Regional Environmental Change manuscript No.
(will be inserted by the editor)

Testing instrumental and downscaled reanalysis time series for temperature trends in NE of Spain in the last century.

M. Turco, R. Marcos, P. Quintana-Seguí, M.C. Llasat

Received: date / Accepted: date

Abstract In the context of climatic temperature studies more often than not a time series is affected by artificial inhomogeneities. To overcome such difficulty we propose a new simple integral methodology which promising results point towards not only the detection of unknown inhomogeneous periods but also to the possibility of reconstructing the uncertain portion of the series. It is based on a parsimonious statistical downscaling (Multiple Linear Regression) of the large scale 20CR reanalysis data. This method is successfully applied upon two long-range temperature series from a couple of centennial observatories (Ebre and Fabra, NE of Spain) which do not have nearby suitable temperature series to compare with. Results of trend analysis point to a clear signal of warming, with a larger rate of increase for the maximum temperature (respect to the minimum one), for the more recent decades (respect to the whole available period) and for the original series (respect to the reconstructed ones).

Keywords statistical dowsncaling · inhomogeneities · multiple linear regression · series reconstruction

M. Turco, GAMA (Meteorological Hazards Analysis Team), Dpt. of Astronomy Meteorology, Faculty of Physics, University of Barcelona, Av. Diagonal 647, Barcelona, E-08028, Spain. Euro-Mediterranean Centre on Climate Change, Impact on soil and coast Division, Via Augusto Imperatore 16, I-73100, Lecce. (mturco@am.ub.es) · R. Marcos, GAMA (Meteorological Hazards Analysis Team), Dpt. of Astronomy Meteorology, Faculty of Physics, University of Barcelona, Av. Diagonal 647, Barcelona, E-08028, Spain.(rmarcos@am.ub.es) · P. Quintana-Seguí, Observatori de l'Ebre (Universitat Ramon Llull - CSIC), Horta Alta 38, Roquetes, 43520, Spain. (pquintana@obsebre.es) · M. C. Llasat, GAMA (Meteorological Hazards Analysis Team), Dpt. of Astronomy Meteorology, Faculty of Physics, University of Barcelona, Av. Diagonal 647, Barcelona, E-08028, Spain.(carmell@am.ub.es)

1 Introduction

A homogeneous climate series is defined as a series whose variations are caused only by changes in weather and climate (*Conrad and Pollak*, 1950). Unfortunately, a time series is often affected by one or more artificial inhomogeneities. Inhomogeneity may be due to various factors: changes in the station environment or in the station itself (e.g. changing from a manual to an automated station), instrument malfunction or observation practices (*Aguilar et al.*, 2003). Some changes may cause spurious (non-climatic) jumps and/or gradual shifts in the data. Regardless of the type and the effect of inhomogeneities, the analysis of a nonhomogeneous series can be misleading. Consequently, it is crucial to determine, assign and adjust any discontinuities in the data.

The first source of information about the homogeneity of the data is the station metadata, but it is often incomplete, not readily available or altogether non-existent. To overcome this difficulty and increase the amount of time series available for climate studies, great effort has been made to develop a number of statistical techniques both to test the homogeneity of the series and to correct the potential inhomogeneities detected (*Peterson et al.*, 1998; *Ducré-Robitaille et al.*, 2003; *Reeves et al.*, 2007). Two main groups of methodologies for homogeneity testing exist, (i) absolute tests (e.g. *Wijngaard et al.*, 2003), when the station is isolated, and (ii) relative tests (e.g. *Vincent*, 1998; *Menne and Williams*, 2009) that use data from neighbouring stations as a regional climate signal. In this case, a significant departure of the tested series from the regional climate signal is assumed to be caused by inhomogeneities. Note that the relative tests are more sensitive and generally preferred over absolute tests but, in contrast, need representative nearby series to be reliably applied.

For example, within the framework of the European Climate Assessment project (ECA, `http://eca.knmi.nl/`), *Wijngaard et al.* (2003) developed a basic homogeneity testing protocol which was based on absolute tests due to the low density of the temperature station network at the time that study was conducted (60 available stations over Europe for the period 1901-1999). The results of these tests are summarized in three classes, e.g. a series could be "useful", "doubtful" or "suspect". The problem of the homogeneity of stations in Europe is important, as evidenced by the fact that with this approach, 94% of the temperature series are tagged "doubtful" or "suspect".

As an alternative approach when there are no neighbouring stations, one can use the reanalysis data as a background field against which other time series can be compared (*Kalnay et al.*, 2006; *Pielke et al.*, 2007a; *Haimberger*, 2007; *Dee et al.*, 2011). However, three main limitations could affect the direct use of reanalysis as a tool to detect inhomogeneities in a local temperature time series. Firstly, the majority of available reanalysis (e.g. ECMWF 40 Year Reanalysis: 1957-2002, *Uppala et al.*, 2005; Japanese 25-year Reanalysis: 1979-2004, *Onogi et al.*, 2007; NCEP/NCAR Reanalysis I: 1948-present, *Kalnay et al.*, 1996) cover only the second half of the twentieth century. Secondly, these reanalysis data may suffer from related changes in the observational network (e.g. at the beginning of the satellite era, *Sturaro*, 2003; *Bengtsson et al.*, 2004). And thirdly, their coarse resolution (around the degree order in latitude and longitude) is unsuitable to perform comparisons with local station time series.

The first and second difficulties can be overcomed by using the recently developed NOAA-CIRES 20th Century Reanalysis V2 (20CR) which spans from 1871 to the present. 20CR used a state-of-the-art data assimilation system and surface pressure observations to generate an atmospheric global dataset. Note that this reanalysis shows high skill for the extratropical tropospheric circulation in the northern hemisphere (*Compo et al.*, 2006). Since only sea level pressure is assimilated, this reanalysis may suffer less than previous reanalysis from some of the most evident discontinuity due to the time-changing observational network, although some concerns have been risen recently about its homogeneity (*Ferguson and Villarini*, 2012) in the early part of last century. Finally, the statistical downscaling methodology usually implemented in regional climate studies (*Wilby et al.*, 2004; *Benestad et al.*, 2008), is able to bridge the scale-gap.

In this study we present a new approach applying a multiple pointwise linear regression (hereafter, MLR), fitted to historical data to capture the empirical relationship between large-scale 20CR variables and observed temperature. While the application of the MLR method is fairly standard in the context of the generation of climate change scenarios (see, e.g. *Maraun et al.*, 2010), as far as the authors know, its application as a downscaling method to test inhomogeneities and reconstruct series is new. Indeed, there are similar approaches, such as one based on Proxy Surrogate Reconstruction (PSG) that uses climate model simulations combined with several proxy and instrumental data (*Graham et al.*, 2007; *Franke et al.*, 2010) or the MLR technique developed by *Vincent* (1998), but they use not MLR-downscaling neither reanalysis fields as a tool to identify inhomogeneities in local time series. Furthermore, several authors (e.g. *Kalnay and Cai*, 2003; *Pielke et al.*, 2007b,a; *Fall et al.*, 2010) compare reanalysis fields to observation using the "observation minus reanalysis" (OMR) approach, but with the main objective of trend analysis and attribution.

Our technique is carefully tested with temperature time series from two stations of the North-East of the Iberian Peninsula: Ebre and Fabra centennial observatories which are regarded as high quality stations in the context of Spanish climate studies. Finally, we analyzed the trend of the original and the reconstructed time series.

This paper is organized as follows: section 2 is a comprehensive description of the data used in the study; section 3 presents the implementation framework and the MLR; in section 4, the MLR is first tested in the period without discontinuities and then, applied to reproduce the doubtful periods; finally, the original and reconstructed time series are analysed. Section 5 summarises the main results.

2 Data description and analysis

2.1 Stations

In this study we use annually averaged daily temperature time series from two high quality stations in Catalonia (Figure 1a), Fabra Observatory (close to Barcelona, lat. 41.42°N, long. 2.12°E and 411 m amsl) and Ebre Observatory (Roquetes, close to Tortosa, lat. 40.82°N, long. 0.49°E and 51 m amsl). Both series span over almost a century (Fabra, 1913–2008; Ebre, 1905-2008), although during the Spanish Civil War the measurements were interrupted in the Ebre Observatory for two years, 1938-1939. There are three main reasons for this choice of stations:

1. The two observatories are reference stations in Spanish climate studies (present in the ECA database, `http://eca.knmi.nl/`) whose temperature series are used, for example, in the report of climatic change in Catalonia (*Llebot*, 2010).
2. In these periods there is no lack of data and metadata is readily available (*Prohom and Herrero*, 2008; *Prohom Duran et al.*, 2009; *Seguí-Grau*, 2003). Note that both observatories are centennial sites devoted to geophysical research and that atmospheric weather variables have been regularly measured there since their foundation (*Curto et al.*, 2009).
3. Both stations are also interesting because of the ongoing controversy over the attribution of detected inhomogeneities to the influence of non-climatic factors or natural phenomena such as volcanic eruption (see *Toreti et al.*, 2010, and reference therein).

The available metadata indicate that at the Fabra Observatory there have been two changes of thermometer, in 1921 and in 1961, along with the use of a non-homologated meteorological shelter during the period 1978-1983 (*Prohom Duran et al.*, 2009). At the Ebre Observatory there is some information about thermometer replacement and a change of position of the station (of around 4 meters), but the time of these changes is unspecified (*Seguí-Grau*, 2003). Probably this replacement has taken place before the 40's; besides, during the suspect period that will be detailed below, there is no change in the station or in its surroundings (G. Solé, personal communication).

To better illustrate the characteristics of the series, a set of indices is calculated. This set was proposed by *Wijngaard et al.* (2003) and consists of: annual average maximum temperature, T_{max}; annual average minimum temperature, T_{min}; annual average diurnal temperature range, DTR=T_{max}-T_{min}; annual average of day-to-day difference (indicated with the suffix i) of DTR, i.e. $vDTR$=DTR_i-DTR_{i+1}.

The use of these variables enables us to retain information on phenomena and sources of inhomogeneity that act upon different aspects of temperature. Therefore, by using this set of tests and variables, four ways of possible inhomogeneity can be explored: diurnal, nocturnal, day-night relationship, day-to-day/day-night relationship. For example, the DTR variable is often more sensitive to the tests of homogeneity, because changes due to station relocation or measurement techniques are generally related to radiation, with different influences on the maximum or minimum temperature. This is apparent in Figure 2, which shows the evolution of T_{max}, T_{min} and DTR for the Fabra Observatory ($vDTR$ does not show any trend or strange behaviour so, for the sake of brevity, is not displayed). In the case of DTR, there

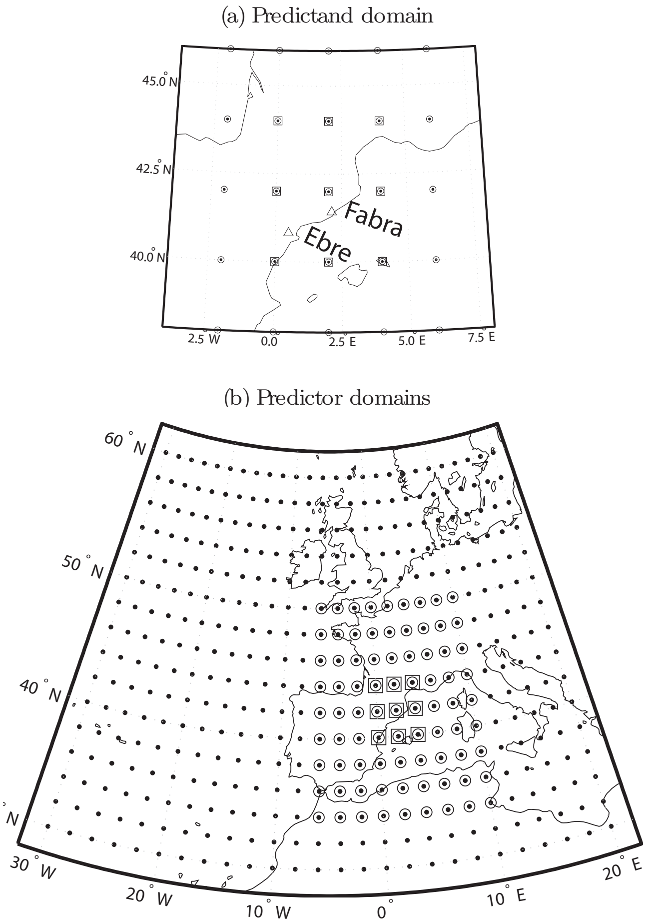

Fig. 1 (a) Region of study (North-East of Spain) and position of the two observatories (Ebre, close to Tortosa and Fabra, close to Barcelona). (b) Predictor domains. Large domain: simple dots. Medium domain: circled dots. Small domain: squared circled dots. Figure 1a zooms on the small domain.

is an odd peak around the early eighties, which can be linked to the period 1978-1983 when the thermometer was in a non-homologated meteorological shelter.

Figure 2 also shows the evolution of T_{max}, T_{min} and DTR for the Ebre Observatory (again in this case the $vDTR$ does not show any particular behaviour and is not displayed). In the case of Ebre, DTR shows a kind of strange behavior around the late eighties-early nineties, in the middle of a global positive trend.

Yet there are no documented variations in neither the station instrumentation nor its environment which could explain that particular behavior. So far then, we can establish a couple of apparent, hereafter called "suspect", periods of inhomogeneity: 1978-1983 for Fabra Observatory and 1986-1992 for Ebre Observatory.

Although there is always some degree of subjectivity in that choice, these "suspect" periods are also hinted

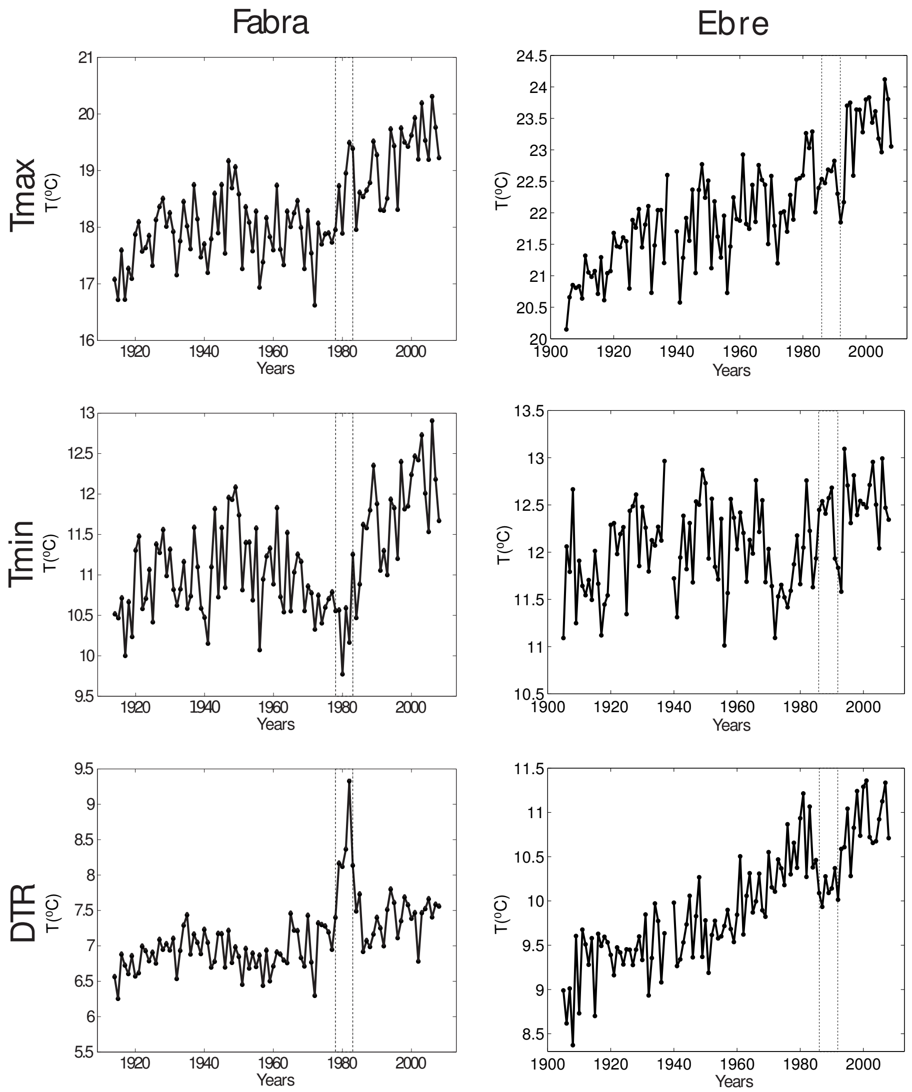

Fig. 2 Evolution of T_{max} (top), T_{min} (center) and DTR (bottom) for Fabra (left) and Ebre (right) Observatories. The anomalous periods are highlighted with vertical dotted lines.

by the set of four homogeneity tests applied to T_{max}, T_{min}, DTR and $vDTR$ following the protocol of *Wijngaard et al.* (2003): the Normal Standard Homogeneity Test (SNHT, *Alexandersson*, 1986); the Buishand Range test (BHR, *Buishand*, 1982); the Pettitt test (PET, *Pettitt*, 1979); and the Von Neumann ratio test (VON, *Von Neumann*, 1941).

However, there are some limitations to this approach: firstly, absolute homogeneity tests can detect inhomogeneities but cannot confirm their type, that is, to confirm which are natural or non-natural; secondly, the positioning of the year of rupture (and sometimes the results of the tests themselves) depends on the length of the series (*Peterson et al.*, 1998), so that the timing of resulting break-points is indicative.

As we can see in Table 1, each of the applied tests shows that all the time series are inhomogeneous (except Ebre $Tmin$ in the SNHT case). Note that several break-point years lie around the “suspect” periods. At Ebre, there are also some break-point years within the 60s. However, in this case, we do not have any metadata that would suggest any artificial changes. Finally, the two changes of thermometer in 1921 and 1961 at Fabra are not highlighted by the homogeneity tests (except for vDTR), because these may have caused smaller shifts than those caused by the use of a non-homologated meteorological shelter. One must bear in mind that the homogeneity tests (SNHT and PET) only give the year of the biggest shift in the series. These break-point years are further considered in Sec. 4.4.

2.2 Reanalysis

In this work we have used 20CR reanalysis as a source of atmospheric information (*Compo et al.*, 2011). To bridge the gap between the global scale and local observations we have applied a downscaling methodology which is described in section 3.

The candidate predictors in this study are the annual mean values of sea level pressure, SLP, air temperatures at 850 hPa, T_{850}, specific humidity at 700 hPa, R_{700}, and the temperature at 2 meters , $T2m$. This choice is consistent with several studies that have analyzed the predictor sensitivity in a temperature statistical downscaling method (e.g. *Huth*, 1999; *Timbal et al.*, 2003; *Fowler et al.*, 2007; *Timbal et al.*, 2009). Several sensitivity tests will be performed to find the best predictor(s) combination (see section 3.2). Besides this, the large-scale predictors are defined on a $2° \times 2°$ grid extending through 3 candidate domains, which are depicted in Figure 1b,

- Small domain: from 40° to 44°N and from 0° to 4°E. It covers the North-East of Spain and the nearby Mediterranean.
- Medium domain: from 34° to 50°N and from 6°W to 10°E. It covers Western Europe and the western Mediterranean.
- Large domain: from 30° to 60°N and from 30°W to 20°E. It covers most of Europe and the adjacent part of the Atlantic Ocean.

The following model calibration shows the best predictor combination and the best domain.

3 Methodology

3.1 Implementation framework

In this section the general strategy to implement the MLR as a homogeneity test is described. It consists in the following steps:

1. In-sample model selection. In this step we select the model using all of the reliable part of the series (i.e. the entire series excluding the "suspect period"). The implicit question is: what are the best statistical model parameters?
2. Out-of-sample test (Figure 3). In this step we want to assess whether the statistical model is capable of reproducing the "reliable" parts of the series by means of a cross-validation (see Section. 3.2 for more details). Here the question is: can the statistical model reproduce the series?
3. Testing the "suspect period" (Figure 4a). In this step we reproduce the temperature in the "suspect period" using all the remaining years to train the statistical model. Here the question is: is the suspect period anomalous?
4. Testing an unknown "suspect period" (Figure 4b). In this step we systematically apply the statistical method reproducing a time-moving-window as if the "anomalous period" was unknown. Here the question is: which is the anomalous period?

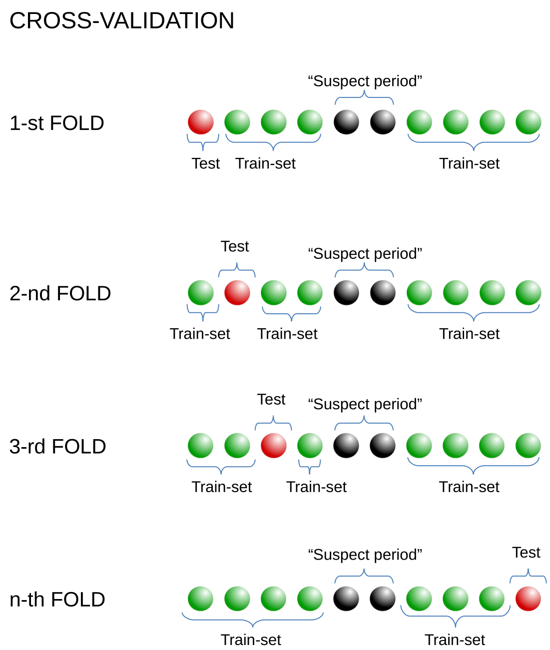

Fig. 3 A schematic view of out of sample test that follows the leave-one-out cross-validation approach. Iteratively, all the single samples from the original sample set are used as the test data, and the remaining samples as the training data. Note that the "anomalous period" is excluded from the validation test.

3.2 Statistical downscaling model

The statistical downscaling model we have implemented is the multiple linear regression (MLR). We have selected this model due to its simplicity and good performance (*Benestad et al.*, 2008), even though we are

Table 1 Results of the SNHT, BHR, PET and VON homogeneity tests applied to the Fabra and Ebre annually averaged daily temperature series for the respective periods 1914-2008 and 1905-2008. The homogeneity test (pvalue $\leq$ 0.05) is shown for each test along with the year in which a break is detected by the SNHT and PET tests.

Fabra	**Break Year (SNHT)**	**Break Year (PET)**	**Homog.? (BHR,PET,VON)**	**Homog.? (SNHT)**
T_{max}	1981	1979	No	No
T_{min}	1986	1986	No	No
DTR	1973	1973	No	No
$vDTR$	1962	1962	No	No
Ebre	**Break Year (SNHT)**	**Break Year (PET)**	**Homog.? (BHR,PET,VON)**	**Homog.? (SNHT)**
T_{max}	1978	1958	No	No
T_{min}	1986	1982	No	Yes
DTR	1970	1958	No	No
$vDTR$	1960	1956	No	No

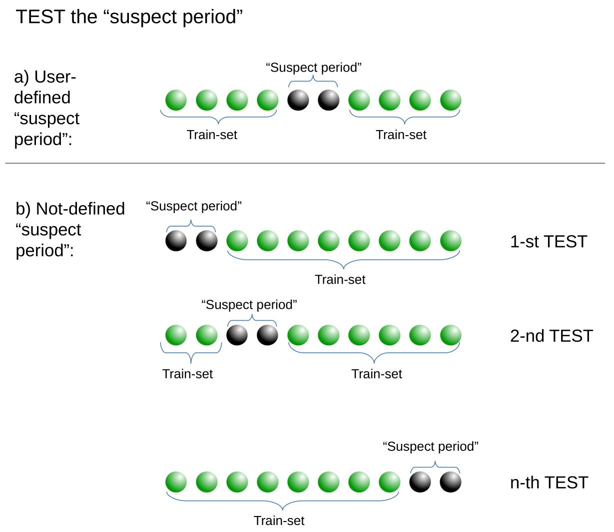

Fig. 4 A schematic view of two approaches to test the "suspect" periods. The first consists in a user-defined "suspect period", while the second is like a cross validation in which, iteratively, all the potential "suspect periods" from the original sample set are used as the test data, and the remaining samples as the training data.

aware the relationship predictor-predictand may not be an exact linear approximation. An extensive analysis of the MLR methods are provided by *Huth* (2002). The simplest MLR has the form:

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{i,j} + \epsilon_i \qquad (1)$$

where, for the *ith* year, Y_i is the response variable (T_{min}, T_{max} and DTR, at Fabra or Ebre station), $X_{i,1}$, ...,$X_{i,p}$ are p predictors (see Section 2.2), and ϵ_i is the residual term. The quantities β_0,...,β_p are unknown coefficients, whose values are determined by ordinary least-squares regression. Different variables (see section 2.2) have been tested as predictors and, to avoid the possible collinearity among them, a stepwise regression was performed (*Wilks*, 2006). This method consists in systematically testing the importance of all potential predictors, adding and removing terms based on their statistical significance (we fixed the level at 5%). Each of the final regression models is the simplest empirical model with the greatest explanatory power. Two stepwise regressions have been evaluated: (1) stepwise screening of the twenty Principal Components (PCs, more than 99% of explained variance) of the predictor variable(s), (2) stepwise screening of gridpoint values. For the stepwise regression using principal components, we have carried out several tests with different numbers of PCs, resulting in minor difference. So we set in twenty the number of PCs to be taken into account. Note that this choice is reasonable also consider the study of *Huth* (2002).

The performance of the model was assessed considering the explained variance (R^2) and the error variance defined as the mean square error (s^2). Finally, the residuals have been tested, with the result that the assumption of residual Gaussianity and zero temporal autocorrelation cannot be refused ($pvalue \leq 0.05$).

It is important to verify the ability of the linear model to perform out-of-sample prediction, i.e. to reproduce the local variable from the knowledge of climatic data outside the period used to test the model. The out-of-sample prediction involves determining the model parameters on one subset of the data (training set), and validating the prediction on the other (testing set). Here a cross-validation is applied (*Von Storch and Zwiers*, 1999), in which a moving window of 1 year is used as the validation data, and the remaining observations as the training data (Figure 3). For example, considering the Fabra series, the first test year is 1914, and the empirical model is calibrated over the period 1915-2008 excluding the suspect 6 years (1978-1983); the second test year is 1915 and is trained with the complementary years (again, excluding the suspect 6 years), and so on. Consequently, a total of 89 (equal to the total length of the series, i.e. 95, minus the suspect 6 years) test periods were considered.

To estimate the uncertainty of the out-of-sample test, we have followed the methodology proposed by *Calmanti et al.* (2007). The practical implementation of this method is summarized here.

1. The residual variance of the calibration period is calculated;
2. Then, 1000 white stochastic series are generated, with variance equal to that calculated previously;
3. Finally the stochastic series are added to the predicted model values, generating an ensemble of 1000 downscaled series.

4 Results

The results are presented following the implementation framework described in Section 3. Furthermore the trend of the original and reconstructed series is studied.

4.1 In-sample model selection

In this section we select the model, i.e. we evaluate several combination of methods (e.g. stepwise regression using principal components, PCs, or using grid-point values), predictor variables (e.g. circulation and/or temperature variables) and domains. The objective in this section is to assess the best setting of the model to reproduce the series in the period without doubtful discontinuities. Table 2 resumes the sensitivity tests carried out. Although all combination were tested (3 predictor sets x 2 methods x 3 domains), for sake of brevity only the most relevant experiments are shown here. Note that for this analysis we have considered the Ebre station for the period 1940-2008, since the years 1938-1939 are missing.

Table 2 Sensitivity tests for the in-sample calibration

Exp.	Predictors	Domain	Regression
$Exp1$	SLP,T_{850},R_{700}	Large	'PCs'
$Exp2$	SLP,T_{850}	Large	'PCs'
$Exp3$	$T2m$	Large	'PCs'
$Exp4$	SLP,T_{850},R_{700}	Large	'Gridpoints'
$Exp5$	SLP,T_{850}	Large	'Gridpoints'
$Exp6$	$T2m$	Large	'Gridpoints'
$Exp7$	SLP,T_{850},R_{700}	Medium	'Gridpoints'
$Exp8$	SLP,T_{850}	Medium	'Gridpoints'
$Exp9$	$T2m$	Medium	'Gridpoints'
$Exp10$	SLP,T_{850},R_{700}	Small	'Gridpoints'
$Exp11$	SLP,T_{850}	Small	'Gridpoints'
$Exp12$	$T2m$	Small	'Gridpoints'

The results of the sensitivity tests are reported in Tables 3 and 4. First of all, the in-sample performances are quite good, with explained variance over 75% in 21 cases out of 36. Generally, the DTR has the lowest R^2. It is apparent that the stepwise regression using grid-point values (experiments 4-6) performs better than the stepwise regression using principal components (experiments from 1 to 3). Among the different set of predictors, the inclusion of humidity (experiments 1-4) improves the results, while similar performances have been obtained using the $T2m$ alone. These results, where the pointwise regression performs better and the inclusion of humidity predictors enhances the effectiveness of the method, are coherent with the studies of *Huth* (2002) and *Timbal et al.* (2009).

Table 3 Sensitivity tests for different sets of predictors (considering the large domain) and statistical models as reported in Table 2. Results for the Fabra and Ebre stations. The model performances are estimated by the explained variance (R^2) and the error variance (s^2). The bolted results highlight the performance of the method implemented.

Fabra	R^2, s^2 (T_{max})	R^2, s^2 (T_{min})	R^2, s^2 (DTR)
$Exp1$	75, 0.18	84, 0.07	34, 0.08
$Exp2$	73, 0.19	79, 0.09	26, 0.09
$Exp3$	79, 0.15	90, 0.05	38, 0.08
Exp4	**87, 0.10**	**88, 0.05**	**68, 0.04**
$Exp5$	85, 0.19	88, 0.05	56, 0.06
$Exp6$	86, 0.10	91, 0.04	55, 0.06
Ebre	R^2, s^2 (T_{max})	R^2, s^2 (T_{min})	R^2, s^2 (DTR)
$Exp1$	68, 0.25	69, 0.08	48, 0.19
$Exp2$	75, 0.20	69, 0.08	59, 0.15
$Exp3$	75, 0.19	77, 0.06	65, 0.13
Exp4	**85, 0.12**	**78, 0.06**	**84, 0.07**
$Exp5$	79, 0.16	69, 0.08	60, 0.14
$Exp6$	77, 0.17	92, 0.02	83, 0.07

Table 4 summarises the experiments with the pointwise regression applied with different predictor domains. Note that experiments 4-6 in table 3 use the same method applied in the experiments of table 4 but consider a larger domain, which generally shows the best performance.

The overall conclusion from these sensitivity tests is that the model presents good performance when it is implemented with this settings:

– method: stepwise regression using gridpoint values;
– predictors: SLP, T_{850}, R_{700};
– domain: large;

As noted by *Huth* (2002), the stepwise regression using gridpoint values performs better than the one using principal components at stations where local drivers such as particular topography have an important role, like is the case of the stations analysed here. The reason could be that the use of gridpoint values allows

Table 4 Sensitivity tests for different sets of predictors and domains as reported in table 2. The statistical method is the pointwise regression. Results for the Fabra and Ebre stations. The model performances are estimated by the explained variance (R^2) and the error variance (s^2).

Fabra	R^2, s^2 (T_{max})	R^2, s^2 (T_{min})	R^2, s^2 (DTR)
$Exp7$	79, 0.16	86, 0.06	45, 0.07
$Exp8$	78, 0.16	86, 0.06	31, 0.85
$Exp9$	83, 1.21	90, 0.04	51, 0.06
$Exp10$	69, 0.23	78, 0.09	22, 0.09
$Exp11$	69, 0.23	79, 0.09	22, 0.09
$Exp12$	75, 0.18	85, 0.06	13, 0.10
Ebre	R^2, s^2 (T_{max})	R^2, s^2 (T_{min})	R^2, s^2 (DTR)
$Exp7$	69, 0.23	72, 0.07	65, 0.13
$Exp8$	69, 0.23	69, 0.08	61, 0.14
$Exp9$	76, 0.18	86, 0.04	65, 0.13
$Exp10$	63, 0.28	89, 0.08	33, 0.23
$Exp11$	63, 0.28	64, 0.09	33, 0.23
$Exp12$	47, 0.38	75, 0.06	75, 0.32

some of the local peculiarities to be explained by the value of a single (or a few) grid-point(s) specifically selected for the site thus providing the best fit. Regarding the predictors, although the combination of SLP, T_{850}, R_{700} performs similarly to the MLR which use the $T2m$ alone, the former set was chosen since the free-atmosphere variables are prognostic, and, in principle, more reliable than diagnostic variables, such as $T2m$. Finally, as expected, the largest domain shows the best performance. In fact, the main component of temperature at annual scale is presumably governed by large circulation systems (*Benestad et al.*, 2008). Consequently, in the next sections we applied the statistical model with the above parameters.

4.2 Out-of-sample test

In this section we want to answer the question: is the model able to reproduce the homogeneous part of the series? To achieve this objective, we have applied the leave-one-out cross validation described in Section 3.2.

Figure 5 shows the out-of-sample validation for the Ebre and Fabra series, respectively. The "suspect" periods are not shown, in order to focus on the remaining part, i.e. the "reliable" one. These figures show the median of the ensemble of the 1000 random realizations together with the 2.5th and the 97.5th percentiles on the ensemble of out-of-sample test. With very rare exceptions, the observed data fall inside the ensemble bands. Note that the average error is around ±5% of the annual value. These results suggest that this methodology is suitable to reconstruct the local temperature series. Figure 5 also shows the reconstruction of the gaps in the Ebre series.

For the sake of brevity, only the selected model results are shown here, but several models (with different predictor combination) were tested. We have seen that while the models selected in the in-sample calibration perform correctly even in out-of-sample mode, the performance of the models that consider various combinations of only two predictors are not satisfactory.

4.3 Testing the "suspect" periods

In the previous sections we have calibrated the statistical model and we have showed that, within certain limits, it is able to reproduce the observed series. Here we test the statistical model in the "suspect" periods.

Figure 6 shows the out-of-sample validation for the Ebre and Fabra series, including the "suspect" periods, focusing on the relevant part of the series (i.e. around the "suspect" periods). For the Ebre observatory, the strange behavior stays inside (or, except for one year, barely exceeds) the uncertainty limits, so the nature of this anomalous period remains unclear. Whereas the Fabra series shows clearly that the odd peak in the DTR series is far beyond the error bands of the method, suggesting the artificial origin of the inhomogeneity. Note that the T_{max} series in the anomalous period is close to the upper error band, whereas the T_{min} is below the lower error band. Finally it is apparent that the DTR magnifies the less evident inhomogeneity with the T_{max} and T_{min} series.

4.4 Test an unknown "suspect period"

The last test performed is the cross validation using a time windows of 5 years for the Fabra series (1914-2008) and for the Ebro series (1905-2004). These values were chosen in order to divide the total duration of the series (95 and 100 years, respectively) into equal parts. Obviously, if we do not exclude an anomalous part of the training period, the performance of the model may worsen. However, Figure 7 (to avoid repetition we show only the DTR of both stations) confirms that this methodology is able to detect the anomalous part of the Fabra series. The result does not change for the Ebre series either: Figure 7 indicates that the "suspect" period falls inside the error bands. Furthermore, the first years of the series (up to around 1920) are close to the lower limit of the error band. This result suggests that the early years of this series (up around the '20s), in which there was a shift of the station (unfortunately the date is unknown) are "suspect". Thus, in the following trend analysis we have reconstructed the anomalous part of

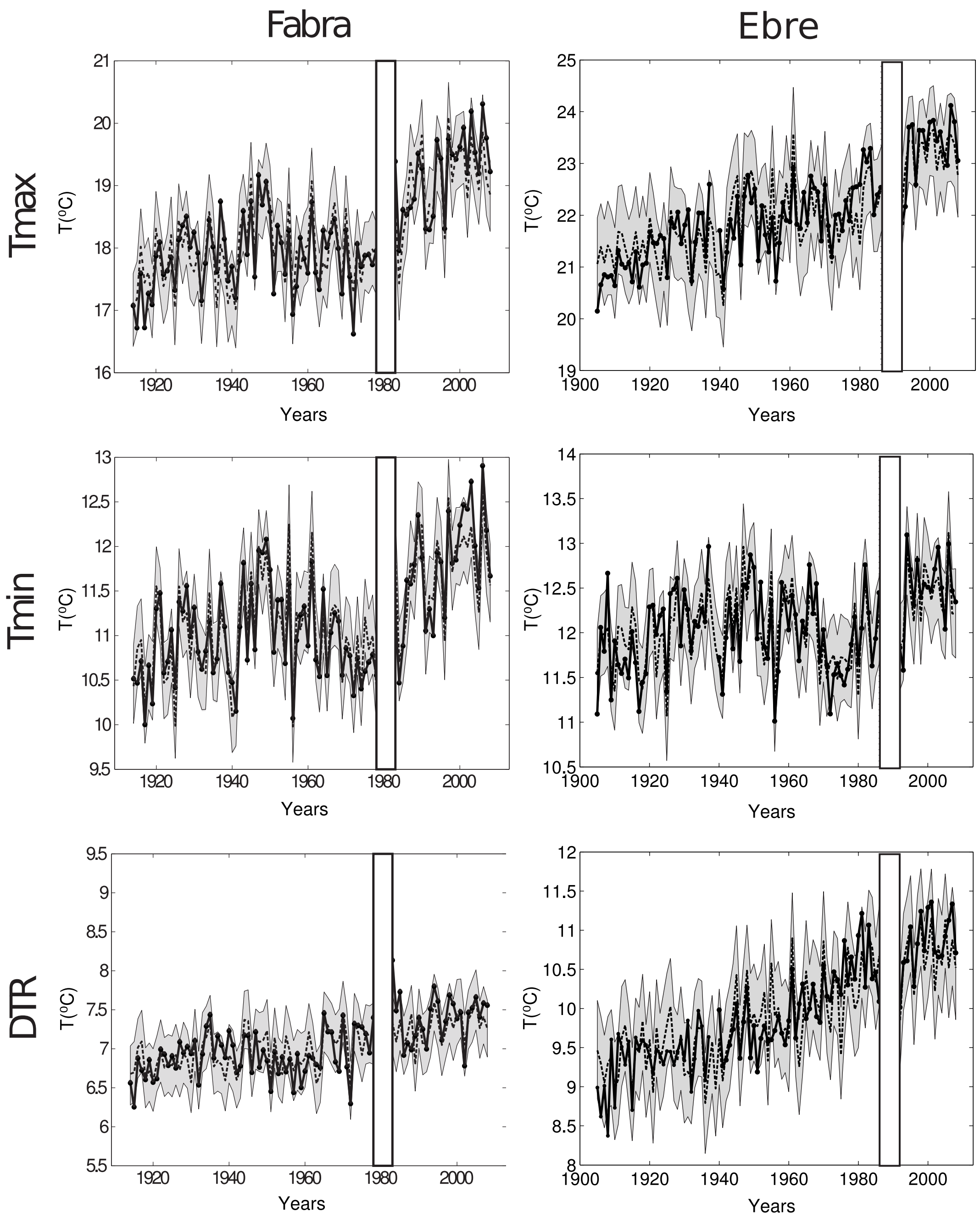

Fig. 5 Out-of-sample validation for T_{max} (top), T_{min} (central) and DTR (bottom) series of Fabra (left) and Ebre (right) observatories. The solid line with filled black points represents the data. The dotted line is the median of 1000 different out-of-sample predictions, whereas the grey bands indicate the 2.5th and 97.5th percentiles of these 1000 values. See the running text for more details.

the Fabra station and the periods 1905-1920 and 1938-1939 (the missing years) for the Ebre observatory.

Finally note that no other anomalous period / break-point years are detected. This suggests that the other changes (like the two changes of thermometer in 1921 and 1961 at Fabra Observatory) have probably caused relatively smaller shifts.

4.5 Trend analysis results

Already by visual inspection of Fig. 1 and Fig. 5, we observe that the annual series of T_{max}, T_{min} and DTR show a positive trend. Note that for all the series, the highest values are in the latest decades. We applied the Mann-Kendall test to both observatories, for the origi-

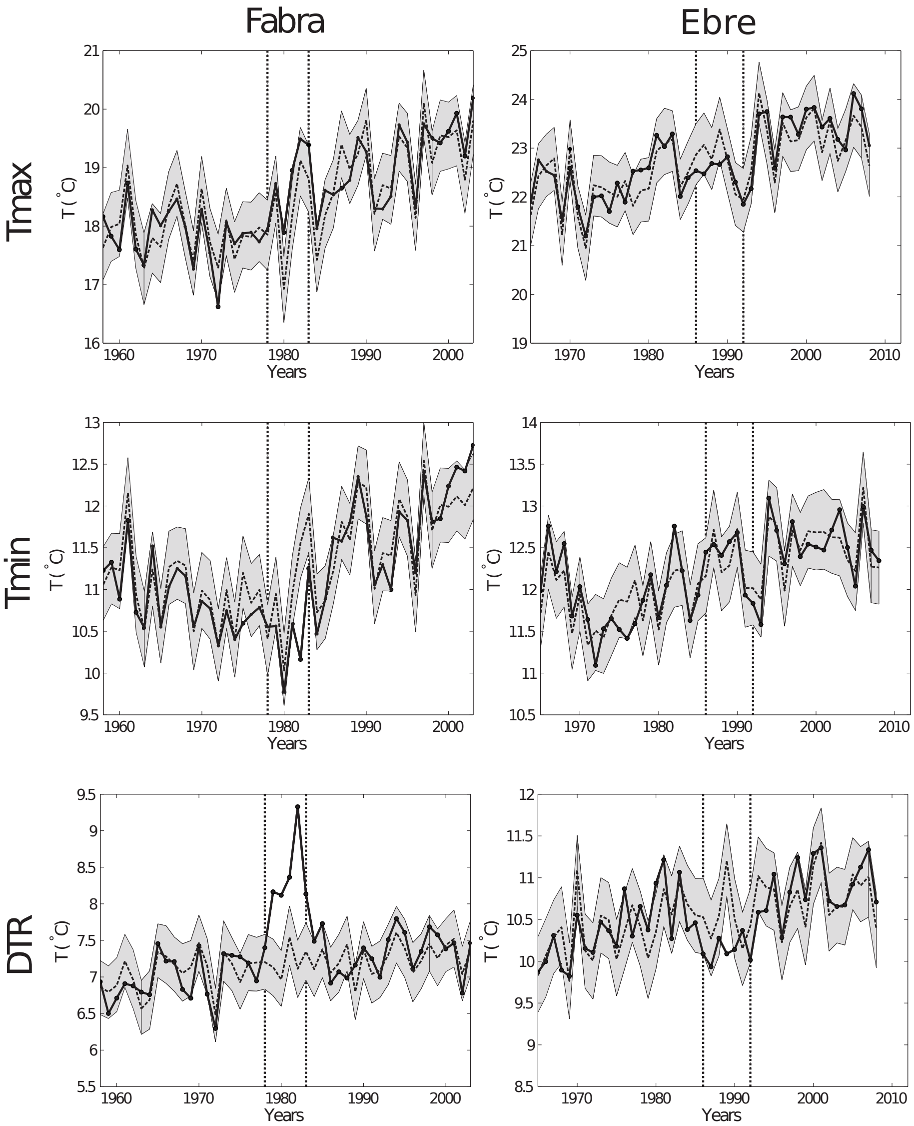

Fig. 6 Testing the "suspect period". Results for T_{max} (top), T_{min} (central) and DTR (bottom) series of Fabra (left) and Ebre (right) observatories. The solid line with filled black points represents the data. The dotted line is the median of 1000 different out-of-sample predictions, whereas the grey bands indicate the 2.5th and 97.5th percentiles of these 1000 values. The "suspect periods" are delimitated by vertical dashed lines.

nal series and for the reconstructed ones (according to the Sec. 4.4). This trend analysis confirms significant positive trends for both stations (Table 5). In addition to analyze the entire period of the records, we also analyze the series over the period 1950-2008, since the early 20th of both instrumental and reanalysis data should be taken with caution. Indeed, some concerns about the homogeneity in the first part of the 20th Century of 20CR could be raised (*Ferguson and Villarini*, 2012). Therefore, even though our experiment proved to be in accordance with the observational data, any trends derived from the early 20th should be taken with caution.

This clear signal of increase is coherent to the observed warming at Mediterranean (*Efthymiadis et al.*, 2011) and global scale (*IPCC*, 2007). Also the increase in DTR is consistent with the studies of *Brunet et al.* (2007) and *Klok and Klein Tank* (2009) which detected a positive trend of DTR over all of Spain and Europe, respectively. Finally, note that the reconstructed series show a lower rate of increase, also considering the anal-

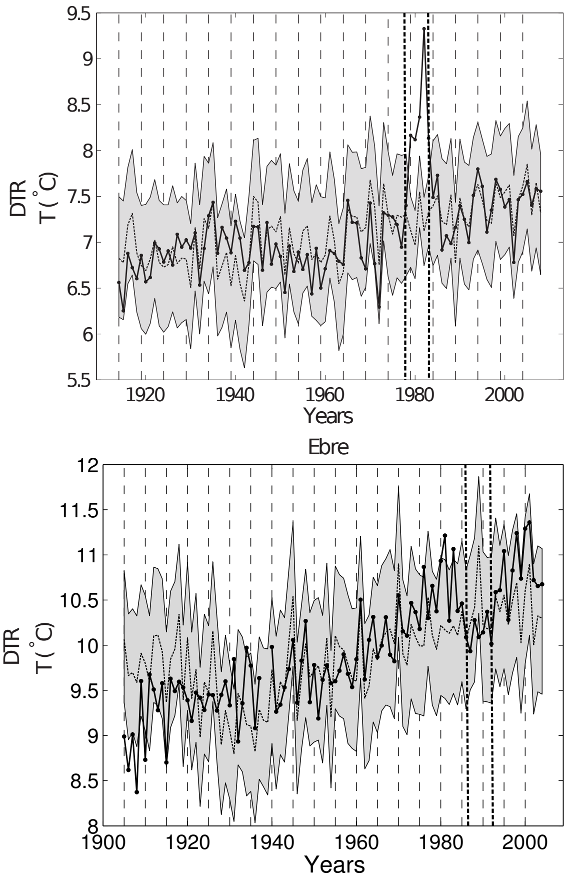

Fig. 7 Testing an unknown "anomalous period". The results for the DTR series of Fabra (top) and Ebre (bottom) observatories. The solid line with filled black points represents the data. The dotted line is the median of 1000 different out-of-sample predictions, whereas the grey bands indicate the 2.5th and 97.5th percentiles of these 1000 values. The potential "suspect" periods are delimited by vertical dashed lines. The defined "suspect periods" are delimitated by bold vertical dashed lines.

ysis for the more recent decades. The results obtained in the second and shorter period should be more reliable.

Observatory	T_{max}	T_{min}	DTR
$Ebre_{1905-2008}$	0.25 (0.16)	0.06 (0.06)	0.19 (0.12)
$Ebre_{1950-2008}$	0.30 (0.21)	0.07 (0.05)	0.23 (0.17)
$Fabra_{1914-2008}$	0.20 (0.17)	0.11 (0.11)	0.09 (0.06)
$Fabra_{1950-2008}$	0.38 (0.30)	0.22 (0.18)	0.15 (0.09)

Table 5 Trend analysis results for T_{max}, T_{min} and DTR (bottom) original and reconstructed (in parenthesis) series of Fabra and Ebre observatories, considering the entire available period of the series (respectively 1914-2008 and 1905-2008) and the period 1950-2008. Units are °C/10y. Pvalue $\leq$ 0.05.

5 Conclusions

The main objectives of this paper were: (i) to test the feasibility of a statistical technique, the Multiple Linear pointwise Regression downscaling technique (MLR), to downscale large scale reanalysis predictor fields (20CR in our case) in order to test the homogeneity of local temperature time series and reconstruct them, and (ii) to analyse the temperature evolution in the last century from two centennial observatories in the North-East of Spain.

The MLR method constitutes an innovative approach which can be complementary to the more standard homogeneity test protocols (*Wijngaard et al.*, 2003). This methodology was applied to two centennial observatories, Fabra and Ebre, regarded as high quality stations for climate studies in Spain, which do not have nearby temperature series to compare with. The temperature series used in this study span the respective periods, 1914-2008 and 1905-2008. For both stations the statistical downscaling model has been calibrated and independently tested over the reliable periods with good results. Specifically, we have positively tested the method to reproduce those values omitted from the calibration. This suggests that the MLR method is able to reconstruct the local temperature series.

The results indicate the feasibility of this method both to detect potentially inhomogeneous regions of the series and to reconstruct them. Indeed, our results are in high agreement with the metadata of both observatories which showed that, while for the Ebre station there are no documented substantial variations in the station environment during its anomalous period, while for Fabra, the thermometer was in a non-homologated shelter between 1978-1983. Finally, we applied this method to reconstruct the interrupted period (1938-1939) of the Ebre Observatory, obtaining a complete series for the period 1905-2008.

The methodology introduced here is flexible and adaptable with other reanalysis preserving the applicability, simplicity and robustness of the technique (*Benestad et al.*, 2008). In this study we made use of the large scale fields from the 20CR, the only available reanalysis covering the entire 20th century. The consistency between the observed data and the reanalysis downscaled values gives some confidence not only to the MLR method, but also on the quality of the 20CR itself. However, since some concerns about its homogeneity in the first part of the 20th Century could be raised, any trends derived from the early 20th should be taken with caution.

For this reason we have tested the series for trends considering the whole available period and the period since 1950. The trend analysis results point to a clear signal of warming, with a larger rate of increase for the maximum temperature (respect to the minimum one), for the more recent decades and for the original series (respect to the reconstructed ones). Regarding the ongoing controversy about the attribution of the nature of the station inhomogeneities, our results show that it is unlikely that they are driven by volcanic factors, especially in the Fabra case. In the Ebre observatory there is no documented change in the station and the suspicious period is inside the error bands. Thus, further studies are recommended to understand the nature of its anomalous period.

In summary, these promising results strongly point towards the possibility of establishing a new procedure both to detect inhomogeneities and reconstruct anomalous periods in a temperature time series without needing nearby stations to compare with. It is the aim of future studies to consolidate the methodology by means of its application to other temperature series, to analyse its applicability to other variables, considering other reanalysis and other statistical downscaling techniques, and explore the reconstruction of monthly means or even daily data. In the case of reconstruction of monthly means, the process should be quite straightforward (developing a model for each month), but with daily data the effect of the autocorrelation of the time series deserves a deeper study.

Acknowledgements This work was supported by esTcena project (Exp. 200800050084078), a strategic action from Plan Nacional de I+D+i 2008-2011 funded by the Spanish Ministry of Medio Ambiente y Medio Rural y Marino. We are most grateful to AEMET, the Ebre Observatory and the Reial Academia de Ciències i Arts de Barcelona for the data and metadata support. Special thanks to Dr. Prohom, Dr. van der Schrier and Mr. Solé for their helpful discussions on the matter.

References

Aguilar, E., I. Auer, M. Brunet, T. C. Peterson, and J. Wieringa (2003), Guidelines on climate metadata and homogenization, *Tech. Rep. WMO/TD–No. 118*, WMO.

Alexandersson, H. (1986), A homogeneity test applied to precipitation data, *Journal of Climatology*, *6*(6), 661–675.

Benestad, R., I. Hanssen-Bauer, and D. Chen (2008), *Empirical-Statistical Downscaling*, World Scientific Publishers.

Bengtsson, L., S. Hagemann, and K. I. Hodges (2004), Can climate trends be calculated from reanalysis data?, *J. Geophys. Res.*, *109*(D11).

Brunet, M., P. D. Jones, J. Sigró, O. Saladié, E. Aguilar, A. Moberg, P. M. Della-Marta, D. Lister, A. Walther, and D. López (2007), Temporal and spatial temperature variability and change over Spain during 1850-2005, *Journal of Geophysical Research*, *112*(D12), 1–28.

Buishand, T. (1982), Some methods for testing homogeneity of rainfall records, *Journal of Hydrology*, *58*, 11–27.

Calmanti, S., L. Motta, M. Turco, and A. Provenzale (2007), Impact of climate variability on alpine glaciers in northwestern Italy, *International Journal of Climatology*, *27*(15), 2041–2053.

Compo, G. P., J. S. Whitaker, and P. D. Sardeshmukh (2006), Feasibility of a 100-Year Reanalysis Using Only Surface Pressure Data, *Bulletin of the American Meteorological Society*, *87*(2), 175.

Compo, G. P., J. S. Whitaker, P. D. Sardeshmukh, N. Matsui, R. J. Allan, X. Yin, B. E. Gleason, R. S. Vose, G. Rutledge, P. Bessemoulin, S. Brönnimann, M. Brunet, R. I. Crouthamel, A. N. Grant, P. Y. Groisman, P. D. Jones, M. C. Kruk, A. C. Kruger, G. J. Marshall, M. Maugeri, H. Y. Mok, Ø. Nordli, T. F. Ross, R. M. Trigo, X. L. Wang, S. D. Woodruff, and S. J. Worley (2011), The twentieth century reanalysis project, *Quarterly Journal of the Royal Meteorological Society*, *137*(654), 1–28.

Conrad, V., and L. W. Pollak (1950), *Methods in Climatology*, 459 pp., Harvard University Press.

Curto, J. J., E. Also, E. Pallé, and J. G. Solé (2009), Sunshine and synoptic cloud observations at Ebro Observatory, 1910-2006, *International Journal of Climatology*, *29*(14), 2183–2190.

Dee, D. P., E. Källén, A. J. Simmons, and L. Haimberger (2011), Comments on reanalyses suitable for characterizing long-term trends, *Bulletin of the American Meteorological Society*, *92*(1), 65–70.

Ducré-Robitaille, J.-F., L. A. Vincent, and G. Boulet (2003), Comparison of techniques for detection of discontinuities in temperature series, *International Journal of Climatology*, *23*(9), 1087–1101.

Efthymiadis, D., C. M. Goodess, and P. D. Jones (2011), Trends in Mediterranean gridded temperature extremes and large-scale circulation influences, *Natural Hazards and Earth System Science*, *11*(8), 2199–2214.

Fall, S., D. Niyogi, A. Gluhovsky, R. A. Pielke, E. Kalnay, and G. Rochon (2010), Impacts of land use land cover on temperature trends over the continental United States: assessment using the North American Regional Reanalysis, *International Journal of Climatology*, *30*(13), 1980–1993.

Ferguson, C. R., and G. Villarini (2012), Detecting inhomogeneities in the Twentieth Century Reanalysis over the central United States, *J. of Geophysical Research*, (117), 11.

Fowler, H. J., S. Blenkinsop, and C. Tebaldi (2007), Linking climate change modelling to impacts studies: Recent advances in downscaling techniques for hydrological modelling, *International Journal of Climatology*, *27*(12), 1547–1578, General Assembly of the European-Geosciences-Union, Vienna, Austria, APR, 2006.

Franke, J., J. F. González-Rouco, D. Frank, and N. E. Graham (2010), 200 years of European temperature variability: insights from and tests of the proxy surrogate reconstruction analog method, *Climate Dynamics*, *37*(1-2), 133–150.

Graham, N. E., M. K. Hughes, C. M. Ammann, K. M. Cobb, M. P. Hoerling, D. J. Kennett, J. P. Kennett, B. Rein, L. Stott, P. E. Wigand, and T. Xu (2007), Tropical Pacific–Mid-Latitude teleconnections in medieval times, *Climatic Change*, *83*(1-2), 241–285.

Haimberger, L. (2007), Homogenization of Radiosonde Temperature Time Series Using Innovation Statistics, *Journal of Climate*, *20*(7), 1377–1403.

IPCC (2007), *IPCC Fourth Assessment Report: Climate Change 2007 (AR4)*, Geneva: IPCC.

Huth, R. (1999), Statistical downscaling in central Europe: evaluation of methods and potential predictors, *Climate Research*, *13*, 91–101.

Huth, R. (2002), Statistical downscaling of daily temperature in central Europe, *Journal of Climate*, *15*(13), 1731–1742.

Kalnay, E., and M. Cai (2003), Impact of urbanization and land-use change on climate, *Nature*, *423*(6939), 528–531.

Kalnay, E., M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, A. Leetmaa, R. Reynolds, M. Chelliah, W. Ebisuzaki, W. Higgins, J. Janowiak, K. C. Mo, C. Ropelewski, J. Wang, R. Jenne, and D. Joseph (1996), The NCEP/NCAR 40-Year Reanalysis Project, *Bulletin of the American Meteorological Society*, *77*(3), 437–471.

Kalnay, E., M. Cai, H. Li, and J. Tobin (2006), Estimation of the impact of land-surface forcings on temperature trends in eastern United States, *J. Geophys. Res.*, *111*(D6).

Klok, E. J., and A. M. G. Klein Tank (2009), Updated and extended european dataset of daily climate observations, *International Journal of Climatology*, *29*(8), 1182–1191.

Llebot, J. (2010), El canvi climtic a catalunya. 2n informe del grup dexperts en canvi climtic de catalunya, pp. 1–32, Generalitat de Catalunya.

Maraun, D., F. Wetterhall, A. M. Ireson, R. E. Chandler, E. J. Kendon, M. Widmann, S. Brienen, H. W. Rust, T. Sauter, M. Themessl, V. K. C. Venema, K. P. Chun, C. M. Goodess, R. G. Jones, C. Onof, M. Vrac, and I. Thiele-Eich (2010), Precipitation Downscaling under Climate Change: Recent Developments to Bridge the Gap Between Dynamical Downscaling Models and the End User, *Reviews of Geophysics*, *48*.

Menne, M. J., and C. N. Williams (2009), Homogenization of Temperature Series via Pairwise Comparisons, *Journal of Climate*, *22*(7), 1700–1717.

Onogi, K., J. Tsutsui, H. Koide, M. Sakamoto, S. Kobayashi, H. Hatsushika, T. Matsumoto, N. Yamazaki, H. Kamahori, K. Takahashi, S. Kadokura, K. Wada, K. Kato, R. Oyama, T. Ose, N. Mannoji, and R. Taira (2007), The JRA-25 Reanalysis, *Journal of the Meteorological Society of Japan*, *85*(3), 369–432.

Peterson, T. C., D. R. Easterling, T. R. Karl, P. Groisman, N. Nicholls, N. Plummer, S. Torok, I. Auer, R. Boehm, D. Gullett, L. Vincent, R. Heino, H. Tuomenvirta, O. Mestre, T. Szentimrey, J. Salinger, E. J. FØrland, I. Hanssen-Bauer, H. Alexandersson, P. Jones, and D. Parker (1998), Homogeneity adjustments of in situ atmospheric climate data: a review, *International Journal of Climatology*, *18*(13), 1493–1517.

Pettitt, A. (1979), A non-parametric approach to the change-point detection, *Applied Statistics*, *28*, 126–135.

Pielke, R., J. Nielsen-Gammon, C. Davey, J. Angel, O. Bliss, N. Doesken, M. Cai, S. Fall, D. Niyogi, K. Gallo, R. Hale, K. G. Hubbard, X. Lin, H. Li, and S. Raman (2007a), Documentation of Uncertainties and Biases Associated with Surface Temperature Measurement Sites for Climate Change Assessment, *Bulletin of the American Meteorological Society*, *88*(6), 913.

Pielke, R. A., C. A. Davey, D. Niyogi, S. Fall, J. Steinweg-Woods, K. Hubbard, X. Lin, M. Cai, Y.-K. Lim, H. Li, J. Nielsen-Gammon, K. Gallo, R. Hale, R. Mahmood, S. Foster, R. T. McNider, and P. Blanken (2007b), Unresolved issues with the assessment of multidecadal global land surface temperature trends, *J. Geophys. Res.*, *112*(D24).

Prohom, M., and M. Herrero (2008), Towards the creation of a climatic database for Catalonia (18th to 21st centuries), *Tethys, Journal of Weather and Climate of the Western Mediterranean*, pp. 3–11.

Prohom Duran, M. J., A. Puertas Castro, and A. Gázquez Picó (2009), La sèrie de temperatura de l'Observatori Fabra (1913-2008): Anàlisi de qualitat, homogeneïtat i tendèndcies, in *XV Jornades de Meteorologia Eduard Fontserè*, pp. 41–50.

Reeves, J., J. Chen, X. L. Wang, R. Lund, and Q. Q. Lu (2007), A Review and Comparison of Changepoint Detection Techniques for Climate Data, *Journal of Applied Meteorology and Climatology*, *46*(6), 900–915.

Seguí-Grau, J. (2003), Análisis de la Serie de Temperatura del Observatorio del Ebro (1894-2002), *Tech. rep.*, Observatori de l'Ebre, Roquetes (Tarragona).

Sturaro, G. (2003), A closer look at the climatological discontinuities present in the NCEP/NCAR reanalysis temperature due to the introduction of satellite data, *Climate Dynamics*, *21*(3-4), 309–316.

Timbal, B., A. Dufour, and B. McAvaney (2003), An estimate of future climate change for western france using a statistical downscaling technique, *Climate Dynamics*, *20*(7-8), 807–823.

Timbal, B., E. Fernandez, and Z. Li (2009), Generalization of a statistical downscaling model to provide local climate change projections for Australia, *Enviromental Modelling & Software*, *24*(3), 341–358.

Toreti, a., F. G. Kuglitsch, E. Xoplaki, P. M. Della-Marta, E. Aguilar, M. Prohom, and J. Luterbacher (2010), A note on the use of the standard normal homogeneity test to detect inhomogeneities in climatic time series, *International Journal of Climatology*, pp. 630–632.

Uppala, S., P. Kallberg, A. Simmons, U. Andrae, V. Bechtold, M. Fiorino, J. Gibson, J. Haseler, A. Hernandez, G. Kelly, X. Li, K. Onogi, S. Saarinen, N. Sokka, R. Allan, E. Andersson, K. Arpe, M. Balmaseda, A. Beljaars, L. Van De Berg, J. Bidlot, N. Bormann, S. Caires, F. Chevallier, A. Dethof, M. Dragosavac, M. Fisher, M. Fuentes, S. Hagemann, E. Holm, B. Hoskins, L. Isaksen, P. Janssen, R. Jenne, A. McNally, J. Mahfouf, J. Morcrette, N. Rayner, R. Saunders, P. Simon, A. Sterl, K. Trenberth, A. Untch, D. Vasiljevic, P. Viterbo, and J. Woollen (2005), The ERA-40 re-analysis, *Quarterly Journal of the Royal Meteorological Society*, *131*(612, Part B), 2961–3012.

Vincent, L. a. (1998), A Technique for the Identification of Inhomogeneities in Canadian Temperature Series, *Journal of Climate*, *11*(5), 1094–1104.

Von Neumann, J. (1941), Distribution of the ratio of the mean square successive difference to the variance, *Annals of Mathematical Statistics*, *13*, 367–395.

Von Storch, H., and F. W. Zwiers (1999), *Statistical analysis in climate research*, Cambridge University Press, Cambridge.

Wijngaard, J., A. Klein Tank, and G. Können (2003), Homogeneity of 20th century European daily temperature and precipitation series, *International Journal of Climatology*, *23*(6), 679–692.

Wilby, R., S. Charles, E. Zorita, and B. Timbal (2004), Guidelines for use of climate scenarios developed from statistical downscaling methods, *Tech. rep.*, IPCC.

Wilks, D. S. (2006), *Statistical Methods in the Atmospheric Sciences*, 627 pp., Academic Press, Burlington, MA.