EFFECT OF DIVERSIFIED PERFORMANCE METRICS AND CLIMATE MODEL WEIGHTING ON GLOBAL AND REGIONAL TREND PATTERNS OF PRECIPITATION AND TEMPERATURE

Christoph Ring, Felix Pollinger, Luzia Keupp, Irena Kaspar-Ott, Elke Hertig, Jucundus Jacobeit and Heiko Paeth

With 9 figures and 5 tables
Received 17 August 2019 · Accepted 7 November 2019

Summary: A main task of climate research is to provide estimates about future climate change under global warming conditions. The main tools for this are dynamic climate models. However, different models vary quantitatively - and in some aspects even qualitatively - in the climate change signals they produce. In this study, this uncertainty about future climate is tackled by the evaluation of climate models in a standardized setup of multiple regions and variables based on four sophisticated metrics. Weighting models based on their performance will help to increase the confidence in climate model projections. Global and regional climate models are evaluated for 50-year trends of simulated seasonal precipitation and temperature. The results of these evaluations are compared, and their impact on probabilistic projections of precipitation and temperature when used as bases of weighting factors is analyzed. This study is performed on two spatial scales: seven globally distributed large study areas and eight sub-regions of the Mediterranean area. Altogether, over 62 global climate models with 159 transient simulations for precipitation and 119 for temperature from four emissions scenarios are evaluated against the ERA-20C reanalysis. The results indicate large agreement between three out of four metrics. The fourth one addresses a new climate model characteristic that shows no correlation to any other ranking. Overall, especially temperature shows a high agreement to the reference data set while precipitation offers better potential for weighting. Because of the differences being rather small, the metrics are better suited for performance rankings than as weighting factors. Finally, there is conformity with previous model evaluation studies: both the model performance and the implications of weighting for probabilistic climate projections strictly depend on the selected region, season and variable. Thus, none of the climate models generally outperforms all others.

Zusammenfassung: Eine Hauptaufgabe der Klimaforschung ist die Bereitstellung zuverlässiger Abschätzungen bezüglich des zukünftigen Klimawandels im Zuge der Globalen Erwärmung. Die wichtigsten Werkzeuge hierfür sind dynamische Klimamodelle. Jedoch erzeugen unterschiedliche Modelle quantitativ – und in einigen Aspekten sogar qualitativ – verschiedene Klimaänderungssignale. In dieser Studie wird diese Unsicherheit über das künftige Klima angegangen. Hierfür werden Klimamodelle in einem standardisierten Versuchsaufbau für unterschiedliche Regionen und Klimavariablen basierend auf vier differenzierten Evaluationsverfahren bewertet. Die Gewichtung der Modelle entsprechend ihrer so erfassten Leistungsfähigkeit erhöht das Vertrauen in Klimamodellprojektionen. Globale und regionale Klimamodelle werden anhand von 50-jährigen Trends in der jahreszeitlichen Entwicklung von Niederschlag und Temperatur bewertet. Die Arbeit wird auf zwei Skalenniveaus durchgeführt: sieben große, global verteilte Gebiete sowie acht Unterregionen des Mittelmeerraums werden zur Modellbewertung verwendet. Insgesamt werden 62 Modelle mit 159 transienten Simulationen des Niederschlags und 119 der Temperatur aus vier Emissionsszenarien auf Grundlage der ERA-20C Reanalyse evaluiert. Die Ergebnisse zeigen hohe Übereinstimmung zwischen drei von vier Metriken. Die vierte Metrik untersucht eine Modellcharakteristik, deren Ergebnisse keinen Zusammenhang mit den übrigen aufweisen. Insgesamt zeigt insbesondere die Temperatur eine hohe Übereinstimmung mit den Beobachtungsdaten, wohingegen der Niederschlag größeres Potential für Gewichtungen bietet. Allerdings fallen die Unterschiede in der Modellbewertung insgesamt gering aus, so dass sich die vier Metriken eher zur Erstellung von Ranglisten als von Gewichtungen anbieten. Generell stimmen die Ergebnisse mit denen früherer Studien überein: sowohl für die Modellleistung als auch die Effekte der Gewichtung probabilistischer Klimaprojektionen gilt, dass sie jeweils von Untersuchungsregion, Jahreszeit und Variable abhängig sind. Entsprechend konnte kein Modell identifiziert werden, dass den anderen Modellen durchweg überlegen ist.

Keywords: Performance evaluation, Mediterranean, global, probabilistic climate projections, model weighting

1 Introduction

Climate change will increase existing or create new risks in future geosystems (IPCC 2013). Dynamical models are the best source of information for planning and adaptation strategies (IPCC 2007; IPCC 2013). A major source for uncertainty in climate prediction derives from the uncertainty about future concentrations of greenhouse gases. To overcome this problem, various idealistic emission scenarios are employed in systematic studies (Nakicenovic et al. 2000; Moss et al. 2010). However, models also have individual deficits due to inadequate resolution or coverage of physical processes (Reichler and Kim 2008; Giorgi et al. 2009; Wang et al. 2014). Both aspects result in inter-model spread, displaying a substantial uncertainty considering the 21st century climate. Hence, reliable climate change projections are one of the most challenging tasks for climate science (Power et al. 2012; Knutti and Sedláček 2012). A popular way to achieve those is the performance-based weighting of models to increase the impact of better performing models in a multi model ensemble.

To assess which models provide the highest reliability concerning future climate change, performance metrics are applied (Stainforth et al. 2005; Hawkins and Sutton 2009). Most of these evaluation approaches concentrate on historic simulations of climate models for the 20th century assuming that high model accuracy or errors in present climate can be transferred to the reliability of future projections (Tebaldi and Knutti 2007; Nikulin et al. 2012). However, there is no ideal way to evaluate climate models so far. Therefore, different evaluation approaches should be applied and models used according to their attested properties (Räisänen and Ylhäisi 2012; Hidalgo and Alfaro 2015; Leduc et al. 2016). Since there is a wide range of climate model evaluation metrics (e.g. Giorgi and Mearns 2002; Perkins et al. 2007; Gleckler et al. 2008; Kumar et al. 2013; Sanderson et al. 2015; Leduc et al. 2016, Ring et al. 2017) which are mostly based on different regions and reference data sets, the synopsis of their results is a challenging task. On the basis of several case studies, Christensen et al. (2007) and Weigel et al. (2010) have demonstrated that choosing the wrong evaluation metrics constitutes a potential new source of uncertainty.

Therefore, the aim of this study is to analyze the results of different performance metrics that have been newly developed in the context of this survey in a standardized setup for the trend of 50 years from 1960-2009 for the historic simulations of 62 models of the Coupled Model Intercomparison Project 3 (CMIP3) and 5 (CMIP5). In contrast to most prior studies, we carry out a very broad and systematic assessment and comparison of the model weighting approaches applied to different climate model ensembles, different regions of the globe, different climate variables and different seasons. In addition, we go one step ahead by transferring the model weights to weighted probabilistic climate predictions with potential effects on the model spread. To get a maximum output of detail, the evaluations are performed for all models and four very different performance metrics in a systematical setup. Based on the metrics results, the models are weighted to increase the impact of better performing models on climate projections. Further, the transferability of metrics to different regional scales is tested. For this, we study seven large regions spread over the globe as well as eight sub-regions of the Mediterranean area. Moreover, the effect of different reference data sets on climate model performance rating is analyzed for all metrics. Thus, we construct a systematical analysis and work out strengths and weaknesses of each applied metric. For both multi model ensembles two future emissions scenarios are considered. In addition to the weighting of single scenario probability density functions (PDFs), a kernel-based combination of both emissions scenarios is applied, considering their mutual uncertainty.

This study is organized in the following manner: in section 2, the study regions are introduced. Data and Methods are described in section 3 and 4. In section 5, the evaluation results are presented. Here, first the individual model performances are assessed, then the focus is set on seasonal and regional patterns and the multi model ensemble differences. Further, in section 5 the individual model results are used as weights to enhance the relative importance of well-performing simulations. This step is done for the time series trend, the single scenario and a multi scenario kernel approach. Finally, in section 6, the results are discussed and compared to those of prior studies. In section 7 we conclude with a brief summary of the lessons learned.

2 Study areas

Figure 1 shows the seven globally distributed study areas and eight Mediterranean sub-regions. In this study, we use the same study areas as Ring et al. (2017) for the model performance evaluation to

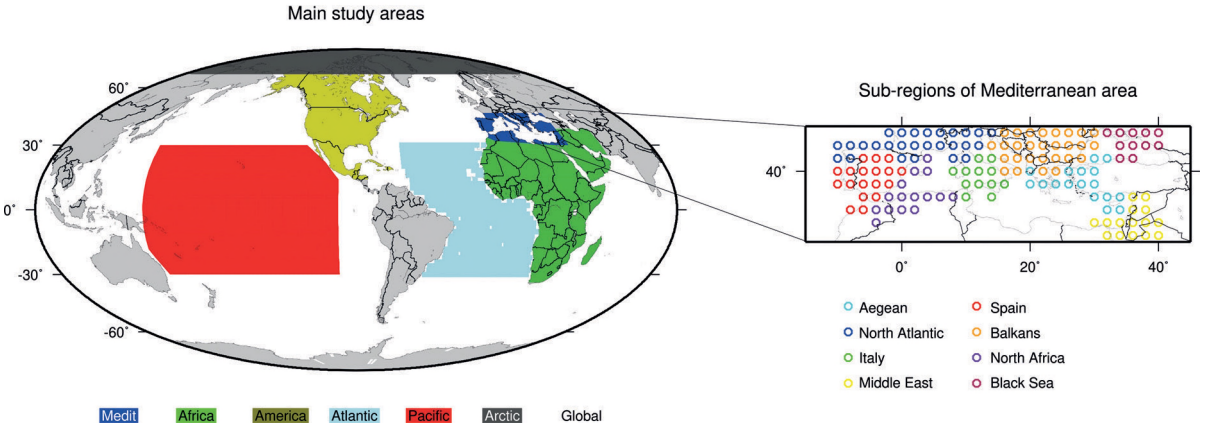

Fig. 1: Overview of the seven large study areas and the eight Mediterranean sub-regions

offer the best basis of comparison. The selection of study areas aims for a coverage of various climatic conditions as well as different challenges considering extent and orography. Areas with land surface are analyzed (North America; Africa; Mediterranean) as well as tropical water surfaces (Atlantic; Pacific). Additionally, two mixed study areas with both water and land surface are analyzed (Globe; Arctic). Using eight sub-regions we focus on the Mediterranean as a hot spot for climate change (Giorgi 2006; Diffenbaugh and Giorgi 2012; Paeth et al. 2016). The sub-regions (Aegean; North Atlantic; Italy; Middle East; Spain; Balkan; North Africa; Black Sea) were selected by means of a principal component analysis of annually aggregated precipitation sums (Ring et al. 2017).

3 Data sets

3.1 Validation data

The main reference data set is the ERA-20C reanalysis compiled by the European Centre for Medium-Range Weather Forecasts (ECMWF) (Poli et al. 2013). Because of the diversity of study areas, the validation data set needs to cover both land and water surfaces for monthly temperature and precipitation for the second half of the 20th century starting 1960. ERA-20C meets all requirements with a global coverage on a 2.5° x 2.5° grid. Even though, ERA-20C is not an observational data set, prior studies attest ERA-20C to constitute a reliable basis for model evaluation in the 20th century (Donat et al. 2016; Dittus et al. 2016). To test the impact of different types of reference data, two weather station based observational data sets are considered as well: E-OBS V12 (Haylock et al. 2008) and CRU TS3.23 (Mitchell and Jones 2005). Both are generally suitable as reference data set (see Tab. 1). However, since they only cover land surface and E-OBS is limited to Europe, we use them for the Mediterranean sub-regions only. Here, several applications of the metrics are carried out to assess the differences in model performances based on each validation data set. For evaluation, all data sets are interpolated to a regular 2° x 2° grid and seasonal precipitation and temperature are calculated.

3.2 Model data

A wide selection of global climate model simulations is employed. For the evaluation, we analyze 20c3m and Historical runs for the time frame of 1960-2009 from CMIP3 and CMIP5, respectively (Randall et al. 2007; Flato et al. 2013). For both multi model ensembles two emissions scenarios are

Tab. 1: Overview of utilized reference data sets

Data set	Temporal coverage	Spatial coverage	Orig. resolution
ERA-20C	1900–2009	Global	1° x 1°
CRU TS3.23	1901–2014	Global (land only)	0.5° x 0.5°
E-OBS V12	1950–2015	Europe (land only)	0.25° x 0.25°

used to assess a wide range of future climate developments. For CMIP3 the Special Report on Emissions Scenarios (SRES) A1B and A2 and for CMIP5 the Representative Concentration Pathways (RCP) 4.5 and 8.5 are considered. Thus, a high and medium emissions scenario is selected for both multi model ensembles. It should be noted that the SRES and RCP scenarios are not interchangeable. For details on the background of SRES and RCP see NAKICENOVIC et al. (2000) and MOSS et al. (2010). Nevertheless, these scenarios offer a suitable gradation of potential future pathways in order to assess a reasonable confidence interval for future climate. Overall, 24 (38) global climate models of CMIP3 (CMIP5) with 54 (105) simulations of precipitation and 57 (62) simulations of temperature are evaluated (see Tab. 2). The number of available ensemble runs depends on variable and scenario. Generally, all available simulations have been used. We are aware that different global climate models or different realizations of one model are not independent from each other. In fact, most are based on similar initial assumption (IPCC 2013). However, this dependence can be neglected since this study aims for a systematical comparison of evaluation metrics based on as many climate simulations as possible for a wide range of different study areas. Although our focus lies on the assessment of global climate models, we include 18 simulations from 3 regional climate models from the Coordinated Regional Downscaling Experiment (CORDEX) framework (JACOB et al. 2014) in this study. These offer very high spatial resolutions and are assumed to achieve an added value compared to global climate models (see Tab. 3). Since the historical CORDEX simulations only start in 1970, we analyze their performance for the shorter period 1970-2009. As for the reference data sets, all climate model data are interpolated to a regular 2° x 2° grid.

Tab. 2: CMIP3 and CMIP5 models used in this study. The numbers indicate how many ensemble simulations of each model (rcp45,pre/ rcp85,pre/ rcp45,temp/ rcp85,temp) are used.

Models (CMIP3)	Models (CMIP5)
BCCR_BCM2.0 (1/1/1/1)	ACCESS1-0 (1/1/1/1)
CGCM3.1(T47) (5/5/5/5)	ACCESS1-3 (1/1/1/1)
CGCM3.1(T63) (1/-/1/-)	BCC-CSM1.1 (1/1/1/1)
CNRM-CM3 (1/1/1/1)	BCC-CSM1.1(m) (1/1/1/1)
GFDL-CM2.0 (1/1/1/1)	CanESM2 (5/5/5/5)
GFDL-CM2.1 (1/1/1/1)	CCSM4 (6/6/6/6)
GISS-AOM (2/-/2/-)	CESM1-BGC (1/1/1/1)
GISS-EH (3/1/3/1)	CESM1-CAM5 (3/3/3/3)
GISS-ER (2/-/5/-)	CMCC-CM (1/1/1/1)
FGOALS-g1.0 (3/-/3/-)	CMCC-CMS (1/1/1/1)
INM-CM3.0 (1/1/1/1)	CNRM-CM5 (1/5/1/1)
IPSL-CM4(LMDZ) (1/1/1/1)	CSIRO-Mk3-6-0 (10/10/10/10)
INGV-SXG (1/1/1/1)	CSIRO-Mk3L-1-2 (3/-/3/-)
MIROC3.2(hires) (1/-/1/-)	EC-EARTH (4/5/-/-)
MIROC3.2(medres) (3/3/3/3)	FGOALS-g2 (1/1/1/-)
MRI-CGCM2.3.2 (4/4/5/4)	FIO-ESM (-/-/1/-)
ECHO-G (3/3/3/3)	GFDL-CM3 (1/1/1/1)
CSIRO-Mk3.0 (1/1/1/1)	GFDL-ESM2G (1/1/1/1)
CSIRO-Mk3.5 (1/1/1/1)	GFDL-ESM2M (1/1/1/1)
ECHAM5/MPI-OM (4/3/4/3)	GISS-E2-H-CC (1/1/-/-)
CCSM3 (7/4/7/4)	GISS-E2-H (16/6/-/-)
PCM (4/4/4/4)	GISS-E2-R-CC (1/1/-/-)
UKMO-HadCM3 (1/1/1/1)	GISS-E2-R (17/5/-/-)
UKMO-HadGEM1 (1(1/1/1/1)	HadGEM2-AO (1/1/1/1)
	HadGEM2-CC (1/1/-/-)
	HadGEM2-ES (4/4/-/-)
	INMCM4 (1/1/1/1)
	IPSL-CM5A-LR (4/4/4/4)
	IPSL-CM5A-MR (1/1/1/1)
	IPSL-CM5B-LR (1/1/1/1)
	MIROC5 (3/3/3/3)
	MIROC-ESM-CHEM (1/1/1/1)
	MIROC-ESM (1/1/1/1)
	MPI-ESM-LR (3/3/3/3)
	MPI-ESM-MR (3/1/2/1)
	MRI-CGCM3 (1/1/1/1)
	NorESM1-ME (1/1/1/1)
	NorESM1-M (1/1/1/1)

4 Methodology

Assigning weights to models within a multi model ensemble requires a detailed evaluation based on their modelling performance compared to reference data, i.e. meteorological observations over recent decades. The assessment of model weights is based on statistical scores that measure the bias between model and reference data with respect to specific climate features, like mean and trend patterns, extremes and spectra of climate variability. Climate models with higher skill scores are assigned a larger weight. The model weights can then be used to

Tab. 3: CORDEX simulations used (one ensemble member each)

Global Model	Regional Model	Resolution
CNRM-CERFACS-CNRM-CM5 (1/1/1/1)	SMHI-RCA4	0.11° x 0.11°
ICHEC-EC-EARTH (1/1/1/1)	SMHI-RCA4	0.11° x 0.11°
ICHEC-EC-EARTH (1/1/1/1)	DMI-HIRHAM5	0.11° x 0.11°
IPSL-IPSL-CM5A-MR (1/1/1/1)	SMHI-RCA4	0.11° x 0.11°
MOHC-HadGEM2-ES (1/1/1/1)	SMHI-RCA4	0.11° x 0.11°
MPI-M-MPI-ESM-LR (1/1/1/1)	SMHI-RCA4	0.11° x 0.11°
CCCma-CanESM2 (1/1/1/1)	SMHI-RCA4	0.11° x 0.11°
CNRM-CERFACS-CNRM-CM5 (1/1/1/1)	SMHI-RCA4	0.11° x 0.11°
CSIRO-QCCCE-CSIRO-Mk3-6-0 (1/1/1/1)	SMHI-RCA4	0.11° x 0.11°
ICHEC-EC-EARTH (1/1/1/1)	SMHI-RCA4	0.44° x 0.44°
ICHEC-EC-EARTH (1/1/1/1)	KNMI-RACMO22E	0.44° x 0.44°
ICHEC-EC-EARTH (1/1/1/1)	DMI-HIRHAM5	0.44° x 0.44°
IPSL-IPSL-CM5A-MR (1/1/1/1)	SMHI-RCA4	0.44° x 0.44°
MIROC-MIROC5 (1/1/1/1)	SMHI-RCA4	0.44° x 0.44°
MOHC-HadGEM2-ES (1/1/1/1)	SMHI-RCA4	0.44° x 0.44°
MPI-M-MPI-ESM-LR (1/1/1/1)	SMHI-RCA4	0.44° x 0.44°
NCC-NorESM1-M (1/1/1/1)	SMHI-RCA4	0.44° x 0.44°
NOAA-GFDL-GFDL-ESM2M (1/1/1/1)	SMHI-RCA4	0.44° x 0.44°

compute weighted ensemble means and probabilistic climate predictions with potential shifts in the mean change and uncertainty. Most of the metrics applied in this study are rather novel statistical approaches to evaluate climate model data output in that sense that they have been used in various scientific and statistical contexts, however, except for the root mean square error metric, their use for rating model performance is tested here for the first time. The metrics strongly differ concerning their complexity and evaluation parameter. We apply them to the trend patterns as well as to spectral time series characteristics. The RMSE is a basic statistical tool which has been frequently used for bias analyses (e.g. Ring et al. 2016), therefore, this metric is considered as a benchmark index. The fingerprinting approaches (FPA) and the harmonic spectrum metric (HM) are used exploratively. To generate a comprehensive knowledge on model performance it is necessary to apply and compare various metrics that, partly, have not been subject to performance evaluation before. The FPA was introduced as a tool of model evaluation by Paeth and Mannig (2013). This approach benchmarks two different types of key model features, the similarity of spatial trend patterns between model and observation and the ability of the model to detect an anthropogenic climate change signal in this trend pattern. The HM metric has been newly developed in the framework of this study. Here, we analyze whether the observed power spectrum, i.e. the relative importance of time scales of climate variability, are reproduced by the models. Both metrics, FPA and HM, have not yet been investigated in the context of model evaluation and weighting. They offer new insights into specific and important aspects of model performance and will, hence, improve the general assessment of current climate models. The RM approach is more common and serves here as a benchmark for the new metrics FPA and HM.

4.1 The Root Mean Square Error metric (RM)

The Root Mean Square Error (RMSE) is a well-known and frequently used statistical skill score. Therefore, it offers a very transparent basis for model evaluation. For each model, every grid point is considered and compared to the observational data equivalent.

$$RM_m = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_{m_i} - y_i\right)^2} \qquad 4.1$$

Here, the RM-skill score (4.1) is calculated for each model m by the RMSE over all grid (n) points x_i for $i = 1, \ldots, n$ and the observational data y_i. We use a regional RMSE Metric (RM) for the climatological trend.

4.2 Harmonic spectrum metric (HM)

The second metric is an explorative approach for climate model evaluation. The harmonic spectrum metric (HM) compares the spectral time se-

ries characteristics of each climate model simulation to the respective validation data set for the study period of n = 50 years (1960-2009). Most other studies on model performance consider grid box based climatic similarities or indices (e.g. PERKINS et al. 2007; RING et al. 2016; KOUTROULIS et al. 2016). For HM, the harmonic time series components are compared with each other. First a Fourier transform is performed. Every time series can be expressed based on the underlying frequencies $\frac{2\pi t}{n}$ with the number of years (n) and the time steps t. Thus, the time series is synonymous to the amplitude C_k with the corresponding phase Φ_k (4.2) (e.g. WILKS 2006, 371ff.).

$$C_k cos\left(\frac{2\pi t}{n} - \Phi_k\right) = A_k cos\left(\frac{2\pi t}{n}\right) + B_k sin\left(\frac{2\pi t}{n}\right) \quad 4.2$$

Here, k stands for each combination of waves or harmonic functions necessary to reproduce the entire time series. Because of the independence of sine and cosine a specific proportion of explained variance R_k^2 for each k can be calculated by C_k^2 and the variance of the original input data s^2 (4.3).

$$R_k^2 = \frac{(n/2)C_k^2}{(n-1)s^2} \quad 4.3$$

Hence, the sum over all R_k^2 equals 1. Now, we consider R^2 as the performance indicator of each simulation. That means R^2 of the respective periodic length or wave should be similar to that of the validation data. As an example: for temperature most simulations R_m^2 show a high explained variance for the longest periodic length (50 years) or wave (n/2 = 25). The same results are found for the validation data R_0^2 indicating the warming trend during the study period 1960-2009. This should result in a high model performance rating. Therefore, we consider seven harmonics covering periodic lengths from 7 to 50 years and calculate the RMSE.

$$HM_m = \sqrt{\frac{1}{7}\sum_{p=1}^{7}\left(R_m^2 - R_0^2\right)^2} \quad 4.4$$

This RMSE is used as the index of similarity or performance metric HM (4.4) for the respective model m with unit Δr^2. Period lengths below seven years are neglected as background noise, i.e. internal or natural climate fluctuations that cannot be reproduced by uninitialized climate model simulations.

4.3 Fingerprinting approaches (FPA)

The last approach (2 metrics) to assess climate model performance is based on the fingerprinting introduced by HASSELMANN (1979) and HEGERL et al. (1996). It is applied by the scalar product of the simulated vector $\vec{x}$ and reference data vector $\vec{y}$.

$$d = \vec{x} \cdot \vec{y} \quad 4.5$$

For both the reference $\vec{y}$ and simulated vector $\vec{x}$ we use the 50-year trend from 1960-2009. Two fingerprint approaches, the optimal and the suboptimal, are considered. In this study, the we use the terminology and interpretation of PAETH and HENSE (2001) and PAETH and MANNIG (2013). For both approaches, the detection variable d is determined to assess the simulation performances. The fingerprinting approaches are considered as filter. The optimal fingerprint (OPT) reduces the impact of the noise component as much as possible and, therefore, provides information about the similarity of the climate change signal. The suboptimal fingerprint (SUB) ignores this aspect and analyses the overall accordance of the climate pattern or vector $\vec{x}$. We use this filter to extract the signal in both validation and model data and estimate its similarity as a performance metric d for the 50-year trend from 1960 to 2009 of the model t_{sim} and observational data t_{obs} with k dimensions depending on the number of grid boxes of the respective region. For the suboptimal fingerprint d_{sub} (4.6) is calculated as

$$d_{sub}(t_{sim}, t_{obs}) = \frac{\sum_{i=1}^{k}\chi_i(t_{sim})\,y_i(t_{obs})}{\sqrt{\sum_{i=1}^{k}\left[\chi_i(t_{sim})\right]^2 \sum_{i=1}^{m}\left[y_i(t_{obs})\right]^2}} \quad 4.6$$

and hence d_{sub} is normalized to [-1,1] indicating high agreement of simulation and observational data for values near to 1. For the optimal fingerprinting approach, the climate signal is filtered and evaluated. However, it is necessary to assess the inverse matrix of natural variability C^{-1} as a filter. Since natural variability is unknown it has to be estimated from historic climate information prior to a dominating anthropogenic climate change signal. Here, we use historic climate simulations with weak anthropogenic forcing as best guess: 50-year trends starting from 1850-1899 to 1900-1949 are considered from all models. Based on these trends (>3600) the covariance matrix C is constructed. Then, a principal component analysis is performed to process the inversion of the covariance matrix C_{ij}^{-1}. The detection variable

d_{opt} (4.7) is then calculated using the leading 8 PCs accounting for >94 % (>72%) explained variance for temperature (precipitation).

$$d_{opt}(t_{sim},t_{obs}) = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k}\chi_i(t_{sim})\gamma_i(t_{obs})C_{ij}^{-1}}{\sqrt{[\sum_{i=1}^{k}\sum_{j=1}^{k}\chi_i(t_{sim})\chi_i(t_{sim})C_{ij}^{-1}][\sum_{i=1}^{k}\sum_{j=1}^{k}\gamma_i(t_{obs})\gamma_i(t_{obs})C_{ij}^{-1}]}} \quad 4.7$$

Here, k is equivalent to the number of PCs. For best comparability the suboptimal performance index d_{sub} is calculated for the PC as well as for the grid box dimension, hence, without data reduction for k x k grid boxes. Since both fingerprinting approaches are based on large spatial climate patterns, the FPA are only used for the global study areas and the entire Mediterranean to avoid random results from the smaller Mediterranean sub-regions.

5 Results

5.1 Evaluation results

Figure 2 shows the 1960-2009 annual trend pattern of different climate model simulations and the ERA-20C validation data set for precipitation and temperature in the entire Mediterranean area (Medit). The best and worst performing simulation are displayed for every metric, RM, HM, SUB and OPT. It should be noted that this result is based on a specific situation that is not necessarily transferable to other combinations of regions or seasons. Nevertheless, we find that not some metrics have the same simulation ranked first and last. However, differences of the skill scores might be small between some simulations.

For precipitation (pre), the validation data set shows a rather strong decrease for most parts of the Mediterranean region with a maximum over the Adriatic Sea. There merely is some marginal increase for single grid cells. This pattern is best matched by the simulation ranked first of RM, SUB and OPT (MRI CGCM2-3-2a R1). Here, we see much similarity with a predominant decrease for most parts with its maximum from Italy to the southern Balkan. Further, we see some increase over southern France. ERA-20C shows a slightly weaker decrease. The pattern is similar to MRI CGCM2-3-2a, however, there is a bias between model and validation data set. This aspect is irrelevant to the fingerprinting metrics while RM finds the smallest deviation to the reference here as well. In contrast, we find a different result for HM. Here, the first ranked IPSL-CM5A-LR R2 shows a considerable increase for the entire study area. This indicates that similarities of harmonic characteristics of the time series of simulation and validation data not necessarily imply the same long-term climate trend pattern as displayed in Fig. 2. RM, SUB and OPT all promote simulations which are visually in agreement with the validation data because the metric based explicitly on the trend pattern. The lowest ranks for all RM, SUB and OPT are assigned to simulations displaying increase in precipitation for large areas. RM even selects the first ranked simulation of HM to be last ranked here.

For temperature (temp), ERA-20C shows a rather homogenous increase that peaks over the Balkan. The first ranked simulations here all show temperature increases for the entire study area as well. In contrast to the precipitation results, all metrics choose different simulations as their first ranked result. RM (CSIRO-Mk3-6-0 R4) and SUB (MPI-ESM-MR R3) offer the highest visual resemblance. Here, even the amount of the increase is on a similar level. For OPT (INGV-ECHAM4 R1) we find a slightly lower increase. Again, the bias remains unimportant for the ranking. For HM (CSIRO-Mk3-6-0 R3) we see very high values from eastern Spain over the southern part of France to eastern boundary of the Alps. For the other study areas and seasons there are similarities in the agreement of the metrics (not shown). RM and the fingerprinting approaches capture the climate pattern better, while HM shows generally different results. Therefore, HM evaluates a new climate model characteristic that does not target the trend pattern in Fig. 2. For HM, SUB and OPT, the highest and lowest ranks are each assigned to different model simulations. This indicates that there is not one single simulation or model neither best nor worst performing in every combination of region and season (situation).

In Fig. 3 we compare the distributions of the results of the model evaluation for all metrics. These values are unprocessed, meaning that they cover different ranges and cannot directly be compared with each other. Major differences exist between the definition and range of them. Thus, interpretation has to be carried out carefully. As RM and HM are based on RMSEs, low values indicate good performances, whereas OPT and SUB are normalized from -1 to 1 (highest correlations for 1).

Fig. 3 shows boxplots of the performance assessments for each season and metric. All seasons are abbreviated for the first letter of the respective three months. Displayed are the results for Medit and Globe for both precipitation and temperature. For

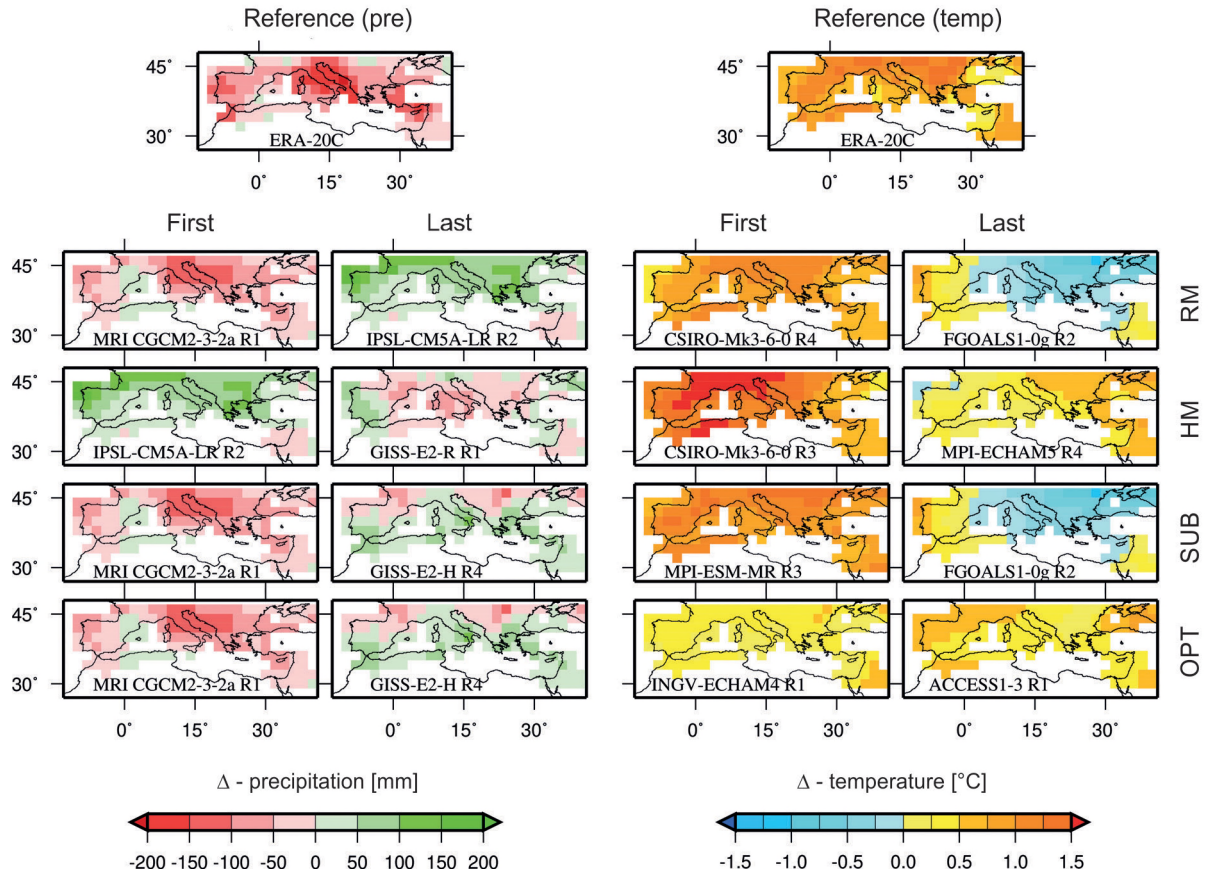

Fig. 2: Comparison of annual precipitation (left) and temperature (right) trend patterns of 1960-2009 for Medit. Displayed are each first and last ranked simulation of the four evaluation metrics: RM, HM, SUB and OPT.

RM, there is a rather stable bias of about 50 mm for Global precipitation. For Medit MAM, JJA and SON the box plots are around 25 mm while DJF, the wettest season, shows the largest extent of error bars on a level of 50 mm as well. For the other study areas, the level of the median depends on the general seasonal precipitation amount as well. The HM results show similar boxplots for all settings between 0.02 and 0.12 Δr^2. For SUB and OPT, we find an overall similar distribution over all seasons and regions centered round 0 with a higher spread for SUB.

For temperature the RM results spread around 0.6 °C for most situations. However, the largest range exists for JJA in Medit. This is the situation with the highest values just like DJF for winter precipitation. Thus, high temperature values offer a potential for more diversified performance ratings. A similar effect can be found for HM. While evaluations results for the Globe are similar over all seasons, study areas with a strong annual cycle show higher capability for different model weights. For SUB and OPT, we see a strong discrepancy in model performances. The overall climate pattern, rated by SUB is much stronger with most values between 0.6 and 0.99. Only DJF shows some weaker - even negative - results. For OPT on the other hand, the results are more similar to those of precipitation. Nevertheless, the general median level is slightly higher around 0.25 for Medit and between 0.25 and 0.6 for the Globe. For the fingerprinting approaches the global view shows the best results while RM and HM are dependent on the respective annual temperature or precipitation maximum.

In Fig. 4 the best mean evaluation result over the respective simulations are displayed for each situation for precipitation and temperature. Since the multi model ensembles comprise different numbers of simulations (see Tab. 2-3) this factor is considered.

Note that SUB and OPT are only calculated for the main study areas and hence the respective boxes of the sub-regions cannot be filled. Further, the evaluation of CORDEX simulations can only be done for the Mediterranean sub-regions. Regarding precipitation and the large study areas, CMIP3 is found most

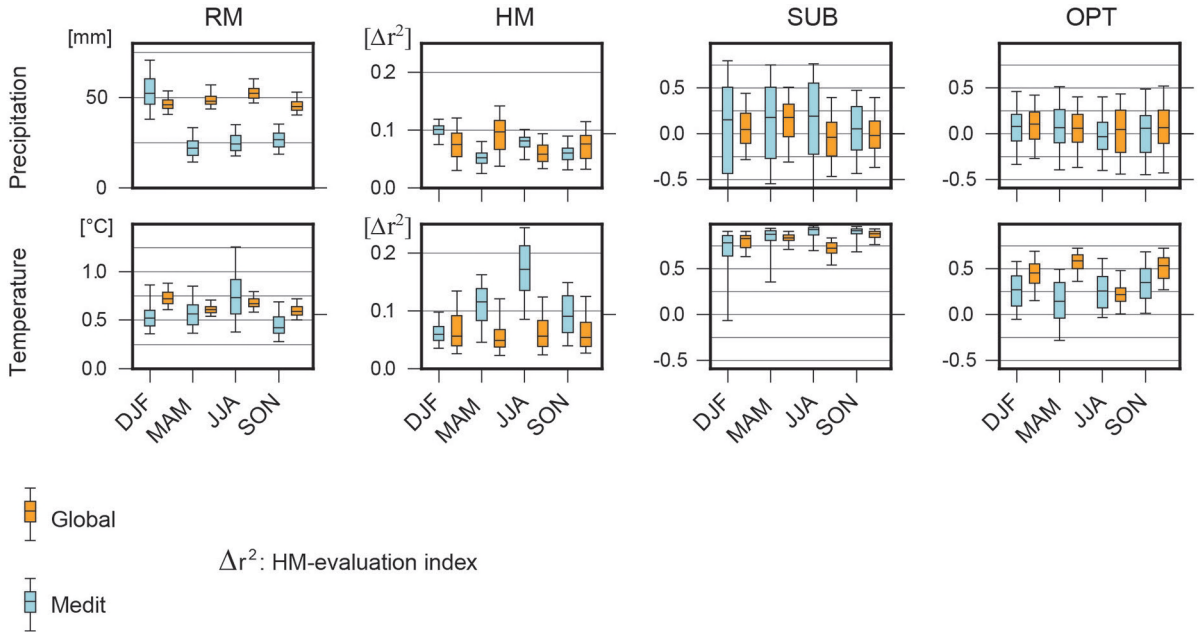

Fig. 3: Mean seasonal evaluation results of each metric for precipitation (top) and temperature (bottom). Boxplots show the median and error bars of the 5th, 25th, 75th and 95th percentile.

frequently. This result is somewhat surprising since the newest model generation is CMIP5. Especially for Medit, CMIP3 outperforms CMIP5 concerning every metric and nearly all seasons. The best performance for CMIP5 can be found for the Arctic. Here, only HM sees best results for all seasons by CMIP3. Generally, it is apparent that the results of RM and HM are rather similar with the evaluation assessments of CMIP3 being mostly higher than those of CMIP5. On the other hand, for SUB and OPT, CMIP5 is superior for most situation. In the sub-regions we see stronger differences between RM and HM. For RM, the majority of situations again show CMIP3 as best performing multi model ensemble, while CMIP5 and CORDEX are predominant for HM.

For temperature, the red colors of CMIP5 and CORDEX are considerably more dominant. Most situations with CMIP3 as best mean evaluation result again are produced by RM and HM for the global regions. However, CMIP5 is found here more frequently as well. The strongest region for CMIP5 is Medit with 13 out of 16 evaluations. On the other hand, for Pacific temperature CMIP3 is best performing in 12 out of 16 situations. The strong CMIP5 and CORDEX performance for Medit continues on the sub-regional scale as well. RM shows 75% of best results by one of the more recent generations of multi model ensembles. In addition, we see a considerable added value of the regional climate models. In 17 out of 64 situations CORDEX outperforms CMIP3 and CMIP5. Overall, it can be concluded that CMIP5 shows better performance in reproducing the correct climate pattern (SUB, OPT) for both precipitation and temperature. For the precipitation bias (RM), CMIP3 shows stronger results for the main and sub-regions. For temperature, CMIP5 seems slightly improved compared to CMIP3. Again, the HM results are difficult to interpret because they appear to offer a unique perspective on climate model performance.

In Fig. 5 the spearman correlation of the final model rankings between all metrics over all seasons is shown for the main study areas. Additionally to RM, HM, SUB and OPT, we include SUB-PX, a pixel based suboptimal fingerprinting approach, to analyze whether the data reduction preprocessing of SUB and OPT is influencing the results. Fig. 5 shows that the model ranking arising from the five metrics is rather similar for most major study areas. We find high correlation coefficients above 0.74 between all three fingerprinting approaches for both precipitation and temperature. Thus for the fingerprinting approaches, models that perform well in simulating the climate change signal show almost equally high results for the general climate pattern (SUB). This is true for both precipitation and temperature. Further, we see positive correlations between the fingerprinting approaches and RM. However, there are some differences depending on the respective region. Temperature correlations are

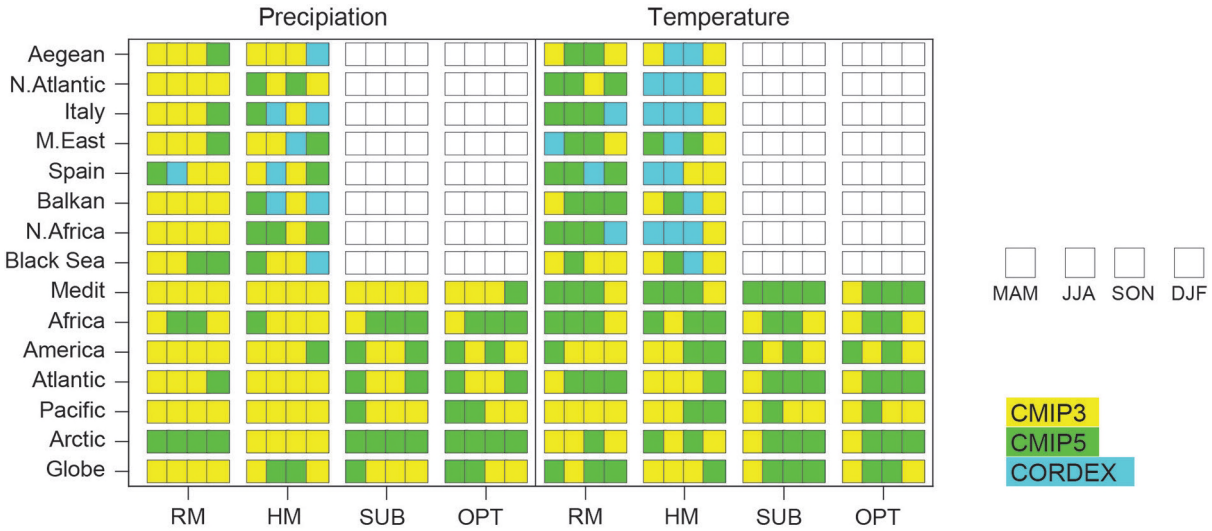

Fig. 4: Comparison of best performing multi model ensembles for all regions and seasons according to mean weight.

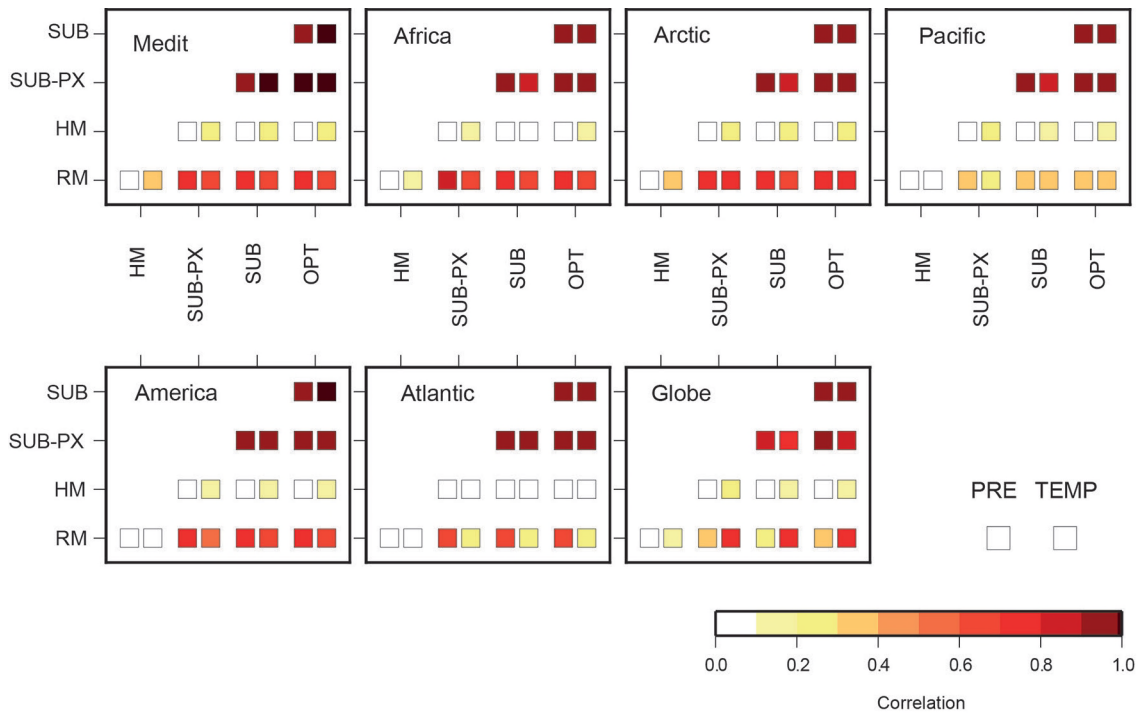

Fig. 5: Spearman correlation coefficients between the rankings of each metric for precipitation and temperature for each region

overall much higher with minimum 0.6 except for the Atlantic region (0.23). Apparently, the precipitation and temperature trend for the continental regions are simulated on a high level based on RM and the fingerprinting approaches. For HM, we find no mentionable correlation whatsoever (typically within +/-0.2). This metric targets an altogether different aspect of the simulation performances than RM and the fingerprinting approach.

5.2 Sensitivity to reference data

The performance metrics show quite high correlations amongst them. However, all evaluation approaches are dependent on the reliability of validation data sets. The results previously discussed are based on the reanalysis ERA-20C. To test their representativeness, all sub-region evaluations have been performed for two further validation data sets.

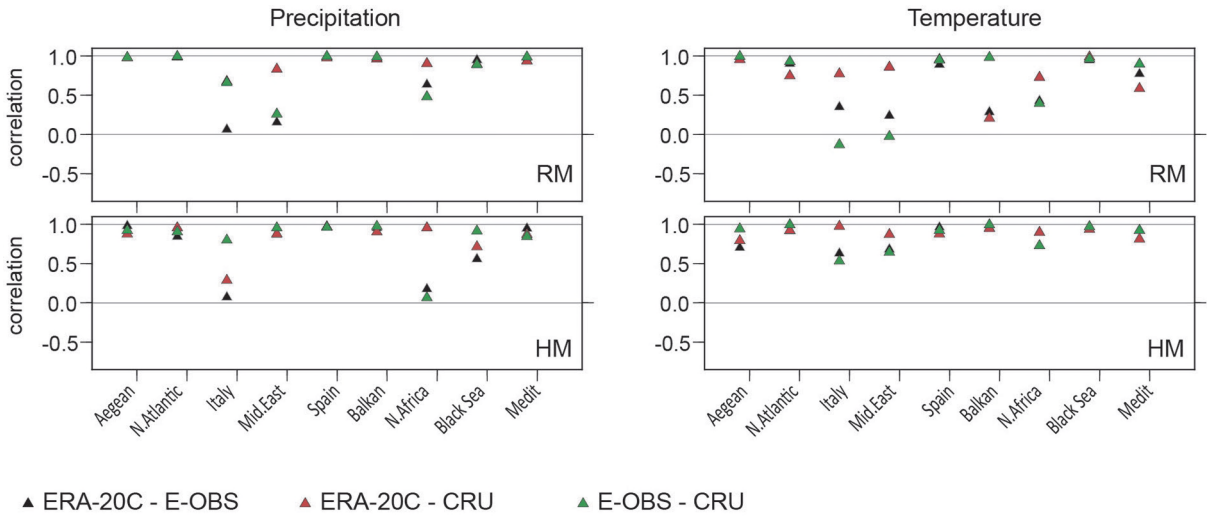

Fig. 6: Spearman correlation coefficients between the DJF evaluation results of all models for different types of validation data sets for the Mediterranean sub-regions

In Fig. 6 the Spearman correlation between the annual rankings from RM and HM for all three possible combinations of reference data sets for the sub-regions are shown. For most situations, the correlations for HM are positive with coefficients of 0.5 or higher. The minimum is 0.05 to 0.2 for precipitation in Italy and North Africa. Regarding most other situations, correlations are substantially stronger around 0.5 for Black Sea or from 0.7 to 0.95 for the remaining ones. The results for temperature are on a similarly high level. Again, the minimum is found for Italy with 0.5 while most of the coefficients from other situations are 0.6 and higher. For the HM metric, we find a low dependency on observational data. The results of the RM correlation analysis show a similar pattern in Fig.6 with almost only positive values. However, the spread is much higher than for HM. Again, the lowest values around zero result for Italy. For the rest of the situations there are exclusively positive correlations. Since the fingerprinting approaches have been solely performed for the major study areas, we only tested the observational dependence for Medit (not shown). This is on a similar level as RM and HM for both temperature and precipitation with values between 0.5 and 1. Overall, Fig. 6 indicates a certain insensitivity of model ranking on the respective observational data set. This is supported by the results of Medit for SUB and OPT (not shown). Here, Spearman correlations are mostly above 0.9. Especially regarding HM, the results are quite stable.

5.3 Weighting impact on multi model uncertainty of future climate change

Finally, to assess the impact of weighting on the model spread (model uncertainty), the weights are applied to the multi model ensemble for climate changes from the end of the 20th to the end of the 21st century. First, it needs to be mentioned that none of the metrics in their original setup are explicitly designed to reduce the ensemble spread. In Fig. 7, the differences between the equally and metric weighted standard deviation and mean are shown for precipitation (1. row) and temperature (2. row) for the major study areas. Changes of standard deviation and mean that are not significant are marked as black or white triangles. Especially for HM and RM but for SUB and OPT as well, the effects of the weightings are rather small. Therefore, a slightly intensified approach for SUB (SUBi) and OPT (OPTi) was applied which leads to higher differences. This is accomplished by applying a threshold at 0.0, meaning that all simulations with negative metric results are assigned a weight of zero and the remaining weights are normalized. This leads to an emphasis on those models with higher similarity while others are neglected. Because of their RMSE-based range of evaluation results, RM and HM are excluded from this modification.

In Fig. 7, the weighting results of the intensified fingerprinting approaches are marked white or colored (for significant changes) as inverted triangles. Obviously, this procedure has a strong effect on the precipitation model weighting. Here, we see

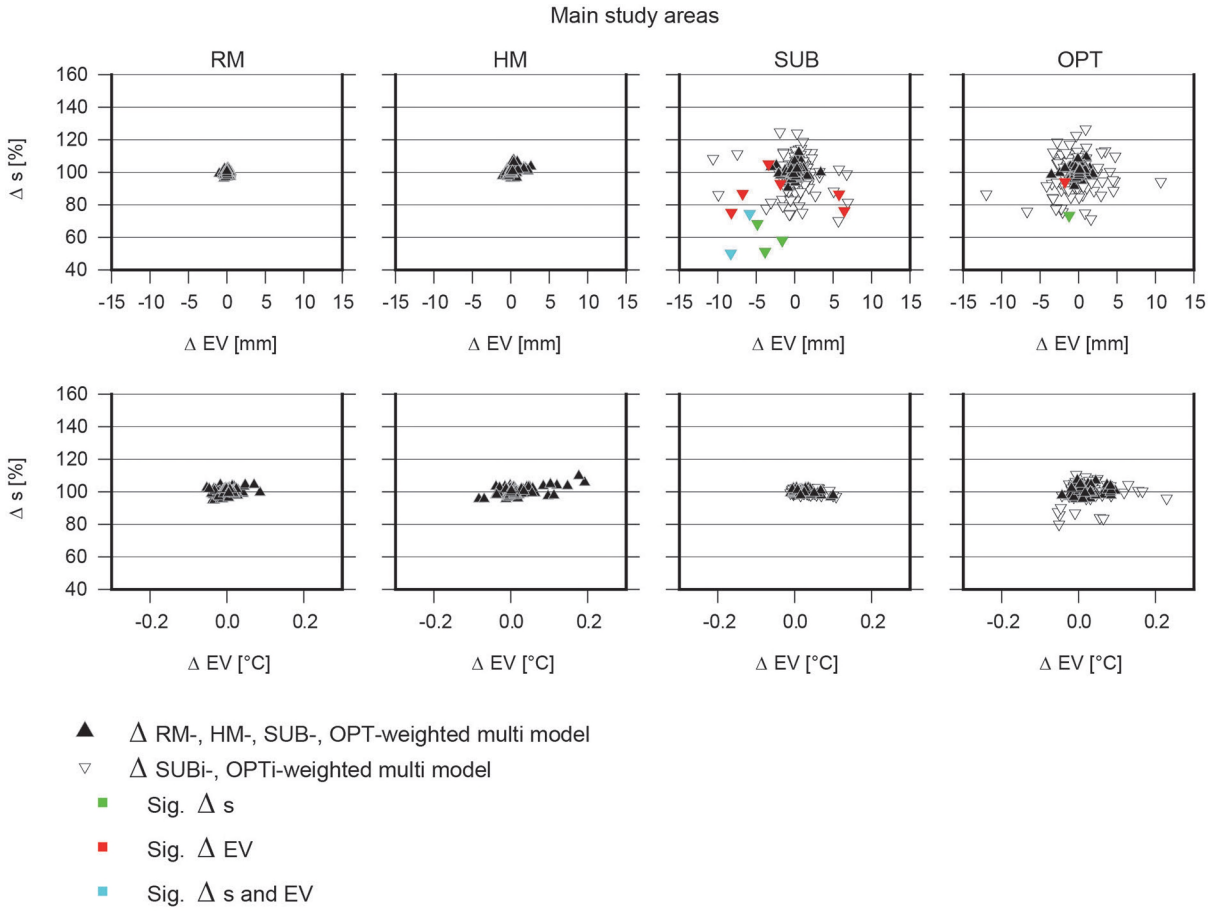

Fig. 7: Summary of the weighting impacts of each metric for main study areas split into precipitation (1. row) and temperature (2. row). Changes are expressed as shifts of the standard deviations on the ordinate (Δ s) and expectation (Δ EV) with respect to the unweighted climate changes for all simulations on the abscissa. The unweighted results would be located in the center with 100 % and 0 mm respectively °C change.

much higher differences between original and metric weighted distributions and even some significant changes. For temperature however, the difference between SUB and SUBi, respectively OPT and OPTi, weighted multi model ensemble is rather small. The reason lies in the very high simulation performance of the climate models for temperature. There are almost no results below the 0.0 threshold. Apparently, the tendency of the weighting effect of the normal and intensified approaches remains unaltered. Thus, SUBi and OPTi appear valid for further investigation. Of course, stronger intensification by shifting the threshold to even higher numbers would be possible as well. However, as this study aims for the comparison of metrics and their results, we decide to not further modify these approaches but to point out their potential. Because of the unstandardized range of evaluation results of RM and HM, a similar approach is not reasonable for both. Since all metrics agree that temperature simulation performance of CMIP3, CMIP5 and CORDEX is on a high level further weighting offers little potential. However, regarding precipitation especially SUBi and OPTi show significant effects on the standard deviations. Further, a shift towards lower standard deviations is discernible for a majority of situations, which can be interpreted as a decrease of uncertainty.

In Fig. 8, Gauss-Kernel-functions (GKF) are applied to the OPTi weighting approach, based on the same standard deviation and mean values of the multi model ensemble for climate changes from the end of the 20th to the end of the 21st century as the results in Fig 8. This is an exemplary presentation of the GKF weighting effect for JJA precipitation and temperature for America, Africa and Pacific for CMIP5. Each plot shows three GFKs: The RCP4.5 (green line), RCP8.5 (red line) and the multi-scenario-Kernel (MSK, blue shading). The MSK has the potential to

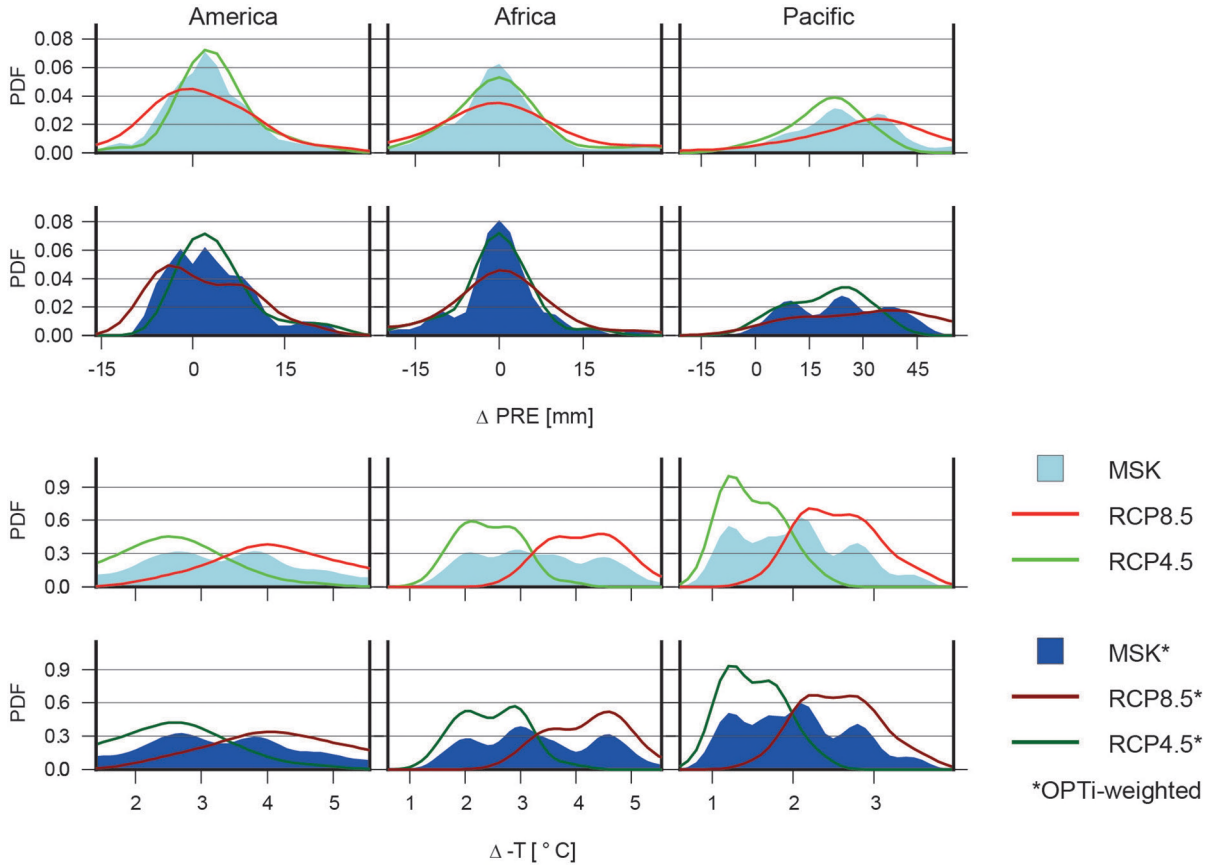

Fig. 8: Single (RCP4.5/RCP8.5) and multi scenario kernel functions (MSK) for change of JJA CMIP5 precipitation (top) and temperature (bottom). The first line shows equally weighted and the second line OPTi-weighted functions.

take two dimensions of uncertainty, the two emissions scenarios (first row) and the weighting impact on the model spread (second row), into account. The equally weighted functions are shown for each situation in the first row (lighter colors) and the OPTi-weighted functions in the second row (darker colors).

For precipitation, most of the MSKs show an increase for Pacific precipitation with a uniform spread between +- 15 mm for America and Africa (first row). Especially for Pacific, the model spread is very high. Both emissions scenarios have their peak at a positive precipitation change, however, RCP4.5 shows a smaller spread overall. The OPTi weighting impact is highly dependent on the respective situation. While there is little effect for Pacific, for Africa and America there is an obvious contraction of the MSK indicating a decrease of model uncertainty, while the maximum remains relatively stable around 0 mm. Against this, for Pacific there is a shift of the expected value towards a smaller precipitation increase with the development of a new local maximum around 8 mm. Additionally, for America and Pacific, multi modal MSKs emerge. The shift from a uni-model PDF within one emissions scenario to different local maxima of higher probability in the MSK approach might become highly important for adaptation strategies.

This effect is even more distinctive for temperature. The strong differences between the RCP4.5 and RCP8.5 scenarios are reflected in Fig. 8. Here, we see multi modal temperature MSKs for every situation of RCP4.5 and RCP8.5. This is similar for CORDEX (not shown), while A1B and A2 have less pronounced local maxima for CMIP3 (not shown). In Fig. 8, we see a range of temperature increases between 1 to 4 °C for Pacific and 1 to 5 °C for America and Africa. As expected, there is little change between the OPTi-weighted and non-intensified functions. This is true for most other situations and all multi model ensembles for temperature.

In Fig. 9 the weighting impact on the ensemble spread in all major study areas for future JJA precipitation and temperature changes are displayed as box plots. Both CMIP3 and CMIP5 show a moderate to strong temperature increase with almost no change

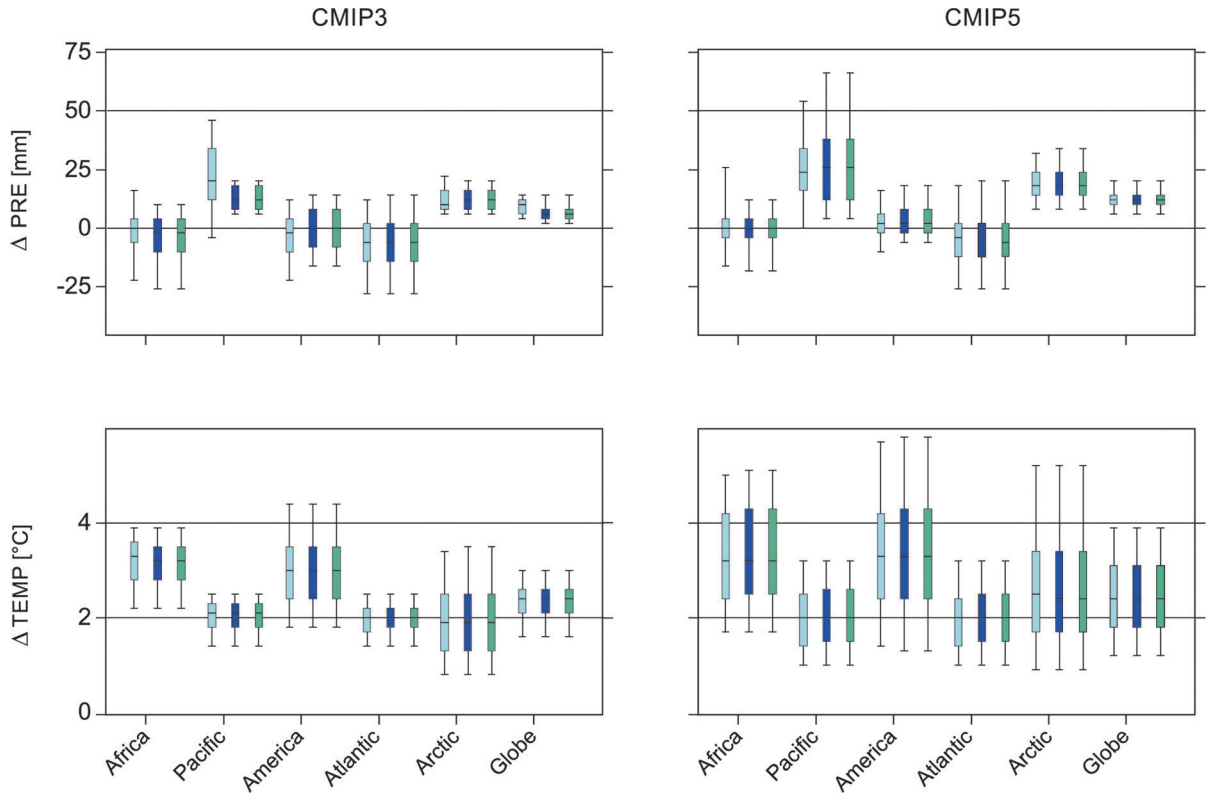

Fig. 9: Summary of the MSK-90 %- confidence intervals of JJA precipitation (top) and temperature (bottom) of CMIP3 and CMIP5 for all main study areas. The unweighted (light blue), SUBi-weighted (dark blue) and OPTi-weighted distributions (green) are displayed. Each Boxplot shows the median and error bars of the 5th, 25th, 75th and 95th percentile.

of spread or mean value related to the weighting. Highest values are found for the Arctic and America reaching 5.2 °C and 5.8 °C. However, here we find the largest range of uncertainty starting at 0.9 °C and 1.3 °C. The other regions spread around a warming between 1 °C and 4.5 °C. For precipitation, Africa, America and the Atlantic show a distribution that is centered around zero. All other regions indicate an increase of precipitation. Although Globe has a very high annual precipitation, the range of uncertainty is lower than for all other regions. The highest uncertainty is found for Pacific. It needs to be mentioned that the impact of SUBi and OPTi weighting remains dependent on the particular situation. However, for some situations such as CMIP3 Pacific or CMIP5 Africa there is a substantial decrease of uncertainty.

6 Discussion

Four different performance metrics have been applied and analyzed in this study. RM and both fingerprinting approaches (FPA) show high consensus in model evaluation of 50-year temperature and precipitation trend patterns. Overall, we conclude that RM and the FPA evaluation metrics are useful for climate model performance rating. Their results indicate comprehensive climate model evaluations fitting in the context of prior studies. RM and HM show high transferability to any regional extent or variable. However, HM turned out to be a generally different approach. Since there are barely any correlations to other metrics we conclude that harmonic time series similarity is an entirely different climate model characteristic than the trend as was found in previous studies regarding the climatological mean (Ring et al. 2017). This conclusion is supported by the general ranking of model performances in Table 4. It illustrates that the mean evaluation results of RM and FPA (mean of SUB and OPT results) are much more similar than those of HM. Considering that RM and FPA both evaluate the climatological trend patterns, this conclusion is somewhat expected. However, Table 4 shows that some of the best ranked models have good results according to all metrics for both precipitation and temperature. However, we assume that there is potential of the HM to add value in climate model evaluation. To support

Tab. 4: Mean ranking over all regional and seasonal situations of precipitation and temperature for all metrics of the major study areas: First 20 (green), last 20 (red), middle range (yellow) and the first and last model of each metric are marked.

	Precipitation				Temperature				
Models	Mean rank	RM	HM	FPA	RM	HM	FPA	Mean rank	Models
GFDL-ESM2M	7,3	14	2	6	5	1	16	7,3	CMCC-CM
CanESM2	8,7	4	6	16	32	3	4	13,0	MIROC3.2(hires)
GFDL-CM3	13,0	10	1	28	1	36	3	13,3	PCM
FGOALS-g1.0	18,3	1	42	12	12	7	24	14,3	CSIRO-Mk3.0
CSIRO-Mk3.0	15,3	10	33	3	4	40	1	15,0	GFDL-CM3
MPI-ESM-MR	13,7	18	22	1	6	15	25	15,3	GISS-ER
INM-CM3.0	19,7	2	52	5	3	5	39	15,7	NorESM1-M
MIROC3.2(hires)	18,3	6	8	41	2	48	2	17,3	CESM1-BGC
BCCR_BCM2.0	15,7	5,0	25	17	15	28	11	18,0	CCSM4
CGCM3.1(T47)	17,0	7	35	9	25	22	8	18,3	HadGEM2-AO
FGOALS-g2	22,0	3	16	47	40	2	14	18,7	UKMO-HadGEM1
GISS-E2-R-CC	20,7	23	32	7	18	12	27	19,0	MIROC-ESM
GISS-E2-H-CC	18,0	29	10	15	41	16	5	20,7	CanESM2
PCM	20,3	39	3	19	8	38	18	21,3	ACCESS1-0
NorESM1-ME	21,0	16	10	37	11	13	40	21,3	FGOALS-g2
MIROC-ESM-CHEM	19,7	17	17	25	33	8	23	21,3	CMCC-CMS
GFDL-CM2.0	21,7	13	41	11	54	4	7	21,7	CGCM3.1(T63)
CGCM3.1(T63)	21,7	22	20	23	37	10	19	22,0	IPSL-CM5A-MR
CNRM-CM5	28,7	7	35	44	47	13	6	22,0	CGCM3.1(T47)
BCC-CSM1.1	21,3	35	9	20	10	40	20	23,3	CCSM3
ECHAM5/MPI-OM	25,3	15	26	35	9	25	37	23,7	MIROC3.2(medres)
MIROC3.2(medres)	26,3	21	54	4	7	36	30	24,3	MIROC5
CCSM3	24,0	32	22	18	21	11	42	24,7	CNRM-CM5
IPSL-CM5A-MR	25,0	31	19	25	13	53	9	25,0	GISS-EH
CNRM-CM3	24,3	36	15	22	24	6	45	25,0	FIO-ESM
MPI-ESM-LR	29,3	30	56	2	45	9	21	25,0	CNRM-CM3
GFDL-ESM2G	28,0	43	5	36	46	18	12	25,3	BCC-CSM1.1
CSIRO-Mk3L-1-2	32,3	9	55	33	21	26	30	25,7	GFDL-ESM2M
CMCC-CM	28,0	33	22	29	36	32	9	25,7	MPI-ESM-MR
IPSL-CM5A-LR	29,3	28	35	25	19	23	38	26,7	GISS-AOM
ACCESS1-0	27,3	48	26	8	29	34	17	26,7	MRI-CGCM2.3.2
GISS-E2-H	28,3	34	30	21	16	24	43	27,7	UKMO-HadCM3
BCC-CSM1.1(m)	31,7	42	7	46	42	21	21	28,0	GFDL-ESM2G
GISS-ER	30,0	25	4	61	53	20	12	28,3	IPSL-CM5A-LR
CESM1-BGC	31,0	37	43	13	17	54	15	28,7	INGV-SXG
CCSM4	31,3	44	21	29	34	31	26	30,3	MPI-ESM-LR
CESM1-CAM5	36,7	27	40	43	14	49	33	32,0	CSIRO-Mk3.5
UKMO-HadCM3	34,0	46	14	42	31	30	35	32,0	GFDL-CM2.1
EC-EARTH	39,0	19	51	47	34	33	29	32,0	BCC-CSM1.1(m)
NorESM1-M	40,0	26	44	50	38	27	35	33,3	GFDL-CM2.0
CSIRO-Mk3.5	32,7	50	34	14	23	50	28	33,7	ECHO-G
GISS-EH	36,0	40	58	10	28	42	32	34,0	MIROC-ESM-CHEM
CMCC-CMS	39,3	47	18	53	30	28	44	34,0	ACCESS1-3
MIROC-ESM	37,0	20	31	60	25	45	34	34,7	NorESM1-ME
HadGEM2-CC	41,7	45	47	33	38	19	48	35,0	BCCR_BCM2.0
IPSL-CM4(LMDZ)	44,7	41	44	49	20	46	40	35,3	CESM1-CAM5
ACCESS1-3	35,3	54	28	24	44	16	51	37,0	ECHAM5/MPI-OM
GISS-AOM	39,7	52	12	55	27	44	46	39,0	FGOALS-g1.0
GFDL-CM2.1	44,7	49	47	38	43	39	47	43,0	IPSL-CM5B-LR
MIROC5	43,0	51	38	40	47	35	50	44,0	CSIRO-Mk3-6-0
MRI-CGCM2.3.2	40,7	57	13	52	52	43	49	48,0	INMCM4
HadGEM2-AO	43,0	55	29	45	49	47	52	49,3	INM-CM3.0
IPSL-CM5B-LR	45,0	58	38	39	50	51	54	51,7	IPSL-CM4(LMDZ)
GISS-E2-R	38,0	24	61	29	51	55	53	53,0	CSIRO-Mk3L-1-2
ECHO-G	50,7	38	57	57	55	52	55	54,0	MRI-CGCM3
HadGEM2-ES	44,7	59	46	29					
INMCM4	43,7	12	60	59					
MRI-CGCM3	54,7	53	53	58					
CSIRO-Mk3-6-0	53,3	60	49	51					
INGV-SXG	57,0	56	59	56					
UKMO-HadGEM1	55,0	61	50	54					

this hypothesis, further studies considering the relevance of harmonic time series evaluation are suggested, especially since there is a strong interest to find different suitable climate model assessment approaches (Gleckler et al. 2008). Contrary to RM, the time series evaluation potentially offers a very strong insensitivity to different reference data types. This advantage over other metrics (Fekete et al. 2004) underlines the importance of further analysis of HM. However, it needs to be noted that using different metrics, study regions, seasons and climate variables often results in other model rankings. Our findings indicate that a most suitable metric can not be identified across all these case studies. On the contrary, we believe that a large spectrum of different metrics tailored to specific climate model characteristics and scientific issues will help putting together the mosaic of climate model performance. Each metric brings its own qualities which will contribute to a more general assessment of the strengths of weaknesses of each model and thus will increase the trust in future projections. A brief summary of advantages and disadvantages of the applied evaluation metrics can be found in Table 5.

Overall, there is some systematic enhancement of model performance from CMIP3 to CMIP5 or CORDEX. Wright et al. (2016), Koutroulis et al. (2016) and Ring et al. (2017) report similar results. However, especially for precipitation, RM and HM show almost identical performances of CMIP3 and CMIP5 with a neglectable advantage for CMIP3. This is in line with Li and Xie (2014) and Grose et al. (2014). In accordance with Flato et al. (2013) all metrics indicate improved temperature simulation of CMIP5 or CORDEX for most analyzed situations. The same tendencies were found by Wright et al. (2016) and Koutroulis et al. (2016). This is especially true for the smaller-scale Mediterranean sub-regions with CORDEX showing a remarkable added value to the multi model ensembles of GCMs (Jacob et al. 2014).

Even though the metrics show good potential for model evaluation, we found noticeable differences in their usability for weighting. RM and HM results rely on measured values based on a RMSE with open range. To some extent, the differences between these values are very small. Thus, differences of the model weights might be too small to generate distinct changes to the future emissions scenarios standard deviations and expected values or PDFs. Furthermore, a reasonable stretch to extend differences between the weights needs to be supported by additional information that cannot be provided by the evaluation alone. Therefore, we consider RM and HM as suitable performance metrics but suggest the fingerprinting approaches as weighting tool. Here, the range of model performances is defined from -1 to 1. On this level, the same problems might appear as for RM and HM. But an introduced threshold at 0, sorting out weaker models led to significant changes in PDFs of future climate change. However, this approach is merely a first conservative adjustment for stron-

Tab. 5: Comparison of applied performance metrics

Metric	Evaluation characteristics	Advantages	Disadvantages
RM	Absolute trend bias	- easy to apply - good accordance to other evaluation results	- rather superficial evaluation
HM	Harmonic time series similarities	- so far unexplored climate characteristic - insensitivity to different observational data sets - no correlation to other evaluation results	- complex approach - relevance of characteristic needs further investigation
SUB	Spatial pattern of observed climate trend	- easy to apply - good accordance to other evaluation results	- rather superficial evaluation
OPT	Spatial pattern of filter climate change signal	- Sophisticated performance metric based	- complex approach - data reduction is necessary - estimation of natural variability

ger model weights. This kind of enhancement of the effects of evaluation metrics (even for RM and HM) has to be considered a topic for further studies. For temperature, SUB and OPT indicate very good performance for almost all models. Therefore, there is a larger potential of improvement for weighting the simulated precipitation. Here, the SUBi and OPTi weighting of PDFs and MSKs leads to both increases and decreases of model spread (uncertainty) over all regional and seasonal situations. Overall, decreased uncertainty clearly prevails. Nevertheless, based on our results, every situation (region, season and scenario) needs to be evaluated individually to get a valid result. Generalizations of results should be avoided. This is true for both, the model evaluation as well as for the weighting impact on the multi model ensemble. A single model which outperforms the others in all or even most situations was not found, a conclusion also confirmed by Gleckler et al. (2008), Power et al. (2012) and Ring et al. (2016). Furthermore, we have to consider, that climate models are not independent. For those models which have multiple realizations, instead of using just one simulation, we evaluated each run and calculated the mean over all runs to reduce their weight in the multi model ensemble and to consider their variability. The independence between different models remains a technical challenge since most models share at least some basic components (Herger et al. 2018). Nevertheless, these climate models are the best source of information of the future climate and the evaluation results from our study still indicate substantial differences across the models.

This study has been performed in context of the COMEPRO framework with distinct regions and model data. This allows for a detailed comparison to the results of different performance metrics. Ring et al. (2017) applied six evaluation metrics based on 2x2 contingency tables (CT) for the 50-year trend and the climatological mean. The results of RM, SUB and OPT fit seamlessly to those of the trend. In fact, they show a high positive correlation with those of the CT approaches. On the other hand, there is no correlation with those of the climatological mean. Interestingly, there is no correlation of the HM results with neither 50-year trend nor mean. This supports the assumption that HM investigates a generally different aspect of climate model performance, underlining the need for further application and investigation.

7 Conclusions

In this study the use of four (RM, HM, SUB, OPT) different performance metrics for state-of-the-art global and regional climate models has been demonstrated. We analyzed their applicability and their results considering one global, six continental-scale study areas and eight smaller sub-regions in the Mediterranean area. Overall, three of four metrics show a high consistency in model rating. The fourth metric turned out to be a promising approach even though its results led to different model ratings. The investigated climate model parameter, spectral similarity of the time series, offers a new perspective on model performance. For the three other metrics, we see a high consistency for model evaluation and rating.

Overall, there is no model outperforming all the others. In fact, for many combinations of global regions and seasons, the older multi model ensemble CMIP3 appears to perform on a similar level as CMIP5. In general, there are only minor differences in model performances. For the sub-regions of the Mediterranean area we found mainly stronger results for the current CMIP5 ensemble and, particularly, the regional climate models of CORDEX. The results of this study underline that to focus on only one model - or even multi model ensemble – is not recommendable without a thorough evaluation of all available simulations. Further, we reach the same conclusion as Ring et al. (2017) that the climate characteristic is of much higher relevance than the type of metric. Since our results allow no obvious preference concerning a best ensemble for most situations, we suggest relying on detailed evaluations using multiple performance metrics to find the best simulation for the particular region and season of interest.

In terms of weighting, the applied metrics showed rather small differences between the original performance values. To achieve stronger effects of model weighting on probabilistic climate assessments, further steps like the introduction of a threshold to create a sub-ensemble is necessary. However, a general statement regarding the type of PDF change, increase or decrease of model spread, is not possible. Again, a detailed evaluation of the respective situation has to be performed for valid results.

Altogether, we see further need for comparing different climate model performance metrics. Especially with detecting the harmonic time series similarities as a new climate model characteristic to be evaluated, research goals in this field should be redefined from evaluation of general model performance to evaluation of specific model characteris-

tics. We further recommend using a wide variety of different evaluation approaches and weighting metrics tailored to specific situations and processes of interest. This study in combination with the results of Ring et al. (2017) offers a comprehensive insight in the performances of different specific characteristics for most state-of-the-art climate models and numerous metrics for a variety of study areas. Further studies could benefit from these results and use or extend the analyzed metrics to generate reliable assessments of potential future climate states.

Acknowledgements

We thank the Program for Climate Model Diagnosis and Intercomparison (PCMDI) and the World Climate Research Program (WCRP) for providing the CMIP3, CMIP5 and CORDEX data sets 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) under grants PA 1194/10-1 and Jucundus Jacobeit

References

Christensen, J. H.; Carter, T. R.; Rummukainen, M. and Amanatidis, G. (2007): Evaluating the performance and utility of regional climate models: the PRUDENCE project. In: Climatic Change 81 (1), 1–6. https://doi.org/10.1007/s10584-006-9211-6

Diffenbaugh, N. S. and Giorgi, F. (2012): Climate change hotspots in the CMIP5 global climate model ensemble. In: Climatic Change 114 (3–4), 813–822. https://doi.org/10.1007/s10584-012-0570-x

Dittus, A. J.; Karoly, D. J.; Lewis, S. C.; Alexander, L. V. and Donat, M. G. (2016): A multiregion model evaluation and attribution study of historical changes in the area affected by temperature and precipitation extremes. In: Journal of Climate 29, 8285–8299. https://doi.org/10.1175/JCLI-D-16-0164.1

Donat, M. G.; Alexander, L. V.; Herold, N. and Dittus, A. J. (2016): Temperature and precipitation extremes in century-long gridded observations, reanalyses, and atmospheric model simulations. In: Journal of Geophysical Research: Atmospheres 121 (19), 11.17411.189. https://doi.org/10.1002/2016JD025480

Fekete, B. M.; Vörösmarty, C. J.; Roads, J. O. and Willmott, C. J. (2004): Uncertainties in precipitation and their impacts on runoff estimates. In: Journal of Climate 17, 294–304. https://doi.org/10.1175/1520-0442(2004)017<0294:UIPATI>2.0.CO;2

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, United Kingdom and New York, NY, USA.

Giorgi, F. (2006): Climate change hot-spots. In: Geophysical Research Letters 33 (8), L08707. https://doi.org/10.1029/2006GL025734

Giorgi, F.; Jones, C. and Asrar, G. (2009): Addressing climate information needs at the regional level: the CORDEX framework. In: WMO Bulletin 58 (3), 175–183.

Giorgi, F. and Mearns, 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: Journal of Climate 15, 1141–1158. https://doi.org/10.1175/1520-0442(2002)015<1141:COAURA>2.0.CO;2

Gleckler, P. J.; Taylor, K. E. and Doutriaux, C. (2008): Performance metrics for climate models. In: Journal of Geophysical Research 113, D6104. https://doi.org/10.1029/2007JD008972

Grose, M. R.; Brown, J. N.; Narsey, S.; et al. (2014): Assessment of the CMIP5 global climate model simulations of the western tropical Pacific climate system and comparison to CMIP3. In: International Journal of Climatology 34 (12), 3382–3399. https://doi.org/10.1002/joc.3916

Hasselmann, K. (1979): On the signal-to-noise problem in atmospheric response studies. In: Shaw, D. B and the Royal Meteorological Society (eds.): Meteorology over the tropical oceans: the main papers presented at a joint conference held 21 to 25 August 1978 in the rooms of the Royal Society, London, by the Royal Meteorological Society, the American Meteorological Society, the Deutsche Meteorologische Gesellschaft and the Royal Society. Bracknell, 251–259.

Hawkins, E. and Sutton, R. (2009): The potential to narrow uncertainty in regional climate predictions. In: Bulletin of the American Meteorological Society 90, 1095–1107. https://doi.org/10.1175/2009BAMS2607.1

Haylock, M. R.; Hofstra, N.; Klein Tank, A. M. G.; et al. (2008): A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006. In: Journal of Geophysical Research: Atmospheres 113, D20119. https://doi.org/10.1029/2008JD010201

Hegerl, G. C.; von Storch, H.; Hasselmann, K.; et al. (1996): Detecting greenhouse-gas-induced climate change with an optimal fingerprint method. In: Journal of Climate 9, 2281–2306. https://doi.org/10.1175/1520-0442(1996)009<2281:DGGICC>2.0.CO;2

Herger, N.; Abramowitz, G.; Knutti, R.; et al. (2018): Selecting a climate model subset to optimise key ensemble properties. In: Earth System Dynamics 9, 135-151. https://doi.org/10.5194/esd-9-135-2018

Hidalgo, H. G. and Alfaro, E. J. (2015): Skill of CMIP5 climate models in reproducing 20th century basic climate features in Central America. In: International Journal of Climatology 35 (12), 3397–3421. https://doi.org/10.1002/joc.4216

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 (2007): 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.

Jacob, D.; Petersen, J.; Eggert, B.; et al. (2014): EURO-CORDEX: new high-resolution climate change projections for European impact research. In: Regional Environmental Change 14 (2), 563–578. http://doi.org/10.1007/s10113-013-0499-2

Knutti, R. and Sedláček, J. (2012): Robustness and uncertainties in the new CMIP5 climate model projections. In: Nature Climate Change 3, 369–373. https://doi.org/10.1038/nclimate1716

Koutroulis, A. G.; Grillakis, M. G.; Tsanis, I. K. and Papadimitriou, L. (2016): Evaluation of precipitation and temperature simulation performance of the CMIP3 and CMIP5 historical experiments. In: Climate Dynamics 47 (5–6), 1881–1898. https://doi.org/10.1007/s00382-015-2938-x

Kumar, S.; Merwade, V.; Kinter, J. L.; et al. (2013): Evaluation of temperature and precipitation trends and long-term persistence in CMIP5 twentieth-century climate simulations. In: Journal of Climate 26, 4168–4185. https://doi.org/10.1175/JCLI-D-12-00259.1

Leduc, M.; Laprise, R.; de Elía, R. and Šeparović, L. (2016): Is institutional democracy a good proxy for model independence? In: Journal of Climate 29, 8301–8316. https://doi.org/10.1175/JCLI-D-15-0761.1

Li, G. and Xie, S.-P. (2014): Tropical biases in CMIP5 multimodel ensemble: the excessive equatorial Pacific cold tongue and double ITCZ problems. In: Journal of Climate 27, 1765–1780. https://doi.org/10.1175/JCLI-D-13-00337.1

Mitchell, T. D. and Jones, P. D. (2005): An improved method of constructing a database of monthly climate observations and associated high-resolution grids. In: International Journal of Climatology 25, 693–712. https://doi.org/10.1002/joc.1181

Moss, R. H.; Edmonds, J. A.; Hibbard, K. A.; et al. (2010): The next generation of scenarios for climate change research and assessment. In: Nature 463, 747–756. 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, U.S.A.

Nikulin, G.; Jones, C.; Giorgi, F.; et al. (2012): Precipitation climatology in an ensemble of CORDEX-Africa regional climate simulations. In: Journal of Climate 25, 6057–6078. https://doi.org/10.1175/JCLI-D-11-00375.1

Paeth, H. and Hense, A. (2001): Signal analysis of the atmospheric mean 500/1000 hPa temperature north of 55N between 1949 and 1994. In: Climate Dynamics 18 (3–4), 345–358. https://doi.org/10.1007/s003820100179

Paeth, H. and Mannig, B. (2013): On the added value of regional climate modeling in climate change assessment. In: Climate Dynamics 41 (3–4), 1057–1066. https://doi.org/10.1007/s00382-012-1517-7

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. In: Global and Planetary Change 151, 144–151. https://doi.org/10.1016/j.gloplacha.2016.03.003

Perkins, S. E.; Pitman, A. J.; Holbrook, N. J.; et al. (2007): Evaluation of the AR4 climate models' simulated daily maximum temperature, minimum temperature, and precipitation over Australia using probability density functions. In: Journal of Climate 20, 4356–4376. https://doi.org/10.1175/JCLI4253.1

Poli, P.; Hersbach, H.; Tan, D.; et al. (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 Report Series 14 (59). Shinfield Park, Reading. https://www.ecmwf.int/en/elibrary/11699-data-assimilation-system-and-initial-performance-evaluation-ecmwf-pilot-reanalysis

Power, S. B.; Delage, F.; Colman, R. and Moise, A. (2012): Consensus on twenty-first-century rainfall projections in climate models more widespread than previously thought. In: Journal of Climate 25, 3792–3809. https://doi.org/10.1175/JCLI-D-11-00354.1

Räisänen, J. and Ylhäisi, J. S. (2012): Can model weighting improve probabilistic projections of climate change? In: Climate Dynamics 39 (7–8), 1981–1998. https://doi.org/10.1007/s00382-011-1217-8

Randall, D. A.; Wood, R. A.; Bony, S.; et al. (2007): Climate models and their evaluation. In: Solomon, S.; Qin, D.; Manning, M.; et al. (eds.) Climate hange 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, United Kingdom and New York, NY, USA.

Reichler, T. and Kim, J. (2008): How well do coupled models simulate today's climate? In: Bulletin of the American Meteorological Society 89, 303–311. https://doi.org/10.1175/BAMS-89-3-303

Ring, C.; Mannig, B.; Pollinger, F. and Paeth, H. (2016): Uncertainties in the simulation of precipitation in selected regions of humid and dry climate. In: International Journal of Climatology 36, 3521–3538. https://doi.org/10.1002/joc.4573

Ring, C.; Pollinger, F.; Kaspar-Ott, I.; et al. (2017): A comparison of metrics for assessing state-of-the-art climate models and implications for probabilistic projections of climate change. In: Climate Dynamics 50 (5–6), 2087–2106. https://doi.org/10.1007/s00382-017-3737-3

Sanderson, B. M.; Knutti, R.; Caldwell, P.; et al. (2015): Addressing interdependency in a multimodel ensemble by interpolation of model properties. In: Journal of Climate 28, 5150–5170. https://doi.org/10.1175/JCLI-D-14-00361.1

Stainforth, D. A.; Aina, T.; Christensen, C.; et al. (2005): Uncertainty in predictions of the climate response to rising levels of greenhouse gases. In: Nature 433, 403–406. https://doi.org/10.1038/nature03301

Tebaldi, C. and Knutti, R. (2007): The use of the multi-model ensemble in probabilistic climate projections. In: Philosophical Transactions of the Royal Society A 365 (1857), 2053–2075. https://doi.org/10.1098/rsta.2007.2076

Wang, C.; Zhang, L.; Lee. S.-K.; et al. (2014): A global perspective on CMIP5 climate model biases. In: Nature Climate Change 4, 201–205. https://doi.org/10.1038/nclimate2118

Weigel, A. P.; Knutti, R.; Liniger, M. A. and Appenzeller, C. (2010): Risks of model weighting in multimodel climate projections. In: Journal of Climate 23, 4175–4191. https://doi.org/10.1175/2010JCLI3594.1

Wilks, D. S. (2006): Statistical methods in the atmospheric sciences. Amsterdam.

Wright, A. N.; Schwartz, M. W.; Hijmans, R. J. and 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. In: Climatic Change 134 (4), 579–591. https://doi.org/10.1007/s10584-015-1552-6

Authors

Dr. Christoph Ring
Dr. Felix Pollinger
Luzia Keupp
Prof. Dr. Heiko Paeth
Universität Würzburg
Institut für Geographie und Geologie
Am Hubland
97074 Würzburg
christoph.ring@uni-wuerzburg.de
felix.pollinger@uni-wuerzburg.de
luzia.keupp@uni-wuerzburg.de
heiko.paeth@uni-wuerzburg.de

Dr. Irena Kaspar-Ott
Prof. Dr. Elke Hertig
Prof. Dr. Jucundus Jacobeit
Augsburg University
Institute of Geography
Alter Postweg 118
86159 Augsburg
Germany
elke.hertig@geo.uni-augsburg.de
irena.kaspar-ott@geo.uni-augsburg.de
jucundus.jacobeit@geo.uni-augsburg.de