DOI: 10.1007/s10584-016-1683-4
The final publication is available at:
http://link.springer.com/article/10.1007/s10584-016-1683-4

Climatic Change manuscript No.
(will be inserted by the editor)

Towards a fair comparison of statistical and dynamical downscaling in the framework of the EURO-CORDEX initiative

A. Casanueva · S. Herrera · J. Fernández · J.M. Gutiérrez

Received: date / Accepted: date

Abstract Both statistical and dynamical downscaling methods are well established techniques to bridge the gap between the coarse information produced by global circulation models and the regional-to-local scales required by the climate change Impacts, Adaptation, and Vulnerability (IAV) communities. A number of studies have analyzed the relative merits of each technique by inter-comparing their performance in reproducing the observed climate, as given by a number of climatic indices (e.g. mean values, percentiles, spells). However, in this paper we stress that fair comparisons should be based on indices that are not affected by the calibration towards the observed climate used for some of the methods.

We focus on precipitation (over continental Spain) and consider the output of eight Regional Climate Models (RCMs) from the EURO-CORDEX initiative at 0.44° resolution and five Statistical Downscaling Methods (SDMs) —analog resampling, weather typing and generalized linear models— trained using the *Spain044* observational gridded dataset on exactly the same RCM grid. The performance of these models is inter-compared in terms of several standard indices —mean precipitation, 90th percentile on wet days, maximum precipitation amount and maximum number of consecutive dry days— taking into account the parameters involved in the SDM training phase. It is shown, that not only the directly affected indices should be carefully analyzed, but also those indirectly influenced (e.g. percentile-based indices for precipitation) which are more difficult to identify.

A. Casanueva (✉) · S. Herrera · J. Fernández
Grupo de Meteorología. Dpto. Matemática Aplicada y Ciencias de la Computación. Univ. de Cantabria, Avda. de los Castros, s/n, 39005 Santander, Spain
Tel.: +34 942201723
E-mail: ana.casanueva@unican.es

J.M. Gutiérrez
Grupo de Meteorología. Instituto de Física de Cantabria. CSIC-Univ. de Cantabria, Avda. de los Castros, s/n, 39005 Santander, Spain

We also analyze how simple transformations (e.g. linear scaling) could be applied to the outputs of the uncalibrated methods in order to put SDMs and RCMs on equal footing, and thus perform a fairer comparison.

Keywords Regional Climate Models · statistical downscaling · EURO-CORDEX · precipitation indices

1 Introduction

Different climate downscaling techniques have been developed since the early 1990s to bridge the gap between the large-scale climate information provided by Global Circulation Models (GCMs) and the regional-to-local scale required for climate impacts assessment (see Maraun et al, 2010, and references therein). Two fundamentally different downscaling techniques have been followed for this purpose: 1) dynamical methods, based on Regional Climate Models (RCMs, Giorgi, 2006; Feser et al, 2011) and 2) Statistical Downscaling Methods (SDMs, von Storch et al, 1993; Wilby and Wigley, 1997). A number of comparison studies have been carried out in the past to assess the relative merits of these two techniques (see e.g. Kidson and Thompson, 1998; Murphy, 1999; Goodess, 2005; Haylock et al, 2006; Schmidli et al, 2007; Tryhorn and DeGaetano, 2011; Hertig et al, 2012; Pizzigalli et al, 2012; Ayar et al, 2015). However, most of these comparisons do not take into account the important differences of these methods when analyzing the results.

RCMs numerically solve the governing equations of the atmosphere in a limited spatial domain, driven by boundary conditions taken from GCMs (or from reanalysis, in the model evaluation phase). Apart from the dynamical core, the RCMs include physical parameterizations for the subgrid processes which occur at spatial scales smaller than the model grid spacing (microphysics, convection, radiation, etc.). In most cases, these parameterizations are tuned based on model evaluation against the available observations for the region of interest (typically gridded temperature and precipitation datasets).

SDMs build on empirical relationships between model variables (predictors) and local point (or gridded) observed predictands of interest. Various conceptually different statistical methods and training approaches have been proposed in the literature to establish these relationships. Under the Perfect Prognosis (PP) approach, the statistical relationships are calibrated in a training phase considering observations for both predictands (historical observations) and predictors (reanalysis data), whereas model (GCM or RCM) predictions are used for the latter under the Model Output Statistics (MOS) approach. On the one hand, the predictors for PP are typically large-scale variables characterizing the circulation for the target area and well represented by both reanalysis and GCMs (see e.g. Brands et al, 2012). A number of methods —including linear and nonlinear regression, weather types, analog re-sampling, and combinations of them— have been proposed to establish the statistical relationships using (daily or monthly) pairwise predictor-predictand time series under this approach. On the other hand, the typical predictor in MOS is directly the variable of interest, which is calibrated against the local observed counterpart. In the climate change context this is typically done using distribution (e.g. mean- or quantile-mapping) corrections —this is usually referred to as *(distributional) bias correction* in the literature.— However, more sophisticated MOS methods also consider circulation predictors and regression or analog techniques to establish the statistical relationships from pairwise time series (Turco et al, 2011), as typically done in weather forecasting applications.

Statistical downscaling methods rely on different assumptions and each of them has several advantages and limitations (Estrada et al, 2013). However, unlike RCMs, SDMs are calibrated in a training phase using some sort of optimization or re-sampling process (or establishing a correction function in bias correction methods) involving the available observations (see e.g. Maraun et al, 2010). As a result, these methods are trained with local observations to reproduce some observed statistics, which are directly affected by the particular calibration process (i.e. optimization, re-sampling, or distribution-mapping process). The affected statistics vary from method to method, thus posing additional constraints for a fair validation and inter-comparison. For instance, the mean is adjusted in standard regression methods —or the mean and variance when considering stochastic or variance inflation variants (McCullagh and Nelder, 1989).— Order statistics are affected by methods suitable for extremes (such as quantile regression, Tareghian and Rasmussen (2013)). The whole distribution is fitted to the observed data —affecting all quantiles of the distribution (Déqué, 2007)— in the case of distributional empirical bias correction methods. Recent studies analyze the transferability of correction approaches to different climate conditions based on more sophisticated cross-validation methods in present climate (e.g. the method is calibrated in the driest/coldest years and validated in the wettest/warmest, on the lines of Gutiérrez et al, 2013; Teutschbein and Seibert, 2013). However, good performance during the calibration period does not guarantee a good performance under changed future conditions (Teutschbein and Seibert, 2012). This is due to the stationarity (time invariance) assumption of the correction, that is not likely to be met under climate change conditions, together with the finite length of the calibration period that may not cover the entire spectrum of the variable of interest (Ehret et al, 2012). Thus, the direct comparison of the different downscaling approaches using indices differently affected by the training process is particularly problematic if the distributions of the training and test subsets are similar in comparison with the future distributions of the climate projections where the methods will be applied.

A fair comparison of RCMs and SDMs has the additional complication of their different spatial representativeness. SDMs provide information at the spatial scale given by the observations (i.e. point stations or grids), whereas RCM results are areal-representative (of the model grid boxes) and, therefore, cannot represent the local variability of point stations (Luo et al, 2013). For this reason, recent studies acknowledge that a fair comparison of RCMs and SDMs requires the use of observational gridded data sets for SDMs calibration and both techniques evaluation (Schmidli et al, 2007; Hertig et al, 2012; Ayar et al, 2015). However, a direct comparison of SDM results for a local station with those for the nearest grid box of an RCM (as e.g. Kidson and Thompson, 1998; Murphy, 1999; Haylock et al, 2006; Tryhorn and DeGaetano, 2011; Pizzigalli et al, 2012) could derive misleading conclusions. The EURO-CORDEX initiative (Jacob et al, 2014) provides an appropriate framework for a fair comparison since a common grid was used for all RCMs and gridded observational products are available over the same grid —such as the European-wide *E-OBS* dataset (Haylock et al, 2008) or the *Spain02 v4* family of EURO-CORDEX-compliant gridded datasets over Spain (Herrera et al, 2015)—. This framework eases the fair comparison of SDMs and RCMs on the same grid, as shown e.g. in Ayar et al (2015).

In the above mentioned studies, SDMs and RCMs were compared without bringing into question whether the indicators considered in the comparison were influenced by the calibration or tuning of the downscaling methods. As far as we know, there is no previous comprehensive comparison study taking this factor into account. In this paper we shed light on this problem and describe an inter-comparison experiment for precipitation over Spain considering eight EURO-CORDEX RCMs at a 0.44° resolution and five PP SDMs trained using the Spain044 gridded observation data in a cross-validation form. The methods considered include an analog resampling technique and four methods based on a Bernoulli (for occurrence) and a Gamma (for amount) distributions, fitted to the data conditioned to circulation in different forms. Therefore, the training process of the SDMs used in this study only affects directly the mean and distribution shape of the precipitation amount, except for the analog method which affects various aspects of the distribution due to its resampling nature. By doing this, we keep the number of parameters affected in the training phase as small as possible, unlike other methods that calibrate the whole distribution. Moreover, in order to analyze the potential impact of the adjustment of these statistics, the comparison is also performed after the application of two basic bias correction methods to both statistical and dynamical downscaling for precipitation frequency and intensity.

This paper is structured as follows. In Section 2 we present the data and methods used. The results are given in Section 3. Finally, the conclusions and summary are presented in Section 4.

2 Data and Methods

2.1 Observational Data

In this work we used precipitation data from the new EURO-CORDEX-compliant gridded daily observational dataset *Spain044* (Herrera et al, 2012, 2015) defined on the 0.44° resolution rotated grid used in the EURO-CORDEX initiative as a common basis for the RCM runs. *Spain044* is part of the *Spain02 v4* products (freely available from `http://www.meteo.unican.es/datasets/spain02`), which are based on a dense network of quality-controlled stations in Spain, covering the period 1971-2008. In order to ensure area-averaged representativeness of the resulting gridbox values, the interpolation method (full monthly 3D thin plate splines plus ordinary kriging on the daily anomalies) was carried out on an auxiliary 0.01° grid, averaging the results afterwards to the final 0.44° resolution grid. Therefore, this dataset is appropriate for the evaluation of the EURO-CODEX RCMs and it is also suitable for statistical downscaling.

2.2 Regional Climate Models

In this work, daily precipitation values from the freely-available RCM simulations within the EURO-CORDEX initiative at 0.44° resolution were downloaded from the ESGF archive (`http://esgf.org/`) in January 2015 (see Table 1). In particular we considered the simulations driven by the ERA-Interim reanalysis (Dee et al, 2011) covering the common period 1990-2008. Notice that this ensemble contains two versions of the WRF model, with different microphysics and radiation schemes but the same convection parameterization. We refer the reader to Table 1 in Kotlarski et al (2014) for further details on the particular model configurations.

Note that 0.44° resolution RCM simulations were considered instead of the state-of-the-art 0.11° runs since previous studies (e.g. Casanueva et al, 2015) have shown limited evidence of added value of the high resolution for this region in this kind of analysis.

Table 1 EURO-CORDEX RCMs used in the study. Codes are used to label RCMs in the figures.

Code	*RCM*	*Institution*	*Reference*
D1	CCLM 4.8.17	COSMO-CLM Community	Rockel et al (2008)
D2	HIRHAM 5	Danish Meteorological Institute, Denmark	Christensen et al (2007)
D3	RACMO 2.2	Royal Netherlands Meteorological Institute, Ministry of Infrastructure and the Environment, Netherlands	Meijgaard et al (2012)
D4	RCA 4	Swedish Meteorological and Hydrological Institute, Sweden	Samuelsson et al (2011)
D5	HadRM 3P	Met Office Hadley Centre, Exeter, UK	Collins et al (2006)
D6	ALADIN 52	Hungarian Meteorological Service, Hungary	Radu et al (2008)
D7	WRF 3.3.1.F	Institut Pierre Simon Laplace / Institut National de l'Environment Industriel et des Risques, France	Skamarock et al (2008)
D8	WRF 3.3.1.G	University of Cantabria, Spain	Skamarock et al (2008)

2.3 Statistical Downscaling Methods

In this study we built on the work done by San-Martín et al (2016) who tested different predictor configurations (both variables and geographical domains) for an ensemble of SDMs in Spain. In particular, we considered the best performing configuration of predictors, formed by sea level pressure (SLP), and temperature and specific humidity at 850 hPa (T850 and Q850, respectively), defined on a geographical domain covering the Iberian peninsula —from 10W to 5E and from 35N to 45N—. Moreover, predictor values at the start and end of the observation period (i.e. data at 00UTC at day D and $D+1$) were included to characterize each particular day D, thus forming a dynamic temporal set up. Predictor values were obtained from the ERA-Interim reanalysis (Dee et al, 2011) data set with 2° x 2° regular latitude-longitude horizontal resolution for the period 1989-2008.

The SDMs used in this work (see Table 2) were those recommended by San-Martín et al (2016) for climate change applications, and included particular configurations of different methodologies: the analog family (AN), weather types (WT), Generalized Linear Models (GLMs) and circulation-conditioned GLMs (GLM-WT). In the present study, the methods were calibrated using either Principal Components (PCs) of the predictor fields or local predictor values in the nearest grid boxes. In the former case, we used 25 PCs, that retain approximately 95% of the variance of the predictor fields. The latter considered the four reanalysis grid boxes nearest to the target location (*Spain044* grid box). The combined method labeled as *S*5 is a version of *S*4 conditioned on 10 Weather Types (WTs) obtained from a classification based on SLP.

All the experiments were accomplished using a k-fold ($k = 5$) cross validation with random sampling, by dividing the total 20-year period in two subsets of 4 years for testing and the remaining 16 years for training the method. This process was repeated five times, leading to five pairs of training and test periods which were considered for all the methods. The resulting test periods were concatenated into a single final downscaled multi-year series for validation. We refer the reader to Gutiérrez et al (2013) and San-Martín et al (2016) for more details regarding the methods and validation framework.

Table 2 Statistical downscaling methods used in the study. Codes are used to label SDMs in the figures. The second column (*CodeSM16*) is the label used by San-Martín et al (2016), who provide full details of the different methods.

Code	*CodeSM16*	*Family*	*Predictors*	*Description*
S1	SM1a	AN	PCs	Nearest analog
S2	SM2c	WT	PCs	100 WTs, simulation from Bernoulli+gamma
S3	SM3a	GLM	PCs	GLM (Bernoulli)+GLM (gamma)
S4	SM3c	GLM	Four nearest gridboxes	GLM (Bernoulli)+GLM (gamma)
S5	SM4b	GLM-WT	Four nearest gridboxes	S4 conditioned on 10 WTs

2.4 Precipitation indices

Table 3 summarizes the precipitation indices that were derived seasonally from daily precipitation amounts (RR). RR1 and SDII account for the mean precipitation regime whereas 90pWET, RX1day and RX5day are related to the tail of the distribution and CDD to the (dry) spells. The performance of the different downscaling methods is illustrated by means of the evaluation of RR, 90pWET, RX1day, RX5day and CDD. Moreover, the mean precipitation frequency (RR1) and the amount/intensity (SDII) are considered to adjust the first moments of the precipitation distribution via simple bias correction methods (Section 2.5).

According to the recommendations from Orlowsky and Seneviratne (2012), 90pWET was derived over the entire period (i.e. for all days in a season for the whole period), while CDD, RX1day and RX5day were calculated for each year and season, considering the interannual median as the final indicator.

Table 3 Precipitation indices used in this study as defined by the Expert Team on Climate Change Detection and Indices (ETCCDI, Sillman and Roeckner, 2008).

ID	*Indicator*	*Units*
RR	Daily precipitation amount	mm/day
RR1	Wet-day frequency	%
SDII	Simple day intensity index (mean wet-day precipitation)	mm/day
90pWET	90th percentile on wet days	mm
CDD	Maximum number of consecutive dry days	days
RX1day	Maximum 1-day precipitation amount	mm
RX5day	Maximum 5-day precipitation amount	mm

2.5 Simple bias correction methods

In order to take into account the effect of model biases (in frequency and amount) in the comparison of SDMs and RCMs, we considered both the raw (statistically and dynamically downscaled) model outputs and different simple bias corrected versions of them. Thus, we can test the potential effect of the training phase for SDMs, which typically adjusts the mean precipitation during the calibration process. Two bias correction methods (*Local Scaling*, LS, and *Frequency Adjustment*, FA) were applied separately to the precipitation indices in Table 3 depending on the different nature of the indices (i.e. intensity- or occurrence-related, respectively). The application of these corrections builds from previous work for RCMs only (Casanueva et al, 2015) and is extended here to SDMs.

The indices 90pWET, RX1day and RX5day were corrected using a multiplicative local scaling (LS) factor obtained as the quotient of the observed and simulated wet-day precipitation:

$$RR_{LS} = RR_{DS}\frac{SDII_{OBS}}{SDII_{DS}} \tag{1}$$

where RR_{DS} represents daily downscaled precipitation. The correction factor changed from season to season for each grid box. The precipitation indices were computed from the resulting RR_{LS} series.

Other precipitation indicators, such as CDD, are more related to precipitation occurrence and the autocorrelation of the precipitation series. This indicator changes as the wet-day threshold (typically 1mm) changes, thus it would be sensitive to changes in the wet-day frequency. The frequency adjustment was applied to the precipitation series by obtaining the *adjusted wet-day threshold* P^* that adjusts the simulated and observed wet-day frequency (i.e. the percentage of wet-days is the same for observations and simulation). For this purpose, P^* was estimated selecting the value of the downscaled precipitation matching the observed wet-day frequency computed with a 1mm threshold ($RR1_{OBS} = F_{OBS}(1mm)$) for each grid box:

$$P^* = F_{DS}^{-1}(F_{OBS}(1mm)) \tag{2}$$

where F is the empirical cumulative density function (CDF), so F_{DS} and F_{OBS} refer to the downscaled and observed CDFs, respectively. Thus, the correction of CDD consists in using P^* (instead of 1mm) as the wet-day threshold in the index calculation.

Note that this correction adjusts the precipitation occurrence, but does not affect the order (and thus, autocorrelation) of the precipitation series, i.e. whether the dry and wet days are located in the correct place. This correction may also affect percentiles on wet days, such as 90pWET. However, previous work analysing this correction shows that the changes in percentiles are very small and in some cases lead to higher biases than the original percentiles (Casanueva et al, 2015).

2.6 Connection between the mean and percentiles

Multiplicative LS correction of a modelled variable X, consists of multipliying at each grid point by a constant λ, to produce a new, corrected variable $Y = \lambda X$, which is expected to match exactly the observed mean μ_O. That is, $\lambda = \mu_O/\mu_X$. This correction is used to mimic the calibration of the mean that occurs during the SDMs training phase. Thus, implicitly, it is useful to determine whether the indicator used for the SDM-RCM comparison would be affected by a calibration of the mean and, then, to analyse the fairness of the comparison.

Wet-day precipitation amount is usually represented by the Gamma distribution, $Ga(\kappa,\theta)$, or its particular exponential distribution case, $Ex(\theta) = Ga(1,\theta)$ (Benestad et al, 2011). Regardless of the probability distribution of X, the quantiles of $Y = \lambda X$, for any positive λ, are accordingly scaled: $Q_Y(p) = \lambda Q_X(p)$. Therefore, all quantiles are linearly scaled along with the mean after LS.

The question remains whether the new quantiles $\lambda Q_X(p)$ better match those of the observations $Q_O(p)$. If the variable from both observations and model results belong to the same Gamma family, multiplicative LS correction provides a perfect correction for all quantiles, and not only for the mean. For example, if both model and observations follow an exponential distribution, which depends on a single scale parameter, a perfect correction would be achieved. The original variable has mean $\mu_X = \theta$ and variance $\sigma_X^2 = \theta^2$. Therefore, after LS: $\mu_{\lambda X} = \lambda\theta$ and variance $\sigma_{\lambda X}^2 = (\lambda\theta)^2$, and the scaled distribution is still exponential with parameter $\lambda\theta$. Adjusting the mean exactly matches the single parameter and, thus, the whole distribution, including all percentiles. Moreover, if the exponential distribution applies to both the observations and model results, the reproduction of the mean (through LS or any other calibration methodology) implies the reproduction of the whole distribution.

In the case of the general Gamma family, the same result applies, as long as the shape parameter, κ, is equal in the observations and model. For reasonably similar shape parameters, LS would tend to bias correct all quantiles, even though the method is devised to correct the mean. Deviations from perfect percentile bias correction therefore indicate different shape parameters of different distribution families between model and observations. Section 3.2 shows the effect of the correction on 90pWET with (statistically and dynamically) downscaled data over Spain.

Note that the correction of percentiles by correcting the mean only holds for distribution families where a scale parameter controls both the mean and the variability. For instance, in the case of temperature, where a Gaussian distribution is commonly considered, mean and variance are independent parameters. The mean can be corrected by additive LS without affecting the variability (the quantiles would be shifted in this case).

Frequency adjustment (FA, Section 2.5) is associated with the wet-day frequency. It is not related to the parameters of the exponential and Gamma distributions, but it changes the precipitation distribution by modifying the number of zero-precipitation values. When P^* is larger than 1mm (the reference wet-day threshold), all values in the range (1mm, P^*) would be considered dry, thus increasing the number of dry days. For the opposite situation ($P^* < 1$mm) the frequency adjustment does not provide an optimal correction since it cannot 'invent' wet days (Bärring et al, 2006). The adjusted threshold P^* would directly have an impact on derived indicators affected by the wet-day definition (e.g. CDD). Note that the wet-day frequency is not an optimized parameter in any of the statistical or dynamical methods considered in this work. Thus, this correction does not resemble any calibration of the considered downscaling methods.

3 Results

3.1 Unfair comparison: Mean precipitation

When looking at the mean precipitation regime, a fair evaluation and comparison of both downscaling techniques on equal footing should be carefully performed. It is important to note that the EURO-CORDEX RCMs have not assimilated any information from *Spain044* observations, whereas the SDMs have been cross-calibrated using them —in particular, GLMs are trained minimizing the distance between the observed and predicted/downscaled daily mean training error.— Therefore, RCMs typically exhibit non-negligible biases (Casanueva et al, 2015), whereas mean precipitation is usually well represented by the different SDMs. This argument, however, should not be used to classify or rank statistical and dynamical techniques as in the recent work from Ayar et al (2015). Every classification of methods will rely on specific criteria, but the fairness of that criteria (i.e. no benefit for any method) is essential.

Comparing SDMs and RCMs in terms of mean precipitation would inevitably favour SDMs, since the mean is an optimized parameter in the SDMs training phase, thus leading to an unfair comparison of downscaling techniques. This is illustrated in the Taylor diagrams (Taylor, 2001) for the statistically and dynamically downscaled mean precipitation fields in the four seasons (Figure 1). In order to give a spatially averaged measure of accuracy avoiding the compensation of opposite sign biases, we use throughout the entire paper the spatially averaged mean absolute error (MAE), which is calculated as the spatial average of the absolute value of the mean temporal errors at each grid box. Each downscaling method is represented by a square (filled with the MAE) using the labels given in Tables 1 and 2. Among the SDMs, the two GLMs (S3 and S4) are almost identical in all seasons (note that the only methodological difference is found in the predictors, i.e. PCs in S3 and nearest grid boxes in S4). S5 (circulation-conditioned GLM) is slightly worse than the other SDMs. Regarding the RCMs, HIRHAM (D2) and RCA (D4) stand out among the others for their worse representation of the spatial pattern. The two WRF versions (D7 and D8) present very similar results in every season. RCMs show the larger spread in performance in summer, probably due to small-scale processes (such as those related to convection) which are more strongly controlled by parameterized physics in summer (Déqué et al, 2005).

As expected, the SDMs largely outperform the RCMs, as the scores are closer to the observations in all seasons. This is an example of an unfair comparison, even though the SDMs have been calibrated at the annual scale and, therefore, may exhibit seasonal biases. However, as shown in Figure 1, this has a small effect on the seasonal spatial patterns. Note that, in this case, performing a fair comparison is difficult, since even the simplest bias correction would adjust the mean precipitation spatial patterns, thus giving optimal results for both RCMs and SDMs. However, a fair comparison of both techniques can be done considering statistics or indicators not affected by the calibration processes, as shown in Sections 3.2 and 3.3.

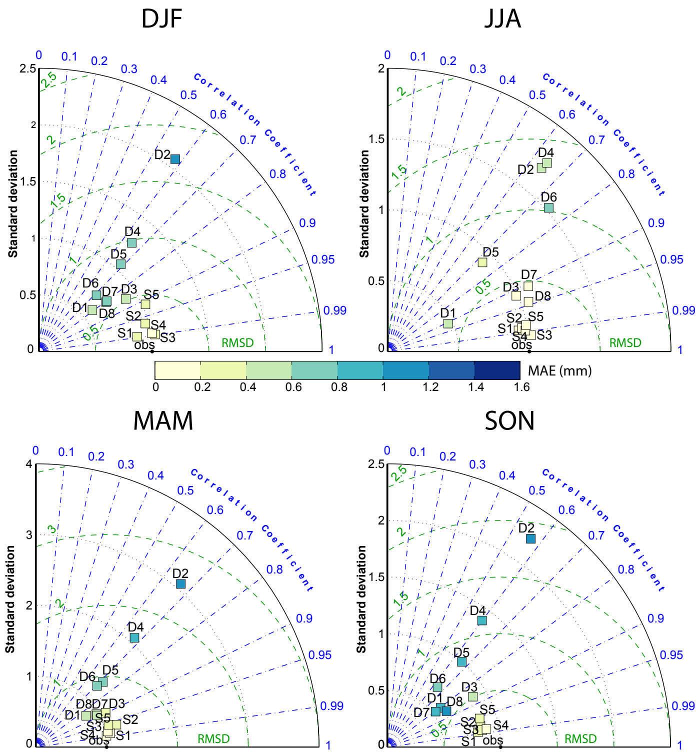

Fig. 1 Taylor diagrams for mean precipitation in Spain in the four seasons. Each square represents either a dynamical (D) or statistical (S) downscaling method, labeled according to the codes in Tables 1 and 2. The diagrams show validation results considering spatial Pearson correlation coefficient (r), centered root mean squared difference (RMSD) and variability (std). Colours inside the squares represent the spatially averaged mean absolute error (MAE). See text for more details.

3.2 Comparing extreme precipitation

Pursuing a fair comparison of RCMs and SDMs, we evaluate precipitation indices which have not been directly optimized during the calibration of the methods (90pWET, RX1day, RX5day and CDD, see Table 3). In this case, results are only shown for winter season (DJF), although the same conclusions also hold for the rest of the seasons. Figure 2 (left panel) shows Taylor diagrams of 90pWET, RX1day and CDD indices for winter. Each arrow represents a different downscaling method linking the validation scores of the original predictions (squares) and the bias-corrected ones (circles; see Section 2.5).

Before any correction, the same conclusion as for mean precipitation (Figure 1) holds for 90pWET (Figure 2a), with better validation scores for the SDMs. Although 90pWET is not an optimized parameter in the SDMs calibration, evaluation results are clearly better than for the RCMs. This can be explained by the relationship that links the mean and the percentiles of a precipitation distribution (Benestad et al, 2012), since the calibration of the mean in the SDMs leads to the adjustment of the percentiles (Section 2.6) and, thus, 90pWET. For this reason, the comparison of SDMs and RCMs in terms of percentile-based indicators would be as unfair as for the mean precipitation.

The local scaling (Section 2.5) is applied to mimic a calibration in the mean in both statistical and dynamical techniques. After this correction, all the methods present comparable results. Results improve not only in terms of spatial correlation and variability, but also in terms of MAE (colors inside the markers in the Taylor diagram). Therefore, RCM biases in mean precipitation are responsible for the worse evaluation results for percentiles and they are able to properly represent percentiles as long as the mean precipitation is adjusted. Negligible changes are found for the SDMs, since good evaluation results were found before the correction.

Similar conclusions apply to RX1day (Figure 2c). Before the correction, SDMs present better scores than the RCMs although S3-S5 exhibit an anomalous large spatial variability. Again, this could be partially explained by the relationship of the tail statistics of the precipitation distribution with the precipitation mean value, since the RX1day indicator would correspond to a percentile at the tail of the distribution. Therefore, a direct comparison of results from both techniques is also unfair in this case. After local scaling, the results of the RCMs become comparable to the SDMs. Similar results were also found for RX5day (not shown).

3.3 Comparing spells

The temporal autocorrelation of the precipitation series is not optimized in the calibration phase of any of the methods, therefore, CDD is a good candidate to provide an example of a fair SDM-RCM comparison. In this case, comparable validation scores are found for winter CDD (Figure 2e) for both downscaling techniques before and after the frequency adjustment (see Section 2.5). Before the correction, specific methods (regardless of the downscaling family) may present similar skill or deficiencies in representing dry spell spatial patterns. After the frequency adjustment, spatial patterns and MAEs improve for the RCMs (in agreement with Casanueva et al, 2015). SDMs show very small changes after the frequency adjustment (mainly a reduction in the spatial variability). This suggests that they present inherent deficiencies in representing dry spells, which cannot be solved by means of a bias correction. Note that the correction does not alter the series autocorrelation, but the wet-day frequency. In particular, S5 shows a completely different behaviour as compared to the other SDMs, whereas the analog method (S1) is the best-performing SDM. Bear in mind that the analog method is an algorithmic method that is based on a resampling of the observations. Therefore, it does not explicitly calibrate the mean or the temporal correlation but, according to the results, they are indirectly quite well captured. This is one advantage of this method, but it also presents some limitations such as the lack of robustness associated to the impossibility of extrapolating future atmospheric conditions (Gutiérrez et al, 2013).

More detailed analyses have been performed to examine the ability of SDMs and RCMs in representing CDD (Figure 3). Before the correction, methods S2-S4 predict longer dry spells than observed (Figure 3, first column). RCMs usually overestimate the number of wet days, and thus underestimate CDD, by frequently simulating light rainfall (Figure 3, second column). The frequency adjustment (Section 2.5) works well for finding optimal thresholds (P^*) greater than 1mm (e.g. D3, D4 and D8 in Figure 3, fourth and sixth columns). However, the excess of dry days leads to close-to-zero wet-day thresholds (see S2-4 in third column in Figure 3). As stated in Section 2.6, the frequency adjustment cannot solve this problem and biases would still be present in the corrected CDD (Figure 3, fifth column), since the procedure cannot invent wet days for too dry methods (Casanueva et al, 2015). Summer precipitation indices in RCMs are affected also by this situation (long dry spells), which can also be seen in winter (e.g. D5).

4 Conclusions

It is nowadays commonly recognized that there are some key factors which must be taken into account for a fair comparison of statistical and dynamical downscaling techniques. Both approaches use observational data in different ways, either explicitly for model fitting/calibration in SDMs (for instance, to fit the parameters of a regression model minimizing the mean squared error), or implicitly for model tuning in RCMs (for instance, to adjust model parameters based on evaluation against observations). Therefore, misleading results can be obtained when comparing the performance of both techniques using scores/indices which might be affected by model fitting. This paper gives insight into a fair comparison of statistical and dynamical downscaling methods.

We analyze RCMs from the EURO-CORDEX initiative compared to previously tested SDMs in continental Spain (San-Martín et al, 2016) for the period 1989-2008. Both the RCM boundary conditions and the SDM predictors are taken from the ERA-Interim reanalysis (Dee et al, 2011). The SDMs calibration is performed using the new EURO-CORDEX compliant gridded observational data set (*Spain044*), therefore the comparison of RCMs and SDMs is accomplished on the same grid, unlike previous studies that interpolate from local stations/grid to RCM grid or vice versa (e.g. Kidson and Thompson, 1998; Murphy, 1999; Haylock et al, 2006).

As expected, we find that SDMs outperform the RCMs with respect to seasonal mean precipitation, with an almost perfect performance in the four seasons. Regarding the derived indicators, 90pWET (90th percentile on wet days) and RX1day (maximum 1-day precipitation amount) appear to be indirectly calibrated by the SDMs, due to their close relationship to the precipitation intensity. A local scaling bias correction method is applied to all statistical and dynamical downscaling methods resembling the calibration phase of the SDMs towards the observations. After this correction, all downscaling methods show comparable skill in reproducing 90pWET, RX1day and RX5day. This confirms that a good representation of mean precipitation also provides good evaluation results for high percentile indicators, regardless of the downscaling technique. This is a result of the usually employed exponential or gamma distribution models for precipitation, as long as the shape parameter is reasonably represented. Thus, the calibration in the mean during the training phase produces also an adjustment of percentile-based indicators and this would inevitably benefit the SDMs in a SDM-RCM comparison (if RCM biases are not removed).

Alternatively, the evaluation of the CDD (maximum number of consecutive dry days) provides a fair comparison of RCMs and SDMs, since the autocorrelation of the precipitation series is not an optimized parameter in the calibration process. Our results show that specific SDMs and RCMs may be more or less skillful regardless of the downscaling technique. A correction in the wet-day frequency produces an improvement in the representation of the CDD spatial pattern although biases might remain high, meaning that the frequency adjustment is not enough to correct deficiencies in the lower part of the distribution in some of the methods.

More efforts devoted to the evaluation of non-optimized parameters, as well as the use of several observational data sets should be considered in a fair SDM-RCM comparison framework. Note that in this work RCMs do not assimilate information from the observational reference, but different results may have been obtained if the observational data set had played a role in the RCM's tuning phase.

Acknowledgements We acknowledge the World Climate Research Programme's Working Group on Regional Climate, and the Working Group on Coupled Modelling, former coordinating body of CORDEX and responsible panel for CMIP5. We also thank the climate modeling groups (listed in Table 1 of this paper) for producing and making available their model output. We also acknowledge the Earth System Grid Federation infrastructure and AEMET and University of Cantabria for the *Spain02* dataset (available at `http://www.meteo.unican.es/en/datasets/spain02`). All the statistical downscaling experiments have been computed using the MeteoLab software (`http://www.meteo.unican.es/software/meteolab`), which is an open-source Matlab toolbox for statistical downscaling. This work has been partially supported by CORWES (CGL2010-22158-C02) and EXTREMBLES (CGL2010-21869) projects funded by the Spanish R&D programme. AC thanks the Spanish Ministry of Economy and Competitiveness for the funding provided within the FPI programme (BES-2011-047612 and EEBB-I-13-06354), JMG acknowledges the support from the SPECS project (FP7-ENV-2012-308378) and JF is grateful to the EUPORIAS project (FP7-ENV-2012-308291). We also thank three anonymous referees for their useful comments that helped to improve the original manuscript.

References

Ayar PV, Vrac M, Bastin S, Carreau J, Déqué M, Gallardo C (2015) Intercomparison of statistical and dynamical downscaling models under the EURO- and MED-CORDEX initiative framework: present climate evaluations. Clim Dyn 46(3-4):1301–1329, DOI 10.1007/s00382-015-2647-5

Bärring L, Holt T, Linderson M, Radziejewski M, Moriondo M, Palutikof JP (2006) Defining dry/wet spells for point observations, observed area averages, and regional climate model gridboxes in Europe. Climate Research 31(1):35–49, DOI 10.3354/cr031035

Benestad RE, Nychka D, Mearns LO (2011) Specification of wet-day daily rainfall quantiles from the mean value. Tellus A 64

Benestad RE, Nychka D, Mearns LO (2012) Spatially and temporally consistent prediction of heavy precipitation from mean values. Nature Climate Change 2(7):544–547, DOI 10.1038/NCLIMATE1497

Brands S, Gutiérrez J, Herrera S, Cofiño A (2012) On the use of reanalysis data for downscaling. Journal of Climate 25:2517–2526

Casanueva A, Kotlarski S, Herrera S, Fernández J, Gutiérrez J, Boberg B, Colette A, Christensen OB, Goergen K, Jacob D, Keuler K, Nikulin G, Teichmann C, Vautard R (2015) Daily precipitation statistics in a EURO-CORDEX RCM ensemble: added value of raw and bias-corrected high-resolution simulations. Climate Dynamics pp 1–19, DOI 10.1007/s00382-015-2865-x

Christensen OB, Drews M, Christensen J, Dethloff K, Ketelsen K, Hebestadt I, Rinke A (2007) The HIRHAM Regional Climate Model. Version 5 (beta). Denmark. Danish Meteorological Institute. Technical Report, Danish Climate Centre, Danish Meteorological Institute

Collins M, Booth BBB, Harris GR, Murphy JM, Sexton DMH, Webb MJ (2006) Towards quantifying uncertainty in transient climate change. Climate Dynamics 27(2-3):127–147

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, Hólm EV, Isaksen L, Kållberg P, Köhler M, Matricardi M, McNally AP, Monge-Sanz BM, Morcrette J, Park B, Peubey C, de Rosnay P, Tavolato C, Thépaut JN, Vitart F (2011) The ERA-Interim reanalysis: configuration and performance of the data assimilation system. Quart J R Meteorol Soc 137:553–597

Déqué M (2007) Frequency of precipitation and temperature extremes over France in an anthropogenic scenario: Model results and statistical correction according to observed values. Global and Planetary Change 57(12):16–26, DOI 10.1016/j.gloplacha.2006.11.030

Déqué M, Jones RG, Wild M, Giorgi F, Christensen JH, Hassell DC, Vidale PL, Rockel B, Jacob D, Kjellström E, Castro Md, Kucharski F, Hurk Bvd (2005) Global high resolution versus limited area model climate change projections over Europe: quantifying confidence level from PRUDENCE results. Climate Dynamics 25(6):653–670, DOI 10.1007/s00382-005-0052-1

Ehret U, Zehe E, Wulfmeyer V, Warrach-Sagi K, Liebert J (2012) HESS Opinions ”Should we apply bias correction to global and regional climate model data?”. Hydrol Earth Syst Sci 16(9):3391–3404, DOI 10.5194/hess-16-3391-2012

Estrada F, Guerrero VM, Gay-García C, Martínez-López B (2013) A cautionary note on automated statistical downscaling methods for climate change. Climatic Change 120(1-2):263–276, DOI 10.1007/s10584-013-0791-7

Feser F, Rockel B, von Storch H, Winterfeldt J, Zahn M (2011) Regional climate models add value to global model data: a review and selected examples. Bulletin of the American Meteorological Society 92(9):1181–1192

Giorgi F (2006) Regional climate modeling: Status and perspectives. Journal de Physique IV (Proceedings) 139(1):101–118, DOI 10.1051/jp4:2006139008

Goodess C (2005) Statistical and regional dynamical downscaling of extremes for European regions. STARDEX Final Management Report Available at http://www.cru.uea.ac.uk/cru/research/stardex

Gutiérrez J, San-Martín D, Brands S, Manzanas R, Herrera S (2013) Reassessing statistical downscaling techniques for their robust application under climate change conditions. Journal of Climate 26:171–188

Haylock M, Cawley G, Harpham C, Wilby R, Goodess C (2006) Downscaling heavy precipitation over the United Kingdom: a comparison of dynamical and statistical methods and their future scenarios. International Journal of Climatology 26(10):1397–1415

Haylock MR, Hofstra N, Klein Tank AMG, Klok EJ, Jones PD, New M (2008) A European daily high-resolution gridded data set of surface temperature and precipitation for 1950-2006. Journal of Geophysical Research-Atmospheres 113(D20), DOI 10.1029/2008JD010201

Herrera S, Gutiérrez J, Ancell R, Pons M, Frías M, Fernández J (2012) Development and analysis of a 50 year high resolution daily gridded precipitation dataset over Spain (Spain02). International Journal of Climatology DOI 10.1002/joc.2256

Herrera S, Fernández J, Gutiérrez (2015) Update of the Spain02 gridded observational dataset for EURO-CORDEX evaluation: Assessing the effect of the interpolation methodology. International Journal of Climatology DOI 10.1002/joc.4391

Hertig E, Paxian A, Vogt G, Seubert S, Paeth H, Jacobeit J (2012) Statistical and dynamical downscaling assessments of precipitation extremes in the mediterranean area. Meteorologische Zeitschrift 21(1):61–77, DOI 10.1127/0941-2948/2012/0271

Jacob D, Petersen J, Eggert B, Alias A, Christensen OB, Bouwer L, Braun A, Colette A, Déqué M, Georgievski G, Georgopoulou E, Gobiet A, Menut L, Nikulin G, Haensler A, Hempelmann N, Jones C, Keuler K, Kovats S, Kröner N, Kotlarski S, Kriegsmann A, Martin E, van Meijgaard E, Moseley C, Pfeifer S, Preuschmann S, Radermacher C, Radtke K, Rechid D, Rounsevell M, Samuelsson P, Somot S, Soussana JF, Teichmann C, Valentini R, Vautard R, Weber B, Yiou P (2014) EURO-CORDEX: new high-resolution climate change projections for European impact research. Regional Environmental Change 14(2):563–578, DOI 10.1007/s10113-013-0499-2

Kidson J, Thompson C (1998) A comparison of statistical and model-based downscaling techniques for estimating local climate variations. Journal of Climate

11(4):735–753
Kotlarski S, Keuler K, Christensen O, Colette A, Déqué M, Gobiet A, Goergen K, Jacob D, Lüthi D, van Meijgaard E, Nikulin G, Schär C C Teichmann, Vautard R, Warrach-Sagi K, Wulfmeyer V (2014) Regional climate modeling on European scales: a joint standard evaluation of the EURO-CORDEX RCM ensemble. Geoscientific Model Development Discussions 7:217–293
Luo Q, Wen L, McGregor JL, Timbal B (2013) A comparison of downscaling techniques in the projection of local climate change and wheat yields. Climatic Change 120(1-2):249–261, DOI 10.1007/s10584-013-0802-8
Maraun D, Wetterhall F, Ireson AM, Chandler RE, Kendon EJ, Widmann M, Brienen S, Rust HW, Sauter T, Themessl M, Venema VKC, Chun KP, Goodess CM, Jones RG, Onof C, Vrac M, Thiele-Eich I (2010) Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. Reviews of Geophysics 48(3):RG3003, DOI 10.1029/2009RG000314
McCullagh P, Nelder JA (1989) Generalized linear models, Monographs on Statistics and Applied Probability. Chapman & Hall, London
Meijgaard Ev, Ulft LHv, Lenderink G, Roode SRd, Wipfler EL, Boers R, Timmermans RMA (2012) Refinement and Application of a Regional Atmospheric Model for Climate Scenario Calculations of Western Europe. Programme Office Climate changes Spatial Planning
Murphy J (1999) An evaluation of statistical and dynamical techniques for downscaling local climate. Journal of Climate 12(8):2256–2284
Orlowsky B, Seneviratne SI (2012) Global changes in extreme events: regional and seasonal dimension. Climatic Change 110(3-4):669–696, DOI 10.1007/s10584-011-0122-9
Pizzigalli C, Palatella L, Zampieri M, Lionello P, Miglietta M, Paradisi P (2012) Dynamical and statistical downscaling of precipitation and temperature in a Mediterranean area. Italian Journal of Agronomy 7(1):2, DOI 10.4081/ija.2012.e2
Radu R, Déqué M, Somot S (2008) Spectral nudging in a spectral regional climate model. Tellus A 60(5), DOI 10.3402/tellusa.v60i5.15501
Rockel B, Will A, Hense A (2008) The Regional Climate Model COSMO-CLM (CCLM). Meteorologische Zeitschrift 17(4):347–348, DOI 10.1127/0941-2948/2008/0309
Samuelsson P, Jones CG, Willén U, Ullerstig A, Gollvik S, Hansson U, Jansson C, Kjellstrm E, Nikulin G, Wyser K (2011) The Rossby Centre Regional Climate model RCA3: model description and performance. Tellus A 63(1), DOI 10.3402/tellusa.v63i1.15770
San-Martín D, Manzanas R, Brands S, Herrera S, Gutiérrez J (2016) Reassessing model uncertainty for regional projections of precipitation with an ensemble of statistical downscaling methods. Journal of Climate , submitted
Schmidli J, Goodess C, Frei C, Haylock M, Hundecha Y, Ribalaygua J, Schmith T (2007) Statistical and dynamical downscaling of precipitation: An evaluation and comparison of scenarios for the European Alps. J Geophys Res 112(10.1029)
Sillman J, Roeckner R (2008) Indices for extreme events in projections of anthropogenic climate change. Climatic Change 86:83–104, DOI 10.1007/

s10584-007-9308-6

Skamarock W, Klemp J, Dudhia J, Gill D, Barker D, Duda M, Wang W, Powers J (2008) A description of the Advanced Research WRF Version 3. Tech. rep., NCAR

von Storch H, Zorita E, Cubasch U (1993) Downscaling of global climate change estimates to regional scales: An application to Iberian rainfall in wintertime. Journal of Climate 6(6):1161–1171, DOI 10.1175/1520-0442(1993)006⟨1161:DOGCCE⟩2.0.CO;2

Tareghian R, Rasmussen PF (2013) Statistical downscaling of precipitation using quantile regression. Journal of Hydrology 487:122–135, DOI 10.1016/j.jhydrol.2013.02.029

Taylor KE (2001) Summarizing multiple aspects of model performace in a single diagram. J Geophys Res 106(D7):7183–7192

Teutschbein C, Seibert J (2012) Bias correction of regional climate model simulations for hydrological climate-change impact studies: Review and evaluation of different methods. Journal of Hydrology 456457:12–29

Teutschbein C, Seibert J (2013) Is bias correction of regional climate model (RCM) simulations possible for non-stationary conditions? Hydrol Earth Syst Sci 17(12):5061–5077, DOI 10.5194/hess-17-5061-2013

Tryhorn L, DeGaetano A (2011) A comparison of techniques for downscaling extreme precipitation over the Northeastern United States. Int J Climatol 31(13):1975–1989, DOI 10.1002/joc.2208

Turco M, Quintana-Segui P, Llasat MC, Herrera S, Gutiérrez JM (2011) Testing MOS precipitation downscaling for ENSEMBLES regional climate models over spain. Journal of Geophysical Research-Atmospheres 116:14

Wilby RL, Wigley TML (1997) Downscaling general circulation model output: a review of methods and limitations. Progress in Physical Geography 21(4):530–548, DOI 10.1177/030913339702100403

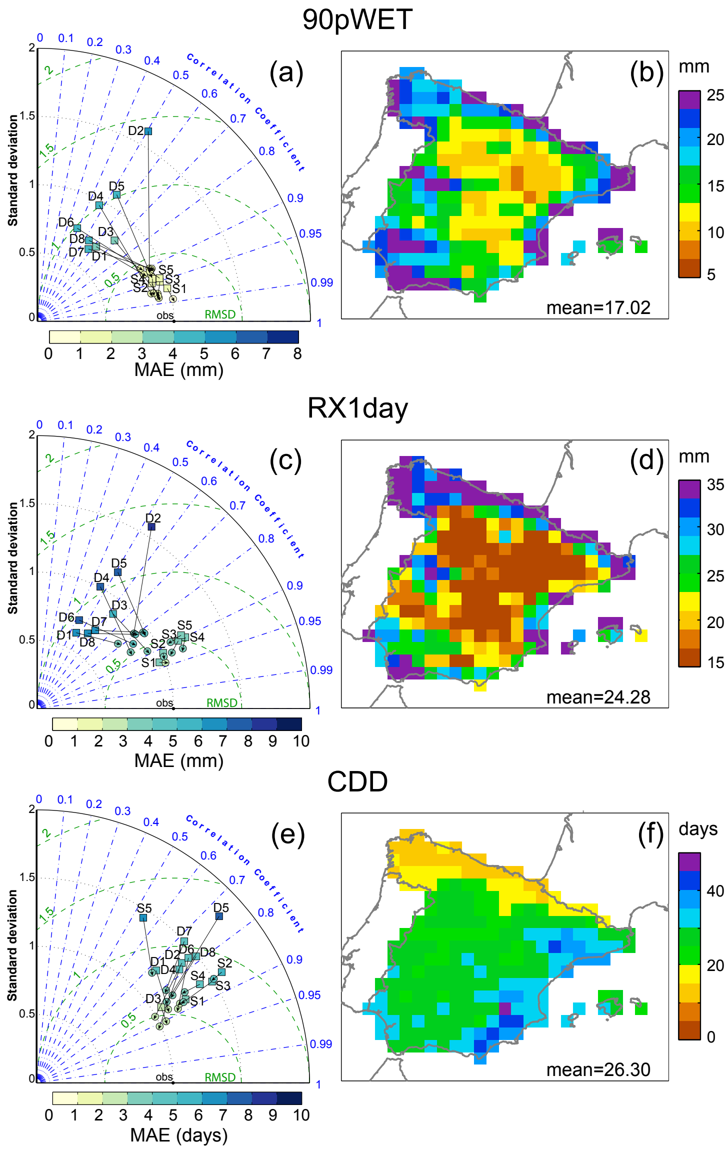

Fig. 2 Observed (*Spain044*) values (right panels) and Taylor diagram for the different statistical and dynamical downscaling methods (left panels) for winter *90pWET* (a-b), *RX1day* (c-d) and *CDD* (e-f). The Taylor diagrams represent the downscaled values before (squares) and after (circles) bias correction. In all cases, symbols are filled with a color corresponding to the MAE (see text).

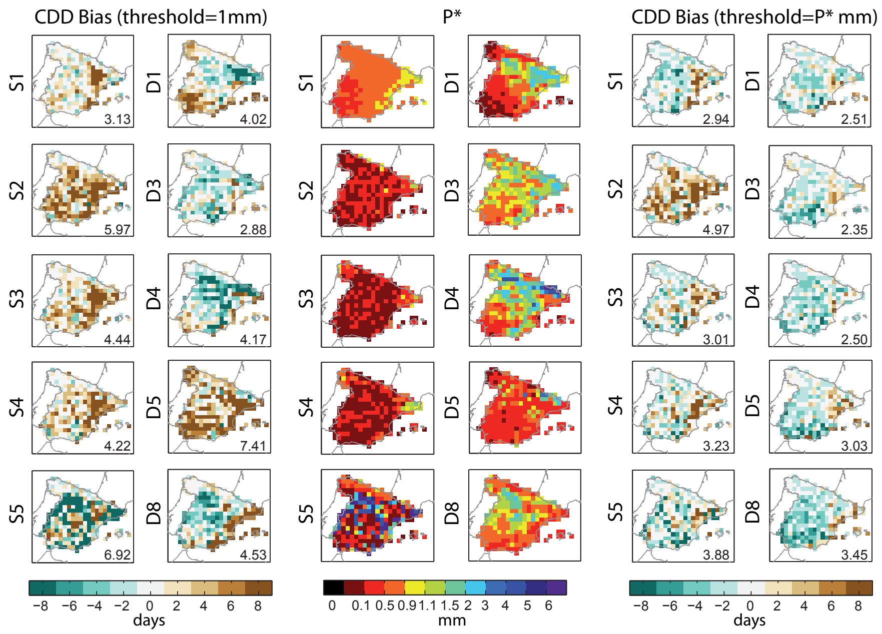

Fig. 3 Biases for CDD before the correction (first and second columns), wet-day adjusted thresholds P^* (third and fourth columns, see Section 2.5) and CDD biases after the correction (fifth and sixth columns) for the SDMs (S1-5) and some representative RCMs, in winter. The numbers inside the figures are the spatially averages MAE's. For a better contrast of spatial differences in P^*, values are presented using a non-linear scale.