How well do CMIP5 Earth System Models simulate present climate conditions in Europe and Africa?

A performance comparison for the downscaling community

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

Abstract The present study assesses the ability of seven Earth System Models (ESMs) from the Coupled Model Intercomparison Project Phase 5 to reproduce present climate conditions in Europe and Africa. This is done from a downscaling perspective, taking into account the requirements of both statistical and dynamical approaches. EC-MWF's ERA-Interim reanalysis is used as reference for an evaluation of circulation, temperature and humidity variables on daily timescale, which is based on distributional similarity scores. To additionally obtain an estimate of reanalysis uncertainty, ERA-Interim's deviation from the Japanese Meteorological Agency JRA-25 reanalysis is calculated. Areas with considerable differences between both reanalyses do not allow for a proper assessment, since ESM performance is sensitive to the choice of reanalysis. For use in statistical downscaling studies, ESM performance is computed on the grid-box scale and mapped over a large spatial domain covering Europe and Africa, additionally highlighting those regions where significant distributional differences remain even for the centered/zero-mean time series. For use in dynamical downscaling studies, performance is specifically assessed along the lateral boundaries of the three CORDEX domains defined for Europe, the Mediterranean Basin and Africa.

S. Brands (✉) · J. M. Gutiérrez
Instituto de Física de Cantabria (UC-CSIC), Santander, Spain
e-mail: brandssf@unican.es

S. Herrera
Predictia Intelligent Data Solutions, Santander, Spain

J. Fernández
Dept. of Applied Mathematics and Comp. Sci.,
Universidad de Cantabria, Santander, Spain

1 Introduction

At the onset of the Coupled Model Intercomparison 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. 2012). 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 (AR5) of the Intergovernmental Panel on Climate Change (IPCC). 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 within the framework of the COordinated Regional Climate Downscaling EXperiment (CORDEX) (Jones et al. 2011). These initiatives use both dynamical and statistical downscaling (SD) 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 (Fowler et al. 2007; Maraun et al. 2010; Winkler et al. 2011a, b). This is done by either running a Regional Climate Model (RCM), driven by a GCM at its lateral boundaries, or by applying

empirical relationships, usually found between large-scale reanalysis data and small-scale station data, to GCM output (Giorgi and Mearns 1991). The basic assumption of applying downscaling methods in this context is that the ESMs should closely reproduce the observed climatology of the large-scale variables used as predictors/drivers in statistical/dynamical schemes (Hewitson and Crane 1996; Timbal et al. 2003; Charles et al. 2007; Plavcova and Kysely 2012).

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 ERA-Interim reanalysis data (Dee et al. 2011). This is hereafter referred to as the 'performance' of the ESMs (Giorgi and Francisco 2000). ERA-Interim is used as reference for evaluating ESM performance, not because it is assumed to be superior to other reanalysis products, but because it is the one used within the COR-DEX initiative (http://wcrp-cordex.ipsl.jussieu.fr). The models' performance is assessed by testing their ability to reproduce the mean and cumulative distribution function of season-specific daily data, hereafter jointly referred to as the 'climatology'.

The study focusses on middle-tropospheric circulation, temperature and humidity variables which are of particular importance for the purpose of downscaling since they are either used as predictor variables in statistical schemes (Cavazos and Hewitson 2005; Sauter and Venema 2011; Brands et al. 2011b) or form the lateral boundaries in dynamical applications (Fernández et al. 2007; Laprise 2008). In order to test ESM performance in different climate regions, we consider a large spatial domain covering Europe and Africa. Specific information for the dynamical downscaling approach is provided by assessing ESM performance along the lateral boundaries of the three domains used in the Euro-CORDEX, Med-CORDEX and COR-DEX-Africa initiatives.

In downscaling studies, reanalysis products are commonly used as a surrogate of observational data. However, reanalyses are known to suffer from biases with respect to observations and consequently can differ significantly over certain regions (see Brands et al. 2012, and references therein). As outlined by Sterl (2004), the difference between two distinct reanalysis datasets is a reasonable estimator of observational uncertainty, especially in case an accepted observational dataset for the variables in question is not available. Albeit seldom assessed in downscaling studies (Koukidis and Berg 2009; Hofer et al. 2012), reanalysis uncertainty is relevant for (1) the evaluation of ESM performance and (2) the applicability of the downscaling methods themselves. With respect to (1),

large differences between JRA-25 and ERA-Interim indicate that ESM performance is sensitive to the choice of reanalysis used as reference for validation and, consequently, cannot be objectively assessed (Gleckler et al. 2008). With respect to (2), calibrating SD-methods and coupling RCMs require the large-scale predictor/boundary data to reflect 'real' atmospheric processes (Fernández et al. 2007; Koukidis and Berg 2009; Hofer et al. 2012). Strictly speaking, downscaling is not applicable in regions where reanalysis uncertainty is large since the latter assumption does not hold. Therefore, apart from assessing ESM performance, we provide a simple estimate of reanalysis uncertainty by calculating the climatological differences between an additional reanalysis product, the Japanese Reanalysis JRA-25 (Onogi et al. 2007), and ERA-Interim. Note that a comprehensive assessment of this issue, which would involve a comparison with observations, is out of the scope of the present paper.

Our results are expected to be of value for the downscaling community because little to no information on the relative performance of the CMIP5-ESMs is available at a time the downscaling community has to decide on which ESMs to rely on. 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. 2012). The additional assessment of reanalysis uncertainty is an update of Brands et al. (2012), who assessed the differences between ECMWF ERA-40 (Uppala et al. 2005) and NCEP/NCAR reanalysis 1 (Kalnay et al. 1996) from a downscaling perspective, and is meant to foster the scientific discussion on this important issue within the downscaling community.

2 Data

The study area considered in this work is shown in Fig. 1. It extends 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.

We use data from the seven ESMs listed in Table 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 use CMIP5 experiment number '3.2 historical' (Taylor et al. 2012). This new generation of control runs is forced by observed atmospheric composition changes of both natural and anthropogenic nature in the
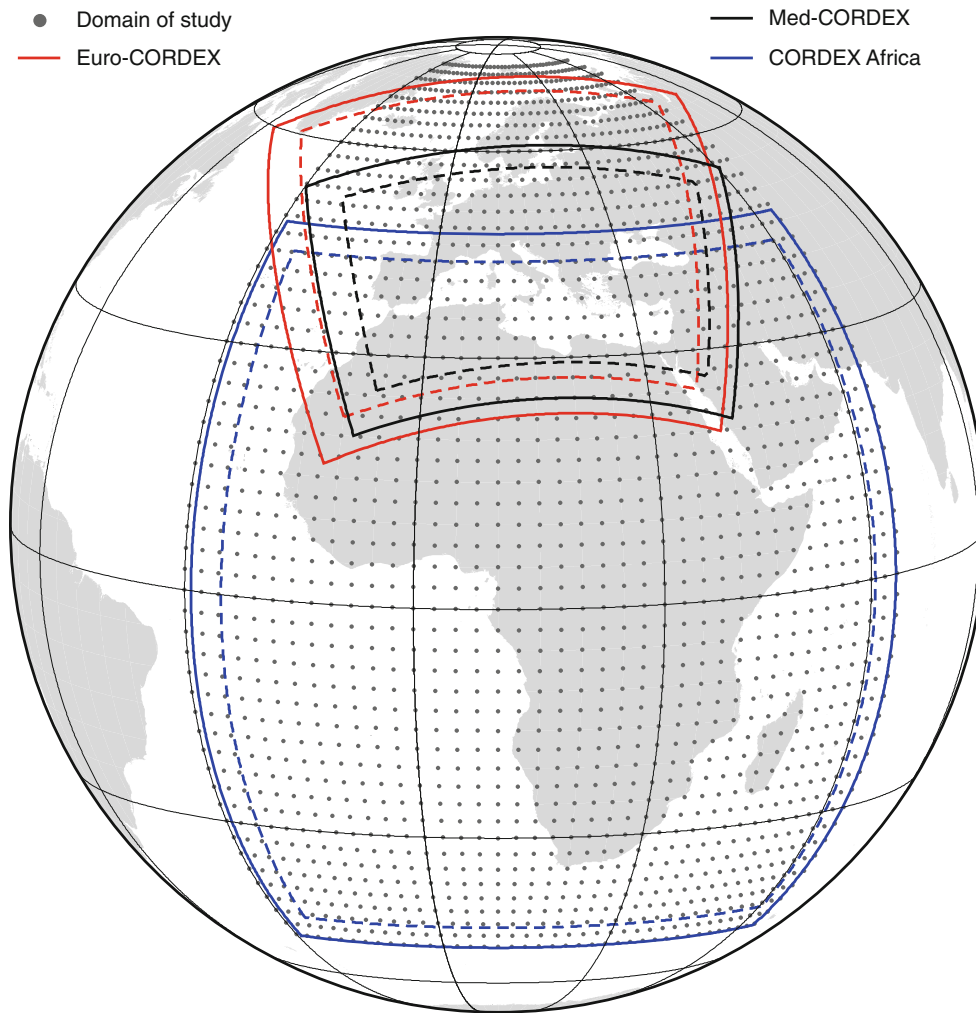
Fig. 1 Geographical domain considered in the study (*black dots*) and lateral boundaries of the Euro-CORDEX, Med-CORDEX and CORDEX Africa domains, *solid* and *dashed squares* refer to the exterior and interior of these boundaries

Table 1 CMIP5 Earth System Models used in this study

Model	Hor. resolution	References
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)

Table 2 Variables used in this study

Code	Name	Height	Unit	Acronyms
Z	Geopotential	500 hPa	$m^2\ s^{-2}$	Z500
T	Temperature	2 m, 850, 500 hPa	K	T2, T850, T500
Q	Specific humidity	850 hPa	$kg\ kg^{-1}$	Q850
U	U-wind	850 hPa	$m\ s^{-1}$	U850
V	V-wind	850 hPa	$m\ s^{-1}$	V850
SLP	Sea-level pressure	Mean sea-level	Pa	SLP

period 1850–2005. 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 for

defining the lateral boundaries in the process of nesting a RCM into a GCM.

As reference dataset for assessing ESM performance, we apply the European Centre for Medium Range Weather Forecasts ERA-interim reanalysis (Dee et al. 2011). As a second quasi-observational dataset, the Japanese

Meteorological Agency JRA-25 reanalysis (Onogi et al. 2007) is used for comparison with ERA-Interim in order to obtain an estimate of reanalysis uncertainty (see Sect. 3 for more details).

Due to distinct native horizontal resolutions (see Table 1), the 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. The period under study is 1979–2005. Daily mean values were used and, when not provided by the original data set, were derived from 6-hourly instantaneous values.

3 Methods

The methodological approach followed in this study is twofold. First, to evaluate the degree of reanalysis uncertainty, atmospheric variables from JRA-25 are validated against those from ERA-Interim. Due to the lack of observational datasets for free-tropospheric variables on daily timescale, the difference between two distinct reanalyses is a reasonable estimator of observational uncertainty. If a close agreement is found, both reanalyses are likely driven by assimilated observations and reasonably reflect reality. On the contrary, in case of considerable differences, at least one reanalysis is dominated by internal model variability rather than observations and therefore does not reflect reality (Sterl 2004). Consequently, validating JRA-25 against ERA-Interim does not yield an 'error' in the sense of one reanalysis being 'better' than the other, but is interpreted as an estimate of reanalysis uncertainty.

Second, ESM performance in present climate conditions is assessed by validating the ESMs listed in Table 2 against ERA-Interim. At this point, the results obtained from the first step allow for testing if the degree of reanalysis uncertainty permits for assessing ESM performance in an objective manner. In case of large reanalysis uncertainties, ESM performance cannot be objectively assessed since it is sensitive to reanalysis choice. On the contrary, in case of negligible reanalysis uncertainties, ESM performance is not sensitive to reanalysis choice and applying JRA-25 as reference for validation would lead to similar results.

The first measure for evaluating reanalysis uncertainty and ESM performance in this study is the mean difference (bias). Since the variability of the applied daily timeseries is much smaller in the tropics than in the mid-latitudes, the bias is normalized by the standard deviation of ERA-Interim to make results comparable (Brands et al. 2011b). This is hereafter referred to as 'normalized bias' or 'normalized mean difference' (when applied to the two reanalyses).

To detect distributional differences, we apply the two-sample Kolmogorov Smirnov test (KS test) to the original time series and to the time series centered to have zero mean, which are obtained by subtracting the seasonal mean from each timestep. For simplicity, the resulting time series will hereafter be referred to as 'centered'. Validating centered time series is equivalent to removing the mean difference and, consequently, permits for detecting distributional differences in higher order moments. Note that comparing centered ESM data to centered ERA-Interim data is one possible solution of correcting the mean error of the ESM, which is commonly done in statistical downscaling studies (Wilby et al. 2004) and recently has also been proposed for the dynamical downscaling approach (Colette et al. 2012; Xu and Yang 2012).

The KS test is a non-parametric hypothesis test assessing the null hypothesis (H_0) that two candidate samples (here: reanalysis and ESM series for a particular gridbox and season of the year) come from the same underlying theoretical probability distribution. It is defined by the statistic:

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

where n is the length of the time series, z^i denotes the ith data value of the sorted joined sample and E and I are the empirical cumulative frequencies from a given ESM (or JRA-25, in case reanalysis uncertainty is assessed) and the ERA-Interim reanalysis, which serves as reference for validation in any case. 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. Decreasing p 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 before estimating the p value of the KS statistic in order to avoid committing too many type I errors (i.e. erroneous rejections of the H_0). Under the assumption that the underlying time series follow a first-order autoregressive process, the effective sample size, n^*, is defined as follows (Wilks 2006):

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

where n is the sample size and p_1 is the lag-1 autocorrelation coefficient.

If not specifically referred to in the text, all of the above mentioned validation measures are applied at the grid-box scale, using season specific time series.

4 Results

In this section we first assess reanalysis uncertainty (by comparing JRA-25 with ERA-Interim) and then evaluate ESM performance (by comparing the ESMs with ERA-Interim). The normalized bias is applied to assess reanalysis differences and ESM errors in the mean of the distribution. Then, to detect reanalysis differences and ESM errors in higher order moments, we apply the KS test to the centered time series. Note that in the latter case the degrees of freedom are reduced by -1, which is a negligible problem since n^* is of the order of several hundreds in any case.

Finally, model performance for the original (i.e. non-transformed) 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 downscaling 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 RCM, which is driven by the ESM at the boundaries of the domain only.

4.1 Reanalysis uncertainty

In Fig. 2, the results of validating JRA-25 against ERA-Interim in boreal winter (DJF, first and second column) and summer (JJA, third and forth column) are mapped for the variables SLP, T2, T850, Q850, U850, V850, T500 and Z500 (from top to bottom). Along the first and third column, the mean difference between JRA-25 and ERA-Interim, normalized by the standard deviation of ERA-Interim (Bias/Std) is mapped. The second and fourth columns display the logarithm to base 10 of the KS statistic's p value (KS pVal), obtained from applying the KS test to the centered time series. Recall that applying centered data at this point permits for detecting reanalysis uncertainties in higher order moments. Values below -1.3 indicate that distributional differences in higher order moments are significant ($\alpha = 0.05$), whereas values exceeding this threshold represent spurious differences (see the white area in the panels). For simplicity, 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 when applying the KS test to the centered time series, thereby indicating
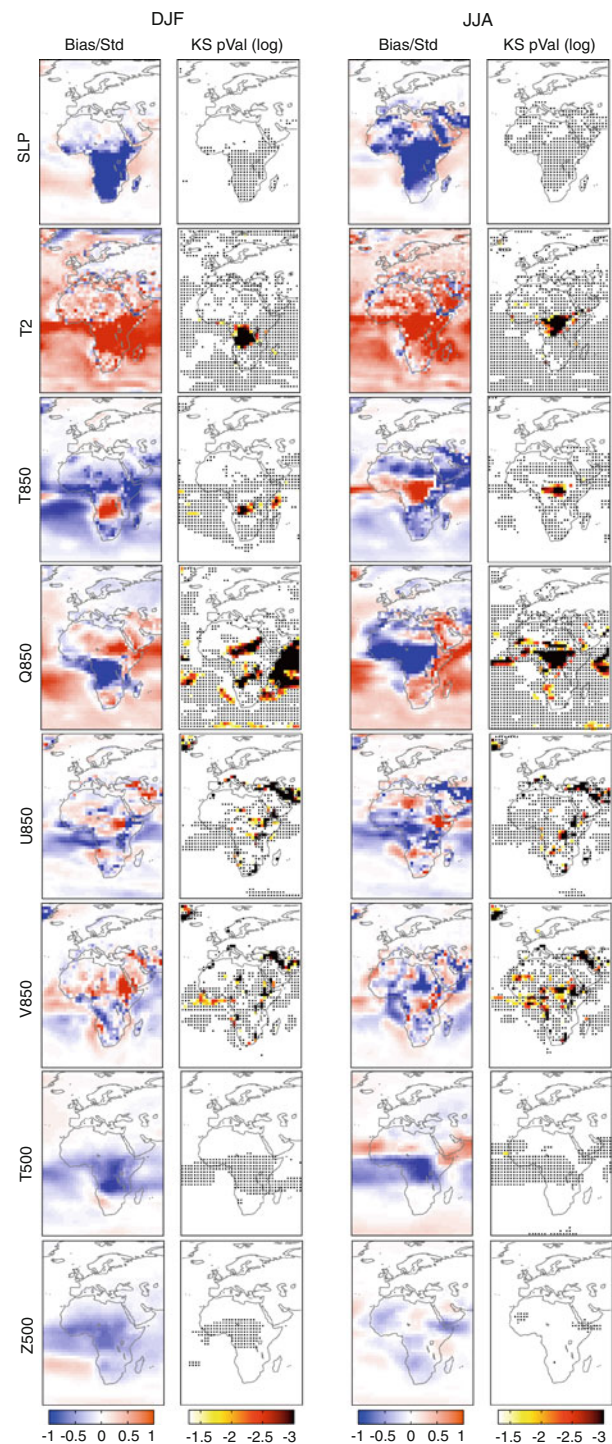
Fig. 2 *Columns* 1+3: Mean differences between JRA-25 and ERA-Interim, normalized by the standard deviation of the latter; *Columns* 2+4: *p* Value (in logarithmic scale) of the KS test applied to the time series from JRA-25 and ERA-Interim, both centered to have zero mean. *Grid boxes* are whitened if the *p* value does 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. *Grid boxes* marked with a *black dot* indicate areas where significant distributional differences for the original reanalysis data are eliminated by using the centered time series

that reanalysis uncertainty is restricted to a shift in the mean of the distribution.

Reanalysis uncertainty for SLP (see row 1 in Fig. 2) is negligible north of 45° N and clearly depends on season in the Northern Hemisphere subtropics (25° N–45° N), where it is more (less) pronounced in JJA (DJF). Over Africa (and especially in JJA), SLP in JRA-25 is much lower than in ERA-Interim, while the opposite is the case over the adjacent ocean areas. Consequently, JRA-25 is characterized by a more pronounced land-sea pressure gradient than ERA-Interim. For the Southern and Northern Hemisphere mid-latitude oceans, reanalysis differences are negligible.

Reanalysis uncertainty for T2 (see row 2 in Fig. 2) is more widespread than for any other variable under study, with JRA-25 being systematically warmer than ERA-Interim. Exceptions from this general result occur over land areas north of 45° N and the northern Arctic Ocean, where differences are negligible or even negative during DJF and MAM (MAM is not shown).

As was the case for SLP, reanalysis uncertainty for T850 (see row 3 in Fig. 2) 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). Along the ascending branch of the Hadley Cell, 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 Hadley Cell.

The tripole difference structure found for T850, as well as its associated seasonality, also appears in Q850 (see row 4 in Fig. 2). Along the ascending branch of the Hadley Cell, JRA-25 is drier than ERA-Interim, while the opposite is the case along the descending branches. 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 (see row 5+6 in Fig. 2), reanalysis uncertainty is generally weaker than for the other variables under study and, in the extratropics, is confined to regions of high orography. 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.

Considerable reanalysis uncertainties for T500 (see row 7 in Fig. 2) are mainly confined to the Tropics. In DJF, JRA-25 is generally colder than ERA-Interim (exception: western South Africa), whereas in JJA it is colder near the Equator but warmer over the semi-arid to arid regions of the Northern Hemisphere.

Finally, although reanalysis uncertainty for Z500 (see row 8 in Fig. 2) 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. In conjunction with higher values in the area of the St. Helen's High, the meridional gradient for Z500 over the South Atlantic is more pronounced in JRA-25 than in ERA-Interim.
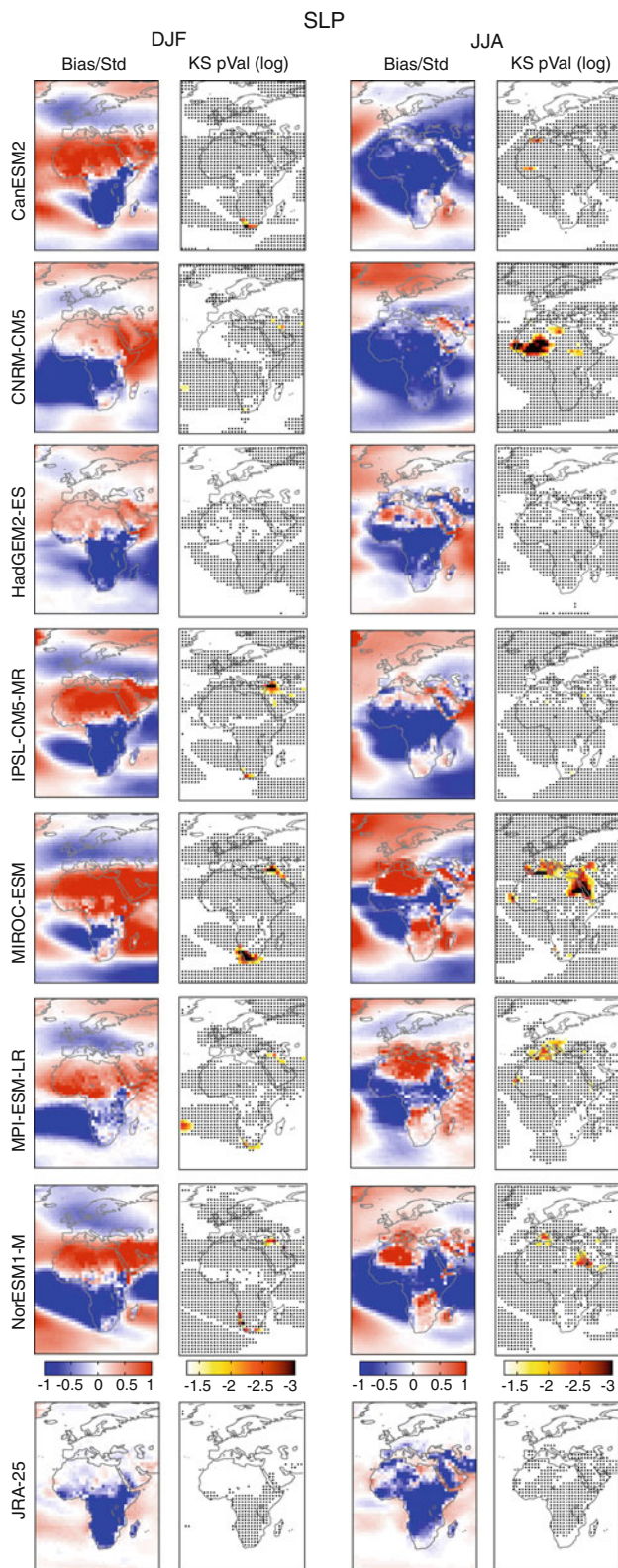
When applying the KS-test to centered/zero-mean data, no significant distributional differences are detected for the case of SLP, T500 and Z500. For T850 and T2, the area of significant distributional differences is reduced to Central Africa (Kongo Basin), where it follows the seasonal march of the Hadley Cell, as was the case for the original data (see Fig. 2, columns 2 and 4). For U850 and V850, this area is 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, significant distributional differences are essentially removed in the extratropics, 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, the mean difference between JRA-25 and ERA-Interim generally exceeds a magnitude of one standard deviation for central-to-south Africa. Even if the data is centered to have zero mean, i.e. if differences in the mean are removed, there remain significant differences in higher order moments. Consequently, it is neither possible to objectively assess ESM performance for central-to-south Africa, nor does the basic assumption of 'real' or 'perfect' large scale data hold in these regions.

In contrast to the tropics, reanalysis uncertainty in the extratropics is generally negligible and the above mentioned problems may consequently be ignored, meaning that ESM performance can be assessed and the basic downscaling assumption can be affirmed.

4.2 Performance maps

Figures 3, 4, 5, 6, 7, 8, 9 and 10 show the results of validating the 7 ESMs listed in Table 1 against ERA-Interim for the case of SLP, T2, T850, Q850, U850, V850, T500 and Z500 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 (KS pVal) obtained from the centered/zero-mean data. For the ease of comparison, the corresponding panels for reanalysis uncertainty (copied from Fig. 2) are displayed at the bottom of each figure.
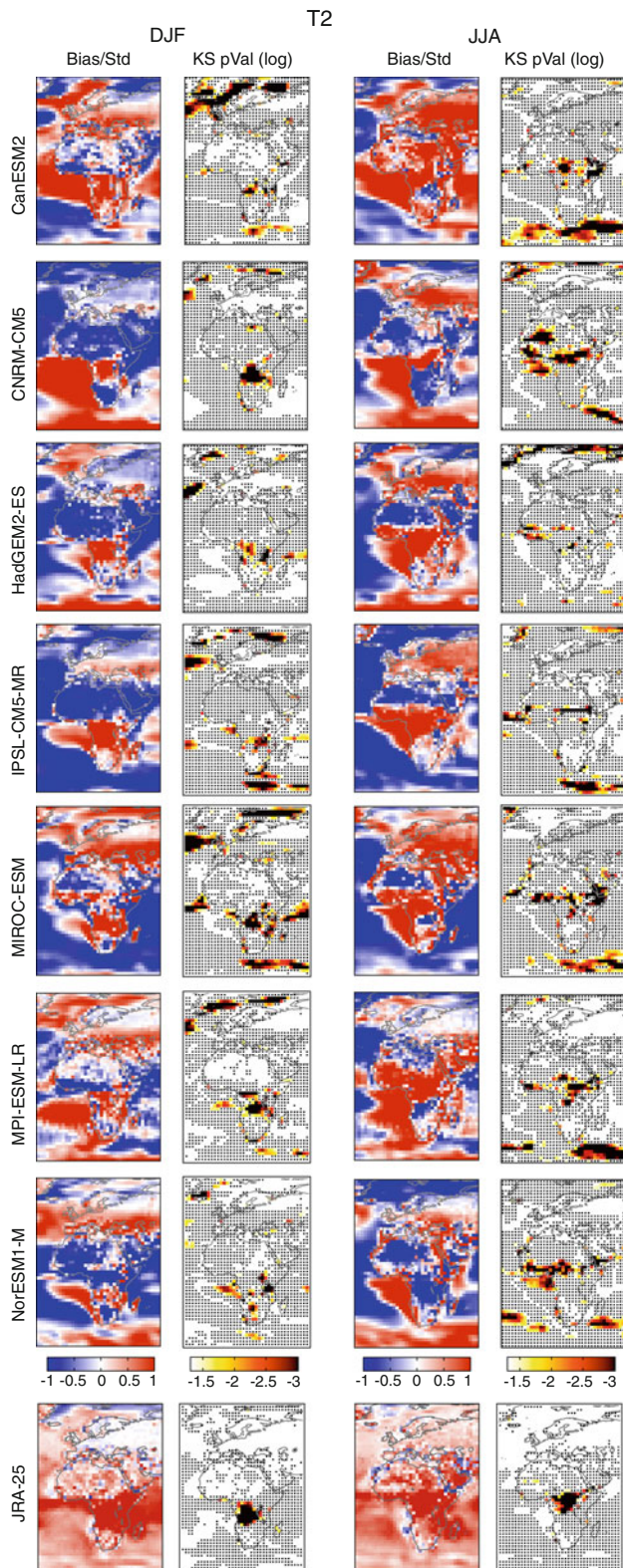
SLP

Fig. 3 *Columns 1+3*: Mean differences (columns 1+3) between the seven ESMs listed in Table 1 and ERA-Interim, normalized by the standard deviation of ERA-Interim; *Columns: 2+4*: *p* Value (in logarithmic scale) of the KS test applied to the time series from the respective ESM and ERA-Interim, both centered to have zero mean. *Grid-boxes* are whitened if the *p* value does not exceed the threshold value of −1.301, i.e. if the distributional differences are not significant (α = 0.05). *Colour darkening* corresponds to increasing (and significant) distributional differences/ESM errors. *Grid boxes* marked with a *black dot* indicate areas where significant ESM errors in the original data are eliminated by using the centered time series; results for SLP. For the ease of comparison, the corresponding panels for reanalysis uncertainty (copied from Fig. 2 are displayed at the bottom of the figure

LR and NorESM1-M (MAM is not shown). In JJA, CanESM2 and CNRM-CM5 suffer from too low SLP values over a large fraction of the land areas. 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 land-sea pressure gradient during the West African rainy season. Over the extratropical North Atlantic, SLP during JJA is systematically overestimated by all ESMs except MPI-ESM-LR and CanESM2, the latter two showing more heterogeneous spatial patterns.
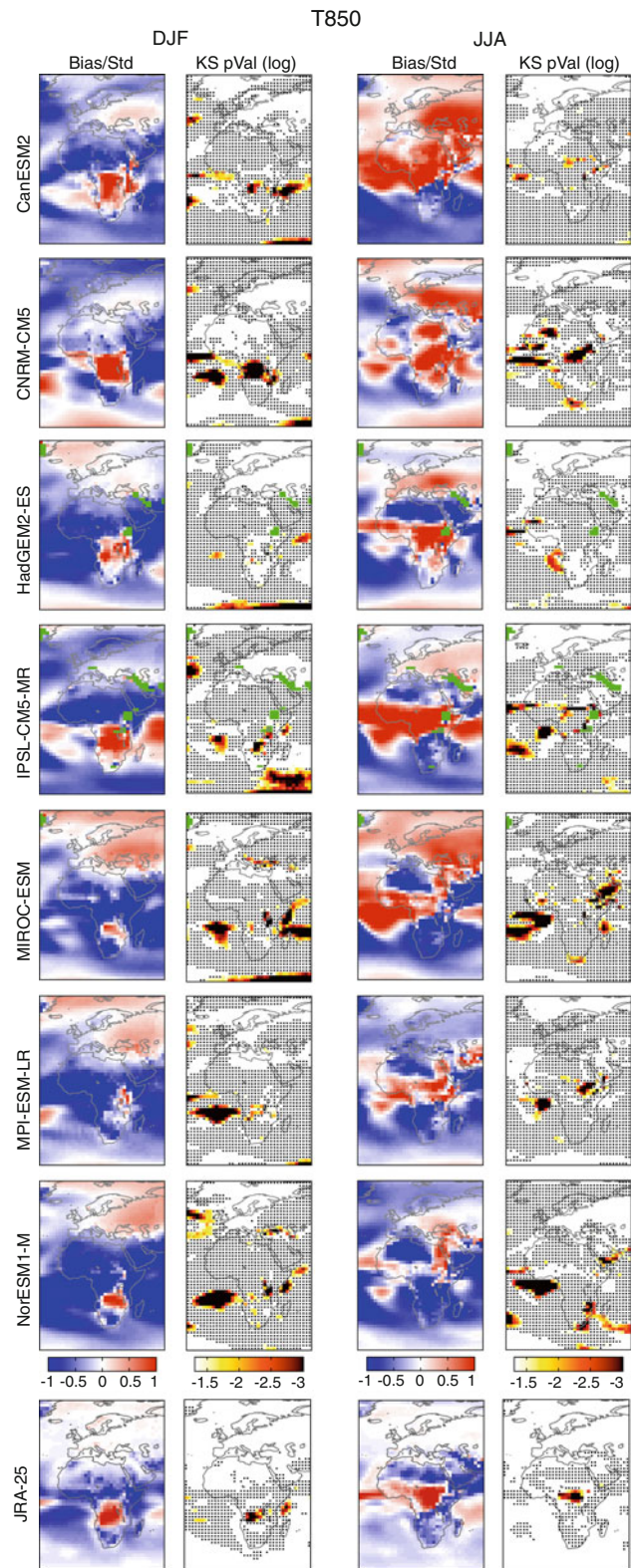
The T2 bias is generally larger and more widespread than at 850 hPa (compare Figs. 4, 5). The aforementioned largely exaggerated meridional pressure gradient during boreal winter and spring is associated with too strong westerlies in the Northern Hemisphere mid-latitudes, which lead to an exaggerated advection of oceanic air masses, resulting in too mild and moist conditions in continental Europe, an effect that extends throughout the whole planetary boundary layer (see Figs. 4, 5, 6 for T2, T850 and Q850 respectively).

During the core of the West African monsoon (JJA), and as revealed by U500 (not shown), a too strong Subtropical Jet, as well as a too weak African Easterly Jet (Cook 1999) are simulated by the ESMs, with NorESM1-M performing best for these features. The monsoonal winds over West Africa, as represented by U850 in JJA, are underestimated over the Sahel but overestimated over the subhumid to humid zones along the Guinea Coast in all ESMs except IPSL-CM5A-MR; the latter underestimating this variable over the entire region (see Fig. 7). Also reflected in U850 is the above mentioned overestimation of the wintertime westerlies in the North Atlantic-European region. In general, the bias for U850 is larger and more widespread than for V850 (compare Figs. 7, 8).

For all ESMs except IPSL-CM5A-MR, a cold bias was found in the middle troposphere (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
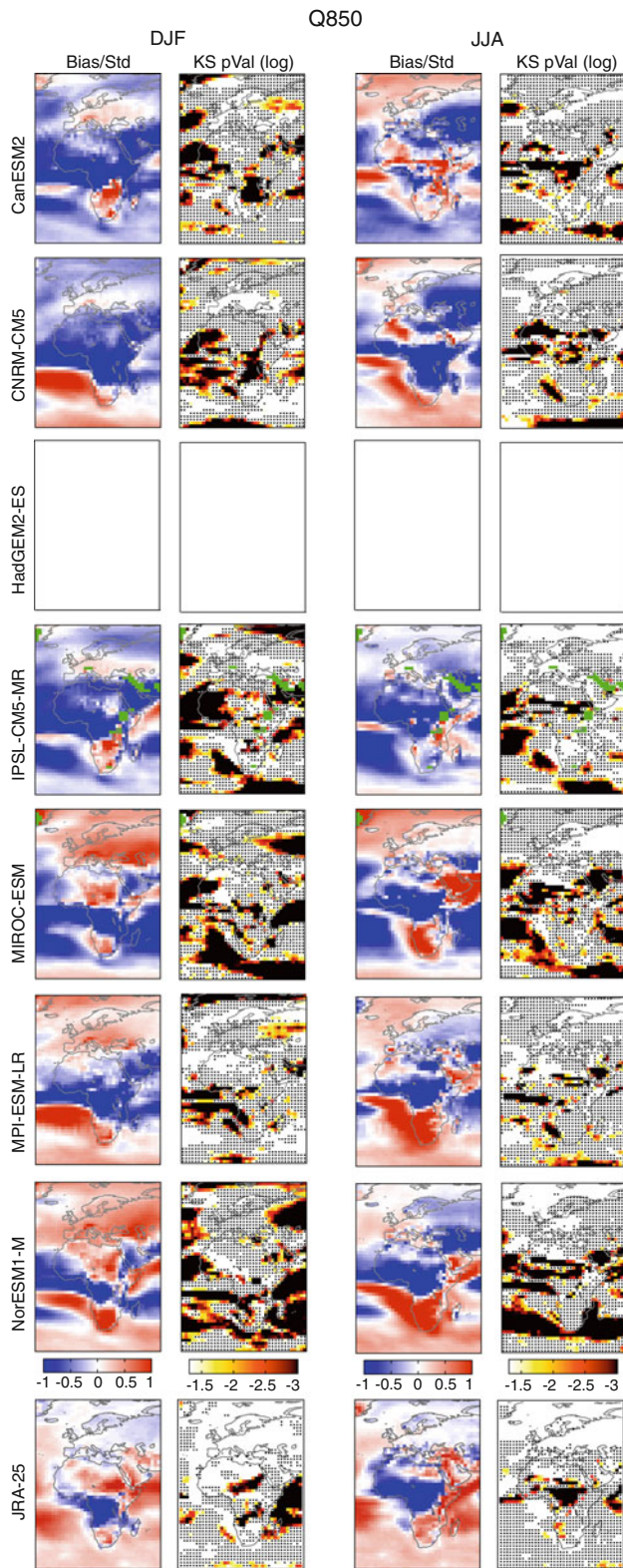
Regarding the ESM error for SLP (see Fig. 3), the meridional pressure gradient in the Northern Hemisphere (NH) extratropics during DJF and MAM is too strong in CanESM2, IPSL-CM5A-MR, MIROC-ESM, MPI-ESM-
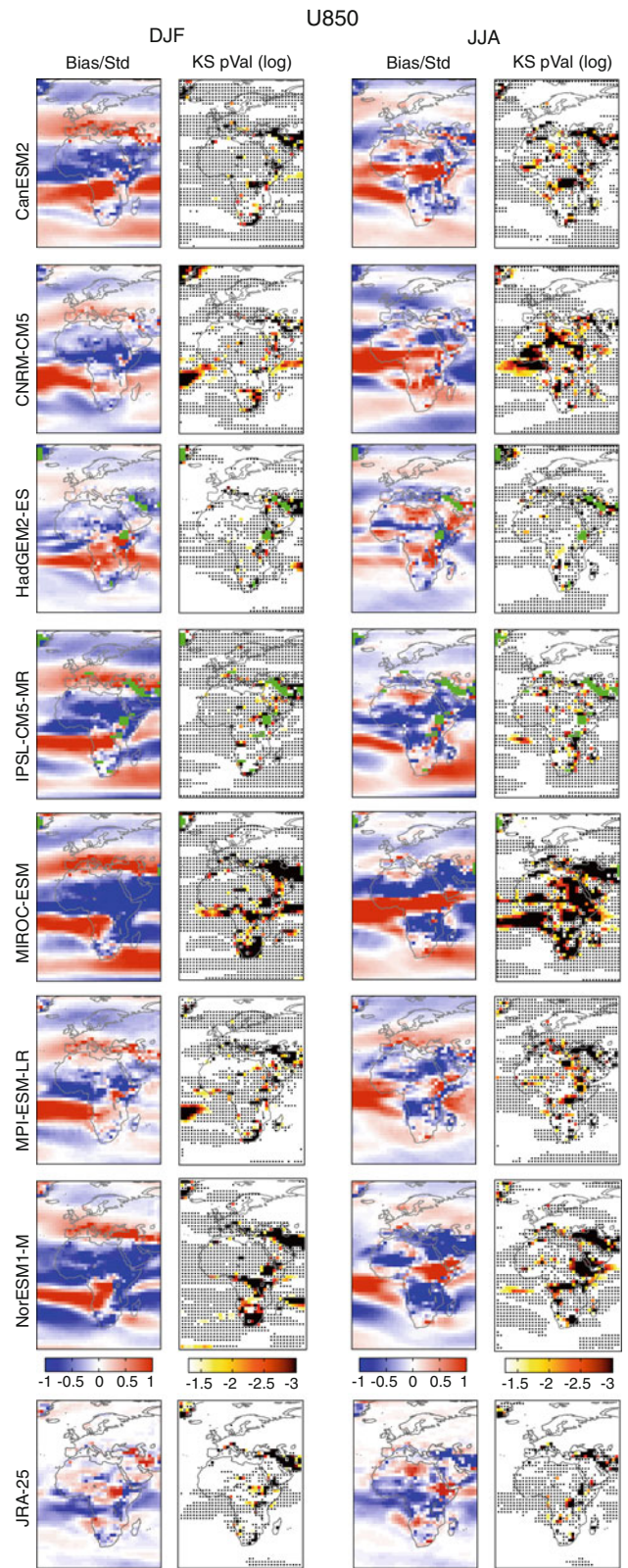
Fig. 4 As Fig. 3, but for T2

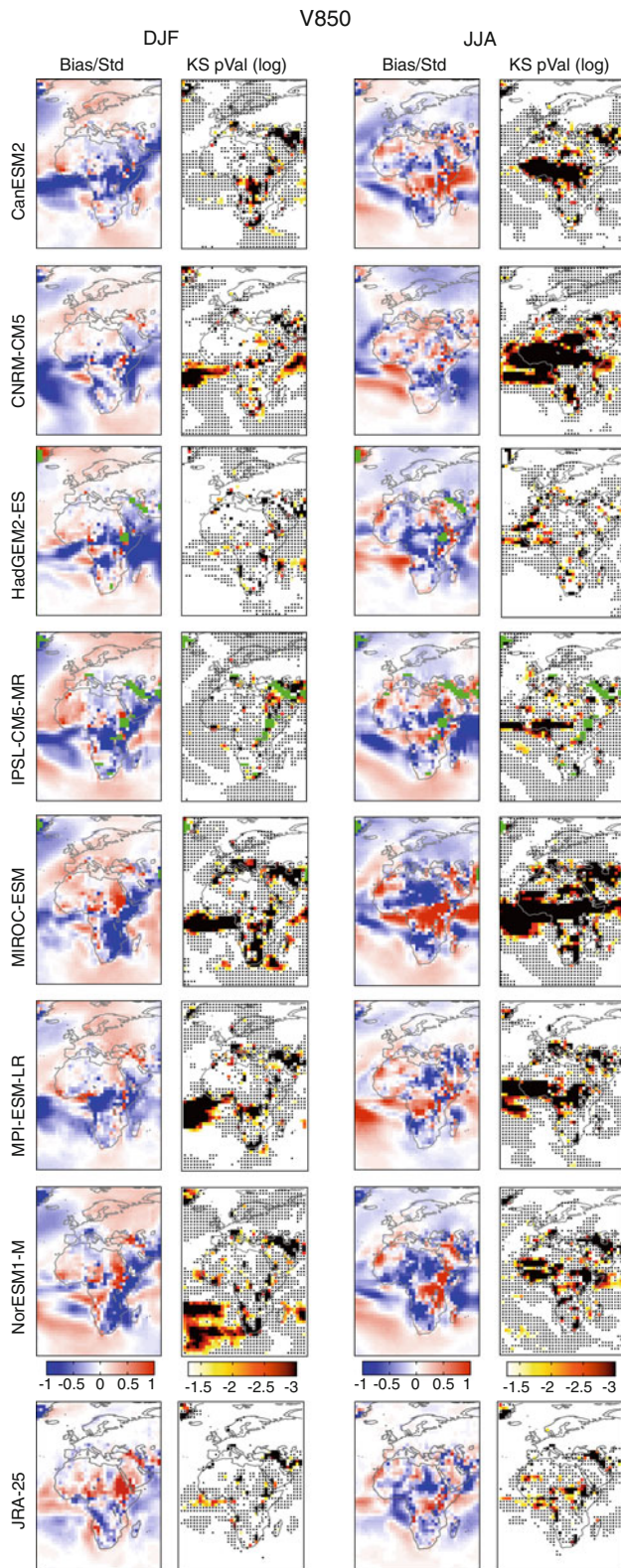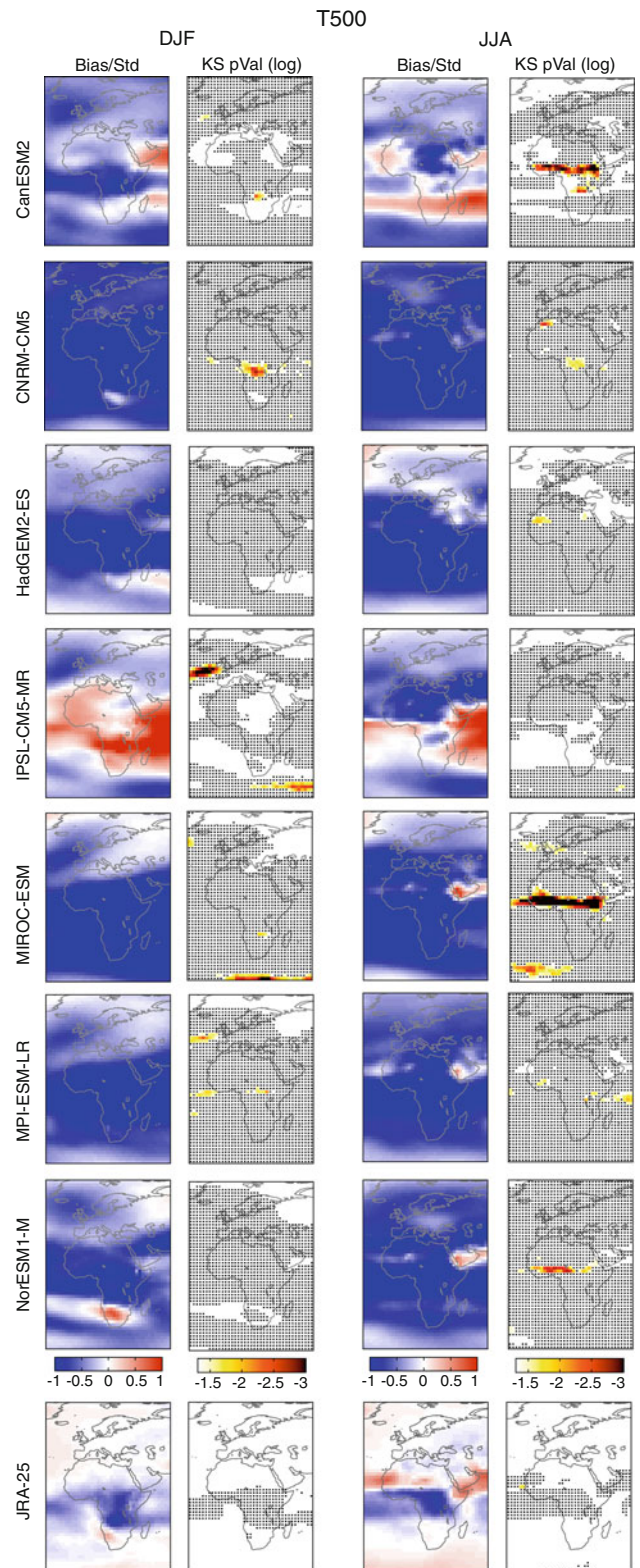**Fig. 5** As Fig. 3, but for T850, *green grid boxes* refer to lack of data at the ESGF-portals

Fig. 6 As Fig. 3, but for Q850, *empty panels* and *green grid boxes* refer to lack of data at the ESGF-portals
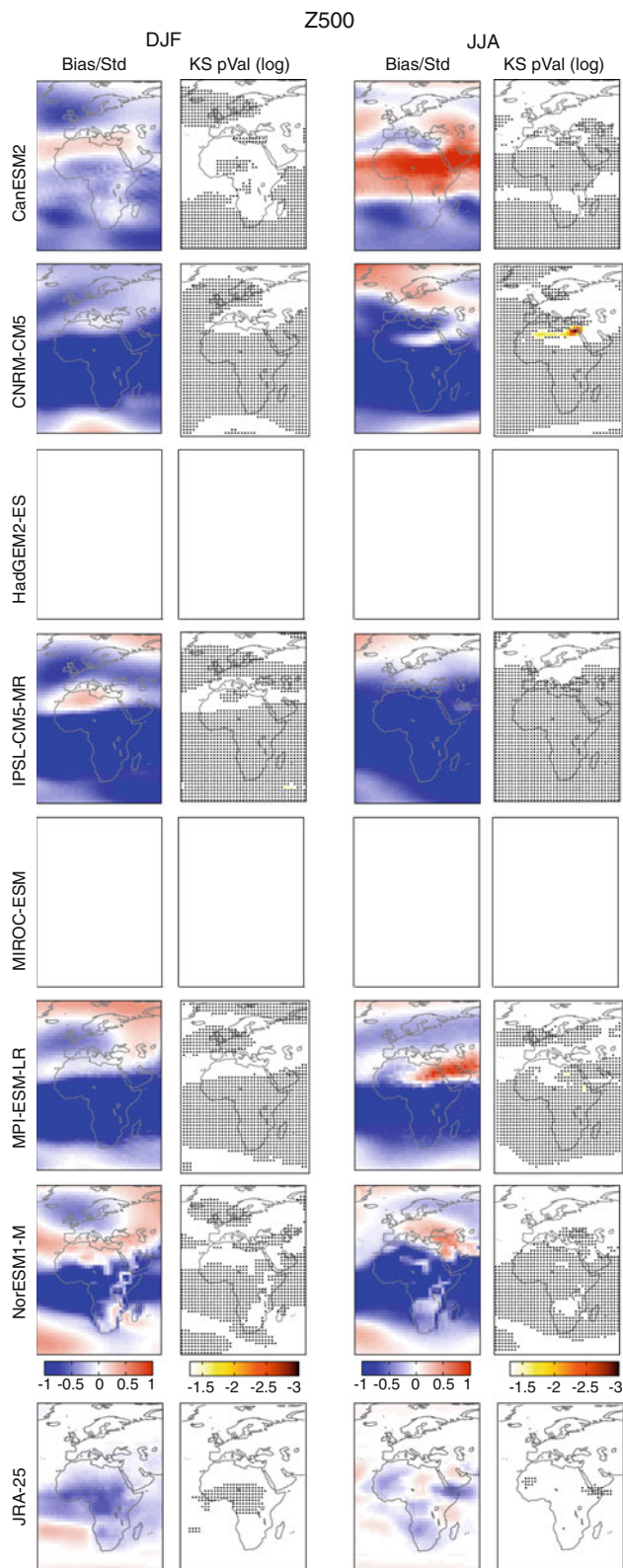
Fig. 7 As Fig. 3, but for U850, *green grid boxes* refer to lack of data at the ESGF-portals

Fig. 8 As Fig. 3, but for V850, *green grid boxes* refer to lack of data at the ESGF-portals

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 ESGF-portals

associated with an underestimation of the geopotential at 500 hPa over the Tropics (see Fig. 10).

Remarkably, one should expect the spatial pattern of the normalized ESM error to be independent from the spatial patterns of the normalized reanalysis difference. However, a considerable agreement between both types of patterns is found in central-to-south Africa, at least for some variables. To mention an example, the pattern of reanalysis uncertainty for T850 (see Fig. 5, JRA-25 is warmer than ERA-Interim over central Africa) is approximately resembled by a warm bias in all of the 7 ESMs under study (compare last row to remaining rows in Fig. 5). This points to a substantial error in the reference data set (ERA-Interim) for this specific region. This error, however, cannot be ultimately deduced from our analyses, since this would require a more thorough verfication against independent station and/or radiosonde data.

For all applied variables, ESM performance largely improves when applying centered time series (see columns 2 and 4 in Figs. 3, 4, 5, 6, 7, 8, 9, 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 latter season not shown), 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 (see Fig. 3, row 3, columns 2 and 4).

In case of the centered 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 outperforms any other ESM for both T850 and T2, the performance of CanESM2 being comparable in case of T850.

Regarding the centered U850 and V850 data (see Figs. 7, 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 using centered data, CanESM2, MIROC-ESM, and NorESM1-M suffer from errors in higher order moments along the ascending branch of the Hadley Cell in JJA (see Fig. 9). In
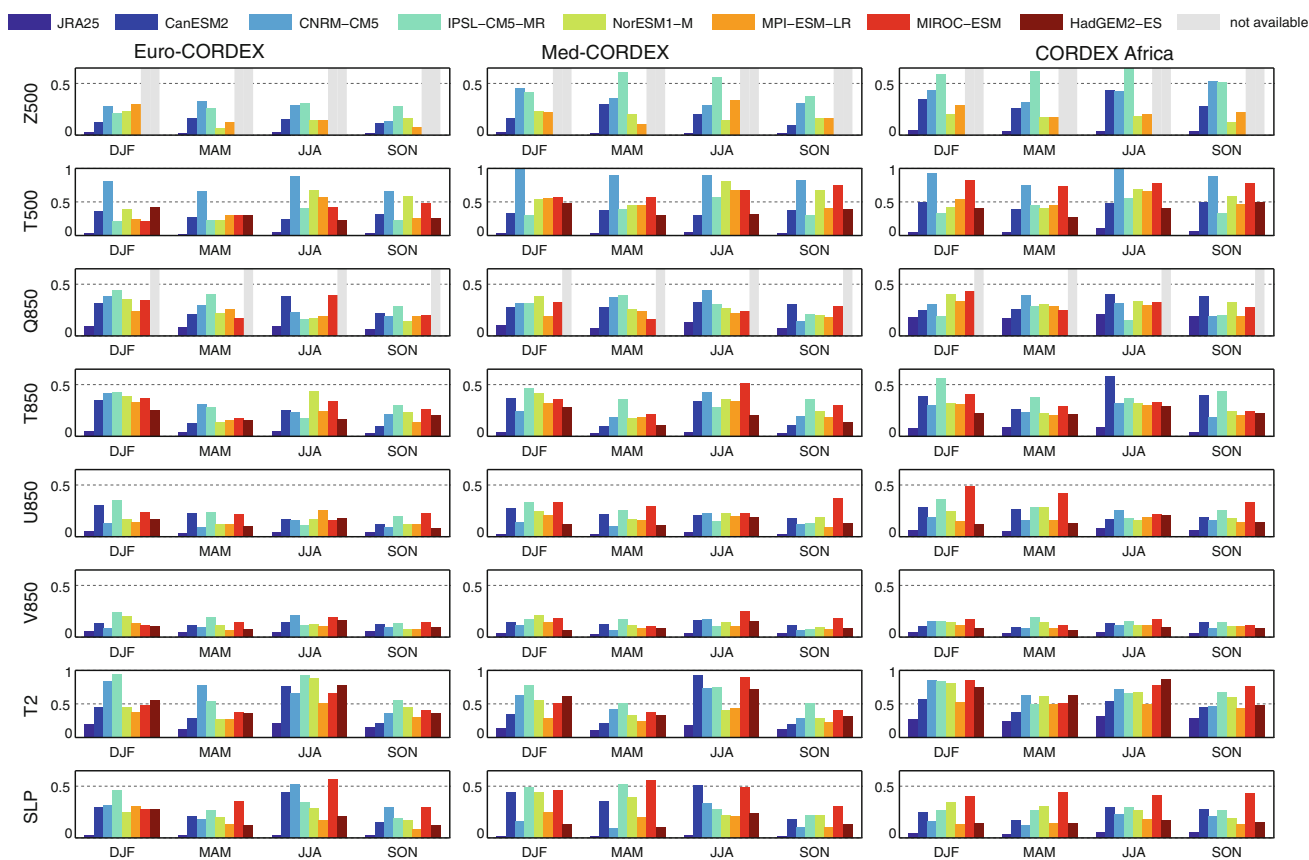
Fig. 11 Median of the absolute normalized mean differences between JRA-25 and ERA-Interim (reanalysis uncertainty, *first bar* in *each panel*) and between the ESMs and ERA-Interim (ESM errors, *remaining bars*) along the lateral boundaries of the three CORDEX domains shown in Fig. 1. *Left* EURO-CORDEX, *middle* Med-CORDEX, *right* CORDEX Africa. Results are shown for all seasons, *grey bars* indicate lack of availability at the ESGF portals. Due to the larger error magnitude, y-axes have been stretched for T2 and T500

IPSL-CM5-MR, this error type appears during DJF between the Azores and the Bay of Biscay while it is virtually absent in HadGEM2-ES.

As shown in Fig. 10, ESM errors for Z500 disappear almost completely for the centered data.

4.3 Performance along the lateral boundaries of the CORDEX domains

Figure 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 in 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 roughly proportional to their respective medians (i.e. the higher the median, the broader the IQR).

It is remarkable that ESM performance along the lateral boundaries of the 3 domains is generally very similar, i.e. the models do not perform systematically worse for any single domain compared to the other two. For any domain under study, ESM performance is best for V850, followed by U850, and is worse for T2 and T500 (note the distinct scaling of the y-axis for the latter two). Intermodel performance differences are smallest for U850 (except over the African domain) and V850 and generally larger for the remaining variables. Also, intermodel performance differences for the Med-CORDEX and CORDEX Africa domains are more pronounced 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. Interestingly, for the CORDEX Africa domain, ESM performance (and reanalysis agreement) along the lateral boundaries is systematically better than in the interior of the domain.

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 comparable to those obtained from validating the ESMs against ERA-Interim in central-to-south Africa. This questions the basic downscaling assumption of 'real' or 'perfect' reanalysis data and hinders the objective evaluation of ESM performance (Gleckler et al. 2008) in these regions.

The reason behind the differences cannot be inferred from our analyses. However, the large differences between JRA-25 and ERA-Interim over central-to-south Africa are consistent with Betts et al. (2009), who found ERA-Interim compared to in-situ station data to be cold-biased over the Amazon basin. Moreover, the cold bias of ERA-Interim over African tropical regions, which was systematically found against JRA-25 and 7 ESMs, indicate that ERA-Interim might not reflect 'real' atmospheric conditions in that area and that, in a strict sense, it should not be applied there for the purpose of downscaling. This should be a warning sign for the CORDEX Africa community, indicating that the errors of the downscaled times series may originate from the driving reanalysis, apart from being caused by SD or RCM errors.

In contrast, reanalysis uncertainty for the Northern Hemispheric extratropics is negligible, which (1) affirms the above mentioned basic downscaling assumption and (2) permits for assessing ESM performance. A largely overestimated meridional pressure gradient was found in 5 out of 7 ESMs during boreal winter and spring, leading to too mild and moist conditions in continental Europe. 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. Consequently, artificial feedback processes in the scenario period resulting from ESM errors in the control/historical period (Raisanen 2007) cannot be ruled out for Europe.

HadGEM2-ES and MPI-ESM-LR generally 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 over southwestern Europe. The systematic superiority of these models questions the paradigm of equiprobable treatment of the driving models in downscaling studies.

For the CORDEX Africa domain, ESM performance and reanalysis agreement along the lateral boundaries is systematically better than in the interior of the domain, which might be one argument against the use of RCM nudging (von Storch et al. 2000). In this context, it is worth mentioning that GCM control runs nudged to reanalysis data (Eden et al. 2012) fail to reproduce the temporal variability of observed precipitation in the tropics (where

reanalysis uncertainty is large) whereas they perform well in the extratropics (where reanalysis uncertainty is low). This indicates that the success of nudging GCMs (and also RCMs) to reanalysis data might critically depend on the degree of reanalysis uncertainty.

The final message is that many of the errors found in the CMIP3-GCMs are still present in current Earth System Models. For instance, the systematic domain-wide 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. Thus, the shortcomings and corresponding recommendations for working with GCM data in the context of downscaling (Wilby et al. 2004) remain valid for the new model generation.

References

Betts AK, Koehler M, Zhang Y (2009) Comparison of river basin hydrometeorology in ERA-Interim and ERA-40 reanalyses with observations. J Geophys Res 114. doi:10.1029/2008JD010761

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 temperatures 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, Gutierrez JM, Herrera S, Cofino AS (2012) On the use of reanalysis data for downscaling. J Clim 25(7):2517–2526 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 11:22893–2290. doi:10.5194/acpd-11-22893-201

Colette A, Vautard R, Vrac M (2012) Regional climate downscaling with prior statistical correction of the global climate forcing. Geophys Res Lett 39. doi:10.1029/2012GL052258

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. Geosci Model Dev 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 DP, Uppala SM, Simmons AJ, Berrisford P, Poli P, Kobayashi S, Andrae U, Balmaseda MA, Balsamo G, Bauer P, Bechtold P, Beljaars ACM, van de Berg L, Bidlot J, Bormann N, Delsol C, Dragani R, Fuentes M, Geer AJ, Haimberger L, Healy SB, Hersbach H, Holm EV, Isaksen L, Kallberg P, Koehler M, Matricardi M, McNally AP, Monge-Sanz BM, Morcrette JJ, Park BK, Peubey C, de Rosnay P, Tavolato C, Thepaut JN, Vitart F (2011) The ERA-Interim reanalysis: configuration and perfor-mance of the data assimilation system. Q J R Meteorol Soc 137(656, Part a):553–597. doi:10.1002/qj.828

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

Eden JM, Widmann M, Grawe D, Rast S (2012) Skill, correction, and downscaling of GCM-simulated precipitation. J Clim 25(11):3970–3984. doi:10.1175/JCLI-D-11-00254.1

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 112(D4). doi:10.1029/2005JD006649

Fowler HJ, Blenkinsop S, Tebaldi C (2007) Linking climate change modelling to impacts studies: recent advances in downscaling techniques for hydrological modelling. Int J Climatol 27(12):1547–1578. doi:10.1002/joc.1556

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 113(D6). doi:10.1029/ 2007JD008972

Hanssen-Bauer I, Achberger C, Benestad R, Chen D, Forland E (2005) Statistical downscaling of climate scenarios over Scan-dinavia. Clim Res 29(3):255–268. doi:10.3354/cr029255

Hewitson BC, Crane RG (1996) Climate downscaling: techniques and application. Clim Res 7(2):85–95. doi:10.3354/cr007085

Hofer M, Marzeion B, Molg T (2012) Comparing the skill of different reanalyses and their ensembles as predictors for daily air temperature on a glaciated mountain (Peru). Clim Dyn 39(7–8):1969–1980. doi:10.1007/s00382-012-1501-2

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 downscal-ing link to CMIP5. CLIVAR Exch Newsl 16:34–40

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

Kalnay E, Kanamitsu M, Kistler R, Collins W, Deaven D, Gandin L, Iredell M, Saha S, White G, Woollen J, Zhu Y, Chelliah M, Ebisuzaki W, Higgins W, Janowiak J, Mo K, Ropelewski C,

Wang J, Leetmaa A, Reynolds R, Jenne R, Joseph D (1996) The NCEP/NCAR 40-year reanalysis project. Bull Am Meteorol Soc 77(3):437–471

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

Koukidis EN, Berg AA (2009) Sensitivity of the statistical down-scaling model (SDSM) to reanalysis products. Atmos Ocean 47(1):1–18. doi:10.3137/AO924.2009

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 dynam-ical models and the end user. Rev Geophys 48. doi: 10.1029/2009RG000314

Maraun D, Osborn T, Rust H (2012) The influence of synoptic airflow on UK daily precipitation extremes. Part II: climate model and E-OBS data validation. Clim Dyn 39(1–2):287–301. doi: 10.1007/s00382-011-1176-0

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 assess-ment. 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 Meteorol Soc Jpn 85(3):369–432. doi:10.2151/jmsj.85.369

Plavcova E, Kysely J (2012) Atmospheric circulation in regional climate models over Central Europe: links to surface air temperature and the influence of driving data. Clim Dyn 39(7–8):1681–1695. doi:10.1007/s00382-011-1278-8

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 Ser A Dyn Meteorol Oceanol 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 Ser A Dyn Meteorol Oceanol 60(3):459–491. doi:10.1111/j.1600-0870.2008.00318.x

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:AS NTFD>2.0.CO;2

Taylor KE, Stouffer RJ, Meehl GA (2012) An overview of CMIP5 and the experiment design. Bull Am Meteorol Soc 93(4):485–498. 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 downscal-ing 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

Uppala S, Kallberg P, Simmons A, Andrae U, Bechtold V, Fiorino M, Gibson J, Haseler J, Hernandez A, Kelly G, Li X, Onogi K, Saarinen S, Sokka N, Allan R, Andersson E, Arpe K, Balmaseda M, Beljaars A, Van De Berg L, Bidlot J, Bormann N, Caires S, Chevallier F, Dethof A, Dragosavac M, Fisher M, Fuentes M, Hagemann S, Holm E, Hoskins B, Isaksen L, Janssen P, Jenne R, McNally A, Mahfouf J, Morcrette J, Rayner N, Saunders R, Simon P, Sterl A, Trenberth K, Untch A, Vasiljevic D, Viterbo P, Woollen J (2005) The ERA-40 re-analysis. Q J R Meteorol Soc 131(612, Part B):2961–3012. doi:10.1256/qj.04.176

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. Geosci Model Dev 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. Elsevier, Amsterdam

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. Geogr Compass 5(6):301–328. doi: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. Geogr Compass 5(6):275–300. doi:10.1111/j.1749-8198.2011.00425.x

Xu Z, Yang ZL (2012) An improved dynamical downscaling method with GCM bias corrections and its validation with 30 years of climate simulations. J Clim 25(18):6271–6286. doi:10.1175/JCLI-D-12-00005.1