How well do CMIP5 Earth System Models simulate present climate conditions?

A performance comparison for the downscaling community

S. Brands · S. Herrera · J. Fernández · J.M. Gutiérrez

Received: date / Accepted: date

Abstract This study assesses the performance of seven Earth System Models (ESMs) from the Coupled Model Intercomparison Project Phase 5 in present climate conditions from a downscaling perspective. Two different reanalyses (ERA-Interim and JRA-25) are used as reference for an objective evaluation of circulation, temperature and humidity variables on daily timescale, which is based on distributional similarity scores. For use in statistical downscaling studies, ESM-performance on the grid-box scale is mapped over a large spatial domain covering Europe and Africa, additionally highlighting those regions where significant distributional differences remain even after correcting the mean error. For use in dynamical downscaling studies, performance is specifically assessed along the lateral boundaries of the 3 CORDEX domains defined for Europe, the Mediterranean Basin and Africa.

Since considerable differences between the reanalyses were found over central to south Africa, ESM-performance cannot be objectively assessed there. For the remaining regions, widespread ESM-errors, like a systematic warm bias in the middle-troposphere, too-strong wintertime westerlies over Europe and a too-weak African Easterly Jet during the monsoon season, were found. Particularly in the tropics, significant distributional differences remain after correcting the mean error. This implies that the limitations and recommendations for working with GCM data in a downscaling context remain valid for the new model generation. HadGEM2-ESM performs overly best and inter-model performance differences along the lateral boundaries of the Euro-CORDEX domain are smaller than for the Med-CORDEX and CORDEX-Africa domains. Thus, choosing the appropriate driving GCM is of particular importance for the latter two projects.

Keywords CMIP5 · Earth System Models · Performance · Present Climate · Downscaling

S. Brands
Instituto de Física de Cantabria (UC-CSIC), Santander, Spain
Tel.: +34-942-20-2064
E-mail: brandssf@unican.es

S. Herrera
Instituto de Física de Cantabria (UC-CSIC), Santander, Spain

J. Fernández Dept. of Applied Mathematics and Comp. Sci., Universidad de Cantabria, Santander, Spain · J.M. Gutiérrez
Instituto de Física de Cantabria (UC-CSIC), Santander, Spain

1 Introduction

At the onset of the Coupled Model Comparison Project Phase 5 (CMIP5), a new generation of General Circulation Models (GCMs) has become available to the scientific community. In comparison to the former model generation, these 'Earth System Models' (ESMs) incorporate additional components describing the atmosphere's interaction with land-use and vegetation, as well as explicitly taking into account atmospheric chemistry, aerosols and the carbon cycle (Taylor et al, 2011). The new model generation is driven by newly defined atmospheric composition forcings —the 'historical forcing' for present climate conditions and the 'Representative Concentration Pathways' (RCPs, Moss et al, 2010) for future scenarios.— The dataset resulting from these global simulations will be the mainstay of future climate change studies and is the baseline of the Fifth Assessment Report of the Intergovernmental Panel on Climate Change (AR5). Moreover, this dataset is the

starting point of different regional downscaling initiatives on the generation of regional climate change scenarios, which are being coordinated worldwide for the first time within the framework of the COrdinated Regional Climate Downscaling EXperiment (CORDEX) (Jones et al, 2011). These initiatives use both dynamical and statistical downscaling approaches to provide high-resolution information over a specific region of interest (e.g. Europe or Africa) at the spatial scale required by many impact studies (Winkler et al, 2011b,a). This is done by either running a Regional Climate Model (RCM), driven by GCM data at its lateral boundaries, or by applying empirical relationships, usually found between large-scale reanalysis- and small-scale station data, to GCM output (Giorgi and Mearns, 1991).

In this study we provide a comprehensive evaluation of the new GCM generation from a downscaling perspective, taking into account the requirements of both statistical and dynamical approaches. To this aim, we test the ability of seven ESMs to reproduce present-day climate conditions as represented by reanalysis data, which is hereafter referred to as 'performance' (Giorgi and Francisco, 2000). Apart from validating 'classical' near surface variables, we focus on middle-tropospheric circulation, temperature and humidity variables, which are of particular importance for the purpose of downscaling (Fernández et al, 2007; Maraun et al, 2010; Sauter and Venema, 2011; Brands et al, 2012). All variables are assessed on daily timescale. To provide information tailored to the statistical approach, ESM-performance at the grid-box scale is mapped on a large spatial domain covering Europe and Africa. Specific information for the dynamical approach is provided by assessing ESM-performance along the lateral boundaries of the 3 domains used in the Euro-CORDEX (Europe, `http://www.euro-cordex.net`), Med-CORDEX (the Mediterranean area, `http://www.medcordex.eu`) and CORDEX Africa (`http://start.org/cordex-africa`) projects. The validation approach used in this paper is based on distributional similarity scores (Brands et al, 2012), taking reanalysis data as the references baseline. Moreover, the degree of reanalysis uncertainty is compared to the errors of the ESMs, thereby detecting those regions where a ranking of the latter is essentially impossible due to considerable reanalysis uncertainty (Gleckler et al, 2008).

Our results are expected to be of value because errors committed on the large scale (i.e. by the global models) are known to drive down the downscaling chain and affect the small-scale outputs (Timbal et al, 2003; Deque et al, 2007; Charles et al, 2007; Brands et al, 2011b) which are commonly used as input variables for impact models (Bedia et al, submitted). Quantification of the model error is important, since poor performance in present climate conditions is associated with outlier-like behavior in the scenario period, i.e. with a considerable deviation of the model's projected signal from some reference multi-model mean (Raisanen, 2007; Knutti et al, 2010). Perhaps most importantly, little to no information on the relative performance of the available driving CMIP5 ESMs is available at a time the downscaling community has to choose on which ESMs to rely on, a lack of knowledge which we intent to fill with the present study. Our approach provides a general overview on ESM-performance on hemispheric to continental scale and, as such, is not meant to replace studies on the synoptic-scale performance (Maraun et al, in print).

2 Data

The study area considered in this work is shown in Fig. 1 and covers the western old World extending from the Arctic to South Africa and from the Central Atlantic to the Ural Mountain Range and Arabic Peninsula, covering the Euro-CORDEX, Med-CORDEX and CORDEX Africa domains.

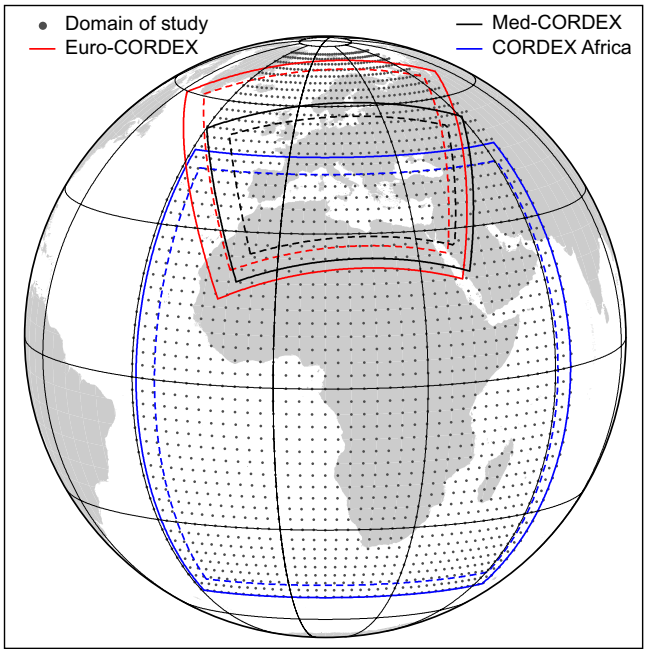

Fig. 1 Geographical domain considered in the study (black dots) and CORDEX exterior (solid) and interior (dashed) domains (in colors) used for the lateral boundary conditions in the Euro-CORDEX, Med-CORDEX and CORDEX Africa domains.

We consider data from the seven ESMs listed in Tab. 1, which were obtained from the Earth System Grid Federation (ESGF) gateways of the German Climate Computing Center (http://ipcc-ar5.dkrz.de), the Program for Climate Model Diagnosis and Intercomparison (http://pcmdi3.llnl.gov), and the British Atmospheric Data Center (http://cmip-gw.badc.rl.ac.uk). Since we evaluate performance in present climate conditions, we considered the CMIP5 experiment number '3.2 historical' (Taylor et al, 2011), which provides simulations of the recent past (1850-2005). This new generation of control runs is forced by observed atmospheric composition changes of both natural and anthropogenic nature. The first historical run of the available ensemble was chosen for the variables listed in Table 2. These variables are standard predictors in statistical downscaling studies (Hanssen-Bauer et al, 2005; Cavazos and Hewitson, 2005), and they are also taken into account (except 2m air temperature, T2) for defining the lateral boundary conditions in the process of nesting a Regional Climate Model (RCM) into a global one. The period under study is 1979-2005 and in case daily mean values where not already provided by the original data, they were calculated upon 6-hourly instantaneous values.

As reference data for validation, we consider the European Centre for Medium Range Weather Forecasts ERA-interim (Dee et al, 2011) and the Japanese Meteorological Agency JRA-25 reanalysis data (Onogi et al, 2007). Due to distinct native horizontal resolutions (see Table 1), both reanalysis and ESM-data were regridded to a regular 2.5° grid by using bilinear interpolation, which is a common step in downscaling- and GCM-performance studies.

Table 2 Variables considered in this study.

Code	Name	Height	Unit
Z	Geopotential	500hPa	$m^2 s^{-2}$
T	Temperature	2m, 850hPa	K
Q	Specific humidity	850hPa	$kg\,kg^{-1}$
U	U-wind	850hPa	$m\,s^{-1}$
V	V-wind	850hPa	$m\,s^{-1}$
SLP	Sea-level pressure	mean sea-level	Pa

3 Methods

The first measure for evaluating reanalysis uncertainty and ESM-performance in this study is the mean difference (bias). Note that the variability of the applied daily variables is much larger in the tropics than in the mid-latitudes and that it additionally varies from one season to another. Thus, to make results comparable, the bias is normalized by the standard deviation of ERA-Interim for each season and grid-box (Brands et al, 2011b).

To detect differences/errors in the higher order moments of the distribution, we apply the two-sample Kolmogorov-Smirnov test (KS-test) on the unbiased/anomaly data, the latter being obtained by subtracting the seasonal mean from each timestep. The KS-test is a non-parametric hypothesis test assessing the null hypothesis (H_0) that two candidate samples (e.g. reanalysis and ESM series for a particular gridbox) come from the same underlying theoretical probability distribution. It is defined by the following statistic:

$$KS\text{–}statistic = \max_{i=1}^{2n} |E(z_i) - I(z_i)| \tag{1}$$

where n is the length of the time-series, E and I are the empirical cumulative frequencies from a given ESM (or JRA25, in case reanalysis uncertainty is assessed) and the ERA-Interim reanalysis, which serves as reference for validation in any case. Moreover, z_i denotes the i-th data value of the sorted joined sample. This statistic is bounded between zero and one, with low values indicating distributional similarity. In this study we use the p-value of this statistic as a measure of distributional similarity. Thus, decreasing values indicate an increasing confidence on distributional differences between both series. Note that a base 10 logarithmic transformation is applied to the p-values in order to better indicate the different significance levels, 10^{-1}, 10^{-2}, 10^{-3}, corresponding to increasing confidences (90, 99, 99.9% respectively) on the dissimilarity of the distributions.

Since the daily time series applied here are serially correlated, we calculate their effective sample size n^* before estimating the p-value of the KS-statistic in order to avoid committing too many type-one errors (i.e. erroneous rejections of the H_0). Under the assumption that the underlying time-series follow a first-order autoregressive process, n^* is defined as follows (Wilks, 2006):

$$n^* = n\frac{1 - p_1}{1 + p_1} \tag{2}$$

where n is the sample size, n^* is the effective sample size, and p_1 is the lag-1 autocorrelation coefficient.

Reanalysis uncertainty is assessed by validating the variables from JRA-25 against those from ERA-Interim. Note that due to the lack of observational datasets for free-tropospheric variables on daily timescale, the difference between two distinct reanalysis is a reasonable estimator of observational uncertainty. If a close agreement is found, the reanalyses are likely driven by the

Table 1 CMIP5 Earth System Models considered in this study.

Model	Hor. Resolution	Reference
CanESM2	$2.8 * 2.8°$	Chylek et al (2011)
CNRM-CM5	$1.4 * 1.4°$	Voldoire et al (2011)
HadGEM2-ES	$1.875 * 1.25°$	Collins et al (2011)
IPSL-CM5-MR	$1.5 * 1.27°$	Dufresne et al (submitted)
MIROC-ESM	$2.8 * 2.8°$	Watanabe et al (2011)
MPI-ESM-LR	$1.8 * 1.8°$	Raddatz et al (2007); Jungclaus et al (2010)
NorESM1-M	$1.5 * 1.9°$	Kirkevag et al (2008); Seland et al (2008)

assimilated observations, while in case of considerable differences at least one of them is dominated by internal model variability rather than observations and, hence, does not reflect reality (Sterl, 2004).

4 Results

In this section we first assess reanalysis uncertainty (comparing JRA-25 with ERA-Interim) and then evaluate ESM-performance (comparing the ESMs with ERA-Interim). The bias is applied to assess reanalysis differences/ESM errors in the mean of the distribution. Then, to detect errors in higher order moments, we apply the KS-test to the anomaly/unbiased data. Note that removing the bias is a common step in statistical downscaling approaches Wilby et al (2004) and, albeit it destroys the non-linear relationships between the atmospheric variables (Laprise, 2008), has been recently proposed for the dynamical approach as well (Zhongfeng and Zong-Liang, in print). Finally, model performance for the original (i.e. uncorrected) data is specifically assessed along the lateral boundaries of the three CORDEX domains defined in Fig. 1, which is of particular interest for the dynamical donwscaling community. Unless RCMs are nudged to the large scale information (von Storch et al, 2000), ESM-performance in the interior of the aforementioned domains is less important for the purpose of dynamical downscaling, since the corresponding atmospheric variability is simulated by the RCMs.

4.1 Reanalysis Uncertainty

In the first and third column of Fig. 2, the normalized mean differences (hereafter BIASstd) between JRA-25 and ERA-Interim are mapped for SLP, T2, T850, Q850, U850, V850, T500 and Z500 (from top to bottom) for boreal winter (first column) and summer (third column). The direction and strength of the bias is given by the figure's colorbar. In the second and forth column, we display the logarithm to base 10 of the KS-statistic's p-value, which we obtain from applying the KS-test to the anomaly (bias-corrected) data. Values below -1.301 indicate significant distributional differences ($\alpha = 0.05$), whereas values above this threshold document that the H_0 of equal distributions cannot be rejected. The latter will hereafter be referred to as 'perfect' distributional similarity. A grid box is marked with a black dot if significant distributional differences for the original data disappear after removing the bias, i.e. in case reanalysis uncertainty is restricted to the mean of the distribution.

Reanalysis uncertainty for SLP is negligible north of $45°N$ and clearly depends on season in the Northern Hemisphere subtropics ($25°N - 45°N$), where it is most (less) pronounced in JJA (DJF). Over Africa (and especially in JJA), SLP from JRA-25 is much lower than SLP from ERA-Interim, while the opposite is the case over the adjacent ocean areas. Consequently, JRA-25 is characterized by a more pronounced sea-land pressure gradient than ERA-Interim. For the Southern and Northern Hemisphere mid-latitude oceans, reanalysis differences are negligible.

Reanalysis uncertainty for T2 is more widespread than for any other variable under study. Except for land areas north of $45°N$ during DJF and MAM (the latter not shown), where difference are negligible at most grid-boxes, JRA-25 is systematically warmer than ERA-Interim.

As was the case for SLP, reanalysis uncertainty for T850 is most pronounced over Africa and negligible over the the Northern-Hemisphere extratropics (with the exception of the Scandinavian Mountains in DJF and Greenland in all seasons). For the Intertropical Convergence Zone (ITCZ), JRA-25 is considerably warmer than ERA-Interim, while the opposite is the case for the large-scale subsidence zones. Interestingly, the resulting meridional tripole structure (JRA-25 colder,JRA-25 warmer,JRA-25 colder) follows the seasonal march of the ITCZ.

The tripole difference structure found for T850, as well as its associated seasonal march, also appears in Q850. At the ITCZ, JRA-25 is dryer than ERA-Interim,

while the opposite is the case at the margins of the Hadley-Cell. Except for central-to-east Europe and the northern North Atlantic, differences for Q850 are remarkable over the whole study area.

For U850 and V850, reanalysis uncertainty is generally weaker than for the other variables under study and in the extratropics is confined to regions of high orography only. During the core of the monsoon season (JJA), U850 and V850 over West Africa are weaker in JRA-25 than in ERA-Interim, while over East-Africa the sign of the difference is more heterogenous.

Finally, although reanalysis uncertainty for Z500 is generally lower than for any other variable under study, considerable differences are found over the tropics and subtropics. Over Africa and the tropical Oceans, and especially during DJF and MAM, Z500 in JRA-25 is lower than in ERA-Interim. This leads to a generalized reduction of the latitudinal height/pressure gradient, which is most pronounced over the South Atlantic in the area of the St. Helen's High.

For SLP and Z500, reanalysis uncertainty can be completely removed by correcting the bias, whereas for T850 and T2, the area of significant distributional differences is reduced to Central Africa (Kongo Basin), where it follows the seasonal march of the ITCZ, as was the case for the original data (see Fig. 2, columns 1 and 3). For U and V at 850, the area of significant distributional differences is largely reduced as well, the remaining areas being confined to high-orography regions and, in case of V850, to the Guinea Coast (with a widespread error in JJA, i.e. during the core of the summer monsoon). For Q850, distributional differences in the extratropics can be essentially removed by correcting the bias, while large areas of significant differences remain over the South Atlantic, Tropical Africa and, with a considerable error magnitude (i.e. low p-value), over the Indian Ocean.

As an anticipated conclusion to bear in mind when interpreting the results of the next section, assessing ESM-performance over central-to-South Africa is virtually impossible due to the large degree of reanalysis uncertainty. In the Northern Hemispheric extratropics, however, a performance-check is generally feasible since reanalysis uncertainty is generally negligible.

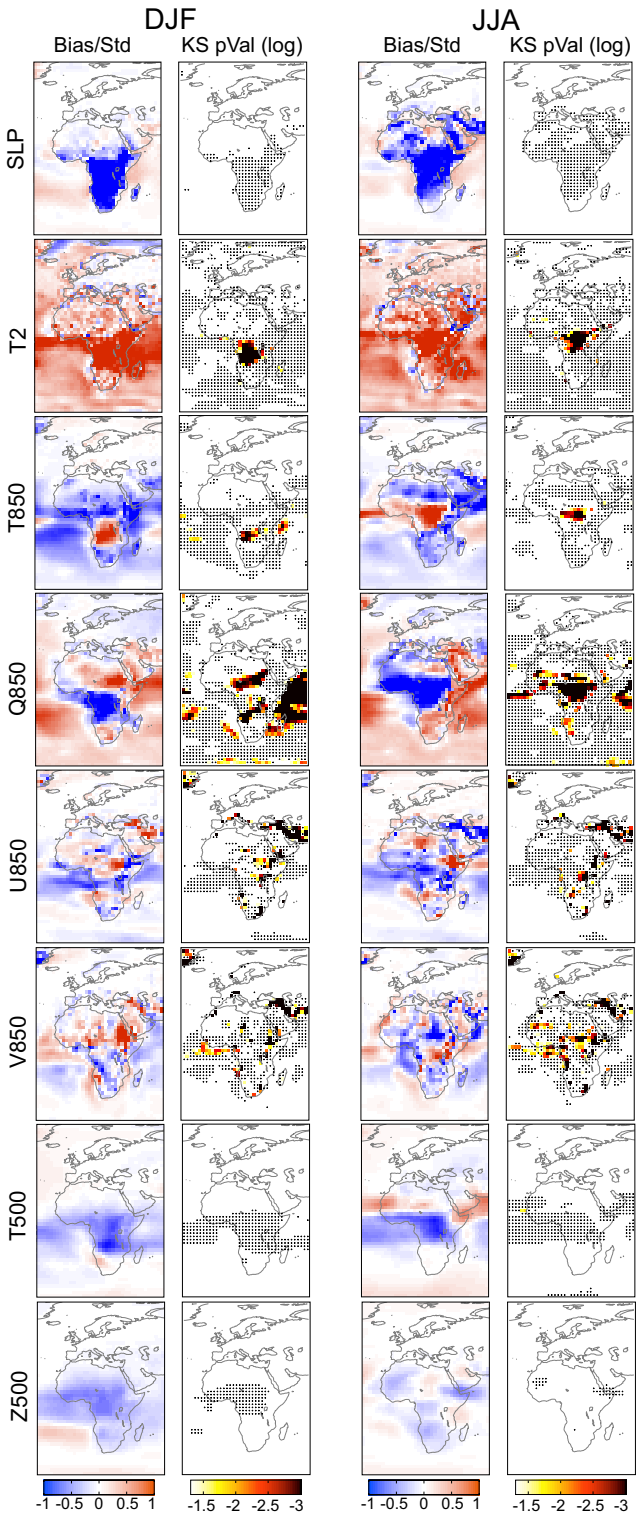

Fig. 2 Columns 1+3: Mean differences between JRA-25 and ERA-Interim, normalized by the standard deviation of the latter; Columns 2+4: P-value of the KS-test applied to the anomaly/unbiased data from JRA-25 and ERA-Interim in logarithmic scale. P-values are whitened if they do not exceed the threshold value of -1.301, i.e. if the distributional differences are not significant ($\alpha = 0.05$). Colour darkening corresponds to increasing (and significant) distributional differences/reanalysis uncertainties. Areas are shaded in black if significant distributional differences for the original reanalysis data are eliminated by removing the bias, results for **all applied variables**

4.2 Performance maps

Fig. 3 to 10 show the results of validating the 7 ESMs listed in Tab.1 against ERA-Interim for the case of SLP, T2, T850, Q850, U850, V850, Z500 and T500 respectively. Columns 1 and 2 (3 and 4) refer to the results for DJF (JJA). For each season we show the bias

normalized by the standard deviation of ERA-Interim Bias/Std, as well as the logarithmic p-value of the KS-statistic as described above.

Regarding the ESM-error for SLP (see Fig. 3), a largely exaggerated Northern-Hemispheric (NH) latitudinal pressure gradient is found for CanESM2, IPSL-CM5A-MR, MIROC-ESM, MPI-ESM-LR and NorESM1-M during DJF and MAM (the latter not shown). CanESM2 and CNRM-CM5 suffer from a negative bias over almost the entire domain except the North Atlantic. For MIROC-ESM, MPI-ESM-LR and NorESM1-M, and in the light of considerable reanalysis uncertainty, both the Sahara Heat Low and the St. Helen's High are too weak during JJA, leading to an underestimation of the westerly/monsoonal winds in the Sahel in these models. During the same season, SLP over the North Atlantic is overestimated in all ESMs except MPI-ESM-LR, the latter showing a slight underestimation.

The largely exaggerated latitudinal pressure gradient during boreal winter and spring is associated with too-strong westerlies in the Northern Hemisphere mid-latitudes, as is reflected by U850 in Fig. 7. This, in turn, is associated with an exaggerated advection of oceanic air masses, leading to too-mild and too-moist conditions in continental Europe, an effect that extends throughout the whole planetary boundary layer (see Fig.4 to 6 for T2, T850 and Q850 respectively). At 2m, the temperature bias is generally larger and more widespread than at 850hPa (compare Fig. 4 to Fig. 5). During the core of the West African monsoon (JJA), and as revealed by U500 (not shown), a two-strong Subtropical Jet, as well as a two-weak African Easterly Jet (Cook, 1999) are simulated by the ESMs, with NorESM1-M performing best for these features. Note that the bias for the zonal winds at 850 hPa is generally greater and more widespread than that of the meridional winds (compare Fig. 7 to Fig. 8). For all ESMs except IPSL-CM5A-MR, a cold bias was found in the middle troposhere (see Fig. 9), which covers a large fraction of the domain under study in any season and, with the exception of CanESM2 and IPSL-CM5A-MR, is associated with an underestimation of the geopotential at 500 hPa over the Tropics (see Fig. 10).

For all applied variables ESM-performance largely improves after removing the bias (see columns 2 and 4 in Fig. 3 to Fig. 10). In case of SLP, errors in higher order moments are detected over the high-orography regions of the Middle-East (for CanESM2, IPSL-CM5-MR and MIROC-ESM in at least one season of the year), over the Red-Sea and adjacent land areas (MIROC-ESM in JJA and SON), the Mediterranean (MIROC-ESM, NorESM1-M and MPI-ESM-LR in JJA), South Africa (CanESM2, IPSL-CM5-MR and MIROC-ESM in SON and/or DJF) and West Africa (CNRM-CM5 in JJA). Best overall performance is yielded for HadGEM2-ES, which, at least in case of SLP, does not suffer from errors in higher order moments at all.

In case of the unbiased T850 data (see Fig. 5), any ESM except CanESM2 and HaDGEM2-ES suffers from significant distributional differences over the tropics, the Southern-Hemisphere subtropics and the North Atlantic, while errors for T2 (see Fig. 4) are more widespread and additionally cover the Southern Hemisphere mid-latitudes. Interestingly, HadGEM2-ES again overly outperforms any other ESM for both T850 and T2, the performance of CanESM2 being comparable in case of T850.

For the unbiased U850 and V850 data (see Fig. 7 and 8), performance is generally better for U850. Errors in higher order moments appear over the tropics and subtropics. Large inter-model differences are found for both variables, with HadGEM2-ES and IPSL-CM5-MR performing clearly better than the remaining ESMs.

Albeit the errors in T500 are largely reduced by removing the bias, CanESM2, MIROC-ESM, and NorESM1-M suffer from errors in higher order moments along the ITCZ in JJA (see Fig. 9). For IPSL-CM5-MR, this error-type appears in DJF between the Azores and the Bay of Biscay.

As shown in Fig.10, ESM errors for Z500 disappear almost completely after removing the bias.

4.3 Performance along the lateral boundaries of the CORDEX domains

Fig. 11 displays the medians (bars) of the samples formed by the absolute normalized differences along the lateral boundaries (LB) of the 3 CORDEX domains shown int Fig 1. From top to bottom (left to right) the results for different variables (LBs) are shown, while the season-specific results are displayed within each panel (see x-axes). For reasons of simplicity, the interquartile ranges (IQRs) are not shown since they are proportional to their respective medians (i.e. the higher the median, the broader the IQR).

It is remarkable that ESM-performance along the lateral bouandaries of the 3 domains is generally very similar, i.e. the models do not perform systematically worse for the Med-CORDEX and CORDEX Africa domain than for Euro-COREX domain. For any domain under study, ESM-performance is best for V850, followed by U850, and is clearly worse for 2T (note the distinct scaling of the y-axis for the latter). Inter-model differences are most pronounced for Z500 and are generally larger for the Med-CORDEX and CORDEX Africa

domains than for the Euro-CORDEX domain. While MPI-ESM-LR and HADGEM2-ES are among the best models in any case, MIROC-ESM and IPSL-CM5-MR generally perform poorer, the remaining ESMs lying in-between in most cases.

5 Discussion and Conclusions

This study has shown that distributional differences between free-tropospheric circulation, temperature and humidity data from JRA-25 and ERA-Interim are generally lower than the differences found for the former generation of reanalysis products (Brands et al, 2012), especially if anomaly/unbiased data are taken into account.

In spite of this general reduction, reanalysis uncertainty over central-to-south Africa is frequently larger than the ESMs' error with respect to ERA-Interim, which violates the basic assumption of observational uncertainty being smaller than the model errors (Gleckler et al, 2008) and hinders an objective assessment of model performance in these regions. This should be taken into account when interpreting the results of recently published studies on the projected precipitation changes in South- and East-Africa (Shongwe et al, 2009, 2011). As model-errors in these regions are difficult to assess due to considerable observational uncertainties, possible artificial feedbacks in the GCM-scenario runs resulting from poor performance in present climate conditions (Raisanen, 2007) cannot be detected. Consequently, for these regions, any projection derived by the delta-method should be seen with caution.

In contrast, reanalysis uncertainty for the Northern Hemispheric extratropics is negligible, which permits for assessing ESM-performance there. A largely overestimated meridional pressure gradient in the North-Atlantic/European sector, leading to too mild and moist conditions in continental Europe, was found in 5 out of 7 ESMs during boreal winter and spring. This is in agreement with van Ulden and van Oldenborgh (2006) and Vial and Osborn (2011) who found serious circulation biases and an underestimation of the frequency and duration of wintertime atmospheric blocking in most CMIP3-GCMs.

The systematic cold bias in the middle troposphere found in this study is consistent with John and Soden (2007), who found similar results for the CMIP3-GCMs. Consequently, the above mentioned artificial feedback processes in the scenario period cannot be ruled out for Europe.

In 5 out of 7 models, the velocity of the monsoonal winds in West Africa (as represented by U850) is underestimated over the Sahel but overestimated over the sub-humid to humid zones along the Guinea Coast, a dipole error-pattern which was not reported for the CMIP3-GCMs (Kim et al, 2008). Similarly, the African Easterly Jet during the monsoon season (as represented by U500 in JJA) is underestimated by all ESMs analysed except NorESM1-M.

HadGEM2-ESM and MPI-ESM-LR outperform the remaining models along the lateral boundaries of the Euro-CORDEX, Med-CORDEX and CORDEX Africa domains, which is in qualitative agreement with Brands et al (2011a), who validated the former versions of these models for southwestern Europe. The systematic superiority of these models questions the paradigm of equiprobable treatment of the driving models in downscaling studies. Finally, ESM-performance at the lateral boundaries is not systematically worse for the CORDEX Africa and Med-CORDEX domain than for the EURO-CORDEX domain. However, choosing the 'right' ESM is more important for the former two domains, since inter-model performance differences are larger than for the latter one.

The final message is that many of the errors found in the CMIP3-GCMs are still present in current Earth System Models. Thus, the shortcomings and corresponding recommendations for working with GCM data in the context of downscaling (Wilby et al, 2004), i.e. taking into account and eventually correcting model errors as a precursor step of climate change impact studies (Bedia et al, submitted), remain valid for the new model generation.

Acknowledgements S.B. would like to thank the 'Consejo Superior de Investigaciones Científicas' for financial support. J.F. acknowledges financial support from the Spanish R&D&I programme through grant CGL2010-22158-C02 (CORWES project). All authors acknowledge and appreciate the free availability of the ERA-Interim and JRA-25 reanalysis datasets, as well as the GCM datasets provided by the ESG web portals.

References

Bedia J, Herrea S, Gutiérrez J (submitted) Dangers of projecting future species distributions with defective present-day baseline climatology. Global Planet Change

Brands S, Herrera S, San-Martin D, Gutierrez JM (2011a) Validation of the ENSEMBLES global climate models over southwestern Europe using probability density functions, from a downscaling perspective. Clim Res 48(2-3):145–161, DOI {10.3354/cr00995}

Brands S, Taboada JJ, Cofino AS, Sauter T, Schneider C (2011b) Statistical downscaling of daily tempera-

tures in the NW Iberian Peninsula from global climate models: validation and future scenarios. Clim Res 48(2-3):163–176, DOI {10.3354/cr00906}

Brands S, Gutiérrez J, Herrera S, Cofiño A (2012) On the use of reanalysis data for downscaling. J Clim DOI {10.1175/JCLI-D-11-00251.1}

Cavazos T, Hewitson B (2005) Performance of NCEP-NCAR reanalysis variables in statistical downscaling of daily precipitation. Clim Res 28:95–107

Charles SP, Bari MA, Kitsios A, Bates BC (2007) Effect of GCM bias on downscaled precipitation and runoff projections for the Serpentine catchment, Western Australia. Int J Climatol 27(12):1673–1690, DOI {10.1002/joc.1508}

Chylek P, Li J, Dubey M, Wang M, Lesins G (2011) Observed and model simulated 20th century arctic temperature variability: Canadian earth system model canesm. Atmos Chem Phys Discuss 11: 22,8932290, DOI {10.5194/acpd-11-22893-201}

Collins WJ, Bellouin N, Doutriaux-Boucher M, Gedney N, Halloran P, Hinton T, Hughes J, Jones CD, Joshi M, Liddicoat S, Martin G, O'Connor F, Rae J, Senior C, Sitch S, Totterdell I, Wiltshire A, Woodward S (2011) Development and evaluation of an Earth-System model-HadGEM2. GMD 4(4):1051–1075, DOI {10.5194/gmd-4-1051-2011}

Cook K (1999) Generation of the African easterly jet and its role in determining West African precipitation. J Clim 12(5, Part 1):1165–1184, DOI {10.1175/1520-0442(1999)012⟨1165:GOTAEJ⟩2.0.CO;2}

Dee et al. (2011) The ERA-Interim reanalysis: configuration and performance of the data assimilation system. Quart J R Met Soc 137(656, Part a):553–597, DOI {10.1002/qj.828}

Deque M, Rowell DP, Luethi D, Giorgi F, Christensen JH, Rockel B, Jacob D, Kjellstrom E, de Castro M, van den Hurk B (2007) An intercomparison of regional climate simulations for Europe: assessing uncertainties in model projections. Clim Chang 81(1):53–70, DOI {10.1007/s10584-006-9228-x}

Dufresne JL, Foujols MA, Denvil S, Caubel A, Marti O (submitted) Climate change projections using the IPSL-CM5 earth system model: from CMIP3 to CMIP5. Clim Dyn

Fernández J, Montávez JP, Saénz J, González-Rouco JF, Zorita E (2007) Sensitivity of the MM5 mesoscale model to physical parameterizations for regional climate studies: Annual cycle. J Geophys Res Atmos 112(D4), DOI {10.1029/2005JD006649}

Giorgi F, Francisco R (2000) Uncertainties in regional climate change prediction: a regional analysis of ensemble simulations with the HadCM2 coupled AOGCM. Clim Dyn 16(2-3):169–182, DOI {10.1007/PL00013733}

Giorgi F, Mearns L (1991) Approaches to the simulation of regional climate change - A review. Rev. Geophys 29(2):191–216

Gleckler PJ, Taylor KE, Doutriaux C (2008) Performance metrics for climate models. J Geophys Res Atmos 113(D6), DOI {10.1029/2007JD008972}

Hanssen-Bauer I, Achberger C, Benestad R, Chen D, Forland E (2005) Statistical downscaling of climate scenarios over Scandinavia. Clim Res 29(3):255–268

John VO, Soden BJ (2007) Temperature and humidity biases in global climate models and their impact on climate feedbacks. Geophys Res Lett 34(18), DOI {10.1029/2007GL030429}

Jones C, Giorgi F, Asrar G (2011) The Coordinated Regional Downscaling Experiment: CORDEX an international downscaling link to CMIP5

Jungclaus JH, Lorenz SJ, Timmreck C, Reick CH, Brovkin V, Six K, Segschneider J, Giorgetta MA, Crowley TJ, Pongratz J, Krivova NA, Vieira LE, Solanki SK, Klocke D, Botzet M, Esch M, Gayler V, Haak H, Raddatz TJ, Roeckner E, Schnur R, Widmann H, Claussen M, Stevens B, Marotzke J (2010) Climate and carbon-cycle variability over the last millennium. Clim Past 6(5):723–737, DOI {10.5194/cp-6-723-2010}

Kim HJ, Wang B, Ding Q (2008) The global monsoon variability simulated by CMIP3 coupled climate models. J Clim 21(20):5271–5294, DOI {10.1175/2008JCLI2041.1}

Kirkevag A, Iversen T, Seland O, Debernard JB, Storelvmo T, Kristjansson JE (2008) Aerosol-cloud-climate interactions in the climate model CAM-Oslo. Tellus A 60(3):492–512, DOI {10.1111/j.1600-0870.2008.00313.x}

Knutti R, Furrer R, Tebaldi C, Cermak J, Meehl GA (2010) Challenges in combining projections from multiple climate models. J Clim 23(10):2739–2758, DOI {10.1175/2009JCLI3361.1}

Laprise R (2008) Regional climate modelling. J Comput Phys 227(7):3641–3666, DOI {10.1016/j.jcp.2006.10.024}

Maraun D, Wetterhall F, Ireson AM, Chandler RE, Kendon EJ, Widmann M, Brienen S, Rust HW, Sauter T, Themessl M, Venema VKC, Chun KP, Goodess CM, Jones RG, Onof C, Vrac M, Thiele-Eich I (2010) Precipitation downscaling under climate change: recent developments to bridge the gap between dynamical models and the end user. Rev Geophys 48, DOI {10.1029/2009RG000314}

Maraun D, Osborn T, Rust H (in print) The influence of synoptic airflow on uk daily precipitation extremes. part II: climate model validation. Clim Dyn

Moss RH, Edmonds JA, Hibbard KA, Manning MR, Rose SK, van Vuuren DP, Carter TR, Emori S, Kainuma M, Kram T, Meehl GA, Mitchell JFB, Nakicenovic N, Riahi K, Smith SJ, Stouffer RJ, Thomson AM, Weyant JP, Wilbanks TJ (2010) The next generation of scenarios for climate change research and assessment. Nature 463(7282):747–756, DOI {10.1038/nature08823}

Onogi K, Tslttsui J, Koide H, Sakamoto M, Kobayashi S, Hatsushika H, Matsumoto T, Yamazaki N, Kaalhori H, Takahashi K, Kadokura S, Wada K, Kato K, Oyama R, Ose T, Mannoji N, Taira R (2007) The JRA-25 reanalysis. J Meteor Soc Jpn 85(3):369–432, DOI {10.2151/jmsj.85.369}

Raddatz TJ, Reick CH, Knorr W, Kattge J, Roeckner E, Schnur R, Schnitzler KG, Wetzel P, Jungclaus J (2007) Will the tropical land biosphere dominate the climate-carbon cycle feedback during the twenty-first century? Clim Dyn 29(6):565–574, DOI {10.1007/s00382-007-0247-8}

Raisanen J (2007) How reliable are climate models? Tellus A 59(1):2–29, DOI {10.1111/j.1600-0870.2006.00211.x}

Sauter T, Venema V (2011) Natural three-dimensional predictor domains for statistical precipitation downscaling. J Clim 24(23):6132–6145, DOI {10.1175/2011JCLI4155.1}

Seland O, Iversen T, Kirkevag A, Storelvmo T (2008) Aerosol-climate interactions in the CAM-Oslo atmospheric GCM and investigation of associated basic shortcomings. Tellus A 60(3):459–491, DOI {10.1111/j.1600-0870.2008.00318.x}

Shongwe ME, van Oldenborgh GJ, van den Hurk BJJM, de Boer B, Coelho CAS, van Aalst MK (2009) Projected changes in mean and extreme precipitation in Africa under global warming. Part I: southern Africa. J Clim 22(13):3819–3837, DOI {10.1175/2009JCLI2317.1}

Shongwe ME, van Oldenborgh GJ, van den Hurk B, van Aalst M (2011) Projected changes in mean and extreme precipitation in Africa under global warming. Part II: East Africa. J Clim 24(14):3718–3733, DOI {10.1175/2010JCLI2883.1}

Sterl A (2004) On the (in)homogeneity of reanalysis products. J Clim 17(19):3866–3873

von Storch H, Langenberg H, Feser F (2000) A spectral nudging technique for dynamical downscaling purposes. Mon Weather Rev 128(10):3664–3673, DOI {10.1175/1520-0493(2000)128⟨3664:ASNTFD⟩2.0.CO;2}

Taylor K, Stouffer R, Meehl G (2011) An overview of CMIP5 and the experiment design. Bull Am Meteor Soc DOI {10.1175/BAMS-D-11-00094.1}

Timbal B, Dufour A, McAvaney B (2003) An estimate of future climate change for western France using a statistical downscaling technique. Clim Dyn 20(7-8):807–823, DOI {10.1007/s00382-002-0298-9}

van Ulden A, van Oldenborgh G (2006) Large-scale atmospheric circulation biases and changes in global climate model simulations and their importance for climate change in Central Europe. Atmos Chem Phys 6:863–881

Vial J, Osborn J (2011) Assessment of atmosphere-ocean general circulation model simulations of winter northern hemisphere atmospheric blocking. Clim Dyn DOI {10.1007/s00382-011-1177-z}

Voldoire A, Sanchez-Gomez E, Salas y Mélia D, Decharme B, Cassou C (2011) The CNRM-CM5.1 global climate model: description and basic evaluation. Clim Dyn DOI {10.1007/s00382-011-1259-y}

Watanabe S, Hajima T, Sudo K, Nagashima T, Takemura T, Okajima H, Nozawa T, Kawase H, Abe M, Yokohata T, Ise T, Sato H, Kato E, Takata K, Emori S, Kawamiya M (2011) MIROC-ESM 2010: model description and basic results of CMIP5-20c3m experiments. GMD 4(4):845–872, DOI {10.5194/gmd-4-845-2011}

Wilby R, Charles S, Zorita E, Timbal B, Whetton P, Mearns L (2004) Guidelines for uses of climate scenarios developed from statistical downscaling methods. supporting material, http://www.narccap.ucar.edu/doc/tgica-guidance-2004.pdf

Wilks D (2006) Statistical methods in the atmospheric sciences, 2 edn. Amsterdam, Elsevier

Winkler JA, Guentchev GS, Liszewska M, Perdinan A, Tan PN (2011a) Climate scenario development and applications for local/regional climate change impact assessments: An overview for the non-climate scientist. Geography Compass 5(6):301–328, DOI 10.1111/j.1749-8198.2011.00426.x, URL `http://dx.doi.org/10.1111/j.1749-8198.2011.00426.x`

Winkler JA, Guentchev GS, Perdinan A, Tan PN, Zhong S, Liszewska M, Abraham Z, Niedzwiedz T, Ustrnul Z (2011b) Climate scenario development and applications for local/regional climate change impact assessments: An overview for the non-climate scientist. Geography Compass 5(6):275–300, DOI 10.1111/j.1749-8198.2011.00425.x, URL `http://dx.doi.org/10.1111/j.1749-8198.2011.00425.x`

Zhongfeng X, Zong-Liang Y (in print) An improved dynamical downscaling method with gcm bias corrections and its validation with 30 years of climate simulations. J Clim DOI {http://dx.doi.org/10.1175/JCLI-D-12-00005.1}

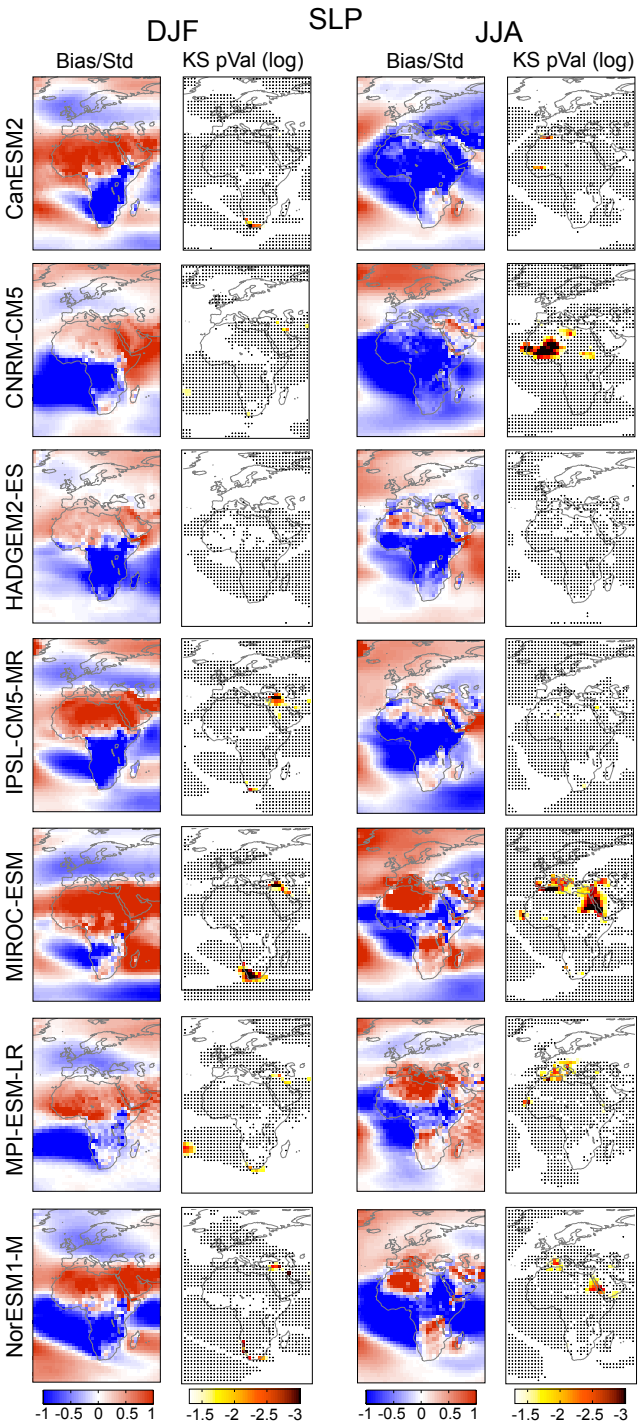

Fig. 3 Columns 1+3: Mean differences (columns 1+3) between the seven ESMs listed in Tab. 1 and ERA-Interim, normalized by the standard deviation of ERA-Interim; Columns: 2+4: P-value of the KS-test applied to anomaly model/reanlysis data data in logarithmic scale. P-values are whitened if they do not exceed the threshold value of -1.301, i.e. if the distributional differences are not significant ($\alpha = 0.05$). Colour darkening corresponds to increasing (and significant) distributional differences/ESM errors. Areas are shaded in black if significant ESM errors in the original data are eliminated by removing the bias; results for **SLP**

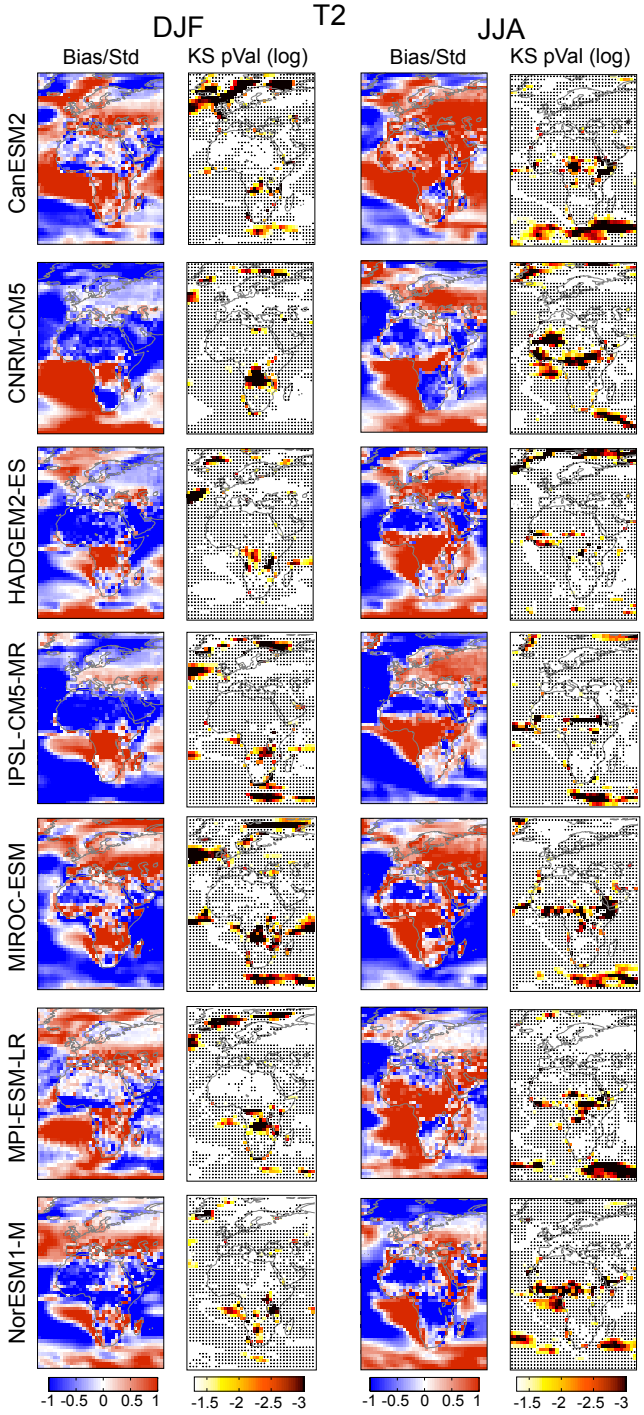

Fig. 4 As Fig. 3, but for **T2**

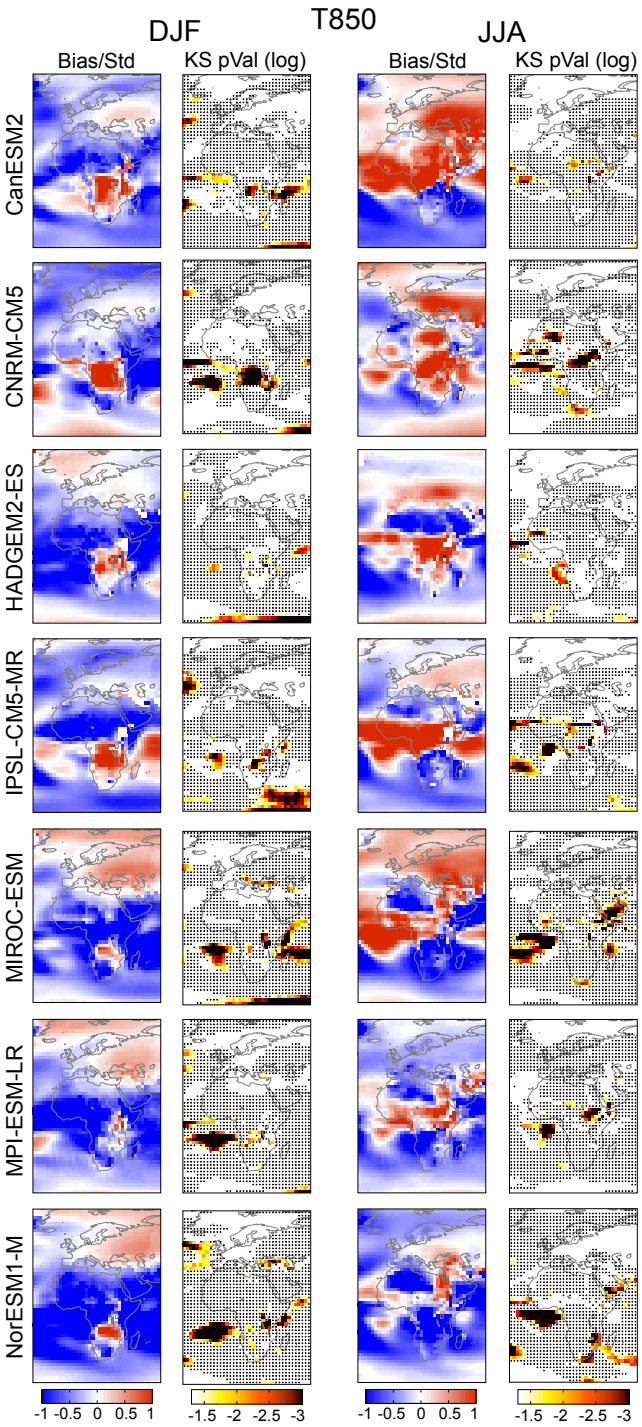

Fig. 5 As Fig. 3, but for **T850**

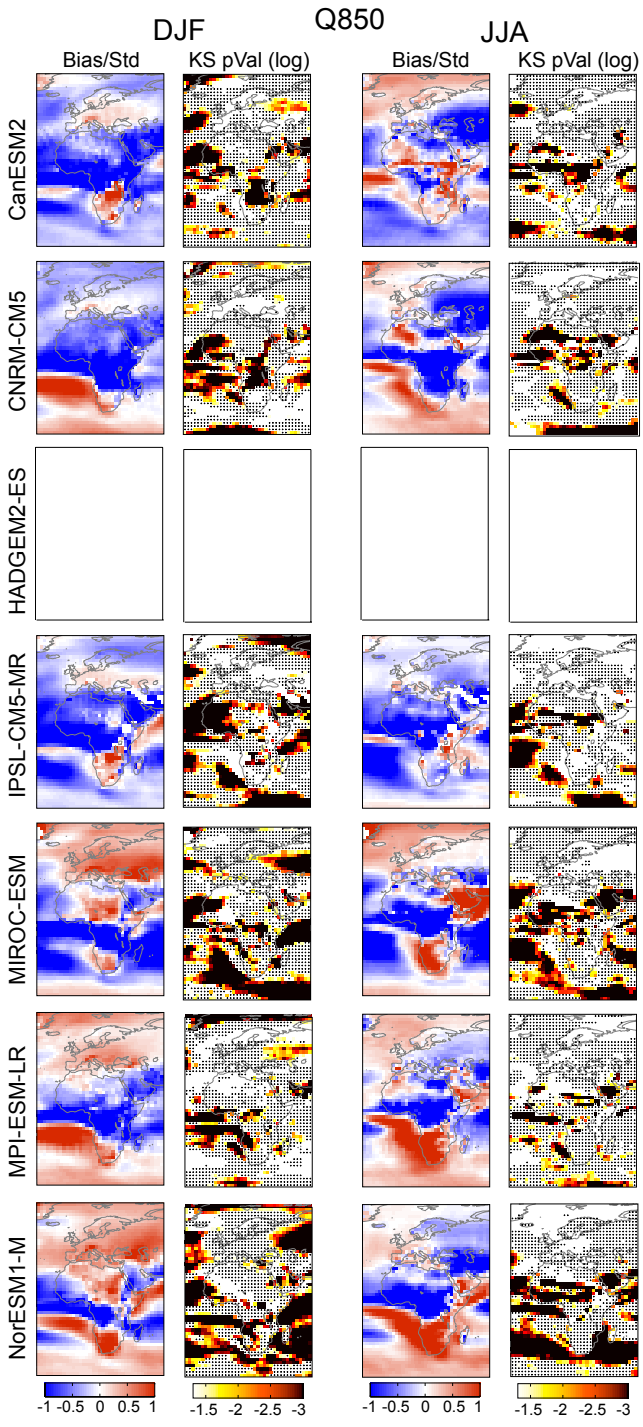

Fig. 6 As Fig. 3, but for **Q850**, empty panels refer to lack of data at the ESG-portals

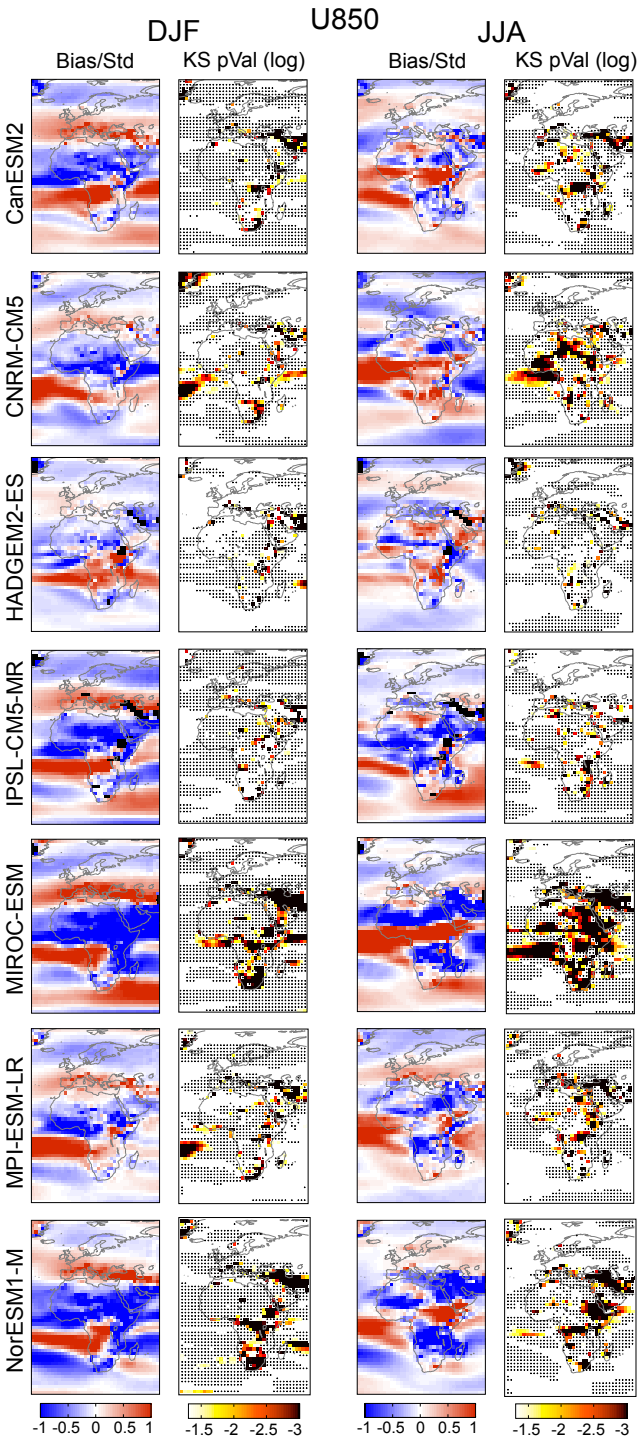

Fig. 7 As Fig. 3, but for **U850**

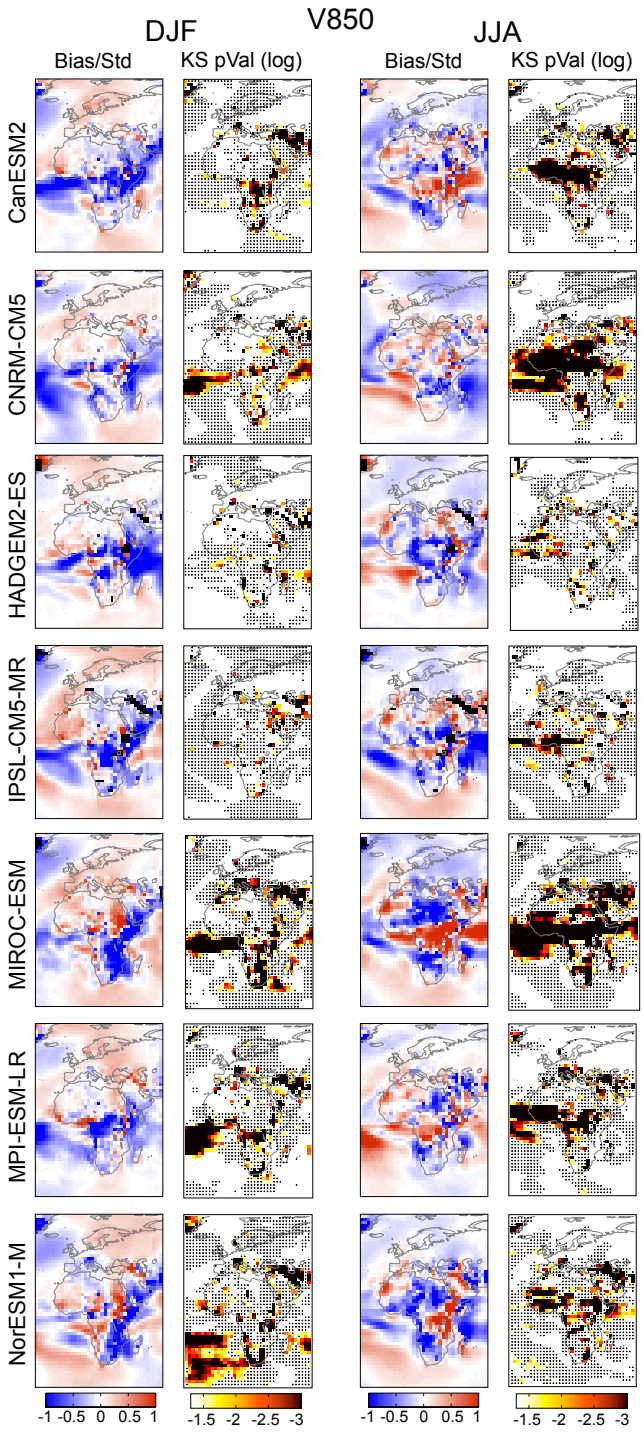

Fig. 8 As Fig. 3, but for **V850**

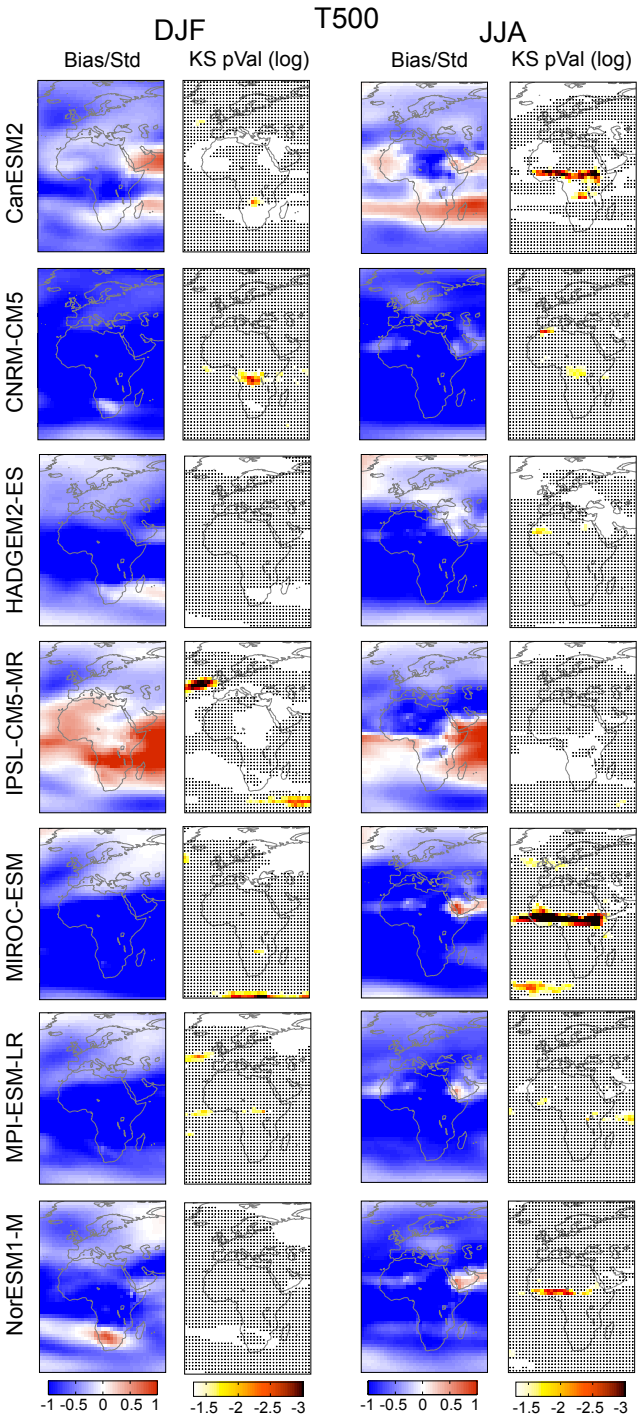

Fig. 9 As Fig. 3, but for **T500**

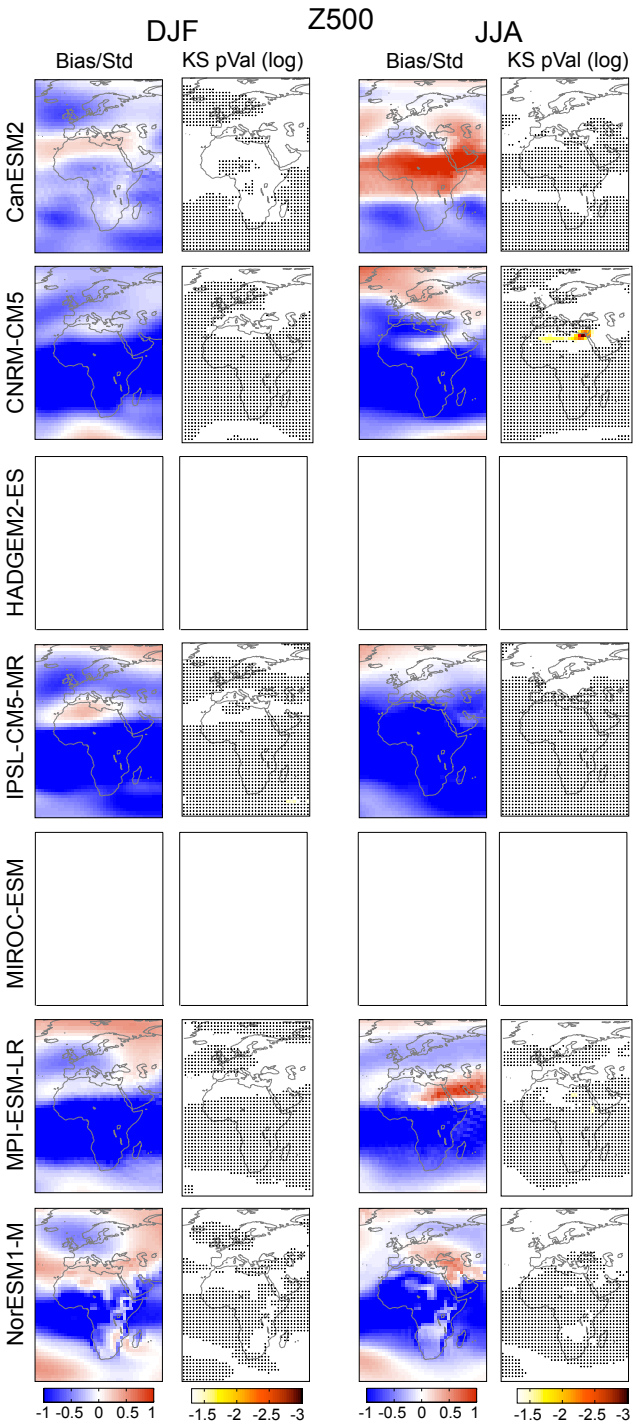

Fig. 10 As Fig. 3, but for **Z500**, empty panels refer to lack of data at the ESGG-portals

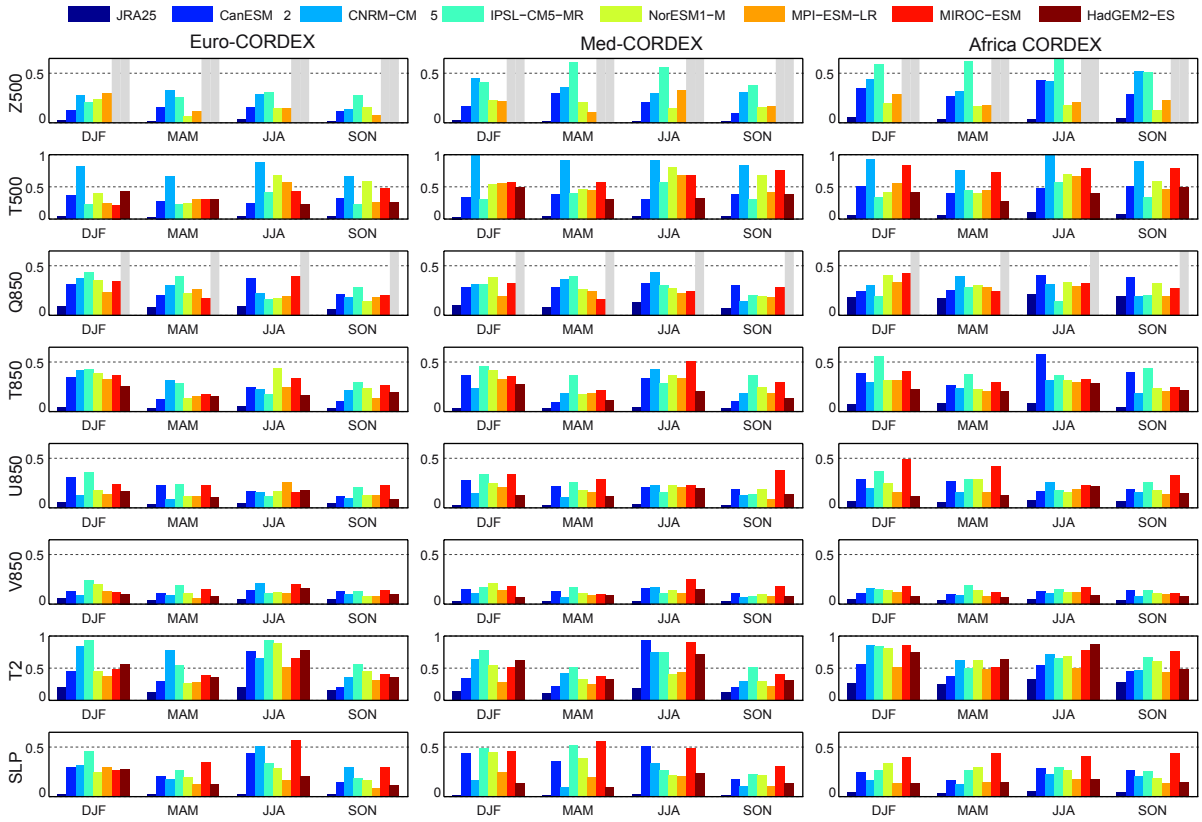

Fig. 11 Median of the normalized mean difference between the seven ESMs and ERA-Interim along the lateral boundaries of the 3 CORDEX domains shown in Fig. 1; results are shown for all seasons