Reassessing Model Uncertainty for Regional Projections of Precipitation with an Ensemble of Statistical Downscaling Methods

D. SAN-MARTÍN

Grupo de Meteorología, Instituto de Física de Cantabria, Consejo Superior de Investigaciones Científicas–Universidad de Cantabria, and Predictia Intelligent Data Solutions SL, Santander, Spain

R. MANZANAS AND S. BRANDS

Grupo de Meteorología, Instituto de Física de Cantabria, Consejo Superior de Investigaciones Científicas–Universidad de Cantabria, Santander, Spain

S. HERRERA

Grupo de Meteorología, Deptartamento de Matemática Aplicada y Ciencias de la Computación, Universidad de Cantabria, Santander, Spain

J. M. GUTIÉRREZ

Grupo de Meteorología, Instituto de Física de Cantabria, Consejo Superior de Investigaciones Científicas–Universidad de Cantabria, Santander, Spain

(Manuscript received 18 May 2016, in final form 15 September 2016)

ABSTRACT

This is the second in a pair of papers in which the performance of statistical downscaling methods (SDMs) is critically reassessed with respect to their robust applicability in climate change studies. Whereas the companion paper focused on temperatures, the present manuscript deals with precipitation and considers an ensemble of 12 SDMs from the analog, weather typing, and regression families. First, the performance of the methods is cross-validated considering reanalysis predictors, screening different geographical domains and predictor sets. Standard accuracy and distributional similarity scores and a test for extrapolation capability are considered. The results are highly dependent on the predictor sets, with optimum configurations including information from midtropospheric humidity. Second, a reduced ensemble of well-performing SDMs is applied to four GCMs to properly assess the uncertainty of downscaled future climate projections. The results are compared with an ensemble of regional climate models (RCMs) produced in the ENSEMBLES project. Generally, the mean signal is similar with both methodologies (with the exception of summer, which is drier for the RCMs) but the uncertainty (spread) is larger for the SDM ensemble. Finally, the spread contribution of the GCM- and SDM-derived components is assessed using a simple analysis of variance previously applied to the RCMs, obtaining larger interaction terms. Results show that the main contributor to the spread is the choice of the GCM, although the SDM dominates the uncertainty in some cases during autumn and summer due to the diverging projections from different families.

1. Introduction

Downscaling methods are nowadays routinely applied to translate the coarse-resolution output from global climate models (GCMs) to the spatial scales required by climate change impact assessment studies (see Winkler et al. 2011, and references therein). However, climate change projections obtained from this approach are intrinsically uncertain and there are many uncertainty sources. These sources can be grouped into 1) ''external'' factors, which the downscaling community has to assume without, in principle, having the possibility to reduce and/or improve them (typically GCM errors, scenario uncertainties, and observational uncertainties) versus 2) ''internal'' factors, which can (and should) be

Corresponding author address: D. San-Martín, Predictia Intelligent Data Solutions SL, Avda. los Castros s/n. I+D S345, 39005, Santander, Spain.
E-mail: daniel@predictia.es

DOI: 10.1175/JCLI-D-16-0366.1

improved to reduce the spread of the climate change projections (see, e.g., Turco et al. 2013). Following this nomenclature, the present study deals with the internal uncertainty sources of statistical downscaling methods (SDMs) applied in perfect prognosis conditions [see Maraun et al. (2010) for definitions], among which the choice of predictors and downscaling methods (applied to the same predictors) are most relevant since the spread stemming from these choices can be even larger than the spread arising from the choice of the driving GCM, the latter usually considered the most important ''external'' uncertainty contributor (see, e.g., Dibike and Coulibaly 2005; Gutiérrez et al. 2013; Hertig and Jacobeit 2013).

To decide which predictor variables and SDMs are suitable for climate change applications, Gutiérrez et al. (2013) proposed to verify the ''goodness'' of the downscaled time series in terms of 1) accuracy (Jolliffe and Stephenson 2003), 2) distributional similarity, 3) variability of the monthly bias values (i.e., the seasonal cycle of the bias), and 4) stationarity of the bias in climate conditions distinct to those used for training and/or calibration (referred to as ''robustness'').

With regard to temperature downscaling, Gutiérrez et al. (2013) argue that a suitable model for climate change applications (note that the term ''model'' hereafter refers to a specific SDM calibrated with a specific predictor combination) should return acceptable results for any of the four aforementioned criteria. They applied a series of standard verification measures for points 1 through 3 above, accompanied by a new statistical test built to measure the fourth point from past observations only (i.e., without the need to apply scenario data from climate models) (Maraun 2012). As a key result, they showed that all tested downscaling methods failed to pass point 4 if a key predictor variable [air temperature at 2 m (2T)] was not considered. If 2T was considered, those methods not passing point 4 returned delta change estimates substantially smaller than those obtained by the methods passing it, showing that the test was indeed able to discard unsuitable methods before actually applying them to scenario data from climate models.

The present study assesses to which degree the validation philosophy presented in the companion paper of Gutiérrez et al. (2013) is transferable to the downscaling of daily precipitation, which undoubtedly is more challenging than simulating temperature alone. To this aim, 12 SDMs from three distinct method families [analog, weather typing, and generalized linear models (GLMs)] are used over the country of Spain. Because of fundamental differences in the precipitation regimes, results are analyzed separately for the Atlantic Ocean and Mediterranean Sea subsectors of this region. After finding the optimal geographical domain and predictor combination following the full spectrum of validation criteria mentioned above, an ensemble of five SDMs suitable for climate change applications is applied to downscale the control and transient future simulations (20C3M and A1B scenarios, respectively) of four GCMs from the ENSEMBLES project participating in CMIP3 (van der Linden and Mitchell 2009). This leads to a 20-member ensemble of local-scale precipitation projections comprising GCM and SDM uncertainty. Overall, a general precipitation decrease is projected to occur in all seasons along the course of the twenty-first century, with an uncertainty or spread smaller in spring and larger in summer and autumn.

In a second working step, the obtained statistical projections are compared with the dynamical solutions available from the ENSEMBLES project, considering the regional climate models driven by almost the same GCMs (Déqué et al. 2007). The two approaches were found to approximately agree on the sign and magnitude of climate change but the spread is clearly larger for the statistical approach. To understand the sources for this kind of uncertainty, a simple analysis of variance is conducted to assess the relative contribution of the SDMs and GCMs to the total uncertainty. Although some sophisticated approaches have been recently proposed for this purpose (Hingray and Said 2014; Hanel and Buishand 2015), the simple Déqué et al. (2012) approach is followed here to allow for a proper comparison with the results obtained with the ENSEMBLES RCMs.

The information produced in this work is part of the Spanish National Climate Change Adaptation Plan (PNACC; freely available at http://www.aemet.es/es/serviciosclimaticos/cambio_climat).

The paper is organized as follows: The region of study and the data used in this work are presented in section 2. Sections 3 and 4 describe the different SDMs and the cross-validation approach followed, respectively. The screening of predictors and geographical domains is presented in section 5, and section 6 presents the assessment of the different SDMs. Section 7 analyzes the future projections obtained applying the ensemble of SDMs to four ENSEMBLES–CMIP3 GCMs and section 8 analyzes the contribution of the global and regional model components to the total uncertainty. Finally, the main conclusions are given in section 9.

2. Region of study and data

a. *Predictand*

The predictand data used in this work are from the Spain02 daily gridded precipitation dataset (Herrera

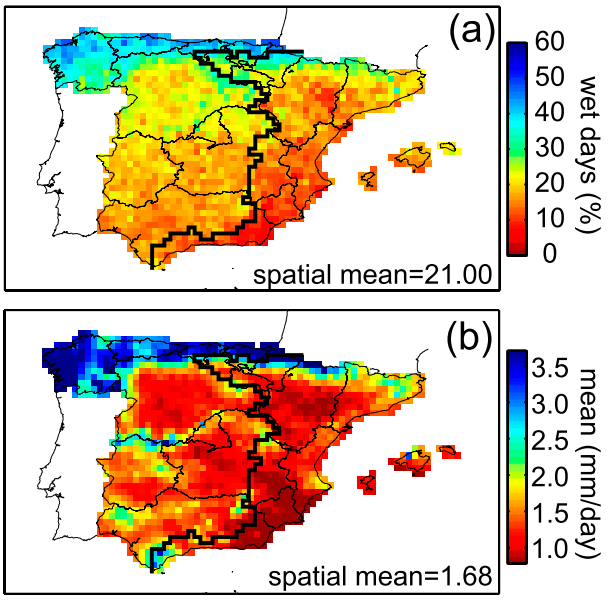

FIG. 1. (a) Percentage of wet days (precipitation ≥ 0.1 mm) and (b) daily mean precipitation of the Spain02 precipitation dataset during 1961–2000. The black line in the maps indicates the water divide between the (left) Atlantic Ocean and (right) Mediterranean Sea hydrological basins, considered as a rough classification in two main precipitation regimes within the country.

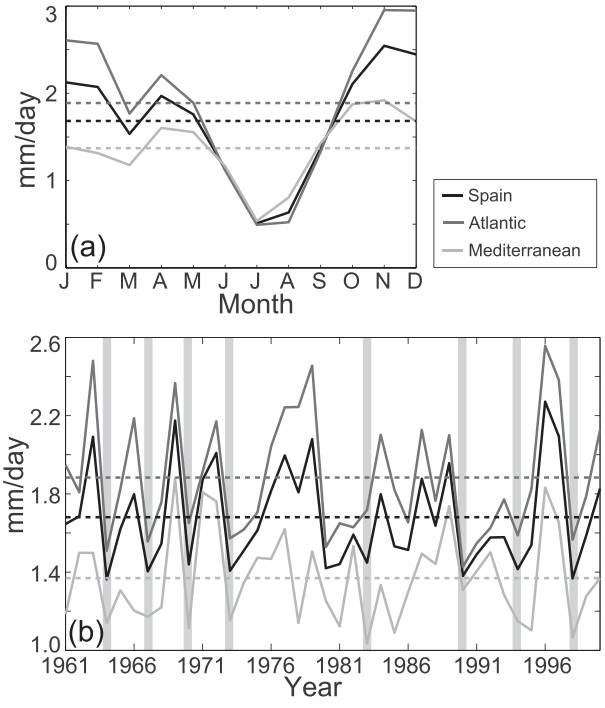

FIG. 2. (a) Intra- and (b) interannual variability of Spain02 precipitation as a whole (black), and the Atlantic (dark gray) and Mediterranean (light gray) regions. Dashed lines indicate the average values in the period 1961–2000. The eight driest years (according to the spatial mean of the pointwise standardized anomalies for the entire region) are indicated with gray shaded bars in (b).

et al. 2012), covering peninsular Spain and the Balearic Islands at a 0.2° resolution with a total of 1445 grid boxes (freely available at http://www.meteo.unican.es/datasets/spain02). Because of the denser network of stations (over 2000) used for its construction, Spain02 outperforms the European-based alternative (E-OBS; Haylock et al. 2008), particularly for the calculation of extreme indicators.

Figure 1 shows the spatial climatological values for the percentage of wet days (precipitation ≥ 0.1 mm; Fig. 1a) and the daily mean for the period 1961–2000 (Fig. 1b), which is the period of study considered in this work. Annual accumulated values range from 1000–2500 mm along the North Atlantic coast to 400–700 mm along the Mediterranean coast (with minimum values of 100 mm in the southeastern region). Figure 2a shows the annual cycle, which gradually changes from a distribution with a predominant rainy season (peaking in November–December) in the Atlantic Ocean region to a bimodal one (peaking in April–May and October–November) in the Mediterranean region. Whereas the Atlantic Ocean area is influenced by frontal systems throughout the year, precipitation along the Mediterranean coast is largely driven by cyclogenesis processes, mainly during September–November (Llasat 2009), resulting in different climates from the Atlantic Ocean to the Mediterranean Sea. This spatial variability provides an ideal test bed for precipitation downscaling studies (von Storch et al. 1993; Trigo and Palutikof 2001; Herrera et al. 2010; Turco et al. 2011). The black line in the maps indicates the water divide between the Atlantic Ocean and the Mediterranean Sea hydrological basins, which is considered in this work as a rough classification of the above precipitation regimes within the area of study.

b. *Historical dry period*

The interannual variability of the precipitation series in the three regions is shown in Fig. 2b. To analyze the stationarity of the different statistical downscaling (SD) techniques under changing climate conditions, the eight driest years (1964, 1998, 1994, 1990, 1970, 1967, 1983, and 1973) were computed according to the spatial mean of the pointwise standardized anomalies for all of Spain (note that interannual variability is very similar for the Atlantic and Mediterranean regions). The spatial mean anomaly for this driest 8-yr period is −21.5% w.r.t. the mean value of the remaining 32 years, with a spatial standard deviation of 7.5% (the results are very similar in the Atlantic and Mediterranean regions with −21.9% and −20.5% mean anomalies and 7.3% and 7.8%

TABLE 1. ERA-40 predictors used in this work. The 2D level refers to the two-dimensional surface variables. The time labels INS, DM, and DA refer to instantaneous (at 0000 UTC), daily mean, and daily accumulated values, respectively.

Label	Name	Units	Level	Time
Z	Geopotential	$m^2 s^{-2}$	500 hPa (Z500)	INS
T	Temperature	K	850 hPa (T850)	INS
Q	Specific humidity	$g kg^{-1}$	850 hPa (Q850)	INS
U	Zonal wind	$m s^{-1}$	500 hPa (U500)	INS
V	Meridional wind	$m s^{-1}$	500 hPa (V500)	INS
W	Vertical wind	$m s^{-1}$	850 hPa (W500)	INS
RV	Relative vorticity	s^{-1}	700 hPa (RV700)	INS
SLP	Sea level pressure	Pa	2D	DM
2T	2-m temperature	K	2D	DM
TP	Total precipitation	mm	2D	DA

TABLE 2. Overview of the GCMs used in this study, taken from the two streams of the ENSEMBLES project (van der Linden and Mitchell 2009). Stream 1 (S1) corresponds to the CMIP3 model versions, whereas S2 indicates new versions developed within the ENSEMBLES project. (Expansions of acronyms are available at http://www.ametsoc.org/PubsAcronymList.)

GCM name	Institution	Run	Stream
BCCR-BCM2.0	Bjerknes Centre for Climate Research, Bergen, Norway	1	S1
CNRM-CM3	Centre National de Recherches Météorologiques, Toulouse, France	1	S1
ECHAM5	Max Planck Institute, Hamburg, Germany	3	S1
HadGEM2	Hadley Centre, Exeter, United Kingdom	1	S2

standard deviations, respectively). This historical dry period will be used in this paper as a surrogate of possible future dry climate conditions in the twenty-first century (note that global and regional simulations project a precipitation decrease of around 20%–30% for the last third of the twenty-first century in Spain; Giorgi and Piero 2008). We want to remark here that although the above historical dry period does not corresponds with a proper (consecutive) climatological period (i.e., may not be representative of the dry climatological conditions projected in future scenarios), the present test has shown to provide useful information about the generalization ability of SDMs in future climate conditions when applied to temperatures in warmer conditions (Gutiérrez et al. 2013). Moreover, this test could be considered a minimum requirement for out-of-sample extrapolation of the SDMs.

c. *Predictors and preprocessing*

A number of predictors typically used to downscale precipitation were considered in this work (Table 1) both from reanalysis (ERA-40; Uppala 2005) and GCMs from the ENSEMBLES–CMIP3 project (Table 2) both for control (20C3M) and transient (A1B) scenario projections (van der Linden and Mitchell 2009). Note that reanalysis uncertainty has not been considered in this paper, since it plays a minor role in this particular region (Brands et al. 2012). Moreover, the predictors considered in this study are reasonably well reproduced (when compared with the corresponding reanalysis data) by the above GCMs over the area of study if the (mean) bias is removed (Brands et al. 2011a). To ensure a consistent definition of these variables among the different datasets, daily instantaneous values (at 0000 UTC) were chosen for the midtropospheric variables, whereas daily aggregated ones were considered for the surface variables. This information was readily available from the above datasets, except the surface aggregated data in the case of ERA-40, which was calculated from the 6-hourly available information. Moreover, relative vorticity (RV) was derived in all cases from the *U* and *V* fields by using the following equation (Pryor et al. 2005):

$$\mathrm{RV} = \frac{dV}{dx} - \frac{dU}{dy}, \tag{1}$$

where dV (dU) is the gradient along V (U) and dx (dy) is the gradient along the longitudes (latitudes).

ERA-40 data were obtained from the ECMWF Meteorological Archival and Retrieval System (MARS) server at their native resolution of 1.125° × 1.125°. Global projection data were obtained from the World Data Center for Climate (WDCC) Climate and Environmental Retrieval and Archive (CERA) database (http://cera-www.dkrz.de/CERA) for the 20C3M (1961–2000) and the A1B (2001–2100) scenarios. The native horizontal resolution of the GCMs ranges from 1.25° to 3.75°. Therefore, all predictor data were regridded on a common regular 2° × 2° lattice by bilinear interpolation. Outliers or "bugs" in GCM fields (particularly for relative humidity) were processed as described in Brands et al. (2011a). Moreover, the mean bias of the GCM was adjusted variable by variable and grid box by grid box by removing the mean annual cycle (monthly means) and adding the one corresponding to the reanalysis data. Note that this correction also introduces some partial adjustment of the variance (that is due to different annual cycle amplitudes in the reanalysis and the GCM). Other authors introduce further adjustments (e.g., in the variance) with some additional benefits (Cheng et al. 2008), but we have tried to keep model preprocessing as simple as possible in this work.

For different configurations of the downscaling techniques described in the next section, we consider both

TABLE 3. Downscaling methods of four different families considered in this work: Analog methods (AM), weather typing (WT), generalized linear model (GLM), and GLMs conditioned on weather types (GLM-WT). Another method is the particular case of a weather generator conditioned to circulation (as given by the method type WT-WG). Methods with a stochastic component are appended with an asterisk to the type code. See the text for further details.

Label	Type	Method and predictor field
M1a	AM	Nearest neighbor (1 analog).
M1b	AM	Mean of five neighbors.
M1c	AM*	One out of 15 neighbors, random selection.
M2a	WT	100 WTs (*k* means), mean of the observations.
M2b	WT*	100 WTs (*k* means), random selection.
M2c	WT-WG*	100 WTs (*k* means), simulation from Bernoulli + gamma fitted distribution.
M3a	GLM*	*n* PCs (95% variance).
M3b	GLM*	Local predictor values in the nearest grid box.
M3c	GLM*	Local predictor values in the four nearest grid boxes.
M3d	GLM*	15 PCs + nearest grid box.
M4a	GLM-WT*	M3b conditioned on 10 WTs (SLP only).
M4b	GLM-WT*	M3c conditioned on 10 WTs (SLP only).

pointwise and/or spatial-wise predictors from the above datasets at nearby grid boxes and/or the principal components (PCs) corresponding to the EOFs (Preisendorfer 1988) of the (joined) standardized predictor fields, respectively; the EOFs are calculated using the ERA-40 data and then the GCM fields are projected accordingly. In this latter case, the total number of PCs is limited to those yielding a fraction of explained variance of 95%, not exceeding a maximum of 30 PCs in any case. The spatial homogeneity of the downscaled series for pointwise predictors is expected to be low, whereas applying PCs should considerably enhance the spatial homogeneity of the results.

d. Regional projections from the ENSEMBLES RCMs

Finally, in order to compare the future regional projections obtained with statistical and dynamical downscaling approaches over Spain, we consider the ensemble of regional climate models (RCMs) in the ENSEMBLES project produced using full boundary conditions from the GCMs (see, e.g., Herrera et al. 2010). To make the statistical and dynamical ensembles as comparable as possible, we have selected a subset of RCMs coupled to the same GCMs used in this study (see Table 2)—with the exception of the HadGEM2 model, which is replaced in the RCM ensemble by the HadCM3Q0, and excluding some badly performing couplings in this region as indicated in Turco et al. (2013). In particular, we consider the following RCM–GCM couplings: ALADIN–ARPEGE, HIRHAM–ARPEGE, CLM–HadCM3Q0, HadRM3Q0–HadCM3Q0, PROMES–HadCM3Q0, RCA–ECHAM5r3, RACMO–ECHAM5r3, HIRHAM–BCM, M-REMO–ECHAM5r3, and RCA–BCM (see Turco et al. 2013, their Table 1). A detailed regional analysis of the climate projections obtained with this ensemble is shown in Turco et al. (2015).

3. Statistical downscaling methods

A number of different statistical deterministic and stochastic precipitation downscaling methods commonly used in the literature to downscale climate change scenarios under the perfect prognosis (PP) approach are analyzed in this paper, considering different configurations (predictors and spatial domains). In all cases, the methods are trained (and cross validated) using predictors from the reanalysis data; afterward, local projections are obtained by applying the fitted or calibrated methods to the predictors simulated by the GCMs. These methods are described in Table 3 and have been classified as follows:

- analog methods (labeled M1),
- weather typing methods (labeled M2),
- generalized linear models (labeled M3), and
- GLMs conditioned on weather types (labeled M4).

The analog method (Lorenz 1969; Zorita and von Storch 1999) is a popular nonparametric downscaling technique based on the assumption that similar local occurrences are expected for similar atmospheric configurations, as measured by the Euclidean distance in this work. This approach is applicable to a wide range of target and predictand variables yielding spatially consistent results at the multiple local sites. This methodology has been applied in a variety of studies to downscale rainfall under climate change conditions (Wetterhall et al. 2005; Brands et al. 2011b; Cubasch et al. 1996; Timbal et al. 2003; Moron et al. 2008; Timbal and Jones 2008; Teutschbein et al. 2011). However, since the AM cannot predict values outside the observed range, it is particularly sensitive to nonstationarities arising in climate change studies (Benestad 2010; Gutiérrez et al. 2013). Table 3 describes three typical configurations of this technique

used in this work (labeled M1a, M1b, and M1c), which consider the closest analog, the mean of the five closest analogs, or a random analog (out of the set of 15 closest ones), respectively. The latter configuration is usually referred to as nearest-neighbor resampling (Beersma and Buishand 2003) and can be considered a stochastic variant of the analog methodology.

The second family of methods used in this study (see M2a through M2c in Table 3) includes three different WT techniques. These methods are also based on the concept of similarity among atmospheric patterns, which are preclassified into a number of homogeneous clusters, or weather types (Gutiérrez et al. 2004; Philipp et al. 2010; Jacobeit 2010). These methods have been applied under climate change conditions in a number of studies (see, e.g., Goodess and Palutikof 1998; Cheng et al. 2011). In this study, the *k*-means algorithm is used to perform a clustering over the historical reanalysis database (considering the joined standardized predictor fields), so each resulting weather type is characterized by a representative pattern (or centroid) with a characteristic local weather given by the corresponding historical observations. The different configurations used in this paper consider different alternatives to provide series of local weather from a particular weather type: The mean of the observations (M2a), a random observation within the subgroup (M2b), or a value simulated from a Bernoulli (for rainfall occurrence) and gamma (for rainfall amount) fitted distribution (within each subgroup); this latter method can be considered as a simple (i.e., including no explicit component for autocorrelation) weather generator (WG) conditioned by circulation (therefore, it is labeled WT-WG). The latter two configurations have a stochastic component and were chosen to avoid the main shortcoming of weather typing techniques, which is the reduction of the variance (Enke and Spegat 1997). Moreover, M2c can simulate predictand values beyond the observed range. A sensitivity experiment to determine the optimum k to be used (keeping a balance between forecast error and predicted variability) was performed, yielding the best results for $k \simeq 100$.

The third family is based on GLMs. These models are an extension of linear regression allowing for nonnormal predictand distributions [see Nelder and Wedderburn (1972) for an introduction], which have been used for downscaling precipitation from global climate change scenarios in a number of studies (Brandsma and Buishand 1997; Fealy and Sweeney 2007; Hertig et al. 2013). The methods considered in this work (see M3a to M3d in Table 3) follow the typical two-stage implementation used to model precipitation in the literature, consisting of a GLM with Bernoulli distribution and logit link for occurrence (equivalent to a logistic regression) and a GLM with gamma distribution and log link for the amount (see, e.g., Coe and Stern 1982; Chandler and Wheater 2002; Abaurrea and Asín 2005). The only difference among the four configurations used in this work is the spatial character of the considered predictors. In M4a the predictor data are the leading PCs, whereas for M3b (M3c) the standardized anomalies at the nearest (four nearest) grid point(s) are used. M3d combines the 15 leading PCs with standardized anomalies at the nearest grid point, in order to account for both spatial and local effects. In all cases, values are simulated for both occurrence and amount from the resulting predicted distributions; in the case of the amount, the shape parameter of the gamma distribution was kept constant.

Note that several extensions of regression methods have been presented in the literature to explicitly include appropriate intersite dependences in the simulation process (Yang et al. 2005). However, in this work we do not evaluate the spatial consistency and/or correlation of the results, and therefore these extensions are not considered.

The fourth and last family (GLM-WT) includes circulation-conditioned versions of the GLM methods M3b and M3c (both using standardized anomalies at the nearest grid points as predictors). In particular, 10 weather types are calculated by conducting the *k*-means algorithm restricted to the circulation variables (SLP, Z, RV, U, and V) included in the predictor pattern. Then, the GLMs are fitted on each weather type using the remaining predictor variables (2T, T, W, and Q). In contrast to the WT family, where 100 weather types are considered, only 10 are used in this case since further discriminating power is provided by the GLMs.

4. Cross-validation procedure

To validate the aforementioned SDMs, we followed the same *k*-fold cross-validation approach introduced in Gutiérrez et al. (2013) (the companion paper). Therefore, the 40-yr period 1961–2000 was randomly split into five ($k = 5$) 8-yr sets (folds). Each of these sets was used once for testing, using the remaining 32 years for training the SD methods. Note that for the case of temperatures in the companion paper, a stratification approach was followed instead to avoid the influence of the existing trends in the resulting folds, so all of them would have the same climatological distribution of the initial sample and, thus, are representative of the climatological period (normal conditions).

The results downscaled for the five test periods were merged into a unique series, covering the whole 40-yr period, which was validated against the observations at

each grid box to evaluate 1) the accuracy, 2) the distributional similarity of the observed and downscaled series, and 3) the robustness of the methods to changing climate conditions. The accuracy (day-to-day correspondence) is the basis of statistical downscaling methods under the PP approach used in this paper. Distributional similarity is required since the (daily) downscaling methods should properly reproduce the observed daily distributions; this also avoids post hoc corrections and/or calibration of the downscaled series. Finally, model stationarity is required to apply the methods in changing climate conditions. To this aim, the following evaluation scores are used in this work:

1) Correlation: To measure the day-to-day correspondence between the downscaled and observed series we used the Spearman rank correlation coefficient, since it is nonparametric and robust to outliers in the series. We computed the correlation both at a gridpoint level and for the regional (Atlantic or Mediterranean region) mean series. Moreover, we computed the correlation for both the daily and the 10-daily aggregated series in order to better capture the accuracy not only for daily precipitation but also for precipitation episodes. Note that, for the stochastic methods, larger correlations could be obtained if only the deterministic component (i.e., the mean of multiple realizations) were validated. However, since we mainly use this score to evaluate the relative improvement of different predictor configurations, we consider the simple case of validating a single stochastic realization.
2) Relative bias: Mean error between the downscaled and observed precipitation series, relative to the precipitation amount at the specific grid-box of Spain02 (expressed in percent).
3) Seasonal bias variability: The standard deviation of the season-specific biases (DJF, MAM, JJA, and SON) is used to measure if the bias is constant/systematic throughout the year (this score is referred to as sigma bias). High values should be avoided since they indicate the need to separately calibrate the methods for each season.
4) Distributional similarity (occurrence) of the binary series of downscaled and observed precipitation occurrences, as defined by a threshold of 0.1 mm. This similarity is measured in terms of the ratio between the relative frequencies of downscaled and observed wet days, as well as the p value of a Z test for the difference between these frequencies, under the null hypothesis that they are equal (denoted as the Z-p value and given in logarithmic scale). Thus, values smaller than -2 indicate a significant difference at a 0.01 level.
5) Distributional similarity (amount) of the downscaled and observed rainy precipitation series, as measured by the two-sample Kolmogorov–Smirnov test (KS test). Under the null hypothesis, both the observed and downscaled time series come from the same underlying distribution. The p value of the test (denoted as the KS-p value, which is given in a logarithmic scale) is used to measure the degree of dissimilarity between the distributions (e.g., values smaller than -2 indicate a significant difference at a 0.01 level). To alleviate the effect of serial correlation on the calculus of the KS-p value, only one every five time steps was considered for the calculation of the KS-p value.
6) Test of stationarity under dry conditions: To statistically test whether or not the performance of a method could vary in changing climate conditions, we consider a test based on dry historical observed periods. In particular we focus on the bias and apply the two-sided Student's t test to determine whether the bias in a historical 8-yr dry test period (see section 2) is significantly different from the biases in random sample of 8-yr test periods, as given by the five test periods of the 5-fold cross-validation. Note that in this case we use the variability of the validation score (the bias in this case) in the five 8-yr sets in order to characterize the random fluctuations of the score in normal conditions.

   The p value from the test (denoted as the dry-p value) is used to quantify the robustness of the methods in changing climate conditions. If the bias in the dry period is significantly larger (or smaller) than that obtained in normal or random conditions (indicated by low p values, e.g., smaller than 0.01), then the method significantly over (or under) estimates the dry period and, therefore, would not be suitable for downscaling transient scenario runs due to the unpredictable consequences of the changing bias. Note that this test is not a sufficient condition for the robustness of the methods in climate change conditions, since the (nonconsecutive) dry period used in this study could not represent the dry periods in a differently forced future climate. The reader is referred to Gutiérrez et al. (2013) (the companion paper) for further details on this test.

Since seasons may change in the future (e.g., more summer-like days) as a consequence of climate change (Ruosteenoja and Räisänen 2013), calibrating the methods separately for each specific season could have uncontrollable effects in the precipitation downscaled from GCM scenario runs (Imbert and Benestad 2005). Therefore, in this work, all the methods were calibrated

TABLE 4. Combinations of predictors considered in this work (see the text for further details on different static and/or dynamic configurations of the predictor set).

Label	Variables
P1	TP
P2	W850
P3	W850 and RV700
P4	SLP, W850, and RV700
P5	SLP, T850, and Q850
P6	SLP, T850, Q850, and Z500
P7	SLP, T850, Q850, U500, and V500
P8	T850, Q850, and Z500
P9	SLP, T850, and Z500
P10	SLP, 2T, Q850, and Z500

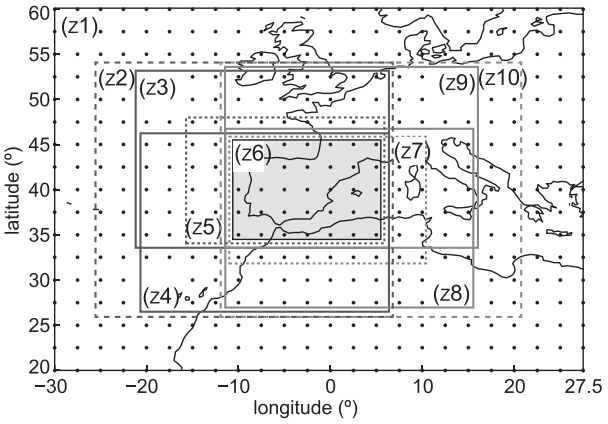

FIG. 3. The 10 geographical domains used in Gutiérrez et al. (2013), with increasing numbering from east to west. The smallest domain is the center domain (Z6), increasing in size toward the extremes (Z1 and Z10).

considering the complete yearly data (i.e., not season specific). However, the above scores were calculated both for the annual and the seasonal downscaled series (i.e., the validation is performed at both an annual and seasonal level). The seasons considered for validation were the standard boreal winter (DJF), spring (MAM), summer (JJA), and autumn (SON).

5. Screening of predictors and geographical domains

An exhaustive screening of the 10 (P1–P10) predictor sets (combinations) listed in Table 4 and the 10 geographical domains (Z1–Z10) shown in Fig. 3—the same used in Gutiérrez et al. (2013)—was carried out in order to find the optimum predictor–domain configuration yielding the best results. Note that other predictor sets were also tested. However, they did not provide any added information to the above combinations and are thus not shown. For the sake of simplicity, only two representative methods—one from the analog (M1a) and the other from the GLM (M4a) families—were considered for this screening. Note that both methods consider spatial-wise predictors and, therefore, their performance depends on the particular geographical domain used.

The choice of the domains used was based on the lessons learned in Timbal and McAvaney (2001), Timbal et al. (2003), Gutiérrez et al. (2004), Brands et al. (2011b), and Gutiérrez et al. (2013), who found that a small areal window covering the region of study was the optimal for downscaling daily data. Regarding the predictor sets, different combinations of the most suitable variables for downscaling precipitation in the area of interest (see Table 1) were defined according to the previous studies found in the literature.

TP was selected as a benchmarking predictor (P1) since several studies have found this variable to yield good results, both using data from reanalysis (Widmann et al. 2003) and, more recently, from RCMs nested into a reanalysis (Turco et al. 2011) and from GCMs nudged to a reanalysis (Eden et al. 2012). However, precipitation is not used directly as a predictor in PP approaches, since it is largely affected by the model orography and parameterizations and therefore it is differently represented in reanalysis and GCM simulations. Moreover, the performance of the GCMs for this variable is normally assumed to be poor [see Trigo and Palutikof (2001) for a study over Iberia].

To account for vertical motion, which is expected to be important during summer and in the Mediterranean region, W850 was considered either solely (P2) or in combination with RV700 (P3) and SLP (P4), taking into account some of the considerations made in an early study (Sauter and Venema 2011). The importance of including humidity into the predictor field (Charles et al. 1999) is reflected in the remaining predictor sets. The P5 predictor set (SLP, T850, and Q850) was found to be optimal in comparable studies conducted in western France (Timbal et al. 2003); note that they used vertically integrated water vapor instead of Q. In this work, this combination was modified by adding midtropospheric circulation variables (Z500 in P6 and U500 and V500 in P7) and by changing Q850 by Z500 (P9), for sensitivity testing purposes. Additionally, the predictive power of SLP (Z500) can be tested by comparison of P7 and P8 (P5 and P6). Finally, differences arising when considering 2T or T850 (Hanssen-Bauer et al. 2005) can be analyzed by comparing P7 with P10; note that using T850 is preferable since GCMs perform better for this variable (Brands et al. 2013).

Following the indications by Gutiérrez et al. (2004) both static and dynamic temporal configurations of the predictor sets listed in Table 4 were tested. For the instantaneous variables (see Table 1), the former

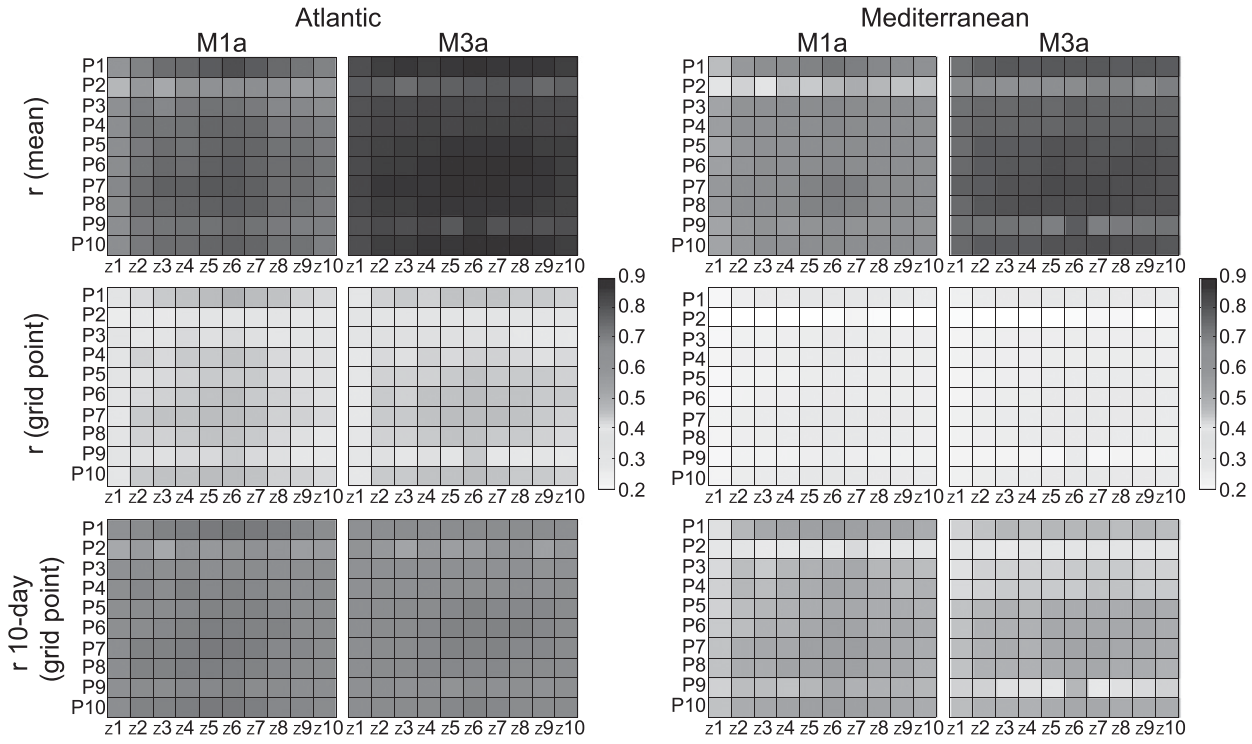

FIG. 4. Validation results from the screening, in terms of accuracy, for the M1a (AM) and M3a (GLM) methods in the (left) Atlantic and (right) Mediterranean regions. The different predictors (domains) are displayed through the y (x) axis. (top) Spearman correlation for the spatial mean of the pointwise daily precipitation series. (middle) Spatial mean of the pointwise Spearman correlations for the daily precipitation series. (bottom) Spatial mean of the pointwise Spearman correlations for the 10-day mean precipitation series.

considers a unique value per day (at 0000 UTC) whereas the latter additionally includes the 0000 UTC values for day $D + 1$, thus providing a window covering the observation period. In contrast to Gutiérrez et al. (2013), the obtained results revealed that the dynamic configuration performed systematically better than the static one for all the scores (this result is still true using 1200 UTC values instead of 0000 UTC ones for the static configurations). Therefore, only the dynamic temporal setup was considered in the following.

Figure 4 shows the annual validation results, in terms of accuracy, from the screening. The first row of Fig. 4 shows the Spearman correlation for the spatial mean of the daily series, whereas the second (third) row of Fig. 4 shows the spatial mean of the pointwise correlations for the daily (10-day mean) series. Note that the validation is performed separately for the Atlantic Ocean and the Mediterranean Sea basins (Fig. 4, left and right panels, respectively). As can be seen, results are more sensitive to the predictor set than to the geographical domain and are generally better in the Atlantic region than in the Mediterranean region. Pointwise correlations are improved by temporal aggregation, increasing from 0.2–0.5 for the daily series to 0.5–0.7 for the 10-day mean values (results for the monthly mean series are only slightly better than in the latter case; not shown). Furthermore, the GLM method (M4a) clearly outperforms the analog approach (M1a) for the spatial mean series, since the stochastic variability of the GLM downscaled series is partially averaged out in this case. A similar result is obtained for the daily and 10-daily correlations using the deterministic estimate (the mean) provided by the GLM (not shown).

In general, the worst-performing predictor sets, for both the Atlantic and the Mediterranean regions, are those including W850 (especially P2), which is contrary to the results of Reichert et al. (1999), and P9, which does not include humidity. The best results are obtained with P1 (precipitation) as well as with P6–P8 and P10, indicating that the reference predictor combination (SLP, T850, and Q850) applied in Timbal et al. (2003) can be slightly improved by including midtropospheric circulation variables (either Z500 or U500 and V500) and that T850 can be substituted by 2T without suffering a notable correlation decrease. In accordance with Timbal et al. (2003), including moisture information (as represented by Q850 in our study) to the predictor field and using relatively small domains maximizes the accuracy.

Figure 5 displays the spatial average of the pointwise biases for the whole year (first column, hereafter referred to as annual bias) and for the four seasons (rows two through five). Again, results are more sensitive to the predictors than to the geographical domain. Furthermore, each method yields overall similar results in both the Atlantic Ocean and the Mediterranean Sea basins. As can be seen, the annual biases for the analog method are systematically negative (dry) for any predictor–domain combination, being larger for P1 and those combinations including W850 (P2–P4), whereas they are almost null in all cases (except for P1) for the GLM method. The last row of Fig. 5 shows the seasonal variability of the bias (sigma bias; see section 4). The larger this score, then the larger the variability of the bias across seasons and, thus, the more unsuitable the method for climate change applications. As shown in the Fig. 5, sigma bias is larger in the Mediterranean region than in the Atlantic region for both the analog and the GLM approaches. Moreover, the largest values are found for those predictor datasets including W850 (P2–P4), or excluding humidity (P9), which suggests again the inadequacy of those combinations. Among the rest of combinations, results are similarly acceptable, with P5 and P6 yielding slightly lower fluctuations for both methods and in both regions. Finally, note that some cases with small annual bias exhibit largely different seasonal biases (even of different sign; see, e.g., P9). Thus, in addition to the bias, the seasonal bias variability should be controlled for the appropriate application of statistical downscaling methods in climate change studies.

Figure 6 shows the validation results for the rest of the scores related to the distributional similarity (see section 4). The first and second rows of Fig. 6 correspond to the ratio of wet days and the corresponding *Z*-*p* value. The third and fourth rows of Fig. 6 show the KS-*p* value (in logarithmic scale) for winter and summer (the seasons presenting the largest problems), respectively. The last row of Fig. 6 shows the *p* value (also in logarithmic scale) from the test of robustness in anomalous dry conditions (dry-*p* value), calculated for the spatial mean bias. As in Figs. 4 and 5, results are more dependent on the predictor combination than on the domain. Furthermore, each method yields overall similar results in both regions. P1 (TP) and P2–P4 (combinations including W850) present problems in the ratio of dry days for the analog approach (note that the occurrence component of the GLM is fitted to data and hence the frequencies are well modeled). Moreover, these combinations, together with P9 (which does not include humidity) lead to distributional problems, both in winter and in summer, for the GLM technique, which seems to be more sensitive to the predictor data than the analog one in terms of distributional similarity. Finally, regarding the robustness to anomalous dry conditions, although results exhibit a considerable variability, they are slightly better in the Mediterranean region than in the Atlantic region. P5 and P6 yield overall the best results for both the analog and the GLM method and in both regions.

The latter results point out the necessity of including Q850 (and excluding W850) among the predictors, since this yields the best results in terms of accuracy and distributional similarity. Thus, the five predictor sets P5–P8 and P10 (P1 is used in this work for benchmarking purposes) perform similarly in terms of accuracy and distributional similarity when defined over a small domain. Moreover, P5 leads to the most robust results under anomalous dry conditions for both methods and together with Z7 domain provides a compromise between having a small sigma bias and a non-significant (at a 99% level) dry-*p* value. Therefore, the particular predictor–domain combination of P5–Z7 was selected as the optimal configuration, which will be used in the following to intercompare the performance of the different downscaling methods.

Performance of the optimal configuration

To further assess the performance of the two reference downscaling methods (M1a and M3a) with the optimal predictor–domain configuration of P5–Z7 at a grid box level, a number of mean and extreme precipitation indicators have been considered [see Table 5; data extracted from the ETCCDI (http://etccdi.pacificclimate.org)]. Figure 7 shows the maps of the resulting cross-validation results. As can be seen from Fig. 7, both methods reproduce accurately the spatial distribution of mean precipitation (PRCPTOT), precipitation intensity (SDII), dry and wet spells (CDD and CWD, respectively) and percentage of rainy days over 20 mm (R20). Furthermore, M1a also describes properly the indices related to extreme precipitation. However, the GLM approach overestimates both the precipitation in the rainiest day (RX1DAY) and the contribution of the top 5% rainy events to the total precipitation (R95PTOT). Moreover, the spatial distribution of the latter is also wrong for this method. Note that previous studies such as those of Fealy and Sweeney (2007) and Hertig and Jacobeit (2013) point out the difficulties in predicting extreme precipitation events with GLMs.

6. Assessment of the SDMs with perfect predictors

Once the optimal configuration of predictors and geographical domain, P5–Z7, was determined, the

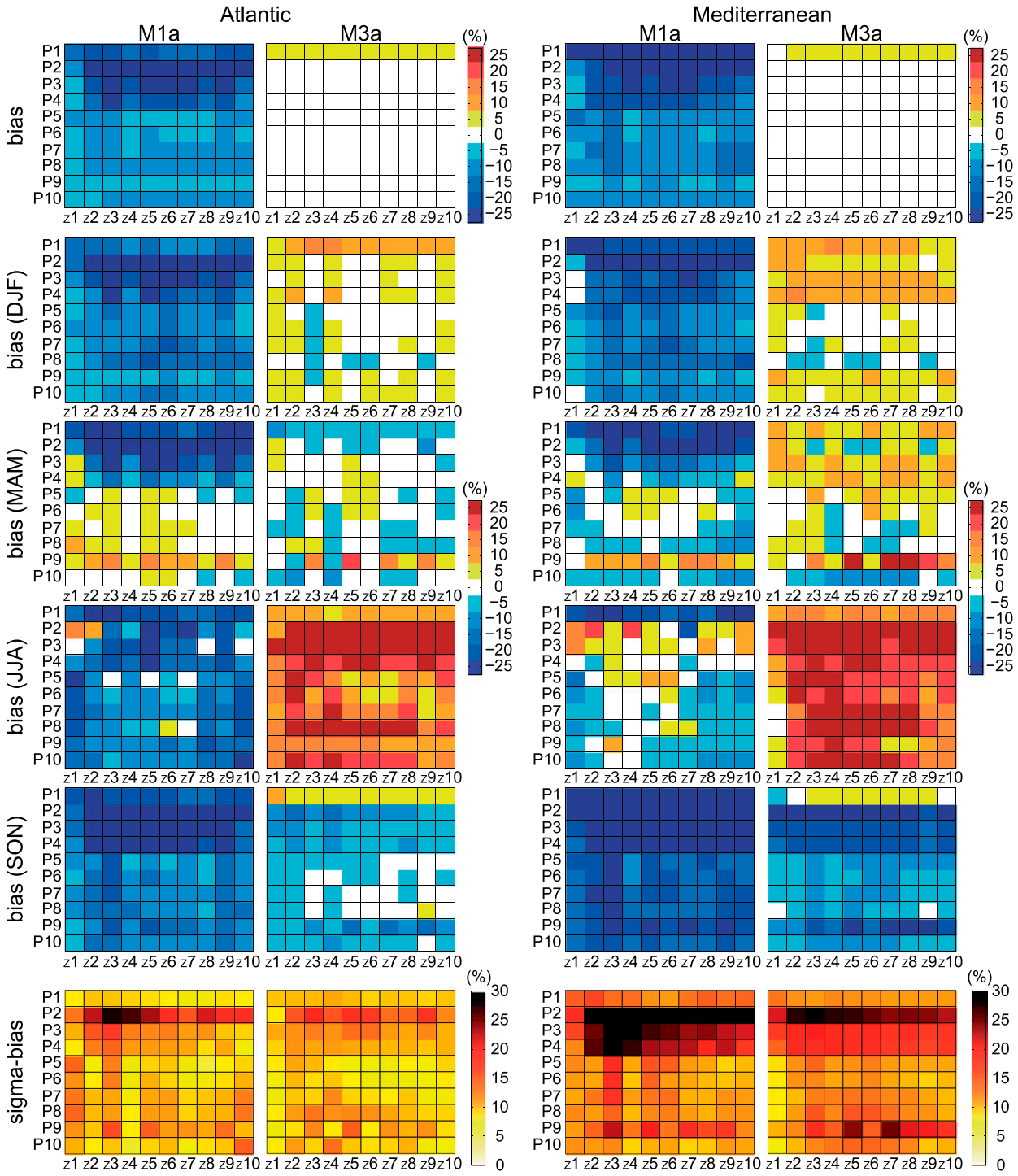

FIG. 5. As in Fig. 4, but for different scores related to the bias. (top) Spatial mean of the pointwise relative biases (%) for the complete series. (middle, rows 2–5) Spatial mean of the pointwise relative biases for each season of the year. (bottom) Standard deviation of the four season-specific spatial mean of the pointwise relative biases (sigma bias). Note that methods are calibrated considering the complete (i.e., not season specific) historical database.

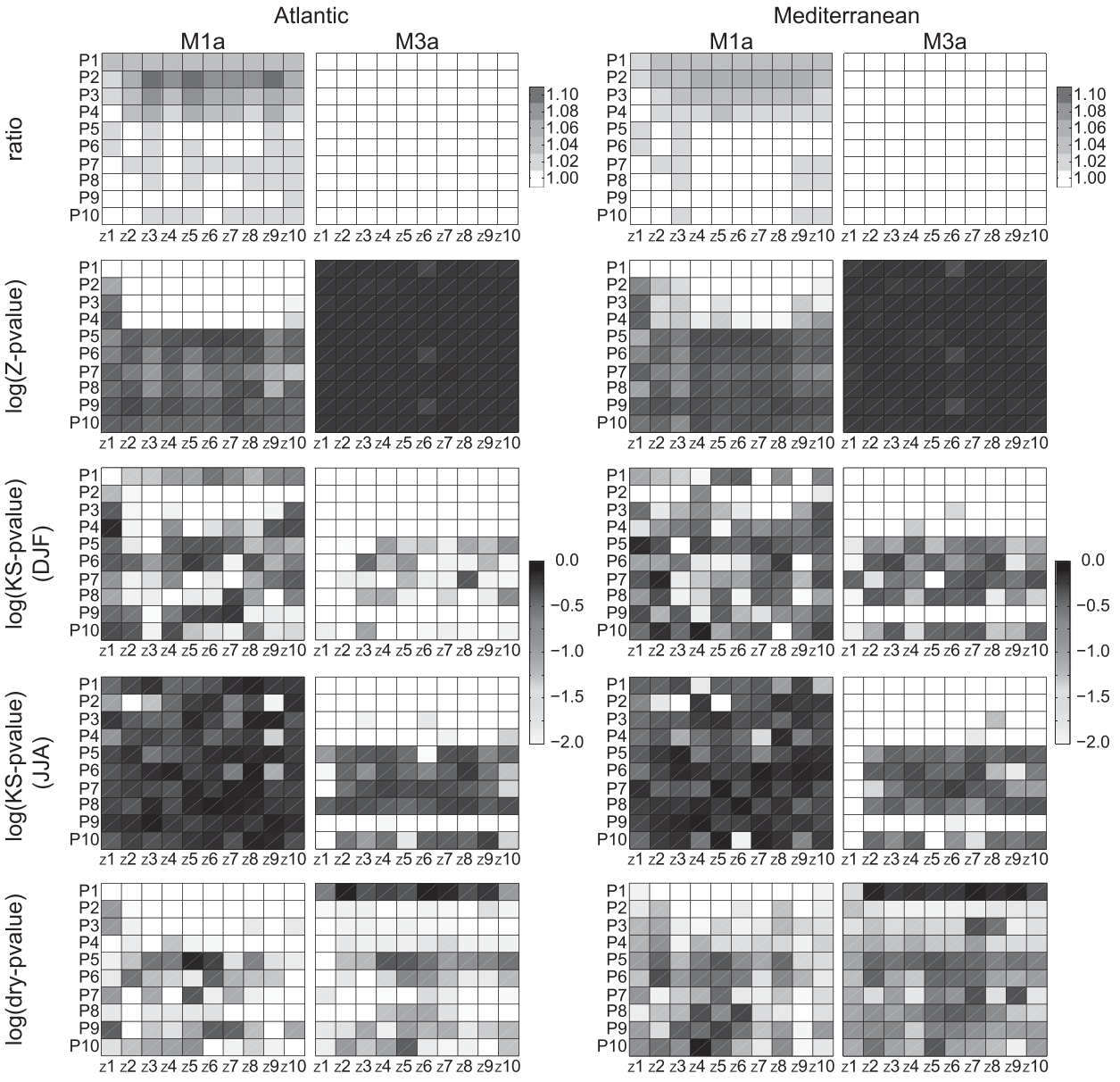

FIG. 6. As in Fig. 4, but for different distributional similarity scores. (top, rows 1 and 2) Ratio of downscaled/observed dry days and the logarithm of the corresponding p value from a Z test for proportions. (middle, rows 3 and 4) Logarithm of the p values from the KS test for the wet-day distributions in DJF and JJA. (bottom) Logarithm of the p value from the test of robustness in anomalous dry conditions.

performance of all the methods listed in Table 3 was first assessed in terms of accuracy (correlation) and distributional similarity for the period 1961–2000. To this aim, the validation scores described in section 4 were computed at each grid box of Spain02 following the same cross-validation procedure as in the screening process. The resulting 1445 pointwise (spatial) scores for each of the methods, representing the performance of the model across the region of study, are represented in Fig. 8 by means of a box-and-whisker plot. The black box covers the interquartile range, whereas the gray line indicates the median and the whiskers the minimum and maximum values. Notable differences between the Atlantic Ocean and the Mediterranean Sea basins are only found for correlation, with higher values in the former region for all methods. For the other scores, results are very similar in both regions. Methods M1b and M2a perform better in terms of accuracy but worse in terms of distributional similarity, failing to predict the frequency and the precipitation distribution of wet days (note that the KS-p values are under 10^{-4} in these cases and

TABLE 5. Mean and extreme precipitation indicators used in this work (see http://etccdi.pacificclimate.org for further details).

Indicator	Units	Description
PRCPTOT	mm	Mean precipitation per day.
SDII	mm	Mean precipitation per wet day.
R20	%	Percentage of days (over the total) with precipitation ≥ 20 mm.
CDD	day	Maximum number of consecutive dry (precipitation < 1 mm) days.
CWD	day	Maximum number of consecutive wet (precipitation ≥ 1 mm) days.
RX1DAY	mm	Precipitation in the rainiest day.
R95PTOT	%	Percentage of precipitation (over the total) in the 5% of rainiest days.

therefore are not shown in Fig. 8). The latter undesired effect is due to a reduction in the predicted variance, since predictions are obtained by averaging a number of observations.

Among the rest of techniques, the two analog alternatives, M1a and M1c, perform similarly well. However, the latter exhibits slightly lower correlations, larger biases, and larger seasonal bias variability (notice its stochastic

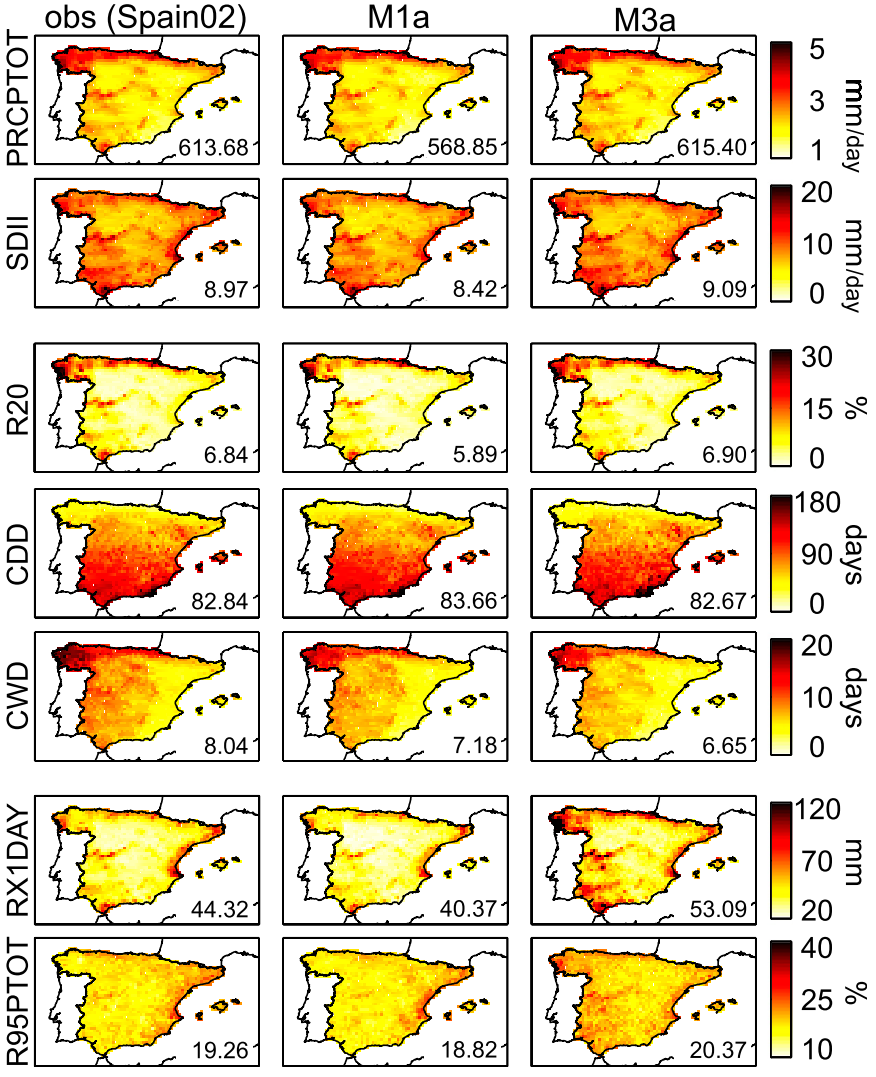

FIG. 7. (left) Observed downscaled mean and extreme precipitation indicators (see Table 5) for the period 1961–2000, considering the optimum predictor–domain configuration P5–Z7 and the (center) M1a and (right) M3a methods. The numbers in the bottom right of the panels show the spatial mean values.

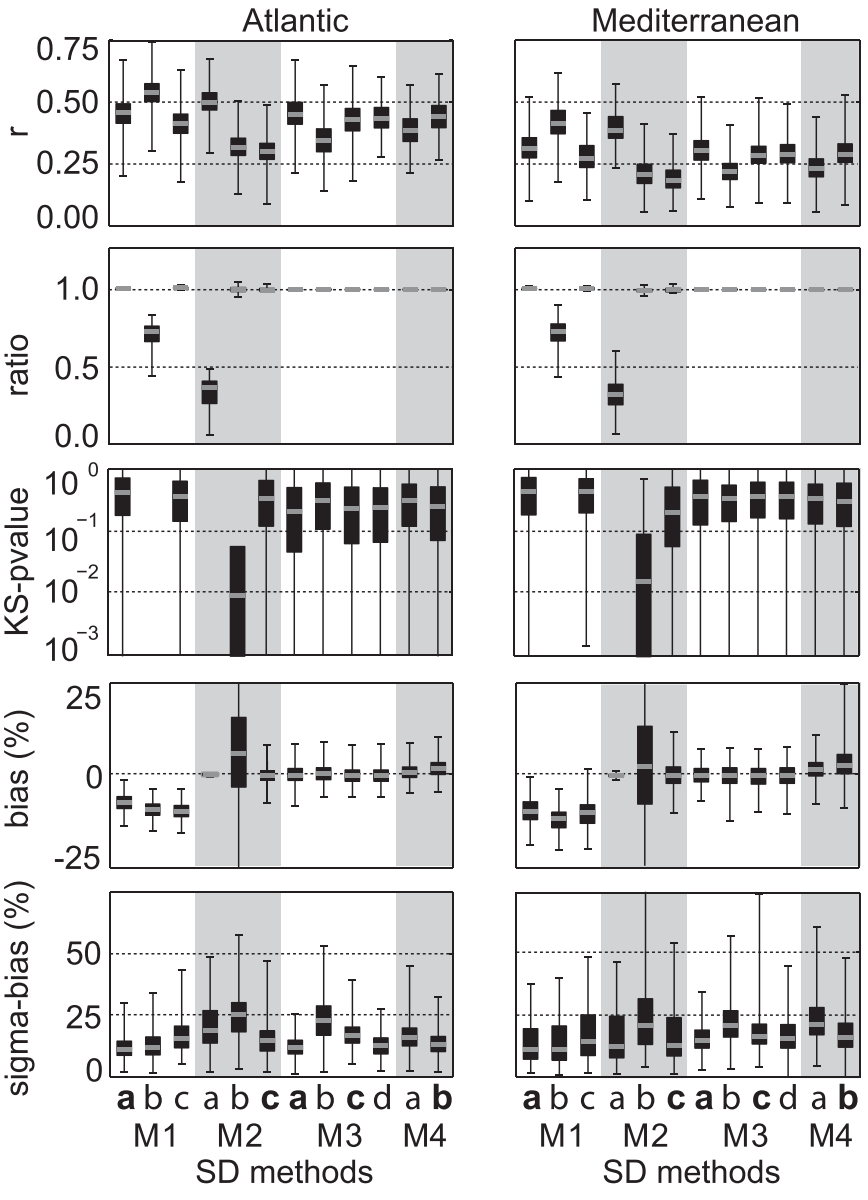

FIG. 8. Pointwise (spatial) results from the validation, in terms of accuracy and distributional similarity, of all the methods in Table 3 for the period 1961–2000, considering the optimum predictor–domain configuration of P5–Z7. The different validation scores considered (see section 4) are displayed in rows. The box-and-whisker plots show the distribution of the scores over the 874 (571) grid points in the (left) Atlantic Ocean and (right) Mediterranean Sea basins. The five techniques finally considered to form the ensemble of downscaling methods (see the text for details) are indicated by boldface labels.

character). Regarding the weather typing techniques, M2b simulates the occurrence slightly better than M2c, but it presents limitations in reproducing the amount of rain in wet days (KS-p values below 0.01) in approximately half of the grid boxes. Furthermore, M2b shows larger bias and larger seasonal bias variability.

Among the GLM techniques, all of them perform overall well. The differences between unconditioned and conditioned (on weather types) approaches (M3 and M4 families, respectively) are smaller than those related to the spatial character of the predictors. In particular, methods considering only the nearest grid box (M3b and M4a) exhibit lower correlations and larger seasonal bias variability than those considering the four nearest grid boxes and/or PCs (M3a, M3c, M3d, and M4b).

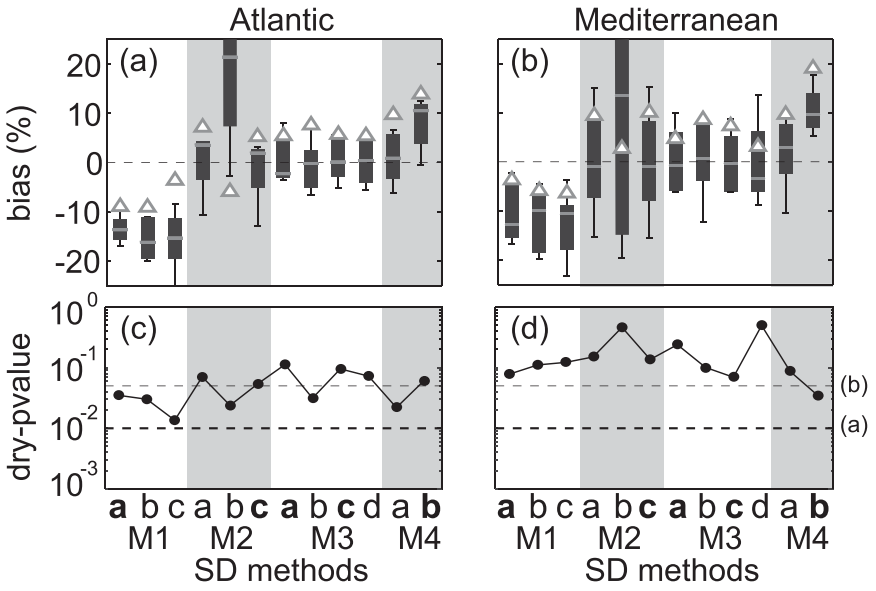

FIG. 9. Results from the two-sided Student's *t* test for robustness for all the methods in Table 3 considering the optimum predictor–domain configuration of P5–Z7. (a),(b) The box-and-whisker plots showing the biases for the five ($k = 5$) 8-yr sets in normal conditions (the gray line corresponds to the mean value). Triangles mark the bias in anomalously dry conditions. (c),(d) The *p* values (a logarithmic scale is used) from the test in anomalously dry conditions. Note that *p* values under 0.05 and 0.01 indicate lack of robustness to changing climate conditions at a 5% and 1% significance level (dashed lines), respectively.

Figure 9 shows the results from the test for the stationarity of the methods, considering the mean value, under anomalously dry conditions. For each method, the bias obtained in anomalously dry conditions (indicated by a triangle) is compared to the biases obtained for the five ($k = 5$) 8-yr sets in normal conditions (represented by the box plots) using the two-sided Student's *t* test (see section 4 for details). The *p* value from the test (dry-*p* value) is shown in Figs. 9c,d in logarithmic scale. Note that *p* values under 0.05 and 0.01 indicate that biases in dry conditions are significantly different from biases in normal conditions at a 5% and 1% significance level, respectively (these two threshold values are marked with a dashed line in Fig. 9).

Contrary to the results in Gutiérrez et al. (2013) for the case of temperatures, where analog and weather typing methods were shown to significantly underestimate warm conditions, the overall results for the case of precipitation are very similar for the different families of statistical downscaling methods for a given predictor configuration (the optimum predictor–domain configuration of P5–Z7 in this case). In particular, none of the techniques exhibits significant differences at a 1% level, although the results for the Atlantic region are slightly worse (there are significant differences at a 5% level for some of the methods). Therefore, in contrast to the case of temperatures, although this test can identify predictor configurations with poor extrapolation capabilities for anomalous dry conditions, it fails to provide any clear indication on the differences observed in the future precipitation projections for different downscaling methods.

7. Downscaling global climate projections

According to the previous validation results for the temporal, marginal, and extrapolation aspects, we selected a reduced number of suitable SDMs (representative of the different families) for downscaling daily precipitation from global climate projections. In particular, we selected an ensemble of five methods M1a, M2c, M3a, M3c, and M4b (indicated by boldface labels on the *x* axis in Figs. 8 and 9) with overall good performance. Note that other alternative selections could be equally considered, as long as deficient methods are discarded in order to properly assess the uncertainty of future climate projections, avoiding the noise introduced by unsuitable models.

The resulting ensemble of five SDMs (calibrated with reanalysis data) was applied to the four GCMs from the ENSEMBLES project shown in Table 2 to obtain a 20-member ensemble of historical (1961–2000, predictors from the 20C3M scenario) and future (2001–2100, predictors from the A1B scenario) regional projections (see section 2c for details on the data preprocessing).

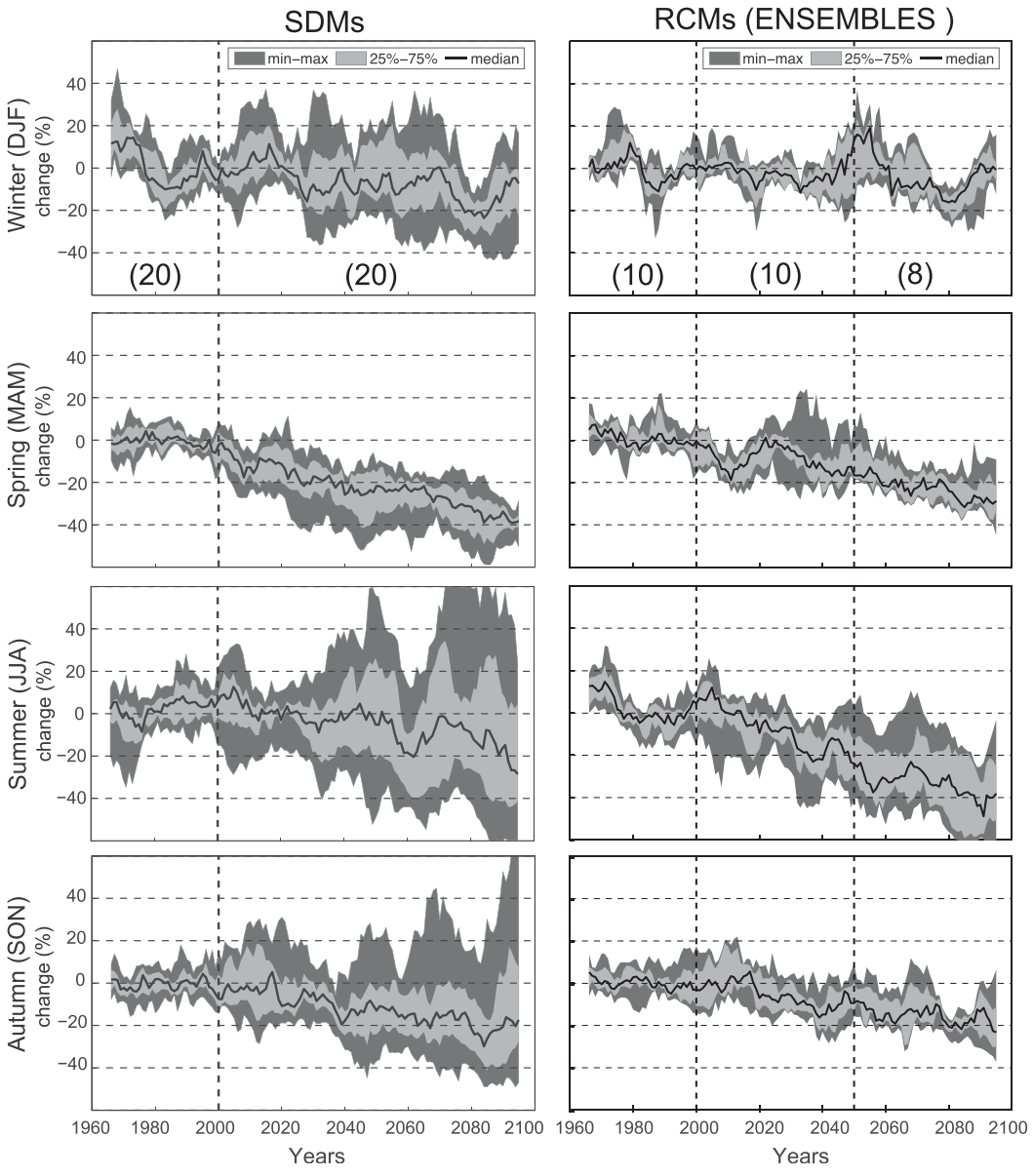

FIG. 10. (top)–(bottom) Spatially averaged seasonal climate change signals (%, w.r.t. the 1961–2000 mean value) for the historical (20C3M scenario, 1961–2000) and transient (A1B scenario, 2001–2100). An 11-yr moving average is applied to smooth the signal. The solid black lines indicate the median of the ensemble, whereas the light (dark) gray shading represents the interquartile (total) ensemble range. (left) The results obtained with the ensemble of five SDMs and four GCMs (20 members) and (right) the ensemble of RCMs from the ENSEMBLES project (with 10 members until 2050 and 8 until 2100). The numbers between parentheses in (top) show the ensemble size in each of the periods.

Figure 10 (left column) shows the projected changes for the spatially averaged seasonal precipitation (for different seasons in the rows of Fig. 10) obtained with the SDM ensemble. The changes are represented as seasonal relative anomalies (in percent) with respect to the corresponding mean value of the historical period 1961–2000. The solid black lines represent the ensemble median and the light (dark) gray shading the interquartile (total) range of the ensemble spread (in this study ''spread'' and ''uncertainty'' have the same meaning). Figure 10 shows a general decrease of the annual precipitation projected along the twenty-first century, with largest decrease magnitude during spring (around −40% at the end of the century, according to the ensemble median) followed by autumn and summer (around −20%, with a larger spread); the smaller signal is obtained for

winter. Overall, these results are in agreement with Giorgi and Piero (2008).

Figure 10 (right) shows the corresponding projected changes from the ensemble of RCMs from the ENSEMBLES project (see section 2d). In general, the trends and the mean signal are similar to the SDM case (Fig. 10, left), with the exception of summer when the RCMs project drier conditions; this could be probably due to the overestimation of summer temperature projections by RCMs as described by Boberg and Christensen (2012). However, the spread (uncertainty) of the statistical downscaling approach is higher than the dynamical downscaling one, with the exception of spring, when both ensembles exhibit quite a similar spread. This difference could be partly attributed to the fact that the size of the SDM ensemble is twice the size of the RCM ensemble. However, since both ensembles are based on a similar set of GCMs, a proper analysis of the relative contribution of the global (GCM) and regional (SDM) model components to the total ensemble spread is required for a comprehensive discussion of this problem.

8. Global and regional model uncertainty components

The contribution of the global and regional model components to the spread (uncertainty) of the climate projections is assessed using a simple analysis of variance approach previously applied to the ensemble of RCMs from the ENSEMBLES project (Déqué et al. 2012). Following the notation in Déqué et al. (2012), let i be the index of SDM ($i = 1, \ldots, 5$), j the index of GCM ($j = 1, \ldots, 4$), and X_{ij} is the response (e.g., winter precipitation change in the Mediterranean region for the 2071–2100 period). Here, the total variance, defined by V here, can be decomposed as

$$V = S + G + SG, \tag{2}$$

where

$$S = \frac{1}{5}\sum_{i=1}^{5}(X_{i\cdot} - X_{\cdot\cdot})^2 \quad \text{and} \quad G = \frac{1}{4}\sum_{j=1}^{4}(X_{\cdot j} - X_{\cdot\cdot})^2 \tag{3}$$

are the terms resulting from SDM alone, and to GCM alone, respectively, and

$$SG = \frac{1}{4}\sum_{j=1}^{4}\frac{1}{5}\sum_{i=1}^{5}(X_{ij} - X_{i\cdot} - X_{\cdot j} + X_{\cdot\cdot})^2 \tag{4}$$

is the interaction term of SDM with GCM. Note that in the above expressions the dot represents the average with respect to the index it replaces. The main advantage for the present study over the original study for RCMs is that all pairs (GCM × SDM combinations) required for (2) to hold are available in this case. Therefore, there is no need to ''fill'' the missing coupling cells in order to account for the unbalanced experimental design when analyzing the variance components.

Figure 11 shows a graphical representation of the magnitudes of the different terms contributing to G and S in (3) for four consecutive time slices (1961–2000, 2011–40, 2041–70, and 2071–2100). For instance, G would correspond to the variance of the mean SDM results (white dots) for the four GCMs shown in the center column of Fig. 11. Figure 11 indicates that the inter-GCM variability is clearly larger than the inter-SDM one in winter and spring, whereas in summer and especially in autumn the results are more similar. A quantitative assessment of this is given in Fig. 12, which shows the fraction of variance (%) explained in the Atlantic and Mediterranean regions by the GCM and the SDM model components, as well as the interaction term (cross-variance), according to (2)–(4). Figure 12 shows that the main contributor to the spread is the choice of the GCM, except for autumn precipitation in the Atlantic region and autumn and summer in the Mediterranean region, where the choice of the SDM dominates the uncertainty during the second half of the twenty-first century. Note that the large spread resulting from the summer and autumn results is largely due to the different projections produced by the two families of techniques used in this study—regression (M3a, M3c, and M4b) and analogs or weather types (M1a and M2c)—which can even disagree in the sign of the (mean) projection (e.g., in summer). This highlights the importance of considering ensembles of different techniques in order to properly sample the uncertainty obtained from SDM projections. Moreover, this also stresses that further research is needed in order to assess the extrapolation capabilities of these techniques. Note that we obtained no indication in this paper that these techniques could be unsuitable for climate change applications. These findings are in agreement with the overall results for Europe from the ENSEMBLES RCMs, but not with the particular results for the Iberian Peninsula (Déqué et al. 2012); only winter and summer seasons were analyzed in that work. The most noticeable difference is the magnitude of the interaction terms, which are larger in the present study. Note that this could be due to the lack of most of the pairs in the GCM–RCM coupling matrix, which could yield an underestimation of the

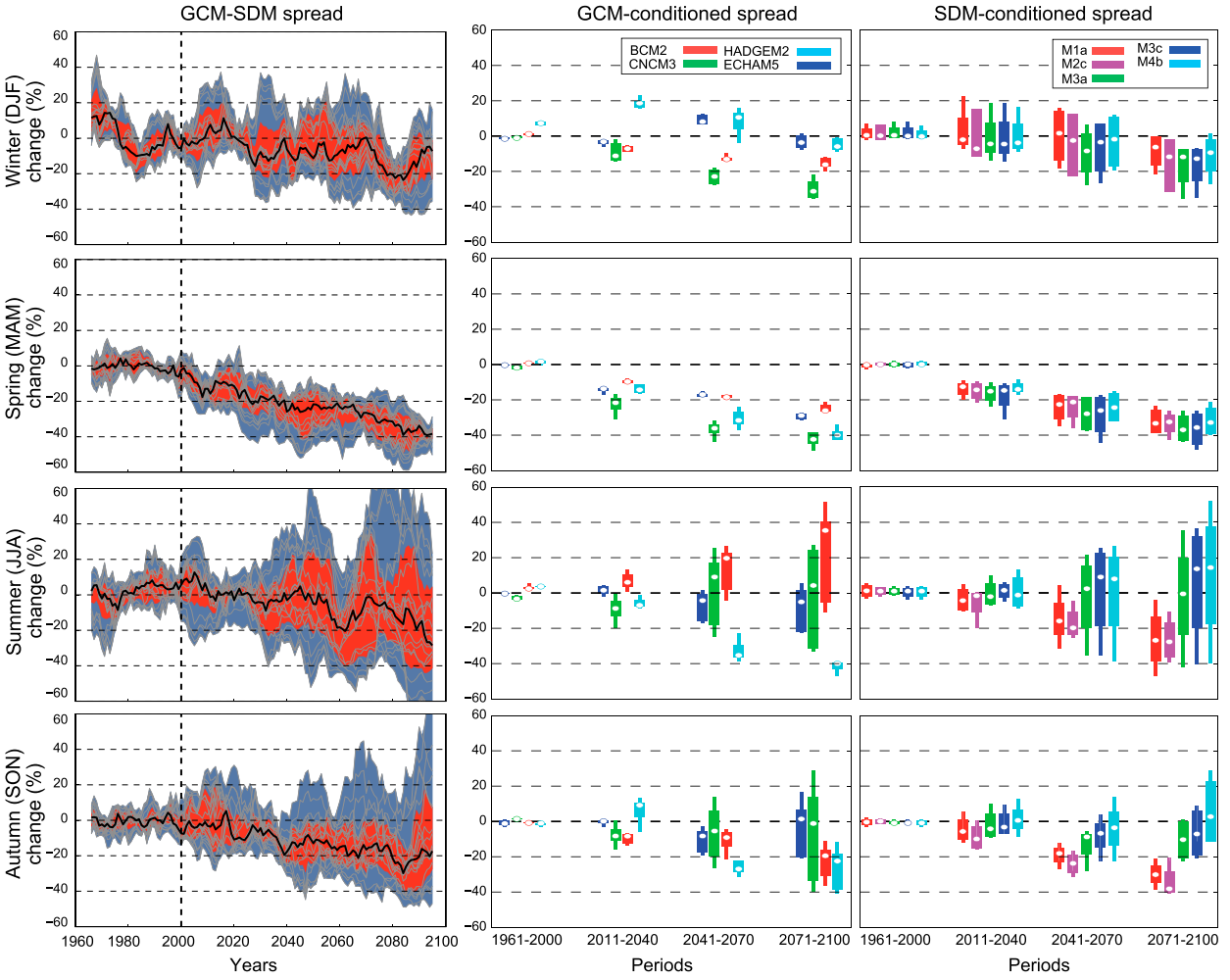

FIG. 11. (left) As in Fig. 10, but including the individual results from the 20 downscaled members (light gray curves). The results for four consecutive time slices (1961–2000, 2011–40, 2041–70, and 2071–2100) conditioned (center) to the GCMs (each box plot represents the variability of the corresponding five SDMs) and (right) to the SDMs (each box plot represents four GCMs), respectively.

interaction terms, even when sophisticated filling methods are used.

9. Conclusions

In the present paper, the performance of state-of-the-art techniques commonly used for statistical downscaling of daily precipitation was assessed, with special focus on their suitability for extrapolating anomalously dry conditions. With this aim, several analog, weather typing, and generalized linear models were intercompared over Spain for the period 1961–2000, following the same structure and methodology introduced in the companion paper of Gutiérrez et al. (2013)—that is, the first part of this work—which performs a similar analysis for the case of temperature.

First, an exhaustive screening of predictor datasets and geographical domains was carried out by considering two illustrative methods. On the one hand, the results (more dependent on the predictors than on the domain considered) point out the necessity of including midtropospheric humidity (in particular Q850) among the predictors, since it yields the best correlations and improves the bias. On the other hand, and in contrast to other previous studies (Reichert et al. 1999), results show that midtropospheric vertical wind velocity (W850) is not an adequate predictor since it leads to poor correlation and serious problems in terms of distributional similarity. The optimum predictor dataset found includes SLP, T850, and Q850, in accordance with the results obtained by Timbal et al. (2003) for western France. Furthermore, the best results are obtained when it is applied over a relatively small domain that covers the area of study. Second, the optimum predictor–domain configuration was used to assess the performance, in terms of accuracy,

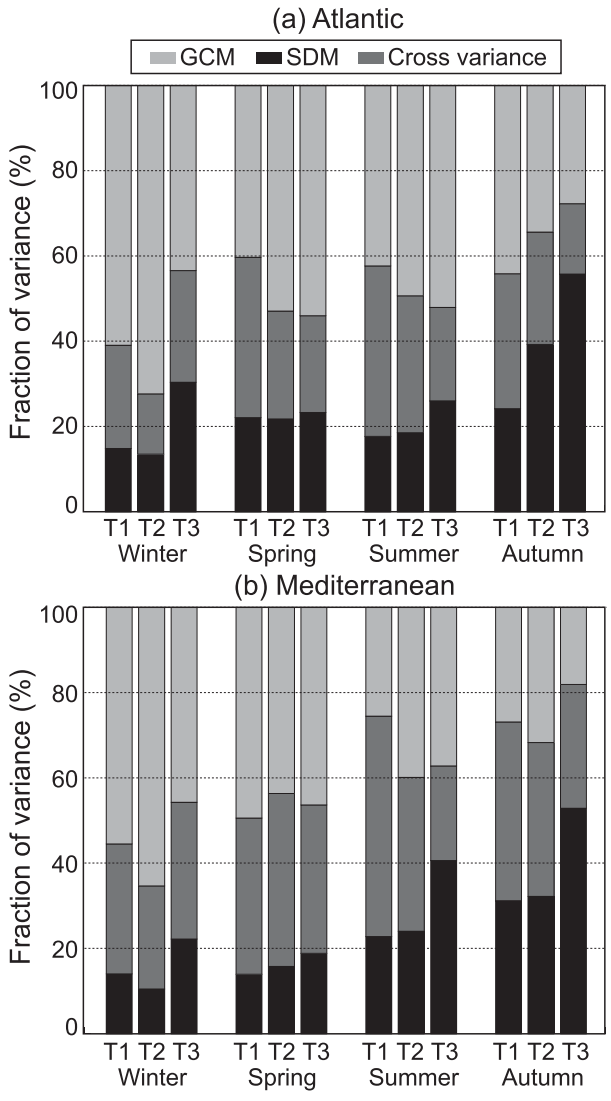

FIG. 12. Fraction of variance (%) explained by the GCM (light gray, upper part of each rectangle) the SDM (black, bottom part), and their interaction (cross-variance, dark gray) terms for the periods 2011–40 (T1), 2041–70 (T2), and 2071–2100 (T3) for the (a) Atlantic and (b) Mediterranean regions.

distributional similarity, and extrapolation of anomalously dry conditions, of all the SDMs considered. Results show important differences among the performance of the different techniques, which are related to various factors, including their stochastic and nonstochastic nature, the spatial character of the predictors considered, etc. Moreover, it was shown that the extrapolation capability for anomalously dry conditions of the different methods is highly dependent on the predictor dataset considered; the same result was also found in Gutiérrez et al. (2013) for temperature. However, in contrast to the case of temperatures, for a given predictor the results of the test for robustness are very similar for the different families of SDMs, and therefore the test fails to provide any indication on the variability of the future precipitation projections for the different SDMs. Therefore, the test can be considered a necessary condition for extrapolation capability, but not a sufficient one for the robust application to future projections.

Deficient methods were discarded and the resulting ensemble of five suitable SDMs was applied to obtain future climate projections from four GCMs from the ENSEMBLES project, obtaining a general decrease of the precipitation projected along the twenty-first century, in particular, with the largest decrease magnitude during spring (around −40% at the end of the century, according to the ensemble median) followed by autumn and summer (around −20%, with a larger spread), and finally by winter. A comparison with the corresponding projected changes from the ensemble of RCMs from the ENSEMBLES project revealed similar trends and mean signals, with the exception of summer, for which the RCMs project drier conditions because of model deficiencies (Boberg and Christensen 2012). However, the spread (uncertainty) of the statistical downscaling approach is higher than the dynamical downscaling one, with the exception of spring, when both ensembles exhibit quite a similar spread.

A quantitative assessment of the GCM and the SDM contribution to the total uncertainty is conducted, with the result that the GCM is the main contributor in most of the cases, except for autumn precipitation in the Atlantic region and autumn and summer in the Mediterranean region, when the SDMs dominate the uncertainty during the second half of the twenty-first century, which corroborates the results from Hertig and Jacobeit (2008) (the uncertainty range arising from the use of different SDMs can even be larger than the one resulting from the application of distinct GCM runs). These findings are in agreement with the overall results for Europe from the ENSEMBLES RCMs, but not with the particular results for the Iberian Peninsula (Déqué et al. 2012). The largest discrepancy is the magnitude of the interaction terms, which are much larger in the present study. This could be due to the lack of most of the pairs in the GCM–RCM coupling matrix, which could yield to an underestimation of the interaction terms, even when sophisticated filling methods are used.

Acknowledgments. This work has been funded by the strategic action for energy and climate change by the Spanish R&D 2008–2011 program "Programa coordinado para la generación de escenarios regionalizados de cambio climático: Regionalización Estadística (esTcena)," code 200800050084078, and the project CGL2015-66583-R

(MINECO/FEDER). The RCM simulations used in this study were obtained from the European Union–funded FP6 Integrated Project ENSEMBLES (Contract 505539). The authors are grateful to the two anonymous reviewers.

REFERENCES

Abaurrea, J., and J. Asín, 2005: Forecasting local daily precipitation patterns in a climate change scenario. *Climate Res.*, **28**, 183–197, doi:10.3354/cr028183.

Beersma, J. J., and T. A. Buishand, 2003: Multi-site simulation of daily precipitation and temperature conditional on the atmospheric circulation. *Climate Res.*, **25**, 121–133, doi:10.3354/cr025121.

Benestad, R. E., 2010: Downscaling precipitation extremes. *Theor. Appl. Climatol.*, **100**, 1–21, doi:10.1007/s00704-009-0158-1.

Boberg, F., and J. H. Christensen, 2012: Overestimation of Mediterranean summer temperature projections due to model deficiencies. *Nat. Climate Change*, **2**, 433–436, doi:10.1038/nclimate1454.

Brands, S., S. Herrera, D. San-Martin, and J. M. Gutiérrez, 2011a: Validation of the ENSEMBLES global climate models over southwestern Europe using probability density functions, from a downscaling perspective. *Climate Res.*, **48**, 145–161, doi:10.3354/cr00995.

——, J. J. Taboada, A. S. Cofiño, T. Sauter, and C. Schneider, 2011b: Statistical downscaling of daily temperatures in the NW Iberian Peninsula from global climate models: Validation and future scenarios. *Climate Res.*, **48**, 163–176, doi:10.3354/cr00906.

——, J. M. Gutiérrez, S. Herrera, and A. S. Cofiño, 2012: On the use of reanalysis data for downscaling. *J. Climate*, **25**, 2517–2526, doi:10.1175/JCLI-D-11-00251.1.

——, S. Herrera, J. Fernández, and J. M. Gutiérrez, 2013: How well do CMIP5 Earth system models simulate present climate conditions in Europe and Africa? *Climate Dyn.*, **41**, 803–817, doi:10.1007/s00382-013-1742-8.

Brandsma, T., and T. A. Buishand, 1997: Statistical linkage of daily precipitation in Switzerland to atmospheric circulation and temperature. *J. Hydrol.*, **198**, 98–123, doi:10.1016/S0022-1694(96)03326-4.

Chandler, R. E., and H. S. Wheater, 2002: Analysis of rainfall variability using generalized linear models: A case study from the west of Ireland. *Water Resour. Res.*, **38**, 1192, doi:10.1029/2001WR000906.

Charles, S. P., B. C. Bates, P. H. Whetton, and J. P. Hughes, 1999: Validation of downscaling models for changed climate conditions: Case study of southwestern Australia. *Climate Res.*, **12**, 1–14, doi:10.3354/cr012001.

Cheng, C. S., G. Li, Q. Li, and H. Auld, 2008: Statistical downscaling of hourly and daily climate scenarios for various meteorological variables in south-central Canada. *Theor. Appl. Climatol.*, **91**, 129–147, doi:10.1007/s00704-007-0302-8.

——, ——, ——, and ——, 2011: A synoptic weather-typing approach to project future daily rainfall and extremes at local scale in Ontario, Canada. *J. Climate*, **24**, 3667–3685, doi:10.1175/2011JCLI3764.1.

Coe, R., and R. D. Stern, 1982: Fitting models to daily rainfall data. *J. Appl. Meteor.*, **21**, 1024–1031, doi:10.1175/1520-0450(1982)021<1024:FMTDRD>2.0.CO;2.

Cubasch, U., H. von Storch, J. Waszkewitz, and E. Zorita, 1996: Estimates of climate change in southern Europe derived from dynamical climate model output. *Climate Res.*, **7**, 129–149, doi:10.3354/cr007129.

Déqué, M., and Coauthors, 2007: An intercomparison of regional climate simulations for Europe: Assessing uncertainties in model projections. *Climatic Change*, **81**, 53–70, doi:10.1007/s10584-006-9228-x.

——, S. Somot, E. Sanchez-Gomez, C. M. Goodess, D. Jacob, G. Lenderink, and O. B. Christensen, 2012: The spread amongst ENSEMBLES regional scenarios: Regional climate models, driving general circulation models and interannual variability. *Climate Dyn.*, **38**, 951–964, doi:10.1007/s00382-011-1053-x.

Dibike, Y. B., and P. Coulibaly, 2005: Hydrologic impact of climate change in the Saguenay watershed: Comparison of downscaling methods and hydrologic models. *J. Hydrol.*, **307**, 145–163, doi:10.1016/j.jhydrol.2004.10.012.

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

Enke, W., and A. Spegat, 1997: Downscaling climate model outputs into local and regional weather elements by classification and regression. *Climate Res.*, **8**, 195–207, doi:10.3354/cr008195.

Fealy, R., and J. Sweeney, 2007: Statistical downscaling of precipitation for a selection of sites in Ireland employing a generalised linear modelling approach. *Int. J. Climatol.*, **27**, 2083–2094, doi:10.1002/joc.1506.

Giorgi, F., and L. Piero, 2008: Climate change projections for the Mediterranean region. *Global Planet. Change*, **63**, 90–104, doi:10.1016/j.gloplacha.2007.09.005.

Goodess, C. M., and J. P. Palutikof, 1998: Development of daily rainfall scenarios for southeast Spain using a circulation-type approach to downscaling. *Int. J. Climatol.*, **18**, 1051–1083, doi:10.1002/(SICI)1097-0088(199808)18:10<1051::AID-JOC304>3.0.CO;2-1.

Gutiérrez, J. M., A. S. Cofiño, R. Cano, and M. A. Rodríguez, 2004: Clustering methods for statistical downscaling in short-range weather forecasts. *Mon. Wea. Rev.*, **132**, 2169–2183, doi:10.1175/1520-0493(2004)132<2169:CMFSDI>2.0.CO;2.

——, D. San-Martín, S. Brands, R. Manzanas, and S. Herrera, 2013: Reassessing statistical downscaling techniques for their robust application under climate change conditions. *J. Climate*, **26**, 171–188, doi:10.1175/JCLI-D-11-00687.1.

Hanel, M., and T. A. Buishand, 2015: Assessment of the sources of variation in changes of precipitation characteristics over the Rhine basin using a linear mixed-effects model. *J. Climate*, **28**, 6903–6919, doi:10.1175/JCLI-D-14-00775.1.

Hanssen-Bauer, I., C. Achberger, R. E. Benestad, D. Chen, and E. J. Forland, 2005: Statistical downscaling of climate scenarios over Scandinavia. *Climate Res.*, **29**, 255–268, doi:10.3354/cr029255.

Haylock, M. R., N. Hofstra, A. M. G. Klein Tank, E. J. Klok, P. D. Jones, and M. New, 2008: A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006. *J. Geophys. Res.*, **113**, D20119, doi:10.1029/2008JD010201.

Herrera, S., L. Fita, J. Fernández, and J. M. Gutiérrez, 2010: Evaluation of the mean and extreme precipitation regimes from the ENSEMBLES regional climate multimodel simulations over Spain. *J. Geophys. Res.*, **115**, D21117, doi:10.1029/2010JD013936.

——, J. M. Gutiérrez, R. Ancell, M. R. Pons, M. D. Frías, and J. Fernández, 2012: Development and analysis of a 50-year high-resolution daily gridded precipitation dataset over Spain (Spain02). *Int. J. Climatol.*, **32**, 74–85, doi:10.1002/joc.2256.

Hertig, E., and J. Jacobeit, 2008: Assessments of Mediterranean precipitation changes for the 21st century using statistical downscaling techniques. *Int. J. Climatol.*, **28**, 1025–1045, doi:10.1002/joc.1597.

——, and ——, 2013: A novel approach to statistical downscaling considering nonstationarities: Application to daily precipitation in the Mediterranean area. *J. Geophys. Res. Atmos.*, **118**, 520–533, doi:10.1002/jgrd.50112.

——, S. Seubert, A. Paxian, G. Vogt, H. Paeth, and J. Jacobeit, 2013: Changes of total versus extreme precipitation and dry periods until the end of the twenty-first century: Statistical assessments for the Mediterranean area. *Theor. Appl. Climatol.*, **111**, 1–20, doi:10.1007/s00704-012-0639-5.

Hingray, B., and M. Said, 2014: Partitioning internal variability and model uncertainty components in a multimember multimodel ensemble of climate projections. *J. Climate*, **27**, 6779–6798, doi:10.1175/JCLI-D-13-00629.1.

Imbert, A., and R. E. Benestad, 2005: An improvement of analog model strategy for more reliable local climate change scenarios. *Theor. Appl. Climatol.*, **82**, 245–255, doi:10.1007/s00704-005-0133-4.

Jacobeit, J., 2010: Classifications in climate research. *Phys. Chem. Earth*, **35**, 411–421, doi:10.1016/j.pce.2009.11.010.

Jolliffe, I. T., and D. B. Stephenson, 2003: *Forecast Verification: A Practitioner's Guide in Atmospheric Science.* 2nd ed. Wiley and Sons, 292 pp.

Llasat, M. C., 2009: High magnitude storms and floods. *The Physical Geography of the Mediterranean*, J. Woodward, Ed., Oxford University Press, 513–540.

Lorenz, E. N., 1969: Atmospheric predictability as revealed by naturally occurring analogues. *J. Atmos. Sci.*, **26**, 636–646, doi:10.1175/1520-0469(1969)26<636:APARBN>2.0.CO;2.

Maraun, D., 2012: Nonstationarities of regional climate model biases in European seasonal mean temperature and precipitation sums. *Geophys. Res. Lett.*, **39**, L06706, doi:10.1029/2012GL051210.

——, and Coauthors, 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.

Moron, V., A. W. Robertson, M. N. Ward, and O. Ndiaye, 2008: Weather types and rainfall over Senegal. Part II: Downscaling of GCM simulations. *J. Climate*, **21**, 288–307, doi:10.1175/2007JCLI1624.1.

Nelder, J. A., and R. W. M. Wedderburn, 1972: Generalized linear models. *J. Roy. Stat. Soc.*, **135A**, 370–384, doi:10.2307/2344614.

Philipp, A., and Coauthors, 2010: Cost733cat—A database of weather and circulation type classifications. *Phys. Chem. Earth*, **35**, 360–373, doi:10.1016/j.pce.2009.12.010.

Preisendorfer, R., 1988: *Principal Component Analysis in Meteorology and Oceanography*. Elsevier, 425 pp.

Pryor, S. C., J. T. Schoof, and R. J. Barthelmie, 2005: Climate change impacts on wind speeds and wind energy density in northern Europe: Empirical downscaling of multiple AOGCMs. *Climate Res.*, **29**, 183–198, doi:10.3354/cr029183.

Reichert, B. K., L. Bengtsson, and O. Akesson, 1999: A statistical modeling approach for the simulation of local paleoclimatic proxy records using general circulation model output. *J. Geophys. Res.*, **104**, 19 071–19 083, doi:10.1029/1999JD900264.

Ruosteenoja, K., and P. Räisänen, 2013: Seasonal changes in solar radiation and relative humidity in Europe in response to global warming. *J. Climate*, **26**, 2467–2481, doi:10.1175/JCLI-D-12-00007.1.

Sauter, T., and V. Venema, 2011: Natural three-dimensional predictor domains for statistical precipitation downscaling. *J. Climate*, **24**, 6132–6145, doi:10.1175/2011JCLI4155.1.

Teutschbein, C., F. Wetterhall, and J. Seibert, 2011: Evaluation of different downscaling techniques for hydrological climate-change impact studies at the catchment scale. *Climate Dyn.*, **37**, 2087–2105, doi:10.1007/s00382-010-0979-8.

Timbal, B., and B. J. McAvaney, 2001: An analogue-based method to downscale surface air temperature: Application for Australia. *Climate Dyn.*, **17**, 947–963, doi:10.1007/s003820100156.

——, and D. A. Jones, 2008: Future projections of winter rainfall in southeast Australia using a statistical downscaling technique. *Climatic Change*, **86**, 165–187, doi:10.1007/s10584-007-9279-7.

——, A. Dufour, and B. McAvaney, 2003: An estimate of future climate change for western France using a statistical downscaling technique. *Climate Dyn.*, **20**, 807–823, doi:10.1007/s00382-002-0298-9.

Trigo, R. M., and J. P. Palutikof, 2001: Precipitation scenarios over Iberia: A comparison between direct GCM output and different downscaling techniques. *J. Climate*, **14**, 4422–4446, doi:10.1175/1520-0442(2001)014<4422:PSOIAC>2.0.CO;2.

Turco, M., M. Quintana-Seguí, C. Llasat, S. Herrera, and J. M. Gutiérrez, 2011: Testing MOS precipitation downscaling for ENSEMBLES regional climate models over Spain. *J. Geophys. Res.*, **116**, D18109, doi:10.1029/2011JD016166.

——, A. Sanna, S. Herrera, M. Llasat, and J. Gutiérrez, 2013: Large biases and inconsistent climate change signals in ENSEMBLES regional projections. *Climatic Change*, **120**, 859–869, doi:10.1007/s10584-013-0844-y.

——, ——, ——, M. C. Llasat, and J. M. Gutiérrez, 2015: Evaluation of the ENSEMBLES transient RCM simulations over Spain: Present climate performance and future projections. *Climate Change and Engineering Geology*, Vol. 1, Engineering Geology for Society and Territory, G. Lollino et al., Eds., Springer, 199–203.

Uppala, S. M., 2005: The ERA-40 Re-Analysis. *Quart. J. Roy. Meteor. Soc.*, **131**, 2961–3012, doi:10.1256/qj.04.176.

van der Linden, P., and J. F. B. Mitchell, Eds., 2009: ENSEMBLES: Climate Change and its Impacts: Summary of research and results from the ENSEMBLES project. Met Office Hadley Centre, 160 pp.

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

Wetterhall, F., S. Halldin, and C. Y. Xu, 2005: Statistical precipitation downscaling in central Sweden with the analogue method. *J. Hydrol.*, **306**, 174–190, doi:10.1016/j.jhydrol.2004.09.008.

Widmann, M., C. S. Bretherton, and E. P. Salathé, 2003: Statistical precipitation downscaling over the northwestern United States using numerically simulated precipitation as a predictor. *J. Climate*, **16**, 799–816, doi:10.1175/1520-0442(2003)016<0799:SPDOTN>2.0.CO;2.

Winkler, J. A., G. S. Guentchev, M. Liszewska, Perdinan, and P.-N. Tan, 2011: Climate scenario development and applications for local/regional climate change impact assessments: An overview for the non-climate scientist. *Geogr. Compass*, **5**, 301–328, doi:10.1111/j.1749-8198.2011.00426.x.

Yang, C., R. E. Chandler, V. S. Isham, and H. S. Wheater, 2005: Spatial-temporal rainfall simulation using generalized linear models. *Water Resour. Res.*, **41**, W11415, doi:10.1029/2004WR003739.

Zorita, E., and H. von Storch, 1999: The analog method as a simple statistical downscaling technique: Comparison with more complicated methods. *J. Climate*, **12**, 2474–2489, doi:10.1175/1520-0442(1999)012<2474:TAMAAS>2.0.CO;2.