and CMIP5 MMEs are estimated with an equally weighted PDF (Tebaldi et al. 2007). We compare this PDF with a metric-weighted PDF in order to assess the change in uncertainty. This evaluation is performed for each region, season and scenario. While most studies are limited to a specific region or a preselection of models (i.e. Sheffield et al. 2013; Miao et al. 2014) we explore an easily transferable metric for various sizes and types of study areas for all available GCMs. We analyze seven regions at continental level and eight sub-regions at sub-continental level located in the Mediterranean Basin to assess the metrics' transferability. The focus is on seven globally distributed regions with different climatic and surface characteristics.

We apply the 2×2 table metrics to determine and compare weights for each model for 50-year trends and climatological means from 1960 to 2009 of seasonal and annual precipitation sums and temperature means. As reference data and for the derivation of model weights, we use the ERA-20C reanalysis (Poli et al. 2016) because they are available over the entire 50-year time period. We expect that in terms of temperature simulation the applied skill scores are higher for CMIP5 compared with CMIP3 (Reichler and Kim 2008; Knutti et al. 2013; Wright et al.

2016; Koutroulis et al. 2016). A general decrease of precipitation in the dry season of the Mediterranean was found by Giorgi and Lionello (2008). However, the pattern of precipitation changes is spatially quite incoherent. Therefore, the assessment of weighting metrics is very challenging (Hewitson and Crane 2006; Hawkins et al. 2016).

Section 2 is dedicated to the considered reference and model datasets. Further, we present the seven large study areas as well as the eight sub-regions of the Mediterranean. The applied 2×2 table metrics and other methods are explained in Sect. 3. Section 4 deals with the results, comprising the derived coefficients of each simulation for all regions and seasons, the correlation between different varieties of the skill scores, and the systematic differences between CMIP3 and CMIP5. In addition, we analyze the impact of the weighting schemes on PDFs of future climate change and extend the analysis to the eight sub-regions of the Mediterranean area to explore the scale effect. In Sect. 5, we summarize and discuss our results, while conclusions are drawn in Sect. 6.

2 Study areas and datasets

2.1 Study areas

The seven large regions and eight Mediterranean sub-regions are depicted in Fig. 1. We aim to cover a broad spectrum of climates as well as the entire globe including water and land surface (Globe). Simulation performance over the oceans is analyzed for the tropical Atlantic (Atlantic) and Pacific Oceans (Pacific), both spanning from 30°S to 30°N , and the polar to subpolar Arctic Ocean north of 75°N (Arctic). The remaining three study areas cover land surfaces only: Africa and the Arabian Peninsula with Iran (Africa), North and Central America (America) and the Mediterranean region (Medit), which is the smallest of the large study areas. All of these study areas have high relevance for the study of climatic phenomena and extremes, i.e. El Niño (Pacific), hurricanes (Atlantic, America) or droughts (Africa, Medit) which stresses the need to reduce uncertainty of climate change. In order to analyze the transferability of our results to smaller scales, we divided Medit into eight sub-regions (Fig. 1). The Mediterranean is a hot spot for future climate change and, thus, regional differences are of high relevance (Giorgi 2006; Diffenbaugh and Giorgi 2012; Paeth et al. 2016). The eight sub-regions have been identified by means of a principal component analysis of annually aggregated precipitation sums for the last century. Thus, we received eight rather homogeneous areas which differ in terms of the amount of precipitation. They are named as follows: North Atlantic, Spain, North Africa, Italy, Balkans, Aegean, Black Sea, Middle East (Fig. 1).

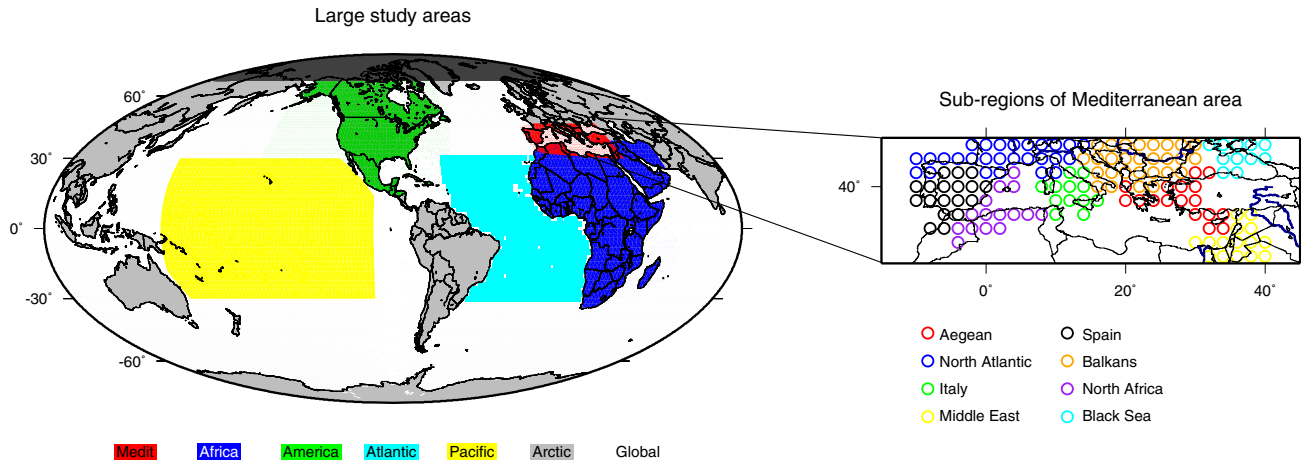


Fig. 1 Overview of the seven large study areas and the eight sub-regions of the Mediterranean area

2.2 Validation data

We use seasonal and annual sums of precipitation and mean temperature from the ERA-20C reanalysis provided by the Centre for Medium Range Weather Forecasts (ECMWF) for validation (Poli et al. 2016). It offers a physically consistent data set with global coverage of temperature and precipitation at a T159 resolution, provided in $1^\circ \times 1^\circ$ grid, for the 20th century without any missing values. Albeit only a substitution for observational data, previous studies confirmed ERA-20C to be a reliable basis even for climatological extremes of precipitation and temperature in the period after 1950 (Donat et al. 2016; Dittus et al. 2016). Furthermore, this study focuses on both land and sea surfaces and, hence, we prefer ERA-20C over other frequently used observational datasets such as e.g. the Climate Research Unit (CRU) dataset. Nonetheless, as a supplement we use CRU TS3.23 (Mitchell and Jones 2005) and the E-OBS V12 dataset (Haylock et al. 2008) for the Medit sub-regions to assess the sensitivity of our results to different reference data sets (Table 1). Both datasets are based on observational data provided by weather stations and are interpolated on a regular $0.5^\circ \times 0.5^\circ/0.25^\circ \times 0.25^\circ$, respectively. CRU has a spatial resolution of $0.5^\circ \times 0.5^\circ$ and E-OBS of $0.25^\circ \times 0.25^\circ$. All validation datasets are preprocessed with a REMAPCON interpolation by the Climate Data Operators (CDO) (Schulzweida et al. 2009) on a $2^\circ \times 2^\circ$ grid to enable best comparability of metrics.

2.3 Model data

In this study, we use 20th century simulations (20C3m/ Historical) and future projections with different emissions scenarios of all available CMIP3 and CMIP5 GCMs. Four emissions scenarios are taken into account: the Special Report on Emissions Scenarios (SRES) A1B, A2 scenarios for CMIP3 (Nakicenovic et al. 2000) and the representative concentration pathways (RCP) RCP4.5 and RCP8.5 scenarios for CMIP5 (Moss et al. 2010).

The impacts of weighting metrics are assessed for the moderate scenarios A1B (CMIP3) and RCP4.5 (CMIP5) as well as for the high-end scenarios A2 (CMIP3) and RCP8.5 (CMIP5) with a substantial increase of emissions (Randall et al. 2007; Flato et al. 2013). They represent a broad coverage of different future climatic pathways. RCP4.5 represents rather low greenhouse gas emissions, A1B intermediate and RCP8.5 and A2 both high emissions. Detailed descriptions are published by Nakicenovic et al. (2000) and Moss et al. (2010). Since for some models only precipitation or temperature were available, we analyze 159 simulations of precipitation and 119 of temperature for the 20th century. For the moderate scenario A1B (RCP4.5) there are 54 (105) simulations of precipitation and 57 (62) for temperature of CMIP3 (CMIP5). A total of 39 (82) simulations of precipitation and 39 (55) simulations of temperature are available for A2 (RCP8.5) scenario. Since most historical simulations of CMIP3 (CMIP5, CORDEX) end in 1999

Table 1 Datasets used for evaluation

	Timeframe	Spatial coverage	References
ERA-20C	1900–2009	Global	Poli et al. (2016)
CRU (TS3.23)	1901–2014	Global (land only)	Mitchell and Jones (2005)
E-OBS (v12)	1950–2015	Europe (land only)	Haylock et al. (2008)

(2005) we use the first years of the A1B (RCP4.5) scenario to complete the 50-year timeframe until 2009. The main reason for this approach is that for each simulation of A2 (RCP8.5) a corresponding simulation of A1B (RCP4.5) is available but not the other way around. Hence, we can calculate a unique weight of each model which can be applied on both emission scenarios. Further, the prescribed forcing of the scenarios is rather similar in the first years of the 21st century.

Tables 2 and 3 show a detailed summary of which model participates in which scenario, which variable and how many simulations are available. All models are preprocessed with an inverse distance interpolation to a common $2^\circ \times 2^\circ$ grid to guarantee the spatial resolution for each simulation (Babak and Deutsch 2009).

In addition to the GCMs, we also apply our analysis to RCMs provided by the CORDEX project for the Medit study area and its eight sub-regions (Giorgi et al. 2009). The regional climate simulations use boundary conditions

Table 2 Climate models of CMIP3 used in this study (Randall et al. 2007)

Models	Precipitation		Temperature	
	Scenario			
	A1b	A2	A1b	A2
BCCR_BCM2.0	1	1	1	1
CGCM3.1 (T47)	2–6	2–6	2–6	2–6
CGCM3.1 (T63)	7	–	7	–
CNRM-CM3	8	7	8	7
GFDL-CM2.0	9	8	9	8
GFDL-CM2.1	10	9	10	9
GISS-AOM	11–12	–	11–12	–
GISS-EH	13–15	10	13–15	10
GISS-ER	16–17	–	16–20	–
FGOALS-g1.0	18–20	–	21–23	–
INM-CM3.0	21	11	24	11
IPSL-CM4 (LMDZ)	22	12	25	12
INGV-SXG	23	13	26	13
MIROC3.2 (hires)	24	–	27	–
MIROC3.2 (medres)	25–27	14–16	28–30	14–16
MRI-CGCM2.3.2	28–32	17–21	31–35	17–21
ECHO-G	33–35	22–24	36–38	22–24
CSIRO-Mk3.0	36	25	39	25
CSIRO-Mk3.5	37	26	40	26
ECHAM5/MPI-OM	38–41	27–29	41–44	27–29
CCSM3	42–48	30–33	45–51	30–33
PCM	49–52	34–37	52–55	34–37
UKMO-HadCM3	53	38	56	38
UKMO-HadGEM1	54	39	57	39

The numbers indicate how many simulations of each model (scenario, variable) are used in this study

Table 3 Climate models of CMIP5 used in this study (Flato et al. 2013)

Models	Precipitation		Temperature	
	Scenario			
	RCP4.5	RCP8.5	RCP4.5	RCP8.5
ACCESS1-0	1	1	1	1
ACCESS1-3	2	2	2	2
BCC-CSM1.1	3	3	3	3
BCC-CSM1.1 (m)	4	4	4	4
CanESM2	5–9	5–9	5–9	5–9
CCSM4	10–15	10–15	10–15	10–15
CESM1-BGC	16	16	16	16
CESM1-CAM5	17–19	17–19	17–19	17–19
CMCC-CM	20	20	20	20
CMCC-CMS	21	21	21	21
CNRM-CM5	22	22–26	22	22
CSIRO-Mk3-6-0	23–32	27–36	23–32	23–32
CSIRO-Mk3L-1-2	33–35	–	33–35	–
EC-EARTH	36–39	37–41	–	–
FGOALS-g2	40	42	36	–
FIO-ESM	–	–	37	–
GFDL-CM3	41	43	38	33
GFDL-ESM2G	42	44	39	34
GFDL-ESM2M	43	45	40	35
GISS-E2-H-CC	44	46	–	–
GISS-E2-H	45–60	47–51	–	–
GISS-E2-R-CC	61	52	–	–
GISS-E2-R	62–78	53–57	–	–
HadGEM2-AO	79	58	41	36
HadGEM2-CC	80	59	–	–
HadGEM2-ES	81–84	60–63	–	–
INMCM4	85	64	42	37
IPSL-CM5A-LR	86–89	65–68	43–46	38–41
IPSL-CM5A-MR	90	69	47	42
IPSL-CM5B-LR	91	70	48	43
MIROC5	92–94	71–73	49–51	44–46
MIROC-ESM-CHEM	95	74	52	47
MIROC-ESM	96	75	53	48
MPI-ESM-LR (ECHAM6)	97–99	76–78	54–56	49–51
MPI-ESM-MR (ECHAM6)	100–102	79	57–58	52
MRI-CGCM3	103	80	60	53
NorESM1-ME	104	81	61	54
NorESM1-M	105	82	62	55

The numbers indicate how many simulations of each model (scenario, variable) are used in this study

of GCMs from CMIP5 (Jacob et al. 2014). RCMs offer advantages such as high spatial resolution (0.11° – 0.44°) but also might be influenced by uncertainty due to affection of coarse-scale systematic errors from GCMs (Giorgi et al.

Table 4 Regional climate models of CORDEX used in this study (Jacob et al. 2014)

Global model	Regional model	Resolution
CNRM-CERFACS-CNRM-CM5	SMHI-RCA4	$0.11^\circ \times 0.11^\circ$
ICHEC-EC-EARTH	SMHI-RCA4	$0.11^\circ \times 0.11^\circ$
ICHEC-EC-EARTH	DMI-HIRHAM5	$0.11^\circ \times 0.11^\circ$
IPSL-IPSL-CM5A-MR	SMHI-RCA4	$0.11^\circ \times 0.11^\circ$
MOHC-HadGEM2-ES	SMHI-RCA4	$0.11^\circ \times 0.11^\circ$
MPI-M-MPI-ESM-LR	SMHI-RCA4	$0.11^\circ \times 0.11^\circ$
CCCma-CanESM2	SMHI-RCA4	$0.11^\circ \times 0.11^\circ$
CNRM-CERFACS-CNRM-CM5	SMHI-RCA4	$0.11^\circ \times 0.11^\circ$
CSIRO-QCCCE-CSIRO-Mk3-6-0	SMHI-RCA4	$0.11^\circ \times 0.11^\circ$
ICHEC-EC-EARTH	SMHI-RCA4	$0.44^\circ \times 0.44^\circ$
ICHEC-EC-EARTH	KNMI-RACMO22E	$0.44^\circ \times 0.44^\circ$
ICHEC-EC-EARTH	DMI-HIRHAM5	$0.44^\circ \times 0.44^\circ$
IPSL-IPSL-CM5A-MR	SMHI-RCA4	$0.44^\circ \times 0.44^\circ$
MIROC-MIROC5	SMHI-RCA4	$0.44^\circ \times 0.44^\circ$
MOHC-HadGEM2-ES	SMHI-RCA4	$0.44^\circ \times 0.44^\circ$
MPI-M-MPI-ESM-LR	SMHI-RCA4	$0.44^\circ \times 0.44^\circ$
NCC-NorESM1-M	SMHI-RCA4	$0.44^\circ \times 0.44^\circ$
NOAA-GFDL-GFDL-ESM2M	SMHI-RCA4	$0.44^\circ \times 0.44^\circ$

2009; Pielke and Wilby 2012; Ayar et al. 2016). Since the 18 simulations of CORDEX (see Table 4) cover a shorter historical timeframe starting not earlier than 1970, the first decade of our 50-year investigation period is missing.

3 Methodology

In this study we explore six evaluation metrics based on a 2×2 table, providing insight in the differences of each metric and its suitability for model weighting. Originally, such contingency tables are used to measure correlations of pairwise nominal data. All metric data can be transformed into nominal data by implementing a threshold value (TV) in order to categorize the data e.g. into two groups: above and below or equal this TV. The applied metrics utilize a pairwise grid box-based comparison between observational and model data. This aspect is similar in Numerical Weather Prediction (NWP) where 2×2 -table approaches are frequently used to estimate success of forecasts (e.g. Ghelli and Primo 2009; Done et al. 2004). However, generally NWP contingency tables are used to verify forecasts of events, e.g. observed tornadoes in terms of hit- or false alarm rates (Wilks 2006; Stephenson 2000). In this study,

the basic idea is transferred to evaluate climate models. In contrast to NWP, we use this 2×2 -table for calculating measures of proximity between the GCMs' regional patterns of precipitation and temperature with those of the observations.

First, we calculated the changes in each dataset for each grid box. This was performed for every region and season. Here, we defined the difference of the mean over the first 15 years (1960–1975) and the mean of the last 15 years (1995–2009) as ‘trend’ to avoid the restrictions of a linear regression. This turned out to be slightly more robust for the sub-regions of Medit than the common regression coefficient. For each region and season the regional mean of the trends was calculated as TV. This was performed separately for each climate model and the evaluation data. Since a change of input variables offers further achievements with little effort we expand our entire study on the 50-year mean. Finally, we compare each grid box for every situation (region, season) for simulation and evaluation data. Since both regional means are calculated individually for model and observational data, all 2×2 -table approaches basically measure the differences of regional distribution.

We consider the regional mean of trend and climatological mean as TV the most appropriate to estimate model performances. As an alternative the median or other percentiles could be used. However, those measures are insensitive to outliers and extreme values, what we consider a disadvantage for our study. Also, mainly due to the relatively small numbers of gridpoints available for the Mediterranean study areas, the use of more extreme percentiles would result in rather similar coefficients. Thus, the 2×2 table is filled as shown in Table 5.

All grid boxes with values of the considered quantity that exceed the respective TV for both datasets are counted as *a*. Each grid box underrunning the TV in both datasets is counted as *d*. The fields *b* and *c* are filled with the counts of grid boxes with a disagreement between simulation and validation data. Therefore, any bias between observational and model data is removed at this point as the contingency table only measures the relative behavior to both individual TVs. As a result, high values for *a* and *d* are equivalent to a high agreement of the respective simulation and the evaluation dataset. Based on these four values *a*, *b*, *c*, *d*, we calculate several skill scores to gain the gross weight of each simulation. The choice of the applied six skill scores is inspired by and named corresponding to Stephenson (2000), expended by the phi-correlation (e.g. Bortz et al. 2008). All metrics in Table 6 are based on the same 2×2 table approach and only differ in terms of the formula and range. For reasons of clarity, we first show the results of the phi-correlation in Sect. 4.1–4.4. The results of all six

Table 5 The basic 2×2 table for the applied metrics

	Evaluation data value ≥ threshold x	Evaluation data value < threshold x	Σ
Simulation value ≥ threshold y	a	b	a + b
Simulation value < threshold y	c	d	c + d
Σ	a + c	b + d	n

All cells (a, b, c, d) are filled based on the over-/underrun of the threshold x (for evaluation data) or y (for simulation)

x regional eval. data mean, y regional sim. data mean

metrics are compared in Sect. 4.5. In the following, simulation performance or quality is understood as a measure of similarity between the spatial patterns of model and evaluation data indicated by the metrics.

For each simulation we get one individual coefficient for every skill score. For models that have more than one simulation, the mean over the coefficients of all simulations is used to produce the model coefficient. To eliminate negative values we used exponential (exp.) coefficients to the basis of e for all simulations. This is necessary for the Log Odds-metric which has a theoretical range from negative to positive infinity. Next, the final (raw) weight is calculated by dividing the respective exp. coefficient by the sum of all exp. coefficients of a specific situation (region, season). Thus, for each model a raw weight can be calculated. Finally, as a weighting approach (WA1) we calculate the final weight for each model of CMIP3 and CMIP5 by dividing each model's raw weight by the sum over all raw weights over the MME. Based on these values the projections of CMIP3 and CMIP5 are weighted and a probability density function (PDF) is estimated. Further, we apply a second approach (WA2) for those metrics with a range starting below zero. Here, all simulations with negative coefficients were assigned zero weight. Sect. 4.1–4.5 are based on WA1, while the results of WA2 are shown in Sect. 4.6.

4 Results

4.1 Annual analysis of phi-coefficients

In Fig. 2, the Phi coefficients for annual precipitation on the left and annual mean temperature on the right for all main study areas are displayed. The black circles show the results for the trend and the white circles those for the climatological mean. On the abscissa the simulations are numbered. The assignment of each number can be taken out of Table 2 for CMIP3 and Table 3 for CMIP5. Since the original Phi coefficients are displayed, a high positive (negative) correlation of model and evaluation dataset is indicated by values close to 1 (−1) and no correlation around 0. Throughout all simulations and regions, the coefficients of the 50-year mean exceed those of the 50-year trend. For the 50-year trend of precipitation, for almost all regions the coefficients are equally spread and centered around zero, except for Globe. Here, we find a mean over all simulations of 0.18. The spread of coefficients is rather small with a minimum of −0.02 for GISS-EH (CMIP3) and a maximum of 0.29 for CGCM3.1 (CMIP3). For the others six regions the spread of coefficients is much wider. Overall, the highest coefficient is 0.44 found for MRI-CGCM2.3.2 (CMIP3) in Medit, while the lowest value is −0.43 (HadGEM2-ES, CMIP5) in the Arctic. Anyway, the means over all simulations are lower than for the Globe with a minimum of −0.04 for America and a maximum of 0.04 for Medit.

For the 50-year precipitation mean, the lowest coefficient is 0.20 for the Atlantic while the highest result is 0.91 for Africa. For all study areas, mean values over all simulations are between 0.58 for the Atlantic and 0.78 for the Globe. Here, the best results overall are shown with a minimum coefficient of 0.67. While the Globe, America and Africa show rather homogenous values, the other regions exhibit a similar spread as for the 50-year trends but on a generally higher level.

For the 50-year trend of temperature, there are higher coefficients across all study areas. Especially the Globe, Medit, Arctic and Atlantic all have simulations with values

Table 6 Overview of the applied 2×2 table metrics

Skillscore	Formula	Range	References
Phi	$\frac{(ad-bc)}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$	[−1,1]	i.e. Bortz et al. (2008)
Chi2 (CHI)	$\frac{(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$	[0,1]	Doolittle (1885)
Heidke (HEI)	$\frac{2(ad-bc)}{(a+c)(c+d)+(a+b)(b+d)}$	[−1,1]	Doolittle (1888) and Heidke (1926)
Gilbert (GSS)	$\frac{a}{a+b+c}$	[0,1]	Gilbert (1884)
Pierce (PIE)	$\frac{ad-bc}{(a+c)(b+d)}$	[−1,1]	Pierce (1884)
Log odds (LOR)	$\log a + \log d - \log b - \log c$	[−∞,∞]	Stephenson (2000)

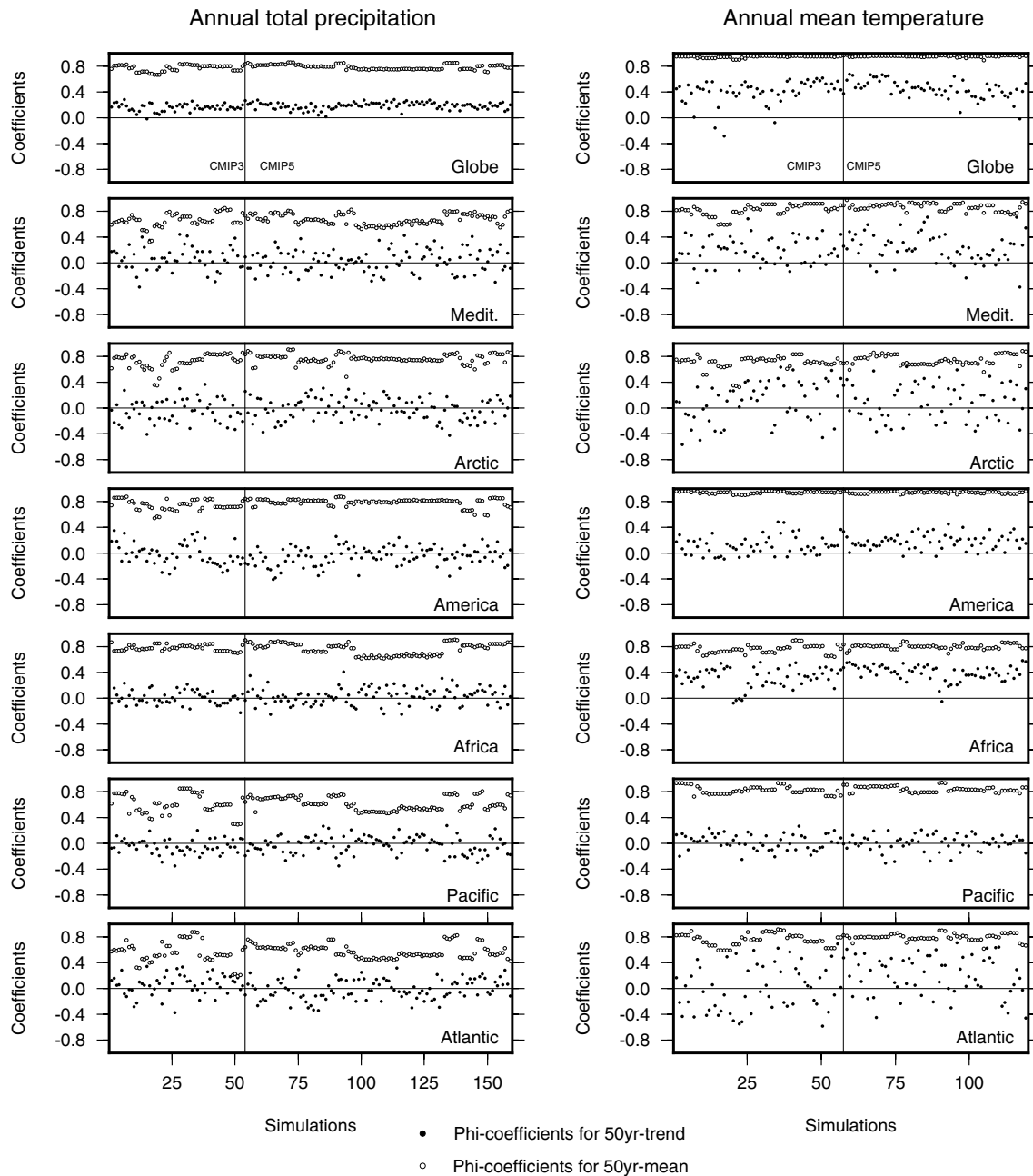


Fig. 2 Results of the phi-correlation metric for annual precipitation (*left*) and mean temperature (*right*) for all main study areas

above 0.64. For America and Africa, the majority of coefficients is positive with a less pronounced spread compared to the other study areas. The distribution of coefficients for the Pacific temperature is almost similar to the one of precipitation. Here, the mean is 0.01 for temperature and -0.03 for precipitation. Further, there are differences in the spread between the other regions as well. While America, Africa and Pacific have a spread of maximally 0.65, the amplitude for the other four regions is 0.97 and higher. The highest Phi coefficients are again achieved for Medit with

0.71 by CSIRO-Mk3-6-0 (CMIP5) while the minimum is -0.58 for CCSM3 (CMIP3) in the Atlantic. The results for the 50-year mean of temperature are very homogeneous for all regions. The minimum mean over all simulations is 0.73 for the Arctic, for which we also see a minimum coefficient of 0.33. For the other regions, there are overall very high coefficients from minimum 0.59 (Atlantic) to maximum 0.97 (Medit, Globe, America). Especially the results for Globe, America, Pacific and Medit are on a constantly high level with means from 0.84 (Medit) to 0.95 (Globe).

Overall, we see strong differences in the evaluation between 50-year trends and means. The evaluation of the means indicates a rather good performance of GCMs for both precipitation and temperature for all study areas. The analysis of the 50-year trend shows a more differentiated result between temperature and precipitation and a higher variability of Phi coefficients among study areas. However, the average coefficients of temperature are on a higher level for both mean and trend.

4.2 Seasonal analysis of phi-coefficients

In Fig. 3, seasonal phi-coefficients for 50-year trends of precipitation and temperature are displayed for the Globe. The seasons are defined by 3 months abbreviated with the first letter of each month (e.g. DJF for December, January and February). As in Fig. 2, the simulation number on the abscissa can be decoded with Tables 3 (CMIP3) and 4 (CMIP5). For precipitation the highest coefficients are found for annual values with a mean of 0.18. However, the annual coefficients are only slightly higher compared to those of maximally 0.16 in SON and minimally 0.10 in

JJA. Overall, we see no simulation exceeding the others in all seasons. Apparently there is a ranking of seasons rather than a ranking of simulations in this context. This impression is even stronger for temperature. The annual mean is highest with 0.42 and the minimum is 0.25 in JJA. DJF and SON are similar to annual, both 0.40, while MAM is intermediate between JJA and the other seasons with 0.35. As for precipitation, no simulation outperforms the others in all seasons. However, there are some simulations with all seasons above 0.4 while others show several negative coefficients. For example, the mean of CNRM-CM3 (CMIP3) over all seasons is 0.01 against 0.51 for CCSM4 (CMIP5). Overall, for temperature a much higher spread is found for all seasons in contrast to quite homogeneous coefficients based on precipitation. For the other study regions, the findings are similar to the ones of the Globe (not shown), even though there is a shift of the mean values from one region to another as seen in Sect. 4.1. The results for 50-year means of both precipitation and temperature are similar in heterogeneity but extend over an even smaller range than that shown in Fig. 3.

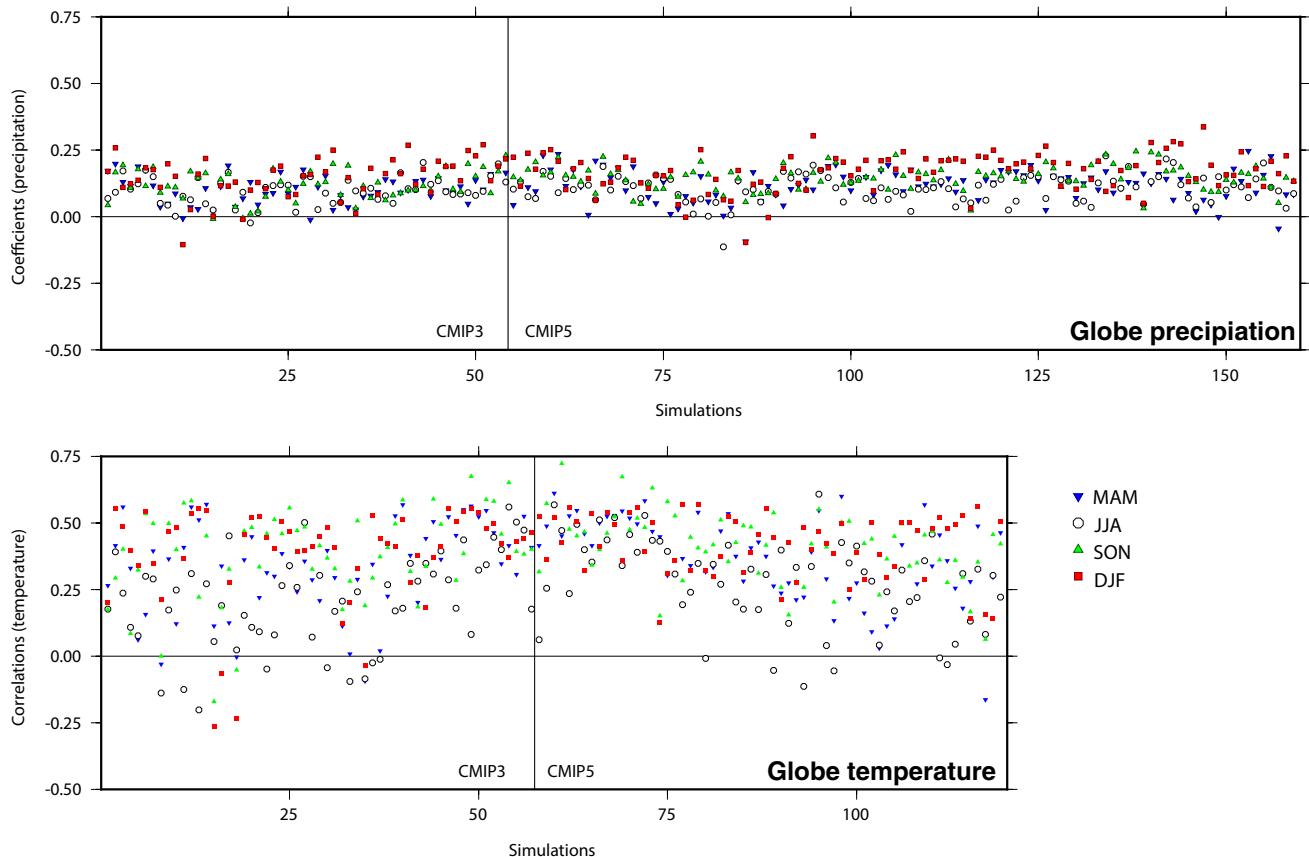


Fig. 3 Comparison of seasonal phi-coefficients for 50-year mean trends of precipitation (*top*) and temperature (*bottom*) for the Globe

4.3 Regional robustness of phi-coefficients

The regional robustness of Phi coefficients for each season is shown in Fig. 4. Here, the analysis of regional robustness focuses on the match of weights between different region. For this, the Spearman rank correlation (von Storch and Zwiers 1999) between all weights of each study areas is depicted for precipitation in the upper left part of each matrix while temperature correlations are shown in the lower right part. In the first row, the 50-year trend correlations are shown, while the results for the means are in the second row. As for the trend, most correlations are on a rather low level. There is an equal ratio between positive and negative correlation coefficients. However, the Globe shows positive correlations for almost every season and region for both temperature and precipitation. We find maximum values between 0.41 and 0.53 for the correlation between Globe and Africa, Pacific as well as Arctic. Most positive values are found in MAM for precipitation and DJF for temperature. However, for most regions the correlation is between -0.2 and $+0.2$ and, hence, reflects a low regional robustness of this metric. Overall, 80 of 105 (76.2%) correlations are positive for precipitation and 67 (63.8%) for temperature. Here, the mean over all fields is

0.11 (0.06) for precipitation (temperature). Even though the phi-coefficients of temperature are generally higher than those of precipitation (see Sect. 4.2), there are no systematic differences in the regional correlation of phi-coefficients between precipitation and temperature for most study areas.

The analysis of the 50-year mean shows 98 positive correlations (93.3%) for precipitation and 94 (90.0%) for temperature. Especially for annual, MAM and SON means there are almost no negative correlations for neither precipitation nor temperature. Overall, there is higher correlation between the rankings of the different regions compared to those of the trend. The mean over all fields is 0.36 for precipitation and 0.24 for temperature.

The ranking of Phi coefficient metric indicates that a simulation that captures the mean of the reference data in one region is likely to perform well in another region as well. In contrast, one cannot expect that the same simulation has a high skill capturing the trend as well.

4.4 Performance of CMIP3 versus CMIP5

Figure 5 shows the mean phi-coefficients from CMIP3 (black) and CMIP5 simulations (gray) for all study areas.

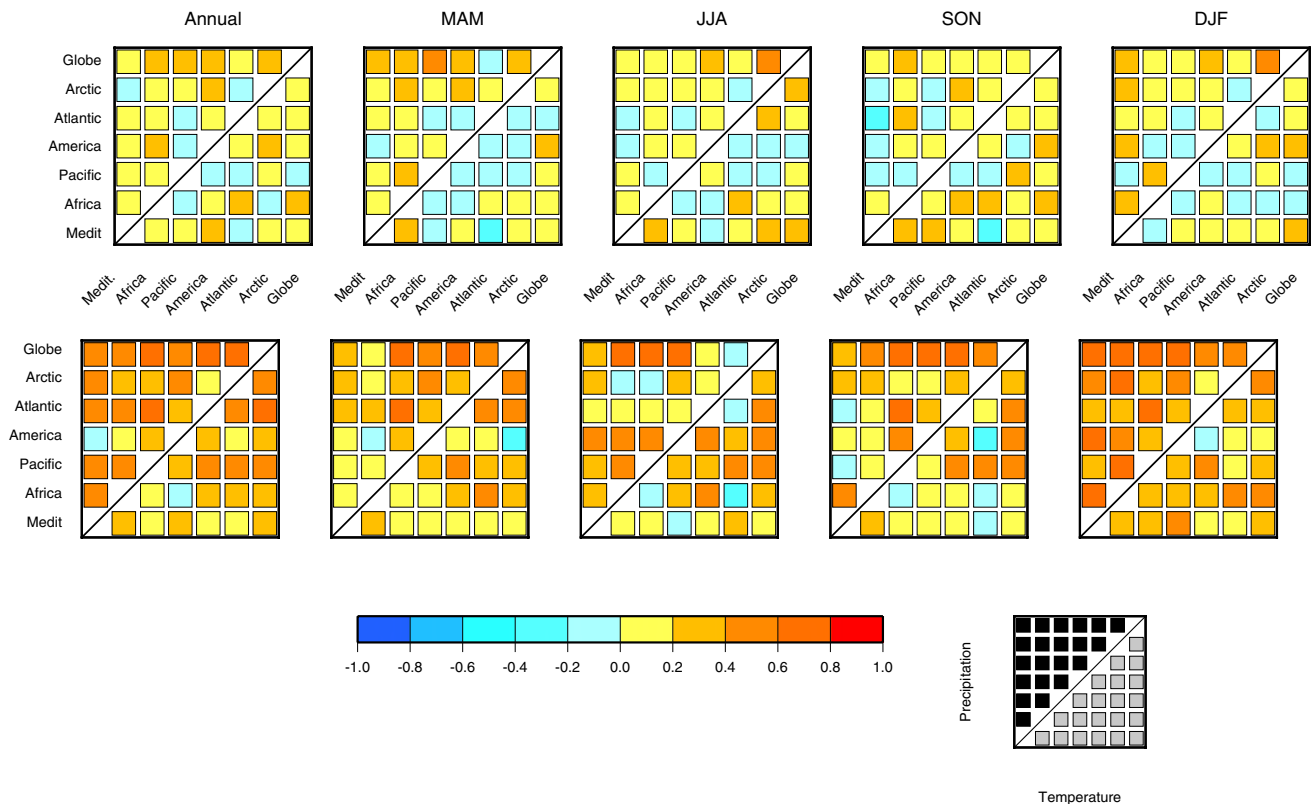


Fig. 4 Spearman correlation of phi-coefficients between all main study areas and for each season for precipitation (*top-left*) and temperature (*bottom-right*), 50-year trends (*top row*) and 50-year means (*bottom row*)

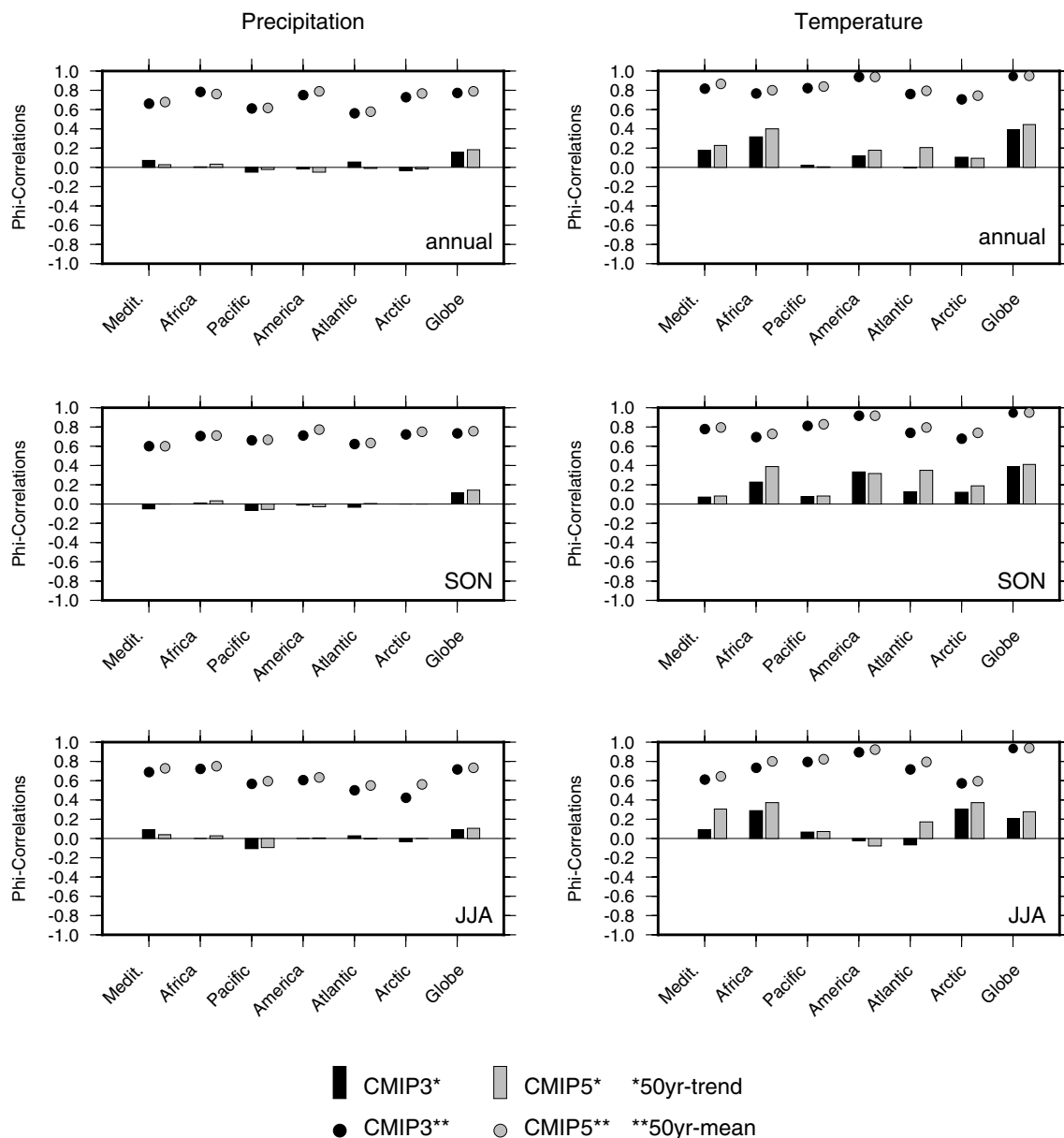


Fig. 5 Comparison of the mean of phi-coefficients between CMIP3 and CMIP5 for all main study areas

The results for annual, JJA and SON precipitation (left) and mean temperature (right) are displayed. The bars show the results for the 50-year trend, while the circles indicate the coefficients based on means. The Phi coefficients are always higher for means than for trends. The MME results for the mean indicate only minor differences between the study areas and seasons. However, for both precipitation and temperature the weakest results are found in the Arctic with a minimum of 0.42 for JJA precipitation and a maximum of 0.74 for MAM temperature (not shown in Fig. 5). For the other regions, most MME mean coefficients are rather high with a maximum of 0.95 for global SON temperature. Furthermore, we notice that

CMIP5 achieves slightly better results than CMIP3 for almost every region and season.

The picture for 50-year trends is different. Here, the highest means of precipitation for both CMIP3 and CMIP5 are found for the Globe. On average, CMIP5 is slightly improved compared with CMIP3. The best result for precipitation is 0.19 for annual values in CMIP5 for the Globe while the minimum is 0.1 for the Pacific JJA means from CMIP3. The other regions show rather weak coefficients from both CMIP3 and CMIP5 within ± 0.1 . A statement on which MME shows better results appears inappropriate given these low values.

For temperature trends, there is high divergence between study areas and seasons. The annual mean values are above 0.3 for both CMIP3 and CMIP5 for Africa and the Globe. The Pacific and Arctic means are below 0.15. In SON, CMIP3 and CMIP5 show positive means for all regions. Furthermore, most regions with high annual mean correlations show good results in SON as well. In JJA the weakest performance of CMIP3 and CMIP5 is found for America. In contrast, the best performance can be identified for the Globe and Africa throughout all seasons. In addition, we find the CMIP5 coefficient mean outperforming the CMIP3 coefficient mean in all seasons and most regions. For example, the Atlantic mean is increased by 0.21 (annual) and 0.22 (DJF).

4.5 Comparison of different metrics

The results so far are solely based on the phi-coefficient. In Fig. 6, six different metrics (see Table 6) based on the same 2×2 table (see Table 5) are compared with each other over all regions and seasons. The results are split into precipitation and temperature for both CMIP3 and CMIP5.

Figure 6 visualizes the rank correlation of the weights of the 50-year trend in the top left part and equivalent of the 50-year mean in the bottom-right part. The diagonal shows the Spearman correlation between the weights for trends and means using the same metric. For 50-year trends we find a very close correlation above 0.95 between the Phi-, Heidke-, Pierce- and Log Odds Ratio-skill scores for both precipitation and temperature. The correlation with the Gilbert metric (GSS) is slightly weaker but still very high with

a correlation of minimally 0.89 for precipitation and 0.93 for temperature with the four metrics mentioned before. However, the results of the Chi2-skill score (CHI) are different. Here the maximum correlation for precipitation is 0.14 and 0.56 for temperature. The reason for this is that CHI is not designed to distinguish between negative and positive coefficients from the 2×2 table because all table entries are squared and accumulated in CHI. This effect is much weaker for temperature because most simulations show a high compliance with the reference data and, hence, positive values prevail in all metrics. As for the other metrics, the actual ranking of models over different regions and seasons is basically the same for five out of four skill scores. The results for 50-year means show for all combinations a spearman correlation in the range from 0.70 to 0.99. Again the highest correlations are found for the combination of Phi, HEI and PIE. CHI now agrees more with the other metrics. The correlations for the same metric between the rankings based on trends and means indicates no connections. Here, all metrics show values within ± 0.12 for both CMIP3 and CMIP5. Therefore, we have to conclude that models with high performance in the simulation of the trend are not equally well performing for the climatological mean. This applies to both precipitation and temperature.

4.6 Impact of weighting on PDFs

Figure 7 illustrates the impact of applied weights on probabilistic projections of precipitation and temperature for CMIP3 and CMIP5 and different emissions scenarios. All weights are based on Phi coefficients from 50-year trends.

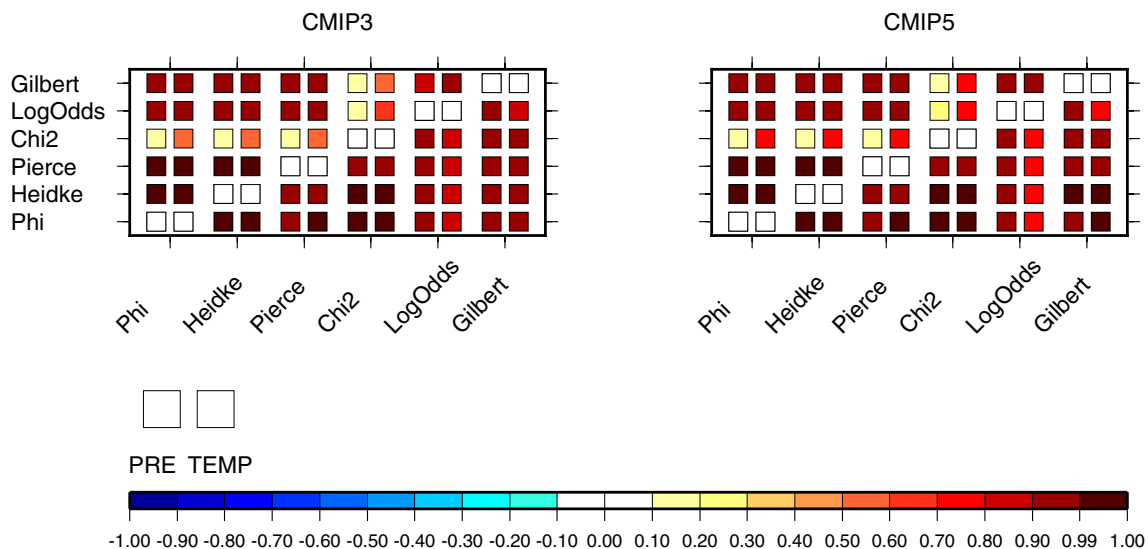


Fig. 6 Spearman correlation between the different skill scores for 50-year trends (*top-left*) and means (*bottom-right*) over all seasons and large study areas for CMIP3 and CMIP5. The *diagonal* shows the

spearman correlation between the weights for trends and means using the same metric

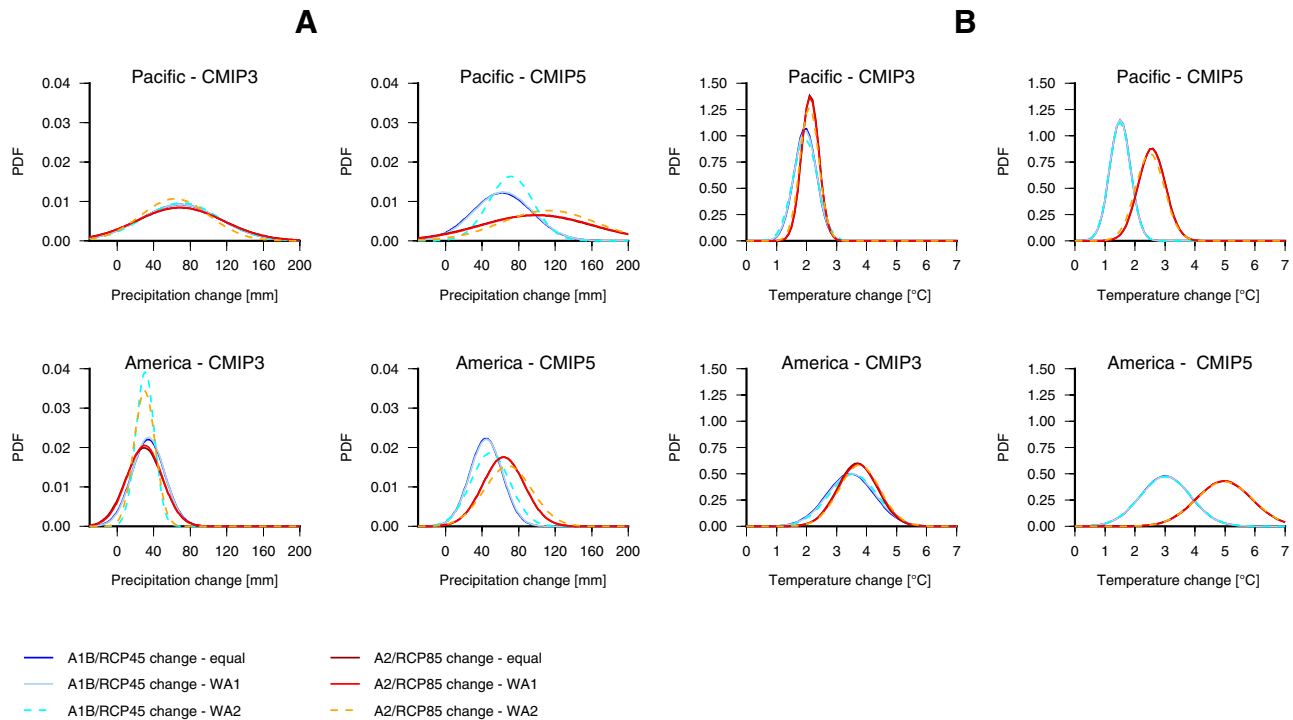


Fig. 7 Impact of the phi-weights on the PDFs of projected changes from the late 20th to the late 21st century for Pacific (*top*) and America (*bottom*). The diagrams for precipitation (A) and temperature (B) include different CMIP3 and CMIP5 emissions scenarios

We show the effects for annual values in the Pacific and America regions. The abscissa displays the changes from the end of the 21st compared to that of the 20th century for either precipitation (A) or temperature (B) while the ordinate shows the PDF values. Please notice the different scales on the ordinates. In the following the spread of PDFs is used as a measure of uncertainty of the projections which is influenced by the used weights. A smaller spread is therefore considered as a reduction while a larger spread indicates an increase of uncertainty. The dark blue (red) lines express the equally weighted changes while the light and dashed blue (red) lines show weighted PDFs. WA1 is based on all models while WA2 only considers models that got a positive phi-correlation while the other models were set to zero weight. Generally, in all regions and seasons there is a considerably higher uncertainty for precipitation than for temperature.

For the Pacific region, both MMEs and all scenarios project a positive change of precipitation and temperature. However, for CMIP3 both scenarios are similar concerning both the width and peak of the PDFs. This is more differentiated for CMIP5. Here, RCP4.5 shows a minor increase of both precipitation and temperature compared to RCP8.5. The WA1 weighted PDFs are almost congruent with the original PDFs. The WA2 weighted PDFs show a slightly decreased spread for CMIP5 and CMIP3 precipitation and an increased spread for CMIP3 temperature. The WA2

weighted PDFs of CMIP5 temperature show no change in standard deviation, however, they are shifted towards lower numbers.

For America, precipitation shows a projected increase with an overall lower spread for both MMEs compared to the Pacific. Here, all unweighted scenarios are rather similar concerning the width of the PDF and the amount of change. The same is true for the WA1 weighted PDFs. Again, the WA2 weighted PDFs show a rather strong decrease of spread for CMIP3. For A1B (A2) the standard deviation of the WA2 PDF is reduced to 54.9% (79.8%) of that of the equally weighted PDF. In contrast, both scenarios in CMIP5 show a slight increase of spread. Temperature in America increases throughout all PDFs. Again, for CMIP3 both scenarios are almost congruent for all PDFs, weighted or not. For CMIP5, the PDFs of RCP4.5 and RCP8.5 are well separated by a mean change of 2 °C. However, both weighting approaches nearly reproduce the unweighted PDF. There is just a minor increase of the standard deviation of the WA2-weighted temperature PDF for RCP8.5.

Other regions and seasons show similarly small differences between the phi-weighted PDFs compared to the equally weighted PDFs. Phi-weighted PDFs (WA1) and the sub-ensemble weighted PDFs (WA2) are basically similar in tendency but WA2 is a more efficient weighting approach. The application of weights based on 50-year

means lead to almost no impact of weighting at all. The underlying coefficients are on a very high and similar level (see Sect. 4.1). This results into rather equal weights for the models with virtually no impact on the PDF shape.

4.7 Transferability to Mediterranean sub-regions

Evaluation metrics and skill scores exhibit a higher applicability and robustness when there are no constraints considering the size or type of study areas. Therefore, we apply the analysis to the eight sub-regions of Medit (see Fig. 1). Figure 8 shows the annual mean phi-coefficients of CMIP3 (black), CMIP5 (gray) and CORDEX (white) simulations for all sub-regions and the entire Medit area. Again, bars represent the results of 50-year trends and circles those of climatological means. Here, there is a wider spread of phi coefficients based on means among the sub-regions compared to the large study areas (see Fig. 5). The minimum is 0.0 for North African annual precipitation from CMIP3. The maximum is 0.85 (CORDEX) for precipitation over Spain. There is a systematic improvement from CMIP3 to CMIP5 and in many regions CORDEX performs best: in five sub-regions and the whole Mediterranean area the RCMs slightly outperform the GCMs.

In general, all MMEs have a higher skill for temperature than for precipitation. For the GCMs, this is consistent with the results of the main study areas (Sect. 4.4). The best results are found for Spain and Medit in CORDEX, both amounting to 0.80. However, the absolute minimum of -0.27 in the Middle East is also given by CORDEX. This area appears to be most challenging for CORDEX simulations since the precipitation skill is also quite small. Both GCM ensembles show rather good results here, with an improvement in CMIP5 compared to CMIP3. In the other sub-regions, CORDEX performs well with highest coefficients in four out of eight study areas.

The coefficients based on 50-year trends are more heterogeneous for temperature than for precipitation. In terms of precipitation, most mean coefficients are rather low

with a maximum in Spain of 0.23 as given by CMIP5. The models' skill seems to be negatively correlated with the annual amount of precipitation, e.g. North Atlantic versus Middle East. There is no apparent improvement from CMIP3 to CMIP5 nor to CORDEX. For temperature, the minimum of -0.04 is in the Black Sea region by CORDEX. The GCMs show rather bad results here as well, even though on a higher level. The maximum of 0.55 in North Africa (CMIP5) is even higher than the maximum of the main study areas (0.45, annual Globe). Further, there is an increase of performance from CMIP3 to CMIP5 in most sub-regions. The strongest increase is found in the North Atlantic region and Italy by more than 0.2. The overall best results are found in Middle East and North Africa, both arid regions with high mean temperature. The results of the other seasons (not shown) show a similar behavior. For the temperature trend, CORDEX has mostly values below both CMIP3 and CMIP5. Here, the reason is the shorter data availability of CORDEX starting 1970 (for an evaluation period of 1960–2009). For an evaluation period of 1970–2009 CORDEX shows stronger results than CMIP3 and CMIP5 in most situations. However, note that a 40-year period (1970–2009) might be severely influenced by natural climate variability in this region which is not captured in GCMs (Paxian et al. 2013).

Figure 9 depicts the spearman correlation of the annual and seasonal weights based on 50-year trends between the sub-regions and the whole Mediterranean region over all simulations including CORDEX. Again, there are mainly low values of correlations for both precipitation (top-left) and temperature (bottom-right). However, they are mostly positive with slightly higher values compared to those in the global regions in most seasons. The maximum value of 0.62 is for precipitation in MAM between Italy and Medit. The minimum of -0.22 is for annual precipitation between Aegean and Spain. Overall, 131 of 180 (72.8%) of correlations of precipitation weights are positive. For temperature there are even 153 (85.0%) positive correlations. The mean over all correlations is 0.09 for precipitation and 0.16 for

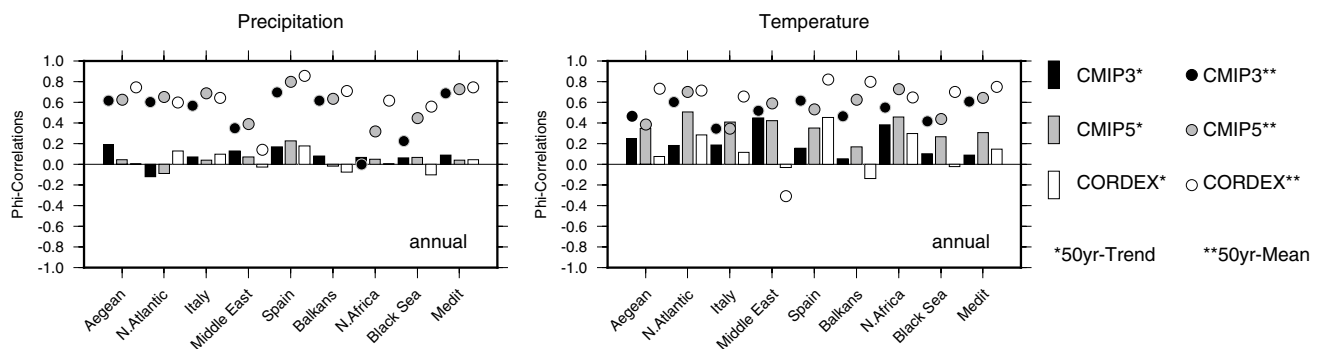


Fig. 8 Comparison of mean phi-coefficients from CMIP3, CMIP5 and CORDEX for the sub-regions of Medit area

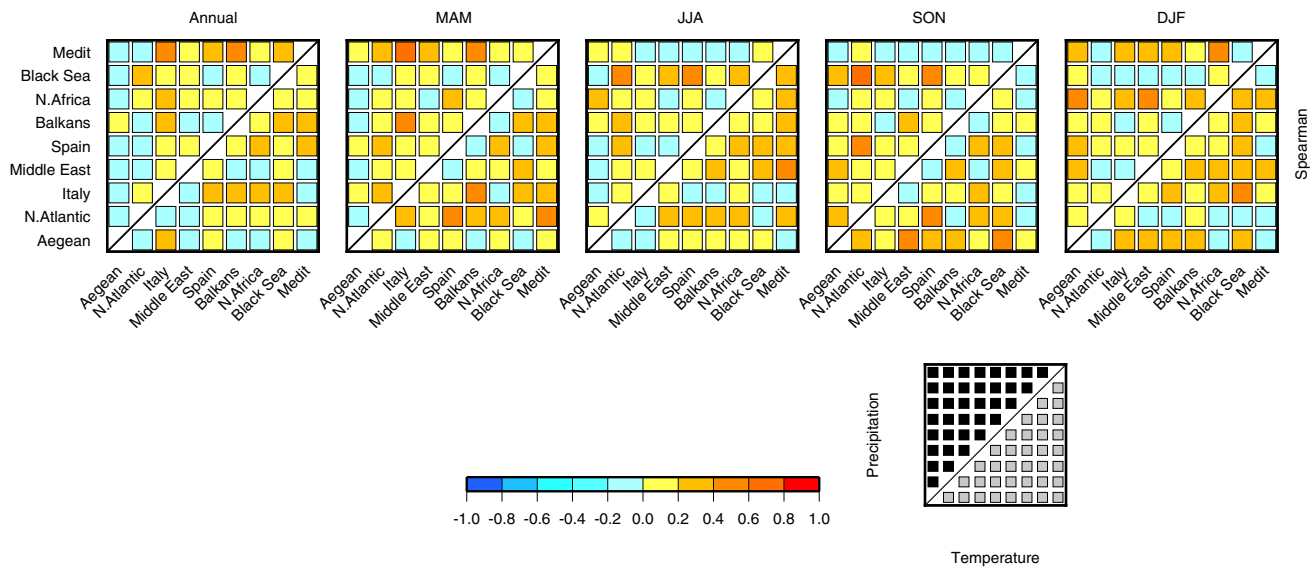


Fig. 9 Spearman correlation of phi-coefficients between the sub-regions of Medit for each season and for precipitation (*top-left*) and temperature (*bottom-right*) 50-year trends

temperature. To sum up, as for the results of the climatological means, the correlation of coefficients between different regions is on the same level as for the main study areas or even slightly higher. The spearman correlations between the regions of the 50-year mean (not shown) are similar to those of the 50-year trend. There are neither stronger correlations nor a noticeable change in the amount of positive values.

4.8 Sensitivity to the reference data

In contrast to observational data such as CRU or E-OBS, ERA-20C covers the entire globe without missing values enabling the analysis of important study areas as Global or Pacific. However, the differences between reference datasets might be a relevant source of uncertainty in model evaluation. Figure 10 shows the 50-year trend for DJF precipitation (left) and temperature (right) for the Mediterranean area for three different datasets. For precipitation, there is strong decrease in most parts of the area. ERA-20C has its maximum decrease over Italy and Portugal. CRU and E-OBS show a similar pattern. However, for E-OBS the maximum over Portugal is much stronger while another maximum is apparent over eastern Turkey. For CRU and E-OBS, there is also a minor increase of precipitation in the North- and South-Eastern parts of the Mediterranean. All in all, there are differences between the datasets but the main tendency of the trend is the same in most cases and for all seasons. Similar results are shown for temperature. Here, Fig. 10 shows for all datasets a rather strong increase in the North-Western part of the area. A slight decrease is

found in the Western part of the study area. This is of larger extent in CRU and E-OBS compared to ERA-20C. Again, the regional distribution and tendency is on a homogeneous level for all datasets and seasons. The accordance of the climatological mean of precipitation and temperature is even on a higher level as for the trend, shown in Fig. 10. For further details, Table 7 shows the Pearson correlation (von Storch and Zwiers 1999) for the Mediterranean area. The pattern of annual precipitation and temperature show positive results for both trend and climatological mean between all evaluation datasets. Best results are found for mean temperature with a maximum correlation of 0.96 for ERA-20C/CRU and CRU/E-OBS. Precipitation results are on a similarly high level with 0.93 (CRU/E-OBS) and 0.90 (ERA20-C/CRU). The trend correlations are between 0.24 (ERA-20C/E-OBS) and 0.56 (CRU/E-OBS). Overall, the correlations show positive and often high correlation for all seasons. Note that for the trend we find noticeable discrepancies between the evaluation datasets. We want to underline that the results of the applied performance metrics are dependent on reliable evaluation data and observational uncertainty must be considered for any further conclusions. Nevertheless, as presented in Fig. 10 and Table 7, we found high consistency between all datasets for the mean and low to mid correlations of the trend for the Mediterranean area. Considering the different characteristics of the study areas and models, ERA-20C therefore turned out to be most suitable as the reference data for this study.

Further, aside from discrepancies between the evaluation datasets, we wanted to assess how climate models perform with respect to different validation data. Therefore, Fig. 11

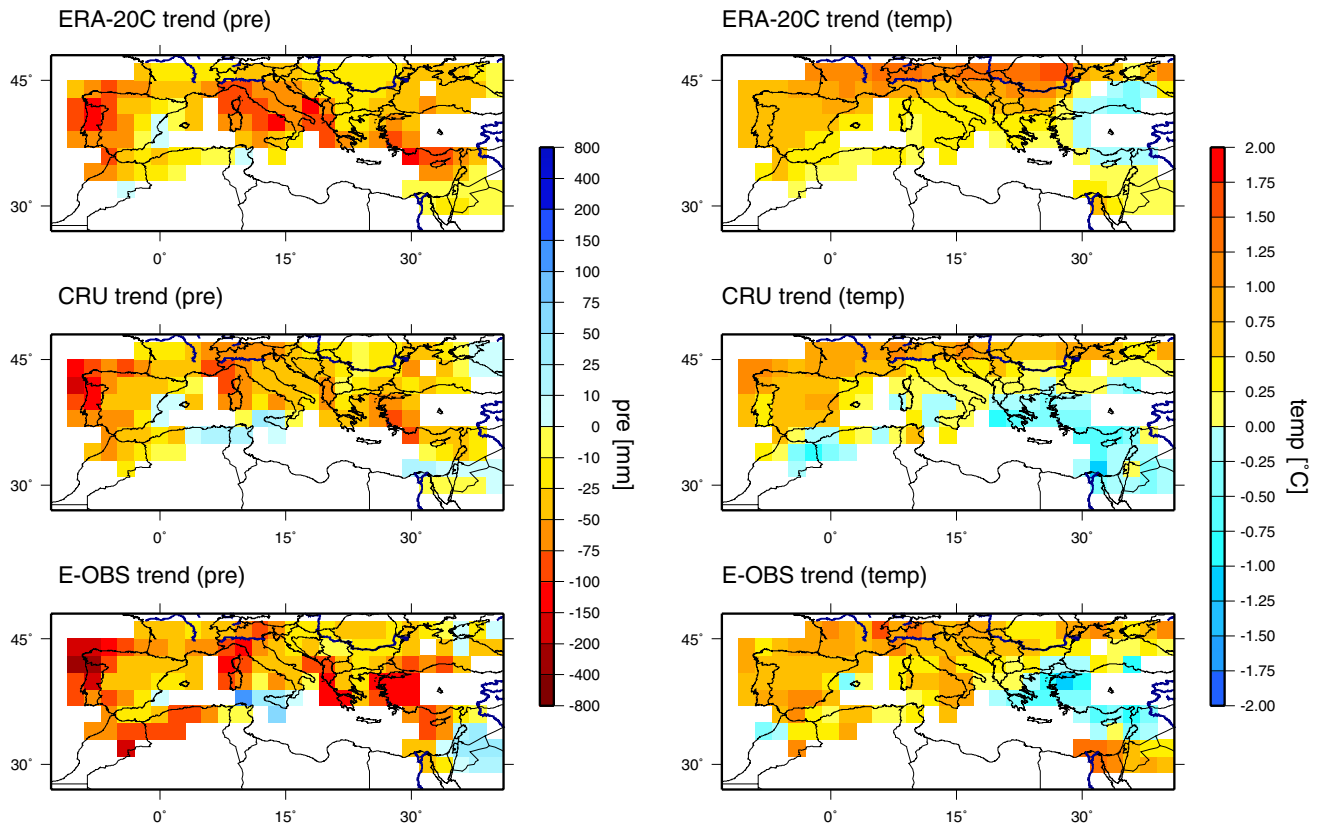


Fig. 10 Comparison of used evaluation datasets. Trend over the period 1960–2009 of precipitation (*left*) and temperature (*right*) for DJF

Table 7 Pearson correlation of the mean and trend pattern (1960–2009) between different evaluation datasets for annual precipitation and temperature for the Mediterranean area

Datasets	Precipitation		Temperature	
	Trend	Mean	Trend	Mean
ERA-20C/CRU	0.32	0.90	0.49	0.96
ERA-20C/E-OBS	0.24	0.82	0.29	0.91
CRU/E-OBS	0.56	0.93	0.36	0.96

illustrates the robustness of phi-skill scores across three different validation datasets of precipitation and temperature for Medit and its sub-regions. We calculated the spearman correlation of phi-coefficients for each sub-region and season between each combination of ERA-20C, CRU and E-OBS. In Fig. 11, the circles refer to 50-year means while the squares stand for 50-year trends. For annual, DJF and JJA precipitation and temperature similar patterns are found: the majority of study areas show medium to high correlations between the weights based on different reference datasets. Overall, the metrics appear to be quite robust for 50-year trends and means. There is a large range of correlations from 0.93 to -0.59 , but in most cases rather high

correlations occur, especially between ERA-20C and CRU. Most values below zero relate to the E-OBS data set. Since E-OBS is considered to be a best-estimate regional dataset this has to be interpreted as a deficiency of CRU or ERA-20C in certain sub-regions, respectively. However, the total amount of negative correlation is low throughout all seasons and regions. Especially for the whole Medit area we find almost all correlations to be positive. This is in line with previous results (Table 7).

5 Discussion

Six skill scores based on the same 2×2 contingency table have been analyzed in this study. They mostly agree with each other, except for the Chi2 approach. Here, the equal treatment of very high and very low coefficients leads to a systematically different ranking of models. Based on our results, the other five metrics are more or less exchangeable. The Log Odds ratio is preferred in meteorological and medical science (Stephenson 2000; Thornes and Stephenson 2001; Paeth et al. 2006). However, we recommend one of the other four, since the Log Odds ratio requires five values or more for each field of the 2×2 table (Stephenson

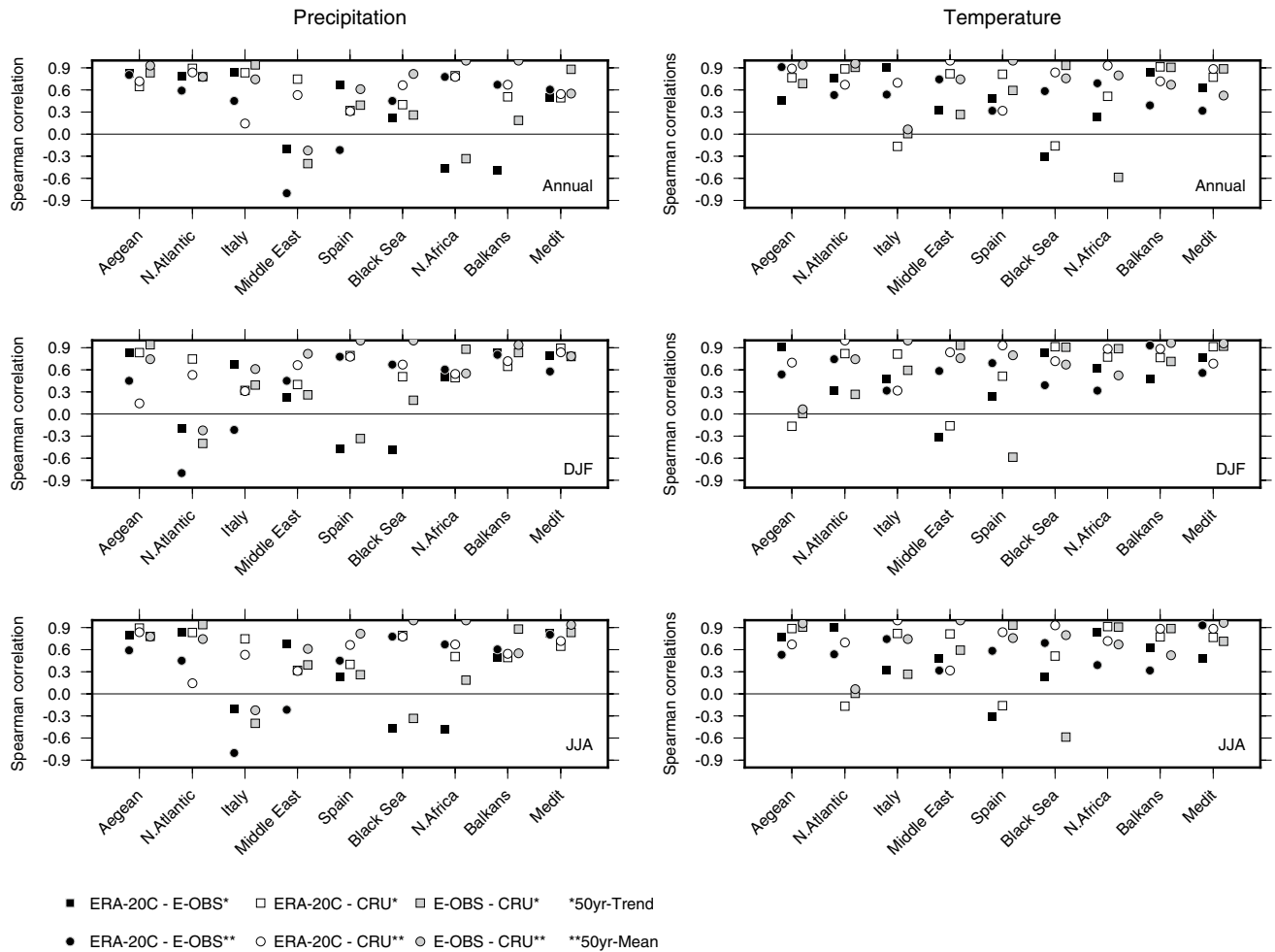


Fig. 11 Spearman correlation of annual, DJF and JJA phi-skill scores between different sets of reference data (ERA-20C, E-OBS, CRU) for each of the sub-regions of Medit and for precipitation (*left*) and temperature (*right*)

2000; Paeth et al. 2006). Zero values might occur in small regions with a 100% accordance between simulation and evaluation data. Here, we have to manually add a small value (e.g. 0.1) to each field to avoid a mathematical error (Gart and Zweifel 1967; Pettigrew et al. 1986). No error or restriction is given for the other metrics. An advantage is that the 2×2 metrics are easily transferable to a wide range of different variables and regions. We analyzed 50-year trends and climatological means over several regions of different sizes. Another important potential of these metrics is that they can easily be extended to the analysis of extremes simply by varying the threshold value in the 2×2 contingency table (i.e. Paeth et al. 2006; Liu et al. 2013). Note though, that Armistead (2013) criticized these categorical metrics for their inability to assess the separate accuracy rates and their sensitivity to the interdependence of different datasets in a 2×2 table. Further, the inapplicability of many metrics to be applied on $k \times k$ ($k \geq 2$) tables is pointed out. However, the mentioned limitations do not effect our

approach. Further, our metrics show robust results and appear flexible and suitable for evaluation of any kind of climate model.

Some systematic enhancement of model performance from CMIP3 to CMIP5 was found for both 50-year temperature trends and means. This is in line with Wright et al. (2016) and Koutroulis et al. (2016). The improvement is less obvious for precipitation as also reported by Li and Xie (2014) and Grose et al. (2014). Precipitation means are characterized by higher skill scores than the trend. This is in line with Perkins et al. (2007), pointing to high modeling quality for temperature over Australia but weaker results for precipitation for CMIP3. Also Huang et al. (2013) found high uncertainty for JJA precipitation in Eastern China.

The rather low regional correlation between the rankings of models based on 50-year trends indicate that there is no model capable of performing equally well in every region. This is confirmed by Power et al. (2012) and Ring et al. (2016) who find rather large discrepancies between model

performances of different regions. However, we saw higher accordance for coefficients of the mean between the major study areas in contrast to the sub-regions. Overall, these results indicate a high conformity with other accepted evaluation metrics (e.g. Perkins et al. 2007; Huang et al. 2013).

Furthermore, the evaluation of study areas at different scales suggests no loss of usability. For the Medit sub-regions, we found higher weights for temperature while the precipitation results remained the same compared to the large study areas. We consider this result to be promising for the applicability of the metrics at smaller scales, being important for aspects of climate impact research. CORDEX simulations were evaluated as well for the sub-regions, since RCMs are expected to provide an added value (Jacob et al. 2014). However, as mentioned before, the first 10 years of our investigation period are not included in the regional simulations. Despite this limitation, the RCMs outperform the GCMs in terms of the climatological mean in most Mediterranean sub-regions (cf. Rummukainen 2010). Yet, CORDEX shows an inferior simulation quality for the temperature trend compared to CMIP3 and CMIP5, while the precipitation trend performance is on a similar level as the one from global climate models.

The unweighted probabilistic multi-model projections show a clear temperature increase of about 2–7 °C, depending on the scenario, for all 14 main or sub-regions with a rather low uncertainty. Miao et al. (2014) assessed a similar range for Northern Eurasia. The MME changes for precipitation are less coherent. Here, we found both positive and negative changes of precipitation amounts. Especially in arid regions like the Mediterranean and most of its sub-regions (i.e. Middle East or Spain) nearly all models simulate a decrease of precipitation. The opposite is true in the Arctic, Pacific and global mean where most simulations indicate an increase of precipitation. Overall, the uncertainty of precipitation is much higher than the one of temperature. These results are in line with most studies published on this topic (e.g. Randall et al. 2007; Flato et al. 2013).

The utility of the 2×2 metrics for improving probabilistic climate projections has been found to be less obvious in most situations (regions, seasons, scenarios). A weighting approach on basis of 50-year means is inadvisable, at least for the large study areas. Here, simulations showed an equally high performance which led to equal weight, while for the 50-year trend we found several rather strong differences in simulation quality. However, our results indicate no homogeneous effect on the PDFs. Both increased and decreased PDF spreads become apparent. Mostly the effects are minor, especially when the WA1 weighting approach is used. For many situations simulations are either performing very well (temperature) or rather weak (precipitation) but on a similar level (see Figs. 2, 3). To intensify the weights,

we also applied a modified metric (WA2) where all models with negative (i.e. ϕ -) coefficients had been ignored. This was found to be a more effective approach, since the general tendencies of the PDFs (increase or decrease) remained the same, for almost all situations, while the impact on the change of uncertainty was enhanced. Of course, changing the impact of weights by selecting a sub-ensemble (here models with positive metric-coefficients) requires sophisticated metrics of model evaluation and well-agreed standards on thresholds.

6 Conclusion

In this study, we applied a simple 2×2 table approach with six different skill scores in order to evaluate state-of-the-art global and regional climate models. Seven regions of large extend and eight smaller sub-regions were tested in terms of 50-year trends and means of annual and seasonal precipitation and temperature. Overall, five of six metrics are quite consistent with each other. These five metrics are equally adequate to determine the ranking of climate models. Our study revealed a considerable improvement of model performance of CMIP5 over CMIP3 models for temperature for both trend and mean for the majority of analyzed seasons and regions. This is in line with Wright et al. (2016). The precipitation trend turned out to be rather challenging for most models of both model generations. However, a main reason for this may be the lack of a pronounced trend of precipitation in the reference data compared to the general increase of temperature (Kumar et al. 2013). The means of both precipitation and temperature were matched by the majority of simulations with a slight improvement from CMIP3 to CMIP5. In the Mediterranean sub-regions, the CMIP5 was still outperformed by CORDEX, even though a shorter timeframe had to be analyzed.

In terms of probabilistic climate projections, we faced the problem that either the majority of models performed equally strong (especially for temperature) or weak (for precipitation). This led to minor effects on the shape and position of PDFs of climate change, when all model contributed to the probabilistic predictions. Here, the Log Odds ratio showed a slightly stronger differentiation of weights in some simulations. Therefore, we applied an additional more stringent sub-ensemble approach. This led to stronger selection, while avoiding an artificial manipulation of the original coefficients. However, since temperature showed high values for most models, the sub-ensemble approach had less impact on temperature than on precipitation. Indeed, for precipitation we could develop a powerful weighting metric. Nevertheless, there is no consistent change in uncertainty over sub-ensembles

for both temperature and precipitation. Instead, we found a high dependence on the variable, region and season.

Altogether, the PDFs showed an expected increase of temperature of 2–7 °C range towards the end of the 21st century with a relatively low range of uncertainty. For precipitation, in most situations the range was centered around 0 but with a rather high range of uncertainty. Thus, the simulation of temperature is at a very high quality level which makes future projections trustworthy and the effect of different emissions scenarios more apparent. In terms of precipitation, only some models reproduce the observed characteristics, yet not systematically in all regions and seasons. However, sound projection of future climate change are required to elaborate appropriate adaptation strategies. For this purpose, further studies on weighting metrics are necessary to improve the quality of probabilistic climate projections. Because of their good transferability and obvious interpretation, the 2×2 table metrics might help as a basis for comparison with other metrics in this context. Another potential to be dealt with by further investigation pertains to the aspect of extreme events which can easily be tackled by varying the thresholds in the 2×2 contingency table.

Acknowledgements We thank the Program for Climate Model Diagnosis and Intercomparison (PCMDI) and the World Climate Research Programme (WCRP) for providing the CMIP3, CMIP5 and CORDEX datasets used in this study. Furthermore, we are grateful for the provided observational data and reanalyses by the Climate Research Unit (CRU), the EU-FP6 project ENSEMBLES, the data providers in the ECA&D project and the European Centre for Medium-Range Weather Forecasts (ECMWF), respectively. This study was conducted within the COMEPRO-Project funded by the Deutsche Forschungsgemeinschaft (DFG).

References

- Armistead TW (2013) H. L. Wagner's unbiased hit rate and the assessment of categorical forecasting accuracy. *Weather Forecast* 28:802–814. doi:[10.1175/WAF-D-12-00047.1](https://doi.org/10.1175/WAF-D-12-00047.1)
- Ayar PV, Vrac M, Bastin S, Carreau J, Déqué M, Gallardo C (2016) Intercomparison of statistical and dynamical downscaling models under the EURO- and MED-CORDEX initiative framework: present climate evaluations. *Clim Dyn* 46:1301–1329. doi:[10.1007/s00382-015-2647-5](https://doi.org/10.1007/s00382-015-2647-5)
- Babak O, Deutsch CV (2009) Statistical approach to inverse distance interpolation. *Stoch Environ Res Risk Assess* 23:543–553. doi:[10.1007/s00477-008-0226-6](https://doi.org/10.1007/s00477-008-0226-6)
- Bishop CH, Abramowitz G (2013) Climate model dependence and the replicate Earth paradigm. *Clim Dyn* 41:885–900. doi:[10.1007/s00382-012-1610-y](https://doi.org/10.1007/s00382-012-1610-y)
- Bortz J, Lienert GA, Boehnke K (2008) *Verteilungsfreie Methoden in der Biostatistik*, 3rd edn. Springer Berlin Heidelberg, Berlin, Heidelberg
- Clark MP, Wilby RL, Gutmann ED et al (2016) Characterizing uncertainty of the hydrologic impacts of climate change. *Curr Clim Change Rep* 2:55–64. doi:[10.1007/s40641-016-0034-x](https://doi.org/10.1007/s40641-016-0034-x)
- Diffenbaugh NS, Giorgi F (2012) Climate change hotspots in the CMIP5 global climate model ensemble. *Clim Change* 114:813–822. doi:[10.1007/s10584-012-0570-x](https://doi.org/10.1007/s10584-012-0570-x)
- Dittus AJ, Karoly DJ, Lewis SC et al (2016) A multiregion model evaluation and attribution study of historical changes in the area affected by temperature and precipitation extremes. *J Clim* 29:8285–8299. doi:[10.1175/JCLI-D-16-0164.1](https://doi.org/10.1175/JCLI-D-16-0164.1)
- Donat MG, Alexander L V., Herold N, Dittus AJ (2016) Temperature and precipitation extremes in century-long gridded observations, reanalyses, and atmospheric model simulations. *J Geophys Res Atmos* 121:11,174–11,189. doi:[10.1002/2016JD025480](https://doi.org/10.1002/2016JD025480)
- Done J, Davis CA, Weisman M (2004) The next generation of NWP: explicit forecasts of convection using the weather research and forecasting (WRF) model. *Atmos Sci Lett* 5:110–117. doi:[10.1002/asl.72](https://doi.org/10.1002/asl.72)
- Doolittle MH (1885) The verification of predictions. *Amer Meteor* J 2:327–329
- Doolittle MH (1888) Association ratios. *Bull Philos Soc Wash* 10:83–96
- Eum H-I, Gachon P, Laprise R (2014) Developing a likely climate scenario from multiple regional climate model simulations with an optimal weighting factor. *Clim Dyn* 43:11–35. doi:[10.1007/s00382-013-2021-4](https://doi.org/10.1007/s00382-013-2021-4)
- Flato G, Marotzke J, Abiodun B et al (2013) Evaluation of climate models. In: *Climate change 2013: the physical science basis. Contribution of working group I to the fifth assessment report of the intergovernmental panel on climate change*. Cambridge University Press, Cambridge, p 866
- Gart JJ, Zweifel JR (1967) On the bias of various estimators of the logit and its variance with application to quantal bioassay. *Biometrika* 54:181. doi:[10.2307/2333861](https://doi.org/10.2307/2333861)
- Ghelli A, Primo C (2009) On the use of the extreme dependency score to investigate the performance of an NWP model for rare events. *Meteorol Appl* 16:537–544. doi:[10.1002/met.153](https://doi.org/10.1002/met.153)
- Gilbert GF (1884) Finley's tornado predictions. *Amer Meteor* J 1:166–172
- Gill PG, Buchanan P (2014) An ensemble based turbulence forecasting system. *Meteorol Appl* 21:12–19. doi:[10.1002/met.1373](https://doi.org/10.1002/met.1373)
- Gillett NP, Annan J, Hargreaves J et al (2015) Weighting climate model projections using observational constraints. *Philos Trans A Math Phys Eng Sci* 373:L02703. doi:[10.1098/rsta.2014.0425](https://doi.org/10.1098/rsta.2014.0425)
- Giorgi F (2006) Climate change hot-spots. *Geophys Res Lett* 33:L08707. doi:[10.1029/2006GL025734](https://doi.org/10.1029/2006GL025734)
- Giorgi F, Lionello P (2008) Climate change projections for the Mediterranean region. *Glob Planet Change* 63:90–104. doi:[10.1016/j.gloplacha.2007.09.005](https://doi.org/10.1016/j.gloplacha.2007.09.005)
- Giorgi F, Jones C, Asrar G (2009) Addressing climate information needs at the regional level: the CORDEX framework. *WMO Bull* 58:175–183
- Gleckler PJ, Taylor KE, Doutriaux C (2008) Performance metrics for climate models. *J Geophys Res* 113:D06104. doi:[10.1029/2007JD008972](https://doi.org/10.1029/2007JD008972)
- Große MR, Brown JN, Narsey S, Brown JR, Murphy BF, Langlais C, Gupta AS, Moise AF, Irving DB (2014) Assessment of the CMIP5 global climate model simulations of the western tropical Pacific climate system and comparison to CMIP3. *Int J Climatol* 34 (12):3382–3399
- Haughton N, Abramowitz G, Pitman A, Phipps SJ (2015) Weighting climate model ensembles for mean and variance estimates. *Clim Dyn* 45:3169–3181. doi:[10.1007/s00382-015-2531-3](https://doi.org/10.1007/s00382-015-2531-3)
- Hawkins E, Smith RS, Gregory JM, Stainforth DA (2016) Irreducible uncertainty in near-term climate projections. *Clim Dyn* 46:3807–3819. doi:[10.1007/s00382-015-2806-8](https://doi.org/10.1007/s00382-015-2806-8)
- Haylock MR, Hofstra N, Klein Tank AMG et al (2008) A European daily high-resolution gridded data set of surface temperature

- and precipitation for 1950–2006. *J Geophys Res* 113:D20119. doi:[10.1029/2008JD010201](https://doi.org/10.1029/2008JD010201)
- Heidke P (1926) Berechnung des Erfolges und der Güte der Windstärkavorhersagen im Sturmwarnungsdienst (Calculation of the success and goodness of strong wind forecasts in the storm warning service). *Geogr Ann Stockholm* 8:301–349
- Hewitson BC, Crane RG (2006) Consensus between GCM climate change projections with empirical downscaling: precipitation downscaling over South Africa. *Int J Climatol* 26:1315–1337. doi:[10.1002/joc.1314](https://doi.org/10.1002/joc.1314)
- 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. *J Geophys Res Atmos* 118:9035–9047. doi:[10.1002/jgrd.50695](https://doi.org/10.1002/jgrd.50695)
- Jacob D, Petersen J, Eggert B et al (2014) EURO-CORDEX: new high-resolution climate change projections for European impact research. *Reg Environ Chang* 14:563–578. doi:[10.1007/s10113-013-0499-2](https://doi.org/10.1007/s10113-013-0499-2)
- Knutti R (2010) The end of model democracy? *Clim Change* 102:395–404. doi:[10.1007/s10584-010-9800-2](https://doi.org/10.1007/s10584-010-9800-2)
- Knutti R, Sedláček J (2012) Robustness and uncertainties in the new CMIP5 climate model projections. *Nat Clim Chang* 3:369–373. doi:[10.1038/nclimate1716](https://doi.org/10.1038/nclimate1716)
- Knutti R, Furrer R, Tebaldi C et al (2010) Challenges in combining projections from multiple climate models. *J Clim* 23:2739–2758. doi:[10.1175/2009JCLI3361.1](https://doi.org/10.1175/2009JCLI3361.1)
- Knutti R, Masson D, Gertzelman A (2013) Climate model genealogy: generation CMIP5 and how we got there. *Geophys Res Lett* 40:1194–1199. doi:[10.1002/grl.50256](https://doi.org/10.1002/grl.50256)
- Koutroulis AG, Grillakis MG, Tsanis IK, Papadimitriou L (2016) Evaluation of precipitation and temperature simulation performance of the CMIP3 and CMIP5 historical experiments. *Clim Dyn* 47:1881–1898. doi:[10.1007/s00382-015-2938-x](https://doi.org/10.1007/s00382-015-2938-x)
- Kumar S, Merwade V, Kinter JL et al (2013) Evaluation of temperature and precipitation trends and long-term persistence in CMIP5 twentieth-century climate simulations. *J Clim* 26:4168–4185. doi:[10.1175/JCLI-D-12-00259.1](https://doi.org/10.1175/JCLI-D-12-00259.1)
- Li G, Xie S-P (2014) Tropical biases in CMIP5 multimodel ensemble: the excessive equatorial Pacific cold tongue and double ITCZ problems*. *J Clim* 27(4):1765–1780
- Liu B, Chen J, Chen X et al (2013) Uncertainty in determining extreme precipitation thresholds. *J Hydrol* 503:233–245. doi:[10.1016/j.jhydrol.2013.09.002](https://doi.org/10.1016/j.jhydrol.2013.09.002)
- Miao C, Duan Q, Sun Q et al (2014) Assessment of CMIP5 climate models and projected temperature changes over Northern Eurasia. *Environ Res Lett* 9:55007. doi:[10.1088/1748-9326/9/5/055007](https://doi.org/10.1088/1748-9326/9/5/055007)
- Mitchell TD, Jones PD (2005) An improved method of constructing a database of monthly climate observations and associated high-resolution grids. *Int J Climatol* 25(6):693–712
- Moss RH, Edmonds JA, Hibbard KA et al (2010) The next generation of scenarios for climate change research and assessment. *Nature* 463:747–756. doi:[10.1038/nature08823](https://doi.org/10.1038/nature08823)
- Nakicenovic N, Alcamo J, Davis G et al (2000) Special report on emissions scenarios†: a special report of Working Group III of the Intergovernmental Panel on Climate Change. Cambridge University Press, USA
- Paeth H, Girmes R, Menz G, Hense A (2006) Improving seasonal forecasting in the low latitudes. *Mon Weather Rev* 134:1859–1879. doi:[10.1175/MWR3149.1](https://doi.org/10.1175/MWR3149.1)
- Paeth H, Vogt G, Paxian A et al (2016) Quantifying the evidence of climate change in the light of uncertainty exemplified by the Mediterranean hot spot region. *Glob Planet Change*. doi:[10.1016/j.gloplacha.2016.03.003](https://doi.org/10.1016/j.gloplacha.2016.03.003)
- Paxian A, Hertig E, Vogt G, Seubert S, Jacobeit J, Paeth H. (2013) Greenhouse gas-related predictability of regional climate model trends in the Mediterranean area. *Int J Climatol* 34:2293–2307. doi:[10.1002/joc.3838](https://doi.org/10.1002/joc.3838)
- Perkins SE, Pitman AJ, Holbrook NJ et al (2007) Evaluation of the AR4 climate models' simulated daily maximum temperature, minimum temperature, and precipitation over Australia using probability density functions. *J Clim* 20:4356–4376. doi:[10.1175/JCLI4253.1](https://doi.org/10.1175/JCLI4253.1)
- Pettigrew HM, Gart JJ, Thomas DG (1986) The bias and higher cumulants of the logarithm of a binomial variate. *Biometrika* 73:425–435. doi:[10.1093/biomet/73.2.425](https://doi.org/10.1093/biomet/73.2.425)
- Pielke RA, Wilby RL (2012) Regional climate downscaling: what's the point? *Eos Trans Am Geophys Union* 93:52–53. doi:[10.1029/2012EO050008](https://doi.org/10.1029/2012EO050008)
- Pierce CS (1884) The numerical measure of the success of predictions. *Science* 4(93):453–454
- Poli P, Hersbach H, Dee DP et al (2016) ERA-20C: an atmospheric reanalysis of the twentieth century. *J Clim* 29:4083–4097. doi:[10.1175/JCLI-D-15-0556.1](https://doi.org/10.1175/JCLI-D-15-0556.1)
- Power SB, Delage F, Colman R, Moise A (2012) Consensus on twenty-first-century rainfall projections in climate models more widespread than previously thought. *J Clim* 25:3792–3809. doi:[10.1175/JCLI-D-11-00354.1](https://doi.org/10.1175/JCLI-D-11-00354.1)
- Randall DA, Bony RA, S. W et al (2007) Climate models and their evaluation. In: Solomon S, Qin D, Manning M et al (eds) *Climate change 2007: the physical science basis. contribution of working group I to the fourth assessment report of the intergovernmental panel on climate change*. Cambridge University Press, Cambridge, p 662
- Reichler T, Kim J (2008) How well do coupled models simulate today's climate? *Bull Am Meteorol Soc* 89:303–311. doi:[10.1175/BAMS-89-3-303](https://doi.org/10.1175/BAMS-89-3-303)
- Ring C, Mannig B, Pollinger F, Paeth H (2016) Uncertainties in the simulation of precipitation in selected regions of humid and dry climate. *Int J Climatol* 36:3521–3538. doi:[10.1002/joc.4573](https://doi.org/10.1002/joc.4573)
- Rummukainen M (2010) State-of-the-art with regional climate models. *Wiley Interdiscip Rev Clim Chang* 1:82–96. doi:[10.1002/wcc.8](https://doi.org/10.1002/wcc.8)
- Sanderson BM, Knutti R, Caldwell P et al (2015) A representative democracy to reduce interdependency in a multimodel ensemble. *J Clim* 28:5171–5194. doi:[10.1175/JCLI-D-14-00362.1](https://doi.org/10.1175/JCLI-D-14-00362.1)
- Schulzweida U, Kornbluh L, Quast R (2009) CDO User's Guide. Climate data operators. Version 1.4.1. In: MPI Meteorol. <https://www.rsmas.miami.edu/users/rajib/cdo.pdf>. Accessed 28 Dec 2016
- Sheffield J, Barrett AP, Colle B et al (2013) North American Climate in CMIP5 experiments. Part I: evaluation of historical simulations of continental and regional climatology. *J Clim* 26:9209–9245
- Stephenson DB (2000) Use of the “odds ratio” for diagnosing forecast skill. *Weather Forecast* 15:221–232. doi:[10.1175/1520-0434\(2000\)015<0221:UOTORF>2.0.CO;2](https://doi.org/10.1175/1520-0434(2000)015<0221:UOTORF>2.0.CO;2)
- Tebaldi C, Knutti R, Allen MR et al (2007) The use of the multimodel ensemble in probabilistic climate projections. *Philos Trans R Soc A Math Phys Eng Sci* 365:2053–2075. doi:[10.1098/rsta.2007.2076](https://doi.org/10.1098/rsta.2007.2076)
- Thornes JE, Stephenson DB (2001) How to judge the quality and value of weather forecast products. *Meteorol Appl* 8:S1350482701003061. doi:[10.1017/S1350482701003061](https://doi.org/10.1017/S1350482701003061)
- Von Storch H, Zwiers FW (1999) *Statistical analysis in climate research*, 1st edn. Cambridge University Press, Cambridge
- Wilkinson JM (2017) A technique for verification of convection-permitting NWP model deterministic forecasts of lightning activity. *Weather Forecast* 32:97–115. doi:[10.1175/WAF-D-16-0106.1](https://doi.org/10.1175/WAF-D-16-0106.1)
- Wilks DS (2006) *Statistical methods in the atmospheric sciences*, 2. Elsevier, Amsterdam

- Woodcock F (1976) The evaluation of yes/no forecasts for scientific and administrative purposes. *Mon Weather Rev* 104:1209–1214. doi:[10.1175/1520-0493\(1976\)104<1209:TEOYFF>2.0.CO;2](https://doi.org/10.1175/1520-0493(1976)104<1209:TEOYFF>2.0.CO;2)
- Wright AN, Schwartz MW, Hijmans RJ, Bradley Shaffer H (2016) Advances in climate models from CMIP3 to CMIP5 do not change predictions of future habitat suitability for California reptiles and amphibians. *Clim Change* 134:579–591. doi:[10.1007/s10584-015-1552-6](https://doi.org/10.1007/s10584-015-1552-6)