Clim Dyn (2015) 45:1395–1413
DOI 10.1007/s00382-014-2388-x

Establishing the skill of climate field reconstruction techniques for precipitation with pseudoproxy experiments

Juan José Gómez-Navarro · Johannes Werner · Sebastian Wagner · Jürg Luterbacher · Eduardo Zorita

Received: 10 June 2014 / Accepted: 20 October 2014 / Published online: 11 November 2014
© Springer-Verlag Berlin Heidelberg 2014

Abstract This study aims at assessing the skill of several climate field reconstruction techniques (CFR) to reconstruct past precipitation over continental Europe and the Mediterranean at seasonal time scales over the last two millennia from proxy records. A number of pseudoproxy experiments are performed within the virtual reality of a regional paleoclimate simulation at 45 km resolution to analyse different aspects of reconstruction skill. Canonical Correlation Analysis (CCA), two versions of an Analog Method (AM) and Bayesian hierarchical modeling (BHM) are applied to reconstruct precipitation from a synthetic network of pseudoproxies that are contaminated with various types of noise. The skill of the derived reconstructions is assessed through comparison with precipitation simulated by the regional climate model. Unlike BHM, CCA systematically underestimates the variance. The AM can be adjusted to overcome this shortcoming, presenting an intermediate behaviour between the two aforementioned techniques. However, a trade-off between reconstruction-target correlations and reconstructed variance is the drawback of all CFR techniques. CCA (BHM) presents the largest (lowest) skill in preserving the temporal evolution, whereas the AM can be tuned to reproduce better correlation at the expense of losing variance. While BHM has been shown to perform well for temperatures, it relies heavily on prescribed spatial correlation lengths. While this assumption is valid for temperature, it is hardly warranted for precipitation. In general, none of the methods outperforms the other. All experiments agree that a dense and regularly distributed proxy network is required to reconstruct precipitation accurately, reflecting its high spatial and temporal variability. This is especially true in summer, when a specifically short de-correlation distance from the proxy location is caused by localised summertime convective precipitation events.

Keywords Precipitation · Palaeoclimate · Climate reconstruction · Regional climate modelling · Proxy · PPE

J. J. Gómez-Navarro (✉)
Physics Institute and Oescher Centre for Climate Change Research, University of Bern, Sidlerstrasse 5, 3012 Bern, Switzerland
e-mail: gomez@climate.unibe.ch

J. J. Gómez-Navarro · S. Wagner · E. Zorita
Institute for Coastal Research, Helmholtz-Zentrum Geesthacht, Max-Plank-Strasse 1, Geesthacht, Germany
e-mail: sebastian.wagner@hzg.de

E. Zorita
e-mail: eduardo.zorita@hzg.de

J. Werner
Bjerknes Centre for Climate Research and Department of Earth Science, University of Bergen, 7800, 5020 Bergen, Norway
e-mail: johannes.werner@geo.uib.no

J. Werner · J. Luterbacher
Department of Geography, Climatology, Climate Dynamics and Climate Change, Justus Liebig University of Giessen, Senckenbergstrasse 1, 35390 Giessen, Germany
e-mail: Juerg.Luterbacher@geogr.uni-giessen.de

1 Introduction

Over the last decade, efforts have been devoted to develop paleoclimate reconstructions targeted at placing current climate change in a longer historical context (Masson-Delmotte et al. 2013 and references therein). Most efforts have been aimed at reconstructing variables such as the North Atlantic Oscillation (Luterbacher et al. 1999, 2001; Cook et al. 2002; Trouet et al. 2009) [see Pinto and Raible (2012) for a review] and temperature (Luterbacher et al. 2004; Xoplaki et al. 2005; Luterbacher et al. 2007; Riedwyl

et al. 2008; Mann et al. 2008; PAGES 2k Consortium 2013; Neukom et al. 2014 among others). Recent work has also been devoted to reconstruct different parts of the hydrological cycle. Gridded atlases of the Palmer Drought Severity Index (PDSI) have been developed for North America (Cook et al. 1999, 2004) and Asia (Cook et al. 2010). In South America (Neukom et al. 2010) undertook a spatially resolved reconstruction for precipitation extending back to 1500 AD by using the Principal Component Regression (PCR) method. Over Europe, precipitation has been reconstructed back to 1500 AD using a variety of methodologies (Pauling et al. 2006; Casty et al. 2007), and currently the so called Old World Drought Atlas is being developed to estimate PDSI evolution back to 1200 AD based on tree ring records (Cook 2013).

Climate reconstructions demand the careful collection and selection of proxy indicators whose evolution can be linked to specific climatic variables through a statistical model. For this reason, a number of statistical techniques have been applied to describe the connections between proxies and the climatic variables (Tingley et al. 2012). Once a statistical link has been identified within the instrumental period and verified with independent data and assuming stationarity, past climate can be reconstructed from the much longer proxy time series. This is the so-called inversion problem [see also discussion in Phipps et al. (2013)]. The approaches may be very different depending on the nature of the proxy indicator, such as tree rings, boreholes, lake sediments, speleothems or documentary evidence. Likewise, there exist a family of methodologies that allow merging and extrapolating various independent local reconstructions in order to create a gridded data set over a large area. The latter approach is generally referred as Climate Field Reconstruction (CFR).

The reliability of climate reconstructions depends to a great extent on the assessment of the skill of the statistical methods. One way of gauging the skill of a reconstruction method is through the concept of pseudo-reality [see Smerdon (2012) for a review]. The idea is to use climate models as a testbed to test the reconstruction method. For this, a number of pseudoproxies are created by contaminating the variables simulated by a climate model with statistical noise to mimic the uncertain relationship between proxy records and local climate. The reconstruction methods are then applied to these data to produce a reconstruction. Finally, this reconstruction is compared with the original simulation, which is perfectly known, and the skill of the method can be so estimated. This approach is known as pseudo proxy experiments (PPE). Albeit with an overall emphasis on temperature, successful applications of PPEs have been reported in the literature. In the following we briefly introduce some experiments specifically conducted for Europe, also using the methods we evaluate in this analysis. Küttel et al. (2007) evaluated the skill of European winter temperature reconstruction over the last 500 years using pseudoproxies from the ECHO-G and HadCM3 climate models. The emphasis was on the effect of the reduction of the number of proxies back in time. They found that the key factor in determining the reconstruction skill is the number of predictors and, particularly, their spatial distribution. Riedwyl et al. (2008, 2009) presented European scale seasonal temperature reconstructions over the past centuries applying and comparing different reconstruction methods. They used PCR and regularized expectation maximization (RegEM) in a PPE using the output from the NCAR Climate System Model (CSM) 1.4 and ECHO-G. The pseudo proxies were created using different white and red noise scenarios to contaminate the climate model output. They concluded that more skilful results are achieved with RegEM, as low frequency variability is better preserved. Smerdon et al. (2010) used the same NCAR global model to assess the skill of two CFR methodologies for global annual temperatures, RegEM and Canonical Correlation Analysis (CCA), whereas Werner et al. (2013) conducted a similar analysis comparing CCA with a Bayesian Hierarchical Model (BHM) over Europe. Generally, the former analysis indicated that RegEM and PCR techniques underestimate the temperature variations to an increasing extent as more noise is added to the proxies, albeit the effect in RegEM is weaker than in PCR. This effect for regression based reconstructions has also been reported in other studies (Smerdon et al. 2010; Tingley and Huybers 2010a; Werner et al. 2013) and is typical for linear regression. The Analog Method (AM) as a CFR method of European temperatures based on long climate simulations has also been evaluated by Franke et al. (2010). They found that a reduction in the number of proxy locations leads to relatively minor changes in the resulting fields, which would imply a still reasonable reconstruction skill for temperature even when the number of proxies is limited.

For the construction of meaningful PPEs it is critical that the model used as a testbed correctly reproduces the relationship between large-scale circulation patterns and small scale features such as orography or land use. Note that there is no need to correctly simulate time series that follow the spatio-temporal evolution of the real climate. Indeed this would likely be impossible due to internal variability, even if a model were able to perfectly represent nature (Gómez-Navarro et al. 2012). Instead, for the PPE to have meaningful results the resemblance to the real world needs rather to be of statistical nature: the canonical relationship between the local variance at the pseudo proxy sites and the large-scale circulation needs to be realistically reproduced/simulated by the climate model. However, GCMs can hardly reproduce these relationships satisfactorily for precipitation, which is the specific focus of this

study. This is due to the fact that precipitation, compared to temperature, is a more complex variable, generally with a skewed distribution and anistropic spatial covariance. Therefore, downscaling techniques have to be applied prior to the implementation of PPEs. Here we use a Regional Climate Model (RCM) as a downscaling tool for palaeoclimate simulations, as described by Gómez-Navarro et al. (2013). Specifically, we use a high-resolution climate simulation spanning the last two millennia as a testbed to conduct a number of precipitation-based PPEs to reconstruct precipitation fields. The construction of a meaningful PPE also requires taking into account the uncertainties in the proxy-climate connection. To account for this uncertainty, different levels and characteristics of statistical noise are added to the simulated series to perturb the original series, making the PPE more realistic.

In this study, a number of PPEs are performed to evaluate the skill of three different CFR techniques to reconstruct precipitation over Europe: CCA, BHM and AM. The paper is structured as follows. Sections 2.1 and 2.2 describe the simulations used as testbeds and PPEs. Section 2.3 describes how the PPE have been designed in terms of network and level of noise. Sections 2.4–2.6 outline the three CFR techniques tested in this study, whereas Sect. 2.7 describes all the statistics employed to assess the skill of the reconstruction techniques. In Sect. 3.1 several skill scores are employed to evaluate the skill of CCA and the AM in a perfect scenario where the pseudoproxies contain no noise at all. Section 3.2 discusses the effect of including different kinds of noise in the proxies. Section 3.3 examines to what extent the AM is sensible to the choice of analog pool, an aspect that is critical to apply this method in real-world reconstructions. Finally, Sect. 3.4 analyses the performance of the BHM method, comparing its capabilities and drawbacks with respect to the other two methods. Section 4 summarises and discusses the main results.

2 Database and methodologies

2.1 Setup of the simulation used as a testbed for the PPEs

The testbed for the PPE performed in this study represents a high-resolution regional climate palaeosimulation. The target variable of the PPEs is seasonally accumulated total precipitation. The simulated domain covers Europe with a horizontal resolution of 45 km. The simulation spans the period 1–1998, being the longest regional transient simulation for Europe to date: a previous simulation covered the last five centuries (Gómez-Navarro et al. 2013, 2014). The Regional Climate Model consists of a modified version of the meteorological model MM5 driven at the boundaries by the ECHO-G model. Both models are driven by estimates of three external forcings: greenhouse gas concentrations in the atmosphere, long-term variations in Total Solar Irradiance (TSI) and variations of Earth's orbital parameters. The effect of volcanic forcing is not included in these two simulations, globally and regionally, due to the lack of reliable estimations of volcanic forcing in the first millennium. The spatial and physical configuration of the ECHO-G and MM5 models is very similar to that employed by Gómez-Navarro et al. (2013). This set-up is able to realistically reproduce the seasonal cycle of precipitation, although some deviations from observations are apparent. The most important relate to the overestimation of the zonal circulation, which leads to a overestimation of precipitation in winter, as well as some biases in areas of complex topography such as the Alps or the coast of Norway, where the 45 km resolution implemented in the RCM is not fine enough (Gómez-Navarro et al. 2013). This simulation is referred hereinafter as the MM5 run.

2.2 Setup of the simulation used for the AM validation

An additional simulation is used in this study to test the ability of the AM to reconstruct precipitation using a pool of analogs from a different simulation. For this, the regional climate model COSCMO-CCLM was used, driven at the boundaries by the Earth Model of the MPI in Hamburg (MPI-ESM). The MPI-ESM model and the respective set-up are described in detail in the study of Jungclaus et al. (2010). As the RCM run outlined in the former section, this simulation includes changes in orbital and greenhouse forcings, as well as variations in the TSI (Flückiger et al. 2002). Addtionally, it also includes the effect of land use changes (Pongratz et al. 2008). The simulated domain consists of a rotated grid with a horizontal resolution of $0.44° \times 0.44°$, which roughly covers the same domain as the MM5 run and with very similar resolution (see Fig. 6). This simulation is, however, shorter and spans the period 1645–2000. Although it improves the skill of the driving GCM, the RCM keeps systematic deviations from the observations. The Mediterranean region shows little precipitation during summer, which is ultimately responsible for the lack of variability discussed below. However over Europe precipitation shows a positive bias year-round. The precipitation bias is mostly pronounced during winter with too much precipitation over central Europe including the Alpine region. This simulation is referred hereinafter as the CCLM run.

2.3 Design of the PPE

The PPEs require a synthetic proxy network to be hypothesised. The construction of this network depends on the focus of the study and can range from being purely random to a more realistic distribution of proxies. For temperature,

the network by Mann et al. (1998), updated by Mann et al. (2008), has been used in the contexts of PPEs (von Storch et al. 2008; Smerdon et al. 2010; Werner et al. 2013 among others). However, there is currently no standard network available for precipitation. In this study we use a simple network consisting of eleven locations distributed over Europe, which are indicated with stars in the Figs. 2, 3, 4, 6 and 7. This is a fundamentally synthetic network, but is based on the approximate location of proxies collected under the umbrella of the ongoing PAGES2kEuroMed initiative (cf. PAGES News April 2014). The number of pseudoproxies, though relatively small, tries to compromise between using locations suitable for producing meaningful CFRs, and using a limited number of pseudoproxies that could mimic the current scarcity of quality-proven hydrologic proxies.

There is a number of ways in which this network can be improved to make it more realistic. First, the location and number of pseudoproxies is just an approximation of the eventual network of real proxies available. Further, it does not consider proxy records that do not cover the full period, but instead the number of series is kept at a constant level of eleven for the whole PPE. These limitations do not invalidate the results of the PPE analysed here, but clearly limit the scope of the resulting conclusions. For example, all results that are specific of certain areas can not be extrapolated to real applications, whereas the distance at which the reconstruction skill becomes negligible, or its seasonal deviations, contains valuable information. The analysis of the sensitivity of the results to the selection of a more realistic grid, as well as the implementation of a number of proxies varying with time, is delayed for future assessments.

Another aspect of the PPE design pertains to the introduction of noise. Real proxy records contain substantial amounts of variations that are not climate-related. This introduces uncertainty in the local reconstruction, which eventually propagates into the CFR. In order to take this additional uncertainty into account and to analyse its impact on the reconstructions of precipitation, we can emulate more realistic proxies by including additive statistical noise to the original series, creating so-called noisy pseudoproxies (Tingley et al. 2012; Smerdon 2012). There are different kinds of noise we can use to contaminate the series. First, we use Gaussian white noise, which is the most simple choice and basically consists of a series of independent and identically distirbuted (IID) serially uncorrelated Gaussian noise added to the original series. The amount of noise introduced depends on the variance of the random process relative to the variance of the original series. Although equivalent, there are different ways to quantify how much noise the series contain: the variance of the noise, signal-to-noise ratio, percent noise by variance or correlation (Smerdon 2012). We use the latter, adding an amount of noise such that the correlation in each location between the original series and the contaminated one is 0.5. This level of noise is the same as employed by (von Storch et al. 2008) in the context of temperature PPE, and correspond to approximately with the level of noise reported in real proxies (Pauling et al. 2006; Dorado-Liñán et al. 2012). Further, it fits within the range of the level of noise employed in other studies for temperature (Xoplaki et al. 2005; Küttel et al. 2007; Smerdon 2012; Werner et al. 2013).

Note that while the use of white noise is widespread, the non-climatic related noise present in real proxies likely contains a certain amount of memory. For instance, in the case of tree-rings, the age-detrending methods may introduce serially correlated errors. Also, biological or chemical processes within the proxy, like infections, mutual competition, fire, etc., may also have life-times that extend beyond the usual time step of one year of tree-ring chronologies Frank et al. (2007). For this reason this study also analyses the effect of contaminating the perfect pseudoproxies with red noise, created with an AR(1) process with a decorrelation time of five years (von Storch and Zwiers 2002). The amount of noise relative to the original series is the same, so that the correlation with the original series is still 0.5.

A last point potentially complicating the reconstruction of hydrological fields is the sensitivity of the proxy to changes of climatic variables. Here we only use an ideal case for precipitation, but a recent study (Bunde et al. 2013) puts into question the connection between proxies and precipitation and argues for other implications such as soil moisture, integrating the effect of precipitation, specific soil characteristics and evaporation. For this study however, we set out to reconstruct precipitation fields based on our synthetic network represented by eleven pseudoproxies.

2.4 Canonical correlation analysis reconstruction

This study analyses the skill of three CFR techniques using PPEs based on the MM5 run. The first one is CCA, taken as one example of the general class of linear CFR methods, which rely on a multivariate linear relationship between the predictor (the proxy time series) and the predictand (the spatially resolved climate fields to be reconstructed) (Luterbacher et al. 2000, 2004; Xoplaki et al. 2005; Pauling et al. 2006; Riedwyl et al. 2008, 2009; Küttel et al. 2010; Smerdon et al. 2010; Werner et al. 2013 among others). The advantages of this method are its low computational cost and intuitive appeal, although some important caveats are discussed below. We briefly describe here the CCA approach.

Following the notation by Smerdon et al. (2010), the n observed time steps of a given climate field $\mathbf{T}$ of spatial

dimension m can be represented by a $m \times n$ matrix, and expressed in terms of the normalised series as

$$\mathbf{T} = \mathbf{M_T} + \mathbf{S_T}\mathbf{T}', \tag{1}$$

where $\mathbf{M_T}$ is a matrix of identical columns which are the average of $\mathbf{T}$ by rows, and $\mathbf{S_T}$ represents a diagonal matrix where each element is the standard deviation in one of the m locations. Analogously, the simultaneous network of proxies in r locations can be described by a $r \times n$ matrix

$$\mathbf{P} = \mathbf{M_P} + \mathbf{S_P}\mathbf{P}'. \tag{2}$$

With this notation, the multivariate linear hypothesis takes the simple form

$$\mathbf{T}' = \mathbf{B}\mathbf{P}' + \epsilon, \tag{3}$$

where $\mathbf{B}$ is the matrix of regression coefficients to be determined by the data, and ϵ represents the variability not explained by a linear relation between the predictand and predictor.

Hence, by combining (1), (2) and (3), we obtain

$$\mathbf{T} = \mathbf{M_T} + \mathbf{S_T}\mathbf{B}\mathbf{S_P}^{-1}(\mathbf{P} - \mathbf{M_P}), \tag{4}$$

which indicates that once $\mathbf{B}$ is estimated from the data during the calibration period, it can be used to estimate $\mathbf{T}$ as a linear combination of elements of $\mathbf{P}$. More importantly, if we assume that this relation holds true beyond the calibration period, it can be used to predict $\mathbf{T}$ for other periods where $\mathbf{P}$ is also known.

It can be demonstrated (von Storch and Zwiers 2002) that the error ϵ in (3) can be minimised if $\mathbf{B}$ is chosen as

$$\mathbf{B} = (\mathbf{T}'\mathbf{P}'^{\dagger})(\mathbf{P}'\mathbf{P}'^{\dagger})^{-1}, \tag{5}$$

where the superscript † denotes the transpose. However, when the sample size number is too small or the dimensionality is too large, the solution of Eq. (5) needs some form of regularization to ensure that the cross-covariance matrix $\mathbf{T}'\mathbf{P}'^{\dagger}$ and the covariance matrix $\mathbf{P}'\mathbf{P}'^{\dagger}$ are properly estimated from the data. This is the inversion problem further discussed by, among others, Smerdon et al. (2010), Tingley et al. (2012), Phipps et al. (2013). This regularization can be achieved by a previous EOF analysis to obtain a reduced-rank representation of $\mathbf{T}'$ and $\mathbf{P}$ by retaining only the leading modes of variability of each variable (von Storch and Zwiers 2002). Then, the EOF patterns are linearly combined to identify pairs of patterns which produce a temporal decomposition of $\mathbf{T}'$ and $\mathbf{P}'$ that simultaneously maximise their correlation but are uncorrelated with the rest of pairs. Once these pairs are established during the calibration, new instances of $\mathbf{P}'$ can be decomposed in the canonical patterns. The corresponding canonical series, scaled by the canonical correlation, can be used to reconstruct a standardised version of the predicted variable. After suitable renormalization, see (4), $\mathbf{T}$ is estimated.

The reduction of rank can be achieved by three truncation selections: the rank reduction of the pseudoproxy matrix; the rank reduction of the target climate matrix; and the choice of the number of canonical coefficients to retain in the search of the canonical pairs. The first two correspond to the number of retained EOFs, whereas the latter must be less than or equal to the smallest of the first two rank reductions (Smerdon et al. 2010). Another decision is whether to calculate the EOFs directly based on the original series, or to normalise them prior to the calculation. The latter approach avoids that locations with higher variability dominate in the calculation of EOFs, which can be problematic in fields with heterogeneous variability such as precipitation. Several tests have been performed to analyse the sensitivity of the reconstruction skill on these modifications. The results, not discussed here for the sake of brevity, indicate that the reconstruction is not very sensitive to these choices in terms of the measures of skill described below. Thus, the set-up employed hereinafter to compare with the other methods consists of retaining 5 EOFs in the climate matrix and 11 EOFs in the pseudoproxy matrix (that is, we do not reduce the rank in the pseudoproxy matrix) and hence use 5 canonical coefficients. Finally the original series, without previous normalization, are employed in the calculation of the EOFs.

2.5 The Bayesian hierarchical method reconstruction

A Bayesian inference on a hierarchy of models (or Bayesian Hierarchical Models, BHM) has also been employed. So far this method has been tested in PPEs for temperature (Werner et al. 2013; Tingley and Huybers 2010a) and precipitation (Gómez-Navarro et al. 2014) and successfully applied to reconstruct Arctic (Tingley and Huybers 2013) and European (PAGES 2k Consortium 2013) temperature anomalies. An advantage of BHM over the CCA is that it performs well even when data availability changes in time and data are in general sparse. More importantly, the underlying theory to assess error estimates is more transparent and more flexible compared to classical statistical methods. However, Bayesian methods are computationally expensive and reconstruction performance critically depends on the suitability of the adopted statistical models. Misspecified models can even cause the failure of the Bayesian estimation using the usual numerical methods (Markov Chain Monte-Carlo sampling), but even then analysing the reasons for this failure provides indications as to which part of the model would need to be improved.

The reconstruction using the BHM is based on the methodology dubbed BARCAST and developed by Tingley and Huybers (2010a, b), also used by Werner et al. (2013) and Tingley and Huybers (2013). In this method, a hierarchy of stochastic models is prescribed to model the climate,

instrumental and proxy data: a stochastic model for the time evolution of the climate variable is used. The climate variables are then used as input in stochastic models for the proxy and instrumental data. In this simple implementation, the models for proxy $\mathbf{W}_{P,t}$ and instrumental $\mathbf{W}_{I,t}$ data are just linear response functions of the climate variable $\mathbf{V}_t$, here the local precipitation, at the site of the instrumental measurements or of the proxy data, respectively. The climate field variable $\mathbf{V}_t \in \mathbf{R}^N$ at all of the N locations (i) at time steps $t \in (1, 1998)$ and the input data are described by

$$\begin{aligned} \mathbf{V}_{t+1} - \mu &= \alpha(\mathbf{V}_t - \mu) + \epsilon_{V,t} \\ \mathbf{W}_{I,t} &= \mathbf{H}_{I,t}(\mathbf{V}_t + \epsilon_{I,t}) \\ \mathbf{W}_{P,t} &= \mathbf{H}_{P,t}(\beta_2 W_{P,t-1} + \beta_1 \mathbf{V}_t + \beta_0 + \epsilon_{P,t}) \end{aligned} \tag{6}$$

Here, we have introduced a change to the original version of Tingley and Huybers (2010a, b). The non-climatic noise in the proxy data may depend on its past state, introducing an autocorrelated error into the system. Assuming an autoregressive process of order 1, the corresponding parameter β_2 is held at zero for the white noise PPEs. For the red noise PPEs, β_2 is estimated in Sect. 3.2. The matrices $\mathbf{H}_{I/P,t} \in \mathbf{R}^{N\times N}$ are diagonal. The entries are one at position (i, i) when an observation in year t at location (i) was made and zero otherwise. The stochastic terms denoted by $\epsilon_{P,t}$ and $\epsilon_{I,t}$ are multivariate normal with a diagonal covariance matrix $\mathbf{1}\tau_P^2$ and $\mathbf{1}\tau_I^2$. They are used to model the local noise in the proxy response function and the errors in the instrumental observations. The interannual climate variability is described by the multivariate normal term $\epsilon_{V,t} \sim N(0, \Sigma)$, where the spatial covariance matrix $\Sigma \in \mathbf{R}^{N\times N}$ is given by

$$(\Sigma)_{i,j} = \sigma^2 \cdot \exp(-\phi|x_i - x_j|) \tag{7}$$

It is clear that this homogeneous covariance structure is an oversimplification of the covariance estimated from observations, and the results of this simplification show up in the reconstruction experiments (Sect. 3.4). Likely, this is one aspect of the model that would require attention in a more realistic statistical model. One possible change to the covariance structure would be the inclusion of spatial inhomogeneities and anisotropy, possibly based on a decomposition of the observed covariance structure of the instrumental period (see Tingley et al. 2012; Werner et al. 2013). While it is still possible to write down the posterior PDF of the covariance matrix without any prescribed structure, it has been suggested (Gelman et al. 2003) that it can be difficult, or even impossible, to achieve convergence in the estimation.

From the model Eq. (6) the conditional probability density functions (PDFs) can be derived for all variables—precipitation depending on the instrumental and proxy data as well as the system parameters [Greek symbols in (6)], but also for all system parameters dependent on the variables. For the conditional PDFs the reader is referred to the appendix in Tingley and Huybers (2010a, b). A Gibbs sampler is then used to iteratively update estimates for the variables and parameters. This procedure is run in parallel in up to five chains. The resulting draws of the conditional PDFs are checked for convergence after 5,000 iteration steps, using convergence measures such as the potential scale reduction factor (Gelman et al. 2003). If convergence was unlikely, the Gibbs sampler was run for an additional 10,000 steps. Following the argumentation in Gelman et al. (2003), single chains were removed afterwards if they converged to non-physical parameters. This may be required if the basin of attraction of the optimal solution is not infinitely big—as is a typical feature of nonlinear optimisation problems. The results from the other chains are then thinned in time to remove the auto correlation present between subsequent draws. The remaining draws form ensembles of reconstruction. Each of the remaining draws is equally probable and self consistent: The spatio temporal dependences of the climate field variable and the dependence of the proxy and instrumental data on it are given by the corresponding parameters of these draws. Each of the draws can be interpreted as a draw from the full (multivariate) joint PDF in the space of all climate variables at each time step and all locations and all parameters or, more close to the language of the climate research community, as equally probable ensemble members. This also means that when trying to compare the variability of the reconstructions to that of the target, the (estimated) distribution of the variability of the individual ensemble members can be used to directly state confidence intervals. This rather direct method of creating impartial uncertainty estimates is one of the main advantages of using Bayesian inference. Another advantage is the possibility of, at least theoretically, implementing arbitrarily complex models, although the analytical and computational effort needed in such cases can be prohibitively high.

In order to apply the model (6) to precipitation data, one obstacle has to be overcome: a joint multivariate normal process is assumed in (6), which clearly is not applicable for precipitation data in many cases—especially in more arid regions, the distribution of seasonal and annual precipitation can be skewed. Thus, the input data needs to be previously transformed. This is done by estimating parameters for a gamma distribution for the local seasonal precipitation at each grid point from the full climate model output. Clearly, this would be unfeasible for short real world instrumental data and underlines the best case nature of this experiment. Then, the data is transformed to a normal distribution using the corresponding probability functions, and the data at the locations and time steps for the proxy and instrumental data in the PPE is selected. Afterwards, the reconstruction is attempted and the results are finally transformed back to the original units of measurement.

2.6 The analog method reconstruction

Although BHM is a non-parametric approach, the version adopted here uses a parametric model. Thus, it shares an important drawback with CCA: their parametric approach based on the IID assumption is roughly valid for temperature. However, for hydrological variables, especially those within climates such as the Mediterranean, this assumption may not be fulfilled. Therefore, non-parametric approaches are in principle more suitable to account for the processes controlling the hydrological cycle. As an example of a simple yet non-parametric and non-linear method, the AM has been applied and tested as well. Although it was first introduced in the 1970s for weather forecasting (Lorenz 1969), it was not widely adopted partly due to the need for sufficiently long observational records to properly sample the large dimensional space of climatic variables. However, a number of variations of the method have been developed to overcome this problem: the number of degrees of freedom can for example be reduced by an EOF analysis or CCA (Zorita and von Storch 1999; Fernández and Sáenz 2003; Xoplaki et al. 2004). One application of this method in climate research has been as a statistical downscaling technique, where the large scale fields of Global Circulation Models are used as predictors and regional variables such as precipitation are the predictands. Used in this way, the method has been shown to produce results comparable to more complex techniques (Zorita and von Storch 1999). However this technique can also be used as an upscaling tool, linking local information of a climatic variable with the corresponding large-scale gridded field. Applied this way, the method can be regarded as a CFR technique. The skill of this approach to reconstruct temperature fields has been recently examined by Franke et al. (2010) and Luterbacher et al. (2010) in a number of PPEs conducted within the virtual world of GCM simulations. Similarly, Schenk and Zorita (2012) used it to generate a high resolution atmospheric reconstruction for Northern Europe based on a realistic regional simulation driven by reanalysis and few local quality controlled series of temperature and sea-level-pressure.

The basic idea of the method is as follows: We assume that a set of observations of the multivariate predictand $\mathbf{T}(t)$ is available over some short time, with concurrent observations of a multivariate predictor $\mathbf{P}(t)$. This predictor is also available at time t_0 where no observations of the predictand, the target field variable, are present. The AM assumes that the value of these unknown $\mathbf{T}(t_0)$ can be approximated by a known value of $\mathbf{T}(t)$ if the predictors $\mathbf{P}(t)$ and $\mathbf{P}(t_0)$ at the target time t_0 and a time t in the observation period are sufficiently similar. The set of values $\mathbf{P}(t)$ with the simultaneous information of the predictand $\mathbf{T}(t)$ is generally denoted the pool of potential analogs. At a given time t_0, the method compares $\mathbf{P}(t_0)$ with all the pool members by calculating a distance measure to be defined later

$$\Delta(t_i) = \mathrm{dist}(\mathbf{P}(t_0), \mathbf{P}(t_i)). \tag{8}$$

The element (or elements) in the pool with a small distance $\Delta(t_i)$ are called the analog(s), $\mathbf{P}(\tilde{t}_i)$. The reconstructed predictand can simply be defined as the value of the predictand at the analog point in time which minimises the distance $\mathbf{T}(t_0) = \mathbf{T}(\tilde{t}_i)$.

Although the basic idea is simple, it can be tailored to fit to different requirements by choosing a suitable distance measure in (8). One of the most simple and intuitive is the Euclidian distance,

$$\mathrm{dist}(\mathbf{P}(t_0), \mathbf{P}(t_i)) = \sqrt{\sum_{j=1}^{r} (P^j(t_0) - P^j(t_i)))^2} \tag{9}$$

where j represents the dimensions of the multivariate predictand. This distance has been compared with more sophisticated metrics by Matulla et al. (2007), and the results indicate that there is no metric that optimises the AM in every aspect. For this reason, we restrict ourselves hereafter to the Euclidian metric defined by (9).

As briefly discussed above, other variants do not directly use the variables $\mathbf{P}$ in the pool, but rather a rank-reduced version, searching the analogs in the EOF space or in the space spanned by canonical pairs identified in the pool (Fernández and Sáenz 2003). These variants are important because a small pool may compromise the skill of CFR due to the potentially large dimensionality of the predictor fields. Nevertheless as we use a very large pool (2000 simulated years), these limitations are absent in this study.

Another modification to the method is to not just select one analog, but to average several analogs to generate a reconstructed field. For example, the N most similar pool members [in the sense of the distance given by (8)] to $\mathbf{P}(t)$ can be used to produce a reconstructions by a weighted average

$$\tilde{\mathbf{T}}(t_0) = \sum_{i=1}^{N} \omega_i \mathbf{T}(\tilde{t}_i) \tag{10}$$

where $\mathbf{T}(\tilde{t}_i)$ denote the predictand fields of the closest analogs, possibly weighted by ω_i. These allow equal importance to all analogs or weights them according to their distance to the target. For the sake of simplicity, we uniformly weight all N closest analogs. Note that an expected counterpart of this averaging procedure is the loss of variance. The diminished variability is a function of the number of elements used to average N, and can be explained by the simple application of the Central Limit Theorem.

Note that an important prerequisite using the AM relates to the availability of having a large pool of realistic observational data to perform the search of analogs. In previous applications of the AM, for instance in weather prediction (Dool 1994) or statistical downscaling (Zorita and von Storch 1999) the pool of analogs consisted of gridded observations of large-scale fields, such as sea-level-pressure or geopotential height. However, in the framework of CFR the pool of analogs consists of fields of simulated precipitation during the last 2000 years (Gómez-Navarro et al. 2014). In this set-up it is important to note that the PPE is constructed on the same data set that is used as a pool for the search of analogs, the MM5 run. This leads to circularity, since the pool of analogs consists of the target that the method tries to reconstruct. This could lead to the overestimation of the skill of the AM compared to the former techniques (indeed, in the case of PPE without noise the AM would always trivially select as analog the original situation, leading to a perfect reconstruction). To circumvent this, the algorithm neglects those analogs whose time step the method is trying to reconstruct. Additionally, a sensitivity experiment was designed to assess the role of the target in the evaluation of the skill of the AM. For this, the CCLM run was used as a target for the PPE, whereas the pool used in the search of analogs still consisted of the MM5 run. The results of this experiment are described in Sect. 3.3.

2.7 Measures of skill

Several metrics were used to evaluate pseudo-reconstructions in past studies. We briefly describe the ones used herein. In all cases the procedure consisted of evaluating the resemblance between the reconstruction based on a limited number of pseudo proxies with the known "target". The first two measures of similarity are the Pearson correlation coefficient between reconstructed and target precipitation at each location evaluated over all the reconstructed period, and the ratio of the reconstructed variance vs. that of the target time series. Both estimators are defined e. g. by (von Storch et al. 2008). A word regarding the choice of correlation measure is due here. It is rather common in the hydrology literature to use the Spearman rank correlation, rather than the Pearson correlation. However, the difference between both measures of dependency between variables is small unless the relation between the variables is strongly non-linear or many outliers in the records exist (McDonald and Green 1960). As we are comparing seasonal series of simulated/reconstructed precipitation, this is not the case, and indeed the difference between the Pearson and Spearman correlation maps is negligible (not shown). Thus, we decided to use the former for the sake of simplicity and for being the most widely measure of dependency used in the literature.

2.7.1 *Reduction of error*

The Reduction of Error (RE) (Cook et al. 1994) measures to what extent the reconstruction better fits the target than a climatological mean derived over a reference period. It is defined as

$$\mathrm{RE} = 1 - \frac{\sum (T_i - \hat{T}_i)^2}{\sum (T_i - \bar{T}_i)^2} \tag{11}$$

where T_i and $\hat{T}_i$ are the target and the reconstructed values respectively for each grid point, respectively. $\bar{T}_i$ denotes the corresponding mean over the reference period. The values of RE range from $-\infty$ to 1. Negative values imply that the reconstruction is worse than a prediction based on the climatology. A positive RE, approaching 1 for a perfect reconstruction, indicates a skilful reconstruction.

2.7.2 *Contingency tables*

An important aspect of a climate reconstruction is its skill in reproducing not only the mean value and low frequency variability, but also the tails of the distributions. These are of major relevance for hydrological variables, since the seasonal precipitation outliers are directly related to important events such as droughts and flooding. In order to evaluate the skill of the reconstruction in reproducing the timing and severity of these events, we use a metric based on contingency tables. For this, we divide the precipitation at each location into three categories, from low to high. The "low" category consists of the lower 10th percentile, the "high" category are values above the 90th percentile of the precipitation series at that grid point. The bulk is considered to be normal. A 3×3 contingency table $\mathbf{C}$ is generated, where each element c_{ij} represents the percentage of times that the reconstruction showed category i when the target was in category j. The matrix $\mathbf{C}$ tends to be diagonal when the reconstruction is able to reproduce the timing of extreme seasons, whereas large off-diagonal entries indicate diminished skill.

This contingency table can be translated into a numerical score by defining a skill matrix $\mathbf{C}$, and defining the score as

$$s = \sum_{i,j=1}^{3} s_{ij} c_{ij} \tag{12}$$

This skill matrix can be defined in a variety of ways. Gandin and Murphy (1992) develop a general framework

to define equitable skill scores, in the sense that they not only penalise the lack of skill, but they also show no preference for forecasting any of the available categories over the others. In our case, we have three categories of events, with probabilities 0.1, 0.8 and 0.1 for "low", the bulk, and "height" events, respectively. Following Gandin and Murphy (1992), there is a free parameter that allows the definition of a whole family of skill matrices which produce desirable properties for the numerical score in Eq. 12. This parameter is the element s_{12}, which can take any value in the interval $-1/2 \leq s_{12} \leq 0$. We select the middle point, $s_{12} = -1/4$, resulting in the matrix

$$\mathbf{S} = \begin{pmatrix} 4.75 & -0.25 & -2.75 \\ -0.25 & 0.0625 & -0.25 \\ -2.75 & -0.25 & 4.75 \end{pmatrix} \tag{13}$$

This metric rewards the reconstruction when it successfully reproduces an extreme event, gives less credit when it reproduces a "normal" event, and penalises the mismatches. There are two types of mismatches, light and severe, with a penalisation that reflects its severity. Finally, the errors are considered symmetric, i. e. forecasting i when j was observed produces the same score as the opposite situation.

Note that although the Pearson correlation and this skill are, in principle, related, they might produce very different results if the ability of the reconstruction to reproduce the tails of the distribution is different to the skill in the middle of the distribution. An idealised example that illustrates such situations can be easily built by cloning a normal-distributed series, and then multiplying all its elements between the 10th and 90th percentiles by -1. The correlations between the modified and the original series is low, whereas the ability to reproduce the tails is indeed perfect. Thus, in order to gain further insight into the behaviour of this skill metric and how it is related to the correlation between reconstruction and target, we performed a simulation study with synthetic series obtained with a random number generator. In each Monte-Carlo step, we generated two series of 2000 standardised normal variables with certain known level of correlation. These two series were then categorised into three states, following the aforementioned criteria of thresholds, and the corresponding skill score was calculated according to Eqs. (12) and (13). For each level of correlation, 1,000 Monte-Carlo experiments were repeated, allowing an estimate of the distribution of the score. Figure 1 depicts the median skill score as a function of the correlation. The shaded area represents the 90 % confidence level based on these simulations. There is a monotonic, although non-linear, relationship between the skill score and the correlation calculated prior to the categorization of the series. The skill score is always lower than the correlation. The difference between both statistics is small in correlations close to 0 or to 1, and it is maximum

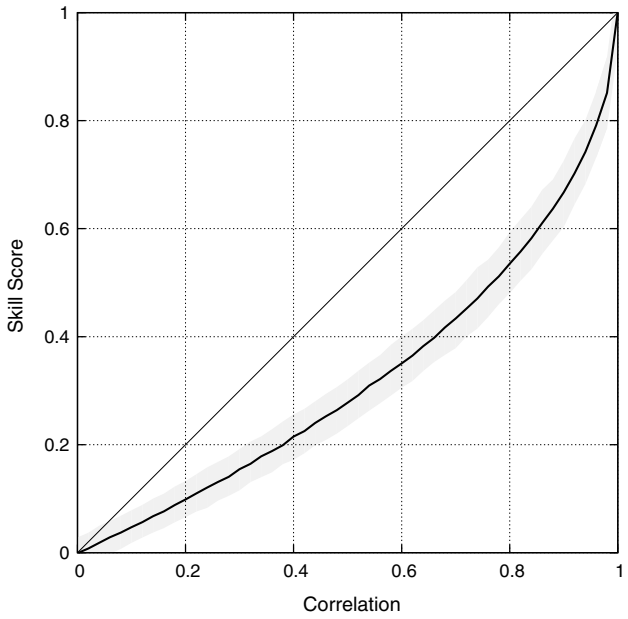

Fig. 1 Monte-Carlo simulation study of the relationship between skill score based in contingency tables and correlation. For each level of correlation, two series of 2,000 standardised normal variables are generated and the skill score is calculated after the categorization of such series. This experiment was repeated 1,000 times to calculate not only the mean skill score but its confidence interval. The figure depicts the median (*solid line*) and the 90 % confidence interval (*shaded*) of these 1,000 simulations for each level of correlation between 0 and 1. The one-to-one line is drawn to facilitate the visualization of the result

in intermediate correlations. Note, however, that this curve is only valid for normal variables. Other distributions, such as the precipitation series we analyse in this study, could, in principle, behave differently. Thus, Fig. 1 should not be considered as a trivial relationship between correlation and skill score valid for all probability distributions. Indeed, the analysis of the skill score allows for the analysis of the ability of the reconstruction to reproduce extreme periods by clustering all "normal" years into the same category.

3 Benchmarking the CFR techniques

In the following sections we assess and compare the skill of the AM and CCA as CFR methods for precipitation. We postpone the BHM evaluation to Sect. 3.4. The reasons for this separation is technical: The latter method is computationally very demanding, which limits the number of grid points the reconstruction can perform for in limited time. Due to these limitations the reconstructions performed with this method have to be done on a coarser resolution grid. The coarser grid is a 225 km spaced sub grid of the original one, excluding a strip around the boarder of the domain

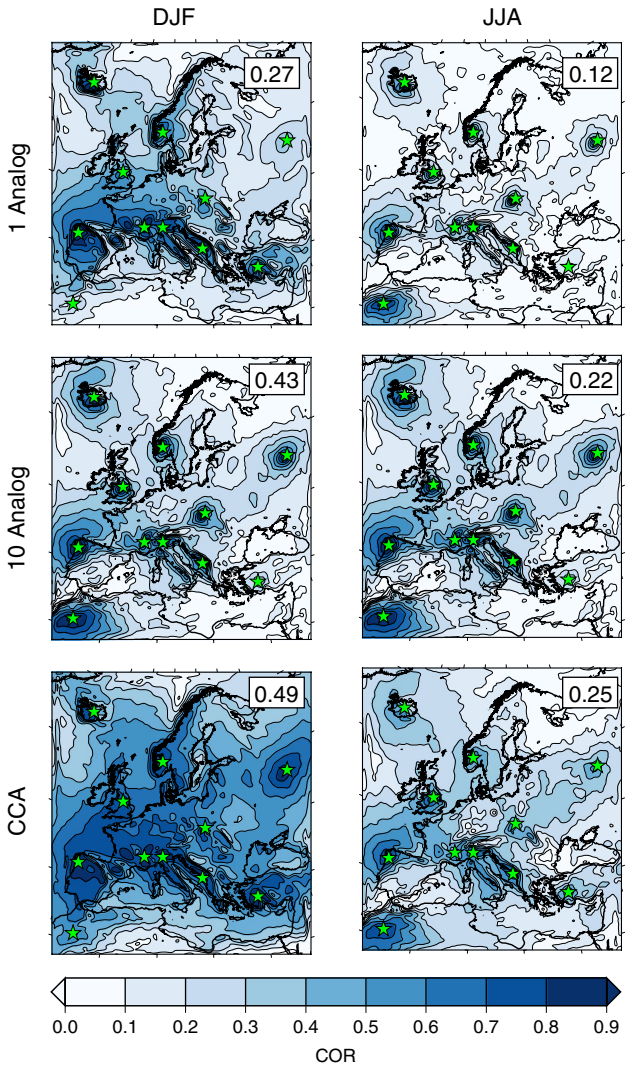

Fig. 2 Pearson correlation between the simulation and the reconstruction performed with pseudoproxies. The *green stars* indicate the location of the pseudoproxies. Reconstructions are performed for winter (*left*) and summer (*right*) with the AM using 1 analog (*top*), 10 analogs (*middle*) and CCA (*bottom*). The number in each *rectangle* indicates the spatial average of the corresponding value

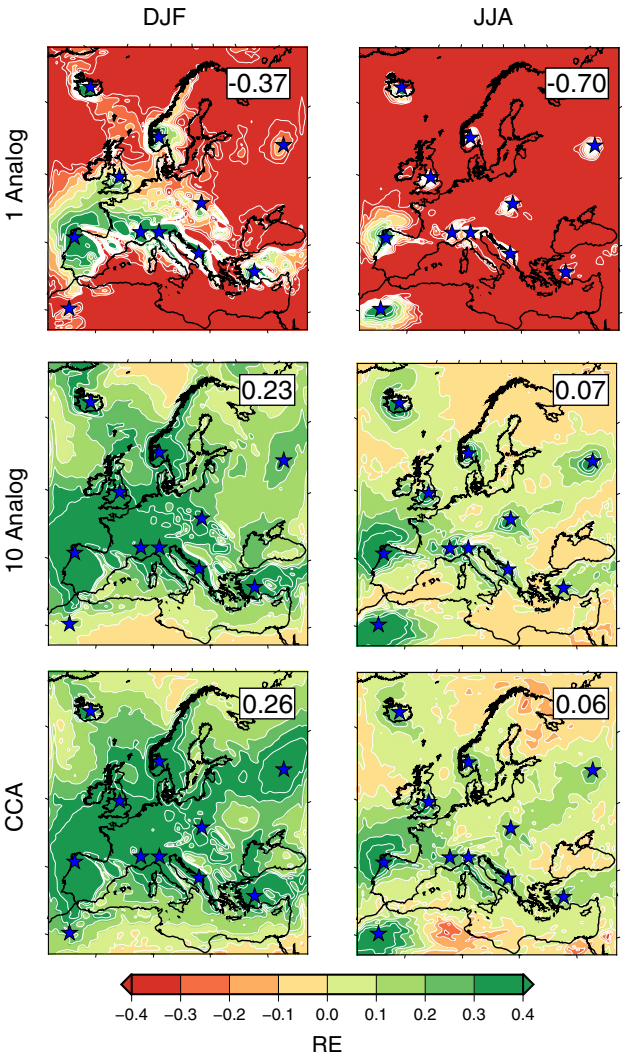

Fig. 3 As Fig. 1, but for the reduction of error (RE) statistics

of the RCM to avoid the boundary effects. This makes the comparison with the former reconstructions (which operate in the original 45 km grid) more difficult. Thus, we decided to evaluate the skill of this technique separately in a dedicated section.

3.1 Performance of the AM and CCA in the idealised case

A number of independent reconstructions based on AM and CCA have been performed for winter and summer based on the same network of pseudoproxies. Both techniques admit a number of variations which allow for the fine-tuning of their performance. Regarding CCA, the EOFs used to regularise can be obtained by using the covariance or the correlation matrix. By using the latter, the over-representation in the EOF analysis of locations where variability is larger is avoided. This is a desirable feature for precipitation, as the variability of this variable is directly related to its mean value, which is highly heterogeneous throughout the domain. By using the correlation matrix, all grid points contribute equally to the EOF patterns. Also the number of EOFs retained in the process can be adjusted using different criteria. Given that the number of pseudoproxies is 11, this puts an upper limit to the number of precipitation EOFs that can be retained for calculating the Canonical pairs. Using too few patterns produces the underestimation of the explained variance, whereas using too many could cause overfitting of the model, leaving too few degrees of freedom for the reconstruction. In this study we tested the performance of reconstructions

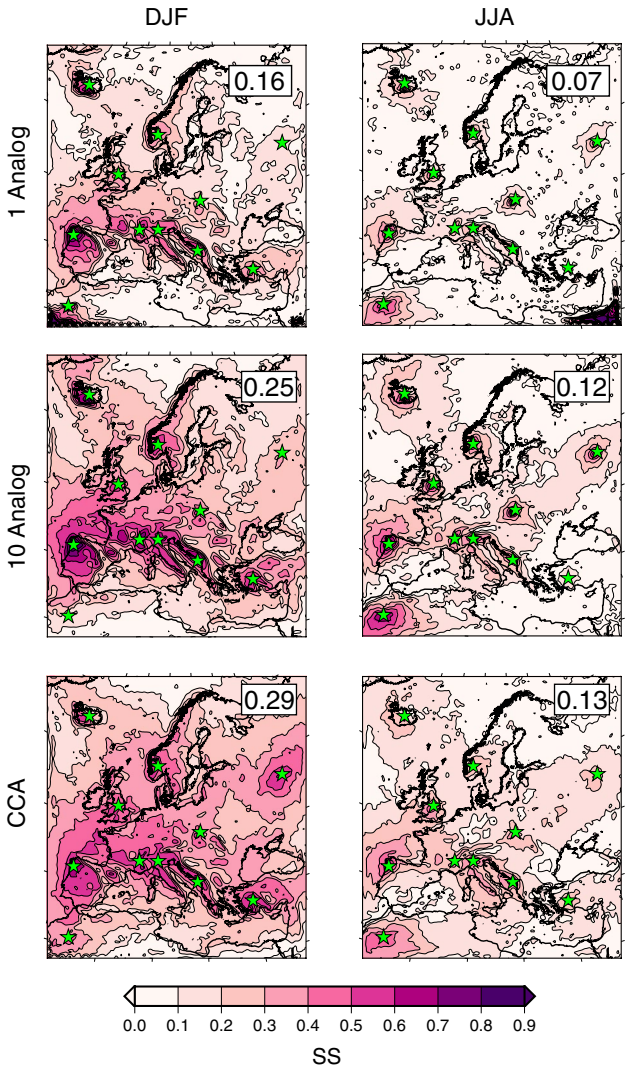

Fig. 4 As Fig. 1, but for the skill score as defined in Eq. (12)

when the 11 first EOFs were employed and when this number was reduced to only 5. We performed these reconstructions with the correlation and covariance matrices, resulting in a total of 4 reconstructions. The results (not shown) indicate that the difference between these 4 approaches is small, so for the sake of brevity only the CCA reconstruction retaining 5 EOFs (which explain roughly 50 % of the total variance) and using the correlation matrix is discussed. Regarding the AM, we tested several values of N in Eq. (10). The values of $N = 1$ and $N = 10$ are discussed here in comparison with the CCA method. The skill of these three reconstruction was evaluated by calculating the correlation (Fig. 2), (RE, Fig. 3) and a skill score (Fig. 4) based on contingency tables as defined by Eqs. (12) and (13). In all cases the pseudoproxies are perfect, i. e. they are the uncontaminated raw grid-cell time series produced by the climate model.

The correlation maps shown in Fig. 2 indicate that the spatially averaged linear relation is not very high (correlations below 0.5 in all cases). High correlations above 0.8 can be found in areas close to pseudoproxies, whereas very low correlations are in remote areas where no proxies are found. This tendency to aggregate skill in pseudoproxy locations has been previously reported among others by Smerdon et al. (2010), Li and Smerdon (2012) and Annan and Hargreaves (2013). Comparing the different methods, they all share a similar spatial structure, strongly influenced by the locations of the proxies. The major differences among the methods relate to the magnitude of correlations. CCA generally shows the highest values, whereas the 1-member analog reconstruction (only the closest analog) performs worst. Figure 2 also illustrates how the AM can easily be tailored to increase the correlation by increasing N. However, by doing this the temporal evolution of the reconstruction is smoother, resulting in the severe loss of temporal variance outlined above (not shown). Indeed, the spatial-averaged ratio of standard deviation between the reconstruction and the target is 0.96 and 0.98 for winter and summer, respectively, when $N = 1$, whereas it drops to 0.49 and 0.39 when $N = 10$. Thus, a trade-off between correlation and variance is established. CCA produces spatially-averaged variance ratios for winter and summer of 0.54 and 0.36, respectively. Hence, the 10-member AM performs in this aspect similar to CCA. However the AM is more flexible, and larger N could be used to increase the correlation at the expense of reducing the variance.

The RE values for winter and summer are depicted in Fig. 3. This parameter evaluates the compromise between the ability of the reconstruction to follow the temporal evolution and its variance through time compared with a reference period (chosen here as the twentieth century). The CCA performs similar to the 10-member analog reconstruction, showing a skillful reconstruction over large parts of the domain, especially in areas close to the proxy locations (indicated with blue stars, see Fig. 3). Likewise, the skill in summer considerably reduces over remote areas with respect to the proxy location. The 1-member AM clearly shows poor skill in most areas, associated to the lack of temporal agreement (Fig. 2).

Finally, the skill score based on categorical assessment is shown in Fig. 4. Recall that, contrary to correlation, this skill emphasises the accuracy of the reconstruction to reproduce the timing of infrequent precipitation seasons, not giving much credit for reproducing "normal" ones. The skill score shows a spatial structure which is very similar to the correlation maps shown in Fig. 2, albeit with lower numerical values as expected from the results of the analysis with the synthetic series shown in Fig. 1. The differences between methods and seasons are also apparent in this figure. If plotted against the correlation, a

strictly monotonic dependency between the two measures is revealed (not shown). Hence Fig. 4 demonstrates that the reconstructions are also able to correctly classify some of the outliers.

3.2 The effect of the inclusion of noise in the PPEs

The reconstructions analysed so far were based on perfect pseudoproxies. In this section we analyse the effect of contaminating the perfect pseudoproxies with statistical noise to make them more realistic. The effect of the introduction of noise is illustrated through the use of Taylor diagrams, which summarise the performance of different methods in a simple graph (Taylor 2001). These diagrams normally depict the performance of a single series with respect to a reference. Here, we use it to represent the spatially averaged skill, so the figure actually shows the spatially averaged correlation, ratio of variance, and centred RMSE. Note that, although the local variations of skill are crucial throughout this study, the spatial structure of the performance (particularly its relation with respect to the distance to the pseudoproxy locations) is very similar to that displayed in Figs. 2, 3 and 4, and omitted here for the sake of brevity.

Figure 5 shows the performance for the three methods discussed above (by rows) in winter and summer (by columns). Each diagram shows the averaged skill for perfect pseudoproxies (blue symbols), white noise pseudoproxies (green symbols) and red noise pseudoproxies (red symbols). Furthermore, we also examine the performance at different temporal scales. For this purpose we apply three time filters to the series prior to the calculation of the skill: unfiltered filter (cross), 31-year running mean (square) and 101-year running mean (triangle). First, as discussed above, the skill in the perfect case (blue crosses) is always higher for winter, indicating that in summer precipitation is more dependent on regional-scale processes, hampering the extrapolation of the proxies information for remote areas. Comparing these three methods, the 1-member analog reconstructions preserves the variance very well, but at the expense of low correlations. Using more analogs, the correlation increases but the variance is underestimated more and more. The skill of the AM becomes increasingly similar in terms of RMSE to CCA.

Regarding the different filtering of the series, it is noteworthy that regardless the CFR method, the nature of the noise (white or red) does not seem to be of relevance for the skill of the reconstructions at interannual scale (red and green crosses mostly overlap in all cases). However, for low-pass filtered precipitation series, the reconstructions with red noise tend to show better correlation but less variability than those contaminated with white noise. If real noise in proxies is indeed better approximated by red noise, this would indicate that the CFR methods will tend to result in higher correlation at low frequencies, but at the expense of a reduced variability at these frequency bands.

3.3 Sensitivity to the choice of the pool of analogs

It can be argued that the comparison performed between the CCA and the AM is biased, because the latter is based on a pool extracted from the same target simulation. The problem is that the quality of the pool is critical for the accuracy of this method. Systematic problems such as biases, underestimation of variance in certain locations, or wrong structure of the spatial covariance, are transferred from the pool to the final reconstruction. Hence, by using the same simulation as the target and as the source of the pool, these systematic problems could not be detected. The result is that the comparison of methods performed above can overestimate the capabilities of the AM method to reconstruct seasonal precipitation.

A simple way to investigate the role of the pool on the performance of the AM is to use a set of analogs that are not extracted from the same simulation as the targets. This can be achieved by attempting to reconstruct a different simulation, performed with a completely different model setup, but using the same pool we used above. For this purpose we use a simulation with the regional CCLM model described above, which spans the period 1645–2000. Both simulations are different: they have been carried out with different regional models, each driven by a different global model, and even encompass different domains. However both share most of the simulation domain (Europe), similar forcings and a similar spatial resolution. In summary, we use the CCLM simulation as an alternative reality to be reconstructed (with different systematic errors) based in the same pool of analogs as before.

The skill of the CFR for the CCLM simulation using the analog pool from the MM5 simulation is illustrated in Fig. 6, which shows the correlations and ratios of standard deviations between the CCLM simulation and its reconstruction based in the same network of pseudoproxies used so far. Here, the 1-member AM with perfect pseudoproxies was implemented, so that this figure can be compared only with the first rows in Figs. 2, 3 and 4. Most results discussed above regarding the correlation remain valid in this experiment. The reconstruction is only skillful close to the areas with proxy information. The ability to reproduce the precipitation field from local proxies is lower in the summer season. The average correlation in winter is similar but slightly higher (0.34 vs. 0.28), although in summer it is lower (0.05 vs. 0.11). As expected, the 10-member AM (not shown) increases the mean correlation in both seasons (0.53 and 0.13 for winter and summer, respectively). However, a different picture emerges when looking at the ratios of the standard deviation. The 1-member AM employed as CFR technique

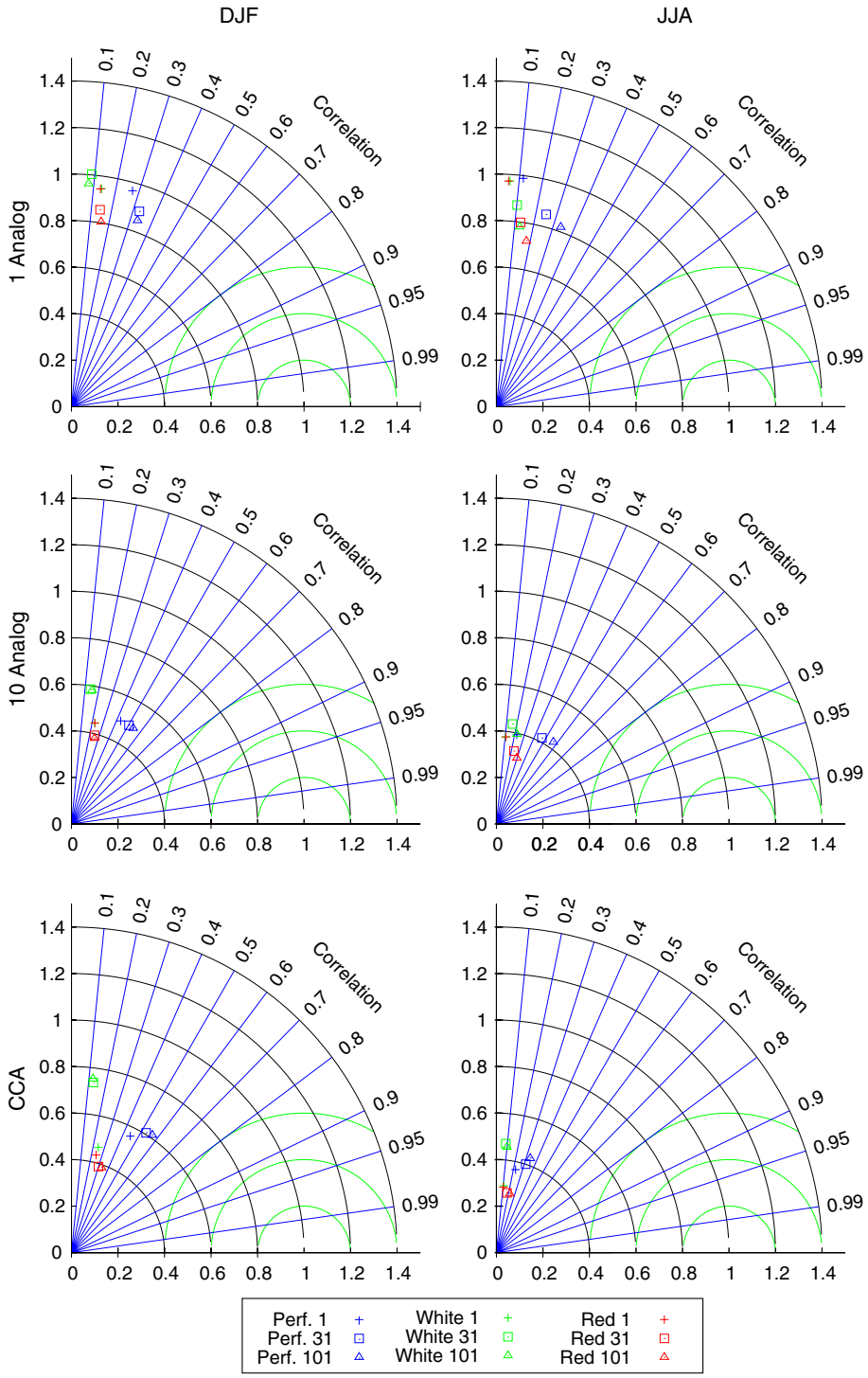

Fig. 5 Skill for the three CFR methods for winter and summer. Each Taylor diagram shows in polar coordinates the spatial-averaged correlation (the cosine of the angle in the polar graph) and the ratio of standard deviations between the reconstruction and the original simulation (radial distance). The distance to the point of coordinates (1, 0) is the spatial-averaged centered RMSE. Three different types of noise [no noise (*blue*), *white* noise (*green*) and *red* noise (*red*)] and three levels of smoothing (the number in the legend denotes the window used to calculate a running mean) are applied to analyse the influence of these factors in the reconstruction skill. See text for details

here preserves the variance of the pool, which may not be the same as the variance of the target. This is clearly shown in the second row of Fig. 6. In winter, the method tends to underestimate variance in the western part of the domain, whereas it overestimates it in the east and over the Mediterranean. This situation becomes worse in summer, where the method strongly overestimates the variability of the CCLM simulation in most areas except in the north of the domain (not shown). As before, when the 10-member AM is used, the overestimation of variance becomes smaller, and the method generally loses variance (not shown). The reason for this overestimation has to be sought in the variance simulated by the CCLM model used here as a testbed. The variance in the MM5 simulation has been evaluated against observations and shown to be realistic (Gómez-Navarro et al. 2013). However, the CCLM simulation shows an anomalous

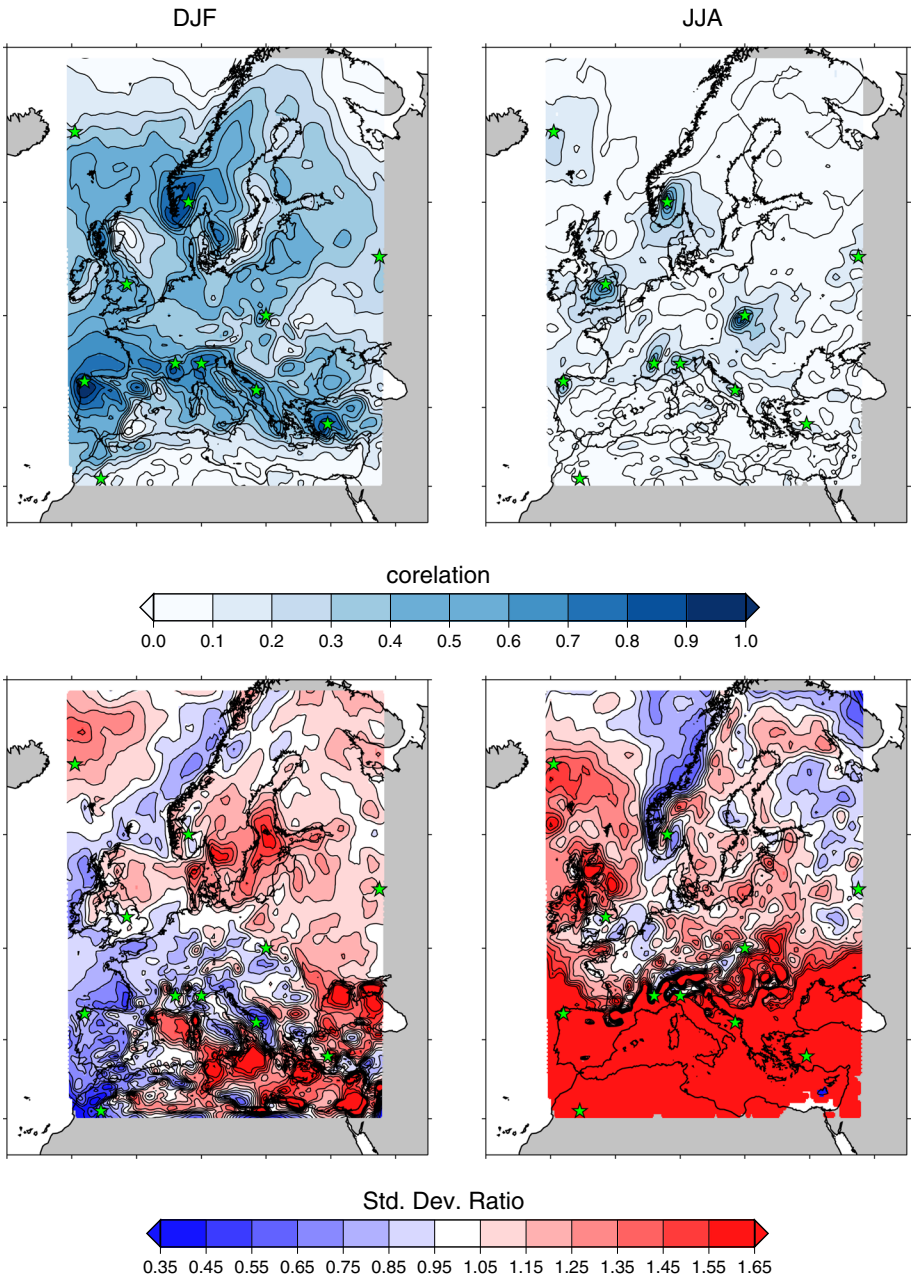

Fig. 6 Pearson correlation (*top*) and ratio of variance (*bottom*) between the CCLM simulation and the reconstruction of this simulation performed with analogs searched within the pool of 2000 years of precipitation simulation employed through this study. The *green stars*indicate the location of the pseudoproxies. Reconstructions are performed for winter (*left*) and summer (*right*) with the AM using 1 analog

and unrealistic smaller variance in the summer precipitation in the southern part of the domain (not shown). The AM is not able to find analogous situations for such unrealistically dry conditions in the pool, which results in this overestimation of the reconstructed variance.

3.4 Performance of the BHM as CFR technique for precipitation

This section assesses the BHM as CFR technique, and compares it with the AM and CCA techniques formerly discussed. As mentioned above, this CFR method has large computational requirements which limit its applicability within the high-resolution dataset we are using as testbed for the PPE. Thus, for this set of experiments we use a coarser resolution version of the original simulation at 225 km resolution. However we still keep the same network of pseudoproxies as in the exercises above in order to facilitate the comparison of the different methods.

The maps of correlation, RE, and skill score of the BHM reconstruction when no noise is included into the original series is shown in Fig. 7, and are the equivalents to Figs. 2, 3 and 4 for this method. Although the coarser resolution is noticeable, the main conclusions drawn from the assessment of the other two techniques still apply here. The areas where the reconstruction is closer to the simulation

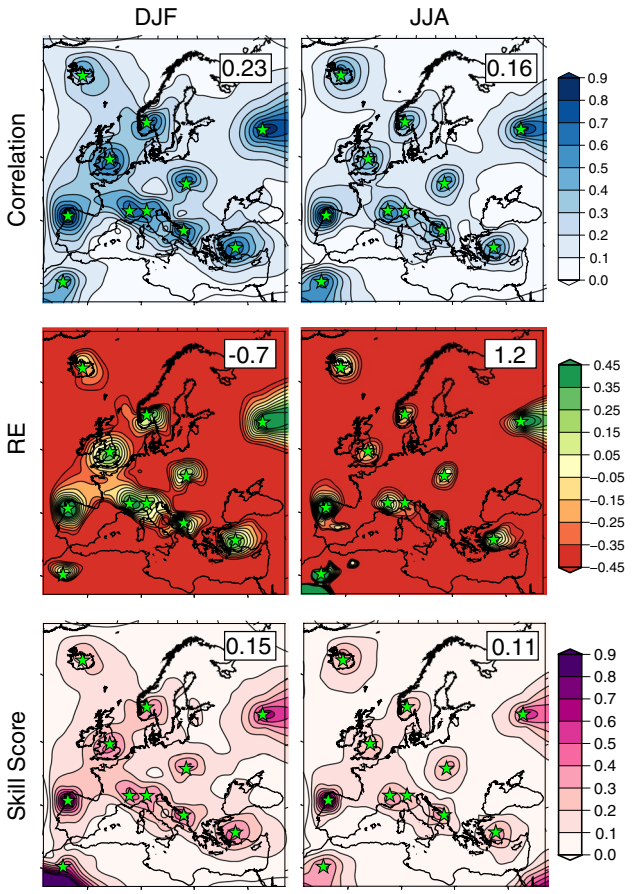

Fig. 7 Correlation (*top*), Reduction of error (*middle*) and skill score (*bottom*) for the BHM reconstructions of winter (*left*) and summer (*right*) precipitation. The figure shows the reconstructions skill in the ideal case, where no noise is included to the pseudoproxies. Thus these results are comparable to Figs. 2, 3 and 4. As in former figures, the number in the *square* indicates the spatial mean of the statistic

are remarkably close to the locations of the pseudoproxies, but the divergence increases with distance from these areas. The average correlation in winter is 0.23, very close to the 1-analog case (0.27), but much lower than 10-analogs or CCA PPEs (correlations 0.43 and 0.49, respectively). The RE shows overall negative skill in both seasons with clear skillful reconstructions in every location of proxy information, indicating that the reconstruction is better than a climatological mean only in the neighbourhoods of the proxy locations. In a similar fashion, the skill score fits very well within the former results, and shows average values of 0.15 and 0.11 for winter and summer, respectively. Although these values are lower than the correlation, it is in good agreement with the expected result for this statistic (see Fig. 1), and indeed is very similar to the 1-analog case in each corresponding season (0.16 and 0.07). Still, the method shows some skill reproducing outliers on the network of proxies, which supports the argument that if a dense network of proxies is available, this method represents a promising CFR tool.

An aspect where the BHM method outperforms CCA and the 10-analogs method is in its ability to preserve the original variance. The maps of ratio of variance (not shown) indicate that the variance of the reconstruction is roughly the same as the original, demonstrating that this method does not suffer from loss of variance. This is another aspect where the 1-analog and BHM reconstructions perform very similar. This fact, together with the inferior abilities of both techniques to reproduce the temporal agreement (see Figs. 2, 3, 4), renders the BHM method roughly equivalent to the 1-analog reconstruction in its overall skill.

As the discussion above outlines, the asymmetry between winter and summer again stands out with this method. The consistency of this results strongly suggests that it is not possible to design a statistical method to extrapolate summer precipitation out of a limited number of proxy locations, regardless of their quality. This is due to the high heterogeneity of this variable, especially noticeable in summer.

Further experiments with noisy pseudoproxies have been performed with this method. The results, not shown for the sake of brevity, indicate that the net effect of noise is to reduce the temporal agreement between the simulation and its reconstruction, albeit it does not affect the skill of the BHM reconstructions to reproduce the original variance. In winter, the spatial-averaged correlations drops from 0.23 to 0.16 and 0.15 when white and red noise is introduced in the proxies, respectively. Similarly, the skill score is reduced to roughly one half. RE however does not change globally so dramatically (from −0.66 to −0.75 and −0.78), although it gets severely reduced in the former skillful locations, rendering the reconstruction better than a constant climatology in very narrow areas around proxy locations. These results can be mostly attributed to the short decorrelation length of precipitation in general and to the omission of a more informed estimate of the structure of the spatial covariance matrix. Clearly, things such as the upslope effect and anticorrelation of windward and leeward sides along the Scandinavian coast or the Alps should be included. Imposing a spatial structure should however be based on physical understanding, as purely statistical patterns might not be stable in time (Raible et al. 2006).

Taking the high computational demand into account, use of BARCAST in a scenario as tested here is unfeasible and will not result in skilful reconstructions for most of the area—at least without the above mentioned extension.

4 Discussion

This study analyses the skill of three completely different approaches to perform CFRs for precipitation. As a first

approximation, reconstructions based on "perfect" proxies are evaluated, in the sense that each one reproduces perfectly the local climatic variability without any source of non-climatic signal. Different measures of the skill are employed, with a focus on the skill of the methods to preserve the original variance and temporal evolution, including the reproducibility of the tails of the distribution. All CFR methods are able to reproduce, to some extent, the evolution of precipitation, but important caveats are apparent. The most important being that, although the CFRs are skilfully close to the locations of the pseudoproxies, precipitation in areas not properly covered by the proxy network can not be reconstructed accurately. This lack of skill is especially problematic in summer. The lack of spatial coherence and the asymmetry between winter and summer is related to the nature of the precipitation regime in Europe. In winter, precipitation is driven to a large extent by large-scale circulation, whereas in summer precipitation is rather modulated by local processes, hampering the extrapolation of precipitation for the entire European domain from a limited set of proxy locations.

Comparing the spatial structure of the different skill measures among the methods, they all share similar patterns, although CCA in general produces higher correlations than the AM and BHM. However, when several situations are averaged to form an analog in the so-called 10-analog variant, the correlation achieved by the AM rises, becoming comparable to CCA. Both the BHM and the 1-member analog version present the advantage of almost perfectly preserving the variance of the original field, whereas the linear assumption implicit in the CCA method destroys an important part of the original variance. The 10-member analog version also underestimates variance, showing a clear trade-off between variance and correlation. Overall, the CCA method is comparable to the 10-analog variant, whereas the BHM resembles the skill of the 1-analog version. Intermediate skill in correlation/preservation of variance can be achieved by considering different numbers of analogs. The RE is a statistic that combines both aspects of the skill, and demonstrates that neither the 10-member analog or the CCA method is clearly better than the other methods, whereas the 1-member AM or BHM seems not to be always the best choice due to the low correlation these methods produce. An important result of the evaluation of the skill of CFR in reproducing extreme events through contingency tables rather than just correlation analysis is that they are at least as reliable in reproducing outliers as they are in "ordinary" situations. This is non-trivial result with implications using the gridded products to analyse the evolution of extreme events, such like impact studies. Overall, none of the three methods (nor variants of the AM) outperforms the others in every aspect, and the compromise between correlation and variance is found across all of our analyses.

The assumption of perfect PPE is unrealistic due to the fact that real proxies include an important amount of non-climatic signal which propagates into the CFR, widening the uncertainty ranges and biases. Thus, we analyse the impact of including two types of noise (white and red) into the pseudoproxies. The level of noise is such that it reduces the correlation with the original series in each proxy location to 0.5. As expected, including noise severely reduces the temporal agreement between reconstruction and target, regardless of the CFR method applied or the season. This has a strong impact on correlation and the skill score, but not so much in the amount of variance explained.

We also compare the reconstruction skill at different frequency bands. At interannual scale, the nature of the noise does not seem to play an important role. However, when temporal low-pass filters are applied to the series before the assessment of their skill, the pseudoproxies contaminated with red noise produced CFR with generally larger correlation. This indicates that if the non-climatic signal embedded in real proxies is red, CFR will tend to show higher skill at low frequencies. This result can be expected as the low-pass filtering for time scales longer than the decorrelation time of the noise results in a residual with a flat spectrum, and thus being more similar to white noise at these time scales and longer.

The role of the pool used in the AM has also been analysed. Here we use a different RCM simulation with the CCLM model as the target pseudoreality, whereas the same 2000-year simulation performed with MM5 is used as pool of analogs. Most of the conclusions derived from comparing the CCA and AM can be confirmed by this analysis. The correlations using the 1-member AM are comparable in winter and summer to the ones obtained when the same simulation was simultaneously used as pool target. This indicates that the nature of the pool does not seem to be critical for the reproduction of the temporal evolution of the reconstruction. However, the comparison of variance demonstrates how the properties of the pool are transferred to the reconstruction. Here, the CCLM simulation shows a strong underestimation of the summer precipitation in Southern Europe, that the pool of the MM5 simulation is not able to reproduce. This highlights the importance of using a realistic pool, and suggests a possible improvement of the method, namely the removal of systematic bias in the pool according to some reliable observational dataset *prior* to the search for analogs.

5 Conclusions and outlook

Our results indicate that the application of CFR techniques to precipitation reconstructions is affected by a number of theoretical limitations hard to overcome with the currently

available network of proxies. The problem arises from the complex behaviour of this variable, which is highly heterogeneous through the annual cycle, but especially in summer. None of the three statistical approaches employed here are able to extrapolate valuable precipitation information to areas away from the proxies sites. Further, it should be noted that this study has to be considered within the cautions of being a theoretical exercise, based mostly on synthetic data generated with climate models and an only partly realistic network of pseudoproxies. Real precipitation can be expected to be even more complex and regional-dependent, which adds a layer of complexity that real reconstructions have to deal with. However, our results also indicate that the CFR of precipitation is a priori possible if a dense network of local proxies is available. Thereby the reliability of real-world reconstruction efforts such as the Old World Drought Atlas (Cook 2013) could be evaluated in further detail, although it would require the use of a more realistic network of proxies than the one we use in this preliminary assessment. Thus, future studies will address the role of such a realistic network of proxies, as well as the impact of the inclusion of series of proxies which do not span the whole period. Finally the interactions between seasons, which can mask the climatic signal extracted from the proxies, is not explicitly taken into account in this study. The design of experiments to gain insight in the role of such processes is feasible within the framework of PPE, but is delayed for future studies. Ultimately, the methods tested in this study, particularly the AM, will be applied to real proxy series with the aim of evaluating the possibilities of performing real precipitation reconstructions for Europe, as well as evaluating their uncertainties.

Acknowledgments This work was funded by the PRIME2 project (priority program INTERDYNAMIK, German Research Foundation). The authors thank the constructive comments by the two anonymous reviewers, as well as the kind support of Dennis Bray reviewing the text of the final version of this manuscript.

References

Annan JD, Hargreaves JC (2013) A new global reconstruction of temperature changes at the last glacial maximum. Clim Past 9(1):367–376. doi:10.5194/cp-9-367-2013

Bunde A, Büntgen U, Ludescher J, Luterbacher J, von Storch H (2013) Is there memory in precipitation? Nat Clim Change 3(3):174–175. doi:10.1038/nclimate1830

Casty C, Raible CC, Stocker TF, Wanner H, Luterbacher J (2007) A European pattern climatology 1766–2000. Clim Dyn 29(7–8):791–805. doi:10.1007/s00382-007-0257-6

Cook E, D'Arrigo R, Mann ME (2002) A well verified multiproxy reconstruction of the winter norh atlantic oscillation index since AD 1400. J Clim 15:1754–1764

Cook ER (2013) The old world drought atlas: Tree-ring reconstructions of past drought over Europe and the Mediterranean basin since 1200 c.e. (invited). In: American Geophysical Union, Fall Meeting 2013, abstract GC12A-05

Cook ER, Briffa KR, Jones PD (1994) Spatial regression methods in dendroclimatology: A review and comparison of two techniques. Int J Climatol 14(4):379–402. doi:10.1002/joc.3370140404

Cook ER, Meko DM, Stahle DW, Cleaveland MK (1999) Drought reconstructions for the continental United States. J Clim 12(4):1145–1162. doi:10.1175/1520-0442<(1999)0121145:DRFTCU>2.0.CO;2

Cook ER, Woodhouse CA, Eakin CM, Meko DM, Stahle DW (2004) Long-term aridity changes in the western United States. Science 306(5698):1015–1018. doi:10.1126/science.1102586

Cook ER, Anchukaitis KJ, Buckley BM, DÁrrigo RD, Jacoby GC, Wright WE (2010) Asian monsoon failure and megadrought during the last millennium. Science 328(5977):486–489. doi:10.1126/science.1185188

Dorado-Liñán I, Gutiérrez E, Andreu-Haylesand L, Heinrich I, Helle G (2012) Potential to explain climate from tree rings in the south of the Iberian Peninsula. Clim Res 55(2):119–134. doi:10.3354/cr01126

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

Flückiger J, Monnin E, Stauffer B, Schwander J, Stocker TF, Chappellaz J, Raynaud D, Barnola JM (2002) High-resolution holocene N2O ice core record and its relationship with CH4 and CO2. Global Biogeochem Cycles 16: doi:10.1029/2001GB001417

Frank D, Büntgen U, Böhm R, Maugeri M, Esper J (2007) Warmer early instrumental measurements versus colder reconstructed temperatures: shooting at the moving target. Quat Sci Rev 26:3298–3310. doi:10.1016/j.quascirev.2007.08.002

Franke J, González-Rouco JF, Frank D, Graham NE (2010) 200 years of European temperature variability: insights from and tests of the proxy surrogate reconstruction analog method. Clim Dyn 37(1–2):133–150. doi:10.1007/s00382-010-0802-6

Gandin LS, Murphy AH (1992) Equitable skill scores for categorical forecasts. Monthly Weather Rev 120(2):361–370. doi:10.1175/1520-0493(1992)120<0361:ESSFCF>2.0.CO;2

Gelman A, Carlin J, Stern H, Rubin D (2003) Bayesian Data Anal, 3rd edn. Chapman and Hall, London

Gómez-Navarro JJ, Montávez JP, Jiménez-Guerrero P, Jerez S, Lorente-Plazas R, González-Rouco JF, Zorita E (2012) Internal and external variability in regional simulations of the iberian peninsula climate over the last millennium. Clim Past 8(1):25–36. doi:10.5194/cp-8-25-2012

Gómez-Navarro JJ, Montávez JP, Wagner S, Zorita E (2013) A regional climate palaeosimulation for europe in the period 1500–1990—part 1: model validation. Clim Past 9(4):1667–1682. doi:10.5194/cp-9-1667-2013

Gómez-Navarro JJ, Werner J, Wagner S, Zorita E, Luterbacher J (2014) Precipitation in the past millennium in europe-extension to roman times. In: Paul A, Schulz M (eds) Integrated analysis of interglacial climate dynamics (INTERDYNAMIC), Springer Briefs in Earth System

Jungclaus JH, Lorenz SJ, Timmreck C, Reick CH, Brovkin V, Six K, Segschneider J, Crowley TJ, Pongratz J, Krivova NA, Vieira LE, Solanki SK, Klocke D, Botzet M, Esch M, Gayler V, Haak H, Raddatz TJ, Roeckner E, Schnur R, Widmann H, Claussen M, Stevens MB, Marotzke J (2010) Climate and carbon-cycle variability over the last millennium. Clim Past 6:723–737. doi:10.5194/cp-6-723-2010

Küttel M, Luterbacher J, Zorita E, Xoplaki E, Riedwyl N, Wanner H (2007) Testing a European winter surface temperature reconstruction in a surrogate climate. Geophys Res Lett 34(7):L07,710. doi:10.1029/2006GL027907

Küttel M, Xoplaki E, Gallego D, Luterbacher J, García-Herrera R, Allan R, Barriendos M, Jones PD, Wheeler D, Wanner H

(2010) The importance of ship log data: reconstructing North Atlantic, European and Mediterranean sea level pressure fields back to 1750. Clim Dyn 34(7–8):1115–1128. doi:10.1007/s00382-009-0577-9

Li B, Smerdon JE (2012) Defining spatial comparison metrics for evaluation of paleoclimatic field reconstructions of the Common Era. Environmetrics 23(5):394–406. doi:10.1002/env.2142

Lorenz EN (1969) Atmospheric predictability as revealed by naturally occurring analogues. J Atmos Sci 26(4):636–646. doi:10.1175/1520-0469(1969)266<36:APARBN>2.0.CO;2

Luterbacher J, Schmutz C, Gyalistras D, Xoplaki E, Wanner H (1999) Reconstruction of monthly NAO and EU indices back to AD 1675. Geophys Res Lett 26(17):2745–2748. doi:10.1029/1999GL900576

Luterbacher J, Rickli R, Tinguely C, Xoplaki E, Schüpbach E, Dietrich D, Hüsler J, Ambühl M, Pfister C, Beeli P, Dietrich U, Dannecker A, Davies T, Jones P, Slonosky V, Ogilvie A, Maheras P, Kolyva-Machera F, Martin-Vide J, Barriendos M, Alcoforado M, Nunes M, Jónsson T, Glaser R, Jacobeit J, Beck C, Philipp A, Beyer U, Kaas E, Schmith T, Bárring L, Jönsson P, Rácz L, Wanner H (2000) Monthly mean pressure reconstruction for the late maunder minimum period (AD 1675–1715). Int J Climatol 20(10):1049–1066. doi:10.1002/1097-0088(200008)20:101049:AID-JOC5213.0.CO;2-6

Luterbacher J, Xoplaki E, Dietrich D, Jones PD, Davies TD, Portis D, Gonzalez-Rouco JF, von Storch H, Gyalistras D, Casty C, Wanner H (2001) Extending North Atlantic oscillation reconstructions back to 1500. Atmos Sci Lett 2(1–4):114–124. doi:10.1006/asle.2002.0047

Luterbacher J, Dietrich D, Xoplaki E, Grosjean M, Wanner H (2004) European seasonal and annual temperature variability, trends, and extremes since 1500. Science 303(5663):1499–1499

Luterbacher J, Liniger MA, Menzel A, Estrella N, Della-Marta PM, Pfister C, Rutishauser T, Xoplaki E (2007) Exceptional European warmth of autumn 2006 and winter 2007: historical context, the underlying dynamics, and its phenological impacts. Geophys Res Lett 34(12):L12,704. doi:10.1029/2007GL029951

Luterbacher J, Koenig SJ, Franke J, van der Schrier G, Zorita E, Moberg A, Jacobeit J, Della-Marta PM, Küttel M, Xoplaki E, Wheeler D, Rutishauser T, Stössel M, Wanner H, Brázdil R, Dobrovolný P, Camuffo D, Bertolin C, Gonzalez-Rouco FJ, Wilson R, Pfister C, Limanówka D, Nordli O, Leijonhufvud L, Söderberg J, Allan R, Barriendos M, Glaser R, Riemann D, Hao Z, Zerefos CS (2010) Circulation dynamics and its influence on European and Mediterranean January–April climate over the past half millennium: results and insights from instrumental data, documentary evidence and coupled climate models. Clim Change 101(1–2):201–234. doi:10.1007/s10584-009-9782-0

Mann ME, Bradley RS, Hughes MK (1998) Global-scale temperature patterns and climate forcing over the past six centuries. Nature 392(6678):779–787. doi:10.1038/33859

Mann ME, Zhang Z, Hughes MK, Bradley RS, Miller SK, Rutherford S, Ni F (2008) Proxy-based reconstructions of hemispheric and global surface temperature variations over the past two millennia. Proceedings of the National Academy of Sciences of the United States of America 105(36):13,252–13,257. doi:10.1073/pnas.0805721105

Masson-Delmotte V, Schulz M, Abe-Ouchi A, Beer J, Ganopolski A, González Rouco JF, Jansen E, Lambeck K, J L, Naish T, Osborn T, B OB, Quinn T, Ramesh R, Rojas M, Shao X, Timmermann A, (2013) Information from paleoclimate archives. In: Stocker TF, Qin D, Plattner GK, Tignor M, Allen SK, Boschung J, Nauels A, Xia Y, Bex V, Midgley PM (eds) Climate change 2013: The physical science basis. Contribution of working group I to the fifth assessment report of the intergovernmental panel on climate change, Cambridge University Press, Cambridge

Matulla C, Zhang X, Wang XL, Wang J, Zorita E, Wagner S, von Storch H (2007) Influence of similarity measures on the performance of the analog method for downscaling daily precipitation. Clim Dyn 30(2–3):133–144. doi:10.1007/s00382-007-0277-2

McDonald JE, Green CR (1960) A comparison of rank-difference and product-moment correlation of precipitation data. J Geophys Res 65(1):333–336. doi:10.1029/JZ065i001p00333

Neukom R, Luterbacher J, Villalba R, Küttel M, Frank D, Jones PD, Grosjean M, Esper J, Lopez L, Wanner H (2010) Multicentennial summer and winter precipitation variability in southern South America. Geophys Res Lett 37(14):L14,708, doi:10.1029/2010GL043680, http://onlinelibrary.wiley.com/doi/10.1029/2010GL043680/abstract

Neukom R, Gergis J, Karoly DJ, Wanner H, Curran M, Elbert J, González-Rouco F, Linsley BK, Moy AD, Mundo I, Raible CC, Steig EJ, van Ommen T, Vance T, Villalba R, Zinke J, Frank D (2014) Inter-hemispheric temperature variability over the past millennium. Nat Clim Change 4(5):362–367. doi:10.1038/nclimate2174

PAGES 2k consortium (2013) Continental-scale temperature variability during the past two millennia. Nat Geosci 6(5):339–346. doi:10.1038/ngeo1797

Pauling A, Luterbacher J, Casty C, Wanner H (2006) Five hundred years of gridded high-resolution precipitation reconstructions over Europe and the connection to large-scale circulation. Clim Dyn 26(4):387–405. doi:10.1007/s00382-005-0090-8

Phipps SJ, McGregor HV, Gergis J, Gallant AJE, Neukom R, Stevenson S, Ackerley D, Brown JR, Fischer MJ, van Ommen TD (2013) Paleoclimate data-model comparison and the role of climate forcings over the past 1500 years. J Clim 26:6915–6936. doi:10.1175/JCLI-D-12-00108.1

Pinto JG, Raible CC (2012) Past and recent changes in the North Atlantic oscillation. Wiley Interdiscip Rev Clim Change 3(1):79–90. doi:10.1002/wcc.150

Pongratz J, Reick C, Raddatz T, Claussen M (2008) A reconstruction of global agricultural areas and land cover for the last millennium. Global Biogeochem Cycles 22(3):GB3018. doi:10.1029/2007GB003153

Raible CC, Casty C, Luterbacher J, Pauling A, Esper J, Frank DC, Buñtgen U, Roesch AC, Tschuck P, Wild M, Vidale PL, Schär C, Wanner H (2006) Climate variability-observations, reconstructions, and model simulations for the Atlantic-European and Alpine region from 1500–2100 AD, predictability and climate risks. In: Wanner H, Grosjean M, Röthlisberger R, Xoplaki E (eds) Climate variability. Springer, Netherlands, pp 9–29

Riedwyl N, Luterbacher J, Wanner H (2008) An ensemble of European summer and winter temperature reconstructions back to 1500. Geophys Res Lett 35(20): doi:10.1029/2008GL035395

Riedwyl N, Kütel M, Luterbacher J, Wanner H (2009) Comparison of climate field reconstruction techniques: application to Europe. Clim Dyn 32(2–3):381–395. doi:10.1007/s00382-008-0395-5

Schenk F, Zorita E (2012) Reconstruction of high resolution atmospheric fields for Northern Europe using analog-upscaling. Clim Past 8(5):1681–1703. doi:10.5194/cp-8-1681-2012

Smerdon JE (2012) Climate models as a test bed for climate reconstruction methods: pseudoproxy experiments. Wiley Interdiscip Rev Clim Change 3(1):63–77. doi:10.1002/wcc.149

Smerdon JE, Kaplan A, Chang D, Evans MN (2010) A pseudoproxy evaluation of the CCA and RegEM methods for reconstructing climate fields of the last millennium. J Clim 24(4):4856–4880. doi:10.1175/2010JCLI4110.1

von Storch H, Zwiers FW (2002) Statistical analysis in climate research. Cambridge University Press, Cambridge

von Storch H, Zorita E, Gonzalez-Rouco JF (2008) Assessment of three temperature reconstruction methods in the virtual reality of a climate simulation. Int J Earth Sci 98(1):67–82. doi:10.1007/s00531-008-0349-5

Taylor KE (2001) Summarizing multiple aspects of model performance in a single diagram. J Geophys Res Atmos 106(D7):7183–7192. doi:10.1029/2000JD900719

Tingley MP, Huybers P (2010a) A bayesian algorithm for reconstructing climate anomalies in space and time. Part I: development and applications to paleoclimate reconstruction problems. J Clim 23(10):2759–2781. doi:10.1175/2009JCLI3015.1

Tingley MP, Huybers P (2010b) A bayesian algorithm for reconstructing climate anomalies in space and time. part II: comparison with the regularized Expectation-Maximization algorithm. J Clim 23(10):2782–2800. doi:10.1175/2009JCLI3016.1

Tingley MP, Huybers P (2013) Recent temperature extremes at high northern latitudes unprecedented in the past 600 years. Nature 496(7444):201–205. doi:10.1038/nature11969

Tingley MP, Craigmile PF, Haran M, Li B, Mannshardt E, Rajaratnam B (2012) Piecing together the past: statistical insights into paleoclimatic reconstructions. Quat Sci Rev 35:1–22. doi:10.1016/j.quascirev.2012.01.012

Trouet V, Esper J, Graham N, Baker A, Scourse JD, Frank D (2009) Persistent positive North Atlantic oscillation mode dominated the medieval climate anomaly. Science 324(5923):78–80. doi:10.1126/science.1166349

Van Den Dool HM (1994) Searching for analogues, how long must we wait? Tellus A 46:314–324. doi:10.1034/j.1600-0870.1994.t01-2-00006.x

Werner JP, Luterbacher J, Smerdon JE (2013) A pseudoproxy evaluation of bayesian hierarchical modeling and canonical correlation analysis for climate field reconstructions over Europe. J Clim 26(3):851–867. doi:10.1175/JCLI-D-12-00016.1

Xoplaki E, González-Rouco JF, Luterbacher J, Wanner H (2004) Wet season Mediterranean precipitation variability: influence of large-scale dynamics and trends. Clim Dyn 23(1):63–78. doi:10.1007/s00382-004-0422-0

Xoplaki E, Luterbacher J, Paeth H, Dietrich D, Steiner N, Grosjean M, Wanner H (2005) European spring and autumn temperature variability and change of extremes over the last half millennium. Geophys Res Lett 32(15):L15,713. doi:10.1029/2005GL023424

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