

Reassessing Statistical Downscaling Techniques for Their Robust Application under Climate Change Conditions

J. M. GUTIÉRREZ, D. SAN-MARTÍN, S. BRANDS, R. MANZANAS, AND S. HERRERA

Instituto de Física de Cantabria (UNICAN-CSIC), Santander, Spain

(Manuscript received 23 November 2011, in final form 30 April 2012)

ABSTRACT

The performance of statistical downscaling (SD) techniques is critically reassessed with respect to their robust applicability in climate change studies. To this end, in addition to standard accuracy measures and distributional similarity scores, the authors estimate the robustness of the methods under warming climate conditions working with anomalous warm historical periods. This validation framework is applied to intercompare the performances of 12 different SD methods (from the analog, weather typing, and regression families) for downscaling minimum and maximum temperatures in Spain. First, a calibration of these methods is performed in terms of both geographical domains and predictor sets; the results are highly dependent on the latter, with optimum predictor sets including near-surface temperature data (in particular 2-m temperature), which appropriately discriminate cold episodes related to temperature inversion in the lower troposphere.

Although regression methods perform best in terms of correlation, analog and weather generator approaches are more appropriate for reproducing the observed distributions, especially in case of wintertime minimum temperature. However, the latter two families significantly underestimate the temperature anomalies of the warm periods considered in this work. This underestimation is found to be critical when considering the warming signal in the late twenty-first century as given by a global climate model [the ECHAM5–Max Planck Institute (MPI) model]. In this case, the different downscaling methods provide warming values with differences in the range of 1°C, in agreement with the robustness significance values. Therefore, the proposed test is a promising technique for detecting lack of robustness in statistical downscaling methods applied in climate change studies.

1. Introduction

Statistical downscaling (SD) methods are nowadays routinely applied for generating local climate change projections from the coarse-resolution outputs of global climate models (GCMs) (Timbal et al. 2003; Haylock et al. 2006; Hewitson and Crane 2006; Timbal and Jones 2008; Benestad 2010; Brands et al. 2011b; Gutzler and Robbins 2011). These methods are based on empirical relationships linking large-scale atmospheric variables (predictors) with some local-scale variables of interest (predictands). Different SD techniques have been proposed to infer these relationships from data under the so-called *perfect prog* approach (Maraun et al. 2010). In this case, reanalysis outputs for a representative period of the past (typically 30 yr) are used as predictors while

simultaneous historical observations at the local scale are used as predictands for model training. Once the optimal model configuration has been found using these (quasi) observed data (Brands et al. 2012), the model is applied to the output of different GCM scenario runs to obtain future projections in different climate change scenarios.

This *perfect prog* downscaling approach is affected by some well-known limiting factors, which are especially relevant when applying it to GCM scenario runs. Some of these factors are usually taken into account when generating climate change projections. For instance, the reanalysis variables selected as large-scale predictors should be well simulated by GCMs, should capture the climate change “signal,” and should have a significant and physically interpretable association with the predictand (Wilby et al. 2004).

However, there are other important limiting factors that have been rarely assessed in earlier studies. First, for the particular choice of predictors made, the statistical downscaling method should provide a stable/stationary statistical relationship between the predictors and the

Corresponding author address: José M. Gutiérrez, Instituto de Física de Cantabria (CSIC-UC), Avenida de los Castros, Santander 39005, Spain.
E-mail: gutierjm@unican.es

predictand in order to remain valid under climate change conditions. This is usually referred to as the *robustness* or *stationarity assumption*, and only a few studies have focused on this problem, using either global or regional climate model outputs as pseudo-observations (Frías et al. 2006; Vrac et al. 2007) or analyzing the stationarity of empirical relationships (Schmith 2008). Second, the downscaled and observed time series should have similar climatological properties (i.e., similar distributions) in order to avoid any form of post hoc correction such as bias correction or more advanced postprocessing techniques such as quantile mapping (Déqué 2007), which if applied must additionally be assumed to be stationary in time (Hagemann et al. 2011). Finally, since future seasonal climates might not exactly correspond to the present ones, the calibration process should not be applied separately for each season—as is common in most SD studies (Maraun et al. 2010)—but for the training period as a whole (Imbert and Benestad 2005; Teutschbein et al. 2011). This requires controlling the seasonal variability of the results, which may be difficult to achieve, since the most informative predictor combination may potentially vary from season to season (Wetterhall et al. 2005, 2007). If some of the above factors are not fulfilled, the results of any SD application should be interpreted with caution, since the choice of the predictors and/or the downscaling methodology can have a large influence on the local climate change scenarios.

In this paper we provide a comprehensive validation framework to test the suitability of common *perfect prog* SD techniques for their applicability in climate change studies, taking into account the abovementioned limitations. The final aim of this work is to find robust downscaling schemes that can be applied under climate change conditions without the necessity of any form of post hoc correction. To this aim, we combine standard accuracy validation scores with additional scores obtained by statistically testing 1) the distributional similarity of the downscaled and observed series and 2) the robustness of the bias to warmer climatic conditions. In the former case, we consider the significance level of the two-sample Kolmogorov–Smirnov test for the null hypothesis of equal downscaled and observed distributions. In the latter case, we compare the bias of the methods in an historical warm period (defined by the eight warmest years in the analysis period) with that obtained in “normal” conditions (characterized by the 8-yr random samples given from a fivefold cross-validation approach).

As an illustrative example, we consider minimum and maximum temperatures in Spain using the publicly available daily gridded dataset Spain02 (Herrera et al. 2012) as predictands. It covers peninsular Spain and the Balearic Islands at a resolution of 0.2° and has been found to

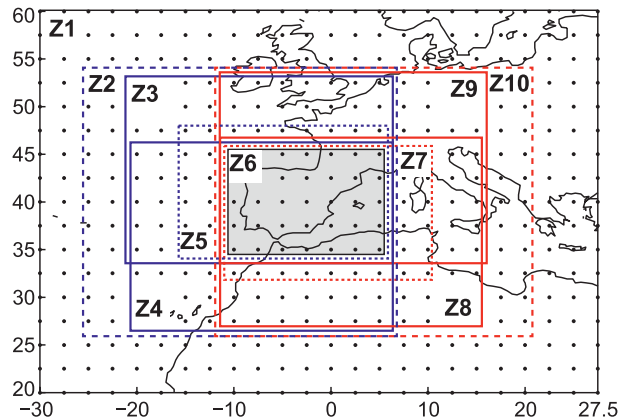


FIG. 1. Different domains used in this study, numbered from east to west, and decreasing in size toward the center: 1) esTcena, 2) W, 3) NW, 4) SW, 5) NWsmall, 6) Iberia, 7) SEsmall, 8) SE, 9) NE, and 10) E.

be of particular interest for impact studies in this region. To obtain general conclusions, we apply an ensemble of the most commonly used statistical downscaling approaches (analogs, weather typing, regression, regression conditioned on weather types) to the most commonly used predictor variables considering both local and spatial predictors, given by the values at the nearest grid box and by the principal components (PCs), respectively. Special focus is given to compare the results of using either free-tropospheric or near-surface temperature as a predictor for the downscaling methods, since there has been some scientific debate on which height level to prefer [see Hanssen-Bauer et al. (2005) for more details].

This work is structured as follows. In section 2, the geographical domains and applied data are presented. Section 3 describes the different statistical downscaling methods. The conducted cross-validation approach, as well as the proposed validation scores, is presented in section 4. Section 5 refers to the screening of the different geographical domains and predictors by using two reference SD methods (analogs and regression using PCs). On the basis of the optimal configuration of domain and predictors, the performances of all SD methods are intercompared in section 6. Finally, some conclusions are given in section 7.

2. Geographical zones and data

The target region of this work is the Iberian Peninsula. Therefore, we defined different predictor areas, Z_1, \dots, Z_{10} , with different sizes, as shown in Fig. 1; note that hereafter Z_i stands for a specific zone. Over this region we considered a number of atmospheric variables (see Table 1) typically used as predictors in temperature downscaling studies (Benestad 2002; Huth 2002; Hanssen-Bauer

TABLE 1. Predictor variables used in this work. Note that Tx and Tn have been only considered for benchmarking purposes only, as their GCM performance for the region of study is poor; see the text for more details.

Code	Name	Levels	Time	Unit
<i>Z</i>	Geopotential	850, 700, 500, 300	0000 UTC	m ² s ⁻²
<i>T</i>	Temperature	850, 700, 500, 300	0000 UTC	K
<i>Q</i>	Specific humidity	850, 700, 500, 300	0000 UTC	kg kg ⁻¹
<i>U</i>	<i>U</i> -wind component	850, 700, 500, 300	0000 UTC	m s ⁻¹
<i>V</i>	<i>V</i> -wind component	850, 700, 500, 300	0000 UTC	m s ⁻¹
SLP	Mean sea level pressure	Mean sea level	Daily mean	Pa
T2m	Daily mean temperature	Model surface	Daily mean	Pa
Tx	Maximum temperature	Model surface	Instantaneous	K
Tn	Minimum temperature	Model surface	Instantaneous	K

et al. 2005; Huth et al. 2008). It has been recently shown that these variables—considering anomalies—are suitable predictors for climate change studies, since their distribution is skillfully reproduced by GCMs in the area under study (see Brands et al. 2011a). The only exceptions are maximum and minimum temperatures (denoted Tx and Tn, respectively), since their use as predictors in climate change downscaling studies has shown to be problematic (Palutikof et al. 1997); thus, these variables are only used as predictors in this study for benchmarking purposes. All these variables were downloaded from the publicly available 40-yr European Centre for Medium-Range Weather Forecasts (ECMWF) Re-Analysis (ERA-40) data (Uppala et al. 2005) with 2.5° resolution for the period 1961–2000 and will be used to test the different SD methods in *perfect prog* conditions focusing on validation measures informative for climate change conditions.

We considered the predictor combinations listed in Table 2, including the typical settings used in downscaling studies; for instance, since the climate change signal is much weaker for circulation variables than for temperature and/or absolute humidity—linked to changes in the radiation budget (Wilby et al. 1998, 2004)—we do not consider predictor datasets including only circulation variables [*Z*, sea level pressure (SLP), *U*, and *V*]. Note also that those combinations marked with a letter *d* (P1, P2, P3, P4, and P6) have been tested with two temporal setups: static and dynamic, as suggested by Gutiérrez et al. (2004). The “static” temporal setup only takes into account 0000 UTC values for the instantaneous variables (*Z*, *T*, *Q*, *U*, and *V*) for day *D*, whereas the “dynamic” temporal approach additionally includes the 0000 UTC values for day *D* + 1, thus providing a window covering the observation period. We want to remark that using 1200 UTC values instead of 0000 UTC values for downscaling Tmax did not improve the results (not shown). Note that hereafter *Pi*, or *Pid*, stands for a specific static, or dynamic, predictor configuration, respectively.

For different configurations of the downscaling techniques (see Table 3), we either consider the standardized

anomalies of the ERA-40 data at nearby grid boxes as predictors or, alternatively, use spatial patterns as given by the PCs of the predictor field (Preisendorfer 1988). In this case, the total number of PCs considered is limited to the leading PCs yielding a fraction of explained variance of 95% (note that a maximum of 30 PCs is not exceeded in any case). In the former case, the spatial homogeneity of the downscaled series is expected to be low, since different predictors are used for each target location; however, in the latter case, the predictors are shared by all locations, which should considerably enhance the spatial homogeneity of the results.

The local target variables of interest in this work (predictands) are the daily 2-m maximum (Tmax) and minimum (Tmin) air temperatures from the recently developed publicly available gridded interpolated observations dataset Spain02 (Herrera 2011; Herrera et al. 2012, freely available online at <http://www.meteo.unican.es/datasets/spain02>). The data come on a regular 0.2° grid and cover the complete time period under study (1961–2000). Figures 2a,b and 2c,d show the corresponding means and standard deviations for Tmax and Tmin, respectively, at each grid box of Spain02, as well as the inter- and intra-annual variability of the spatial mean anomalies (Figs. 2e,f). Note that Tx (Tn) hereafter refers to the maximum (minimum) temperatures as predictors,

TABLE 2. Tested predictor combinations, ranked by decreasing complexity; the combinations marked by the letter *d* have been tested with both the static and dynamic temporal setup.

Code	Predictor variables
P1, P1d	SLP, T850, Q850, U500, V500
P2, P2d	SLP, T850, Q850, Z500
P3, P3d	SLP, T850, Q850
P4, P4d	SLP, T850
P5	SLP, T2m
P6, P6d	T850
P7	T2m
P8	Tx
P9	Tn

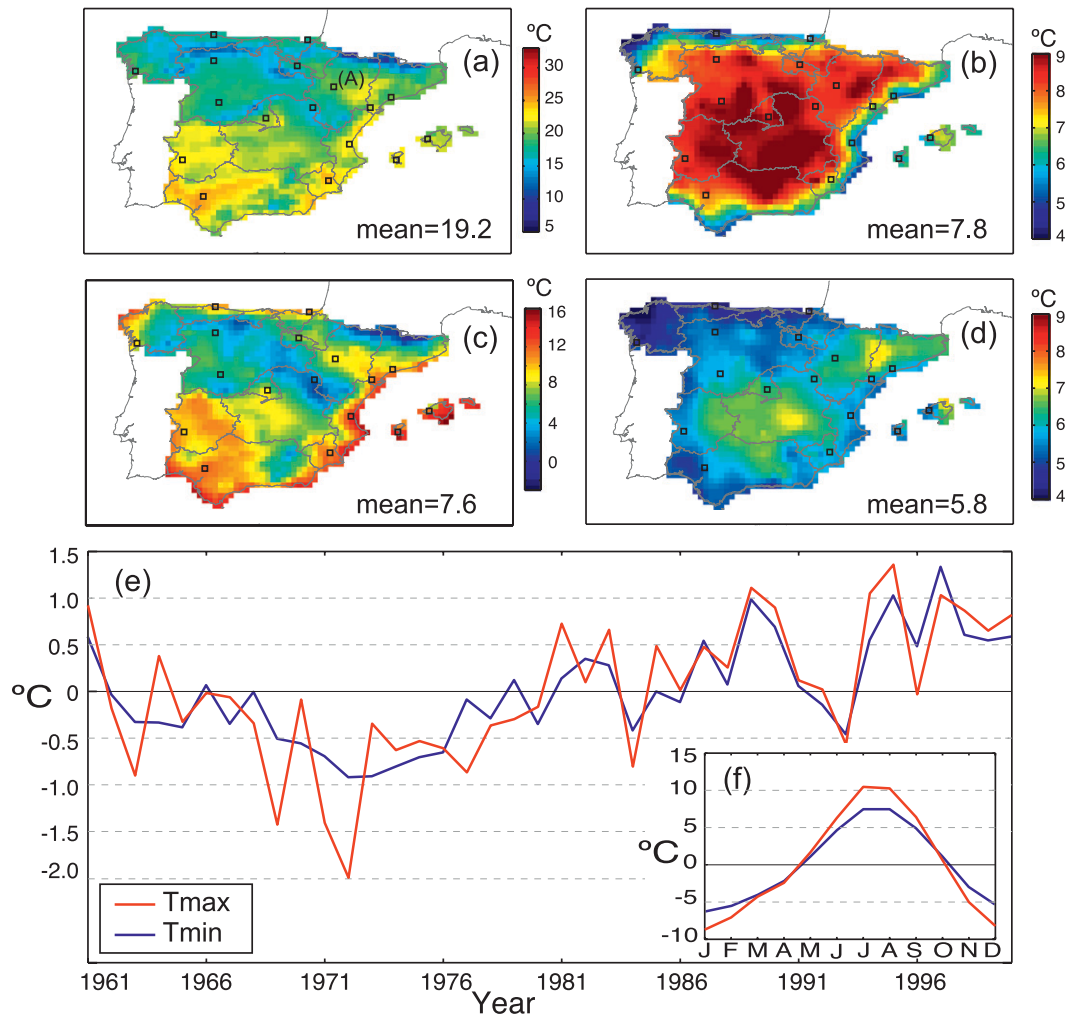


FIG. 2. Daily mean and daily standard deviation of the Spain02 daily dataset for (a),(b) maximum and (c),(d) minimum temperatures for the period 1961–2000. The boxes in these figures show the location of the 17 representative grid points used in the study; the point labeled by (A) in (a) will be referred to for illustrative purposes in section 5. Also shown are the (e) inter- and (f) intra-annual variability of the spatial mean values for these variables; note that in these cases, anomalies with regard to the annual mean value are shown.

whereas Tmax (Tmin) will be used as abbreviation for the predictands, respectively.

Because of the differing spatial extent of the different climatic regions in the area under study, we will consider the 17 grid boxes shown in Fig. 2 for calculating spatial averages, since this will impede that the results are dominated by the larger climatic regions. Note that the time series associated with these grid boxes are very close to those of 17 high-quality observed time series, which are publicly available from the Spanish Meteorological Agency (AEMET; <http://www.aemet.es>) and, thus, the interpolation error of the interpolation/gridding scheme is minimized in this case. This will be important when considering warm anomalous periods in section 4c, with magnitudes around 1°C, where spurious warm spatial

patterns may arise in regions with sparse data due to the interpolation method.

3. Downscaling methods

In this paper we intercompare a number of different statistical downscaling methods, including the most popular ones used for climate change applications. These methods are described in Table 3 and have been classified according to the following categories:

- M1: Analog methods (AM);
- M2: Weather typing methods (WT);
- M3: Multiple linear regression, from PCs, point-wise, or both (LR); and

TABLE 3. Downscaling methods of four different families considered in this work: AM—Analog methods, WT—weather typing, LR—linear regression, and LR-WT—regression conditioned on weather types.

Code	Type	Method and predictor field
M1a	AM	Nearest neighbor (1 analog)
M1b	AM	Mean of 5 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	100 WTs (k means), simulation from Gaussian distribution
M3a	LR	Linear regression with n PCs (95% variance)
M3b	LR	Local predictor values in the nearest grid box
M3c	LR	15 PCs + nearest grid box
M4a	LR-WT	M3c conditioned on 10 WTs (all predictors)
M4b	LR-WT	M3b conditioned on 10 WTs (all predictors)
M4c	LR-WT	M3b (T , Q) conditioned on 10 WTs (SLP only)

- M4: Linear regression conditioned on weather types (LR-WT).

The first group of downscaling schemes (M1a to M1c) includes three different versions of the analog method (AM), which was introduced in the atmospheric sciences by Lorenz (1963, 1969) and compared with other SD techniques by Zorita and von Storch (1999). In this study, the Euclidean distance was used to obtain the analogs from the predictor field (Matulla et al. 2008). The technique labeled M1a is based on the nearest analog, whereas M1b and M1c consider the 5 and 15 nearest analogs, respectively. M1b uses the mean of the corresponding observed values as the target value, whereas M1c randomly selects one of them (Brandsma and Buishand 1998; Beersma and Buishand 2003). These three configurations have been chosen after a sensitivity analysis (with regard to the number of analogs, the applied distance measures) and roughly reflect the different methodological approaches. On the one hand, the optimum configuration for M1b was selected after comparing the results obtained for 5, 10, 15, and 20 analogs, obtaining similar results (correlation), but progressively underestimating the variance. On the other hand, 15 analogs for M1c was shown to yield a reasonable tradeoff between a decreasing correlation and an increasing and thus more realistic ratio of modeled to observed variance (see, e.g., Timbal and McAvaney 2001).

Because of its conceptual simplicity and applicability to any predictand variable, the AM is still widely used as a benchmark method in statistical downscaling applications (Brands et al. 2011a; Pons et al. 2010; Teutschbein et al. 2011; Timbal et al. 2003; Timbal and Jones 2008; Wetterhall et al. 2005). However, its main drawback is

its inability to extrapolate unobserved values and, hence, its tendency to underestimate warming in climate change conditions. A possible correction for this issue has been recently suggested by Benestad (2010), but it has not been considered in this study.

The second group of downscaling methods contains three different classification or weather typing techniques (M2a–M2c) based on the k -means clustering algorithm, which was applied to the atmospheric state vector formed by all the considered predictors standardized at a gridbox level to avoid biased results due to different scales (Gutiérrez et al. 2004). M2a and M2b are modifications of the abovementioned analog method, with the search space being quantized into weather types (WTs). Weather types are first calculated applying the k -means method (obtaining their corresponding “centroids”), and then each day is assigned to the closest WT (closest centroid). This consequently reduces the computational cost and allows for an interpretation of the results in terms of frequencies of the different WTs. A sensitivity study revealed an optimum number around 100 WTs, obtained as the threshold value where both the correlation and variance of the results saturate, allowing us to keep the size of each group large enough to guarantee robust results [see Huth et al. (2010) for a detailed overview of classification techniques in the atmospheric sciences]. M2a estimates the downscaled value as the mean of the observations corresponding to the particular weather type, whereas M2b picks one value at random within those in the corresponding WT. M2c combines the k -means weather typing approach with a Gaussian weather generator in order to avoid using the empirical WT distribution and to partially overcome the analog method’s limitation to extrapolate values unobserved in the past. In the training period, each observed temperature time series is partitioned into 100 subseries corresponding to 100 WTs. The parameters of the Gaussian distribution are then fitted to each of these subseries and are used for randomly generating temperature series conditioned to the corresponding weather type in the independent test periods.

The third group of methods contains three different versions of multiple linear regression (M3a–M3c) (Benestad 2002, 2005; Huth 2002, 2004). On the one hand, PCs are used as predictors—considering those explaining 95% of the variance (with a maximum of 30 PCs)—making up the “global” predictor setup M3a. On the other hand, the standardized values from the nearest grid box are applied, making up the “local” predictor setup M3b. Note that we also tested the performance when considering several neighboring grid boxes, but similar results were obtained. Finally, we combine both the global (15 PCs) and local (nearest gridbox values)

predictors, obtaining the mixed predictor configuration M3c. The comparison of these three setups will allow us to assess the performance of spatial versus pointwise predictors. Note that this family of methods has extrapolation capabilities and hence may be more robust in climate change conditions.

The fourth group of methods (M4a–M4c) is a combination of weather typing (M2) and multiple linear regression (M3). As in M2, a k -means clustering is first applied to determine a number of WTs. As a result of a previously applied sensitivity study, 10 WTs were considered for this family of methods. Note that although a higher number of WTs was considered for M2 methods in order to increase accuracy, for the M4 family accuracy is provided by the regression step rather than by the weather typing step and, thus, these methods can work with fewer WTs—actually, they should do so in order to prevent working with too small a sample when adjusting the regression model. In the first two cases (M4a and M4b), the clustering is performed upon all predictor variables, while in the third case (M4c), it is performed on SLP (representing circulation) only. Afterwards, a linear regression is computed conditioned on each weather type, considering either both the local and global predictor info (M4a) or the local predictor information (M4b and M4c) only. For M4c, the regression step is limited to those variables that have not been used in the clustering process (temperatures and humidity). The idea behind the method M4c is that temperature and humidity values some hundreds of kilometers away do not physically affect the predictand at a given location. Hence, they are excluded from the clustering of the large-scale data but included as regressors (from the grid box that is nearest to the location of the predictand). Moreover, this avoids the redundancy of using the same variables for clustering and then for regression conditioned on the resulting clusters.

To obtain the optimum configuration of these methods, different combinations of the geographical zones (Fig. 1) and predictors (Table 2) are tested in the following sections.

4. Cross-validation scheme and validation scores

To appropriately assess and compare the performance of different SD methods, a cross-validation approach is considered to avoid model overfitting. The most popular and simple of these approaches is data splitting, which considers independent data for training (e.g., 80% of the available data) and validation/test (e.g., the remaining 20%). To avoid spurious effects of the particular partition performed, the process needs to be repeated several times, which leads to more robust average scores and

additionally permits for the application of statistical inference in order to estimate confidence intervals of the results. However, in this case, the test subsets for the different realizations may overlap, thus providing non-independent results. To avoid this problem, we consider a nonoverlapping test set selection, namely a k -fold cross-validation approach (Markatou et al. 2005), which is commonly used in the machine learning community to compare the performance of different models. The available data (n years in our study) is divided into k non-overlapping data subsets, each of which contains n/k elements. Each data subset is then used as a test set, with the remaining data acting as a training set in each case. Thus, the resulting k scores are obtained from independent test samples, allowing for a proper statistical inference.

In our case, we consider five replicas (fivefold cross-validation) each containing 8 years for testing, and 32 years for training. To circumvent statistical artifacts potentially arising from trends (see Fig. 2e), we considered a stratified regular sampling where the first test sample was formed by the years 1961, 1966, 1971, 1976, 1981, 1986, 1991, and 1996, the second by the years 1962, 1967, etc., and so on. Note that with this approach we keep an 80%/20% balance in the training/test data, typically used in this type of studies.

Finally, in order to take into account future seasonal shifts as projected by GCM scenario runs, no season-specific models have been considered in this work.

a. Accuracy

Accuracy validation scores assess the correspondence of the simulated and observed day-to-day temperature sequences, which is the basis of the statistical downscaling approach. The Pearson correlation coefficient is used in this paper for this purpose, although there are other popular measures, such as the root-mean-square error (RMSE). Note that correlation (r) and RMSE are related by the equation $RMSE = \sqrt{\sigma_p^2 + \sigma_o^2 - 2r\sigma_p\sigma_o} + b^2$ (Murphy 1988), where b is the bias and σ_p and σ_o are the standard deviations of the prediction and observation, respectively. Thus, since the bias of the statistical downscaling methods was found to be relative low (see section 5), the correlation can be seen as a standardized version of the RMSE, the latter not being shown in this paper. To assess the seasonal dependence of the results, correlation coefficients are calculated both for the annual and season-specific time series.

b. Distributional consistency

Distributional consistency scores evaluate the downscaling methods' capability to reproduce the distribution of the target time series. The most popular scores are the bias (mean difference) and the ratio of variances. In

addition, some studies have focused on the higher-order moments of the distribution (skewness, kurtosis; Huth et al. 2003), trying to obtain a more complete description of distributional similarity. Note that the observed distribution should be reproduced by any SD method applied in a climate change context in order to avoid the post hoc correction of the downscaled time series—such as bias removal, quantile mapping, or output rescaling (Déqué 2007)—which would require the additional assumption of the error being constant under climate change conditions.

In this paper we apply the classical two-sample Kolmogorov–Smirnov (KS) test to evaluate the hypothesis that the observed and downscaled time series come from the same underlying distribution. We computed the KS statistic and the corresponding p values at a gridpoint level. Note that low p values indicate significant distributional dissimilarities between the observed and downscaled series. To avoid the effect of serial correlation on the analysis, we consider time series formed by every tenth day only. Alternatively, we could have modified the test considering the effective sample size (Wilks 2006), but since the length of the series is long enough we preferred using the standard test. As was the case for the correlation coefficient, we applied this test both to the annual and to season-specific time series in order to assess the seasonal dependence of the results.

Besides the KS p value, the annual bias of the downscaled series, as well as the standard deviation of the resulting seasonal biases (σ_{bias} , indicator of the seasonal dependence of the bias), is calculated as additional distributional similarity score. Both the bias and σ_{bias} should be kept small, since large errors are likely to nonlinearly propagate in future climate conditions (Raisanen 2007).

c. Robustness/stationarity to climate change conditions

To test the robustness of the downscaling methods to changing climate conditions (and hence the hypothesis of model stationarity), in this paper we present a test to determine whether or not the performance of a given downscaling method in a historical warm period is significantly different from the performance in a normal/random period, measured in terms of the bias. If the bias in the former case is significantly different, then the method fails to properly predict the warming signal and it is prone to miss the warming signal in future climate change projections. This is done by comparing the biases obtained in the 5-fold test periods with the bias obtained in a “warm” test period, defined by the eight warmest years in the period 1961–2000 on the basis of the maximum temperatures, considering the spatial mean of the standardized anomalies at the 17 high-quality grid boxes of Spain02 as reference value. The resulting

years were 1995, 1989, 1994, 1997, 1961, 1990, 1998, and 2000, in decreasing rank order. Applying the analysis to the minimum temperatures leads to an identical ranking of the warm years, with the exception of the coldest one. Thus, to keep consistency of the results, we decided to use the same period for both variables. The resulting warm anomalies for Tmax and Tmin, with respect to the remaining 32 years, have a spatial mean value of $+0.97^\circ$ and $+0.75^\circ\text{C}$, respectively, and thus can be taken as surrogates of a possible moderate warming allowing to test the methods in conditions similar to those projected by scenario runs for the next few decades.

To quantify whether the bias in the warm period, b_w , is significantly different from the five biases obtained in the normal test periods, $b_k, k = 1, \dots, 5$, (the fivefolds of the cross-validation process) we apply a standard t test to the mean difference $\bar{d} = (1/5)\sum_{k=1}^5 d_k = (1/5)\sum_{k=1}^5 (b_w - b_k)$ in order to test whether this difference is significantly different from zero. Thus, we consider the following test statistic (Dietterich 1998):

$$t = \frac{\sqrt{5}\bar{d}}{\sqrt{\text{var}(d)}}; \quad \text{var}(d) = \frac{1}{4} \sum_{k=1}^5 (d_k - \bar{d})^2, \quad (1)$$

which follows a t distribution with four degrees of freedom. Although it has been recently reported that this approach (k -fold cross validation) may slightly overestimate the variance (Markatou et al. 2005), we apply this conservative procedure in order to minimize the type-1 errors (false detection of significant differences) (DeGroot and Schervish 2002). Note also that although five samples could be considered an insufficient number to estimate the sample variance, the k -fold cross-validation approach has shown to provide similar values to the more computationally intensive leave-one-out cross validation, especially when the size of the test data becomes large (Markatou et al. 2005), as is the case in our study.

Therefore, we will consider the p value corresponding to a two-sided hypothesis test with null hypothesis $H_0 \equiv \bar{d} = 0$ from (1) as a measure of robustness of the SD methods in climate change conditions. Low values (e.g., below 0.05) document a significant difference of the bias in warm conditions compared to the bias in “normal” conditions. Large values, in turn, indicate an only spurious difference between both bias types, the associated SD method being robust to warmer climate conditions.

As an illustrative example, Fig. 3 shows the application of this test to the analog method M1a (based on the closest analog; see Table 3) for minimum temperatures, considering two different predictor sets (see Table 2): P5 (SLP+T2m, left column) and P3 (SLP+T850+Q850, right column) over the domain Z8 (see Fig. 1). The figure shows (first row) a comparison of the biases in

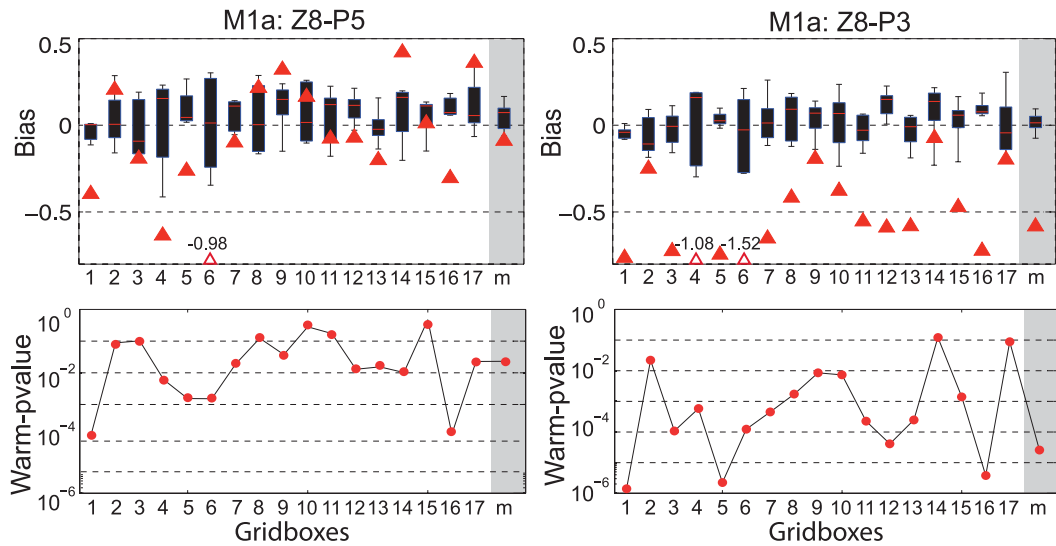


FIG. 3. Robustness of the analog method (M1a) for minimum temperature using two different predictor sets: (left) SLP+T2m and (right) SLP+T850+Q850 considering the 17 stations shown in Fig. 2, indicated along the x axes of each subplot, and the mean of these stations, shaded and labeled as m . The first row shows a comparison of the biases in normal conditions (corresponding to the fivefold cross-validation approach, shown by the box plots) and in warm conditions (red triangles). The second row shows the significance level (p values) of these differences, as obtained from a t test; note that in this case, logarithmic coordinates are used in the y axis.

normal climatic conditions (as represented by the fivefold cross validation and visualized by the box-and-whisker plots) and in the warm period (red triangles) for each of the 17 representative grid boxes shown in Fig. 2a and for their mean (shaded in the figure). Note that whereas for SLP+T850+Q850 (right column) the magnitude of the biases is clearly smaller for the warm conditions than for the normal ones (i.e., the warming is underestimated by this predictor combination), the results for SLP+T2m (left column) are more favorable at most of the grid points. This is qualitatively shown in the figures in the second row, showing the significance level (p values) corresponding to these differences, as obtained from a t test. Thus, this test allows for estimating the statistical significance of these differences and provides a quantitative measure of robustness. Moreover, the results for the spatial mean (labeled by m and shaded in the above figures) are representative of the behavior found for the set of stations, so the corresponding p values can be used for comparison purposes for the area under study as a whole. In the following sections we will follow this approach to characterize the robustness of the methods (and their configurations) in warming conditions.

5. Selection of geographical domains and predictors

This section is dedicated to a screening of the different domains (see Fig. 1) and predictor combinations (see

Table 2) in order to find optimal configurations for downscaling maximum (Tmax) and minimum (Tmin) temperatures. Two commonly used downscaling methods, the nearest neighbor analog method (M1a) and multiple linear regression on PCs (M3a), are applied in this screening process (see Table 3). For validation purpose, the downscaled series corresponding to the five non-overlapping test periods of the cross-validation approach (see section 4) are joined into single continuous 40-yr series that are then evaluated with the abovementioned scores (sections 4a–4c). To avoid spurious effects of serial autocorrelation on the test results, only every tenth time step of these joined series was considered for validation. For simplicity, the results for the individual grid boxes (we considered the 17 high-quality grid boxes shown in Fig. 2) are averaged to obtain a single quantitative measure, except in the case of the robustness test in the warm period, which is applied to the time series of the daily spatial mean biases. Since the 10 domains displayed in Fig. 1 are fully combined with the 14 predictor sets listed in Table 2, the two methods were tested for 140 different configurations.

The dynamic temporal predictor setup (recall: 0000 + 2400 UTC values) was found to generally outperform the static one (recall: 0000 UTC values only) for downscaling Tmax, while the opposite is true for downscaling Tmin. Hence, for the sake of simplicity, Fig. 4 shows the results of the dynamic predictor combinations (P1d, P2d, P3d, P4d, P5, P6d, P7, and P8) for Tmax, and of the

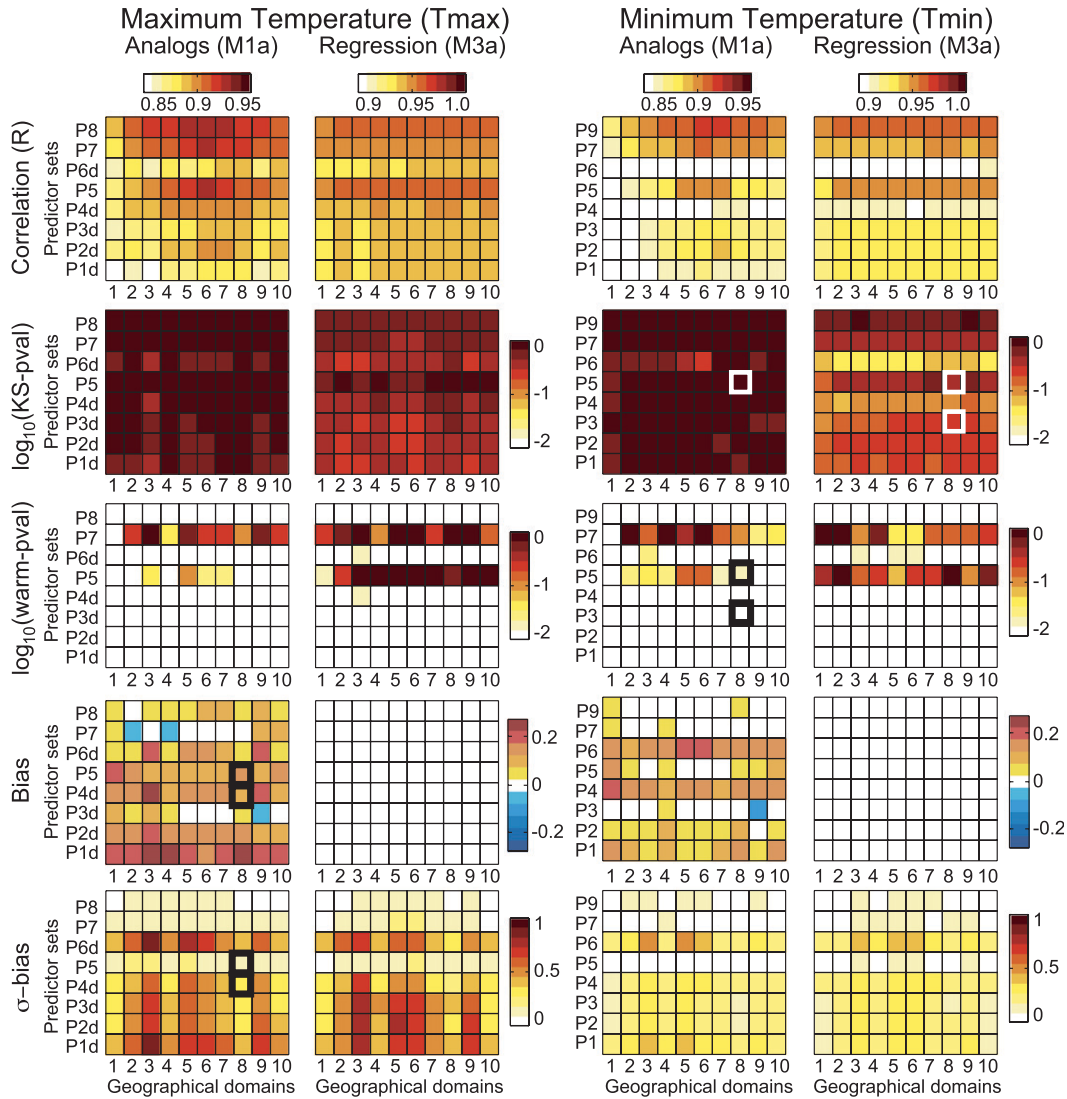


FIG. 4. Calibration results for the 10 domains (x axes of each subplot) and 8 predictor combinations (y axes) of each subplot; the letter d indicates dynamical configuration of the corresponding predictors (see the text for more details). Shown are (top) the Pearson correlation, (second row) KS p value, (third row) warm p value, (fourth row) bias, and (bottom) bias seasonal variability. The first column corresponds to Tmax and the second to Tmin. White/black marked cells are used in the text for illustrative purposes.

static combinations (P1, P2, P3, P4, P5, P6, P7, and P9) for Tmin. Along the columns, the results of the two applied methods are displayed for Tmax (columns 1 and 2) and Tmin (columns 3 and 4), respectively. Along the rows, the following validation scores are shown: the Pearson correlation coefficient (R), p values of the Kolmogorov–Smirnov test for distributional similarity (KS – p value), p values of the robustness test for warm climate conditions (warm – p value), the bias of the complete time series (Bias), and the intraseasonal variability of the bias (σ bias), the latter being defined as the standard deviation of the seasonal biases. In each matrix subplot, the results for all

possible combinations of domains (along columns) and predictor sets (along rows) are shown. Note that the geographical domains have been numbered from east to west, with smaller domains lying in the center and bigger ones at the margins of the x axis (see Fig. 1).

The results are more sensitive to the predictor choice than to the applied geographical domain, although in the case of the analog method (M1a) better results are generally obtained with smaller domains. In particular, information on the near-surface temperature (in terms of T2m and Tx or Tn) generally yields the best results. The correlation and KS p values are highest in these cases,

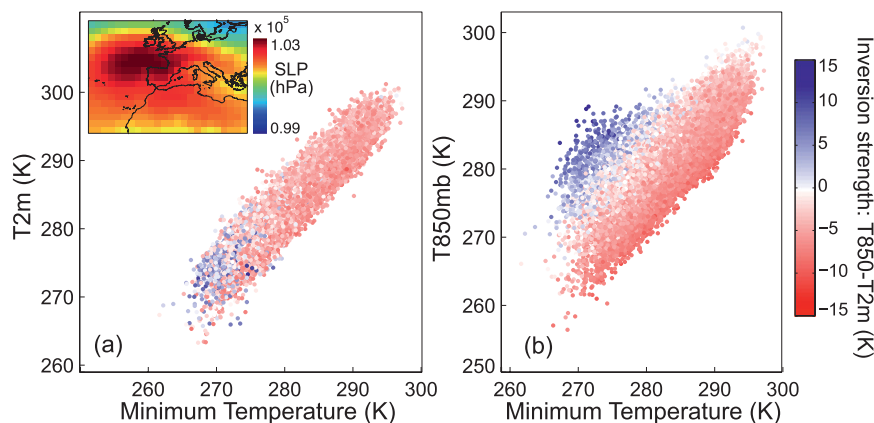


FIG. 5. Analysis of the effect of temperature inversion on the relationship of minimum temperature observations (x axis) vs two predictors: (a) T2m and (b) T850. Colors indicate inversion strength, defined as the temperature difference between T850 and T2m. The values correspond to an illustrative grid box labeled as (A) in Fig. 2a. The inset in (a) shows a typical situation of temperature inversion, obtained as the weather type with highest inversion frequency out of a set of 25 weather types obtained applying the k -means algorithm to SLP.

while the bias and its associated seasonal variability are negligible. Moreover, the warm p values are larger in these cases, indicating a robust behavior in warming climate conditions. With p values lower than 0.01 in most of the cases, the remaining predictor combinations are clearly less robust.

At a seasonal scale, the results are poorest in winter, especially for Tmin, with low correlation values and significant distributional inconsistencies (see more details in the sections below). Moreover, a more pronounced error is found when excluding surface temperature predictors, with a systematic overestimation of low temperature values. As an explanation for this problem we found that T850 does not appropriately discriminate cold episodes related to temperature inversion in the lower troposphere/boundary layer. To characterize this problem we defined the inversion strength as the temperature difference between T850 and T2m [a similar approach was used in Pavelsky et al. (2011)] and studied the relationships between minimum temperature and the predictors focusing on this variable. Figure 5 illustrates this analysis for a particular point by plotting the minimum temperature observations (x axis) versus the closest T2m (Fig. 5a) and T850 (Fig. 5b) predictor values. This figure shows that whereas the cold episodes with strong inversions are appropriately captured by T2m, exhibiting a good linear relationship with Tmin, they correspond to high T850 values, destroying the linear correlation with Tmin. These events have an annual frequency of approximately 4% and typically occur in winter, associated with stable conditions with high surface pressure (see the inset in Fig. 5a for a typical situation, obtained as the weather type with highest inversion frequency,

from the set of 25 weather types obtained applying the k -means algorithm to SLP).

As a general result, the best configuration of predictors and geographical domains found to robustly downscale both Tmin and Tmax is predictor P5 (SLP and T2m) in combination with domain Z8 [southeast (SE)]. This configuration will be used to compare the performance the different statistical downscaling methods in section 6.

As an extension to these general calibration results, more detailed information including a comparison to earlier studies is given in the next three subsections. Alternatively, these subsections may be skipped, in which case the reader should directly proceed to the full comparison of the SD methods (see section 6).

a. Accuracy (correlation)

The results for the Pearson correlation coefficient (first row in Fig. 4) are generally better for Tmax than for Tmin. Moreover, correlation decreases in both cases if near-surface temperature information is excluded from the predictor field. This underlines the predictive power of the latter and gives confidence in the strategy adopted by the Norwegian downscaling community, which exclusively uses T2m for temperature downscaling in many studies (see, e.g., Benestad 2002, 2011). Among the lower free-tropospheric fields (i.e., at 850 hPa), Q —in combination with T —plays an important role for downscaling Tmin whereas it does not improve the results for Tmax. This finding is consistent with Timbal et al. (2003) and Brands et al. (2011b), who applied a version of the analog method for western France and the northwestern Iberian Peninsula. For both Tmax and Tmin, information on middle-tropospheric fields (500 hPa) does not provide

an added value to the above mentioned predictors. Multiple linear regression using PCs (M3a) outperforms the nearest neighbor analog method (M1a). For the latter method, small domains generally perform better than larger ones, which is consistent with Gutiérrez et al. (2004). The largest domain covering the whole European–North Atlantic sector performs worse in any case.

Similar results are obtained when analyzing the season-specific time series (see Fig. 6). Highest correlations are found in autumn (>0.9) and lowest in winter (<0.6 for some predictor–domain combinations).

b. Distributional similarity (KS p values)

In contrast to the results for accuracy (see former section), the results for distributional similarity (in terms of the KS p value) are better for the nearest neighbor analog method (M1a) than for regression using PCs (M3a), particularly in the case of Tmin.

In agreement with the accuracy results, distributional consistency is generally best for autumn and poorest for winter, where, in the case of downscaling Tmin with M3a, significant distributional inconsistencies are found for all combinations of predictors and domains. Figure 7 shows the areas where distributional dissimilarities for Tmin are significant at a test level of 5% (black areas). Results are shown for two different predictor combinations (marked by white boxes in Fig. 4), putting emphasis on the effect of including/excluding surface temperature information. P3 combines SLP with T850 and Q850, while P5 combines SLP with T2m (see Table 2). Both predictor combinations are applied on the same geographical domain (Z8). The first column corresponds to the analog method (M1a) for which significant distributional inconsistencies are virtually absent in any season and for both combinations of predictors (only the results for one of the combinations are shown). The second and third columns show the results for linear regression with PCs (M3a) applied to the just mentioned predictor combinations. Although the area of significant inconsistencies can be considerably reduced by using T2m (i.e., P8) instead of T850 and Q850 (i.e., P3), results for the winter season are far from being satisfactory.

In case of Tmax, domains extended to the south and/or east (e.g., domain 8, SE) yield the best performance, as they allow for solving the problem of systematic distributional inconsistencies in winter (not shown).

c. Robustness in climate change conditions (warm p values)

One of the most surprising results obtained in this study is that related to the robustness of the downscaling methods in anomalous warm periods. In particular, Fig. 4 (third row) shows that the only combinations of

predictors with no significant differences between the bias in warm and normal conditions are those considering T2m. For instance, as we have briefly described previously, Fig. 3 shows the robustness of the analog method (M1a) for Tmin with different predictors (P5 on the left and P3 on the right), but the same geographical domain (Z8). Note that they differ in the use of T2m or T850 and Q850 in addition to SLP, respectively (see Table 2). This figure shows a comparison of the biases for normal conditions as represented the fivefold cross validation (box plots) with the biases for the warm period (red triangles), considering the time series of the spatial mean over the 17 stations shown in Fig. 2. Obviously, P5 leads to more robust results than P3, a result consistently found for all applied SD methods.

d. Bias and seasonal bias variability

Unlike the bias for multiple regression on PCs, the bias of the nearest neighbor analog method (M1a) is especially sensitive to the predictor and domain choice (see fourth row in Fig. 4). Varying the predictor combination for a given domain, or conversely changing the domain while keeping the predictor combination constant, may lead to considerable modifications in the magnitude of the bias. For seasonal bias variability (σ bias, fifth row), however, results are more sensitive to the predictor choice, again obtaining better results when using near-surface (instead of free) tropospheric temperature predictors. Figure 8 gives an illustrative example of the seasonal bias variability for Tmax, applying the nearest neighbor analog method with two different predictor sets: P5 (left column) and P4d (right column) on the same domain Z8 (the corresponding spatial mean results are indicated by the black boxes in Fig. 4). Biases for the complete time series are shown in the first row (annual), while the season-specific ones are shown in rows 2 to 5. Note that although the bias for the complete time series is smaller for P4d than for P5, the opposite is the case for the season-specific results, the latter being more important if working in a climate change context, in which it is important to keep validation results constant throughout all seasons of the year.

6. Intercomparison of the downscaling techniques

In this section, a full comparison of the 12 SD methods listed in Table 3 is given for both Tmax and Tmin, based on the results obtained in the former section (i.e., using the predictor–domain configuration P5-Z8; P5: SLP and T2m, Z8: SE Iberia). Figure 9 shows the results for Tmax (first column) and Tmin (second column) for the 17 high-quality grid boxes of Spain02. Note that instead of providing mean values, box-and-whisker plots of the 17

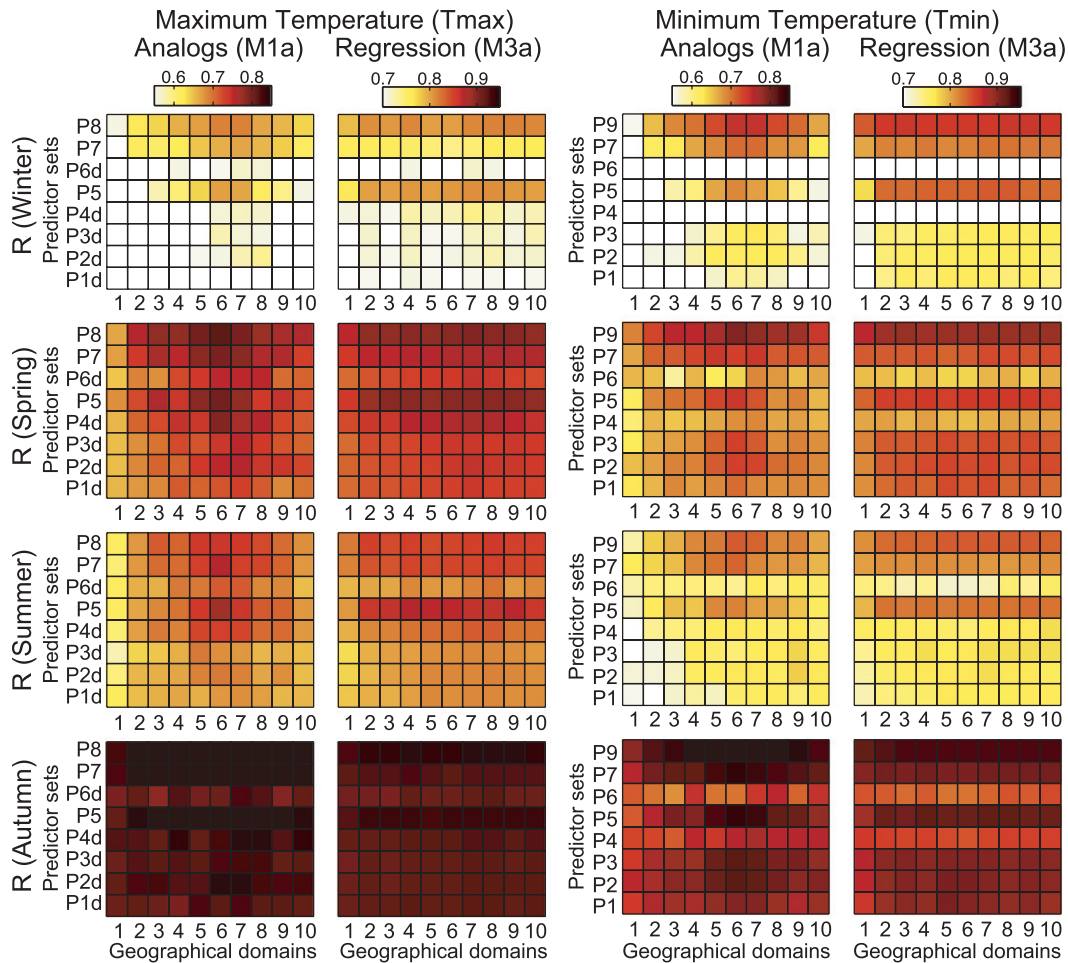


FIG. 6. Correlation, as in Fig. 4 (first row), but for all seasons (in rows). For the sake of comparison, the same color bar has been used for the seasonal panels (both for maximum and minimum temperatures) of a particular method (analog from 0.55 to 0.85 and regression from 0.7 to 0.95 correlations values, respectively).

corresponding validation scores are given in this section, which allows analyzing the spatial variability of the results. Since distributional inconsistencies were found in section 5 particularly for the winter season, we show the KS p values for both the complete series (annual) and the winter-specific ones.

The overall performance of the different SD methods is very similar for both target variables. With the exception of method M1b, the family of analog methods exhibits a good performance, with reasonable correlations (although smaller than for the rest of methods) and optimum distributional consistency results (particularly in winter). Although a systematic warm bias is found for this family (with median values around 0.2°C), the seasonal variability of the bias is small, as compared with the rest of the methods, and hence M1a and M1c could be suitable for climate change applications.

For the family of weather typing methods (M2), overall results are best for the Gaussian variant (M2c), yielding highest KS p values particularly for Tmin. However, the bias variability is too large, particularly for Tmax, so these methods have to be carefully used in climate change conditions. Therefore, M2c is the only weather typing method that could be suitable for climate change applications, particularly for Tmin. Note that in spite of its stochastic nature, it yields reasonable correlation coefficients of at least 0.65, due to the weather typing component.

The family of regression methods (M3) exhibits a good overall performance, with the exception of the technique relying only on the predictor from the closest reanalysis grid box (M3b), which suffers from significant distributional inconsistencies and a large seasonal variability of the bias. Methods M3a and M3c (based on PCs or PCs combined with predictors from the closest reanalysis

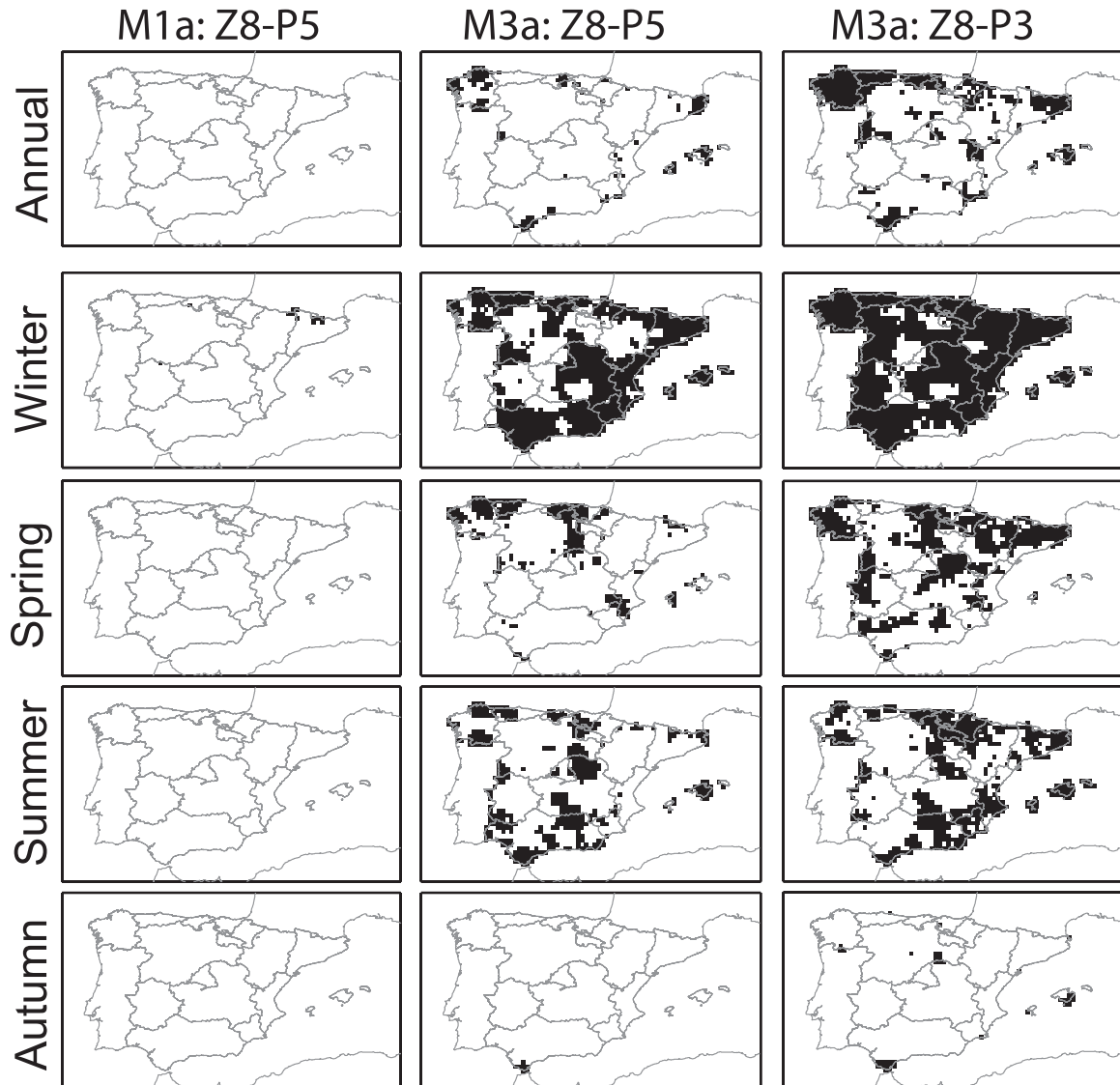


FIG. 7. Distributional similarity ($KS p$ value) of the downscaled and observed series for T_{min} and two different predictors (P5 and P3) applied of the same domain (Z8). (left) Results of the nearest neighbor analog method (M1a) for any of the two predictor combinations. Also shown are results for linear regression using PCs for (middle) including and (right) excluding surface temperature information ($T2m$) from the predictor field. Regions where significant distributional differences between the downscaled and observed series were detected at a test level of 5% are shown in black.

grid box) exhibit high correlation values, good distributional consistency (with the exception of winter for T_{min}), and small biases in all seasons. Therefore these methods could be suitable for climate change studies.

In the case of regression conditioned to weather types (family M4), and in contrast to the M3 family, performance is better when using predictors from the nearest reanalysis grid box (M4b and M4c) than when using PCs (M4a). Note that this is a reasonable result, since the weather types already provide spatial information and, thus, the PCs become redundant in the regression phase. However, the overall performance of the conditioned

regression methods (M4) family is worse than that of the simple/nonconditioned regression (M3), and only method M4c could be considered to be suitable for climate change studies. Note that in the latter case, the circulation predictor (SLP) is used for weather typing and the regression is based on the $T2m$ temperature values.

Finally, Fig. 10 shows the results for testing the robustness of the methods under climate change conditions considering both historical warm periods, used as surrogate of future warming (Figs. 10a–d), and future projections downscaled from a state-of-the-art GCM (the ECHAM5 model), considering the warming signal

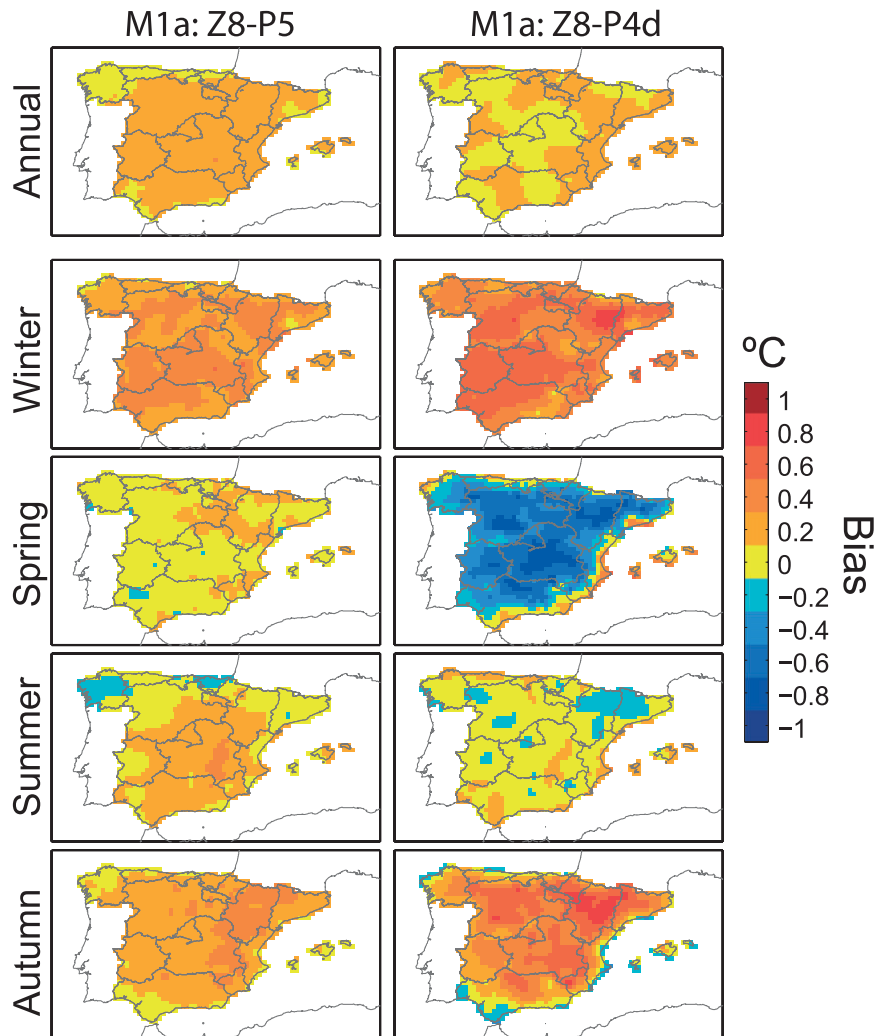


FIG. 8. Annual (first row) and seasonal (in rows) biases for the same downscaling method (analog, M1a) and geographical region (Z8), but with two different predictor sets: (left) P5 and (right) P4d.

for 2071–2100 (A1B scenario) with regard to 1971–2000 (20C3M scenario) (Figs. 10e,f). In this case, following the results from Figs. 3 and 4, the mean temperature of the 17 stations is considered for the analysis. The first row shows the box-and-whisker plots corresponding to the fivefold test periods (indicating normal climate conditions), together with a red triangle indicating the bias of the warm period. Differences between warm and normal periods can be visually established from this figure. The second row shows the statistical significance of these differences, as given by the p values obtained from (1); note that three significance levels (a) 0.01, (b) 0.05, and (c) 0.1 are indicated with the dashed lines in the figures. No significant differences are found for regression and regression conditioned on weather types (except for M4a), indicating their robustness to warmer climate

conditions. Significant differences with p values smaller than 0.01 are found for all weather typing techniques (M2, with the exception of M2b for T_{min} , which exhibits a large bias variance in normal periods) and also for analog techniques M1b and M1c for T_{min} . Moreover, all the analog techniques exhibit significant differences at the level 0.05. In case of the nearest neighbor analog method (M1a), the relative bias differences for the warm period (with respect to the lower bound of the interquartile range; i.e., the 25th percentile of the normal periods) are below 0.1°C (slightly higher for T_{min} than for T_{max}), which is less than 10% of the warm anomaly. However, these differences may nonlinearly propagate in future climate conditions, as given by GCM projections, that are considerably warmer than those considered in this study, so

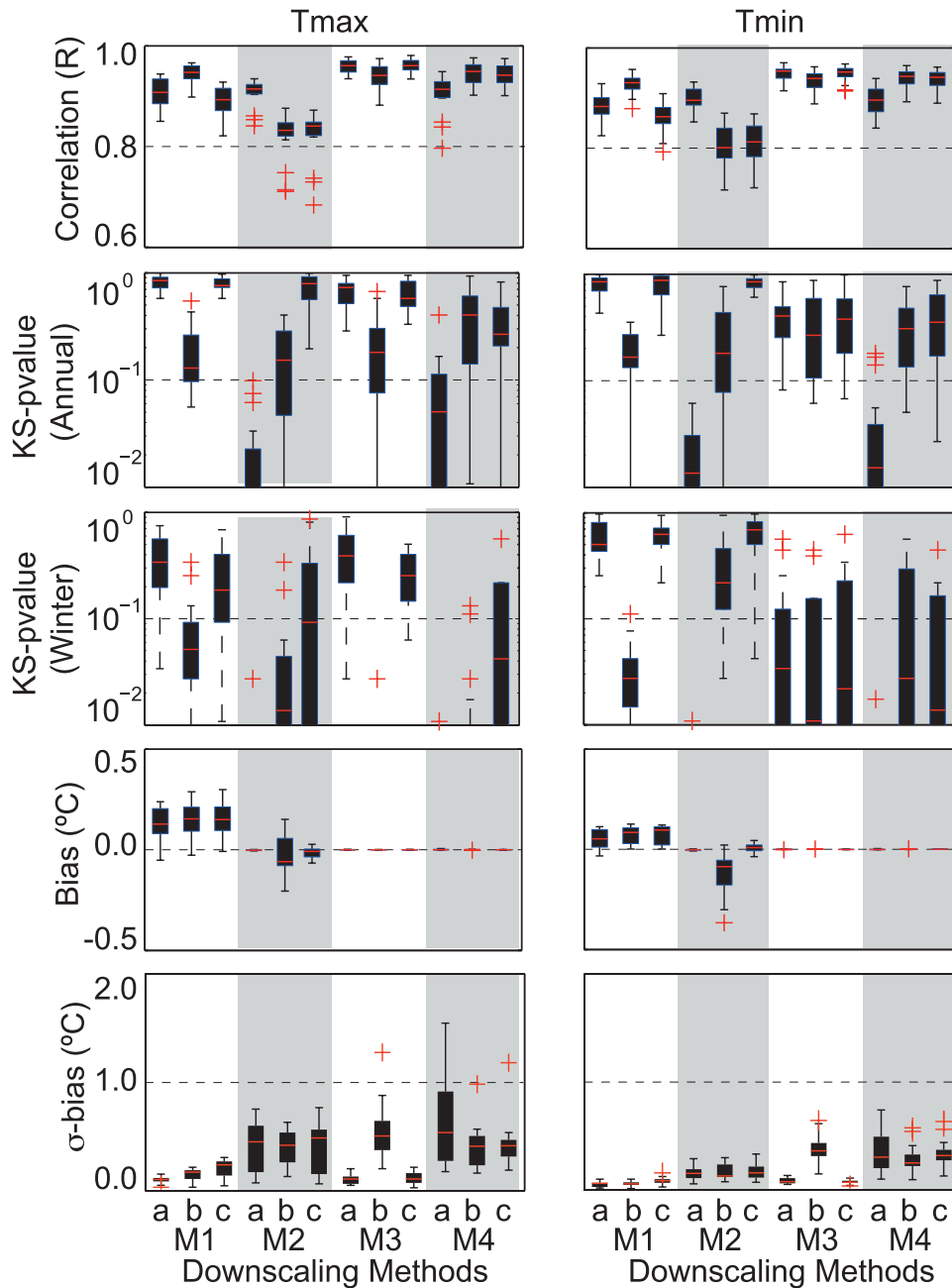


FIG. 9. Performance of the 12 SD methods for (left) Tmax and (right) Tmin according to (top) the Pearson correlation coefficient, (rows 2 and 3) the KS p values for annual and winter series, and (rows 4 and 5) the annual bias and bias seasonal variability; all methods (displayed along the x axes of each subplot) were configured using the same optimal combination of predictors and domain (P5: SLP and T2m; Z8: SE Iberia); see text for details on the construction of the box-and-whisker plots.

the downscaling method may critically underestimate the warming signal.

To test this possibility, we consider a state-of-the-art GCM, the ECHAM5 model by the Max Planck Institute of Meteorology, Germany (Roeckner 2008), and compute the warming signal in the late twenty-first century as the

difference of temperatures in the period 2071–2100 (A1B scenario) and the control period 1971–2000 (20C3M scenarios). Figures 10e and 10f show the warming signal for maximum and minimum temperatures, respectively, as projected by several statistical downscaling methods. Note that, depending on the method, warming values

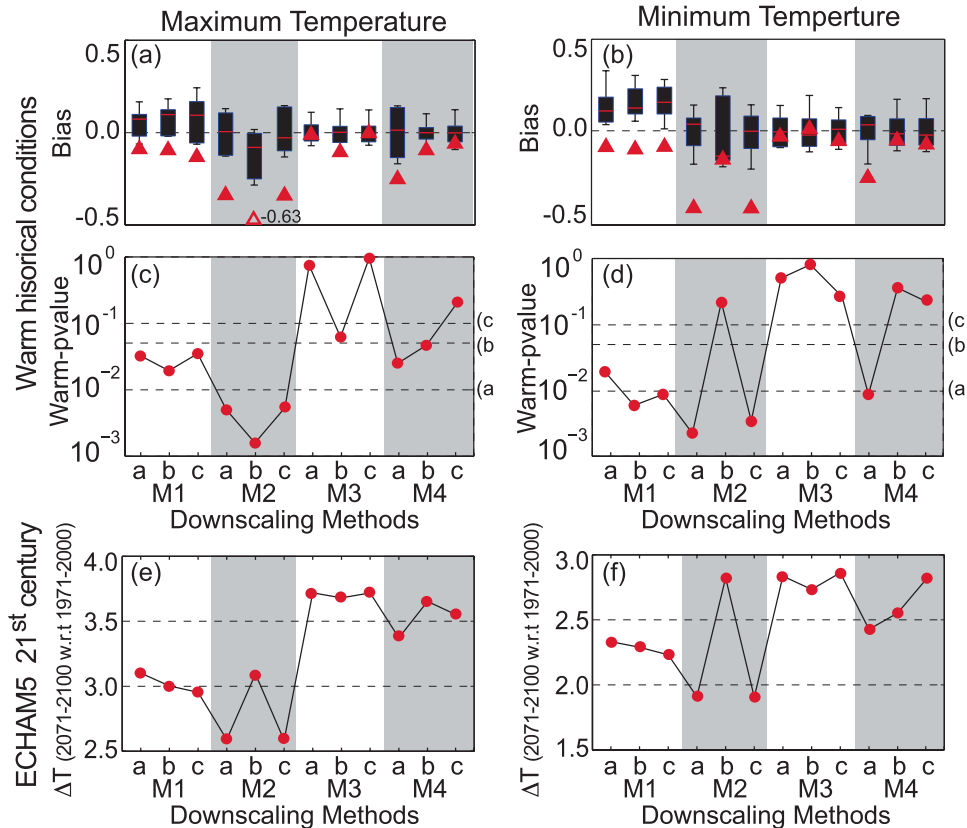


FIG. 10. Robustness of the SD methods (along the x axis of the figures) for (left) T_{\max} and (right) T_{\min} under (top) warm climate conditions. The box-and-whisker plots for the five k -fold normal test periods, together with a red triangle indicating the bias of the warm test period. (middle) The statistical significance of these differences, as given by the p values obtained from (1). (bottom) The warming signal in the late twenty-first century (defined as the difference of temperatures in the period 2071–2100 and the control period 1971–2000, considering A1B and 20C3M projections, respectively) for the ECHAM5 (run 3) model. (c),(d) $\alpha = 0.01, 0.05$, and 0.1 is indicated by a, b, and c, respectively.

range from 2.5° to 3.7°C and from 2° to 3°C , respectively, with a variability of 30%. Note also that these differences are in good agreement with the values given in Figs. 10c and 10d, so the methods failing the historical warm period test are those leading to smaller climate change signals. Therefore, if we consider only the robust statistical downscaling methods given by the test proposed in this paper, the variability of the warming signal would be greatly reduced, leading to robust mean increments of 3.6° and 2.9°C for maximum and minimum temperature, respectively.

7. Conclusions

To determine the suitability of statistical downscaling methods for climate change studies we propose a validation framework using three criteria: accuracy (based on correlation), distributional consistency (based on a two-sample Kolmogorov–Smirnov test),

and stationarity under global warming (based on a t test for a historical warm period), building on a k -fold cross-validation scheme. Note that the first two criteria are currently being used in similar studies to assess the reliability of statistical downscaling methods (see, e.g., Bürger et al. 2012), whereas the latter is a novel approach to assess the robustness of statistical downscaling methods.

Concerning the most suitable predictors and geographical domains for climate change studies, the result of an intercomparison validation analysis of different combinations of factors has shown that 2-m air temperatures are preferable to free-tropospheric temperatures (in particular, temperature at 850 hPa) since, if the latter are applied, results are not reliable and are nonrobust to warming climate conditions for any of the applied methods. An explanation of this result is also provided, related to temperature inversion episodes in the lower troposphere, with high pressure and low

surface temperatures, the latter being systematically overestimated when using T850 as predictor.

The proposed validation framework was applied to a number of downscaling methods commonly used for downscaling temperature, including analog methods, weather typing techniques, multiple linear regression, and regression conditioned on weather types. Overall, regression methods are most appropriate for climate change studies, although they fail to reproduce the observed winter distribution of minimum temperature. Weather typing methods are less appropriate for climate change studies, as they significantly underestimate the temperatures in moderately warmer conditions. Analog methods best reproduce the observed distributions, but they significantly underestimate the observed values in warm periods, although with magnitude smaller than 10% for a warm anomaly close to 1°C. This underestimation is found to be critical when considering the warming signal in the late twenty-first century (differences of the period 2071–2100 with respect to 1971–2000 for A1B and 20C3M scenarios, respectively), as given by a state-of-the-art GCM, the ECHAM5–MPI model. In this case, the different warming values resulting from the statistical downscaling methods—ranging from 2.5° to 3.7°C and from 2° to 2.9°C, for maximum and minimum temperature, respectively—are in good agreement with the robustness significance values, so the methods detected to be nonrobust are those leading to wrong climate change signals with low values. For instance, critical differences of approximately 1°C are found when comparing analog and regression methodologies. Therefore, the proposed test for robustness based on warm historical periods provides an objective criterion for discarding non robust statistical downscaling techniques for climate change future projections. This is the case for the analog and pure weather typing methods, which should not be used for climate change projection of temperatures in the Iberian Peninsula.

Note that analyzing the uncertainty due to different GCMs is out of the scope of this paper and here we just present some evidence of the suitability of the robustness test in warm historical conditions to detect nonrobust methods when applied to future climate change projections.

Finally, note that the configurations considered in this paper are of quite general nature and better performance could be obtained for each particular algorithm with some further adaptation for the particular application at hand.

Acknowledgments. This work has been funded by the Spanish I+D+i 2008–11 Program: A strategic action for energy and climate change (ESTCENA, code

200800050084078) and the project CGL2010-21869 (EXTREMBLES). S.B. was supported by a JAE PREDOC grant (CSIC, Spain). The authors would like to especially thank the three anonymous reviewers who helped to considerably improve this manuscript.

REFERENCES

- Beersma, J., and T. Buishand, 2003: Multi-site simulation of daily precipitation and temperature conditional on the atmospheric circulation. *Climate Res.*, **25**, 121–133.
- Benestad, R. E., 2002: Empirically downscaled temperature scenarios for northern Europe based on a multi-model ensemble. *Climate Res.*, **21**, 105–125.
- , 2005: Climate change scenarios for northern Europe from multi-model IPCC AR4 climate simulations. *Geophys. Res. Lett.*, **32**, L17704, doi:10.1029/2005GL023401.
- , 2010: Downscaling precipitation extremes: Correction of analog models through PDF predictions. *Theor. Appl. Climatol.*, **100**, 1–21, doi:10.1007/s00704-009-0158-1.
- , 2011: A new global set of downscaled temperature scenarios. *J. Climate*, **24**, 2080–2098.
- Brands, S., S. Herrera, D. San-Martín, 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.
- Brandsma, T., and T. Buishand, 1998: Simulation of extreme precipitation in the Rhine basin by nearest-neighbour resampling. *Hydrol. Earth Syst. Sci.*, **2**, 195–209.
- Bürger, G., T. Q. Murdock, A. T. Werner, S. R. Sobie, and A. J. Cannon, 2012: Downscaling extremes—An intercomparison of multiple statistical methods for present climate. *J. Climate*, **25**, 4366–4388.
- DeGroot, M. J., and M. J. Schervish, 2002: *Probability and Statistics*. 3rd ed. Addison-Wesley, 816 pp.
- 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 Planet. Change*, **57**, 16–26.
- Dietterich, T., 1998: Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput.*, **10**, 1895–1923, doi:10.1162/089976698300017197.
- Friás, M. D., E. Zorita, J. Fernández, and C. Rodríguez-Puebla, 2006: Testing statistical downscaling methods in simulated climates. *Geophys. Res. Lett.*, **33**, L19807, doi:10.1029/2006GL027453.
- Gutiérrez, J., A. Cofiño, R. Cano, and M. Rodríguez, 2004: Clustering methods for statistical downscaling in short-range weather forecasts. *Mon. Wea. Rev.*, **132**, 2169–2183.
- Gutzler, D. S., and T. O. Robbins, 2011: Climate variability and projected change in the western United States: Regional downscaling and drought statistics. *Climate Dyn.*, **37**, 835–849, doi:10.1007/s00382-010-0838-7.
- Hagemann, S., C. Chen, J. O. Haerter, J. Heinke, D. Gerten, and C. Piani, 2011: Impact of a statistical bias correction on the

- projected hydrological changes obtained from three GCMs and two hydrology models. *J. Hydrometeorol.*, **12**, 556–578.
- Hanssen-Bauer, I., C. Achberger, R. Benestad, D. Chen, and E. Førland, 2005: Statistical downscaling of climate scenarios over Scandinavia. *Climate Res.*, **29**, 255–268.
- Haylock, M. R., G. C. Cawley, C. Harpham, R. L. Wilby, and C. M. Goodess, 2006: Downscaling heavy precipitation over the United Kingdom: A comparison of dynamical and statistical methods and their future scenarios. *Int. J. Climatol.*, **26**, 1397–1415, doi:10.1002/joc.1318.
- Herrera, S., 2011: Desarrollo, validación y aplicaciones de Spain02: Una rejilla de alta resolución de observaciones interpoladas para precipitación y temperatura en España. Ph.D. dissertation, Universidad de Cantabria, 129 pp. [Available online at www.meteo.unican.es/tesis/herrera.]
- , J. Gutiérrez, R. Ancell, M. Pons, M. 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.
- Hewitson, B. C., and R. G. Crane, 2006: Consensus between GCM climate change projections with empirical downscaling: Precipitation downscaling over South Africa. *Int. J. Climatol.*, **26**, 1315–1337, doi:10.1002/joc.1314.
- Huth, R., 2002: Statistical downscaling of daily temperature in central Europe. *J. Climate*, **15**, 1731–1742.
- , 2004: Sensitivity of local daily temperature change estimates to the selection of downscaling models and predictors. *J. Climate*, **17**, 640–652.
- , J. Kysely, and M. Dubrovsky, 2003: Simulation of surface air temperature by GCMs, statistical downscaling and weather generator: Higher-order statistical moments. *Stud. Geophys. Geod.*, **47**, 203–216, doi:10.1023/A:1022216025554.
- , S. Kliegrova, and L. Metelka, 2008: Non-linearity in statistical downscaling: Does it bring an improvement for daily temperature in Europe? *Int. J. Climatol.*, **28**, 465–477, doi:10.1002/joc.1545.
- , C. Beck, and O. E. Tveito, 2010: Classifications of atmospheric circulation patterns—Theory and applications: Preface. *Phys. Chem. Earth*, **35** (9–12), 307–308, doi:10.1016/j.pce.2010.06.005.
- Imbert, A., and R. 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.
- Lorenz, E., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.*, **20**, 130–141.
- , 1969: Atmospheric predictability as revealed by naturally occurring analogues. *J. Atmos. Sci.*, **26**, 636–646.
- Maraun, D., 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.
- Markatou, M., H. Tian, S. Biswas, and G. Hripcsak, 2005: Analysis of variance of cross-validation estimators of the generalization error. *J. Mach. Learn. Res.*, **6**, 1127–1168.
- Matulla, C., X. Zhang, X. L. Wang, J. Wang, E. Zorita, S. Wagner, and H. von Storch, 2008: Influence of similarity measures on the performance of the analog method for downscaling daily precipitation. *Climate Dyn.*, **30**, 133–144, doi:10.1007/s00382-007-0277-2.
- Murphy, A., 1988: Skill scores based on the mean square error and their relationship to the correlation coefficient. *Mon. Wea. Rev.*, **116**, 2417–2424.
- Palutikof, J., J. Winkler, C. Goodess, and J. Andresen, 1997: The simulation of daily temperature time series from GCM output. Part I: Comparison of model data with observations. *J. Climate*, **10**, 2497–2513.
- Pavelsky, T., J. Boé, A. Hall, and E. Fetzer, 2011: Atmospheric inversion strength over polar oceans in winter regulated by sea ice. *Climate Dyn.*, **36**, 945–955, doi:10.1007/s00382-010-0756-8.
- Pons, N. R., D. San-Martín, S. Herrera, and J. M. Gutiérrez, 2010: Snow trends in northern Spain: Analysis and simulation with statistical downscaling methods. *Int. J. Climatol.*, **30**, 1795–1806, doi:10.1002/joc.2016.
- Preisendorfer, R., 1988: *Principal Component Analysis in Meteorology and Oceanography*. 1st ed. Elsevier, 425 pp.
- Raisanen, J., 2007: How reliable are climate models? *Tellus*, **59A**, 2–29, doi:10.1111/j.1600-0870.2006.00211.x.
- Roeckner, E., 2008: ENSEMBLES ECHAM5-MPI-OM SRESA1B run3, daily values: CERA-DB_ENSEMBLES_MPEH5_SRA1B_3_D. World Data Center for Climate. [Available online at http://cera-www.dkrz.de/WDCC/ui/Compact.jsp?acronym=ENSEMBLES_MPEH5_SRA1B_3_D.]
- Schmith, T., 2008: Stationarity of regression relationships: Application to empirical downscaling. *J. Climate*, **21**, 4529–4537.
- 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:10.1002/joc.2256.
- Timbal, B., and B. McAvaney, 2001: An analogue-based method to downscale surface air temperature: Application for Australia. *Climate Dyn.*, **17**, 947–963.
- , 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.
- Uppala, S., and Coauthors, 2005: The ERA-40 Re-Analysis. *Quart. J. Roy. Meteor. Soc.*, **131**, 2961–3012, doi:10.1256/qj.04.176.
- Vrac, M., M. L. Stein, K. Hayhoe, and X.-Z. Liang, 2007: A general method for validating statistical downscaling methods under future climate change. *Geophys. Res. Lett.*, **34**, L18701, doi:10.1029/2007GL030295.
- 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.
- , —, and —, 2007: Seasonality properties of four statistical-downscaling methods in central Sweden. *Theor. Appl. Climatol.*, **87**, 123–137, doi:10.1007/s00704-005-0223-3.
- Wilby, R., H. Hassan, and K. Hanaki, 1998: Statistical downscaling of hydrometeorological variables using general circulation model output. *J. Hydrol.*, **205**, 1–19.
- , S. Charles, E. Zorita, B. Timbal, P. Whetton, and L. Mearns, 2004: Guidelines for use of climate scenarios developed from statistical downscaling methods. 27 pp. [Available online at <http://www.narccap.ucar.edu/doc/tgca-guidance-2004.pdf>.]
- Wilks, D., 2006: *Statistical Methods in the Atmospheric Sciences*. 2nd ed. Elsevier, 648 pp.
- 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.