Tellus (2005), 57A, 448–463
Printed in UK.

TELLUS

Operational predictability of monthly average maximum temperature over the Iberian Peninsula using DEMETER simulations and downscaling

By M. DOLORES FRÍAS[1,*], JESÚS FERNÁNDEZ[2], JON SÁENZ[2] and CONCEPCIÓN RODRÍGUEZ-PUEBLA[1], [1]*Dept. Física General y de la Atmósfera, Universidad de Salamanca, Salamanca, Spain;* [2]*Dept. Física Aplicada II, Universidad del País Vasco, Bilbao, Spain*

(Manuscript received 26 March 2004; in final form 26 October 2004)

ABSTRACT

The multi-model ensemble for seasonal to interannual prediction developed in the European Union project DEMETER has been used to quantify the predictability of monthly average maximum temperature that could be achieved operationally over the Iberian Peninsula. Statistical downscaling based on canonical correlation analysis is applied to increase the spatial resolution available from the global models. The downscaling is based on empirical connections between the North Atlantic sea level pressure and monthly average maximum temperature over the Iberian Peninsula.

The maximum temperature estimated from the multi-model ensemble and the single models is compared to the observations. The statistical downscaling model skill is characterized by means of the correlation, variance fraction and the Brier skill score. The results suggest the following: the downscaling model works properly when driven by observed large-scale fields in terms of the correlation and the variance fraction scores, despite some problems owing to sample degeneracy; the predictability is almost limited to February, which is one of the initialization months of the DEMETER ensemble, and it is lost when this month is not considered as starting month. This result is supported by the fact that the areally averaged reproducibility is lower during non-initialization months. In any case, the analysis of the variance test performed reveals that the monthly average maximum temperature is scarcely predictable. Finally, the results also support the advantage of using a multi-model ensemble approach instead of single models participating in DEMETER.

1. Introduction

General circulation models (GCMs) are able to reasonably simulate the behaviour of large-scale atmospheric fields, such as sea level pressure (SLP). They simulate more or less properly low-order zonal-wavenumber features, but they are not as accurate when it comes to the simulation of high-order zonal-wavenumber features. The GCM restrictions in low scales can be partly explained by their limited horizontal resolution (von Storch et al., 1993).

Regional information is requested in many hydrological or ecological studies making it necessary to downscale the GCM's outputs. For this purpose, different downscaling techniques have been developed as tools for interpolating large-scale information into a local or regional scale (Wilby and Wigley, 1997; Zorita and von Storch, 1997; Murphy, 1999; Benestad, 2001). Downscaling techniques are also common tools in operational weather forecasting (Billet et al., 1997) and their use with the DEMETER models is a natural continuation of the DEMETER project. In this paper, statistical downscaling is used to forecast maximum temperature over a mid-latitude region, as is the Iberian Peninsula, with large spatial contrasts of topography which are difficult to include in the GCM parametrizations.

GCM predictions are sensitive to initial conditions and uncertainties in the model parametrization (Boer et al., 1992; Palmer and Anderson, 1994; Barnett, 1999). In fact, the global climate system is not completely predictable in terms of the existing initial conditions. The problem of the sensitivity to initial conditions is usually overcome through the use of an ensemble of predictions starting from different initial conditions (Toth and Kalnay, 1993; Palmer et al., 1998). However, uncertainties also arise from an incomplete representation of reality in the formulation of different models. Thus, different models (even highly truncated ones) starting from identical initial conditions will yield different results after some forecast time (Palmer, 1999). Therefore, in order to obtain results not dependent on the model formulation, a multi-model ensemble system has been developed over the last few years by running global seasonal forecasts in hindcast mode. Several European meteorological centres

*Corresponding author.
e-mail: dfrias@usal.es

Table 1. Models participating in the EU project DEMETER

	Atmosphere	Ocean
CERFACS	ARPEGE (T63, 31 levels)	OPA 8.2 (2.0 × 2.0, 31 levels)
ECMWF	IFS (T95, 40 levels)	HOPE (1.4 × 0.3–1.4, 29 levels)
INGV	ECHAM-4 (T42, 19 levels)	OPA 8.1 (2.0 × 0.5–1.5, 31 levels)
LODYC	IFS (T95, 40 levels)	OPA 8.2 (2.0 × 2.0, 31 levels)
Météo-France	ARPEGE (T63, 31 levels)	OPA 8.1 (182GP × 152GP, 31 levels)
Met-Office	HadCM3 (2.5x3.75, 19 levels)	HadCM3 (1.25 × 0.3–1.25, 40 levels)
MPI	ECHAM-5 (T42, 19 levels)	MPI-OM1 (2.5 × 0.5–2.5, 23 levels)

contribute to the DEMETER European Union project (Development of a European Multi-model Ensemble system for seasonal to inTERannual prediction), which comprises a total of seven global atmosphere–ocean coupled models, each running from an ensemble of nine initial conditions (Palmer et al., 2004). The data used in this paper correspond to the models developed in these European centres (see Table 1 for details): the European Centre for Research and Advanced Training in Scientific Computation, France (CERFACS); Météo-France, Centre National de Recherches Météorologiques, France (CNRM); the European Centre for Medium-Range Weather Forecasts (ECMWF); Istituto Nazionale de Geofisica e Vulcanologia, Italy (INGV); Laboratoire d'Océanographie Dynamique et de Climatologie, France (LODYC); Max-Planck Institut für Meteorologie, Germany (MPI); the UK Meteorological Office (UKMO). The DEMETER hindcasts are started from February, May, August and November initial conditions and each hindcast is integrated for six months. This set of quasi-independent models with different initial conditions allows us to account for the influence of errors in both initial conditions and model parametrizations. In a previous European project, PROVOST (PRediction Of climate Variations On Seasonal Time-scales), the ability of a multi-model ensemble to produce more reliable probability forecasts of seasonal climate anomalies than single-model ensembles was demonstrated (Palmer and Shukla, 2000). The multi-model capability was also obtained in other studies in which data from different models were taken into account (Doblas-Reyes et al., 2000; Stefanova and Krishnamurti, 2002).

We summarize the findings quoted above and state them as follows to confront our problem, which is to analyse the feasibility of performing operational seasonal forecasts on the basis of the results from the DEMETER project. Therefore, our starting hypothesis will be the following.

(i) A multi-model ensemble prediction system, like that used in the DEMETER project, is able to represent to some extent the large-scale features of the climate system and provide usable seasonal forecasts of these large-scale fields (SLP in our case).

(ii) Surface variables – the monthly average maximum temperature (Tmax) in this paper – show a clear relationship with large-scale atmospheric flows around the studied area (the Iberian Peninsula in this paper).

(iii) Amongst the plethora of downscaling techniques available, we selected canonical correlation analysis (CCA) as the tool to provide a proper transfer function from the large-scale field to the regional information needed. Considering that the forecast is aimed at seasonal/monthly values of maximum temperature and that this variable is quite normally distributed, it is appropriate to use linear techniques rather than other non-linear techniques, such as analogues, best suited for other non-normally distributed variables. Previous studies (Zorita and von Storch, 1999; Fernández and Sáenz, 2003) have shown that those non-linear techniques are not significantly better than CCA for monthly values of precipitation. Therefore, their use in the analysis of monthly averaged maximum temperature is not a priori justified in this work.

With these assumptions in mind, the main objective of this paper is to analyse the feasibility of performing operational seasonal forecasts of maximum temperature over the Iberian Peninsula on the basis of the results from the DEMETER project. This feasibility will be measured in terms of a quantitative assessment of the monthly/seasonal predictability of maximum temperature over the Iberian Peninsula by means of the combined strategy of numerical multi-model ensemble seasonal forecasts plus a downscaling technique. The methodology and results of this paper are sensitive to the planning underlying this virtual experiment. We quantify the predictability by applying the downscaling model in hindcast mode. We consider that the statistical downscaling model (SDM) is formulated using observational data alone. Afterwards, the numerical seasonal forecast method (DEMETER in this case) starts and the SDM is run in hindcast mode during the validation period. The comparison of observations and statistically downscaled Tmax yields a measure of predictability. We also explicitly test the advantages of using a downscaling model by comparing downscaled values of Tmax with direct model output. This test is evaluated in terms of correlation and variance fraction skills.

In order to proceed this way, the observational record during the validation period is not used to fit the SDM. We consider that, in order to assess the predictability, having this validation data independent of the formulation is a more stringent approach than alternative formulations, such as cross-validation schemes, usually applied during analysis of the whole observational record (Moron et al., 2001).

The outline of the paper is as follows. The data sets considered for this study are described in Section 2. In Section 3 we present the procedure to downscale maximum temperature, which takes into account the relationships between observed Tmax and SLP based on CCA. This section includes the calibration and validation steps. The results obtained by applying the statistical model to the DEMETER output are presented in Section 4; the estimated values using the multi-model ensemble and single models are evaluated against the observations using different statistical tests. The most significant findings and conclusions are presented in Section 5.

2. Data sets

This work makes use of instrumental, reanalysis and model data from the DEMETER project.

Station data for Tmax were provided by the Meteorological Institutes of Spain and Portugal and correspond to 55 stations evenly distributed over the Iberian Peninsula from 1950 to 2001. Their locations are shown in the right panels of Fig. 1.

Monthly SLP was considered as predictor and it was drawn from three different sources: the National Center for Environmental Prediction/National Center for Atmospheric Research (NCEP/NCAR) Reanalysis (NNR; Kalnay et al., 1996) from 1950 to 2001; the ECMWF ERA-40 from 1980 to 2001 (Simmons and Gibson, 2000; Uppala, 2002); and from the models of the DEMETER project, available from 1980 to 2001 (for brief details on the models, see Table 1). The seven $\times$ nine ensemble members provided by this project were used individually, and also the full set of these models has been employed to construct the multi-model ensemble mean. The DEMETER mean monthly maximum temperature was also considered. These reanalysis and model data are defined on the same regular latitude–longitude grid of resolution $2.5^\circ \times 2.5^\circ$, covering the area 60°W–20°E and 30°–65°N, corresponding to the North Atlantic area (see Fig. 1, left panels).

Both SLP and Tmax data were divided into two periods of time. The first period spans the interval from 1950 to 1979, which is used to train the SDM using SLP data from NNR and observed Tmax data. The second period covers the interval 1980–2001 and is used to validate the downscaling model. It covers the interval when the DEMETER project data are available. In this case, SLP data from both ERA-40 and DEMETER models and Tmax data from observations and DEMETER models are analysed. The ERA-40 data, considered in a different period of time from the NNR data used to formulate the downscaling model, are a proxy to a ‘perfect-prog’ model. This is meant to yield a measure of the best forecast available, limited by the accuracy of the downscaling model alone. The comparison between observed Tmax and those estimated from DEMETER and the downscaling model yields an estimation of the predictability that can be achieved on the basis of state-of-the-art GCMs plus a CCA-based downscaling model.

Monthly SLP and Tmax anomalies were calculated from the corresponding monthly climatology of NNR and instrumental data from the period 1950–1979. These climatologies were also used to compute anomalies during the second period (1980–2001); the NNR climatology was used to compute SLP anomalies for the ERA-40 and DEMETER data and the 1950–1979 observed Tmax climatology for the Tmax data from 1980 to 2001. Consequently, the bias existing between different models and reanalysis will be converted to a temporally constant, spatially variable bias in the downscaled temperature field. Therefore, none of the main validation scores (correlation and variance

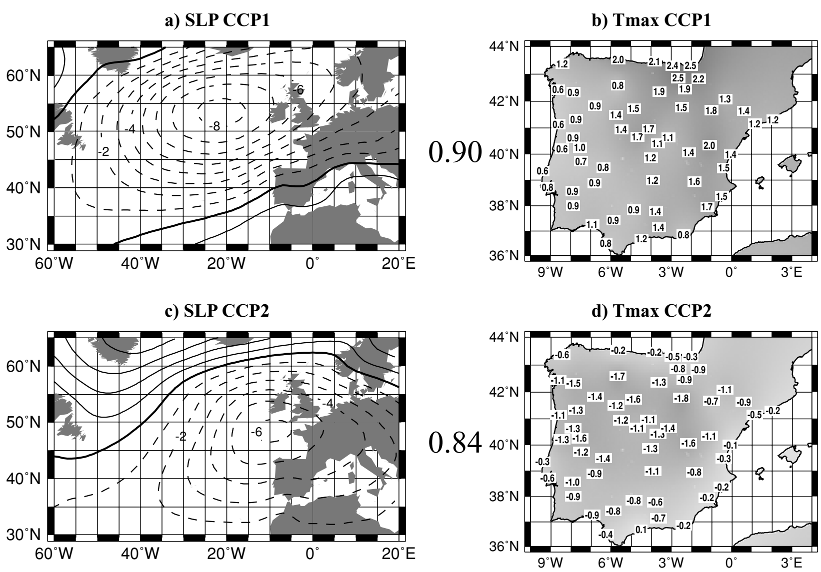

Fig 1. The two leading CCPs corresponding to SLP (left; units are hPa) and monthly average of maximum temperature (right; units are K) in February. The numbers between the maps indicate the canonical correlation corresponding to each pair.

fraction) used in this paper will be affected by this definition of the anomalies.

3. Downscaling model

In this study we have used CCA to obtain the relationship between T_{max} [$T_{max}(t, y)$] over the Iberian Peninsula and SLP [SLP(t, x)] corresponding to the North Atlantic area. As already mentioned in the introduction, Tmax is quite normally distributed. As such, the use of other non-linear techniques, such as analogues (Zorita and von Storch, 1999) or analogues in the CCA space (Fernández and Sáenz, 2003), best suited for non-normally distributed variables, such as precipitation, cannot be justified a priori. Thus, the anomaly fields are expressed as a linear combination of their canonical correlation patterns (CCPs) $p_k(x)$ and $q_k(y)$

$$\mathrm{SLP}(t, x) \approx \sum_k a_k(t) p_k(x) \tag{1}$$

$$T_{\max}(t, y) \approx \sum_k b_k(t) q_k(y), \tag{2}$$

where $a_k(t)$ and $b_k(t)$ are their canonical correlation coefficients (CCCs).

The CCA is not applied directly to Tmax and SLP data. A principal component truncation was made to pre-filter the data. Afterwards, the CCA is applied to uncorrelated variables (the leading principal components) with less spatial noise and still accounting for almost all of the variance of the original field (Barnett and Preisendorfer, 1987).

As a result of the CCA, a_k and b_k are optimally correlated and if SLP is known at a particular time t' (output from a GCM, as in all of the DEMETER runs), an estimation of Tmax is possible using a simple linear regression between both canonical expansion coefficients (von Storch et al., 1993)

$$a_k(t') = \mathrm{SLP}(t', x)\, p_k^{\mathrm{A}}(x) \tag{3}$$

$$T_{\max}(t', y) = \sum_k \rho_k a_k(t') q_k(y), \tag{4}$$

where ρ_k denotes the regression coefficient (which can be proved to be equal to the canonical correlation) and $p_k^{\mathrm{A}}(x)$ are the adjoint canonical patterns (von Storch and Navarra, 1995). This simple statistical method is a good choice in this type of study when there is a strong relationship between large and regional information and the regional scale variable is almost normally distributed. It is also one of the methods most used for statistical downscaling. The ability of this approach to predict Tmax over the Iberian Peninsula is evaluated in the next section.

3.1. Calibration

The CCA method was applied on a monthly basis to the first few principal components obtained from Tmax observations and SLP data from NNR for a 30-yr period (1950–1979). These data were previously detrended on a grid-point basis. The trend is recovered by projecting the SLP field with its original trend in eq. (3). The number of principal components considered to pre-filter each data set was determined by applying the North et al. (1982) test and a Monte Carlo test based on the congruence coefficient (Cheng et al., 1995). In accordance with these tests, two or three modes were selected, depending on the month, for Tmax and three or four modes for the SLP field (see Table 2). The chosen modes are stable, non-degenerated with the rejected ones and they describe more than 70% of the total variance in every case. For instance, in February, four and two modes were considered for SLP and Tmax, respectively, which account for 92% and 88% of the total variance of each field.

CCA was applied to the previously truncated fields. The high canonical correlation obtained for the leading modes justifies the use of this linear downscaling model. In some cases, the figures of canonical correlation yield very low values for the third canonical modes (see Table 2) but these patterns were considered according to the North et al. test. In any case, these pairs of patterns with low values of the canonical correlation ρ_k will have little influence on the downscaled anomalies because it is the weight applied in the downscaling model [see eq. (4)]. The first two CCPs corresponding to SLP and Tmax in February (the best correlated month) are shown as an example in Fig. 1. Figures 1a and 1b show that the maximum influence of the SLP field on the Iberian Peninsula Tmax field occurs in the north-east. For the second pattern, this influence takes place over the central zone of the Peninsula (Figs. 1c and d). Canonical correlations are 0.90 for the first pair and 0.84 for the second. The CCPs resemble the spatial signature of several atmospheric variability modes of the Northern Hemisphere, which were identified in

Table 2. PCA truncation of the original fields, fraction of total variance retained for each field and canonical correlations for each month

	N EOFs SLP,Tmax	Variance(%) SLP,Tmax	CC1	CC2	CC3
Jan	4,2	91,76	0.84	0.37	–
Feb	4,2	92,88	0.90	0.84	–
Mar	3,3	82,90	0.86	0.55	0.34
Apr	3,2	78,72	0.83	0.43	–
May	3,3	75,91	0.77	0.57	0.10
Jun	4,3	84,84	0.88	0.65	0.20
Jul	4,3	84,72	0.75	0.65	0.46
Aug	4,2	83,73	0.75	0.09	–
Sep	4,2	85,86	0.82	0.71	–
Oct	4,3	87,90	0.92	0.77	0.40
Nov	4,2	88,79	0.80	0.53	–
Dec	4,2	90,80	0.79	0.69	–

several previous studies by using different statistical approaches such as correlation analysis of 500-hPa height (hereafter W&G Wallace and Gutzler, 1981), rotated EOF analysis of 700-hPa height (Barnston and Livezey, 1987, hereafter B&L; Bell and Halpert, 1995), the gradient of SLP between Gibraltar and Iceland (Jones et al., 1997) and the North Annular Mode (NAM), which is a planetary scale pattern of climate variability characterized by an out-of-phase relation or seesaw in the strength of the zonal flow along 55°N and 35°N [it is also known as the Arctic Oscillation (AO); Thompson and Wallace, 2001]. In particular, over the Iberian Peninsula the teleconnection indices associated with Tmax are the East Atlantic (EA), North Atlantic Oscillation (NAO) and AO. The indices corresponding to these teleconnection patterns used in this paper are defined as follows. The AO Index is defined as the leading principal component of the SLP field north of 20°N (Thompson and Wallace, 2000). The NAO index is defined as the normalized pressure difference between Gibraltar and south-western Iceland (Jones et al., 1997). The EA index is defined by two different methods. The original method (Wallace and Gutzler, 1981) defines the EA index as

$$\mathrm{EA} = \frac{1}{2} Z^*(55^\circ\mathrm{N}, 20^\circ\mathrm{W}) - \frac{1}{4} Z^*(25^\circ\mathrm{N}, 25^\circ\mathrm{W}) - \frac{1}{4} Z^*(50^\circ\mathrm{N}, 40^\circ\mathrm{E}),$$

where Z^* is the normalized monthly 500-hPa geopotential height anomaly. Finally, a second version of the EA index is also computed from rotated EOF analysis of the 700-hPa geopotential height field and distributed by the Climate Prediction Center (CPC; Bell and Halpert, 1995). Table 3 depicts the correlation coefficients obtained between the CCCs and the teleconnection indices in November, December, January, February, March and April, when the influence of these patterns is higher. Correlation values are significantly lower for other months (not shown) and only values higher than 0.50 are shown in this table in order to make the analysis of this matrix easier. There is evidence that the EA teleconnection patterns, defined by both W&G and B&L, are the most significantly associated with Tmax over the Iberian Peninsula, which agrees with previous results computed during a longer period (49 yr; Sáenz et al., 2001, and references therein). The correlation values between the canonical expansion coefficients and EA index shown in this table are clearly significant at the 95% confidence level, according to a Monte Carlo test, taking into account the reduction of the degrees of freedom in the series owing to temporal autocorrelation. From these results it is possible to investigate the sources of interannual temperature variability over the Iberian Peninsula considering the variability of the transport of sensible heat by mean circulations, without the variability of eddy sensible heat fluxes playing any significant role in this problem (Sáenz et al., 2001). This interpretation in terms of the EA pattern influence is more straightforward and physically based than others that consider the non-linear response to the NAO index (Castro-Díez et al., 2002).

Table 2 depicts the canonical correlation results for every month. The relationship obtained between Tmax and SLP corroborates the appropriate use of SLP to estimate Tmax over the Iberian Peninsula using this CCA-based statistical downscaling model.

3.2. *Validation*

The next issue is to validate the SDM. Two different data sets (the NNR and ERA-40) were considered as predictors to obtain the maximum temperature estimated with the downscaling model.

To estimate the error owing to the downscaling method, the same NNR series used in the formulation of the CCA model (1950–1979) are used as predictor. The series estimated in this way represent the maximum predictability of the model only conditioned by the number of modes selected in the pre-filtering principal component analysis (PCA). Figure 2 shows the correlation coefficients between Tmax observed and Tmax estimated using the downscaling model for the months of an extended winter (November, December, January and February). The stations with a significant correlation at the 95% confidence level are surrounded by a circle. The local significance level at each

Table 3. Correlation coefficients between SLP CCCs for the calibration period and teleconnection indices: EA_W&G (East Atlantic from Wallace & Gutzler); EA_CPC (East Atlantic from Climate Prediction Center); AO (Arctic Oscillation from Climate Prediction Center); NAO_CRU (North Atlantic Oscillation from Climate Research Unit); NAO_CPC (North Atlantic Oscillation from Climate Prediction Center)

	December		January		February		March		April	
	CCC1	CCC2	CCC1	CCC2	CCC1	CCC2	CCC1	CCC2	CCC1	CCC2
EA_W&G	–	−0.82	0.82	–	−0.75	–	0.82	–	0.71	–
EA_CPC	–	0.81	−0.54	–	0.83	–	−0.51	–	−0.58	–
AO	0.56	–	–	–	–	−0.57	–	−0.57	–	0.50
NAO_CRU	0.57	–	–	–	–	0.67	–	−0.48	–	0.67
NAO_CPC	–	–	–	–	–	–	–	−0.64	–	–

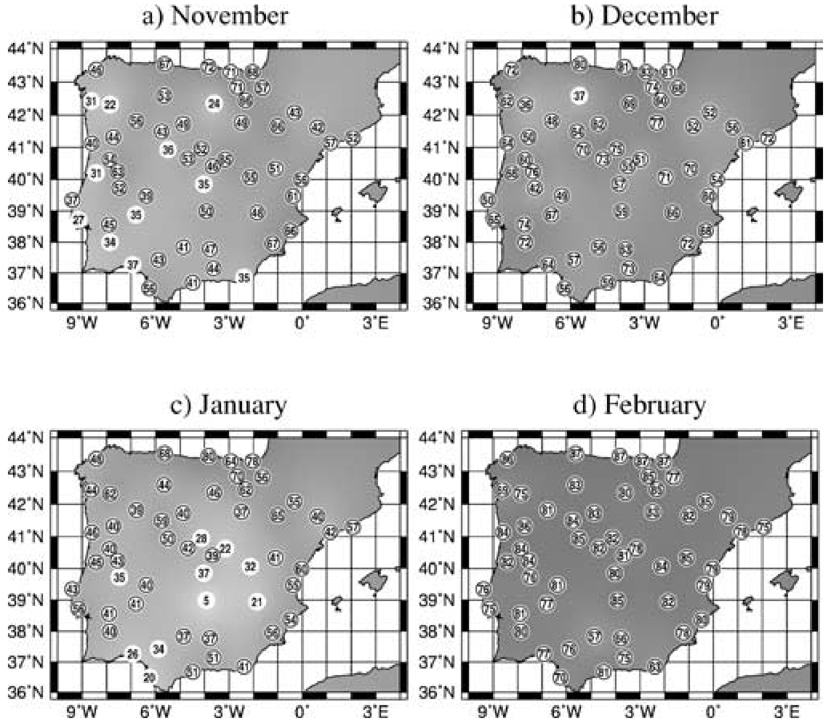

Fig 2. Correlation maps (×100) between observed maximum temperature at each site and the output of the statistical downscaling model using NNR 1950–1979 data as predictor.

site was obtained by means of a Monte Carlo test on the correlation coefficient of synthetic time series taking into account the autocorrelation of each series. For each observatory, the lag-1 autocorrelation is computed and used to generate many samples of an autoregressive AR(1) random time series. The correlation coefficients from these samples are used to create a numerical histogram. The values in this histogram are then used to compute the bounding values, which pass the null hypothesis 'the correlation coefficient is different from 0', depending on the confidence level required (95% in our case). Every map shows significant correlation coefficients but higher values are obtained in December (Fig. 2b) and February (Fig. 2d), where the values are significant at 95% in most cases. The other months present similar values with significant correlations in most of the stations (except March, which shows only 30% of significant correlations). The fact that the downscaling model itself using data from the reanalysis is not able to obtain values significant at 95% in all the observing sites comes from the quite truncated CCA applied to avoid problems derived from overfitting. An analysis of the global field significance (Livezey and Chen, 1983) of these maps is presented in Section 4.

ERA-40 data are taken into account in the initial conditions used to produce the models included in the DEMETER project and, more importantly, they represent an independent data set which would be a 'perfect model'. For this reason, the ERA-40 SLP is used in this part of the study to validate the downscaling model in a different period and with a different data set to that used for calibration. Figure 3 shows the results using the ERA-40 SLP anomalies from 1980 to 2001 as predictor for the extended winter months. For these ERA-40 reanalysis data but for the period of time which is going to be used with the DEMETER models, the values of correlation are lower than those in Fig. 2. In November and January (Figs. 3a and c), the correlation coefficients are lower than in December (Fig. 3b) and February (Fig. 3d), and it is in February when the map shows significant correlation at 95% in most stations of the Iberian Peninsula. We have obtained very good agreement between results derived by using predictor data from ERA-40 and NNR for the period 1980–2001 (not shown). The values shown in Fig. 3 for ERA-40 and those from the NNR data merit some comments regarding the low correlations obtained in several sites in some months between observed and downscaled temperature even when using reanalysis data. We believe that this is due to the fact that the eigenvectors are quite degenerated because the calibration sample is small. In order to show this, a projection of the NNR, ERA-40 and multi-model ensemble data during the validation period is shown against the eigenvalue order for January in Fig. 4. It can be seen that the variance of each projection does not always decrease with an increasing EOF order, as should be expected if the sample were large enough to maintain a low degeneracy.

The results of the statistical model validation for the extended winter shown, and for the rest of the year (not shown), indicate that the statistical model is mostly well designed and that the results obtained are consistent with observations, as expected when reanalysis data are used as large-scale input to the downscaling model.

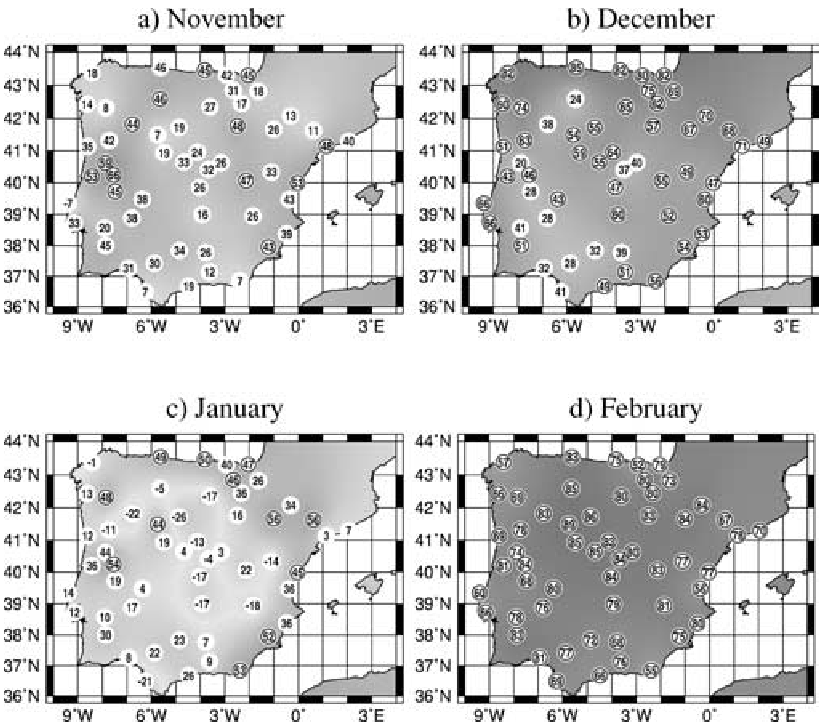

Fig 3. As in Fig. 2, but using ERA-40 1980–2001 data as predictor.

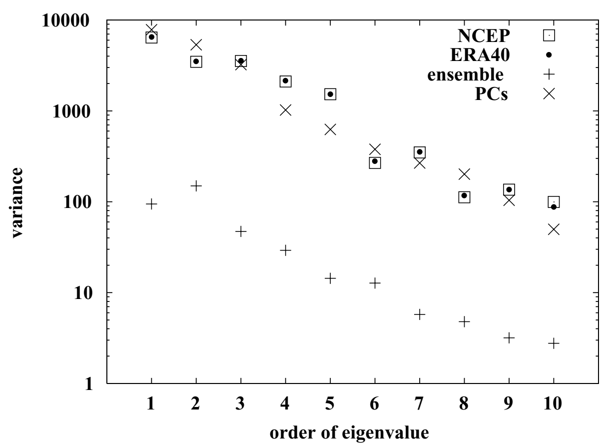

Fig 4. The $\log_{10}\sigma^2$ of the projection of validation data (NNR, ERA-40, multi-model ensemble) and of the NNR principal components for the period from 1950–1979 against order of eigenvalue in January.

4. Forecast based on predictors from DEMETER

In this section a comparison of the statistical downscaling estimates and observed Tmax anomalies is analysed when the models of the DEMETER project are taken as predictors. One of our interests in performing this study was to quantify the advantages of using ensemble predictions instead of single models. Therefore, the results obtained from the multi-model ensemble mean are contrasted with the estimations by the individual models. The analysis was carried out for the period 1980–2001, which is the time made available by all the models in the DEMETER project to date.

Figure 5 shows the spatial distributions of the correlation between observed and estimated Tmax using the multi-model ensemble mean as predictor for each month. The starting months of the DEMETER hindcasts (February, May, August and November) are shown in the first column. The skill varies depending on the month and location. The best correlation values were obtained in February and June (Figs. 5a and e) with higher correlation coefficients towards the west and central part of the Iberian Peninsula. These values are hardly significant in any significant amount of points for all the months, except February, which seems globally predictable over the entire Iberian Peninsula.

To quantify this point to some extent, an approximation to the field significance was tested using the Livezey and Chen (1983) technique to the 95% confidence level. The test was performed for the data sets used for verification (NNR for the calibration period 1950–1979 and ERA-40 for the DEMETER period 1980–2001), for the results downscaled from the DEMETER multi-model ensemble-mean data and for the direct multi-model ensemble-mean Tmax output. Table 4 illustrates the percentage of locally significant correlation coefficients between observed and estimated maximum temperature for each month. The effective number of spatial degrees of freedom is much less than the number of stations owing to the large correlation

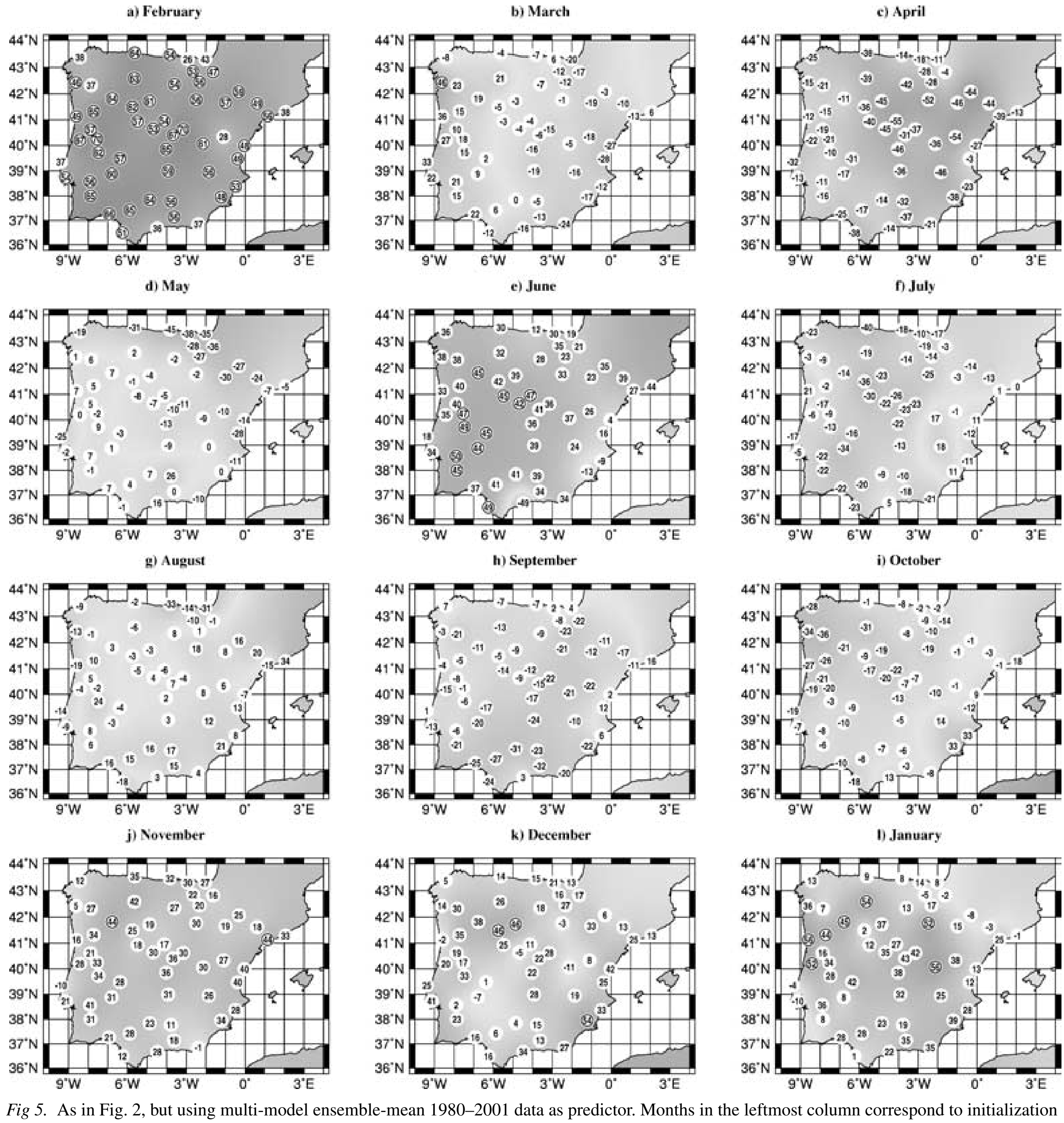

Fig 5. As in Fig. 2, but using multi-model ensemble-mean 1980–2001 data as predictor. Months in the leftmost column correspond to initialization months of the DEMETER runs.

decay distance of the maximum temperature. For this reason, the percentage of successfully passed local significance tests required for global significance rises. The short sample length and the small number of degrees of freedom (around two to three) obtained through standard methods (Bretherton et al., 1999) do not allow a reliable estimation of the spatial degrees of freedom. As an example, Table 4 marks with an asterisk the percentages which are significant to a 95% confidence level considering three degrees of freedom. Conversely, the months not marked with an asterisk do not show enough field significance to pass the test (e.g. April and May using the ERA-40 as predictor outside the calibration period, or all months but February and June when using the data from DEMETER as predictor).

The reduction in the percentage of significant stations in several months when changing from the calibration period to the validation period (ERA-40) is remarkable. This reflects the

Table 4. Percentage of stations significant to 95% for each month. Field significant values on a 95% after Livezey and Chen (1983) are marked with an asterisk. Results for the two SDM input data sets used for verification (NNR and ERA-40), the results from the DEMETER multi-model ensemble-mean data (DEMETER) and the direct multi-model ensemble-mean Tmax output (DEMETER direct output) are shown

	NNR + SDM	ERA-40 + SDM	DEMETER + SDM	DEMETER direct output
Jan	83*	20*	12	20*
Feb	100*	100*	83*	89*
Mar	30*	45*	1	0
Apr	74*	14	0	0
May	90*	5	0	41*
Jun	98*	87*	18*	0
Jul	83*	36*	0	27*
Aug	70*	41*	0	18*
Sep	100*	72*	0	1
Oct	100*	98*	0	5
Nov	80*	21*	1	9
Dec	98*	74*	5	0

effect of the SDM, while the step from ERA-40 to DEMETER shows the reduction owing to the use of the ensemble prediction system, where, as was already visually evident, only February maintains a reasonable skill.

ANOVA methods (Wilks, 1995) also confirm the previous results. A simple linear regression between the estimated and observed maximum temperature significantly explains the observed maximum temperature at most of the stations at the 95% level for the NCEP and ERA-40 predictors. However, the predictability for the DEMETER predictors is limited to February and, to a much lesser extent, June. For the rest of the months, the variability of unpredictable noise is much greater than the variability of the signal.

The lack of predictability of all months but February may be because the relationship between Tmax and large-scale SLP patterns is lower during these months and, also, because it reflects a seasonal dependence suggested by many authors (Halpert and Ropelewsky, 1992; Wilby, 1993; Palmer and Anderson, 1994) and confirmed to reach maximum predictability during winter and spring seasons in the framework of the PROVOST project multi-model ensemble (Pavan and Doblas-Reyes, 2000). The good results obtained in February are due to two reasons. First, it is the month best represented by the downscaling model, as shown in Figs. 2d and 3d and the canonical correlations on Table 2. Secondly, it is one of the months in which the DEMETER hindcasts start from initial conditions. The influence of the starting date can be observed by comparing Figs. 5a, d, g, j and 6a–d where February, May, August and November data from the hindcasts starting in November, February, May and August are considered. That is, in Fig. 6, the results are computed with model runs started four months earlier. There are no points showing correlation values significant at a 95% confidence level. This means, basically, that long-lead prediction is quite limited over the area. The lack of predictability beyond the one-month threshold is directly inherited from the lack of predictability of the large-scale predictor by the GCMs at these latitudes. The agreement among the different GCM estimates of SLP over the Atlantic can be measured through the reproducibility (Yang et al., 1998; Hassan et al., 2004). The reproducibility is a measure of the strength of the signal contained in the ensemble. A null reproducibility is related to a full disagreement among the ensemble members and, thus, a signal with no common variability. The maximum reproducibility (1) is reached when all members agree and there is no internal variability of the ensemble; the experiment is fully reproducible. The areally averaged reproducibility of the SLP field entering the downscaling model is computed separately for each month. Figure 7 reveals, as expected, that the starting months are the most reproducible, with decaying reproducibility as the forecast time is increased. February shows the highest values, although 0.18 is still a low value. Even in February, the ensemble of GCM simulations shows a great variability in the estimation of the predictor on the Atlantic area.

In order to discuss the improvement of the multi-model ensemble with respect to the individual models of the DEMETER project, statistical downscaling is applied to predictors selected from the individual models. We obtained that the correlation coefficients between estimated and observed Tmax depend on the month, the model and the initial condition. In general, the correlation is higher in February for most of the models and initial conditions; hence, the comparison is focused on this month. Figure 8 shows the correlation maps between observed and estimated Tmax using the multi-model ensemble mean (Fig. 8a) and the different individual models (Figs. 8b–h) as predictors in February. For each of the different model runs provided by the DEMETER project, the estimated series in Fig. 8 correspond to the mean of the series estimated for every initial condition of one single model. The figure shows that the correlation coefficients are significant at 95% in nearly every station of the different maps corresponding to most of the individual models and the multi-model ensemble mean. Some exceptions correspond to the INGV (Fig. 8d) and the MaxP (Fig. 8h) models which show no significant correlation coefficients. The CERFACS model works better than the multi-model ensemble for this particular month in terms of the correlation skill.

Although the results are not shown for the other months, we obtained larger differences among models and initial conditions, which reveals that the estimated Tmax using the proposed statistical downscaling technique is very sensitive to the model formulation and the initial conditions used in each model run.

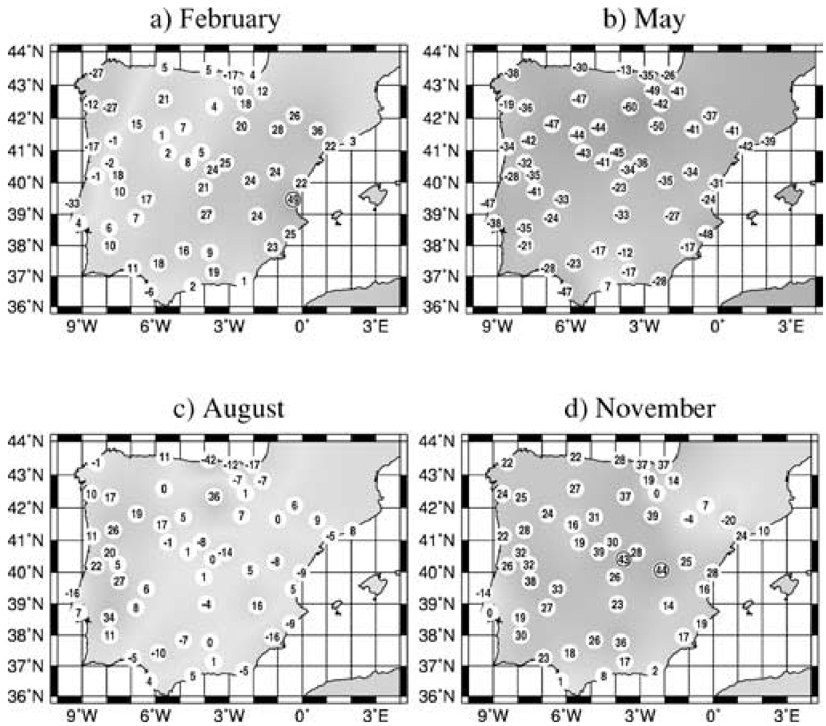

Fig 6. Correlation maps between observed maximum temperature and statistical model output for February, May, August and November using the multi-model ensemble mean as predictor with the starting month four months earlier, in November, February, May and August, respectively.

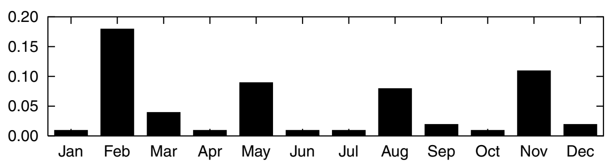

Fig 7. Reproducibility of the SLP as defined by Yang et al. (1998). For each month, the reproducibility averaged over the Atlantic region (60°W–20°E/30°N–65°N) is shown.

Although the multi-model ensemble enhances prediction capabilities with respect to single models, there are some disadvantages owing to the variability reduction as a consequence of averaging models and ensemble members. Figure 9 shows the variance ratio of estimated to observed maximum temperature for the multi-model ensemble mean and the single-model ensemble means for February. There is an underestimation of variance in the case of the multi-model ensemble mean (Fig. 9a) in relation to the observations, especially in the southern part of the Iberian Peninsula and on the Mediterranean coast, as a consequence of the averaging of models with anomalies that almost cancel each other out to obtain the predictor field in the SDM. The corresponding maps for the individual DEMETER models (Figs. 9b–h) also reflect variance reduction but the factor is lower than that corresponding to the multi-model ensemble, owing to the smaller number of elements in the single-model ensembles. Averaging (Figs. 9b–h) causes an underestimation of the Tmax observed variance, especially in the south and on the Mediterranean coast. However, in many circumstances, when one initial condition is considered in a single model, the results indicate variance overestimation with respect to the Tmax observed variance, particularly in the north-west of the Iberian Peninsula. It seems initially strange that a SDM might overestimate the variance when not averaging the members of the ensemble. The reason is that some models create stronger anomalies than those in the observations. Therefore, they are able to produce [through eq. (3)] time expansion coefficients with variances greater than in the case of observed data.

Some of the stations in the interior of the Iberian Peninsula agree in variance ratio and fit better the variability of the observations in February. Figure 10 shows the performance of the estimated Tmax series for the multi-model ensemble mean in contrast with the observations for three stations located in different parts of the Iberian Peninsula: Valladolid, located in the central-west zone of the Iberian Peninsula; Sevilla, in the south; Castellón, in the east. The box and whiskers represent the DEMETER models spread, where the whiskers show the minimum and maximum estimated values and the boxes the lower and upper quartiles. For clarity, only a 4-yr period from 1989 to 1992 is shown. In general, the estimated Tmax using the multi-model ensemble mean underestimates the variability of the observations. Nevertheless, the values are closer in the winter months, where the observed variability is better represented. The spread of the individual models is smaller in stations located along the Mediterranean coast (e.g. Castellón, in Fig. 10). However, in general, the results obtained with individual models tend to overestimate or underestimate the real value of maximum temperature and those results corresponding to the multi-model mean more

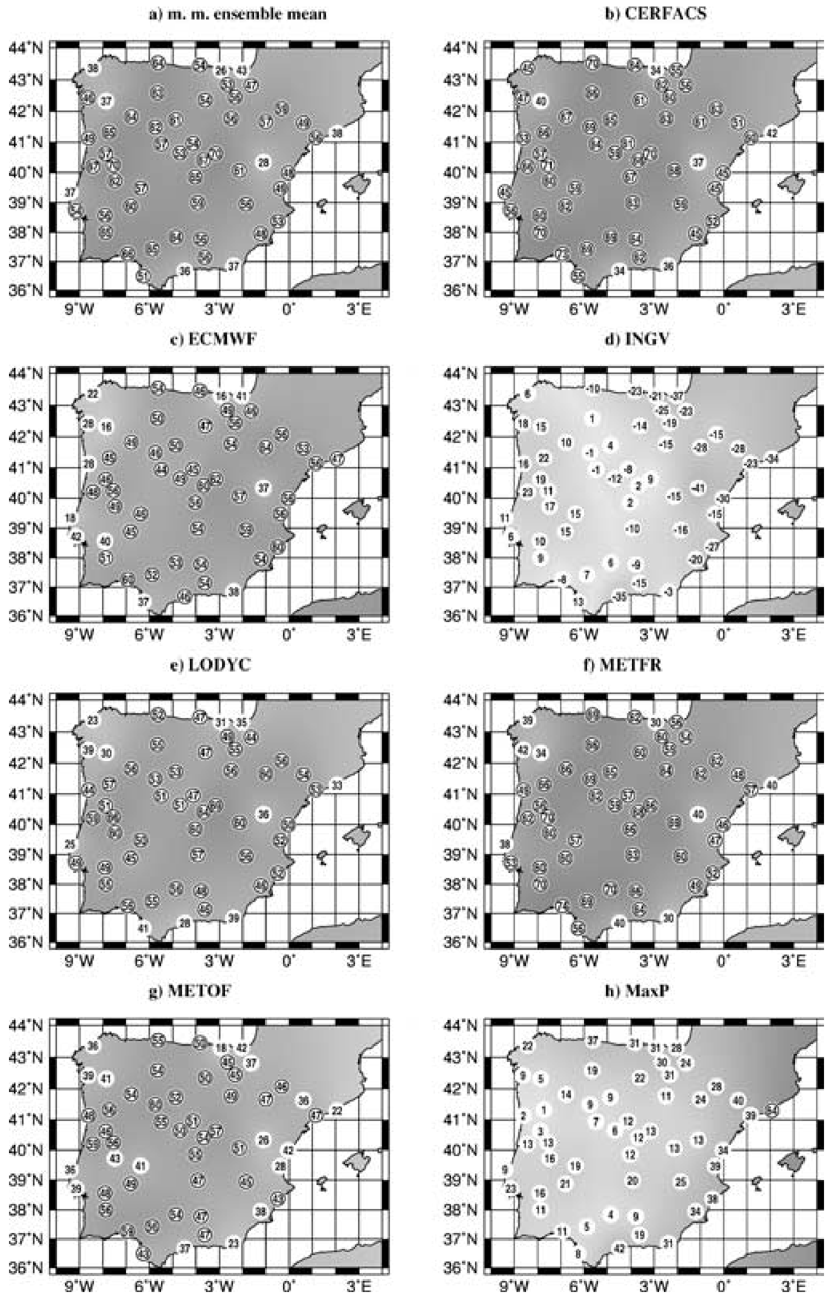

Fig 8. Spatial distribution of correlation ($\times 100$) between observed and estimated maximum temperature in February using multi-model ensemble mean and DEMETER models as predictors: (a) multi-model ensemble mean; (b) CERFACS; (c) ECMWF; (d) INGV; (e) LODYC; (f) METFR (Météo-France); (g) METOF (UKMO); (h) MaxP (MPI).

closely approximate the observations in spite of the underestimation of the observed variability.

In order to see if the use of a downscaling model leads to a real improvement in results over direct model output of monthly average Tmax from the DEMETER runs, a bilinear interpolation was applied to the direct gridded model output data in order to compute Tmax at the observing sites. Then, a comparison of observed monthly average Tmax and interpolated direct model output fields was carried out. The influence of the starting month previously shown in the downscaled results is higher on results from the direct Tmax outputs from the DEMETER project. For these starting months, the correlation values between models and multi-model ensemble-mean Tmax and the observed data are higher than those obtained with the downscaling model, especially in February for all cases and in May for the multi-model ensemble mean and in most of the DEMETER models. The percentage of significant stations using the multi-model ensemble mean (Table 4) is globally significant for January, February (starting month), May (starting month), July and August (starting month). The globally significant correlation obtained in June with the SDM is not significant in the direct output, though. For the rest of the year, the correlation values

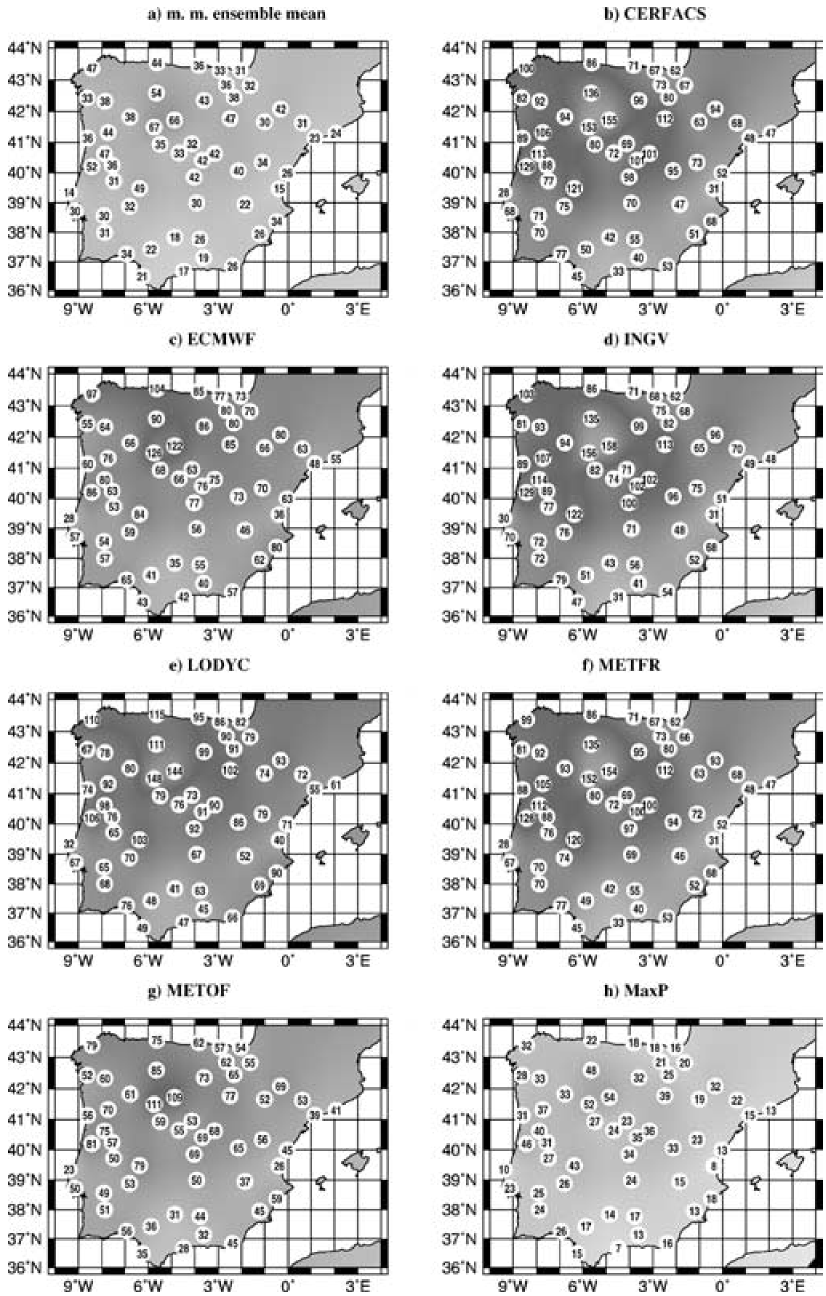

Fig 9. Ratio of estimated and observed maximum temperature variances ($\times 100$) in February using multi-model ensemble-mean and DEMETER models as predictor: (a) multi-model ensemble mean; (b) CERFACS; (c) ECMWF; (d) INGV; (e) LODYC; (f) METFR; (g) METOF; (h) MaxP.

are low and not significant. However, the variance ratio of the interpolated Tmax outputs from the DEMETER project to the variance of the observations is lower than those estimated from the downscaling model for every month. Therefore, according to one of the skill scores used (correlation), direct model output performs better than statistically downscaled data for the starting months. However, this is not true for the second skill score (variance fraction), where statistically downscaled data outperforms direct model output.

The simplest seasonal forecasting system that can be arranged is that based on the seasonal cycle of any climatological element. The skill of the estimated maximum temperature series with the multi-model ensemble mean and the individual models by DEMETER is analysed using a common measure of forecast verification, the Brier skill score (BSS) relating to a forecast based on climatology (Wilks, 1995). First developed by Brier (1950), it is widely used in the evaluation of ensemble forecasts with a single model (Molteni et al., 1996) and with a multi-model ensemble (Palmer et al., 2000; Stefanova and Krishnamurti, 2002). The BSS assesses the probability of occurrence of a chosen event, in our case, that the predicted Tmax is a fraction of the standard deviation over or under the climatology. A

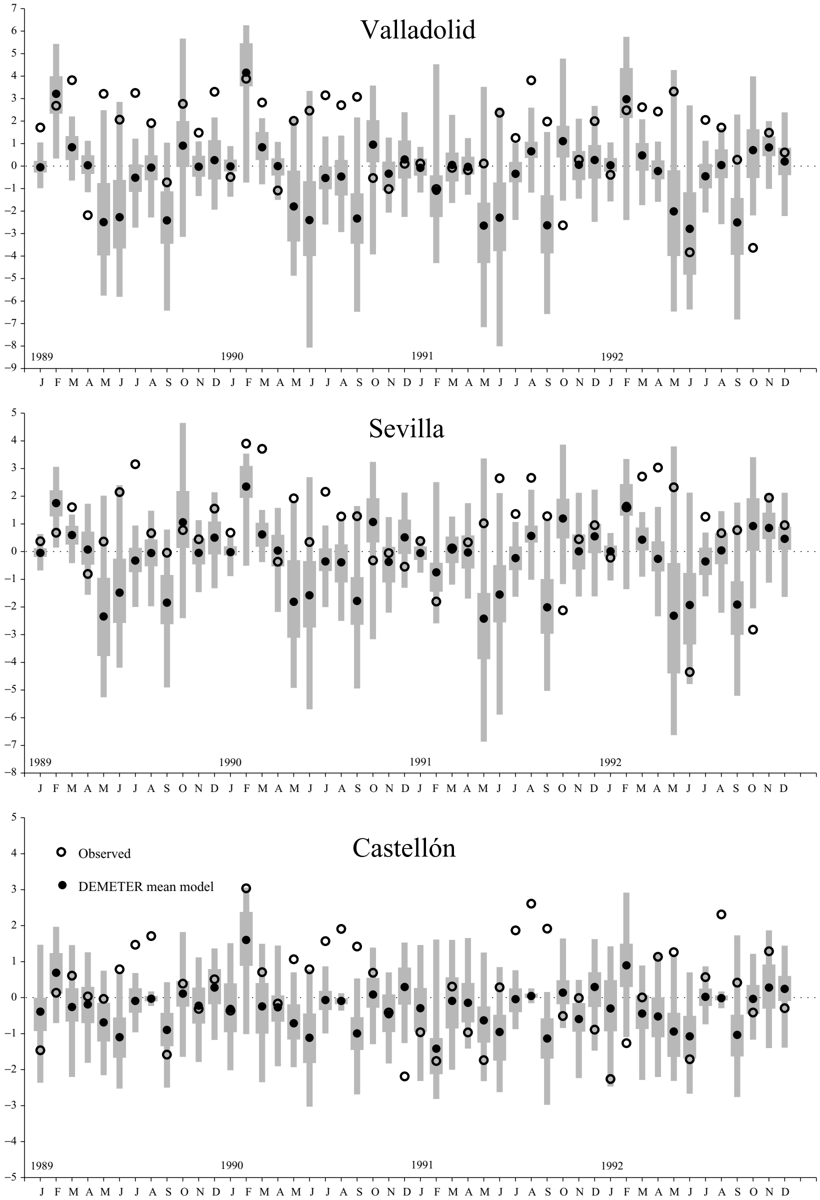

Fig 10. Values of observed and estimated maximum temperature using multi-model ensemble mean and DEMETER models as predictor from 1989 to 1992: (a) Valladolid; (b) Sevilla; (c) Castellón.

value lower than or equal to zero indicates a forecast worse than the reference forecast, in our case the climatology, and a BSS larger than zero indicates the opposite, with a value of 1 corresponding to a perfect forecast. Table 5 shows the mean of the BSS obtained from each station for the estimated series using the multi-model ensemble mean and the individual models for six events (anomaly above 1, 0.6 and 0.4 standard deviation and below −0.4, −0.6 and −1 standard deviation) in February, which was the best predicted month. As the validation anomaly time series are centred on the calibration mean, a possible bias could affect the BSS results. Therefore, the series were centred to calculate the BSS in the estimate of the probability of these events for each station. According to BSS values, the models shown in Table 5 forecast better than the climatology the events below −0.4 and −0.6 standard deviation where the BSS is higher than zero. The values are different depending on the model and the event, but, in general, the multi-model ensemble mean presents positive values especially in the most extreme events, and INGV and MaxP are the models which present the lowest BSS values, with forecasts being worse than those derived from other models in most cases. If we analyse the BSS in different parts of the Iberian Peninsula (not shown), the multi-model ensemble and the individual models show positive and higher values in the central and western parts of the Iberian Peninsula and worse forecasts than climatology on the Mediterranean coast. This is in agreement with the values shown in Fig. 8, where the correlations are higher in these central and western zones and, in general, lower on the Mediterranean coast. This probably reflects the fact

Table 5. Values of BSS for the multi-model ensemble mean and DEMETER models in February for six events (anomaly exceeding 1.0σ and 0.4σ and anomaly below -1.0σ, -0.6σ and -0.4σ)

	$>1.0\sigma$	$>0.6\sigma$	$>0.4\sigma$	$<-0.4\sigma$	$<-0.6\sigma$	$<-1.0\sigma$
Ensemble	0.10	−0.20	−0.14	0.52	0.41	0.14
CERFACS	0.03	−0.12	0.08	0.54	0.44	−0.19
ECMWF	−0.19	−0.26	−0.04	0.54	0.43	−0.06
INGV	−1.10	−1.13	−0.90	0.13	0.10	−0.46
LODYC	−0.03	−0.05	−0.14	0.37	0.28	−0.11
METFR	0.34	0.11	0.16	0.45	0.40	−0.18
METOF	−0.25	−0.17	−0.07	0.36	0.34	−0.09
MaxP	−1.02	−0.62	−0.30	0.34	0.21	−0.40

that Tmax over some areas, such as the Mediterranean, are more dependent on local factors than on the large-scale atmospheric patterns.

5. Conclusions

A quantitative assessment of the predictability of monthly/seasonal maximum temperature over the Iberian Peninsula considering the combination of a multi-model ensemble of coupled atmosphere–ocean GCMs and a CCA-based downscaling model was carried out in hindcast mode for the period 1980–2001.

The results show that the downscaling model works well in general, despite some errors which can be traced to sample degeneracy owing to the small sample length available. Moreover, there is still a part of the maximum temperature linked to unpredictable noise and other possible independent predictors, as revealed by the ANOVA performed.

In general, the application of the downscaling model using the DEMETER simulations indicates that predictability is a function of observation site and month. There are several reasons that can explain this fact. First, observed (reanalysis) and DEMETER ensemble-mean main modes of variability do not coincide. Therefore, when DEMETER anomalies are projected on to the corresponding spatial adjoint patterns, the results do not reflect sufficiently high canonical expansion coefficients. It follows that these coefficients are not always valid to reconstruct observed anomalies. Secondly, during some parts of the year, the relationship between large-scale and regional information is weaker than in others. This makes the downscaling model itself perform very differently in different months.

In any case, for the Iberian Peninsula, quantitatively valuable predictability is very small and is limited to the initialization months, especially in February, or, by extension, to the winter season when the results are analysed by seasons.

Variance of the predicted series using model ensemble is always much lower than observed, and this can be explained by two reasons. First, the downscaling method itself produces a loss in the variance fraction as a consequence of the PCA truncation, which cannot be recovered, even with a perfect model, unless the model simulates too high anomalies (therefore, it would not be perfect). Secondly, the ensemble nature of the prediction, when averaging its different members, systematically produces anomalies in the SLP (predictor) field much weaker than observed ones. Therefore, in the analysis of the downscaled data and their relationship with observed anomalies, this loss of variance must be explicitly taken into account as being due to the averaging process of the members of the ensemble.

Single-model ensembles were also analysed. Correlation results are slightly lower than for the multi-model ensemble, except for the INGV and MaxP models, which show poorer agreement with observations than the rest of the models.

In general, the Tmax series estimated with the multi-model ensemble mean underestimate the observations, but in winter months the forecasts better represent the observed variability, especially in stations located in the interior of the Iberian Peninsula. With individual models, the approximation to the observations is lower and tends to overestimate or underestimate the observations because the spread, in general, is high.

The direct DEMETER outputs show more influence of the starting month in the correlation values with the observed data, but they are not able to reproduce the real variability, showing variance ratio values even lower than those from the SDM estimated values. This is probably due to the inability of the coarse-resolution GCMs to capture regional variability details.

The comparison of the DEMETER multi-model ensemble prediction system plus a CCA-based downscaling model with a seasonal prediction system based on the climatology shows that predictability is higher than the climatology in events below −0.4 and −0.6 standard deviation. It is the same for most of the DEMETER models except for INGV and MaxP models, which again show worse results. In general, the predictability of extreme events is higher in the western and central parts of the Iberian Peninsula and lower on the Mediterranean coast.

6. Acknowledgments

The authors thank the data providers for this study (ECMWF, NCEP/NCAR and the National Meteorological Institutes of Spain and Portugal) and are grateful for the financial support provided by projects REN2002-04584-C04-04-CLI and REN2002-04584-CO4-02-CLI, funded by the Spanish Ministry of Science and Technology. CR-P and MDF are grateful for support from the Regional Government of Castile and Leon through project SA027/02. JF and JS are grateful for grant 9/UPV 00060.310-15343/2003, from the University of the Basque Country. MDF has a grant from the Ministry of Science and Technology and JF has a grant from the Department of Education, Universities and Research of the Basque Government. The authors thank E. Zorita and F. J. Doblas for their helpful comments and suggestions. The authors are grateful to the two anonymous reviewers for their comments, which improved the manuscript, and to the SCI of the University of Salamanca for the English revision.

References

Barnett, T. P. 1999. Comparison of near-surface air temperature variability in 11 coupled global climate models. *J. Climate* **12**, 511–518.

Barnett, T. P. and Preisendorfer, R. W. 1987. Origins and levels of monthly and seasonal forecast skill for United States air temperature determined by canonical correlation analysis. *Mon. Wea. Rev.* **115**, 1825–1850.

Barnston, A. G. and Livezey, R. E. 1987. Classification, seasonality and persistence of low-frequency circulation patterns. *Mon. Wea. Rev.* **115**, 1083–1126.

Bell, G. D. and Halpert, M. S. 1995. *NOAA Atlas No. 12. Interseasonal and Interannual Variability: 1986 to 1993*. Climate Analysis Center. U.S. National Oceanic and Atmospheric Administration.

Benestad, R. E. 2001. A comparison between two empirical downscaling strategies. *Int. J. Climatol.* **21**, 1645–1668.

Billet, J., DeLisi, M. and Smith, B. G. 1997. Use of regression techniques to predict hail size and the probability of large hail. *Wea. Forecasting* **12**, 154–164.

Boer, G. J., Arpe, K., Blackburn, M., Déqué, M., Gates, W. L. et al. 1992. Some results from an intercomparison of the climates simulated by 14 atmospheric general circulation models. *J. Geophys. Res.* **97**, 12 771–12 786.

Bretherton, C. S., Widmann, M., Dymnikov, V. P., Wallace, J. M. and Bladé, I. 1999. The effective number of spatial degrees of freedom of a time-varying field. *J. Climate* **12**, 1990–2009.

Brier, G. W. 1950. Verification of forecasts expressed in terms of probabilities. *Mon. Wea. Rev.* **78**, 1–3.

Castro-Díez, Y., Pozo-Vázquez, D., Rodrigo, F. S. and Esteban-Parra, M. J. 2002. NAO and temperature variability in southern Europe. *Geophys. Res. Lett.* **29**, doi:10.1029/2001GL014042.

Cheng, X., Nitsche, G. and Wallace, J. M. 1995. Robustness of low-frequency circulation patterns derived from EOF and rotated EOF analyses. *J. Climate* **8**, 1709–1713.

Doblas-Reyes, F. J., Déqué, M. and Piedelievre, J. P. 2000. Multi-model spread and probabilistic seasonal forecast in PROVOST. *Q. J. R. Meteorol. Soc.* **126**, 2069–2087.

Fernández, J. and Sáenz, J. 2003. Improved field reconstruction with the analog method: searching the CCA space. *Clim. Res.* **24**, 199–213.

Halpert, M. S. and Ropelewsky, C. F. 1992. Surface temperature patterns associated with the Southern Oscillation. *J. Climate* **5**, 577–593.

Hassan, A. S., Yang, X.-Q. and Zao, S.-S. 2004. Reproducibility of seasonal ensemble integrations with ECMWF GCM and its association with ENSO. *Meteorol. Atmos. Phys.* **86**, 159–172.

Jones, P. D., Jónsson, T. and Wheeler, D. 1997. Extension to the North Atlantic Oscillation using early instrumental pressure observations from Gibraltar and South-West Iceland. *Int. J. Climatol.* **17**, 1433–1450.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D. et al. 1996. The NCEP/NCAR 40-year Reanalysis Project. *Bull. Am. Meteorol. Soc.* **77**, 437–471.

Livezey, R. E. and Chen, W. Y. 1983. Statistical field significance and its determination by Monte Carlo techniques. *Mon. Wea. Rev.* **111**, 46–59.

Molteni, F., Buizza, R., Palmer, T. and Petroliagis, T. 1996. The ECMWF ensemble prediction system: Methodology and validation. *Q. J. R. Meteorol. Soc.* **122**, 73–119.

Moron, V., Ward, M. N. and Navarra, A. 2001. Observed and SST-forced seasonal rainfall variability across tropical America. *Int. J. Climatol.* **21**, 1467–1501.

Murphy, J. 1999. An evaluation of statistical and dynamical techniques for downscaling local climate. *J. Climate* **12**, 2256–2284.

North, G., Bell, T., Cahalan, R. and Moeng, F. 1982. Sampling errors in the estimation of empirical orthogonal functions. *Mon. Wea. Rev.* **110**, 699–706.

Palmer, T. N. 1999. A nonlinear dynamical perspective on climate prediction. *J. Climate* **12**, 575–591.

Palmer, T. N. and Anderson, D. L. T. 1994. The prospects for seasonal forecasting-a review paper. *Q. J. R. Meteorol. Soc.* **120**, 755–793.

Palmer, T. N. and Shukla, J. 2000. Editorial to DSP/PROVOST special issue. *Q. J. R. Meteorol. Soc.* **126**, 1989–1990.

Palmer, T. N., Gelaro, R., Barkmeijer, J. and Buizza, R. 1998. Singular vectors, metrics and adaptive observations. *J. Atmos. Sci.* **55**, 633–653.

Palmer, T. N., Brankivic, C. and Richardson, D. 2000. A probability and decision-model analysis of PROVOST seasonal multi-model ensemble integrations. *Q. J. R. Meteorol. Soc.* **126**, 2013–2033.

Palmer, T. N., Alessandri, A., Andersen, U., Cantelaube, P., Daveyand, M. et al. 2004. Development of a European multi-model ensemble system for seasonal-to-interannual prediction DEMETER. *Bull. Am. Meteorol. Soc.* **85**, 853–872.

Pavan, V. and Doblas-Reyes, F. J. 2000. Multi-model seasonal hindcasts over the Euro-Atlantic: skill scores and dynamic features. *Climate Dyn.* **16**, 611–625.

Sáenz, J., Rodríguez-Puebla, C., Fernández, J. and Zubillaga, J. 2001. Interpretation of interannual winter temperature variations over southwestern Europe. *J. Geophys. Res.* **106**, 20 641–20 652.

Simmons, A. J. and Gibson, J. K. 2000. The ERA-40 project plan. *ERA-40 Project Report Series 1*, ECMWF, Reading.

Stefanova, L. and Krishnamurti, T. N. 2002. Interpretation of seasonall climate forecast using Brier skill score, the Florida State University Superensemble, and the AMIP-I data set. *J. Climate* **15**, 537–544.

Thompson, D. W. J. and Wallace, J. M. 2000. Annular modes in the extratropical circulation. Part I: month-to-month variability. *J. Climate* 1000–1016.

Thompson, D. W. J. and Wallace, J. M. 2001. Regional climate impacts of the northern hemisphere annular modes. *Science* **293**, 85–89.

Toth, Z. and Kalnay, E. 1993. Ensemble forecasting at NMC: the generation of perturbations. *Bull. Am. Meteorol. Soc.* **74**, 2317–2330.

Uppala, S. 2002. ECMWF reanalysis 1957–2001, ERA-40. *ERA-40 Project Report Series 3*, ECMWF, Reading.

von Storch, H. and Navarra, A. eds. 1995. *Analysis of Climate Variability. Applications of Statistical Techniques*. Springer-Verlag, Berlin.

von Storch, H., Zorita, E. and Cubasch, U. 1993. Downscaling of global climate change estimates to regional scales: an application to Iberian rainfall in wintertime. *J. Climate* **6**, 1161–1171.

Wallace, J. M. and Gutzler, D. S. 1981. Teleconnections in the geopotential height field during the Northern Hemisphere winter. *Mon. Wea. Rev.* **109**, 784–812.

Wilby, R. L. 1993. Evidence of ENSO in the synoptic climate of the British Isles since 1880. *Weather* **48**, 234–239.

Wilby, R. L. and Wigley, T. M. L. 1997. Downscaling general circulation model output: a review of methods and limitations. *Prog. Phys. Geography* **21**4, 530–548.

Wilks, D. S. 1995. *Statistical Methods in the Atmospheric Sciences*. Academic Press, San Diego.

Yang, X.-Q., Anderson, J. L. and Stern, W. F. 1998. Reproducible forced modes in AGCM ensemble integrations and potential predictability of atmospheric seasonal variations in the extratropics. *J. Climate* **11**, 2942–2959.

Zorita, E. and von Storch, H. 1997. A survey of statistical downscaling techniques. *GKSS 97/E/20*.

Zorita, E. and von Storch, H. 1999. The analog method as a simple statistical downscaling technique: Comparison with more complicated methods. *J. Climate* **12**, 2474–2489.