Description and validation of a two-step analogue/regression downscaling method

J. Ribalaygua[1], L. Torres[1], J. Pórtoles[1], R. Monjo[1], E. Gaitán[1], M.R. Pino[2]

[1] Fundación para la Investigación del Clima.

C/ Tremps 11, Madrid 28040, Phone: +34915210111,

e-mail: fic@ficlima.org

[2] Instituto de Medio Ambiente. Facultad de Ciencias de la Salud. Universidad de San Jorge.

Abstract This study describes a two-step analogue statistical downscaling method for daily temperature and precipitation. The first step is an analogue approach: the "*n*" days most similar to the day to be downscaled are selected. In the second step, a multiple regression analysis using the "n" most analogous days is performed for temperature, whereas for precipitation the probability distribution of the "*n*" analogous days is used to define the amount of precipitation. Verification of this method has been carried out for the Spanish Iberian Peninsula and the Balearic Islands. Results show good performance for temperature (BIAS close to 0.1ºC and Mean Absolute Errors around 1.9ºC); and an acceptable skill for precipitation (reasonably low BIAS except in autumn with a mean of -18%, Mean Absolute Error lower than for a reference simulation, i.e. persistence, and a well-simulated probability distribution according to two non-parametric tests of similarity).

Keywords: Analogue, regression, statistical downscaling, Spain, verification

1 Introduction

The most powerful tools for constructing future climate projections are General Circulation Models (GCMs) (Huebener et al. 2007). GCMs currently operate at spatial resolutions of around 200 km and this coarse resolution makes their climate information unsuitable as input for impact models (von Storch 1994; Mearns et al. 1997). The latter are essential for designing adaptation policies that seek to minimize negative impacts of climate change and exploit positive ones. To solve this mismatch, in the last decades considerable effort has been put into the development of different strategies in order to infer high-resolution information from low-resolution variables, i.e., 'sensibly projecting the large-scale information on the regional scale' (von Storch et al. 1993). All these strategies fall into the overall denomination of downscaling techniques.

There are two main downscaling approaches (Murphy 1999; Fowler et al. 2007). In so-called dynamical downscaling (Giorgi et al. 2001; Christensen et al. 2007), high-resolution fields are obtained by nesting a Regional Climate Model (RCM) into the GCM (Giorgi et al. 1994; Jones et al. 1997), or using a GCM with variable resolution (stretching technique) (Déqué and Piedelievre 1995). In the statistical approach (Wilby et al. 2004; Imbert and Benestad 2005), high-resolution predictands are obtained by applying relationships identified in the observed climate between these predictands and large-scale predictors to GCM output.

Both approaches have advantages and disadvantages, and both neccesitate assumptions that cannot be verified for the climate change context (Giorgi et al. 2001). They therefore contribute to the uncertainty cascade leading to the final climate simulations. Several criteria can be used to assist in the selection of the most suitable approach depending on the application (Wilby et al. 2004).

The need to consider climate scenario uncertainties (due to uncertainties in the initial condition fields, in the forcing scenarios, in the climate sensitivity of GCMs, in the downscaling skill and so on) in a risk assessment framework leads to the need for probabilistic climate projections. In this context, statistical methods seem to provide a good downscaling option. Their relatively minor needs in terms of both GCM driving data and computational resources are very relevant for processing the growing number of available GCM simulations. In addition, when very high resolution (local) information is demanded, statistical methods can perform better than dynamical ones at least for the present day (Van der Linden and Mitchell 2009), due to the still coarse resolution of current nested or stretched models (and to the fact that RCMs do not use local observations which implicitly capture local meteorological characteristics). The higher diagnostic capability of statistical methods applied at the local scale is generally accepted in the meteorological operational forecasting framework, where statistical reinterpretation systems are the main tool for obtaining local information.

During the last few decades, long-term statistics of climate have experienced relatively small changes compared to inter-annual variability. This variability offers an indirect way to assess the stability in a future climate context of statistical relationships used for downscaling. In this context, two requirements can be identified for statistical downscaling performance. The first requirement is that performance should be good at different time scales (daily, seasonal, annual, decadal...) (Wilby and Wigley 1997). The second requirement is that almost all of the predictor situations that appear in the GCM future climate must be within the applicability range of the statistical relationships determined for the calibration period of the method.

In this study, a novel statistical downscaling method will be described and verified. According to the two requirements mentioned above, verification assesses the performance of the method at different time scales, when the low-resolution input data are extracted from an observed reference dataset (ERA40 Reanalysis).

2 Study area and datasets

2.1 Study area

The area of study was Peninsular Spain and the Balearic Islands. The predominant climate in this area is Mediterranean, although an oceanic climate (in the north) and a mountain climate (in the Sierra Nevada and Pyrenees), among others, can be found. (Capel-Molina 2000). There is a strong spatial and seasonal contrast for both temperature and precipitation. Daily temperatures, for example can have extreme values ranging from below -15ºC, in high valleys in the northeast in winter to over 44ºC in the southwest and southeast of the peninsula. Extreme monthly precipitation can have values ranging from 0 mm in July (in parts of the southeast peninsula), to almost 1000 mm in northern areas, and even at times in the east of the peninsula, where heavy precipitation is normally recorded on a few days (Martín-Vide 2004).

2.2 Surface observations dataset

A group of stations belonging to the Spanish Meteorological Agency (AEMET) has been used. Figure 1 shows the spatial distribution of the 5,273 precipitation (Fig. 1a) and 1,866 temperature (Fig, 1b) stations used. Only stations with at least 2,000 daily records within the common period with ERA40 (1958-2000) have been used.

2.3 Atmospheric dataset: ERA40 Re-Analysis

For verification of the methodology we have downscaled the ECMWF (European Centre for Medium-Range Weather Forecasts) ERA-40 Re-Analysis (http://www.ecmwf.int/research/era/do/get/era-40) for the 1958-2000 period. It has a reduced Gaussian grid with approximately uniform 125km spacing. In this verification procedure, the original temporal and spatial resolutions of ERA40 have been "relaxed" to those used by one of the GCMs to be downscaled (i.e., ECHAM5, Roeckner et al 2006). Thus only 00Z (and not 06Z, 12Z and 18Z, also provided by ERA40) information has been used, with a spatial resolution of 1.8º lat x 1.8º lon. The geographical limits of the atmospheric window used are 31.500ºN to 55.125ºN latitude and 27ºW to 14,625ºE longitude. This window has been defined trying to cover both the geographic area under study as well as the surrounding areas which have a meteorological influence on the peninsula. Likewise, two interior sub-windows have been defined (Figure 2) with different weights assigned to the grid points depending on their influence on the study area.

3 Methodology

3.1. General and theoretical Considerations

In our opinion, the development of a statistical downscaling methodology and the selection of predictors should be carried out based on theoretical considerations and taking into account the final use of the methodology. Four basic ideas should be always kept in mind:

1. The stationarity problem: in a climate change scenario, the relationships between predictors and predictands could change. Thus predictors should be physically linked to the predictands, because these linkages will not change. Also, they should take into account all the physical forcings of these predictands.

2. The characteristics and limitations of the GCMs: the methodologies to be developed will be finally applied to GCM outputs. Therefore, the predictors selected should be well simulated by the GCMs. Moreover, the temporal and spatial resolution of the GCM should also be considered.

3. The statistical tool must be sufficiently "non-linear" to handle the strongly non-linear relationships that link predictors with most local surface weather predictands;

4. For climate change applications it is preferable not to use any seasonal stratification in the selection of predictors, at least in some circumstances: in a climate change scenario, the climatic characteristics of the calendar seasons may change. Thus predictors / predictands relationships detected in a population of "present-day" days belonging to a particular season, with specific climatic characteristics, would not be applicable for future days if the climatic characteristics of that season have changed.

According to these ideas, some general and theoretical considerations regarding the selection of predictors have been identified:

According to idea 1: The selection of predictors should be undertaken based on theoretical considerations, rather than using empirical analyses which could result in non-physically based relationships that may be not applicable in the future due to the stationarity problem. The predictors should be physical forcings of the predictands, or at least, should be physically linked to the predictands. Furthermore, the identified relationships between predictors and predictands should be those which best reflect the physical links between them - again in order to assure so far as possible the stationarity of these relationships. If these requirements are fulfilled, a good diagnostic capability should be obtained at the daily scale. Thus this daily skill should be

analysed since it is required to ensure the stability of statistical relationships for the future.

According to idea 2: The predictors should be field variables, rather than point values, because the former are more reliably simulated by GCMs.

According to idea 2: The predictors should be free-atmosphere rather than boundary layer variables because the former are more reliably simulated by GCMs.

According to idea 2: The predictors should be variables that are well simulated by GCMs. The downscaling method presented here has been adapted to be used for the production of daily operational meteorological forecasts. Many predictors are used in the operational version because it has been shown that they all improve the forecasting skill. But some of them cannot be used in climate simulations because, although they have proven to be well simulated for the next few days by operational Numerical Models, and therefore to be useful in meteorological forecasting, they are too dependant on initial conditions to be well simulated by GCMs for the next decades.

According to idea 2: Working with coarser temporal and / or spatial detail than those provided by GCMs means that some information is not used. Many of the physical forcings of the predictands can only be captured working at temporal and spatial scales that are as small as possible. This could be especially relevant for the simulation of some extreme precipitation events. For these reasons, in our opinion, we should work at daily and synoptic scales, because these are the scales at which the GCMs provide information.

According to idea 3: The statistical method should include strategies to take into account the non-linearity of the relationships between many of the predictors and predictands.

According to idea 4: We think that it is preferable not to make any seasonal stratification in the definition of the predictors/predictands relationships. According to sensitivity analyses performed with the downscaling method presented here, seasonal stratification does not improve the skill, hopefully because the relationships it uses correctly reflect the physical links between predictors and predictands, i.e., they are not just empirical relationships.

The statistical downscaling method has been developed and the predictors selected taking into account the conceptual framework presented above.

The method estimates high-resolution surface meteorological fields for a day "*x*" (the problem day), in two steps: the first step is an analogue technique (Zorita et al. 1993); in the second step, high-resolution surface information is estimated in a different way for precipitation (using a probabilistic approach) and temperature (using multiple linear regression).

Similar two step approaches have been applied in operational forecasting (Woodcock 1980; Balzer 1991). For climate change applications, Enke and Spekat (1997) adopted a similar technique, but where the first step of analogue stratification is replaced by stratification using a predefined clustering of atmospheric patterns. Analogue techniques can be considered as a special form of the clustering approach, where a specific type is determined for each problem day, containing the *n* most analogous days. This strategy greatly reduces the variability within a predefined cluster, which includes days with quite different atmospheric configurations. As a result, analogue techniques generally offer higher diagnostic capability regarding high resolution effects than do predefined clustering schemes.

3.2. First step: the analogue technique

In the first step, the *n* most similar days to day "*x*", identified on the basis of their low-resolution atmospheric fields, are selected from a reference dataset. The skill of the method depends on the spread and quality of the atmospheric and surface reference datasets and, in particular, on the measure used to determine the similarity between days (Matulla et al, 2008). Consequently, according to the ideas mentioned above, the similarity measure must contain diagnostic capability regarding high-resolution precipitation fields (i.e., low-resolution atmospheric fields considered to be similar according to the measure must be associated with similar high-resolution precipitation fields). Thus the similarity measure must assess the likeness of as many as possible precipitation physical forcings (see idea 1) associated with the low resolution atmospheric configurations of the days being compared. In addition to diagnostic capability, the predictor variables of the measure must be reasonably well simulated by GCMs (see idea 2).

Some statistical methods entail strongly automated procedures to select the best predictors and to adjust the optimum predictors/predictand relationships (Hewitson and Crane, 1994; Wilby and Wigley, 1997). This is not, however, easy for analogue techniques for which calibration entails a laborious task of testing different combinations of predictors and similarity measures. Nevertheless, this allows the selection of predictors and similarity measures under theoretical considerations, with the aim of capturing physical forcings between predictors and predictands in order to guarantee the stationarity of the relationships (see idea 1, which we always keep in mind).

The similarity measure between two days must be a scalar magnitude (to allow ordering) and summarises the resemblance of these two days with regard to their predictor fields.

Different algorithms which have traditionally been used to assess similarity between fields were tested in the calibration process: Pearson correlation coefficients and several Euclidean and pseudoeuclidean distances. Similarity measures were required to not only deal with the general pattern of the days being compared, but also with the values of the corresponding individual points of both fields. For the latter requirement, Pearson correlation coefficients perform worse than Euclidean distances and thus provide lower precipitation diagnostic capabilities. The good performance of Euclidean distances is supported by the analogue technique literature (Martin et al. 1997; Kruizinga and Murphy 1983).

The similarity between two days is calculated by determining (and standardising) independently those days likeness with respect to each of the four final predictors fields. The unlikeness of days x_i and x_j regarding each predictor field "*P*", is calculated as a pseudoeuclidean distance with:

$$D_P(x_i, x_j) = \sqrt{\frac{\sum_{k=1}^{N} \left(P_{ik} - P_{jk}\right)^2 \cdot W_k}{\sum_{k=1}^{N} W_k}} \qquad (1)$$

where P_{ik} is the value of the predictor "*P*" of the day x_i, at the grid point *k*; W_k is the weighting coefficient of the *k* grid point. And *N* is the number of the atmospheric grid points.

Once $D_P(x_i , x_j)$ has been calculated, it must be standardized. The standardization is carried out by substituting $D_P(x_i , x_j)$ by $cent_P$, which is the closest centile of the

reference population of Euclidean distances among predictor fields "*P*", to the $D_P(x_i, x_j)$ value. The centile values are previously determined, independently for each "*P*" predictor field, over a reference population of more than 3 000 000 values of D_P. The reference population is calculated by applying equation 1, with the same W_k values, to randomly selected pairs of days. If the closest value to $D_P(x_i, x_j)$ is $cent_{i,j,P}$, it means that about the $cent_{i,j,P}$ % of the 3 000 000 D_P values are lower than $D_P(x_i, x_j)$. The use of centile instead of the original distance D_P allows consideration of dimensionless and initially equally weighted variables for each predictor "*P*" in the measure.

After the four $D_P(x_i, x_j)$ independent calculation and standardization (determination of the closest four $cent_{i,j,P}$), the final similarity ($sim_{i\,j}$) measure between days x_i and x_j is given by the inverse of a weighted average of the $cent_P$ for the four "*P*" predictors (Eq. 2).

$$sim_{i,j} = \left(\sum_{P=1}^{4} w_P cent_{i,j,P} \right)^{-1} \qquad (2)$$

where w_P is the weighting coefficient of the predictor field "*P*". The four predictors are:

- *spd1000*: geostrophic wind speed at 1000 hPa
- *dir1000*: geostrophic wind direction at 1000 hPa
- *spd500*: geostrophic wind speed at 500 hPa
- *dir500*: geostrophic wind direction at 500 hPa

These predictors were selected based on theoretical considerations according to the ideas mentioned in 3.1: they can be derived from only Z1000 and Z500 (which are reasonably well simulated by GCMs, Brands et al, 2010); they are physically linked to precipitation (1000 hPa wind is related to topographical forcings of precipitation, and Z1000 and Z500 to dynamical forcings, according to ω Equation, Holton 2004); they capture most of the precipitation forcings (although convective forcings are only implicitly considered); and they are spatial fields, not grid-point values.

In addition to different algorithms and predictors, different combinations of w_P and W_k coefficients were also tested. The W_k coefficients are required in order to consider the greater influence on Iberian precipitation of wind features closer to the Peninsula. W_k coefficients can be different for each predictor. The combination of $N \times W_k$, found to be more efficient is shown in Figure 2. The four predictors were finally equally weighted ($w_P = 0.25$).

As previously explained, only 00Z information was used ("relaxing" ERA40 time resolution down to that offered by most GCMs). We performed several tests and finally we used the average of 00Z and 24Z (i.e., 00Z of the next day) fields for precipitation and maximum temperature, and 00Z fields only for minimum temperature.

3.3. Second step

3.3.1. Temperature: multiple linear regression analysis.

The estimation procedure for temperatures requires, after selection of the *n* analogous days described above (n = 150 for temperature), further diagnosis using multiple linear regression. Although predictor/predictand relationships determined in this second step are linear, an important part of the non-linearity of the links between free atmosphere variables and surface temperatures is reduced with the first step (analogue) stratification, which selects the most similar days with respect to precipitation and

cloudiness (two of the variables which introduce most non-linearity in the relationships). Linear regression performs quite well for the estimation of surface maximum and minimum temperatures due to the near-normal statistical distribution of these variables. It is important to remember that when using linear regression the predictand quantity is bound to have essentially the same statistical distribution as the predictor(s) variable(s) (Bürguer 1996). In this regard, potential predictors should possess close-to-normal distributions.

The multiple linear regression is performed independently for each surface point, and uses forward and backward stepwise selection of predictors. There are four potential predictors:

1. 1000/500 hPa thickness above the surface station.
2. 1000/850 hPa thickness above the surface station.
3. A sinusoid function of the day of the year.
4. And a weighted average of the station mean daily temperatures of the ten previous days.

Both thicknesses are used to include the strong relationship between lower troposphere and surface temperatures (a meteorological factor). The sinusoid function of the day of the year is used to consider the number of sunlight hours and its effect on the warming/cooling of the surface air (a seasonal factor). And the ten days temperature weighted average is used to account for the soil thermal inertia influence (a soil memory factor).

The non-linear influence of other important meteorological factors, such as cloudiness, precipitation and low troposphere wind speed, is considered through the first-step of analogue stratification. The regression is performed for a population of *n* days which present very similar precipitation, and subsequently very similar cloudiness, conditions.

For each station (and each problem day) the regression is performed twice using as predictands maximum and minimum temperatures. Thus two diagnostic equations are calculated (using the predictand and predictor values of the *n* analogous days population) and applied to estimate both daily temperatures for each station and problem day.

3.3.2 Precipitation: probabilistic approach

Every problem day (x_i) has ***n*** analogues (a_{ij}) each with a certain similarity (sim_{ij}) (n=30 for precipitation). Each analogue (a_{ij}) has an observed precipitation (ρ_{ij}) and an estimated probability (π_{ij}) according to Eq. 3.

$$\pi_{ij} = \frac{sim(a_j, x_i)}{\sum_{k=1}^{n} sim(a_k, x_i)} \tag{3}$$

Thus each problem day (x_i) has ***n*** pairs of [ρ_{ij}, π_{ij}], and a preliminary estimate of precipitation (p_i) can be obtained by combining the ***n*** pairs according to equation 4.

$$p_i = \sum_{j=1}^{n} \rho_{ij}\pi_{ij} \tag{4}$$

Since it is calculated as an average this preliminary estimate greatly smooths the extreme values of precipitation and underestimates the number of dry days.

In order to overcome this limitation, we designed another approach. Consider the pool containing all n analogs of every day in one particular month ($n \times m$ analogues). The aim of our approach is to generate surrogate precipitation time series over this month that follow a similar sample distribution as the precipitation observed in that pool of analogues.

In our approach, we first pool all n analogues for all m days in a particular month. Each member of this pool is characterized by a precipitation amount ρ_{ij} and a probability π_{ij}. In a first step, the elements of the pool are ranked by decreasing precipitation (ρ_{ij}). We then define groups within this pool, as follows. The element with the highest precipitation in the pool becomes the first element of the first group. The first group is then filled by subsequently including elements of subsequent lower ranks until the sum of their probabilities adds up to unity. The element with the following rank is then placed in the second group and the same procedure is repeated until the second group is filled. In this manner all elements of the pool can be ascribed to a group. It can be demonstrated that the number of groups defined in this manner is equal to the number of days m. A new set of m precipitation values, p_h', is obtained by weight-averaging the precipitation of each element in group h (weighted by their relative probabilities), according to equation 5. These new values (p_h') are ranked as well.

$$p_h' = \sum_{k|\Sigma\pi_k=1} \rho_k \pi_k \tag{5}$$

Finally, the first guess p_i obtained by equation 4 is replaced by the new value p_h' that has the same rank as p_i, so that the highest p_h' is associated with the day (x_i) with the highest p_i; the second highest p_h' with the day with the second highest. A simple example of the whole procedure is presented in appendix A.

Proceeding this way, the probability distribution of the m new precipitation values (p_h') is similar to the probability distribution of $n \times m$ values of precipitation (ρ_k) - as desired (see discussion in 5.2). This method allows an empirical distribution of rain amounts for each day of the month to be constructed without assuming any a priori hypothesis about the probability distribution of each month (or assuming a particular associated analytical probability function such as the gamma function).

3.4 Verification of the methodology

Verification of the methodology was carried out by comparing simulated and observed series for the three variables (maximum and minimum temperature and precipitation). Daily Mean Absolute Error (MAE, Stauffer and Seaman 1990) and BIAS (Yu et al. 2006) are used as error measures. For precipitation, the Ranked Probability Score (RPS, Wilks, 2005) is used to compare both preliminary and final precipitation estimates with two reference predictions: climatology and persistence. The Pearson correlation is also used to compare mean simulated and observed values as well as 95th percentiles for the three variables.

In addition, observed and simulated probability distribution functions (PDF) of daily precipitation are compared for each month using two non-parametric goodness-of-fit tests: the Kolmogorov-Smirnov test with *bootstraping* (Marsaglia et al. 2003; Sekhon 2010), which has been used in earlier climatic studies (Abaurrea and Asín 2005), and the Anderson-Darling test (Scholz and Stephens, 1987).

All the statistical analyses were carried out using R, a free software environment for statistical computing and graphics (R Development Core Team, 2010).

4 Results

4.1 Simulation of precipitation

The similarity measure used in analogues selection was adjusted in order to seek the highest predictive capability for precipitation. Thus the probabilistic prediction obtained from the selected analogues provides good results on the daily timescale. For example, categorizing the precipitation into three classes (<0.1 mm, between 0.1 and 10 mm, and greater than 10 mm), the average (for all stations) annual RPS value obtained is 0.08, that represents a Skill Score of 0.46 compared to persistence and 0.23 compared to climatology. For summer, the RPS for simulated precipitation is similar to that for persistence due to the low number of rain days in summer (Fig. 3a). The final precipitation estimate, obtained in the second step of the method, also has a daily MAE smaller than that for "climatology" and "persistence" (Fig. 3b). The average relative bias is practically negligible in all seasons, except autumn when it is around -18% and even around -30% over the eastern peninsula (Fig. 4).

The number of dry days is also correctly simulated. The average bias is less than two days (4% of the total number of dry days).

The 95th percentile can be used as an indicator of high precipitation since it is obtained from the upper tail of the distribution. Thus we have analysed the seasonal 95th percentile of daily precipitation and estimated the correlation between the simulated and observed seasonal 95th percentiles time series (Fig. 5a). Winter is the season best simulated, with correlations of over 0.8 in the southeast peninsula, whereas summer gives correlations of less than 0.4 for Mediterranean areas.

4.2 Simulation of temperature

As regards the simulation of temperature, the largest average bias is found for maximum winter temperature, i.e., almost 0.2ºC, with values of 0.1ºC or lower for other seasons and for minimum temperature (Fig. 6). The daily mean absolute error (MAE) is around 1.8ºC for both maximum and minimum temperature, although the latter varies more between seasons with an average MAE of 1.6ºC in summer and almost 1.9ºC in winter.

The spatial distribution of the errors show that the BIAS for maximum temperature depends on the season: in winter it is nearly always negative (around -0.2ºC); in autumn it is virtually zero; in spring slightly positive (+0.1ºC) in inland areas and somewhat negative on the coast, reaching -0.2ºC in the gulf of Valencia (Fig. 7); and in summer it is generally higher and positive inland and in the southeast of the Iberian Peninsula. For minimum temperature, the average bias in summer is slightly negative on the east coast and in some areas of the south and north coast, and for the remaining seasons very low, with the exception of the southeast (in autumn) and the northeast (in winter).

The MAE for maximum temperature is spatially quite homogeneous and ranges between 1.8 and 2.0ºC. The areas with highest MAE are inland in the southeast and north of Extremadura where it reaches up to 2.2ºC in summer. As regards minimum temperature, the southeast inland area is once again the area with the highest MAE, also in summer - this time around 2ºC. The remaining geographical areas range between 1.6 and 2ºC, although in winter the area with MAE of 2ºC has a wider extension.

As regards the seasonal 95th percentile of daily maximum and minimum temperatures (Fig. 5b and 5c), the highest temporal correlation is around 0.8 in autumn and spring for maximum temperature and around 0.7 in winter for minimum temperature. Summer shows a low correlation (0.4) in the eastern Peninsula for both maximum and minimum temperatures.

5 Discussion

5.1 Advantages and limitations

The methodology presented in this study, as in other statistical approaches, shows disadvantages compared to dynamic downscaling: (1) historical observations of the studied variables are needed; (2) they have possible spatial or inter-variable inconsistencies; and (3) there may be a possible problem of non-stationarity in the relationships between predictors and predictands particularly due to weak physical linkages.

The main advantages of the statistical approaches are two. The first is the low computational cost, which allows the downscaling of many GCM outputs and several greenhouse gas emission scenarios in order to quantify uncertainties (Van der Linden and Mitchell 2009). The second is that specific information is provided for point locations with observations, and in these observations the microclimatic features of these points are implicit. This local detail is relevant as the same future climate may bring changes with respect to the current climate which could be quite different for points which are a few km apart. This supposition has been confirmed with the results obtained when local future climate scenarios are produced using this methodology. Dynamic approaches typically provide spatial resolutions of up to 25 km, which are still insufficient to resolve topography with enough detail and to show differences in the projected changes for points located close together.

Regarding this particular statistical approach, it presents good verification results that are consistent with other studies when comparisons with other downscaling methods are considered (Goodess et al 2011, Brunet et al. *2008*;Van der Linden and Mitchell 2009).

In our opinion, these good verification results are due to some particular characteristics of this methodology: (1) predictors selection is based on theoretical considerations, trying to reflect the physical linkages between predictors and predictands, which to some extent reduces the stationarity problem; (2) it operates at the maximum spatial and temporal resolution offered by GCMs; (3) it considers the full range of data variability (we are not, for example, working with principal components); and (4) it performs linear analysis on the population of analogues, which reduces to a large extent the non-linearity of the relationships between predictors and predictands (see 3.3.1).

It should be mentioned that, though analogue methods ensure to some extent spatial and inter-variable consistency, the second step performed here could reduce this consistency.

As regards limitations, poorer results have been highlighted in the simulation of precipitation in autumn in the Mediterranean area, and for this variable on a daily timescale (although in general the monthly and seasonal scales are well simulated). This limitation may be associated with the insufficient spatial and temporal resolution used (i.e., that offered by GCMs), as they cannot resolve atmospheric structures which are small in size and/or have a short life cycle, such as the convective structures

responsible for heavy precipitations in Spain, especially in Mediterranean areas in autumn.

5.2 Precipitation simulation

Regarding the methodology (see 3.3.2), the second step for precipitation includes a pooling and ranking of the $n \times m$ precipitation amounts corresponding to each group of m problem days (with n analogous for each problem day). Proceeding this way, we obtain a probability distribution for the m final precipitation values (according to equation 5) which is more similar to the probability distribution of the $n \times m$ values of precipitation. To evaluate this, we used the Anderson-Darling test for the final precipitation estimates compared with the $n \times m$ amounts. The test gave a p-value for each group of m days of each station time-series, and all these p-values were averaged for each station. The final estimated precipitation (as well as observations) passes the Anderson-Darling test in almost all cases, with a significant p-value >0.05, while the preliminary estimates do not pass this test (Fig. 8 upper panel).

We also performed the Anderson-Darling test for several simulations of precipitation compared with observed precipitation. Results show that the final estimated precipitation is closer to observed precipitation than both the precipitation of the first analogue and the preliminary estimate (Fig. 8 lower panel).

The use of linear analysis for the second step of precipitation estimation is a subject of debate. For temperature, the second step consists of multiple regression, where the CDF shape of predictors is somewhat similar to the CDF shape of the predictand. In the case of precipitation, for a suitable multiple regression we need to find some predictors whose probability distribution has a similar shape (Bürguer, 1996). However, physically linked predictors (mainly moisture and instability) have very different probability distributions compared to precipitation, and thus linear relationships can not really be identified.

Although the preliminary precipitation estimates obtained by averaging the analogues precipitation amounts provide good results for mean values, they underestimate both the number of dry days and the heavy rainfall amounts. Thus the aim of the probabilistic approach is to obtain precipitation time-series that properly represent the characteristics of the precipitation regime.

Daily precipitation simulated by numerical models generally has very high MAE in comparison with other variables (Hamill 1999; McBride and Ebert 2000), thus some authors use verification methods such as Ranked Probability Scores (Hersbach 2000; Weigel et al. 2006). In this study, however, we have focused on verification of daily precipitation on rainy days, it is thus necessary to compare the PDFs of observed and simulated daily precipitation. In this respect, two non-parametric tests were performed, the Kolmogorov-Smirnov test with a *bootstrap* treatment (KS; Marsaglia et al. 2003, Sekhon 2010), and the Anderson-Darling test (Scholz and Stephens 1987). These tests showed that in general the distribution of simulated precipitation is not significantly different from that observed (p-value > 0.05). The results are similar for every month, with slight advantages for winter months in comparison to summer (Fig. 9).

The simulated daily precipitation presents a substantial MAE and BIAS for autumn in areas with a Mediterranean climate influence, with a clear underestimation of precipitation (Fig. 3 and 4). This is probably due to the difficulty found in the realistic simulation of deep convection, which is typical of the Mediterranean (Lazier et al. 2001; Herrmann 2008). This convection produces very intense precipitation due to persistent convergences of humidity and to the orthographic characteristics of the

eastern peninsular coast (Gibergans et al. 1995; Chastagnaret and Gil-Olcina 2006). Given the synoptic resolution of the GCMs (in general around 2 degrees), it is difficult to capture with precision such mesoscale phenomena; the availability of higher spatial resolutions would allow consideration of other physical predictors in order to improve results – for example, in operational forecasting the convergence of humidity is often used (Jansa et al. 2000).

Future improvements of the methodology for precipitation would be desirable. The inclusion of other physical forcings (humidity and instability) should be tested. The effectiveness of this inclusion will probably be related to an increase in GCM resolution

We have tested the inclusion of humidity as an additional predictor for the second step using different approaches, but so far verification results did not improve significantly. In addition, GCMs currently have problems simulating moisture (Hu et al. 2005; John and Soden 2007), partly due to low spatial resolution (Räisänen 2007). Therefore we decided not to include humidity in the final version of the methodology. Nevertheless, we consider that there are physical links between one day's humidity field and its synoptic configuration (related to air trajectories, temperature, etc.).

5.3 Daily temperature and data accuracy

Initially, maximum and minimum temperatures were simulated from the ERA40 atmospheric configuration at 00Z. This gave larger MAE for maximum temperature than for minimum (2.0°C versus 1.8°C). The reason seemed to be that 00Z information allows good simulation of minimum temperature (which usually occurs around 05Z in this area) but the use of 12Z information is more suitable for maximum temperature (which usually occurs around 14Z). Thus the temporal resolution of the predictors is important in order to minimise errors in simulating temperature.

However, since for many climate models 12Z information is not available, we decided to use the average of the 00Z and 24Z (00Z of next day) to simulate maximum temperature. This reduced MAE to similar values as obtained for minimum temperature, around 1.8 °C (Fig. 7).

The recording precision of temperature observations also clearly affects the mean absolute error of the simulation. If, for example, the MAE obtained averaging over all the stations (many of which show precision of 1ºC) is compared to the results averaging over only those stations with good precision (<0.5ºC), the MAE can be reduced by 3 to 5 tenths of a degree (Fig. 10). For this comparison, we have analysed stations with at least 60% of the series with good precision (295 stations), and those with at least 90% of the series with good precision (only 22 stations).

5.4 Method of verification and temporal evolution

In order to achieve the effect of cross-validation, the methodology includes an "auto-restriction" when simulating past climate: in the first step, the 5 previous days and the 5 subsequent days of the problem day are excluded from the group from which the most analogous days have been selected. However, it is necessary to prove that the methodology also allows for adequate simulation of a climatic period "training" on a different period (a pure cross-validation scheme, from here on referred to as "Cross"). In this respect, Figure 11 shows that the "auto-restriction" gives verification results similar to those obtained when simulating half of the data with the corresponding other

half (Cross); the periods used for this test are 1958-1974 (period 1) and 1975-2000 (period 2).

Note that the "auto-restriction" method gives a MAE which is slightly lower than for the Cross method, owing to the fact that the selection of analogues is more effective the longer the "training" period used.

Another relevant aspect considered in Figure 11 is the capacity to adequately simulate one period from another. That is to say, it is possible to simulate a warmer and slightly drier period (1975-2000) despite training on a relatively cold and wet period (1958-1974), and vice-versa (Brunet et al. 2001; Lopez-Bustins et al. 2008). Thus it is expected that the method described in this paper should be able to adequately simulate anticipated climate changes.

Good performance in the simulation of climate evolution can also be analysed using the Pearson correlation of time series of seasonal precipitation and temperature. In this respect, it has been estimated that the median correlation for seasonal precipitation is R = 0.7 (Fig. 12), whereas for temperature it is R = 0.8 (Fig. 13).

The area which shows most skilful simulation of the temporal evolution of precipitation is the southwest peninsula, particularly in winter, with a correlation of R > 0.9; this is probably due to the fact that the origin of precipitation in this area is largely frontal. In contrast, lower correlations are obtained in the southeast peninsula, the Ebro Valley and the Balearic Islands, especially in summer (R < 0.3), owing to the scarcity of precipitation (Fig. 12a). Figure 12b shows the observed and simulated time series in winter and summer for a station with R equal to the median of all the stations, and with data spanning at least 30 years.

As regards the temporal simulation of temperature, the spatial distribution of the correlations is more homogeneous, especially in spring. The seasons showing poorer simulation are summer and winter in the southeast and southwest respectively, with correlations lower than 0.6 (Fig. 13). Figure 13b shows the observed and simulated seasonal time series for a station with R equal to the median of all the stations, and with at least 30 years of data.

It can also be seen that seasonal simulation of the 95th percentile of precipitation and temperature gives a temporal correlation with observations (Fig. 6) which is spatially coherent with that for the seasonal averages (Fig. 12 and 13).

6 Conclusions

This paper presents a two step statistical downscaling methodology which allows good simulation of past climate on the Spanish peninsula and the Balearic Islands on a local scale, based on ERA40 reanalysis data. The mean absolute error (MAE) for daily precipitation is lower than for two reference simulations (persistence and climatology), whereas for minimum and maximum daily temperature, MAE is around 1.8ºC – however, MAE for temperature is between 3 and 5 tenths of ºC lower if only those stations showing good recording precision (<0.5ºC) are considered.

The bias obtained was generally insignificant, except for autumn precipitation in Mediterranean areas. The reason for this arises from the difficulty in simulating deep convection, which is typical of the Mediterranean, owing to the spatial resolution used (i.e., that offered by the GCMs) Despite this, the PDFs of simulated daily precipitation are not significantly different from observed, at least according to the Kolmogorov-Smirnov (p-value = 0.4, with *bootstraping*) and Anderson-Darling (p-value = 0.3) tests.

Finally, the temporal evolution of climate is also well simulated, both for precipitation (with a correlation of about R = 0.7) and for temperature (with a correlation of about R = 0.8). It is also shown that the method is capable of satisfactorily simulating the period 1975-2000 when trained on the period 1958-1974, and vice-versa.

Acknowledgements

This study was partly supported by the Ministry of Science and Innovation funding under the GENCEI project (contract no. CGL2005-06600-C03-03, 2006-2008). The authors thank the Spanish Meteorology Agency (Agencia Estatal de Meteorología – AEMET) for providing the observed data set and the European Centre for Medium-Range Weather Forecasts (ECMWF) for offering the ERA-40 reanalysis data (http://data-portal.ecmwf.int/data/d/era40_daily). We also thank Clare Goodess (Climate Research Unit, East Anglia University) for her help.

Appendix

An example will help to understand the proposed method and its justification (see point 3.3.2). To simplify, in this example m=4, n=4, and all the analogous days will have the same probability π_{ij}=1/4. Table A.1 shows the supposed observed precipitation (ρ_{ij}) of each analogous day (a_{ij}), and the preliminary precipitation estimate (Eq. 4).

Within these 4 problem days together, there would be a probability of one day with precipitation over 50 mm of 25% (for x_1) + 25% (for x_2) + 50% (for x_4), so it is expected that one day of those 4 has a precipitation over 50 mm. But no preliminary precipitation estimate reaches that amount, due to smoothing in the average. Likewise, the probability of no rain would be 50% (for x_1) + 50% (for x_2) + 100% (for x_3), so it is expected that two of those 4 days are dry, while the preliminary precipitation estimate suggested only one, again due to averaging and smoothing.

To solve this problem, we pooled all $n \cdot m$ analogues (n analogues for each of all m days in the month) to construct a sample distribution, and obtained the final precipitation amount estimation from this joint probability distribution (Eq. 5). Sorting the $m \cdot n$ observed precipitation amounts (ρ_{ij}), the m final precipitation amounts are obtained by averaging each of the m groups of n sorted analogues. This way, one day over 50 mm and two dry days are obtained, and these final precipitation amounts are assigned to each of the problem days according to their preliminary precipitation amount estimates (see table A.2).

The final precipitation distribution is much more similar to the $m \cdot n$ analogue observed precipitation distribution than the preliminary precipitation distribution was. And the extremes (high values and dry days) of that $m \cdot n$ analogue distribution are much better represented.

References

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

Balzer K (1991) Recent improvements in medium-range local weather forecasting in the Deutscher Wetterdienst. In: Lectures of the WMO training workshop on the interpretation of NWP products in terms of local weather phenomena and their verification, pp. 225-229

Brunet M, Aguilar E, Saladíe O, Sigró J, López D (2001) The Spanish Temperature Series. Time variations and trends over the last 150 years. Geophysical Research Abstracts 3: GRA3 5333 76.

Brunet M, Casado MJ, De Castro M, Galán P, López JA, Martín JM, Pastor A, Petisco E, Ramos P, Ribalaygua J, Rodríguez E, Sanz I, Torres L (2008). Generación de escenarios regionalizados de cambio climático para España. Centro de Publicaciones, Secretaría General Técnica. Ministerio de Medio Ambiente y Medio Rural y Marino. Spanish Meteorology Agency (AEMET)

Bürguer G (1996) Expanded Downscaling for Generating Local Weather Scenarios. Clim Res 7:118-28

Capel-Molina JJ (2000) El Clima de la Península Ibérica. Ed. Ariel, Barcelona.

Chastagnaret G, Gil-Olcina A (2006) Riesgo de inundaciones en el Mediterráneo Occidental. Collection de la Casa de Velásquez 95:115-130. ISSN 1132-734

Christensen JH, Carter TR, Rummukainen M, et al. (2007) Evaluating the performance and utility of regional climate models: the PRUDENCE project. Clim Change 81:1-6

Déqué M, Piedelievre JP (1995) High resolution climate simulations over Europe. Clim Dyn 11:321-339

Enke W, Spekat A (1997) Downscaling climate model outputs into local and regional weather elements by classification and regression. Clim Res 8:195-207

Fowler HJ, Blenkinsop S, Tebaldi C (2007) Linking climate change modelling to impacts studies: recent advances in downscaling techniques for hydrological modelling. Int J Clim 27: 1547-1578

Gibergans Baguena J, Llasat MC, Martin Vide J (1995) Precipitaciones extremas en el área mediterránea. Riegos y drenajes XXI, 11:27-34. ISSN 0213-3660

Giorgi F, Shields Brodeur C, Bates GT (1994) Regional climate change scenarios over the United States produced with a nested regional climate model. J Climate 7:375-399

Giorgi F, Francisco R (2001) Uncertainties in the prediction of regional climate change. Global Change and Protected Areas 9:127-139

Goodess CM, Anagnostopoulou C, Bárdossy A, Frei C, Harpham C, Haylock MR, Hundecha Y, Maheras P, Ribalaygua J, Schmidli, J., Schmith T, Tolika K, Tomozeiu R, Wilby RL (2011) An intercomparison of statistical downscaling methods for Europe and European regions – assessing their performance with respect to extreme temperature and precipitation events. Climate Research Unit (University of East Anglia, UK) Research Report, in press.

Hamill TM (1999) Hypothesis Tests for Evaluating Numerical Precipitation Forecasts. Wea Forecasting 14: 155–167

Hersbach H (2000) Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. Wea Forecasting 15:559–570

Herrmann JM (2008) Relevance of ERA40 dynamical downscaling for modeling deep convection in the Mediterranean Sea. Geophys Res Let 35:L04607

Hewitson BC, Crane RG (eds) (1994) Neural nets applications in geography. Kluwer Academic Publishers. Dordrecht.

Holton J (2004) An Introduction to Dynamic Meteorology, fourth ed. Academic Press, New York, 535 pp.

Huebener H, Cubasch U, Langematz U, Spangehl T, Niehörster F, Fast I, Kunze M (2007) Ensemble climate simulations using a fully coupled ocean–troposphere–stratosphere general circulation model. Phil Trans Roy Soc A 365: 2089–210. doi:10.1098/rsta.2007.2078.

Hu H, Oglesby RJ, Marshall S (2005) The Simulation of Moisture Processes in Climate Models and Climate Sensitivity. J. Climate, 18, 2172–2193. doi: 10.1175/JCLI3384.1

Imbert A, Benestad R (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

Jansa A, Genoves A, Garcia-Moya JA (2000) Western Mediterranean cyclones and heavy rain. Part 1: Numerical experiment concerning the Piedmont flood case. Met Apps 7: 323-333

John VO, Soden BJ (2007) Temperature and humidity biases in global climate models and their impact on climate feedbacks. Geoph Res Let 34: L18704, doi:10.1029/2007GL030429

Jones RG, Murphy JM, Noguer M, Keen B (1997) Simulation of climate change over Europe using a nested regional-climate model. II: Comparison of driving and regional model responses to a doubling of carbon dioxide. Q J R Meteorol Soc 123:265-292

Kruizinga S, Murphy AH (1983) Use of an Analogue procedure to formulate objective probabilistic temperature forecasts in the Netherlands. Mon Wea Rev 111:2245-2254

Lazier JR, Pickart RS, Rhines PB (2001) Deep convection. In: Lazier (ed) Ocean Circulation and Climate. Observing and Modelling the Global Ocean. London: Academic Press.

Lopez-Bustins JA, Martin-Vide J, Sanchez-Lorenzo A (2008) Iberia winter rainfall trends based upon changes in teleconnection and circulation patterns. Glob Planet Change 63: 171–176

Marsaglia G, Tsang WW, Wang J (2003) Evaluating Kolmogorov's distribution. J Stat Softw 8:18

Martin E, Timbal B, Brun E (1997) Downscaling of general circulation model outputs simulation of the snow climatology of the French Alps and sensitivity to climate change. Clim Dyn 13:45-56.

Martín-Vide J (2004) Spatial distribution of a daily precipitation concentration index in peninsular Spain. Int J Clim 24: 959-971

Matulla C, Zhang X, Wang XL, Wang J, Zorita JE, Wagner S, von Storch H (2008): Influence of similarity measures on the performance of the analog method for downscaling daily precipitation, Climate Dynamics, 30,133-144

McBride JL, Ebert EE (2000) Verification of quantitative precipitation forecasts from operational numerical weather prediction models over Australia. Wea Forecasting 15:103–121

Mearns, LO (1997) On the statistical evaluation of climate model experiments - Comment. Clim Change 37:443-448

Murphy J (1999) An evaluation of statistical and dynamical techniques for downscaling local climate. J Climate 12:2256-2284

R Development Core Team (2010): R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org. Accessed 6 July 2011

Räisänen, J. (2007) How reliable are climate models?. Tellus A, 59: 2–29. doi: 10.1111/j.1600-0870.2006.00211.x

Roeckner E, Brokopf R, Esch M, Giorgetta M, Hagemann S, Kornblueh L (2006) Sensitivity of Simulated Climate to Horizontal and Vertical Resolution in the ECHAM5 Atmosphere Model. J Climate 19: 3771–3791.

Scholz FW, Stephens MA (1987) K-sample Anderson-Darling Tests, Journal of the American Statistical Association 399: 918–924

Stauffer DR, Seaman NL (1990) Use of four-dimensional data assimilation in a limited-area mesoscale model. Part I: Experiments with synoptic-scale data. Mon Wea Rev 118: 1250-1277

Sekhon JS (2010) Matching: Multivariate and Propensity Score Matching with Balance Optimization. R package version 4.7-11. URL http://CRAN.R-project.org/package=Matching. Accessed 6 July 2011

van der Linden P, Mitchell JFB (eds.) 2009: ENSEMBLES: Climate Change and its Impacts: Summary of research and results from the ENSEMBLES project. Met Office Hadley Centre, UK. 160pp. http://ensembles-eu.metoffice.com/docs/Ensembles_final_report_Nov09.pdf (see. p 68) Accessed 10 February 2012

von Storch V, Zorita E, Cubasch U (1993) Downscaling of Global Climate Change Estimates to Regional Scales: An Applicattion to Iberian Rainfall in Wintertime. Am Meteorol Soc 6:1161-1171

von Storch V (1994) Inconsistencies at the interface of climate impact studies and global climate research. Max Planck Institute for Meteorology Technical Report 122.

Weigel AP, Liniger MA, Appenzelle C (2006) The Discrete Brier and Ranked Probability Skill Scores. Monthly Weather Review 118:135-124

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

Wilby RL, Wedgbrow CS, Fox HR (2004) Seasonal predictability of the summer hydrometeorology of the River Thames, UK. J Hydrol 295:1-16

Wilks DS (2005) Statistical Methods in the Atmospheric Sciences Chapter 7, San Diego: Academic Press

Woodcock F (1980) On the use of analogues to improve regression forecasts. Mon Wea Rev 108:292-297

Yu S, Eder B, Dennis R, Chu S-H, Schwartz SE (2006) New unbiased symmetric metrics for evaluation of air quality models. Atmos Sci Let 7:26–34

Zorita E, Hughes J, Lettenmaier D, Storch Hv (1993) Stochastic downscaling of regional circulation patterns for climate model diagnosis and estimation of local precipitation. Max Planck Institute for Meteorology Technical Report 109.

Tables

Table A.1. Preliminary precipitation estimates (Eq. 4) using m=4 problem days and n=4 analogous days

Problem day (x_i)	Analogous days (a_{ij})	Analogue probability (π_{ij})	Analogue precipitation (ρ_{ij})	Preliminary precipitation (p_i)
x_1	a_{11}	0.25	0	13.75
	a_{12}	0.25	50	
	a_{13}	0.25	0	
	a_{14}	0.25	5	
x_2	a_{21}	0.25	0	16
	a_{22}	0.25	0	
	a_{23}	0.25	4	
	a_{24}	0.25	60	
x_3	a_{31}	0.25	0	0
	a_{32}	0.25	0	
	a_{33}	0.25	0	
	a_{34}	0.25	0	
x_4	a_{41}	0.25	70	36.25
	a_{42}	0.25	60	
	a_{43}	0.25	10	
	a_{44}	0.25	5	

Table A.2. Final precipitation estimates (Eq. 5) using the $n \times m$ analogues of Table A.1, sorted by analogue precipitation (ρ_{ij}).

Analogous days (a_{ij})	Analogue probability (π_{ij})	Analogue precipitation (ρ_{ij})	Final precipitation (p_h')	Assigned to problem day (x_h)
a_{41}	0.25	70	60	x_4
a_{24}	0.25	60		
a_{42}	0.25	60		
a_{12}	0.25	50		
a_{43}	0.25	10	6	x_2
a_{14}	0.25	5		
a_{44}	0.25	5		
a_{23}	0.25	4		
a_{11}	0.25	0	0	x_1
a_{13}	0.25	0		
a_{21}	0.25	0		
a_{22}	0.25	0		
a_{31}	0.25	0	0	x_3
a_{32}	0.25	0		
a_{33}	0.25	0		
a_{34}	0.25	0		

Figure captions

Fig. 1 Meteorological stations used for this study; a) precipitation stations, b) temperature stations.

Fig. 2 Atmospheric windows and grid-point weighting for each atmospheric level.

Fig. 3 Box plots of RPS and MAE for daily precipitation, for all stations, compared with two reference simulations: climatology and persistence.

Fig. 4 Spatial distribution of BIAS for precipitation: a) absolute BIAS, b) relative BIAS.

Fig. 5 Spatial distribution of the Pearson correlation for simulated and observed seasonal 95^{th} percentile time series, for: a) daily precipitation, b) daily maximum temperature, and c) daily minimum temperature.

Fig. 6 Box plots for BIAS and daily MAE for maximum and minimum temperature for all stations.

Fig. 7 Spatial distribution of BIAS and daily MAE for maximum and minimum temperature: a) BIAS of maximum temperature, b) BIAS of minimum temperature, c) MAE of maximum temperature, d) MAE of minimum temperature.

Fig. 8 a) Box plots of mean P-values, for all stations, for the Anderson-Darling test on the indistinguishability of the simulated $m \times n$ values of rainfall (for each of m = 30 problem days, with n = 30 analogues for each one), compared to the m: observed values, preliminary and final estimated values of precipitation. b) The same test for comparing observed values with several simulations: value of the first analogue, preliminary and final precipitation estimates. For each station, the mean P-value is the average for every calculated P-value for each m-days period.

Fig. 9 Box plot of P-values, for all stations, for two non-parametric tests of the indistinguishability between the observed and simulated probability distribution functions, for precipitation amounts on wet days: a) Kolmogorov-Smirnov test with *bootstrap*ing, and b) Anderson-Darling test.

Fig. 10 Box plots of monthly MAE according to data recording precision: a) MAE for daily maximum temperature for all stations (black) and for those with intermediate precision (at least 60% of the data with precision higher than 0.5ºC) (blue); b) The same as a) but for better precision (at least 90% of the data with precision higher than 0.5ºC (green); c) and d) The same as a) and b) respectively, but for minimum temperature.

Fig. 11 Box plots of the comparison between BIAS and MAE according to two verification methods (original and Cross), and two periods (Period 1 and Period 2), for all stations: a) BIAS of precipitation; b) BIAS of maximum temperature; c) MAE of monthly precipitation; d) MAE of daily maximum temperature.

Fig. 12 Correlation of simulated and observed time series of seasonal precipitation: a) Spatial distribution of the correlation for the four seasons; b) seasonal precipitation (winter and summer) time series for a station with correlation equal to the median (R = 0.7; station is Requejada Reservoir, AEMET code 2232),.

Fig. 13 Correlation of simulated and observed time series for seasonal maximum temperature: a) spatial distribution of the correlation for the 4 seasons; b) seasonal maximum temperature time series for a station with a correlation equal to the median (R = 0.8; station is Almazán, AEMET code 2045

Figure captions

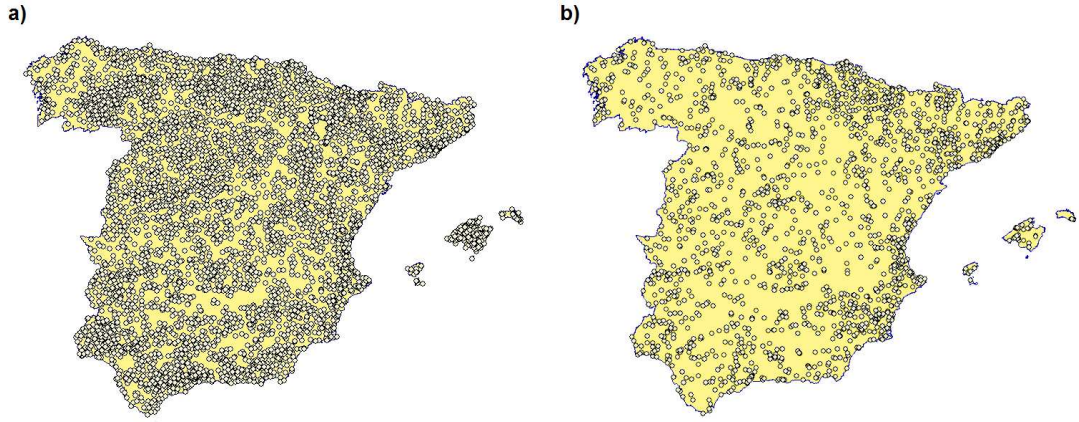

Fig. 1 Meteorological stations used for this study; a) precipitation stations, b) temperature stations.

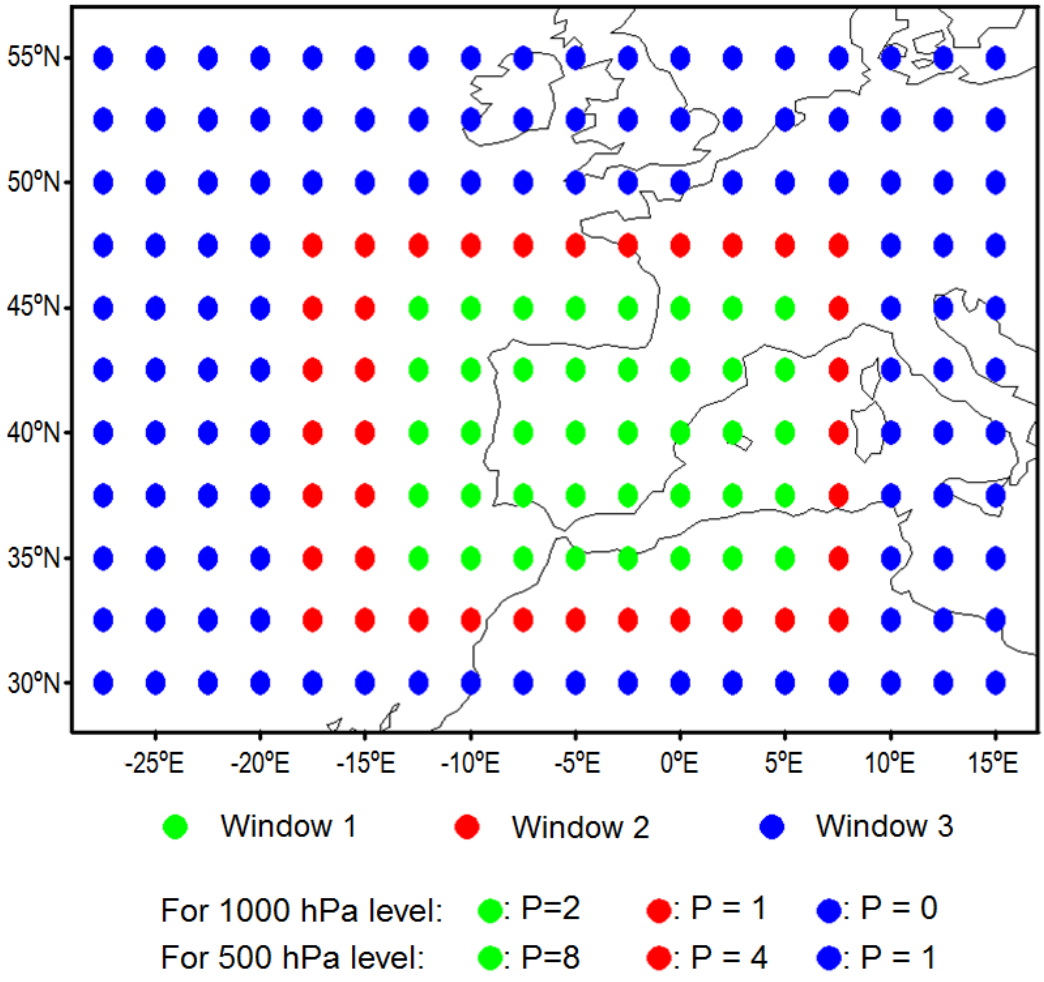

Fig. 2 Atmospheric windows and grid-point weighting for each atmospheric level.

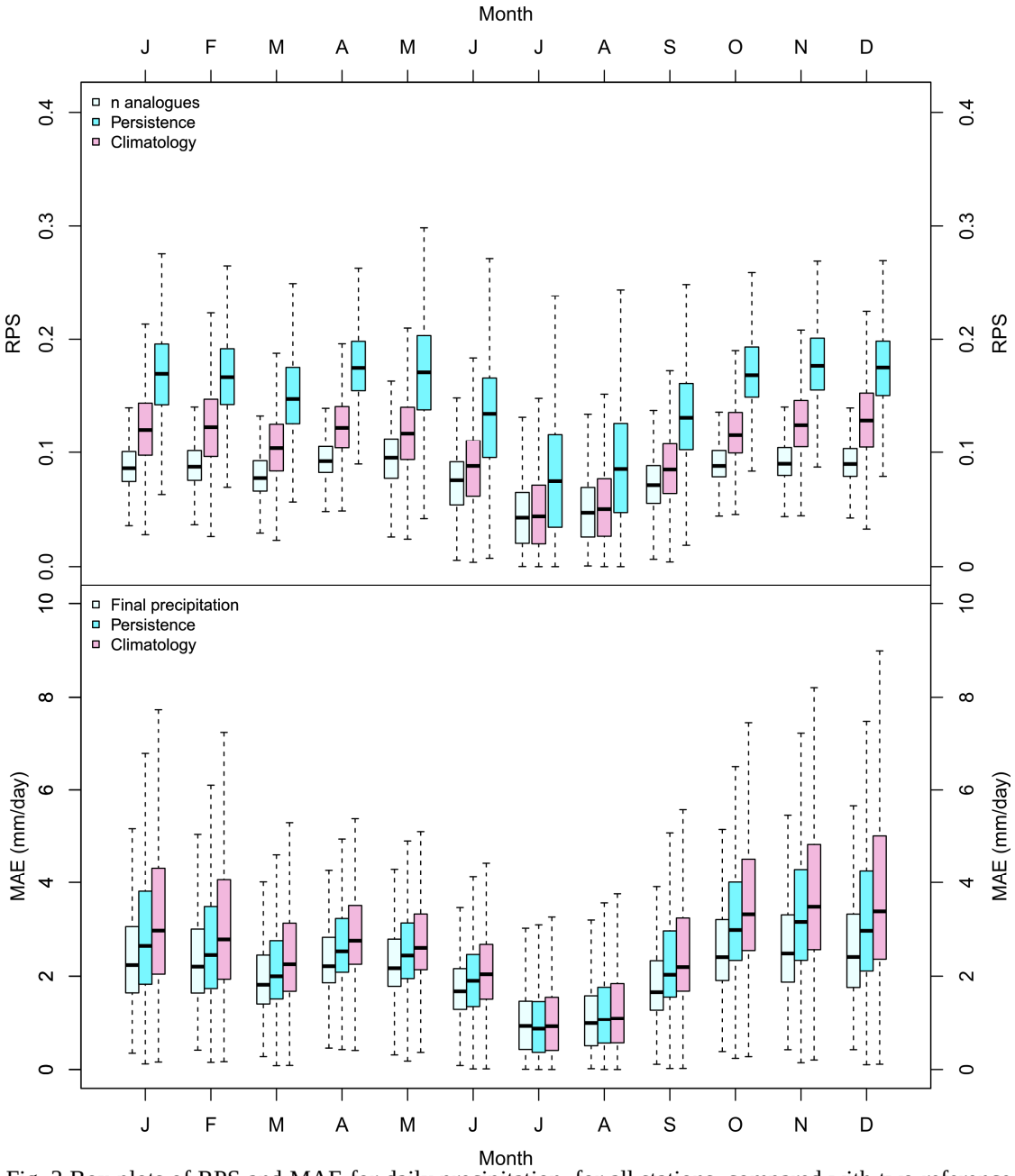

Fig. 3 Box plots of RPS and MAE for daily precipitation, for all stations, compared with two reference simulations: climatology and persistence.

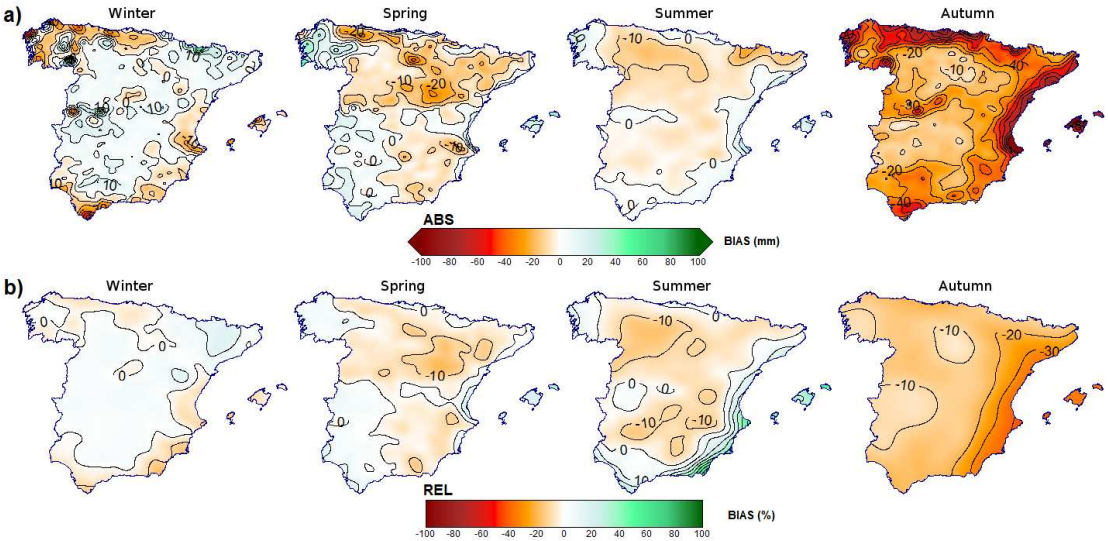

Fig. 4 Spatial distribution of BIAS for precipitation: a) absolute BIAS, b) relative BIAS.

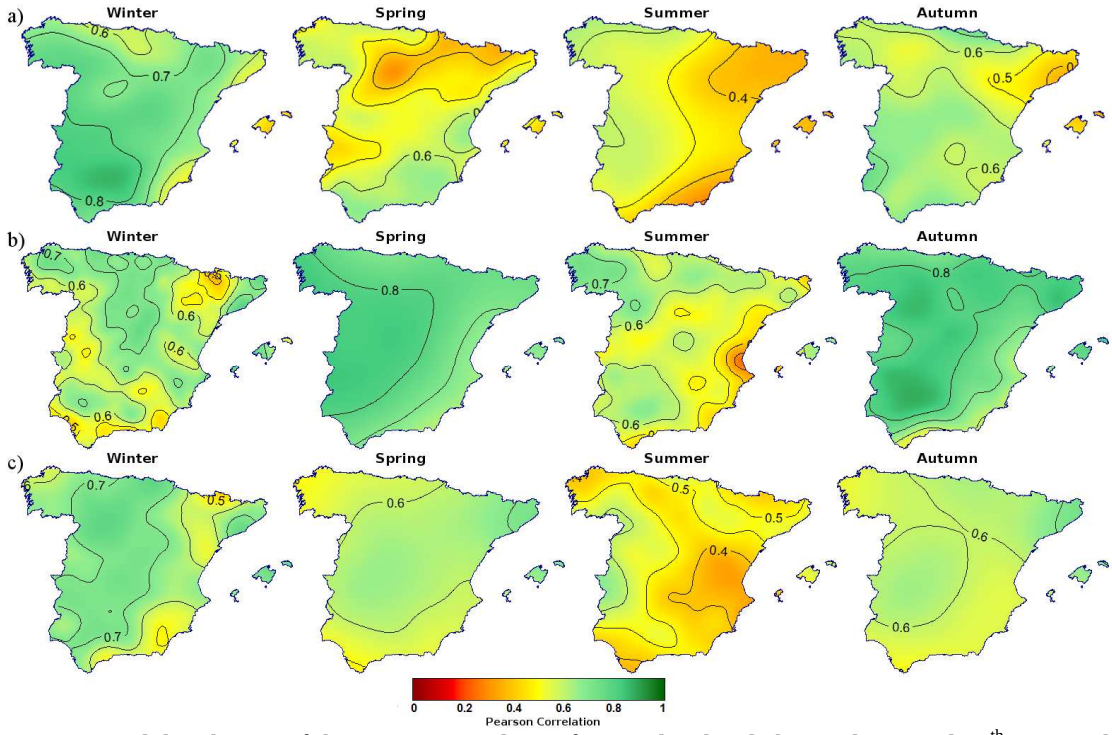

Fig. 5 Spatial distribution of the Pearson correlation for simulated and observed seasonal 95th percentile time series, for: a) daily precipitation, b) daily maximum temperature, and c) daily minimum temperature.

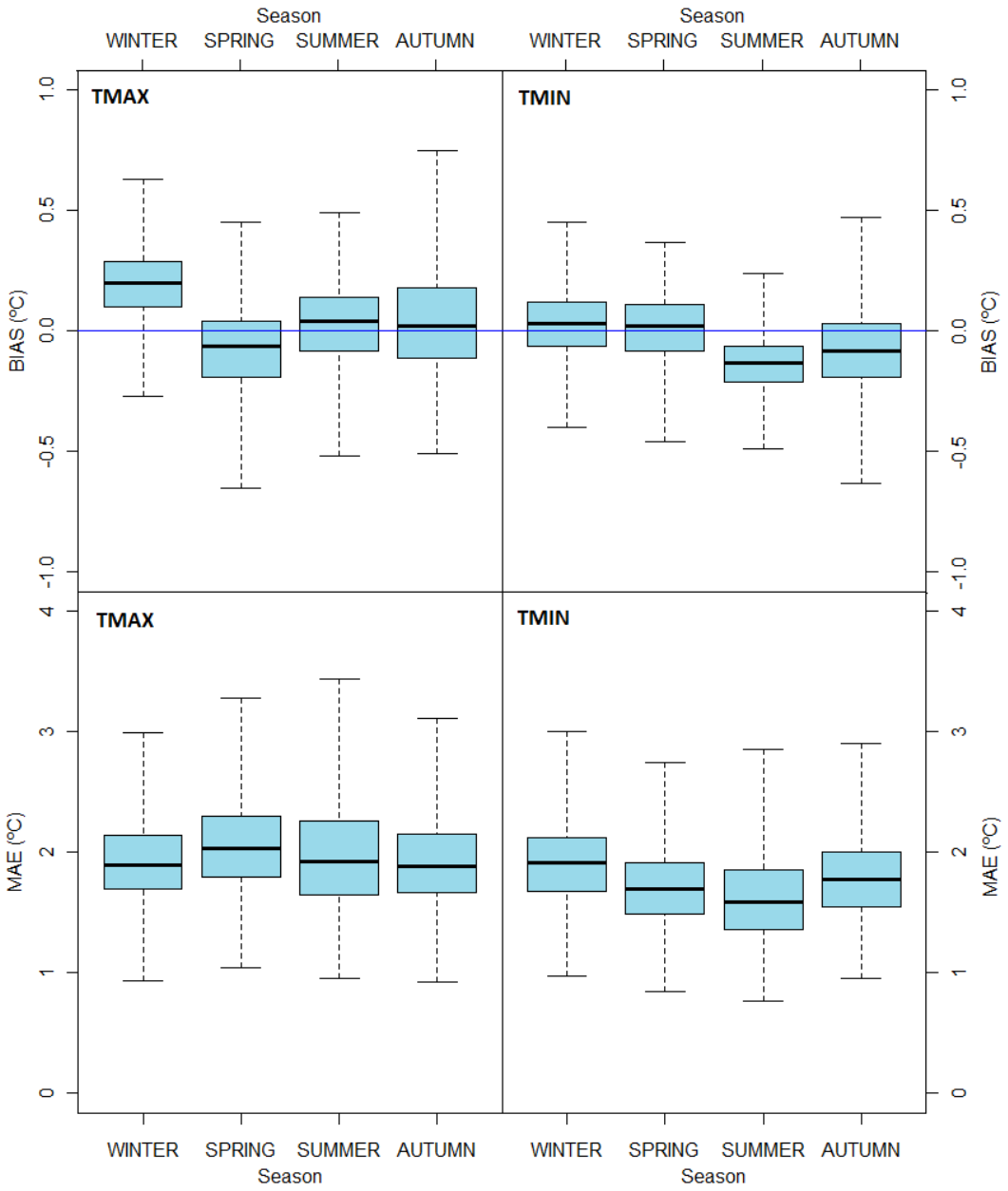

Fig. 6 Box plots for BIAS and daily MAE for maximum and minimum temperature for all stations.

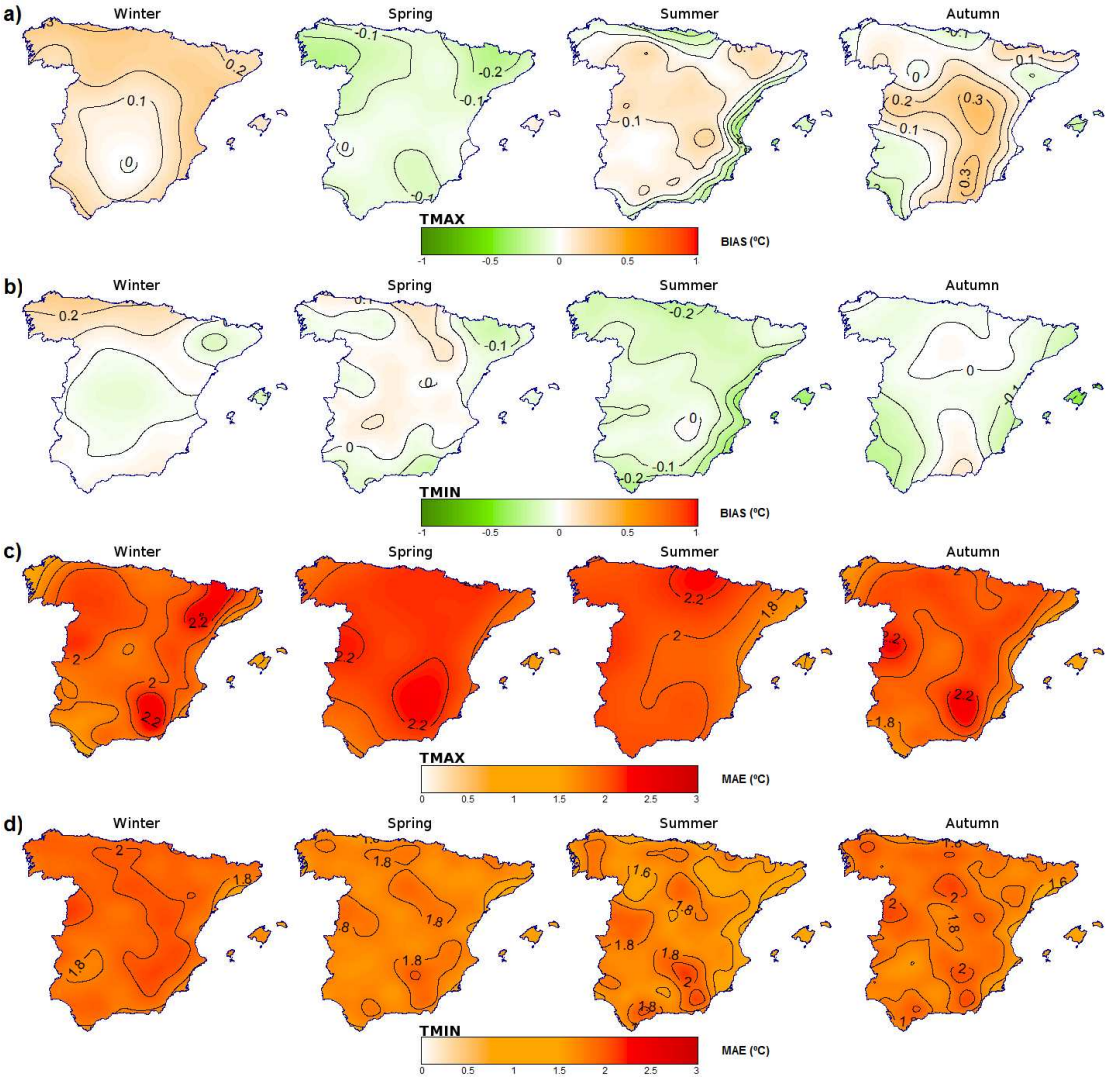

Fig. 7 Spatial distribution of BIAS and daily MAE for maximum and minimum temperature: a) BIAS of maximum temperature, b) BIAS of minimum temperature, c) MAE of maximum temperature, d) MAE of minimum temperature.

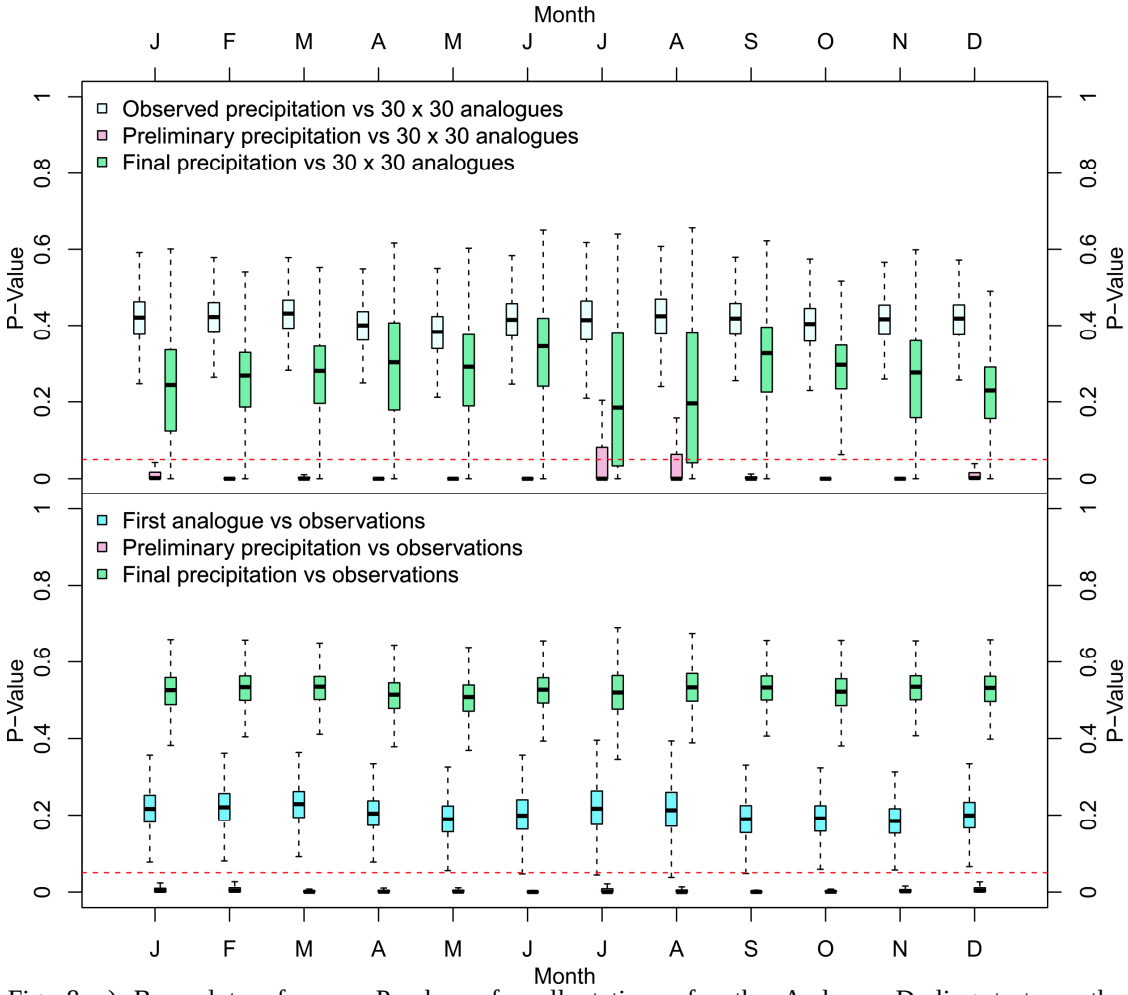

Fig. 8 a) Box plots of mean P-values, for all stations, for the Anderson-Darling test on the indistinguishability of the simulated $m \times n$ values of rainfall (for each of m = 30 problem days, with n = 30 analogues for each one), compared to the m: observed values, preliminary and final estimated values of precipitation. b) The same test for comparing observed values with several simulations: value of the first analogue, preliminary and final precipitation estimates. For each station, the mean P-value is the average for every calculated P-value for each m-days period.

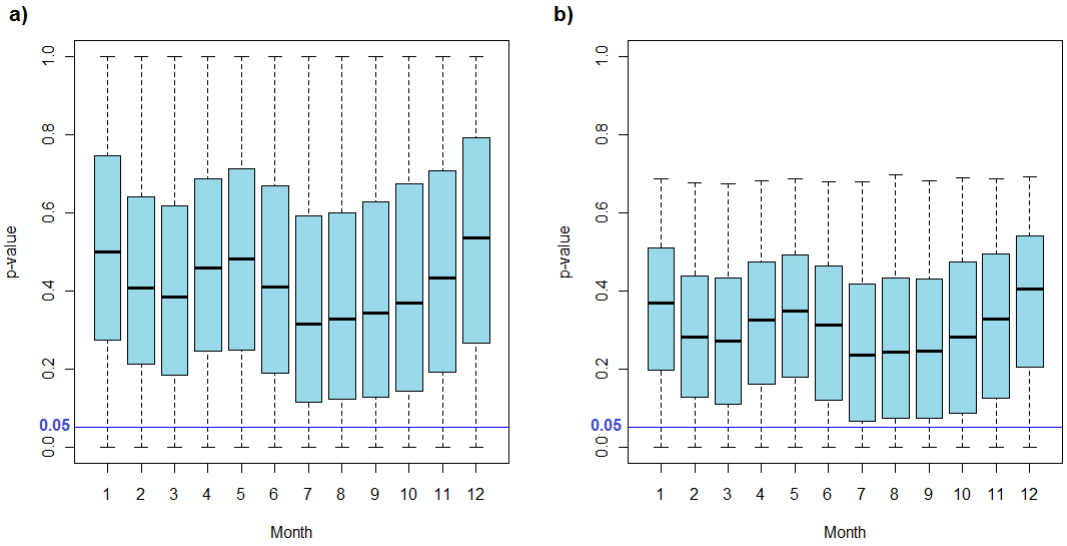

Fig. 9 Box plot of P-values, for all stations, for two non-parametric tests of the indistinguishability between the observed and simulated probability distribution functions, for precipitation amounts on wet days: a) Kolmogorov-Smirnov test with *bootstrap*ing, and b) Anderson-Darling test.

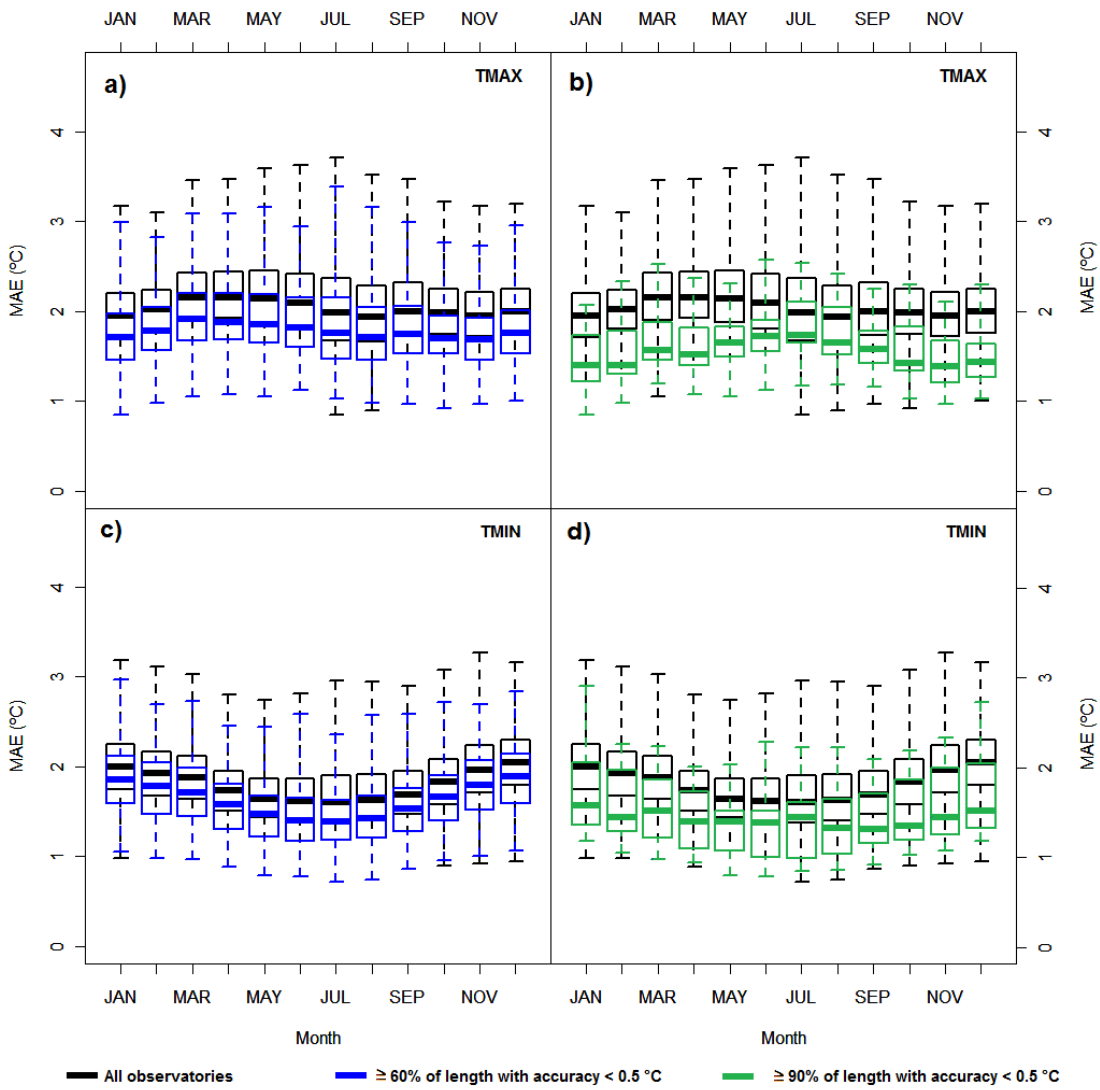

Fig. 10 Box plots of monthly MAE according to data recording precision: a) MAE for daily maximum temperature for all stations (black) and for those with intermediate precision (at least 60% of the data with precision higher than 0.5ºC) (blue); b) The same as a) but for better precision (at least 90% of the data with precision higher than 0.5ºC (green); c) and d) The same as a) and b) respectively, but for minimum temperature.

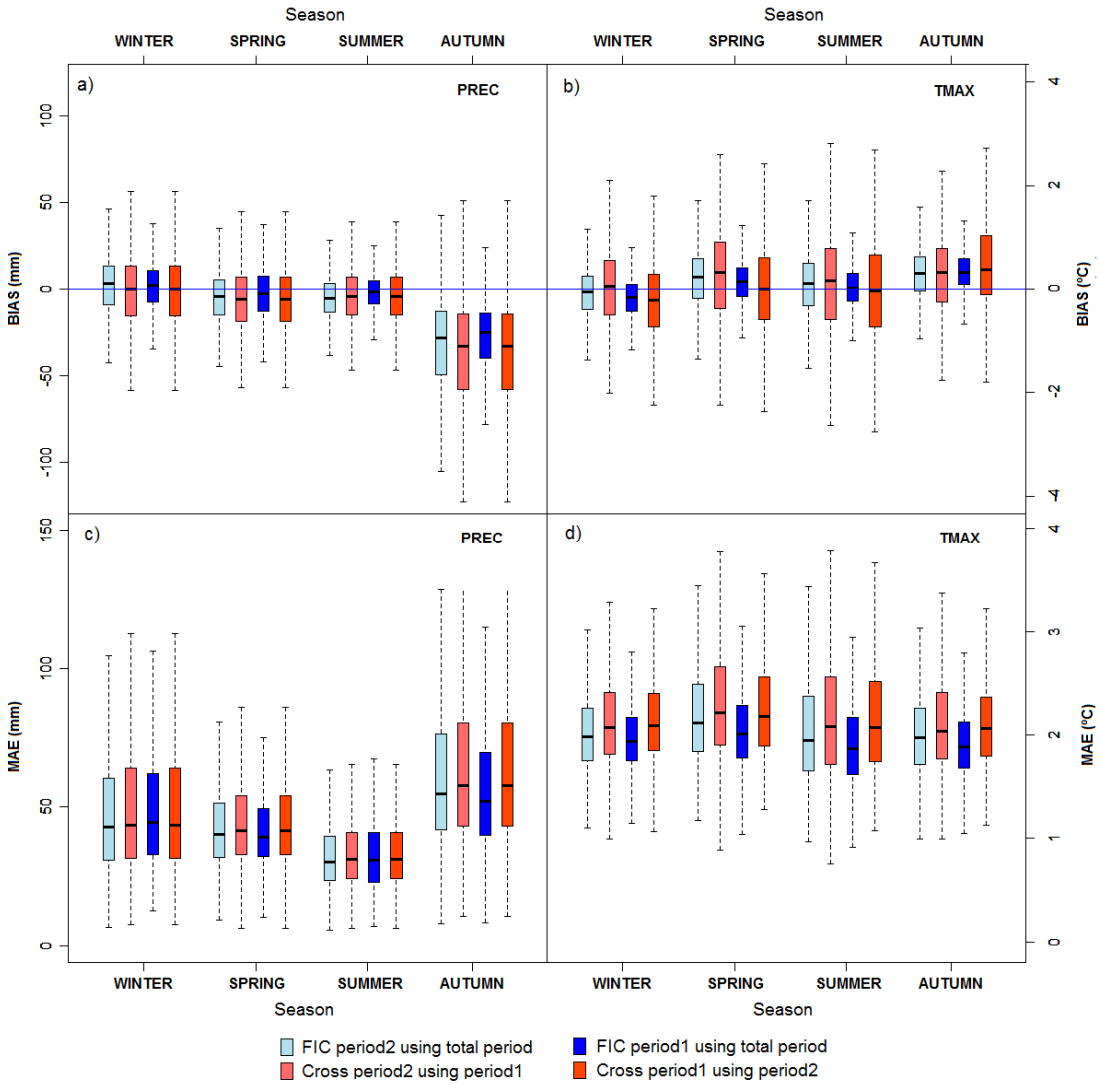

Fig. 11 Box plots of the comparison between BIAS and MAE according to two verification methods (original and Cross), and two periods (Period 1 and Period 2), for all stations: a) BIAS of precipitation; b) BIAS of maximum temperature; c) MAE of monthly precipitation; d) MAE of daily maximum temperature.

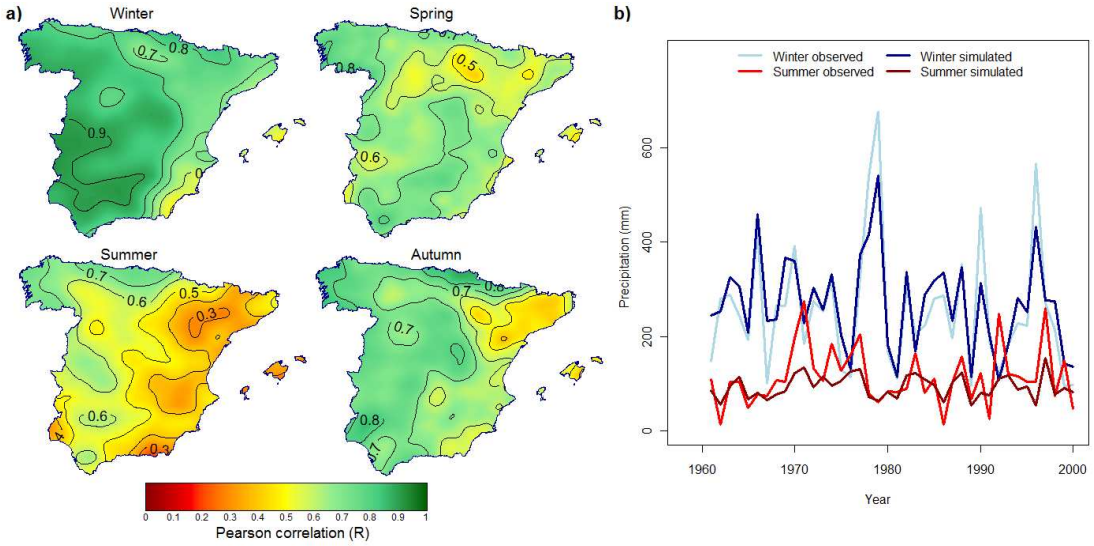

Fig. 12 Correlation of simulated and observed time series of seasonal precipitation: a) Spatial distribution of the correlation for the four seasons; b) seasonal precipitation (winter and summer) time series for a station with correlation equal to the median (R = 0.7; station is Requejada Reservoir, AEMET code 2232),.

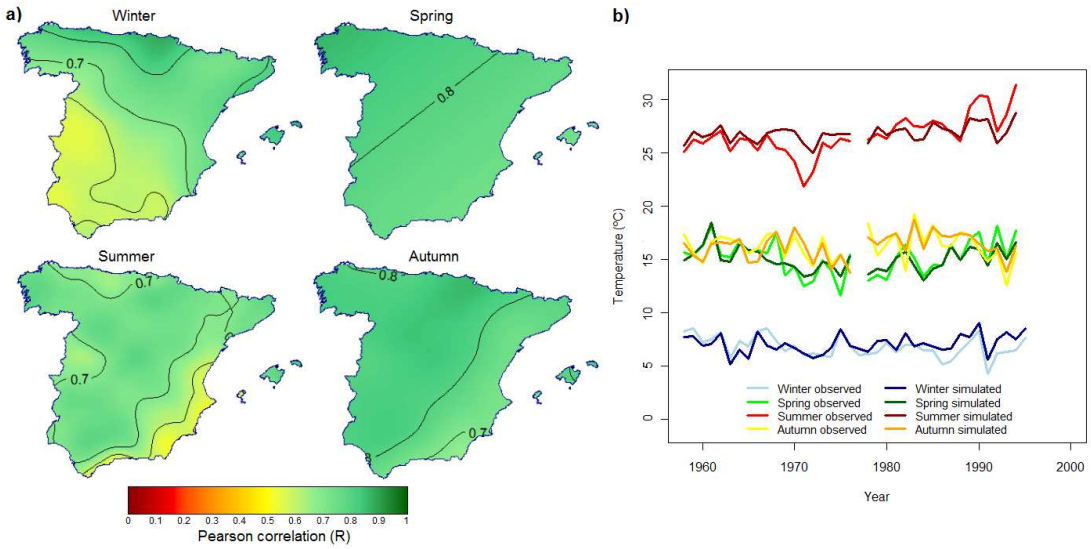

Fig. 13 Correlation of simulated and observed time series for seasonal maximum temperature: a) spatial distribution of the correlation for the 4 seasons; b) seasonal maximum temperature time series for a station with a correlation equal to the median (R = 0.8; station is Almazán, AEMET code 2045