Vol. 48: 145–161, 2011
doi: 10.3354/cr00995

CLIMATE RESEARCH
Clim Res

Published August 30

*Contribution to CR Special 27 'Climate change in the NW Iberian Peninsula'*

OPEN ACCESS

# Validation of the ENSEMBLES global climate models over southwestern Europe using probability density functions, from a downscaling perspective

**S. Brands*, S. Herrera, D. San-Martín, J. M. Gutiérrez**

**Instituto de Física de Cantabria (CSIC – Universidad de Cantabria), 39005 Santander, Spain**

ABSTRACT: In this study we analyzed the performance of 12 state-of-the-art global climate models (GCMs) from 2 different model generations used in the ENSEMBLES project (a European Commission-funded climate-change research project) over southwestern Europe. For this purpose, we assessed the similarity of the simulated and quasi-observed (reanalysis) probability density functions for circulation, temperature, and humidity variables at various pressure levels, which we chose from a statistical-downscaling point of view. Our main goals were to assess which GCM variables can be reliably used as predictors for downscaling, and which GCMs perform especially well over the region under study. Results showed that specific humidity is as reliably reproduced as circulation and temperature variables, and that overall performance is best for the Hadley Centre's HADGEM2 model. Secondary goals were to estimate the skillful scale of the models, and to measure the added value of bias correction, a post-processing step commonly used in practice. We found that all models lack performance at the scale of individual grid boxes, indicating that they are not robustly skillful at their smallest scale. We also found that model performance generally improves after removing monthly bias. However, model errors at higher-order moments, which cannot be removed by simply correcting the bias, were common in some models.

KEY WORDS: Global Climate Models · Evaluation · Model performance measure · Climate change · Downscaling · Iberian Peninsula · Monte Carlo methods · Multi-model

## 1. INTRODUCTION

'All downscaling approaches will only be as accurate as the available GCM predictors' (Wilby et al. 1998, p. 17).

Nowadays, statistical downscaling (SD) is a sound and mature field that provides several techniques to use coarse-resolution global climate models (GCMs) or atmosphere-ocean GCMs (AOGCMs) on regional to local scales (Hewitson & Crane 1996, Wilby & Wigley 1997, Zorita & von Storch 1999, Maraun et al. 2010). These methods link the large-scale output of GCMs (predictors) with simultaneous local historical observations (predictands) in the region of interest.

Selecting appropriate large-scale predictor(s) is a key task of the SD approach. The choice depends on the area under study (Cavazos & Hewitson 2005), the predictand to be downscaled (Haylock et al. 2006), and the underlying data sets (Timbal et al. 2003). To date, most SD studies have been applied to mid-latitude climates. For these regions, some spatially robust predictors have been identified when working under optimal conditions (Cavazos & Hewitson 2005), i.e. taking the predictor data from quasi-observations which are typically represented by reanalysis data (Hewitson & Crane 1996, Wilby et al. 2004, Sauter & Venema in press).

However, little is known about how the predictive power of SD models trained on reanalysis data is affected, when they are applied to GCM data (Randall et al. 2007). In this case, the predictor choice made under optimal conditions has to be re-evaluated with respect to the following criteria:

*Email: brandssf@unican.es

(1) The applied GCMs should successfully reproduce the statistics of the reanalysis data set used to calibrate/train the statistical model.

(2) The predictor–predictand relationship, identified with reanalysis/observational records, should be time-invariant/stationary (Wilby 1997, Frías et al. 2006).

In the present work we focus on the first point and refer to this problem as 'model performance', i.e. the 'ability of AOGCMs to reproduce different aspects of present-day climate' (Giorgi & Mearns 2002, p. 1142), as characterized in reanalysis data. Due to a lack of general consensus, there currently exist many different GCM performance measures, which can be classified as those including (1) a single and (2) several of the time-series characteristics relevant for climate modeling (Giorgi & Mearns 2002, Räisänen et al. 2010). These are: climatological mean state (Randall et al. 2007, Gleckler et al. 2008), frequency of extreme events (Kharin & Zwiers 2000), seasonal cycle (Errasti et al. 2010), low frequency variability (Benestad 2003, Santer et al. 2008), and interannual variability (Gleckler et al. 2008). However, an optimal metric of overall model performance probably does not exist, as the usefulness of any validation approach depends on the intended application (Gleckler et al. 2008).

The main goal of the present study was to assess the ability of 12 state-of-the-art GCMs — from 2 different model generations used in the ENSEMBLES project, a European Commission-funded climate-change research project (van der Linden & Mitchell 2009) — to reproduce the probability density functions (PDFs) of large-scale variables taken from reanalysis data.

To this aim we used the PDF score proposed by Perkins et al. (2007) and Maxino et al. (2008). As the complete simulated PDF is validated, we evaluated the models' performance to reproduce the mean state as well as the frequency of extreme events. However, we neither validated low frequency nor interannual variability and our results have to be seen in this context.

Secondary goals of the present study were to estimate the 'skillful scale' (Grotch & MacCracken 1991, Benestad et al. 2008) of the models and to check the added value of a simple monthly bias correction, a post-processing step often applied to raw GCM data in order to cope with limited model performance (e.g. Demuzere et al. 2009).

The present study can be seen as a GCM performance guide, tailored to the downscaling and impact-assessment community working in southwestern Europe. In Section 2, the data used and the area under study are presented. Section 3 introduces some discussion about GCM validation from a downscaling perspective. Section 4 describes the validation and bias-correction procedure. Results are shown in Section 5, while the discussion and conclusions are in Section 6.

## 2. STUDY AREA AND DATA

In downscaling studies, the areal extent for which information is needed for the large-scale predictor variables may vary from a single grid-box (e.g. Abaurrea & Asín 2005) to a domain of subcontinental scale (e.g. Brands et al. 2011, this CR Special). We chose a domain that widely surrounds the northwestern Iberian Peninsula (Fig. 1). It extends from 30° N to 60° N and from 20° W to 5° E and is hereinafter called 'southwestern Europe'.

Daily GCM 20th-century control-run data were obtained from the ENSEMBLES Stream 1 and Stream 2 projects (Niehörster et al. 2008, van der Linden & Mitchell 2009). As for the variables to be validated, we chose geopotential (Z), temperature (T), relative humidity (R), specific humidity (Q), zonal (U) and meridional (V) wind components, as well as mean sea-level pressure (MSLP) (Table 1). Pressure levels at 850, 700, and 500 hPa were considered for all 3D variables. For Z, the 1000 hPa level (Z1000) was additionally included for the sake of comparison with MSLP. Although both Z1000 and MSLP may be readily available in repositories of reanalysis products, this is not necessarily the case for GCM databases. Consequently, if a multi-model ensemble of GCMs is applied (e.g. for downscaling purpose), both variables are often interchangeably used, assuming that they are equally well reproduced by a given GCM.

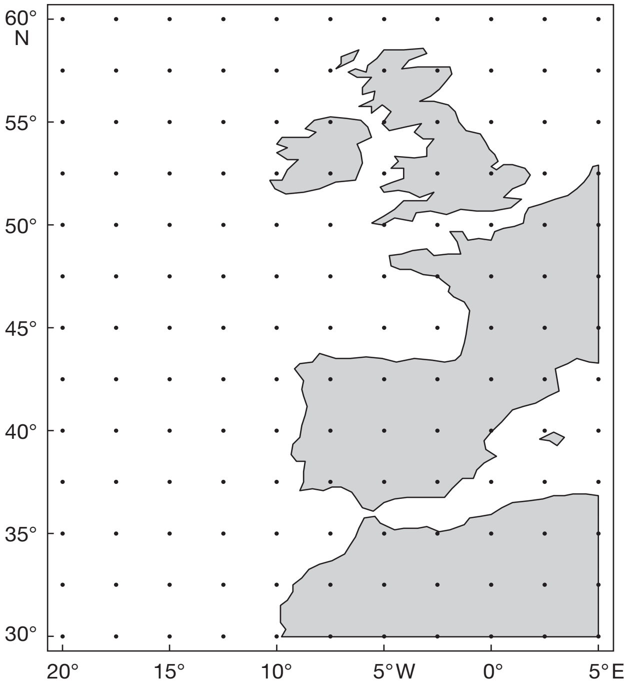


Fig. 1. Study area: southwestern Europe. Dots indicate the common 2.5° × 2.5° grid used in the study

Table 1. Variables analyzed in the present work. Inst.: instantaneous records

| Variable | Description | Pressure levels (hPa) | Units | Temporal aggregation |
|---|---|---|---|---|
| Z | Geopotential | 1000, 850, 700, 500 | $m^2 s^{-2}$ | Inst. at 00:00 h UTC |
| T | Temperature | 850, 700, 500 | K | Inst. at 00:00 h UTC |
| U | Zonal wind component | 850, 700, 500 | $m s^{-1}$ | Inst. at 00:00 h UTC |
| V | Meridional wind component | 850, 700, 500 | $m s^{-1}$ | Inst. at 00:00 h UTC |
| R | Relative humidity | 850, 700, 500 | % | Inst. at 00:00 h UTC |
| Q | Specific humidity | 850, 700, 500 | $kg kg^{-1}$ | Inst. at 00:00 h UTC |
| MSLP | Mean sea-level pressure | Sea level | Pa | Daily mean value |

Except MSLP, which is a daily mean value, all data are instantaneous records at 00:00 h UTC. Table 2 shows an overview of the GCMs and acronyms used; for detailed information about initial conditions, model physics, and external forcings, see the references listed in Table 2.

In accordance with the multi-model ensemble strategy (Randall et al. 2007), and in spite of common model components (Jun et al. 2008) in the ensemble used, each member (model) is assumed to be independent. We worked with fully coupled GCMs, i.e. we did not validate atmosphere-only GCMs.

The data span a 30 yr period from 1 January 1969 to 31 December 1998 and were obtained from the CERA database of the World Data Center for Climate, Hamburg (http://cera-www.dkrz.de/CERA/). In total, 6 GCMs were chosen from each stream of the ENSEMBLES project. Stream 1 models were those used in the Fourth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC-AR4), whereas Stream 2 models were developed within the ENSEMBLES project. If various runs of the same GCM were available, we selected only the first. Due to limited data availability, some of the GCM predictor variables could not be obtained for every model and/or pressure level and had to be excluded from the validation procedure. For an overview of the validated variables, see Table 3.

The main difference between the 2 generations of models is that Stream 2 GCMs include anthropogenic land-use change models, whereas Stream 1 models do not, the only exception being HADGEM (Niehörster et al. 2008, van der Linden & Mitchell 2009). Moreover, as outlined by Niehörster et al. (2008), all GCMs are driven by anthropogenic forcing, while natural external forcing due to the solar cycle and episodic great-volcano eruptions are not taken into account. Although the anthropogenic forcing agents are slightly different between some GCMs (e.g. forcing of sulfate aerosols is taken into account by all models except EGMAM and EGMAM2; Huebener & Koerper 2008, Niehörster et al. 2008), we think that this factor is probably negligible for the results reported in the present paper. A detailed analysis with varying forcing configurations would be needed to fully address this issue, but this is beyond the scope of the present paper.

As our study is written from a downscaling point of view, we chose the European Centre for Medium-Range Weather Forecasts (ECMWF) ERA-40 reanalysis as a reference dataset for validation, while being well aware of possible quality problems, especially concerning relative humidity (Ben Daoud et al. 2009). Reanalysis data represent the only quasi-observational data resource that offers a wide range of predictor vari-

Table 2. Overview of the global climate models (GCMs) used in the present study, taken from the 2 streams of the ENSEMBLES project. Stream 1: model versions from the Fourth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC-AR4); Stream 2: new versions developed within the ENSEMBLES project

| GCM name | Acronym | Stream | Institution | Source |
|---|---|---|---|---|
| BCCR-BCM2 | BCM2 | 1 | Bjerknes Centre for Climate Research, Norway | Drange (2006) |
| CNRM-CM3 | CNCM3 | 1 | Centre National de Recherches Météorologiques, France | Royer (2006) |
| ECHO-G | EGMAM | 1 | Freie Universität Berlin, Germany | Niehörster (2008) |
| IPSL-CM4 | IPCM4 | 1 | Institut Pierre Simon Laplace, France | Dufresne (2007) |
| METO-HC-HadGEM | HADGEM | 1 | Met Office Hadley Centre, UK | Johns (2008) |
| MPI-ECHAM5 | MPEH5 | 1 | Max Planck Institute for Meteorology, Germany | Roeckner (2007) |
| CNRM-CM33 | CNCM33 | 2 | Centre National de Recherches Météorologiques, France | Royer (2008) |
| ECHO-G2 | EGMAM2 | 2 | Freie Universität Berlin, Germany | Huebener & Koerper (2008) |
| IPSL-CM4v2 | IPCM4V2 | 2 | Institut Pierre Simon Laplace, France | Dufresne (2009) |
| METO-HC-HadCM3C | HADCM3C | 2 | Met Office Hadley Centre, UK | Johns (2009a) |
| METO-HC-HadGEM2 | HADGEM2 | 2 | Met Office Hadley Centre, UK | Johns (2009b) |
| MPI-ECHAM5C | MPEH5C | 2 | Max Planck Institute for Meteorology, Germany | Roeckner (2008) |

Table 3. List of variables (by pressure level in hPa) available for the 12 models. See Tables 1 & 2 for abbreviations. x: available

| Model | Variable | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | T | | | Q | | | R | | | Z | | | U | | | | V | | | MSLP |
| | 850 | 700 | 500 | 850 | 700 | 500 | 850 | 700 | 500 | 850 | 700 | 500 | 1000 | 850 | 700 | 500 | 850 | 700 | 500 | sea level |
| MPEH5 | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x |
| MPEH5C | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x |
| HADGEM | x | | x | | | | x | x | x | x | x | x | | x | | | x | | x | |
| HADGEM2 | x | x | x | x | x | x | x | x | x | x | x | x | | x | x | x | x | x | x | x |
| HADCM3C | x | x | x | | | | x | x | x | x | x | x | x | x | x | x | | x | x | x |
| CNCM3 | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x |
| CNCM33 | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x |
| BCM2 | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x |
| EGMAM | x | x | x | x | x | x | | | | x | x | x | x | x | x | x | x | x | x | x |
| EGMAM2 | x | x | x | x | x | x | | | | x | x | x | x | x | x | x | x | x | x | x |
| IPCM4 | x | x | x | x | x | x | | | | x | x | x | | x | x | x | x | x | x | x |
| IPCM4V2 | x | x | x | x | x | x | | | | x | x | x | | x | x | x | x | x | x | x |

ables at the different levels needed for an SD study (e.g. Wilby et al. 2004), especially in the climate-change context (Hewitson & Crane 2006). Ideally, GCMs should additionally be validated against radiosonde (Ben Daoud et al. 2009) and/or satellite (Brogniez & Pierrehumbert 2007) data to estimate observational uncertainties.

ERA-40 data were obtained from the ECMWF MARS server at their native resolution of 1.125° × 1.125°. In contrast, the native horizontal resolution of the GCMs varies between 1.25 and 3.75° and comes on regular or Gaussian grids. Therefore, all predictor data were regridded on a common regular 2.5° × 2.5° lattice by using bilinear interpolation. Gleckler et al. (2008) highlighted that validation results are sensitive to the choice of grid size in the case of precipitation, while for MSLP they are not. Consequently, if spatially redundant variables (like MSLP) are evaluated, as is the case in the present study, sensitivity to different grid sizes is probably negligible. Outliers were defined as values above or below 10 times the interquartile range (IQR) and subsequently set to 'not a number' values. In accordance with Huth (2005) and Ben Daoud et al. (2009), we found negative values and values well above 100 % for R in both the reanalysis and GCM control-run data, as obtained from the ECMWF MARS server and CERA database (see Section 2). Before the regridding process, we corrected them to 0 and 100 % respectively.

For the sake of simplicity, we refer to 'observations' when talking about ERA-40 reanalysis data and to 'simulations' when focusing on GCM data.

## 3. VALIDATING GCMs FROM A DOWNSCALING PERSPECTIVE

Although there exist plenty of GCM validation studies with a regional focus (Giorgi & Mearns 2002, Perkins et al. 2007, Errasti et al. 2010), they are of limited practical value for the downscaling community and, in particular, the SD community. The main reason is that most of these studies focus only on surface variables (e.g. Perkins et al. 2007) but do not validate middle tropospheric variables, which are commonly used as predictors in the downscaling process. Moreover, most studies work with monthly averages, which are of practical value on a daily timescale only if weather generators are used (Semenov & Stratonovitch 2010). Finally, if a wide range of daily variables is evaluated, validation is usually restricted to a single model or model family (e.g. Ringer et al. 2006).

Consequently, a study assessing which of the predictor variables identified in optimal downscaling conditions, i.e. by using reanalysis data (Cavazos & Hewitson 2005, Sauter & Venema 2011), are reliably reproduced by state-of-the-art GCMs, is still missing, to our knowledge. This is an important issue, since GCM errors are transmitted through the downscaling scheme and affect the downscaled time series (Chen et al. 2006, Brands et al. 2011), as well as subsequent impact studies (Beaumont et al. 2008).

Among the predictors to be validated, we included 2 variables, Q and R, which on the one hand have considerable predictive power for downscaling in optimal conditions (Cavazos & Hewitson 2005), but on the other hand are assumed to be poorly reproduced by the current GCM generation (Maraun et al. 2010). This is an important issue, since humidity variables should be included as predictors in order to yield realistic climate-change signals (Charles et al. 1999, Hewitson & Crane 2006). In contrast, the use of circulation predictors alone yields downscaled projections that should be treated with caution, as the circulation's simulated response to greenhouse gas forcing is negligible (Wilby et al. 1998, Goodess & Jones 2002)—at least with predictions until the middle of the 21st century.

## 4. VALIDATION METHODS

To assess GCM performance, we compared simulated and observed PDFs for each grid box, using the reliability score proposed by Perkins et al. (2007) and Maxino et al. (2008).

For a given grid box, both ERA-40 and GCM data were classified in 200 discrete bins, spanning the whole range of both series. The probability density for each class and for both simulated and observed time series was estimated by kernel density smoothing (Wilks 2006). In accordance with Perkins et al. (2007) and Maxino et al. (2008), we used MatLab's *ksdensity* function for computation. Gaussian kernels and a width parameter optimized for normal distributions were considered. Simulated and observed probability densities were then compared for each bin, retaining each pair's minimum. The resulting sample of minima was summed. The total yields a value of 1 for a perfect fit of both PDFs and 0 in the case of no overlap:

$$\text{PDF score} = \frac{1}{n}\sum_{i=1}^{n}\min(\text{PDF}f_i - \text{PDFo}_i) \qquad (1)$$

where PDF score = GCM – performance metric, suggested by Perkins et al. (2007); PDF$f_i$ = stimulated normalised probability density for the $i$th bin; and PDFo$_i$ = observed normalised probability density for the $i$th bin.

Besides providing the summarized validation results for the entire domain, the PDF score was mapped at each grid box for a set of key predictor variables, permitting a geographical interpretation of the results. This ensures that different users can check the GCM performance for a specific region of interest.

As errors along the whole distribution are taken into account, the PDF score is preferable to classic validation measures, e.g. bias, if overall GCM performance is to be assessed. However, because of its absolute nature, no information is given about the 'direction' of the error, i.e. if the distribution of the GCM data is displaced to the left or to the right. Furthermore, the PDF score is similar to the 'linear error in probability space' (Potts et al. 1996), with respect to both calculation and interpretation. Thus, it gives very little weight to errors committed at the tails of the distribution, which are of crucial importance for impact studies and the corresponding adaption strategies (Easterling et al. 2000). Therefore, we recommend to use the PDF score as a first-guess validation measure of a model's performance and to calculate complementary metrics (e.g. Goodess 2005), tailored to the specific purpose of the study.

To correct the raw (original) output of a given GCM, its monthly bias was removed using the following technique. At each gridbox, the ERA-40 monthly mean was first added to each timestep of the GCM time series. Then, the GCM's own monthly mean was subtracted from each timestep of the resulting time series. PDF scores were calculated as well for these corrected data and the following Monte Carlo test was applied to reveal the significance of the performance increase after bias correction.

For a given variable, season, pressure level, and grid box, 1000 synthetic PDF scores were generated by bootstrapping (Efron 1982) both the observed and uncorrected (= biased) GCM series 1000 times and then calculating the PDF scores upon these 1000 pairs of resampled time series. The percentage of these artificially generated PDF scores exceeding the PDF score calculated from the corrected (unbiased) GCM data can be interpreted as the p-value of the test. The bias-corrected data are assumed to be significantly better than the raw data at a 95% confidence level if this percentage is <5%.

When simultaneously applying individual statistical hypothesis tests at a great number of geographical locations, the significance level to be employed is much lower than the significance level which actually is intended to be met (Katz 1992). To take into account this so-called problem of 'multiplicity' (Tukey 1977), we repeated the above-mentioned hypothesis test, assuming the improvement through bias correction to be significant if none of the above-mentioned bootstrap estimates of the PDF score for raw data exceeded the PDF score for the bias-corrected data. This is equivalent to a significance level <0.01% at the grid-box scale. As the results of both procedures were very similar, we decided to focus on the significance level of 5% at the grid-box scale.

## 5. RESULTS

Figs. 2 & 3 show a summary of the PDF scores for each GCM, variable, and pressure level on the whole domain. Each subplot corresponds to a specific variable and pressure level, e.g. Z at 500 hPa (Z500). GCMs are displayed on the *x*-axes and PDF scores on the *y*-axes.

For a given subplot, 4 pairs of bars and error bars were plotted for each GCM, corresponding to the validation results for winter (DJF), spring (MAM), summer (JJA), and autumn (SON). Each bar stands for the median of a sample of 143 PDF scores (= 143 grid boxes) and thus describes the central tendency of the validation results. The associated error bar corresponds to the IQR of the sample and hence refers to the spatial dispersion of the validation results. These error bars can be interpreted as a measure of the 'skillful scale' (Grotch & MacCracken 1991, Benestad et al. 2008) of the GCMs: the smaller the lower tails, the less skillful is the GCM at the grid-box scale.

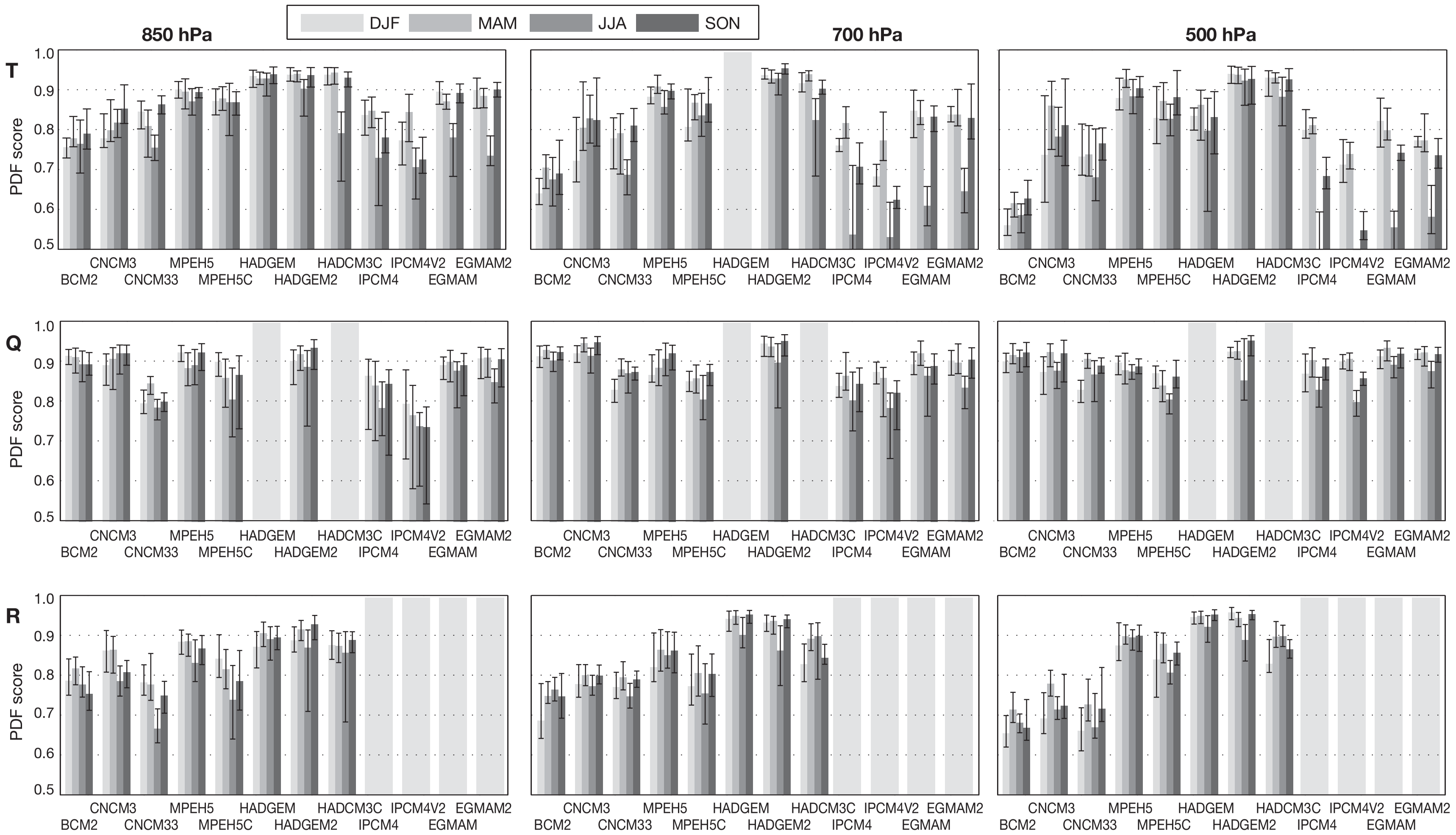


Fig. 2. Probability density function (PDF) score by season for temperature (T), specific humidity (Q), and relative humidity (R) (rows) at different pressure levels (columns) for 12 global climate models (GCMs). Bars indicate the median PDF scores over the whole domain, and error bars show the corresponding interquartile range (IQR). DJF: winter, MAM: spring, JJA: summer, SON: autumn. Grey-shaded blank columns = no data available in the CERA database

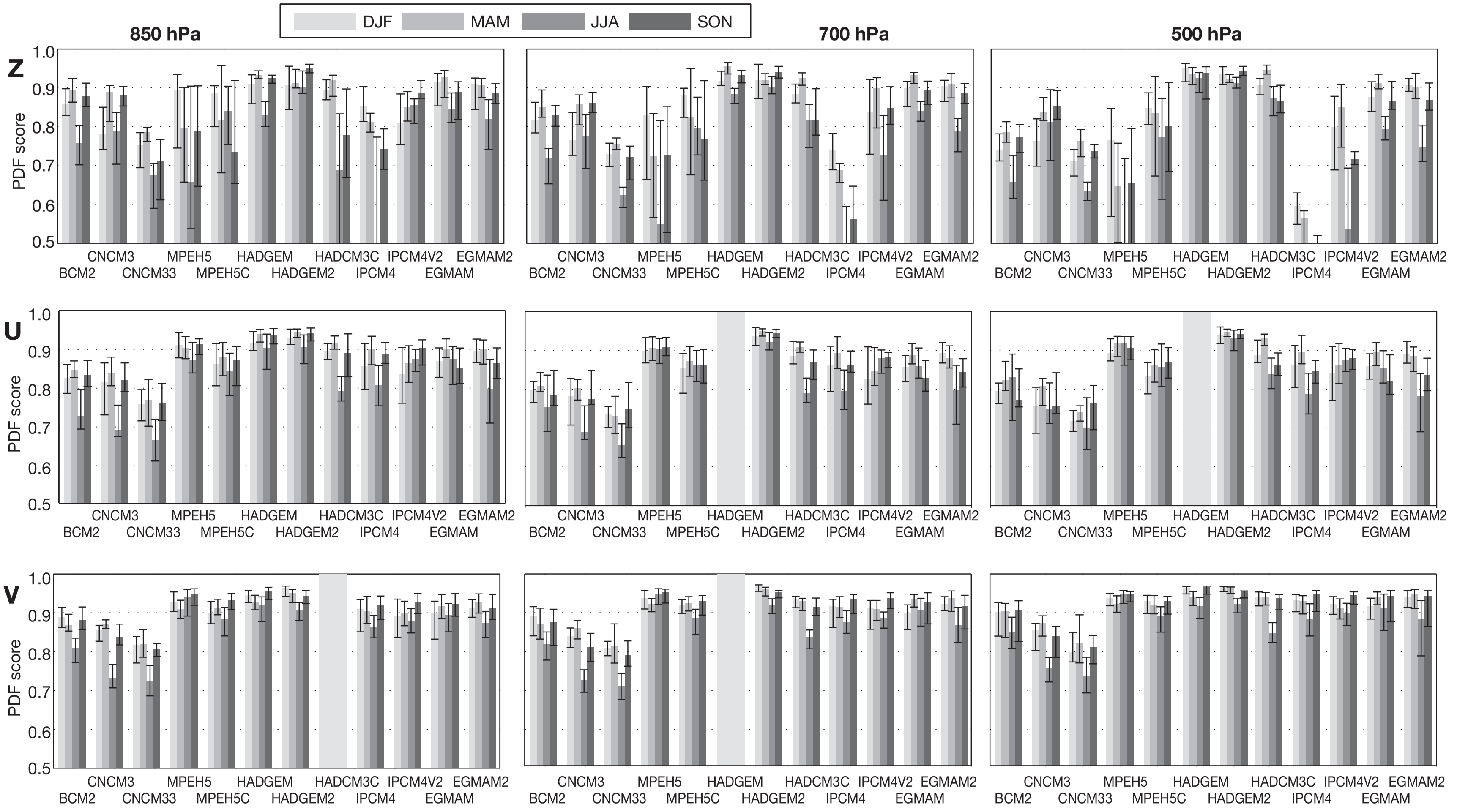


Fig. 3. Probability density function (PDF) score by season for geopotential (Z), zonal wind component (U), and meridional wind component (V) (rows) at different pressure levels (columns) for 12 global climate models (GCMs). Bars indicate the median PDF scores over the whole domain, and error bars show the corresponding interquartile range (IQR). DJF: winter, MAM: spring, JJA: summer, SON: autumn. Grey-shaded blank columns = no data available in the CERA database

Looking at both median values (data bars) and dispersions (error bars) of the PDF scores, Q was as reliably or better reproduced than T and Z (Figs. 2 & 3). GCM performance for R was, in general, lower than for Q (Fig. 2) and was most robust for U and V (Fig. 3).

If a model performed comparatively worse for Z and T at 850 hPa, its error grew with height. These negative-outlier models were MPEH5 and IPCM4 for Z and BCM2, IPCM4, and IPCM4V2 for T (Figs. 2 & 3).

Besides the height dependence, performance also varied with season, with summer results being the worst among all seasons for most models, variables, and pressure levels (Figs. 2 & 3). Remarkably, the seasonal dependence/sensitivity of the performance with the season was most pronounced for the just-mentioned negative-outlier models.

MSLP and Z1000 were reproduced nearly identically by each GCM (Fig. 4). HADCM3C, which performed comparatively well for most middle-tropospheric predictors (Figs. 2 & 3), failed to reproduce both MSLP and Z1000 in summer. This is consistent with the model's comparatively low PDF scores for Z850 (Fig. 3).

In terms of median PDF score (bars in Figs. 2 to 4), HADGEM2 outperformed or equalled any other individual GCM in at least 3 of 4 seasons for virtually all variables and heights. Results for HADGEM were comparable, but all other GCMs showed equivalent performance only in particular cases.

However, when focusing on the range of the PDF scores, and hence referring to the spatial dispersion of the GCM's reliability, results were less favorable for HADGEM and HADGEM2, especially for the summer season. In particular for the humidity variables, but also for T500, T850, and V850, the performance of both models considerably diverges towards lower values, indicating limited reliability at individual grid boxes. Moreover, every single GCM suffered from low outliers well below the 25th percentile (lower limit of the error bars/IQR) at the grid-box scale. To visualize this problem, the validation results for predictor variables frequently used in SD studies (MSLP, Z500, T850, and Q850) were mapped (Fig. 5).

Fig. 5 shows the PDF scores for the uncorrected (original) and bias-corrected (unbiased) GCM data at each grid box. A grid box is marked with a black dot if the PDF score for the corrected data was significantly (α = 5%) higher than for its uncorrected counterpart, i.e. if the bias-correction procedure led to a significant improvement in model performance.

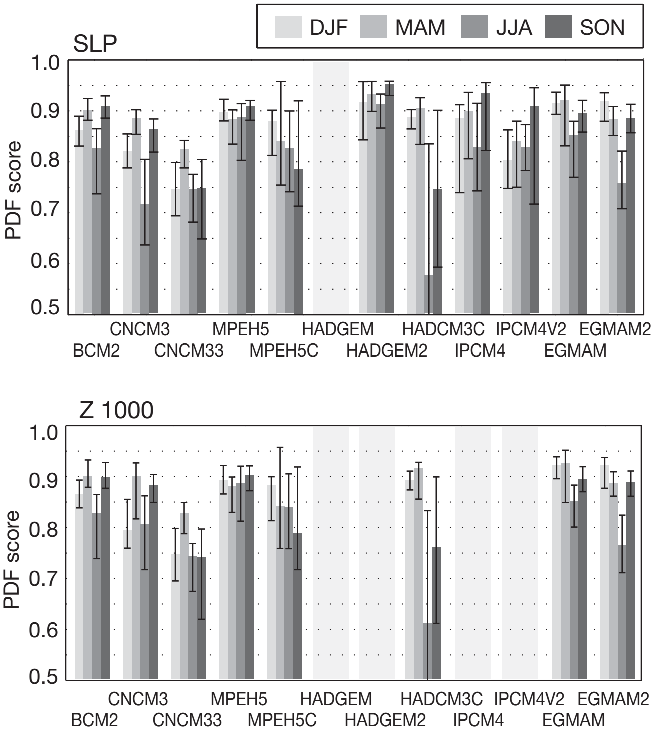


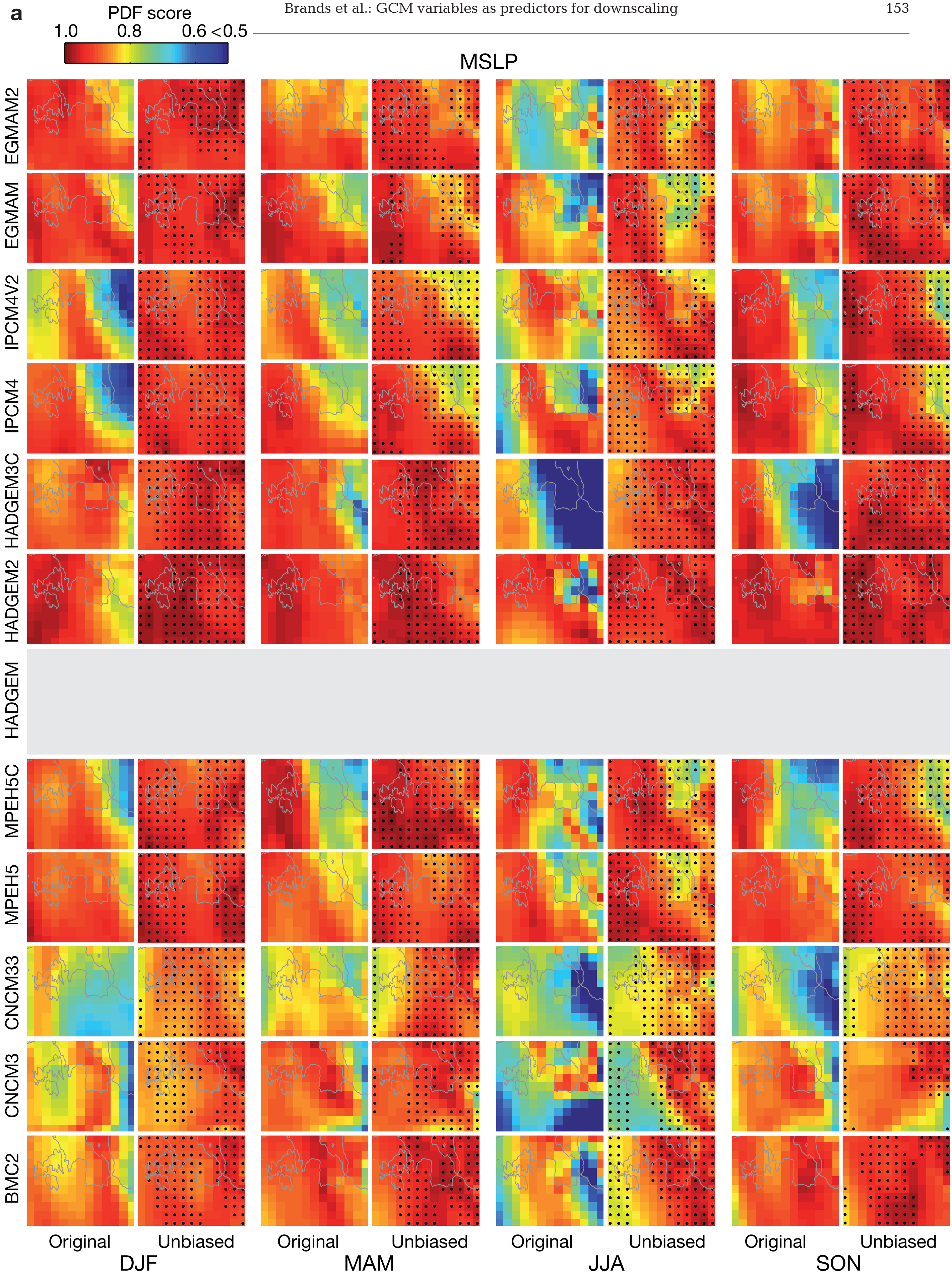
Fig. 4. Probability density function (PDF) score by season for mean sea-level pressure (MSLP) and geopotential at 1000 hPa (Z1000) for 12 global climate models (GCMs). Bars indicate the median PDF scores over the whole domain, and error bars show the corresponding interquartile range (IQR). DJF: winter, MAM: spring, JJA: summer, SON: autumn. Grey-shaded blank columns = no data available in the CERA database

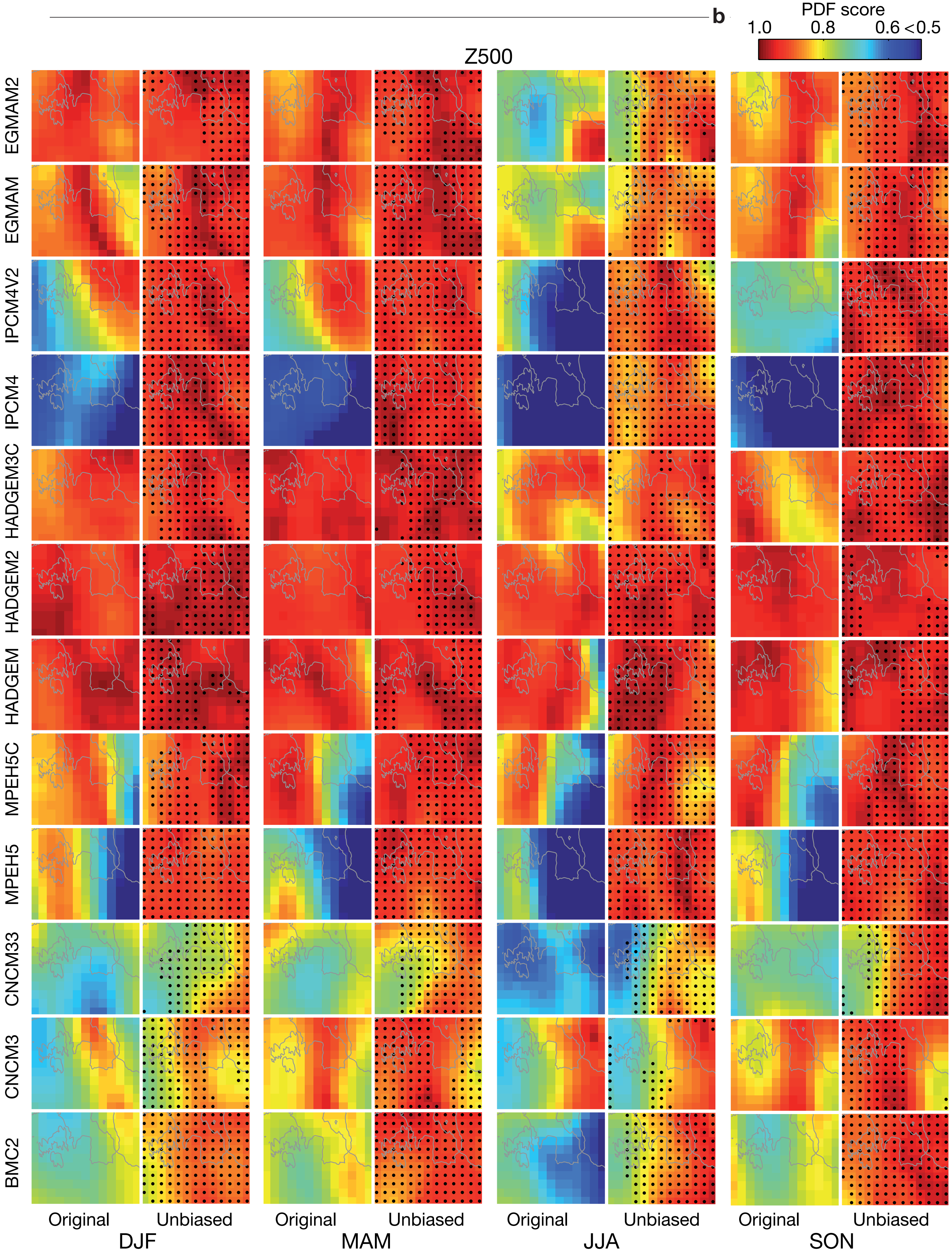
Focusing on the uncorrected data (maps with 'original' in Fig. 5), the lowest PDF scores have the tendency to cluster over southern Spain and/or northwestern Africa. For HADGEM and HADGEM2, this clustering is especially pronounced in summer.

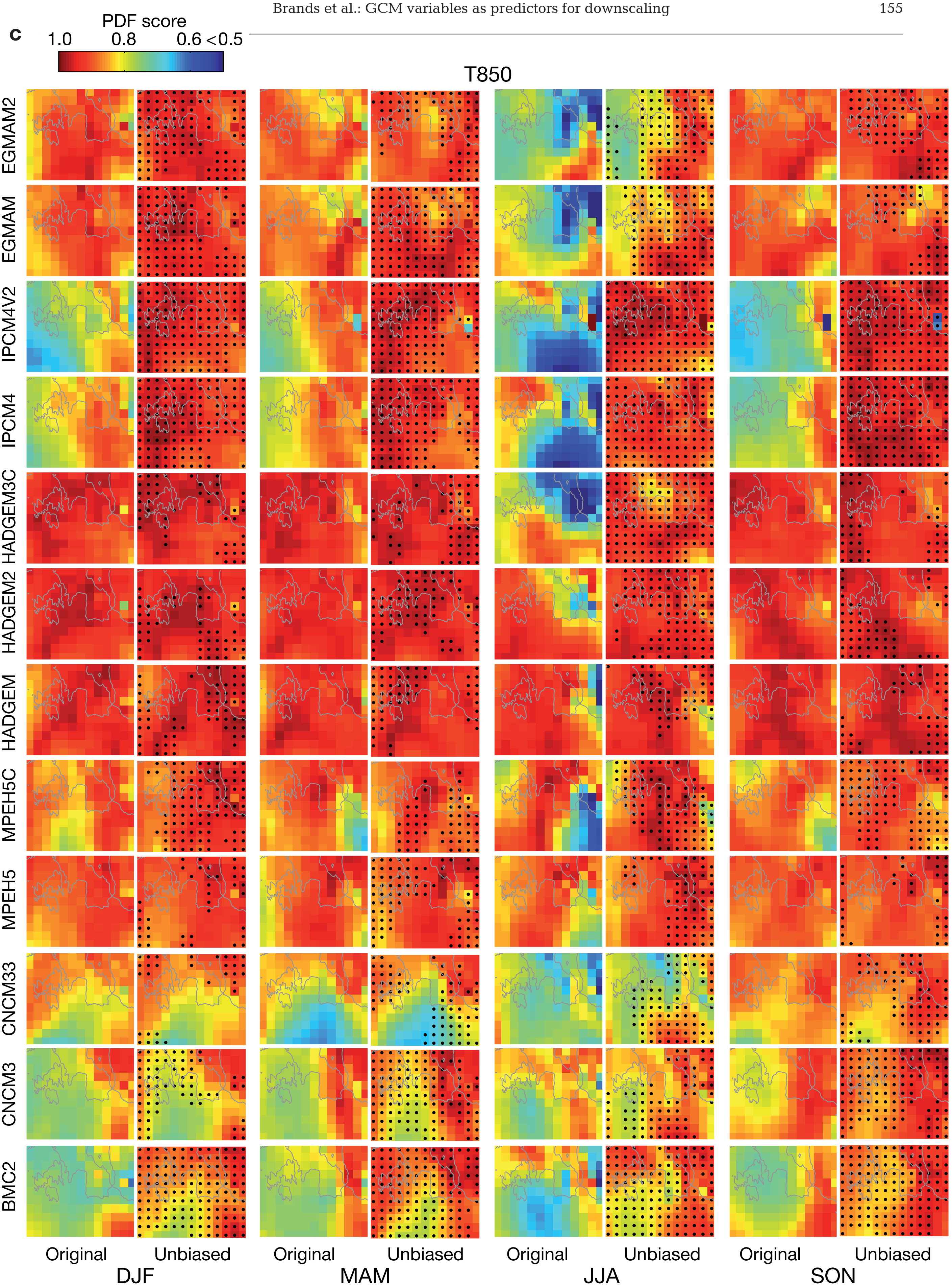
Obviously, the PDF scores are spatially correlated, in particular for Z500 and MSLP. For Z500, MSLP, and Q850, performance roughly decreased from north to south in all seasons. In contrast, for T850, performance increased in the same direction during all seasons except JJA.

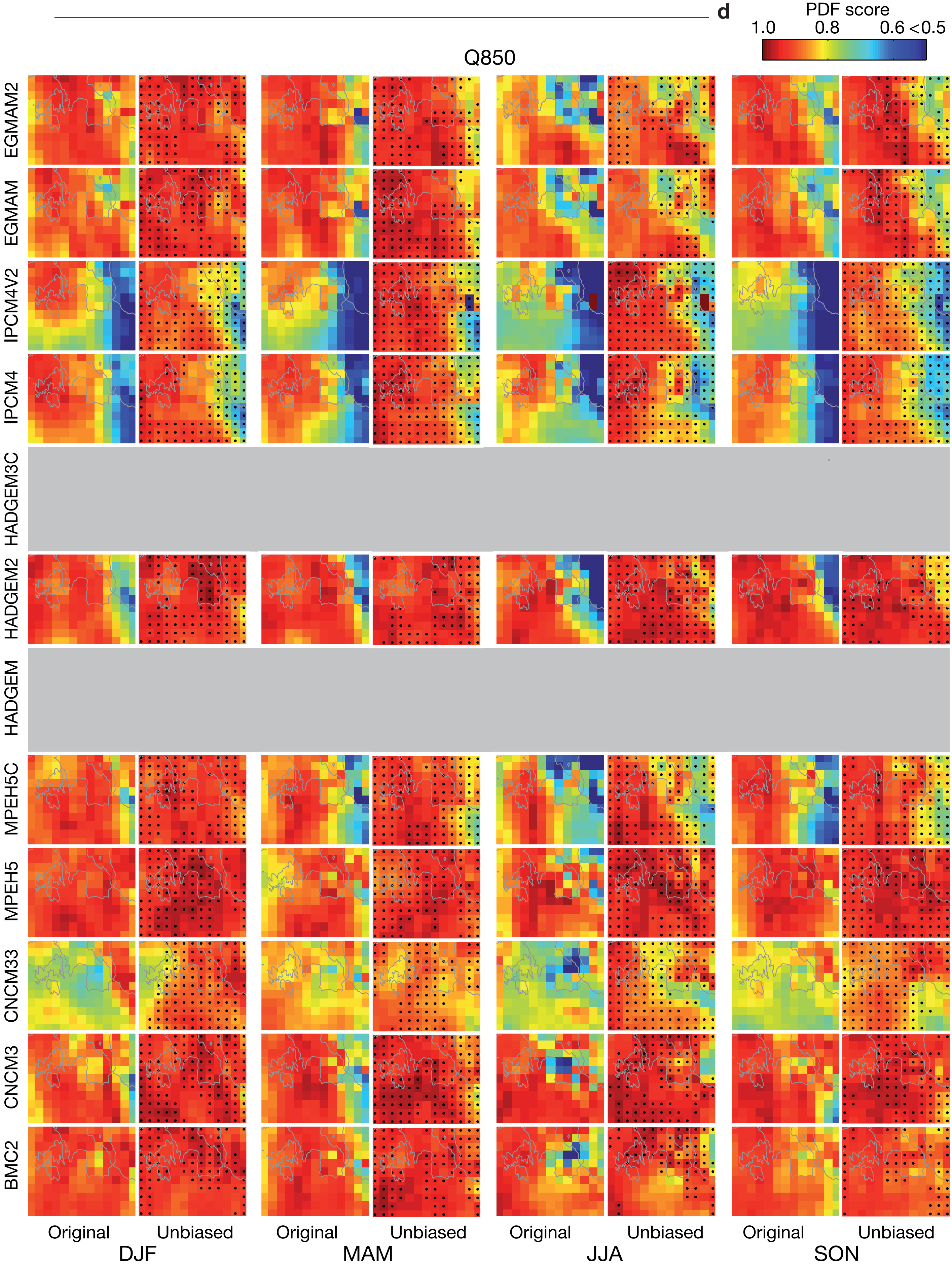
In general, bias correction significantly enhanced model performance (maps marked 'unbiased' in Fig. 5). However, this is not the case if the GCM errors lie in higher-order moments, which can be shown by comparing the JJA results for Z500, as simulated by

Fig. 5. (Next 4 pages) Mapped probability density function (PDF) score for uncorrected (original) and monthly bias-corrected (unbiased) (a) mean sea-level pressure (MSLP), (b) geopotential at 500 hPa (Z500), (c) temperature at 850 hPa (T850) and (d) specific humidity at 850 hPa (Q850) data as simulated by 12 global climate models (GCMs). A grid box is marked with a black dot if the PDF score for the corrected data is significantly (α = 5%) higher than for its uncorrected counterpart. DJF: winter, MAM: spring, JJA: summer, SON: autumn. Grey-shaded blank columns = no data available in the CERA database


a
PDF score
1.0
0.8
0.6
<0.5
MSLP
EGMAM2
EGMAM
IPCM4V2
IPCM4
HADGEM3C
HADGEM2
HADGEM
MPEH5C
MPEH5
CNCM33
CNCM3
BMC2
Original
Unbiased
Original
Unbiased
Original
Unbiased
Original
Unbiased
DJF
MAM
JJA
SON


b
PDF score
1.0
0.8
0.6
<0.5
Z500
EGMAM2
EGMAM
IPCM4V2
IPCM4
HADGEM3C
HADGEM2
HADGEM
MPEH5C
MPEH5
CNCM33
CNCM3
BMC2
Original
Unbiased
DJF
Original
Unbiased
MAM
Original
Unbiased
JJA
Original
Unbiased
SON


c
PDF score
1.0
0.8
0.6
<0.5
T850
EGMAM2
EGMAM
IPCM4V2
IPCM4
HADGEM3C
HADGEM2
HADGEM
MPEH5C
MPEH5
CNCM33
CNCM3
BMC2
Original
Unbiased
Original
Unbiased
Original
Unbiased
Original
Unbiased
DJF
MAM
JJA
SON


d
PDF score
1.0
0.8
0.6
<0.5
Q850
EGMAM2
EGMAM
IPCM4V2
IPCM4
HADGEM3C
HADGEM2
HADGEM
MPEH5C
MPEH5
CNCM33
CNCM3
BMC2
Original
Unbiased
DJF
Original
Unbiased
MAM
Original
Unbiased
JJA
Original
Unbiased
SON

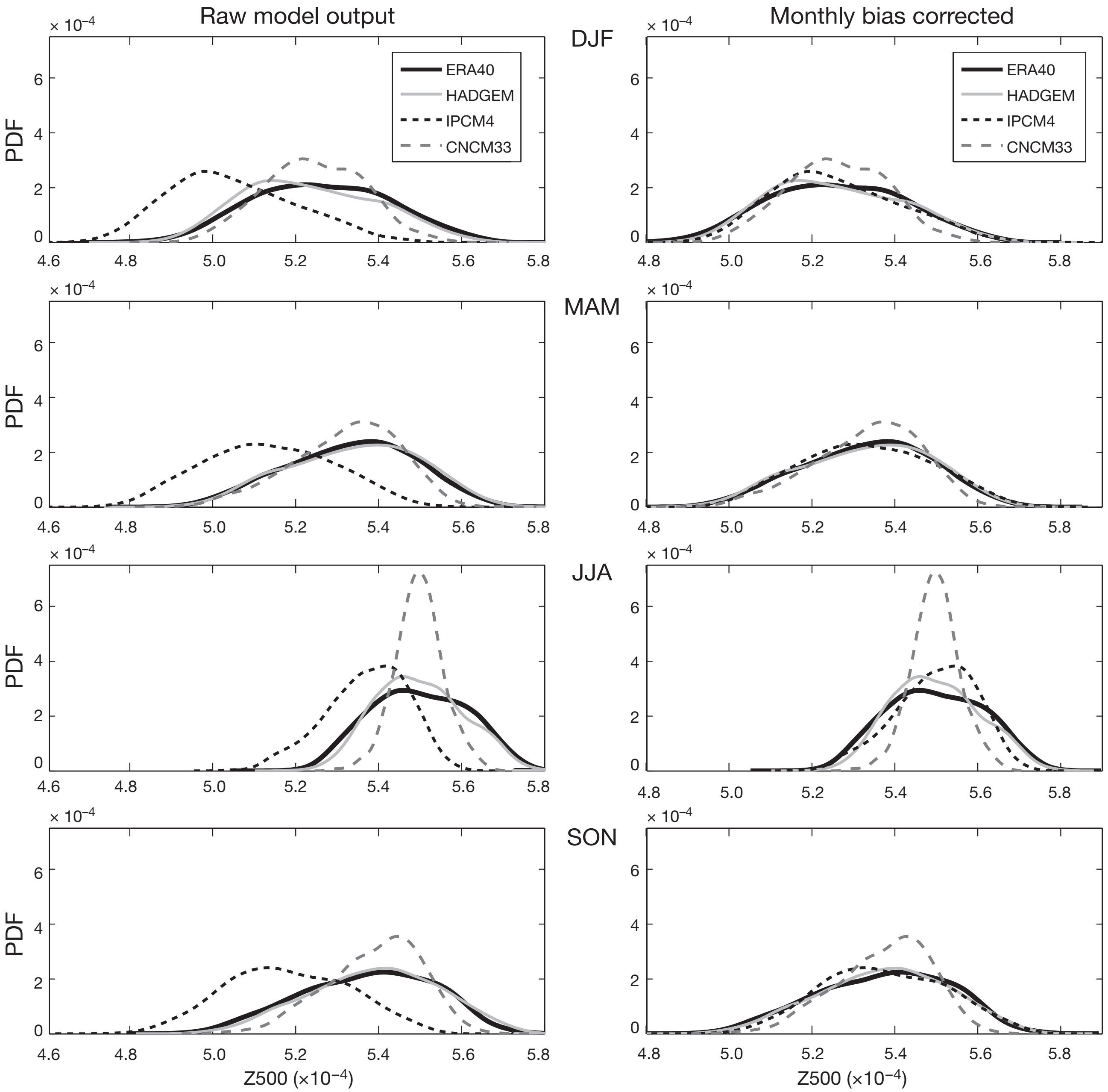


Fig. 6. Comparison of probability density functions (PDFs) for ERA-40, HADGEM, IPCM4, and CNCM33 data for geopotential at 500 hPa (Z500) at the northeastern extreme of the domain. PDFs were calculated for the original (left) and monthly bias-corrected (right) global climate model (GCM) data by season. Errors for CNCM33 are in higher-order moments and therefore cannot be corrected by removing the bias. DJF: winter, MAM: spring, JJA: summer, SON: autumn

MPEH5 and MPEH5C, to those simulated by CNCM3 and CNCM33 (Fig. 5b). The former models lacked performance in the southern part of the domain, while the latter had problems in the northern and western parts. After bias correction, the PDF scores in these problematic zones can be significantly improved for MPEH5 and MPEH5C. However, this was not the case for CNCM3 and CNCM33, indicating that they erroneously simulate higher-order moments.

Another example is shown in Fig. 6, which shows the PDFs of Z500 for ERA-40, HADGEM, IPCM4, and CNCM33 at the northeastern extreme of the domain, for both original and bias-corrected data. While bias correction improved the fit of IPCM4's PDF, its effect was negligible for CNCM33 and HADGEM: the former overestimates the kurtosis and underestimates the variance, while the latter shows a good fit for the uncorrected data that cannot be improved by removing the bias.

Consequently, if scientists decide to correct GCM output, removing the bias is not sufficient in every case, and more complex correction methods, like 'quantile mapping' (Themeßl et al. 2011), should be taken into account.

To compare the overall reliability of a given GCM, PDF scores of Z, T, Q, U, and V at 850, 700, and 500 hPa, as well as of MSLP, were calculated for each grid box and then joined into a single sample for both original and unbiased data. The medians and IQRs of these joined samples are visualized in Fig. 7.

In terms of median PDF score, and considering the uncorrected data (data bars in Fig. 7 upper panel), HADGEM2 performed best and CNCM33 worst in all seasons. In summer, performance was worst and negative deviations towards low PDF values (error bars in Fig. 7 upper panel) were most pronounced for all models. For that season, negative deviations were highest for IPCM4, followed by MPEH5 and IPCM4V2, and lowest for HADGEM2. With median and IQR values for the PDF scores being almost identical, results for EGMAM and EGMAM2 were very similar in every season of the year.

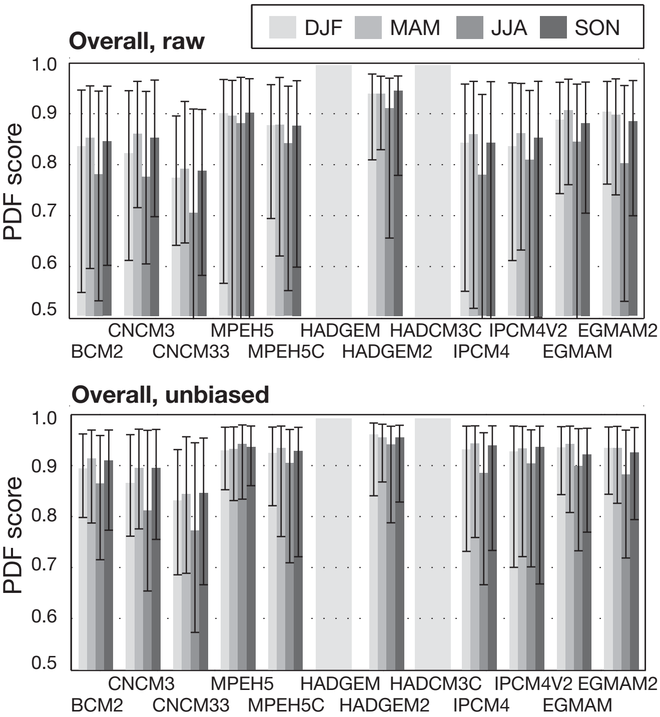


Fig. 7. Overall performance in terms of probability density function (PDF) score of 12 global climate models (GCMs) for the uncorrected (upper panel) and monthly bias-corrected (lower panel) data by season. Bars indicate median values, and error bars show the corresponding interquartile range (IQR). DJF: winter, MAM: spring, JJA: summer, SON: autumn. Grey-shaded blank columns = no data available in the CERA database

When compared to the performance of the original (uncorrected) data, the performance of the unbiased data (Fig. 7 lower panel) generally improves, which leads to more uniform PDF scores across the models. However, due to the errors in higher-order moments that could not be corrected by bias removal, CNCM3, CNCM33 and, to a lesser degree, BCM2 exhibit poorer overall performances.

In 2 out of 4 cases, the Stream 1 version of a given model performed better than its Stream 2 correspondent (MPEH5 vs. MPEH5C and CNCM3 vs. CNCM33), while in the remaining 2 cases (EGMAM vs. EGMAM2 and IPCM4 vs. IPCM4V2), nearly identical values are yielded (Fig. 7 upper panel). However, in order to state a general conclusion about the 'added value' of the Stream 2 GCM generation, model intercomparison should be based on a wider range of regions and validation measures.

## 6. DISCUSSION AND CONCLUSIONS

In the present study, the ability of ENSEMBLES GCMs to reproduce the PDFs of observed predictor variables used in SD studies was assessed for southwestern Europe.

The PDF score suggested by Perkins et al. (2007) and Maxino et al. (2008) has proven to be applicable for the validation of a wide range of GCM variables/distributions. In spite of its limited sensitivity for errors at the extreme tails of the PDF, the PDF score detects gross errors in the distribution's form and consequently can be recommended as a measure of overall model performance.

As a first result, Q was as reliably simulated as Z and T. As a predictor in SD studies, Q should be preferred to R, as the latter suffers from data-quality problems, such as negative values and values well above 100 percent, and is less reliably reproduced. The inferior performance for R, when compared to Q, can be explained by R's dependence on T, described by the Clausius-Clapeyron equation. Thus, if a model lacks performance for T, it will equally do for R (see Fig. 2).

Performance for MSLP and Z1000 was very similar and comparatively high. Consequently, these variables are mutually interchangeable. The variables Z, T, Q, and MSLP cannot be unrestrictedly recommended for downscaling from GCM scenario runs, as particularly worse-performing 'outlier' models are present for each of them, at least for some seasons of the year and/or pressure levels. In turn, for U and V, performance was comparatively robust throughout all GCMs.

Badly performing grid boxes clustered in northwestern Africa and southern Spain, especially in the case of

MSLP. This may reflect serious problems in the models' capacity to reproduce the Saharan Heat Low, a key synoptic feature for the region under study. A more detailed assessment of the physical causes of this error is recommended for future research.

For the region under study, in an overall comparison, HADGEM2, followed by MPEH5, yielded the best performance. CNCM3 and CNCM33 performed worst and suffered from frequent errors in higher-order moments. This highlights the need for GCM validation studies on a daily timescale and indicates that, for some models and/or geographical areas, and from the PDF point of view, bias correction alone is not sufficient to significantly enhance model performance.

As an update to Grotch & MacCracken (1991), even the best overall performing models are not robustly skillful at their smallest scale, which underlines the importance of validating, and optionally correcting, GCM data before downscaling it. This issue is of particular importance for downscaling techniques working with a single or only a few grid boxes in the context of climate change.

In comparison to Perkins et al. (2007), who validated GCM precipitation and maximum and minimum temperature for 12 regions over Australia with essentially the same PDF score, our results neither support the notable performance of ECHO-G (EGMAM in the present study), nor that of IPSL (IPCM4 in the present study). Comparison to Maxino et al. (2008), who conducted a study similar to Perkins et al. (2007), leads to a similar conclusion: IPCM4, showing notable performance over Australia, is not outstanding at all over southwestern Europe. This underlines the well-known sensitivity of GCM performance to the area under study, an issue which has already been stated by Perkins et al. (2007), Maxino et al. (2008), Gleckler et al. (2008), and Knutti et al. (2010).

In comparison to Errasti et al. (2010), who validated the PDF of monthly mean values with essentially the same score for a very similar geographical domain, our results support the outstanding mean performance of HADGEM and MPEH5, as well as the only-moderate performance of CNCM3. In addition, our validation results for MPEH5 show remarkable spatial spread, leading to substantial performance decreases at individual grid boxes.

In contrast to Errasti et al. (2010), BCM2 was not found to have outstanding performance in our study. First, this may be explained by the different time resolution underlying the validation. In our study (daily timescale), errors of higher-order moments were detected for BCM2 (e.g. for Q850), which cannot be identified by validating monthly mean values (Errasti et al. 2010). Second, surface variables, e.g. 2 m air temperatures, which were validated by Errasti et al. (2010), are highly dependent on the reanalysis' and GCM's surface orography and land-sea mask. Therefore, they may be substituted by pressure-level variables like temperatures at 1000 hPa. Third, and most important, the overall ranking is highly sensitive to the choice of variables (Gleckler et al. 2008, Knutti et al. 2010). This is a crucial point, since a general agreement on the most important variables (and pressure levels) to be validated, as well as on how to reduce redundant information, is still lacking (Knutti et al. 2010). Thus, while the comparatively high PDF scores for BCM2 in simulating MSLP found by Errasti et al. (2010) do match our results (Fig. 5), the same model's performance is not outstanding at all for most variables of the free troposphere (Figs. 2 & 3).

*Acknowledgements.* We especially thank the 2 anonymous reviewers who helped to considerably improve former versions of our manuscript. We also thank all authors of the MeteoLab Toolbox (www.meteo.unican.es/software/meteolab), which has shown to be very useful to post-process and validate GCM data sets. Acknowledgements go to the World Data Center for Climate (WDCC) for providing the ENSEMBLES Stream 1 GCM data (www.mad.zmaw.de/wdc-for-climate/) and the European Centre for Medium-Range Weather Forecasts (ECMWF) for offering the ERA-40 reanalysis data (http://data-portal.ecmwf.int/data/d/era40_daily). This work was funded by the strategic action for energy and climate change by the Spanish R+D 2008–2011 'Programa coordinado para la generación de escenarios regionalizados de cambio climático: Regionalización Estadística' (esTcena), code 2008 00050084078, and the CICYT project CGL2010-21869.

## LITERATURE CITED

➤ Abaurrea J, Asín J (2005) Forecasting local daily precipitation patterns in a climate change scenario. Clim Res 28:183–197

➤ Beaumont LJ, Hughes L, Pitman AJ (2008) Why is the choice of future climate scenarios for species distribution modelling important? Ecol Lett 11:1135–1146

➤ Ben Daoud B, Sauquet E, Lang M, Obled C, Bontron G (2009) Comparison of 850-hPa relative humidity between ERA-40 and NCEP/NCAR re-analysis: detection of suspicious data in ERA-40. Atmos Sci Lett 10:43–47

➤ Benestad RE (2003) What can present climate models tell us about climate change? Clim Change 59:311–331

Benestad RE, Hanssen-Baur I, Chen D (2008) Empirical-statistical downscaling. World Scientific, Singapore

➤ Brands S, Taboada JJ, Cofiño AS, Sauter T, Schneider C (2011) Statistical downscaling of daily temperatures in the northwestern Iberian Peninsula from general circulation models: validation and future scenarios. Clim Res 48: 163–176

➤ Brogniez H, Pierrehumbert RT (2007) Intercomparison of tropical tropospheric humidity in GCMs with AMSU-B water vapour data. Geophys Res Lett 34:L17812 doi:10.1029/2006GL029118

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

➤ Charles SP, Bryson BC, Whetton PH, Hughes JP (1999) Validation of downscaling models for changed climate conditions: case study of southwestern Australia. Clim Res 12: 1–14

Chen D, Achberger C, Räisänen J, Hellström C (2006) Using statistical downscaling to quantify the GCM-related uncertainty in regional climate change scenarios: a case study of Swedish precipitation. Adv Atmos Sci 23:54–60

Demuzere M, Werner M, van Lipzig NPM, Roeckner E (2009) An analysis of present and future ECHAM5 pressure fields using a classification of circulation patterns. Int J Climatol 29:1796–1810

Drange H (2006) ENSEMBLES BCCR-BCM2.0 20C3M run1, daily values. CERA database. World Data Center for Climate, Hamburg. Available at: http://cera-www.dkrz.de/WDCC/ui/Entry.jsp?acronym=ENSEMBLES_BCM2_20C3M_1_D

Dufresne JL (2007) ENSEMBLES IPSL-CM4 20C3M run1, daily values. CERA database. World Data Center for Climate, Hamburg. Available at: http://cera-www.dkrz.de/WDCC/ui/Compact.jsp?acronym=ENSEMBLES_IPCM4_20C3M_1_D

Dufresne JL (2009) ENSEMBLES STREAM2 IPSLCM4_v2 20C3M run1, daily values. CERA database. World Data Center for Climate, Hamburg. Available at: http://cera-www.dkrz.de/WDCC/ui/Compact.jsp?acronym=ENSEMBLES2_IPCM4v2_20C3M_1_D

Easterling DR, Meehl GA, Parmesan C, Changnon SA, Karl TR, Mearns LO (2000) Climate extremes: observations, modeling and impacts. Science 289:2068–2074

Efron B (1982) The jackknife, the bootstrap, and other resampling plans. Society for Industrial and Applied Mathematics, Philadelphia, PA

Errasti I, Ezcurra A, Saénz J, Ibarra-Berastegi G (2010) Validation of IPCC AR4 models over the Iberian Peninsula. Theor Appl Climatol 103:61–79

Frías MD, Zorita E, Fernández J, Rodríguez-Puebla C (2006) Testing statistical downscaling methods in simulated climates. Geophys Res Lett 33:L19807 doi: 10.1029/2006GL027453

Giorgi F, Mearns LO (2002) Calculation of average, uncertainty range, and reliability of regional climate changes from AOGCM simulation via the 'Reliability Ensemble Averaging' (REA) method. J Clim 15:1141–1158

Gleckler PF, Taylor KE, Doutriaux C (2008) Performance metrics for climate models. J Geophys Res Atmos 113:D06104 doi: 10.1029/2007JD008972

Goodess CM (2005) STARDEX, downscaling climate events. University of East Anglia, Norwich. Available at: www.cru.uea.ac.uk/projects/stardex/reports/STARDEX_FINAL_REPORT.pdf

Goodess CM, Jones PD (2002) Links between circulation and changes in the characteristics of Iberian rainfall. Int J Climatol 22:1593–1615

Grotch S, MacCracken M (1991) The use of general circulation models to predict regional climatic change. J Clim 4:286–303

Haylock MR, Cawley GC, Harpham C, Wilby RL, Goodess CM (2006) Downscaling heavy precipitation over the United Kingdom: a comparison of dynamical and statistical methods and their future scenarios. Int J Climatol 26:1397–1415

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

Hewitson BC, Crane RG (2006) Consensus between GCM climate change projections with empirical downscaling: precipitation downscaling over South Africa. Int J Climatol 26:1315–1337

Huebener H, Koerper J (2008) ENSEMBLES STREAM2 EGMAM2 20C3M run1, daily values. CERA database. World Data Center for Climate, Hamburg. Available at: http://cera-www.dkrz.de/WDCC/ui/Entry.jsp?acronym=ENSEMBLES2_FUBEMA2_20C3M_1_D

Huth R (2005) Downscaling of humidity variables: a search for suitable predictors and predictands. Int J Climatol 25: 243–250

Johns TC (2008) ENSEMBLES METO-HC-HADGEM1 20C3M run1, 6h and 12h instantaneous values. CERA database. World Data Center for Climate, Hamburg. Available at: http://cera-www.dkrz.de/WDCC/ui/Compact.jsp?acronym=ENSEMBLES_HADGEM_20C3M_1_6H12H

Johns TC (2009a) ENSEMBLES STREAM2 METO-HC-HADCM3C 20C3M run1, daily values. CERA database. World Data Center for Climate, Hamburg. Available at: http://cera-www.dkrz.de/WDCC/ui/Entry.jsp?acronym=ENSEMBLES2_HADCM3C_20C3M_1_D

Johns TC (2009b) ENSEMBLES STREAM2 METO-HC-HADGEM2AO 20C3M run1, daily values. CERA database. World Data Center for Climate, Hamburg. Available at: http://cera-www.dkrz.de/WDCC/ui/Compact.jsp?acronym=ENSEMBLES2_HADGEM2_20C3M_1_D

Jun MR, Knutti R, Nychka DW (2008) Spatial analysis to quantify numerical model bias and dependence. J Am Stat Assoc 103:934–947

Katz RW (1992) Role of statistics in the validation of general circulation models. Clim Res 2:35–45

Kharin VV, Zwiers FW (2000) Changes in the extremes in an ensemble of transient climate simulations with a coupled atmosphere-ocean GCM. J Clim 13:3760–3788

Knutti R, Furrer R, Tebaldi C, Cermak J, Meehl GA (2010) Challenges in combining projections from multiple climate models. J Clim 23:2739–2758

Maraun D, Wetterhall F, Ireson AM, Chandler RE and others (2010) Precipitation downscaling under climate change. Recent developments to bridge the gap between dynamical models and the end user. Rev Geophys 48:RG3003 doi: 10.1029/2009RG000314

Maxino CC, McAvaney BJ, Pitman AJ, Perkins SE (2008) Ranking the AR4 climate models over Murray-Darling Basin using simulated maximum temperature, minimum temperature and precipitation. Int J Climatol 28:1097–1112

Niehörster F (2008) ENSEMBLES EGMAM 20C3M run1, daily values. CERA database. World Data Center for Climate, Hamburg. Available at: http://cera-www.dkrz.de/WDCC/ui/Compact.jsp?acronym=ENSEMBLES_FUBEMA_20C3M_1_D

Niehörster F, Fast I, Huebener H, Cubasch U (2008) The stream one ENSEMBLES projections of future climate change. ENSEMBLES Tech Rep 3. ENSEMBLES Project Office, Met Office, Exeter. Available at: http://ensembles-eu.metoffice.com/tech_reports.html

Perkins SE, Pitman AJ, Holbrook NJ, McAneney J (2007) Evaluation of the AR4 climate models' simulated daily maximum temperature, minimum temperature, and precipitation over Australia using probability density functions. J Clim 20:4356–4376

Potts JM, Folland CK, Jolliffe IT, Sexton D (1996) Revised 'LEPS' scores for assessing climate model simulations and long-range forecasts. J Clim 9:34–53

Räisänen J, Ruokolainen L, Ylhäisi J (2010) Weighting of model results for improving best estimates of climate change. Clim Dyn 35:407–422

Randall DA, Wood RA, Bony S, Colman R and others (2007) Climate models and their evaluation. In: Solomon S, Qin D, Manning M, Chen Z and others (eds) Climate change 2007: the scientific basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, p 589–662

Ringer MA, Martin GM, Greeves CZ, Hinton TJ and others (2006) The physical properties of the atmosphere in the new Hadley Centre Global Environmental Model (HADGEM1). Part II: aspects of variability and regional climate. J Clim 19:1302–1326

Roeckner E (2007) ENSEMBLES ECHAM5-MPI-OM 20C3M run1, daily values. CERA database. World Data Center for Climate, Hamburg. Available at: http://cera-www.dkrz.de/WDCC/ui/Compact.jsp?acronym=ENSEMBLES_MPEH5_20C3M_1_D

Roeckner E (2008) ENSEMBLES STREAM2 ECHAM5C-MPI-OM 20C3M run1, daily values. CERA database. World Data Center for Climate, Hamburg. Available at: http://cera-www.dkrz.de/WDCC/ui/Compact.jsp?acronym=ENSEMBLES2_MPEH5C_20C3M_1_D

Royer JF (2006) ENSEMBLES CNRM-CM3 20C3M run1, daily values. CERA database. World Data Center for Climate, Hamburg. Available at: http://cera-www.dkrz.de/WDCC/ui/Compact.jsp?acronym=ENSEMBLES_CNCM3_20C3M_1_D

Royer JF (2008) ENSEMBLES STREAM2 CNRM-CM33 20C3M run1, daily values. CERA database. World Data Center for Climate, Hamburg. Available at: http://cera-www.dkrz.de/WDCC/ui/Compact.jsp?acronym=ENSEMBLES2_CNCM33_20C3M_1_D

Santer BD, Thorne PW, Haimberger L, Taylor KE and others (2008) Consistency of modelled and observed temperature trends in the tropical troposphere. Int J Climatol 28:1703–1722

Sauter T, Venema V (in press) Natural three-dimensional predictor domains for statistical precipitation downscaling. J Clim doi: 10.1175/2011JCLI4155.1

Semenov M, Stratonovitch P (2010) Use of multi-model ensembles from global climate models for assessment of climate change impacts. Clim Res 41:1–14

Themeßl MJ, Gobiet A, Leuprecht A (2011) Empirical-statistical downscaling and error correction of daily precipitation from regional climate models. Int J Climatol 31:1530–1544

Timbal B, Dufour A, McAvaney B (2003) An estimate of future climate change for western France using a statistical downscaling technique. Clim Dyn 20:807–823

Tukey JW (1977) Some thoughts on clinical trials, especially problems of multiplicity. Science 198:679–684

Uppala SM, Kållberg PW, Simmons AJ, Andrae U and others (2005) The ERA-40 re-analysis. Q J R Meteorol Soc 131:2961–3012

van der Linden P, Mitchell JFB (eds) (2009) ENSEMBLES: climate change and its impacts: summary of research and results from the ENSEMBLES project. ENSEMBLES Project Office, Met Office Hadley Centre, Exeter. Available at: http://ensembles-eu.metoffice.com/docs/Ensembles_final_report_Nov09.pdf (~50 MB)

Wilby RL (1997) Non-stationarity in daily precipitation series: implications for GCM down-scaling using atmospheric circulation indices. Int J Climatol 17:439–454

Wilby RL, Wigley TML (1997) Downscaling general circulation model output: a review of methods and limitations. Prog Phys Geogr 21:530–548

Wilby RL, Hassan H, Hanaki K (1998) Statistical downscaling of hydrometeorological variables using general circulation model output. J Hydrol (Amst) 205:1–19

Wilby RL, Charles SP, Zorita E, Timbal B, Whetton P, Mearns LO (2004) Guidelines for use of climate scenarios developed from statistical downscaling methods: supporting material of the Intergovernmental Panel on Climate Change. Task Group on Data and Scenario Support for Impacts and Climate Analysis, Rotherham

Wilks DS (2006) Statistical methods in the atmospheric sciences, 2nd edn. Elsevier Academic Press, Amsterdam

Zorita E, von Storch H (1999) The analog method as a simple statistical downscaling technique: comparison with more complicated methods. J Clim 12:2474–2489

*Submitted: August 9, 2010; Accepted: January 17, 2011*

*Proofs received from author(s): August 8, 2011*