Vol. 24: 199–213, 2003 | CLIMATE RESEARCH Clim Res | Published September 19

Improved field reconstruction with the analog method: searching the CCA space

Jesús Fernández*, Jon Sáenz

Depto. Física Aplicada II, Universidad del País Vasco, Apdo. 644, 48080 Bilbao, Spain

ABSTRACT: The analog based downscaling method is revisited in an application to precipitation data for the northern coast of the Iberian Peninsula. Analog situations of a large-scale predictor are searched in the historical record and the regional-scale predictand is reconstructed by using the analog records found. The usual approach is insensitive to the predictand variable, and we present a new approach using Canonical Correlation Analysis (CCA), which consists in projecting the predictor field onto the spatial patterns obtained in a CCA between the predictor and predictand variables and searching for analogies in this dimension-reduced predictor. This approach is tested against the usual analog search, based on the projection onto the patterns derived from Principal Component Analysis (PCA), and the more commonly used linear CCA downscaling technique. In a projection space of the same dimension, the new approach performs better in reconstructing the precipitation (based on correlation and variance skill scores) than the PCA approach. The CCA linear method yields a similar correlation skill by comparison to our new approach, but reconstructs a much lower fraction of the variance. The non-normality of the probability density function inherent to the precipitation data is partly lost by the linear method, whereas it is preserved by the analog methods. A sensitivity analysis on several parameters of the analog search was also conducted. The improvement of the CCA approach over analogs seems to be related to the identification in the predictor field of the areas most closely connected to the predictand.

KEY WORDS: Analogs · CCA · Downscaling · Precipitation · Cantabrian Coast

1. INTRODUCTION

Increasing interest has been devoted in the last decade to the downscaling of the information offered by Global Circulation Models (GCMs) for use on regional scales. The main reason for this effort is the lack of reliability of GCMs in representing regional climates (von Storch et al. 1993). Two general approaches have been developed to downscale the large-scale information: (1) A dynamical approach consisting of a physical interpolation of the GCM coarse grid to a finer grid; this is usually accomplished by Limited Area Models, which are nested into the GCM and obtain the required boundary conditions from it (Giorgi & Mearns 1991). (2) A statistical approach, in which an empirical relationship between a large-scale variable and a variable affected by local factors is used to obtain reliable small-scale information; a wide variety of statistical techniques have been used for this purpose.

There is a large number of papers comparing different empirical downscaling techniques (e.g. Biau et al. 1999, Huth 1999, Zorita & von Storch 1999, Tang et al. 2000). There are also some comparisons of the dynamical and empirical approaches (Wilby & Wigley 1997, Kidson & Thompson 1998, Murphy 1999). The main advantage of the statistical approaches is their low computational cost, while their weak point is that they assume that the underlying physical relationship between the variables is stationary, a doubtful assumption in an altered climate. They also require a strong statistical relationship between the variables. On the other hand, the dynamical approach is designed to reproduce altered climates by means of simplified physical equations but, in the smaller scales, still requires empirical parameterizations (keeping them subject to the problem of the stationarity). The computational effort of this approach is much greater. Empirical downscaling techniques can be classified as linear

*Email: chus@wm.lc.ehu.es

(e.g. linear regression and CCA) or non-linear (e.g. analog based techniques, neural networks and weather typing methods), allowing different degrees of complexity.

The analog search method is straightforward in reproducing non-linear relationships between the variables. Its computational costs are low compared with other non-linear techniques such as neural networks, which need to be iteratively trained. In short-term forecast applications (Lorenz 1969, Ruosteenoja 1988, van den Dool 1989) analogs are searched in the past history of the predictand variable and the forecast relies on the similar evolution of the present situation and the past analog one. Several ways for searching analogs can be applied (Barnett & Preisendorfer 1978, Toth 1991). The easiest one is a simple search of a single analog in a pool of historic cases. It can be improved, e.g. by searching for similarities in the present situation and in the evolution which gave rise to it. But, even reducing the number of degrees of freedom (d.f.) by characterizing the atmosphere by its flow at a single level, the analogs found are mediocre, and they do not perform better than a persistence model. This is why analog search was rejected long ago as a model for short-range weather forecast. For downscaling purposes the analog scheme must be changed to include local-scale variables, as applied by Dehn (1999), Zorita & von Storch (1999), Timbal & McAvaney (2001), Timbal et al. (2003). In this case the approach is to look for an analog in a large-scale field (supposed to be reliably predicted by GCMs) and then use the local target field simultaneous to the large-scale analog to reconstruct the local-scale field (Zorita & von Storch 1999). No errors from leading time extrapolation arise in this kind of analog search. However, there are 2 main sources of error in any statistical downscaling technique: (1) The predictor field does not explain all of the variability of the predictand field, and (2) errors inherited from the simulated large-scale variable.

The main problem with the use of the analog method is the need for a huge pool of historic cases to find good analogs (van den Dool 1994). The quality of the analogs can be improved by employing a larger library for the search, but this can only be achieved when studying an artificial climate, such as that created by a long GCM integration (e.g. Luksch & von Storch 1999). Observational data are insufficient for finding good analogs, and this problem can only be overcome by selecting a reduced set of relevant d.f. and searching for similarities on it.

The high dimensionality of the atmospheric phase space is usually reduced through the use of principal component analysis (PCA) (Luksch & von Storch 1999, Zorita & von Storch 1999, Timbal & McAvaney 2001). This classical approach will be referred to here as PCA-Analog Downscaling Model (DM, hereafter). The analogs selected by this method only take into account the predictor field, and they would be the same for any predictand related to the same predictor. For instance, even though the physical mechanisms of temperature and precipitation variability are different over our area of interest (Sáenz et al. 2001a,b), the large scale analogs selected for reconstructing both variables would be the same. The selection of the analog cases of the predictor should involve the corresponding predictand field. One such approach (Fraedrich & Rückert 1998) uses a quadratic metric with free coefficients that are iteratively fitted to optimize the forecast error. Unfortunately, the method loses the simplicity of the analog method and becomes similar to a neural network. Since the length of the observation records is insufficient to find good analogs, van den Dool (1994) proposed the construction of artificial analogs by linearly combining all records, with weights determined by a least squares procedure. This really involves the predictand field, but the analog method merely becomes a multivariate linear regression model, losing its non-linear properties.

Another drawback of the analog method is its inability to reconstruct downscaled results beyond the limits of the calibration library. This would be a problem if the downscaled variable were affected by a trend breaking the limits of the observation records. This problem also depends on the size of the library. The greater the number of different climates covered by the library, the higher will be the probability of finding more extreme analogs.

This study proposes a new approach to the usual analog DM by finding the analogs in the space of the CCA temporal expansion coefficients (we will refer to this method as CCA-Analog DM). This study compares the downscaling skill of the CCA-Analog DM with the PCA-Analog DM and with the linear *plain* CCA DM, to find a relevant non-linear part in the relationship between predictor and predictand fields.

The use of an empirical downscaling technique requires a strong relationship between a large-scale variable and the regional scale target variable. The relationship between large-scale atmospheric circulation over the Atlantic Ocean and precipitation over SW Europe (see Zorita et al. 1992, von Storch et al. 1993, Trigo et al. 1999, Ulbrich et al. 1999) has been selected to test the CCA approach for searching analogs. We use sea level pressure (SLP) to characterize Atlantic circulation; geopotential height is less adequate, due to its increase with global warming, which is not reflected in the circulation (Zorita et al. 1995). The downscaling target region is the northern coast of the Iberian Peninsula (Cantabrian Coast, Fig. 1), which has a behavior of precipitation variability that is different from the rest of

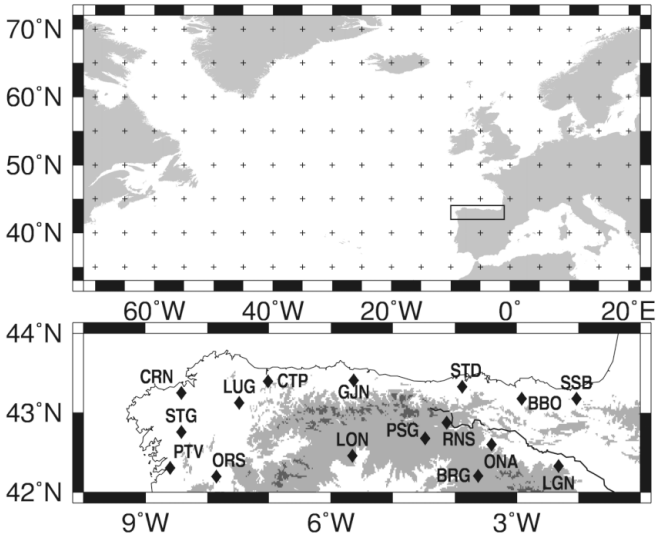

Fig. 1. Atlantic NCAR SLP grid and Cantabrian Coast UCM stations. Lower panel: light gray is >800 m, dark gray is >1600 m elevation. Ebro River course is depicted at lower right

the Iberian Peninsula (Rodríguez-Puebla et al. 1998, Serrano et al. 1999, González-Rouco et al. 2000, Sáenz et al. 2001b). The relationship between SLP and Atlantic circulation is stronger during the rainy winter months, while small-scale convective storms dominate precipitation over the area during the summer season. Therefore, only winter data have been analyzed.

Precipitation time scales are shorter than the monthly scale used in this study. Nevertheless, monthly total rainfall amounts are driven by the monthly average circulation as shown by the significant canonical correlations obtained in previous studies (Zorita et al. 1992, von Storch et al. 1993, González-Rouco et al. 2000). The processes linking precipitation to circulation are non-linear; however, there must be a linear part which is captured by the CCA linear DM. The monthly average of these processes may lead to a near-linear net relationship, which is well represented by a linear DM such as CCA. On the other hand, the monthly average could keep a significant non-linear part which is only captured by a non-linear method such as the analogs used in this work.

We test our technique with observed data sets described in Section 2 and further details of the downscaling methods are given in Section 3. The results obtained with the common CCA linear DM, a sensitivity test on several parameters of the usual PCA approach to the analogs, and a comparison of the new approach with the other 2 techniques are discussed in Section 4. An appendix points out several important issues in the selection of the pre-filtering for the CCA space.

2. DATA

The monthly SLP field used is the NCAR 5° × 5° analyses (Jenne 1975). The main reason for using these old analyses instead of the new Reanalysis at a 2.5° × 2.5° resolution is that the old one starts in 1899 giving half a century more data. The missing values in this data set have been filled by González-Rouco (1997) using a step by step multiple regression procedure. The selected predictor region is 70° W to 20° E and 35° to 70° N, designed to include the main features of the large-scale circulation modes governing precipitation over the Iberian Peninsula (Zorita et al. 1992, von Storch et al. 1993, Ulbrich et al. 1999, González-Rouco et al. 2000, Sáenz et al. 2001b).

Monthly accumulated precipitation data over the Cantabrian Coast were taken from the station homogenized precipitation data set compiled by González-Rouco et al. (2001) at the *Universidad Complutense de Madrid* (UCM). A total of 16 stations from this data set are in the region limited by 10° to 1° W and 42° to 44° N (Fig. 1). The time period covered by this data set is 1899–1989.

Monthly precipitation with much higher resolution (a 0.5° × 0.5° grid, 59 grid points over the Cantabrian Coast) is provided by the Climate Research Unit (CRU), University of East Anglia (New et al. 2000) but this data set has been disregarded on the basis of Fig. 2, which shows the deseasonalized winter monthly time series in Lugo (dotted line) and at the nearest grid point of the CRU dataset (solid line). The variability during the first half of the century is very small and this situation appears at all grid points over the western part of the selected area. The quality of the CRU analyzed data on our target area is not high enough for reliable downscaling in the following sections.

In order to remove any seasonal bias from the records of our analog library, the data sets were deseasonalized by removing monthly climatology. December, January and February monthly anomalies were selected. The common time period covers 1899–1989 on a monthly basis (273 time records, 91 years × 3 records yr^{-1}). This period has been split into two: 1899–1960 and 1961–1989. The latter period is withheld as a known future to validate the downscaling models while the former is used to perform the CCA, and as the library for the analog search.

The leading variability modes according to an S-mode PCA using the anomaly field covariance matrices are shown in Figs. 3 & 4. Degeneracy of the eigenvalues is assessed through North et al. (1982) sampling error bars. The variance fraction retained at several truncation levels is shown on the graphs along with the degenerate multiplets. The first and second SLP empir-

ical orthogonal functions (EOFs) resemble the North Atlantic Oscillation (NAO) and East Atlantic (EA) patterns (Fig. 3). The leading precipitation EOFs show, respectively, a zonal and a meridional gradient (Fig. 4).

3. METHODOLOGY

We used 2 different statistical downscaling techniques concerned with finding a relationship between 2 fields. The first (or left) one is a large-scale field

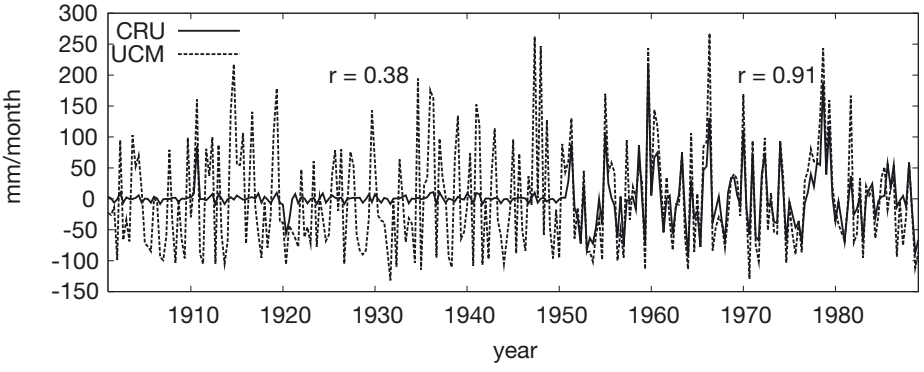

Fig. 2. Deseasonalized monthly precipitation anomaly according to the CRU (grid point 43°15′N, 7°15′W) and UCM (Lugo station, 43°15′N, 7°28′W) data sets; r: correlation between both series up to, and after 1950

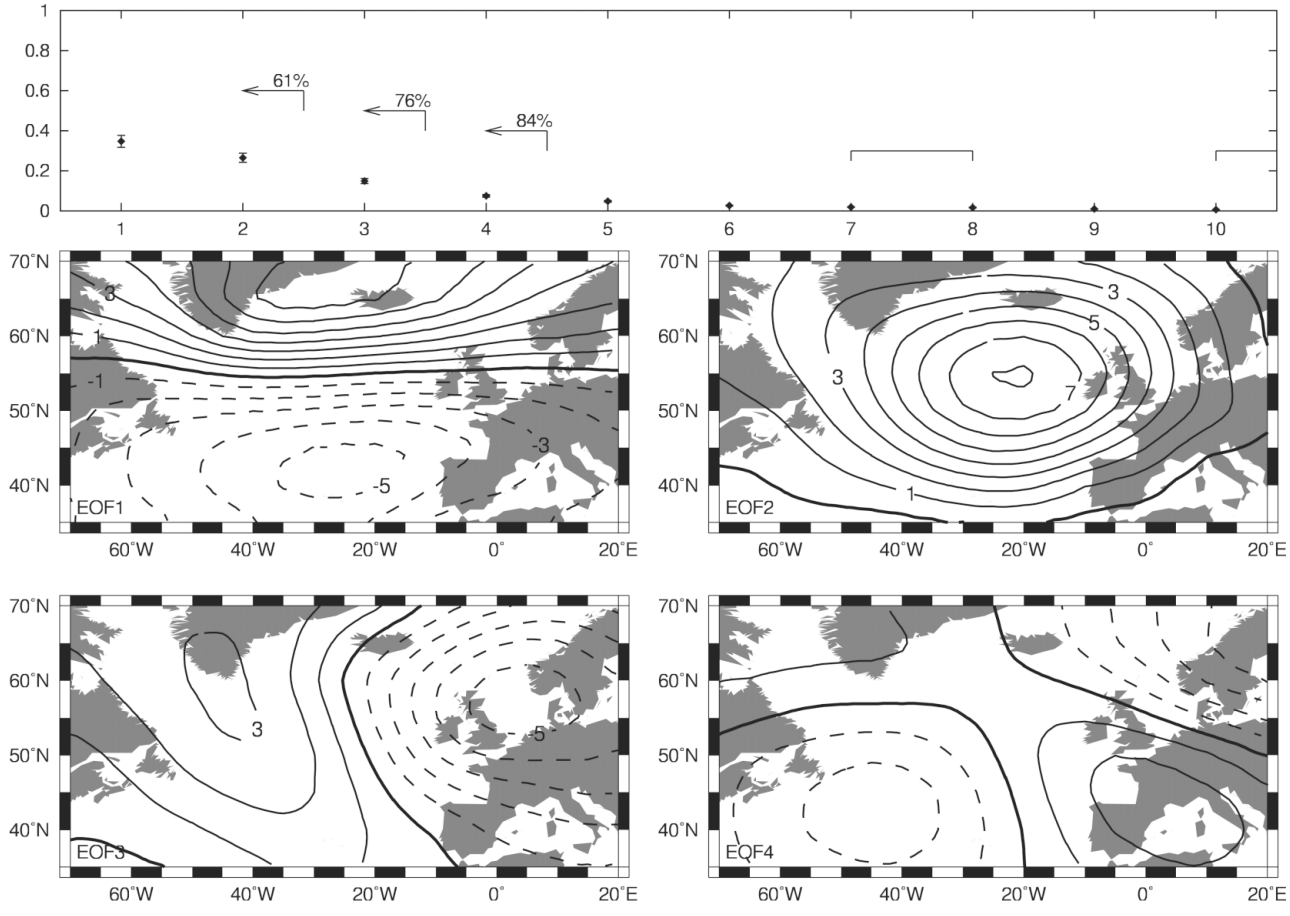

Fig. 3. Variance fraction explained by the 10 leading EOFs from the NCAR SLP (top). Degenerate multiplets (due to overlapping error bars) are joined by lines. Spatial loadings for the 4 leading EOFs (bottom panels). Units: hPa

$slp(t,x_j)$ defined over a large scale spatial domain x_j, j = 1...N_x for each time t. The second (or right) one $pre(t,y_k)$ is restricted to a regional scale domain y_k, k = 1...N_y. The statistical relationship found is then used to reconstruct the regional scale field solely from the large-scale data.

3.1. Canonical correlation analysis

We used CCA as a standard baseline technique for downscaling as it has been widely used for this purpose (von Storch et al. 1993, Huth 1999, González-Rouco et al. 2000). CCA searches for spatial projection patterns, $p_l(x_j)$ and $q_l(y_k)$, such that their temporal expansion coefficients, $a_l(t)$ and $b_l(t)$, show the maximum possible correlation. The fields $slp(t,x_j)$ and $pre(t,y_k)$ are, then, linearly decomposed as:

$$slp(t,x_j) = \sum_l a_l(t)p_l(x_j) \quad pre(t,y_k) = \sum_l b_l(t)q_l(y_k) \qquad (1)$$

Since the canonical correlation patterns are not orthogonal, left and right adjoint spatial patterns (p_l^A and q_l^A) are defined in order to solve the previous equations for the temporal expansion coefficients (von Storch & Navarra 1995). For example,

$$a_l(t) = \sum_j slp(t,x_j)p_l^A(x_j) \qquad (2)$$

solves for the SLP left expansion coefficient with a similar equation for the right field. The same equations would could project any future fields $slp(t',x_j)$ or $pre(t',y_k)$ (with t' out of the temporal range used to derive the CCA spatial patterns) and obtain the canonical coordinates of those fields.

The maximum correlation property allows the use of this technique as a DM. The SLP future field (provided by a GCM or, as in this study, by withholding a part of the available observed data) projected onto the adjoint SLP patterns yields temporal expansion coefficients which are used along with the precipitation canonical patterns $q_l(y_k)$ to reconstruct the regional field. Minimizing the squared error through a linear model between the expansion coefficients gives the following expression for the reconstructed precipitation field:

$$\widetilde{pre}(t',y_k) = \sum_l \rho_l a_l(t')q_l(y_k) \qquad (3)$$

where $\rho_l = corr(a_l,b_l)$ is the canonical correlation.

To filter out spatial noise a PCA-based pre-filtering has been carried out on each CCA decomposition (Barnett & Preisendorfer 1987). That is, the coordinates of the fields that enter the CCA calculation are the leading principal components instead of the real grid point coordinates.

3.2. Analog search

One of the key aspects of the analog search is the selection of the similitude measure to select the past analogs. Looking for analogs in real space is usually inadequate due to the high number of d.f. (van den Dool 1994). With the Atlantic area selected for this study and the 5° × 5° resolution of the NCAR data set, a real phase space search would take place in a 152-dimensional space. Most of these grid points are highly correlated or involve noise; moreover, this 152-dimensional space is mostly empty. Taking an SLP measurement precision of 0.1 hPa, a pressure range from 995 to 1035 hPa and the spatial resolution of the NCAR data set the 273 records at our disposal only fill 0.45% of the phase space.

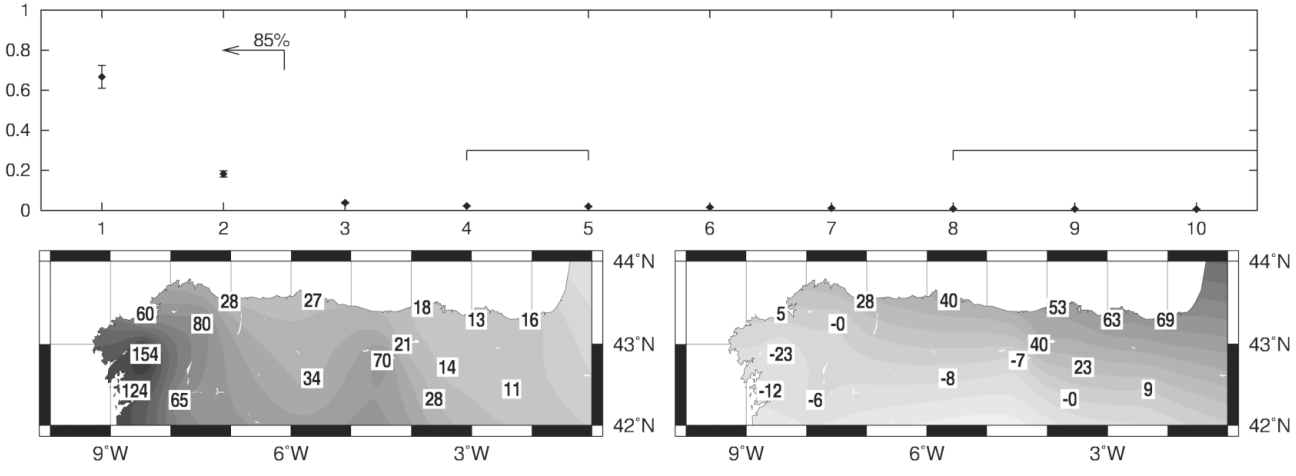

Fig. 4. Variance fraction explained by the 10 leading EOFs from the UCM precipitation data set (top). Degenerate multiplets (due to overlapping error bars) are joined by lines. Spatial loadings for EOF1 (bottom left) and EOF2 (bottom right). Units: mm/month

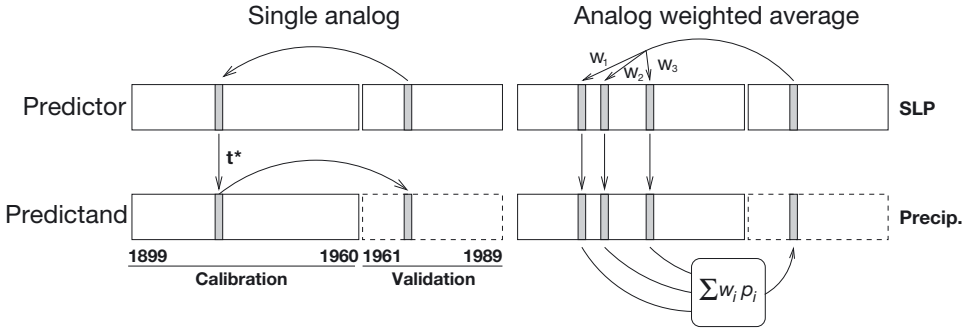

Fig. 5. Single analog (left) and weighted average (right) analog search scheme for downscaling. Analog records for the validation period 1961–1989 are searched in the historical records (1899–1960) of the large-scale variable. The local scale values associated to the analog records found are used to reconstruct the local variable directly (single analog search) or by averaging several analogs (weighted average analog search)

The usual method for the reduction of the dimensionality of the search space is the projection onto the space spanned by the leading EOFs obtained from an S-mode PCA. That is, given the first N_ε EOFs ($e_p(x_j)$, $p = 1...N_\varepsilon$), the coordinates in the PCA-space γ_p of any given spatial pattern of anomalies with real-space coordinates $\Gamma(x_j)$ are:

$$\gamma_p = \sum_{j=1}^{N_x} \Gamma(x_j)\, e_p(x_j) \qquad \forall p = 1...N_\varepsilon \tag{4}$$

and the similitude of this pattern to the anomaly field at each time t is measured by the euclidean distance in this space:

$$d_\Gamma(t) = \sqrt{\sum_{p=1}^{N_\varepsilon} (\alpha_p(t) - \gamma_p)^2} \tag{5}$$

where $\alpha_p(t)$ are the standardized (zero mean, unit variance) principal components (PCs) of the SLP field $slp(t,x_j) = \sum_p \alpha_p(t) e_p(x_j)$. The use of standardized coordinates gives a meaning to the distances we work with in the reduced phase space. The distances are measured in standard deviations and give information on the goodness of an analog, as it is possible to estimate whether a given analog pattern is near or far from the base case, depending on the numeric value of the distance. On the other hand, the standardized distance assigns the same importance in the analog search to all EOFs, even though the variance of the SLP field accounted for by each EOF decreases when increasing their order. The effect of both PC scalings (standardized and variance-carrying) on the DM skill is tested in the intercomparison in Section 4.2.1.

The analog DM consists in finding the time $t^\star$ in our analog library for which the distance $d_\Gamma(t^\star)$ is minimum, and to reconstruct the predictand with its own value at time $t^\star$ (Fig. 5). This downscaling model will be referred to as PCA-Analog DM.

Several historical cases may approach the minimum distance. Selection of a single analog would disregard other analogs of nearly equal quality, which could give rise to different precipitation situations. One way to take into account these close analogs is to use the n closer circulations to the base case and average their precipitation counterparts to reconstruct the field. A simple average would tend to reduce the quality of a good analog if the remaining $n-1$ were less good. In order to avoid this problem a weighted average is employed (Fig. 5) using weights which decrease with the square of the distance to the base case. The effect of these weightings is tested in Section 4.2.1.

3.3. The CCA approach to analogs

The projection onto the PCA truncated space not only reduces the dimensionality of the data but also removes high-frequency spatial noise. This allows a search for patterns with similarities in their main characteristics, while disregarding small local differences.

The PCA is not the only procedure to obtain projection patterns that reduce the dimensionality of phase space and noise. The CCA patterns described in the previous section also constitute a valid projection space. Noise in the CCA was filtered with the Barnett & Preisendorfer (1987) PCA pre-filter. With the use of the CCA space as projection space the analogs are searched in a space with a topology that takes into account the predictand field. The euclidean distance between a base case $\Gamma(x_j)$ and the SLP field at time t in this case is:

$$d_\Gamma^{CCA}(t) = \sqrt{\sum_{l=1}^{N_C} \left(a_l(t) - \sum_j \Gamma(x_j) p_l{}^A(x_j) \right)^2} \tag{6}$$

where $N_C = min(N_\varepsilon^{slp}, N_\varepsilon^{pre})$; N_ε^{slp} and N_ε^{pre} are the number of EOFs retained in the CCA pre-filtering of the SLP and precipitation fields, respectively. The selection of N_ε^{slp} and N_ε^{pre} is crucial in order to obtain

analogs different from those obtained by projection on the PCA-space (see Appendix 1).

4. VALIDATION OF THE DOWNSCALING MODELS

The skill of the downscaling techniques is measured by 2 different scores: (1) The correlation between the original anomaly field and the field reconstructed during the validation period; this correlation skill is insensitive to the level of reconstructed variability and only accounts for the agreement in the peaks of the reconstructed precipitation series. (2) To account for the variability reproduced by the downscaling models, a variance skill is defined as the ratio of the variance of the reconstructed series during the validation period to the variance of the original series during this period. The latter skill score should equal 1 for the models to reproduce the right level of variability.

4.1. Linear CCA DM

The CCA applied to NCAR analyses data and UCM precipitation stations over the Cantabrian Coast during the calibration period (1899–1960) yields the spatial patterns shown in Fig. 6. The 4 leading SLP EOFs and the 2 leading precipitation EOFs have been retained in the PCA pre-filtering in order to optimize the stability of the CCA patterns. The stability has been assessed by a Monte Carlo test on the congruence coefficient (Richman 1986, Cheng et al. 1995) of patterns obtained from 200 random subsamples with 90 temporal records each (out of a total of 186 in the calibration period). This PCA pre-filter selection (4, 2) shows non-overlapping North et al. (1982) sampling error bars and, consequently, no degenerate multiplet is likely to be broken in this truncation.

The first pair (Fig. 6, left) of canonical correlation patterns (CCPs) has a canonical correlation of 0.86 and the second pair has 0.68. The first SLP CCP describes a counterclockwise circulation around a center south of Ireland and explains 20% of the total variance of the field. The associated precipitation CCP (Fig. 6, bottom left) shows decreasing precipitation anomalies from west to east on the Cantabrian Coast. This precipitation pattern explains 56% of the total variance of the precipitation anomaly field. The relation can be physically explained through the advection of moist (dry) air by the corresponding anomalous circulation during positive (negative) phases. The SLP part of the second CCA mode (Fig. 6, top right) shows 2 opposite-sign centers. A positive anomaly maximum is located in the center of the northern Atlantic Ocean, while a negative weaker center is located over Italy. The induced anomalous circulation over the Cantabrian Coast is north to south and accounts for 13% of the variance of the SLP field. A precipitation gradient decreasing from the coast to the interior is found for the precipitation part of this mode, explaining 15% of the variance.

The results are consistent with those found for the entire Iberian Peninsula (von Storch et al. 1993, Zorita & Storch 1999, González-Rouco et al. 2000).

The correlation between the precipitation anomaly field reconstructed by the CCA downscaling model and the observations is shown in Fig. 7. The correlation for the calibration period (1899–1960) is shown as an upper limit to the predictability of the precipitation by this

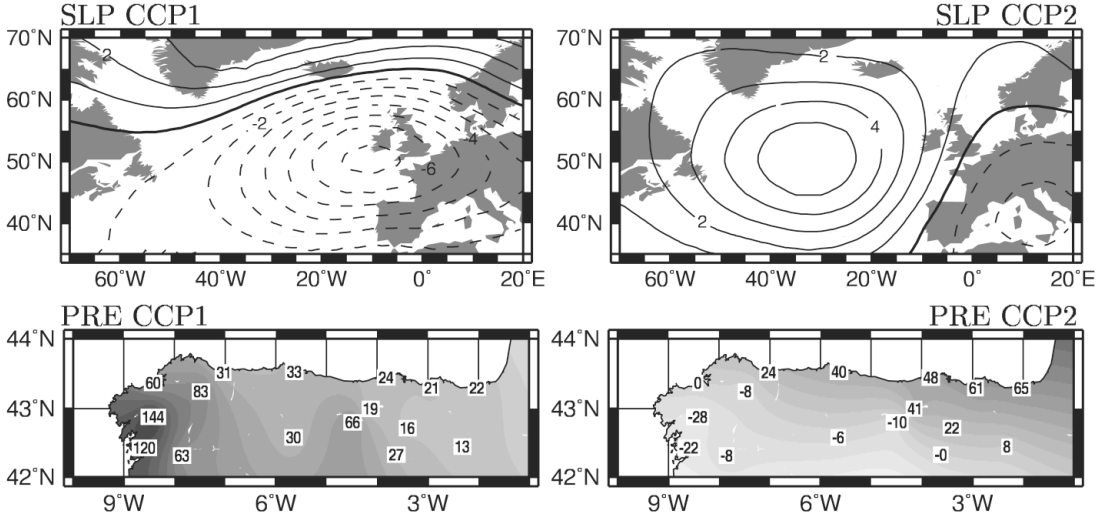

Fig. 6. Canonical correlation patterns obtained using the period 1899–1960 for the NCAR SLP (hPa) and the UCM precipitation (mm/month). The first (left) and second (right) pairs show canonical correlations 0.86 and 0.68, respectively

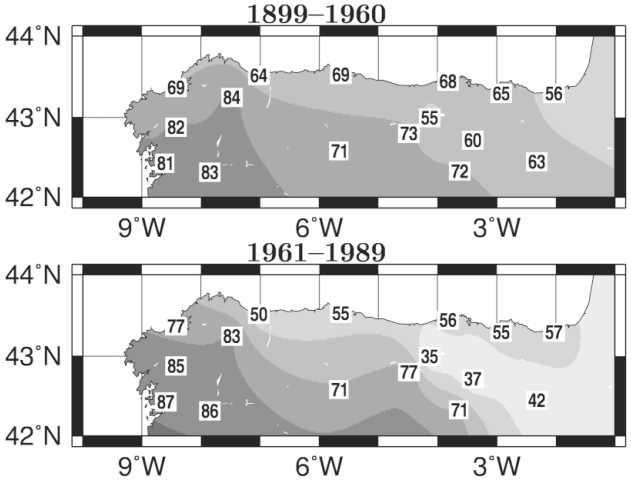

Fig. 7. Correlation (100) between observed and CCA DM reconstructed precipitation for 1899–1960 and 1961–1989 (i.e. correlation skill)

method, since the canonical expansion coefficients themselves were used for this reconstruction without temporal extrapolation. Higher correlations are found for the western part of the area (where the variability of precipitation is also higher). During the validation period a decrease in correlation is found in the basin of the Ebro River (values <0.5, which is the threshold for a 95% significance level), possibly related to the mountain ranges (see Fig. 1), which surround the basin and block the westerly circulation associated with the first CCA mode (Fig. 6, top left) and the northerly circulation associated with SLP CCP2 (Fig. 6, top right).

4.2. Non-linear Analog DMs

4.2.1. PCA-Analogs

PCA is the most common statistical tool for dimensionality reduction in analog searches. The 4 leading SLP EOFs explain a large and disjoint—according to North et al. (1982) sampling error bars—part of the total variance of the field (Fig. 3) so this number seems reasonable for our PCA-Analog search.

Sensitivity analysis of the PCA-Analog DM was based on the correlation and variance skills. These skills are plotted for each of the UCM stations in Figs. 8 & 9. Fig. 8 shows the correlation skill for one analog (n = 1) and for averages of the 3 and 6 nearest analogs in a PCA-space with a dimension $N\varepsilon$ = 4. The lower part of Fig. 8 illustrates the CCA-Analog approach and will be discussed in Section 4.2.2. Some stations are predicted well (western part of the area), whereas predictions for Reinosa, Oña and Logroño (RNS, ONA and LGN) are poor. The correlation skill grows with the number of averaged analogs but improves little when more than 3 are used.

Fig. 9 shows the variance skill using a number n = 1, 3 and 6 of averaged patterns. The level of reproduced variability is good for n = 1 (no average). The averaging of the patterns smoothes the extreme values. The (weighted) mean value at each grid point is always less than or equal to the value of the patterns being averaged, thus the variance of the mean pattern is also less. The opposite effect is observed in the variance skill with respect to the correlation skill: the bigger the n,

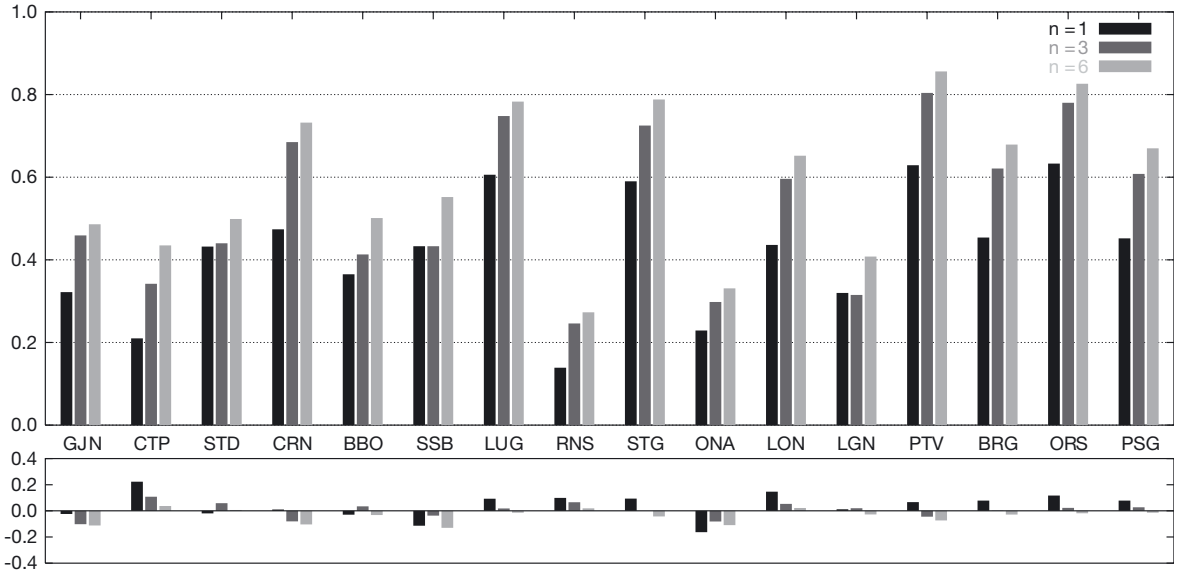

Fig. 8. Correlation skill of the PCA-Analog DM at each station while the smoothing average runs through n = 1, 3, 6 (top panel). The bottom panel shows the improvement when using the projection of the analog search over the CCA space. PCA-Analog uses N_ε = 4 and CCA-Analog N_ε^{slp} = 4 and $N_\varepsilon^{pre} = N_C$ = 2. See Fig. 1 for abbreviations

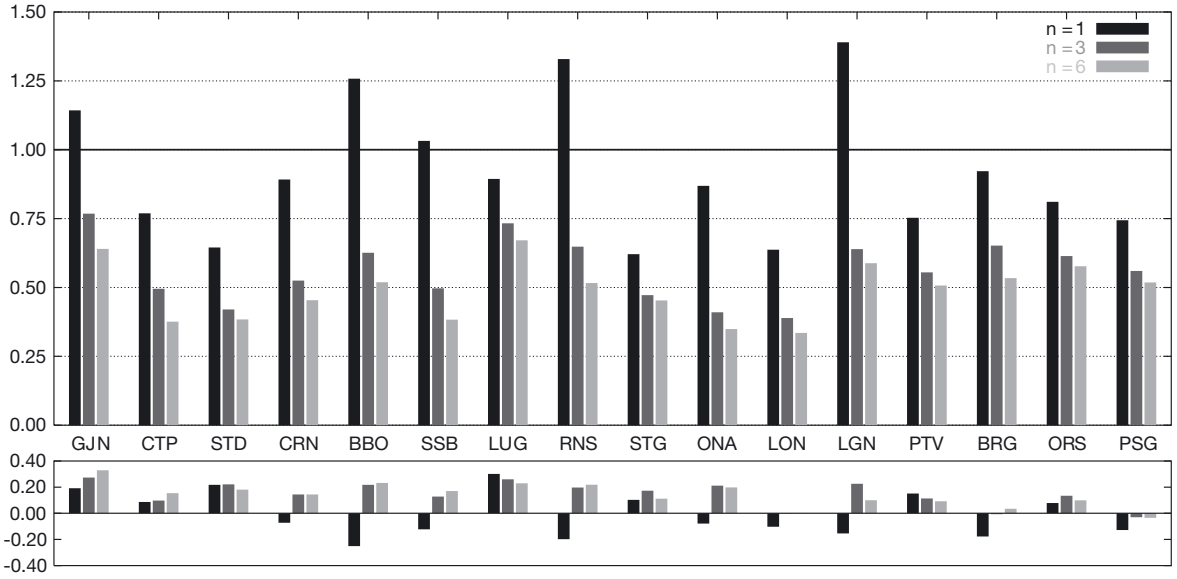

Fig. 9. Variance skill of the PCA-Analog DM at each station while the smoothing average runs through n = 1, 3, 6 (top panel). The bottom panel shows the skill difference CCA-Analog minus PCA-Analog. PCA-Analog uses N_ε = 4 and CCA-Analog N_ε^{slp} = 4 and $N_\varepsilon^{pre} = N_C$ = 2. See Fig. 1 for abbreviations

the smaller is the variance skill (cf. Figs. 8 & 9). An intermediate value of n = 3 was therefore used in the analyses described below.

Averaging should be avoided for the purpose of obtaining the correct level of variability in the reconstructed field. Variance with a good fit is inherent to the analog methods because the predictand is built with elements of the observed variable. Thus, even though the predictor does not explain all of the variability of the predictand, the variance is well reproduced and includes actual precipitation noise in addition to the signal. This avoids artificially inflating the variance by adding noise, which is the usual approach in other downscaling techniques (von Storch 1999). The variance may even be overestimated (Fig. 9). This means that the analogs found are, on average, further removed from the mean state.

The sensitivity analysis for the dimensionality of the phase space is shown in Fig. 10, along with the effects of the standardized coordinates. The correlation skill is similar for standardized and variance-carrying coordinates in a low-dimensional phase space, but the use of standardized coordinates is not recommended when high order EOFs are included. When the distance along each dimension is weighted according to the variance explained by that mode (Fig. 10, right) the correlation skill initially oscillates but attains a stable value with the 14th EOF. Even though only the 4 leading EOFs are stable for explaining the SLP variability according to our Monte Carlo test, the first 10 to 13 EOFs seem to be related to the precipitation over the area under study.

The same stabilization occurs in the variance skill at about 10 dimensions (Fig. 11; variance skill ±SD for 16 stations). The reconstruction with the standardized metric (Fig. 11, left) shows similar skill for the first few

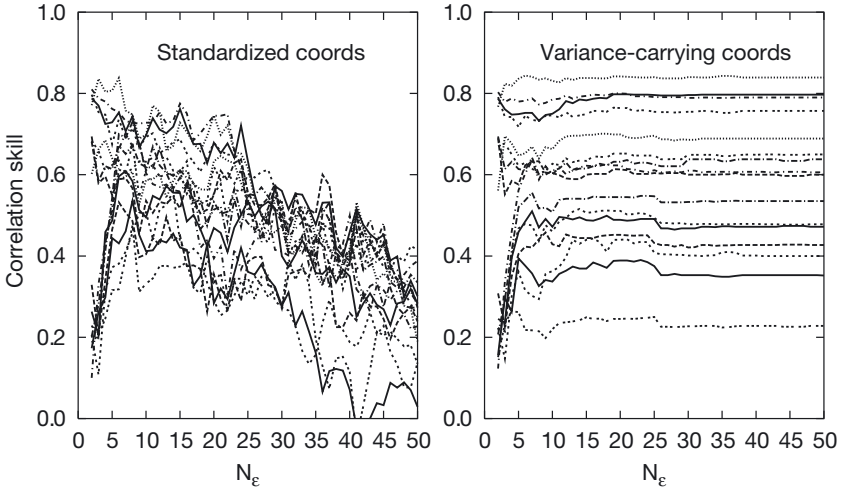

Fig. 10. Evolution of the correlation skill at each station with varying degrees of freedom of the search space for the PCA-Analog DM with weighted average of n = 3 analogs. Each line corresponds to one of the 16 stations. Correlation skill with standardized PCs (left) and with variance-scaled dimensions (right)

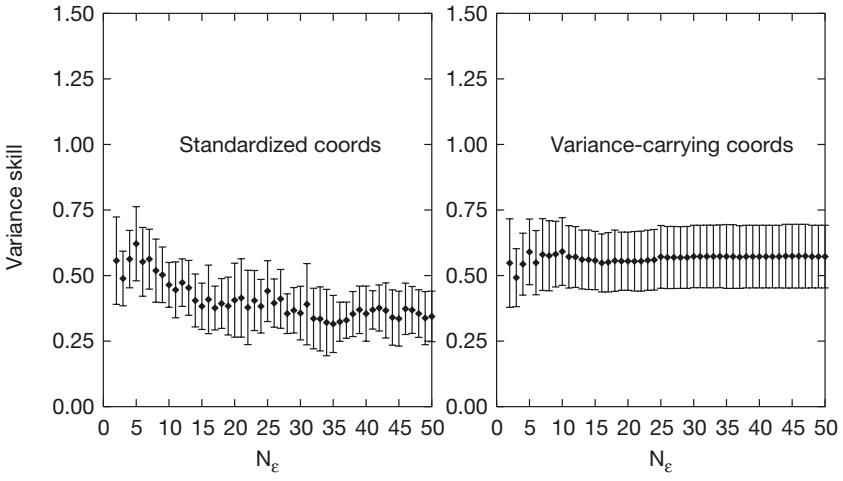

Fig. 11. Average evolution of the variance skill with varying degrees of freedom of the search space for the PCA-Analog DM with weighted average of n = 3 analogs (mean ± SD of the 16 stations). Variance skill with standardized PCs (left) and with variance-scaled dimensions (right)

dimensions, but declines with growing dimensionality. This effect is, however, due to the smoothing (n = 3) since for n = 1 the average variance skill is nearly 1 at all dimensionalities (data not shown).

4.2.2. CCA-Analogs vs. PCA-Analogs

A stable selection of CCA patterns is reached with the PCA pre-filter N_ε^{slp} = 4, N_e^{pre} = 2 (Section 4.1., Figs. 8 & 9). In Fig. 8 the bars represent the CCA-Analog correlation skill minus that of the PCA-Analog. Thus, positive bars mean that CCA-Analog DM improves over the PCA-Analog DM. No clear improvement in correlation skill is evident with the new method. The selection of single analogs (n = 1, black bars on Fig. 8) seems to be better with the new approach since 0.1 in correlation skill is gained in several stations and 0.2 in Cervera del Pisuerga (PSG). For averaged searches (n = 3 or 6) the skill is similar in both CCA-Analog and PCA-Analog DM.

The lower part of Fig. 9 shows CCA-Analog minus PCA-Analog results for variance skill difference; as variance skill should be close to 1, negative bars also represent an improvement if the corresponding PCA-Analog reconstructed variance skill is >1. For instance, the negative bars in Bilbao, Reinosa and Logroño (BBO, RNS, LGN) correct the variance overestimated by PCA-Analog DM for these stations, and most underestimates are corrected as well.

In the comparison carried out in Figs. 8 & 9 the PCA-Analog DM searches for analogs with 4 d.f. while the CCA-Analog DM only uses 2. The information about the circulation over the North Atlantic is being condensed into 2 d.f. They are chosen to reproduce the precipitation on the Cantabrian Coast with a skill similar to that reached by the PCA-Analog DM with 4 d.f.

The comparison of the PCA-Analog DM and the CCA-Analog DM should take place in a search space of the same dimension where the search library equally fills the phase space by both methods. Thus, in Fig. 12 the PCA-Analog DM has been reduced to 2 d.f. by projection over the 2 leading EOFs. In this case the skill improvement of the CCA-Analog DM is apparent. But, in this comparison, the predictor signal entering each DM is different. While the CCA-Analog DM makes use of 84% of the variance of the predictor field, the PCA-Analog DM has only 61%.

Since the reconstruction of the precipitation is determined by SLP analogs selection, the explanation of the CCA-Analog improvement should be in the SLP analog selection differences. Fig. 13 shows the time correlation at each grid point of the NCAR SLP predictor field and the analog reconstructed field (in this case the predictor itself is reconstructed by the analogs found, and compared with its own base patterns) by means of the ordinary PCA-space analog search and the CCA-space search during the independent period 1961–1989. Correlations are very high over a wide area in the case of the PCA-space search. In the CCA-space search, however, the high correlation area is reduced to an elongated region over the east Atlantic. This means that the PCA search uses patterns similar to a base case over the whole area, while the CCA search only considers the high correlation area shown in Fig. 13, resulting in a less reliable SLP field. However, our efforts are not directed towards reconstructing the SLP field itself, but the precipitation induced by this field. The analog patterns selected by the CCA-Analog DM provide a more accurate reproduction of the circulation which determines precipitation over the Cantabrian Coast, even though these patterns do not fit the predictor as well as the patterns derived from the PCA-space approach.

The domain for the analog search is important and sometimes regarded as a parameter that needs to be optimized (Timbal & McAvaney 2001). The search projecting onto the CCA-space performs this domain optimization automatically. The domain is weighted with high loading factors where the circulation is interesting for the precipitation. A manual selection of the domain is usually bounded to rectangular (Timbal &

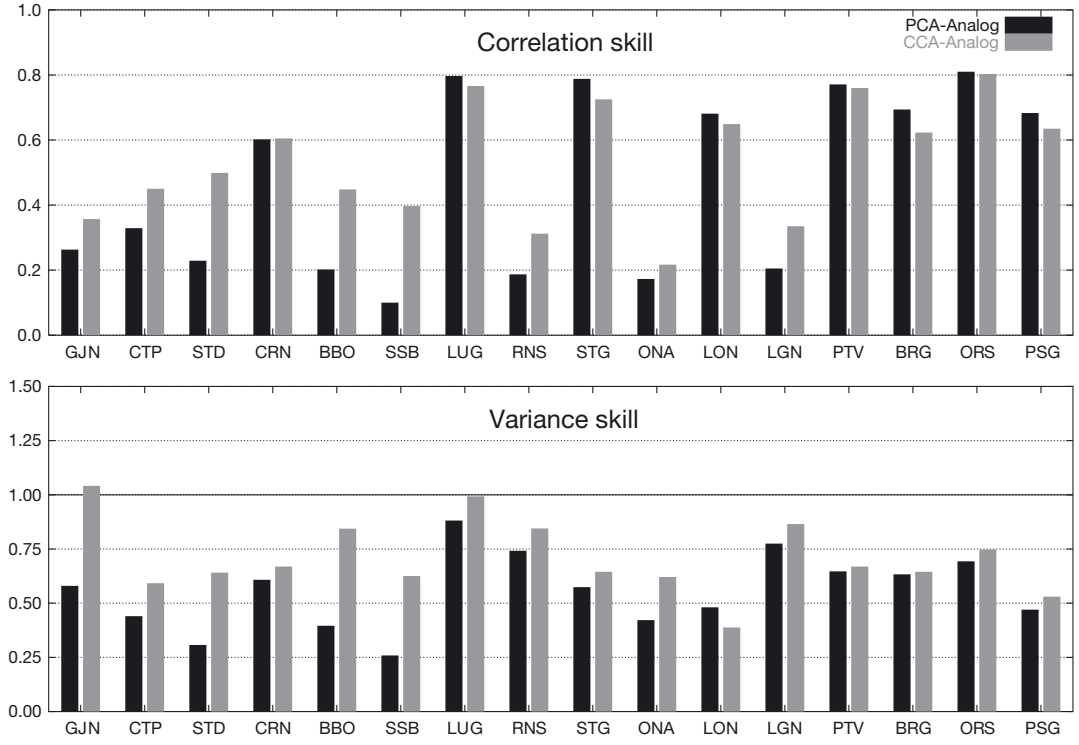

Fig. 12. Correlation (top) and variance (bottom) skills at each station for PCA-Analogs (N_ε = 2) and CCA-Analogs (N_ε^{slp} = 4, N_ε^{pre} = 2). See Fig. 1 for abbreviations

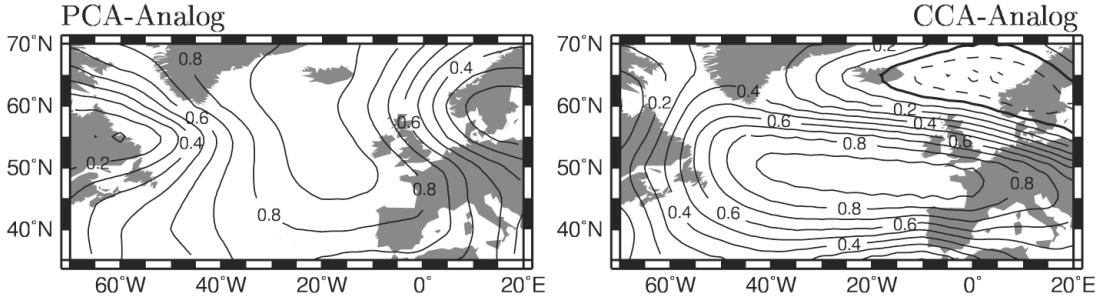

Fig. 13. Correlation patterns between the SLP reconstructed by PCA- and CCA-Analog DMs, and the original NCAR SLP. Analog DMs use: n = 3, $N_\varepsilon = N_C$ = 2, N_ε^{slp} = 4 and N_ε^{pre} = 2

McAvaney 2001) or circular (van den Dool 1989) regions, although systematic selection of irregular regions is not a completely unexplored field (Barnett & Preisendorfer 1978).

We tested whether this domain optimization occurs and the CCA-Analog approach is less sensitive to the selection of the predictor domain than the PCA approach. Two other domains were selected (Fig. 14): (1) A wider domain, D1, including the previous domain D0 used throughout this study. (2) A smaller domain, D2, centered over the area with high SLP correlation in the CCA-Analog approach (Fig. 13, right). The performance of each DM on each domain has been summarized in 2 numbers: the average correlation skill and the average variance skill, computed over the 16 precipitation stations (Table 1). The correlation skill of a smaller area is similar to that obtained for a larger area (D2 compared to D0) if the former covers the areas with certain skill for prediction of the latter. This is true for both PCA- and CCA-analog DMs. The CCA-Analog DM also provides a good skill for an even larger area such as D1. However, the correlation skill of the PCA-Analog DM decreases substantially in this larger predictor area. The selection of D2 was motivated by the results obtained for D0, which is why the information contained in D0 can be obtained from D2. The CCA

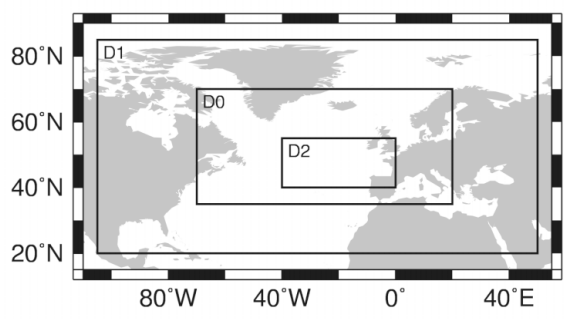

Fig. 14. Domains used to test the sensitivity of the results to the predictor area. The small domain D2 is located over the higher loadings in Fig. 13 (right)

method selects the domain of interest automatically, performing better than the PCA approach in all 3 cases.

The average variance skill behavior (vs in Table 1) is different. The PCA approach gives very similar skills for the 3 domains. As long as the area of interest for the precipitation is included in the search domain, the explained variance is extracted from the data and the non-relevant information does not reduce the skill, as was the case for the correlation skill. In the CCA approach the skills are better and the domain D0 performs best.

Table 1. Average correlation (cs) and variance (vs) skills over 16 precipitation stations for 3 different domains D0, D1 and D2 (cf. Fig. 14)

	PCA-Analog		CCA-Analog	
	cs	vs	cs	vs
D0	0.47	0.56	0.54	0.71
D1	0.25	0.58	0.43	0.64
D2	0.46	0.59	0.55	0.65

4.3. Linear vs. non-linear CCA approaches

The correlation skill of the non-linear technique is systematically lower than that obtained by the CCA DM (Fig. 15). As both techniques yield very similar skills, the CCA-Analog DM does not seem to be capturing any extra non-linear elements in the relationships between the variables, at least on this time scale (see also Zorita & von Storch 1999). The variance skill is improved by the non-linear method, as the analog methods reproduce the variance better than a raw CCA DM.

Due to the particular behavior of daily precipitation (which often has a value of zero), its probability density function (PDF) is highly skewed. When monthly total precipitation is used, the PDF tends to be more normal.

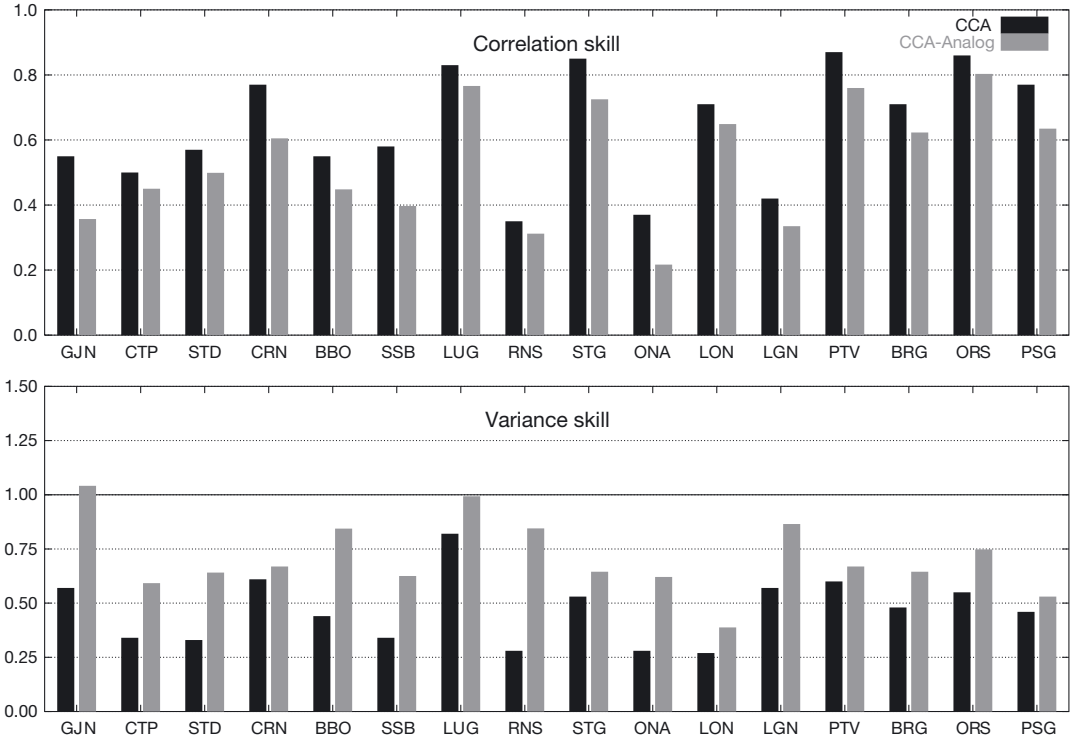

Fig. 15. Correlation (top) and variance (bottom) skills at each station for the CCA linear and the CCA-Analog DM . See Fig. 1 for abbreviations

The χ^2 statistic measures the deviation of a PDF from normality (Wilks 1995). χ^2 is essentially the mean squared distance to a normal PDF with the same mean and variance.Using the monthly total precipitation is insufficient to normalize the PDFs (Fig. 16). The PCA-Analog and CCA-Analog reconstructed precipitation PDFs show a level of nonnormality similar to that of the observed precipitation. The PDF of the CCA DM reconstructed precipitation is more normal than that of the observed precipitation (note the logarithmic scale), even though it is not statistically significant at the 5% significance level. The statistic values at Gijón and A Coruña (GJN and CRN) are closely reproduced by the linear model, because observed precipitation is more normally distributed at these 2 stations.

GCMs represent the most useful approach to get information about the future evolution of the large-scale climate. But, as long-term climate prediction is not an initial value problem, temporal correlation of a long-range simulation with observed variables cannot be expected. The same is true for the downscaled information. The information derived from GCMs needs to be interpreted in a probabilistic way. The correctness of multiyear averages, variance, or trends implies the conservation of the PDF. If a downscaling model does not preserve the non-normality of the PDF (as shown in the case of the CCA DM for the Cantabrian Coast), any change in the PDF due to forcing (e.g. greenhouse gases, volcanism, SST tropical forcing) will be obscured by the normalization of the PDF that has resulted from downscaling. No clear conclusions can be drawn from the CCA DM downscaled precipitation PDF in a climate change experiment since the climate change effects on the PDF and the DM-induced normalization are merged in the change of the PDF.

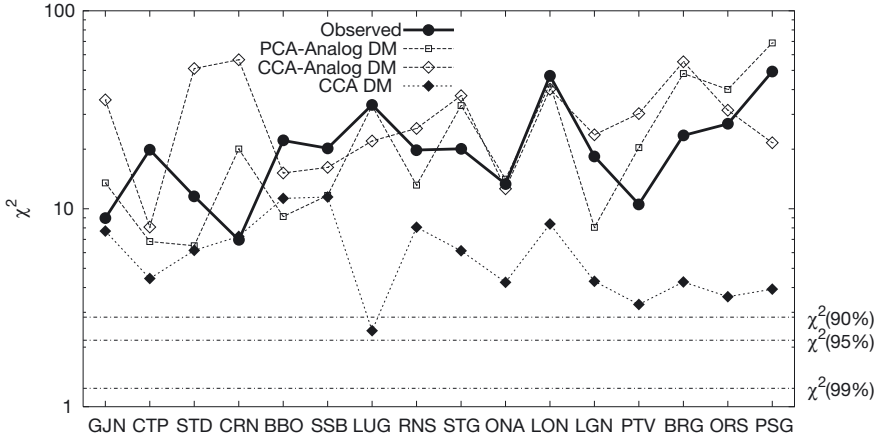

Fig. 16. χ^2 normality test and confidence level thresholds of normal distribution for observed and reconstructed precipitation data using the linear CCA DM, the PCA-Analog DM and the CCA-Analog DM. See Fig. 1 for abbreviations

5. CONCLUSIONS

A new phase space for the selection of atmospheric analogs when used for downscaling has been proposed. The standard PCA technique for dimensionality reduction of the phase space is insensitive to local fields in downscaling. The approach presented here consists in a projection onto the space spanned by the leading CCA spatial patterns of the predictor, and this should select circulations induced by the predictor field relevant to the local predictand field.

The new approach finds d.f. of the North Atlantic circulation with greater relevance to precipitation over the Cantabrian Coast. When working with a reduced number of dimensions in the phase space, the directions selected by the CCA-Analog DM are more accurate and yield higher predictive skills than the standard approach (projection onto the EOFs).

This improvement by the CCA-Analog DM method derives from the automatic selection of the large-scale domains of interest for the precipitation. The low sensitivity of the correlation and variance skills to different predictor regions supports this idea.

The PCA truncation in the PCA-Analog DM is sometimes carried out as a routine truncation of the noise, and a high number of EOFs is considered, in order to make sure that nearly 100% of the variance is taken into account (Luksch & von Storch 1999, Timbal & McAvaney 2001). The quality of the analogs depends on this truncation, because the volume of the phase space is being increased, while the number of patterns in the search library is maintained constant. These authors locate their truncations in the stable part of Fig. 10 (right panel), but this procedure does not guarantee maximum skill or better analogs. Moreover, for about 8 d.f. the search in standardized coordinates (with very low skill for high order truncations) shows average correlation skills never attained by the search in variance-carrying coordinates and with less dispersion between stations (data not shown).

The results in correlation skill of the CCA-Analog DM are similar to those of the CCA DM and there is no gain in using the non-linear (more costly) model. The analog methods have better variance skill. Even after smoothing, the level of reproduced variability is higher than the linearly reconstructed one, and the average variance skill is very close to one when no smoothing is applied (but this reduces the correlation skill).

The linear CCA DM fails to reproduce the non-normality of the precipitation

PDFs, and the downscaled precipitation has more normal distribution than observed. This normalization could erroneously be interpreted as a decrease in negative precipitation anomalies and an increase in positive ones. The analog models better maintain this non-normality, which is more suitable for assessing GCM climate change downscaled precipitation. But the analog methods do not simulate possible trends leading to extreme values larger than those observed during the calibration period.

Acknowledgements. We thank E. Zorita for the time we enjoyed at GKSS Forschungzentrum (Geesthacht, Germany) during the spring of 2001 when we began to work with analogs. We are also grateful to J. F. González-Rouco, who supplied the UCM precipitation data and the interpolated NCAR data. We thank the NCEP/NCAR for providing the SLP data. The overland high-resolution precipitation data was supplied by the Climate Impacts LINK Project (UK Department of the Environment Contract EPG 1/1/16) on behalf of the Climatic Research Unit, University of East Anglia. J.F. was supported by a grant from the Departamento de Educación, Universidades e Investigación, Gobierno Vasco. Financial support was also provided by the Ministerio de Ciencia y Tecnología, project number REN2002-04584C04-04. The comments by 3 anonymous reviewers improved the final version of this paper.

LITERATURE CITED

Barnett TP, Preisendorfer RW (1978) Multifield analog prediction of short-term climate fluctuations using a climate state vector. J Atmos Sci 35:1171–1787

Barnett TP, Preisendorfer RW (1987) Origins and levels of monthly and seasonal forecast skill for United States air temperature determined by canonical correlation analysis. Mon Weather Rev 115:1825–1850

Biau G, Zorita E, von Storch H, Wackernagel H (1999) Estimation of precipitation by kriging in the EOF space of the sea level pressure field. J Clim 12:1070–1085

Cheng X, Nitsche G, Wallace JM (1995) Robustness of low-frequency circulation patterns derived from EOF and rotated EOF analyses. J Clim 8:1709–1713

Dehn M (1999) Application of an analog downscaling technique to the assessment of future landslide activity—a case study in the Italian Alps. Clim Res 13:103–113

Fraedrich K, Rückert K (1998) Metric adaption for analog forecasting. Physica A 254:379–393

Giorgi F, Mearns LO (1991) Approaches to the simulation of regional climate change: a review. Rev Geophys 29:191–216

González-Rouco JF (1997) Modelo de predicción de la precipitación peninsular en climas perturbados. PhD thesis, Universidad Complutense de Madrid, Madrid

González-Rouco JF, Heyen H, Zorita E, Valero F (2000) Agreement between observed rainfall trends and climate change simulations in the southwest of Europe. J Clim 13: 3057–3065

González-Rouco JF, Jiménez JL, Quesada V, Valero F (2001) Quality control and homogeneity of precipitation data in the Southwest of Europe. J Clim 14:964–978

Huth R (1999) Statististical downscaling in central Europe: evaluation of methods and potential predictors. Clim Res 13:91–101

Jenne RL (1975) Data sets for meteorological research. Technical Note NCAR-TN/IA-111, National Center for Atmospheric Research, Boulder, Colorado. USA

Kidson JW, Thompson CS (1988) A comparison of statistical and model-based downscaling techniques for estimating local climate variations. J Clim 11:735–753

Lorenz EN (1969) Atmospheric predictability as revealed by naturally occurring analogues. J Atmos Sci 26:636–646

Luksch U, von Storch H (1999) An empirical approach for estimating macroturbulent heat transport conditional upon the mean state. J Atmos Sci 56:2070–2080

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

New M, Hulme M, Jones P (2000) Representing twentieth-century space-time climate variability. Part II: Development of 1901–96 monthly grids of terrestrial surface climate. J Clim 13:2217–2238

North G, Bell T, Cahalan R, Moeng F (1982) Sampling errors in the estimation of empirical orthogonal functions. Mon Weather Rev 110:699–706

Richman MB (1986) Rotation of principal components. J Climatol 6:293–335

Rodríguez-Puebla C, Encinas AH, Nieto S, Garmendia J (1998) Spatial and temporal patterns of annual precipitation variability over the Iberian Peninsula. Int J Climatol 18:299–316

Ruosteenoja K (1988) Factors affecting the occurrence and lifetime of 500 mb height analogues: a study based on a large amount of data. Mon Weather Rev 116:368–376

Sáenz J, Zubillaga J, Rodríguez-Puebla C (2001a) Interannual variability of winter precipitation in northern Iberian Peninsula. Int J Climatol 21:1503–1513

Sáenz J, Zubillaga J, Rodríguez-Puebla C (2001b) Interannual winter temperature variability in the north of the Iberian Peninsula. Clim Res 16:169–179

Serrano A, García JE, Mateos VL, Cancillo ML, Garrido J (1999) Monthly modes of variation of precipitation over the Iberian Peninsula. J Clim 12:2894–2919

Tang B, Hsieh WW, Monahan AH, Tangang FT (2000) Skill comparisons between neural networks and canonical correlation analysis in predicting the Equatorial Pacific sea surface temperatures. J Clim 13:287–293

Timbal B, McAvaney BJ (2001) An analogue-based method to downscale surface air temperature: application for Australia. Clim Dyn 17:947–963

Timbal B, Dufour A, McAvaney B (2003) An estimate of future climate change for western France using a statistical downscaling technique. Clim Dyn 20:807–823

Toth Z (1991) Intercomparison of circulation similarity measures. Mon Weather Rev 119:55–64

Trigo IF, Davies TD, Bigg GR (1999) Objective climatology of cyclones in the Mediterranean region. J Clim 12: 1685–1696

Ulbrich U, Christoph M, Pinto JG, Corte-Real J (1999) Dependence of winter precipitation over Portugal on NAO and baroclinic wave activity. Int J Climatol 19: 379–390

van den Dool HM (1989) A new look at weather forecasting through analogues. Mon Weather Rev 117:2230–2247

van den Dool HM (1994) Searching for analogues, how long must we wait? Tellus A 46:314–324

von Storch H (1999) On the use of 'inflation' in statistical downscaling. J Clim 12:3505–3506

von Storch H, Navarra A (eds) (1995) Analysis of climate variability: applications of statistical techniques. Springer, Berlin

von Storch H, Zorita E, Cubasch U (1993) Downscaling of global climate change estimates to regional scales: An application to Iberian rainfall in wintertime. J Clim 6: 1161–1171

Wilby RL, Wigley TML (1997) Downscaling general circulation model output: a review of methods and limitations. Prog Phys Geogr 21(4):530–548

Wilks DS (1995) Statistical methods in the atmospheric sciences. Academic Press, San Diego

Zorita E, von Storch H (1999) The analog method as a simple statistical downscaling technique: comparison with more complicated methods. J Clim 12:2474–2489

Zorita E, Kharin V, von Storch H (1992) The atmospheric circulation and sea surface temperature in the North Atlantic area in winter: their interaction and relevance for Iberian precipitation. J Clim 5:1097–1108

Zorita E, Hughes JP, Lettemaier DP, von Storch H (1995) Stochastic characterization of regional circulation patterns for climate model diagnosis and estimation of local precipitation. J Clim 8:1023–1042

Appendix 1. PCA pre-filter selection

The PCA pre-filter proposed by Barnett & Preisendorfer (1987) for the CCA calculation performs a PCA on both fields and carries out the CCA over a reduced set of the leading principal components to filter out the noise. Apart from filtering the noise, the equations to solve are simpler and the canonical correlation patterns in the PCA-space coordinates are orthogonal.

The first point to take into account when dealing with this pre-filter is that the total number of canonical patterns is limited by the PCA truncation selected. The truncation level will be referred to as (N_ε^{slp},N_ε^{pre}); N_ε^{slp} is the number of EOFs retained in the left field (SLP) and N_ε^{pre} the number retained in the right one (precipitation, in our example). The maximum number of canonical patterns after a (N_ε^{slp},N_ε^{pre}) truncation is $N_C = min(N_\varepsilon^{slp},N_\varepsilon^{pre})$. In our example, a truncation level of (4, 2) has been used and the maximum number of canonical patterns is therefore $N_C = 2$.

The canonical patterns in the EOF space can be obtained through a singular value decomposition (SVD) of the cross-covariance matrix of the PCs of both fields ($C_{\alpha\beta}$, which is a $N_\varepsilon^{slp} \times N_\varepsilon^{pre}$ matrix)

$$C_{\alpha\beta} \overset{SVD}{=} L\Sigma R^T \tag{7}$$

where the L ($N_\varepsilon^{slp} \times N_C$) and R ($N_\varepsilon^{pre} \times N_C$) columns are orthonormal vectors corresponding to the left and right canonical patterns expressed in EOF-space coordinates and Σ is a diagonal matrix containing the decreasing canonical correlations. This is evident considering the patterns $\Im$ and $\Re$ and the diagonal correlations matrix r of the usual eigenvalue approach (e.g. see Zorita et al. 1992).

$$C_{\alpha\beta}C_{\beta\alpha}\Im = \Im r^2 \tag{8}$$

$$C_{\beta\alpha}C_{\alpha\beta}\Re = \Re r^2 \tag{9}$$

Substituting the SVD decomposition:

$$C_{\alpha\beta}C_{\beta\alpha}\Im = L\Sigma R^T R\Sigma L^T\Im = L\Sigma^2 L^T\Im \tag{10}$$

And this is equal to $\Im r^2$ if, and only if, $\Im = L$ and $r = \Sigma$ due to the uniqueness of the eigenvalues of a symmetric matrix once they are constrained by their orthonormality. To return to real space coordinates it is necessary to reconstruct the canonical patterns (which will no longer be orthonormal vectors) using the EOF matrices E ($N_{slp} \times N_\varepsilon^{slp}$) and F ($N_{pre} \times N_\varepsilon^{pre}$)

$$P = EL;\ Q = FR \tag{11}$$

where P ($N_{slp} \times N_C$) and Q ($N_{pre} \times N_C$) are the standard canonical correlation patterns. The coordinates in the PCA or in the CCA space for our analog search are obtained by projecting the field using the E or P matrices, respectively. These matrices are related through the matrix L, which has orthonormal columns. The matrix L is $N_{slp} \times N_C$, so if $N_{slp} < N_{pre}$, then $N_C = N_{slp}$ and L is a square orthogonal matrix. In this case the PCA-space and the CCA-space are related through a rigid rotation (L being the rotation matrix) which leaves the distances between any pair of points unchanged, and the analogs found by both methods are exactly the same. As a result of this, a truncation with $N_{slp} > N_{pre}$ must be selected in order to obtain different analogs when using the CCA approach. The L matrix acts, in this latter case, by merging the N_ε^{slp} PCA coordinates into a smaller number ($N_C = N_{pre}$) of directions interesting for precipitation purposes.

Editorial responsibility: Hans von Storch, Geesthacht, Germany

Submitted: January 20, 2003; Accepted: July 9, 2003
Proofs received from author(s): September 8, 2003