Clim Dyn (2009) 33:707–722
DOI 10.1007/s00382-009-0561-4

Adaptation of the optimal fingerprint method for climate change detection using a well-conditioned covariance matrix estimate

Aurélien Ribes · Jean-Marc Azaïs · Serge Planton

Received: 11 April 2008 / Accepted: 18 March 2009 / Published online: 12 May 2009

Abstract The "optimal fingerprint" method, usually used for detection and attribution studies, requires to know, or, in practice, to estimate the covariance matrix of the internal climate variability. In this work, a new adaptation of the "optimal fingerprints" method is presented. The main goal is to allow the use of a covariance matrix estimate based on an observation dataset in which the number of years used for covariance estimation is close to the number of observed time series. Our adaptation is based on the use of a regularized estimate of the covariance matrix, that is well-conditioned, and asymptotically more precise, in the sense of the mean square error. This method is shown to be more powerful than the basic "guess pattern fingerprint", and than the classical use of a pseudo-inverted truncation of the empirical covariance matrix. The construction of the detection test is achieved by using a bootstrap technique particularly well-suited to estimate the internal climate variability in real world observations. In order to validate the efficiency of the detection algorithm with climate data, the methodology presented here is first applied with pseudo-observations derived from transient regional climate change scenarios covering the 1960–2099 period. It is then used to perform a formal detection study of climate change over France, analyzing homogenized observed temperature series from 1900 to 2006. In this case, the estimation of the covariance matrix is only based on a part of the observation dataset. This new approach allows the confirmation and extension of previous results regarding the detection of an anthropogenic climate change signal over the country.

Keywords Anthropogenic climate change · Detection · Optimal fingerprints · Covariance matrix estimation

A. Ribes (✉) · S. Planton
CNRM-GAME, Météo France-CNRS, 42 av G Coriolis, 31057 Toulouse, France
e-mail: aurelien.ribes@cnrm.meteo.fr

S. Planton
e-mail: serge.planton@meteo.fr

J.-M. Azaïs
Université de Toulouse, UPS, IMT, LSP, 118 Route de Narbonne, 31062 Toulouse, France
e-mail: azais@cict.fr

1 Introduction

According to the IPCC third assessment report (IPCC 2001), "detection is the process of demonstrating that an observed change is significantly different (in a statistical sense) than can be explained by natural internal variability". As a consequence of this definition, detection is mainly a statistical issue. A first methodology for the detection of a model-predicted signal in observational data has been proposed by Hasselmann (1979, 1993). This method, commonly referred to as the "optimal fingerprint" method, is based on a maximization of the signal to noise ratio, with a classical approach of statistics.

The "optimal fingerprint" method was first applied to climate data to detect a signal of change in the global surface temperature by Hegerl et al. (1996), and subsequently to other parameters, such as the free atmosphere temperature and oceanic data sets (Barnett et al. 2001; Tett et al. 2002). The issue of the scale on which the climate change signal can be detected has also been investigated (Stott and Tett 1998; Zwiers and Zhang 2003), and there have been some successful detection studies at a regional scale (Stott 2003).

The main difficulty when using the "optimal fingerprint" is that it requires to know with a good accuracy the expected direction of climate change and the covariance matrix associated with the internal climate variability of the climate vector used. This is particularly challenging when the detection is applied at the sub-regional scale (Spagnoli et al. 2002).

On the one hand, the expected signal of climate change is classically taken from climate simulations using either general circulation models (GCMs) or regional climate models (RCMs) according to the scale of the analysis. Although various sources of uncertainties are associated to this "guess-pattern", detection studies usually assume that this direction of change is known.

On the other hand, the covariance matrix of the observations (namely the "covariance matrix" associated to the internal variability) is never known exactly, and has to be estimated. When studying the climate, the estimation of this covariance matrix C can be based directly on observations, as in Spagnoli et al. (2002), or on pseudo-observations generated by climate model simulations without external forcing, as in Hegerl et al. (1996, 1997), and many other detection-attribution studies since. The choice of pseudo-observations can be justified by several reasons. Firstly, the matrix C denotes the only internal climate variability, whereas observations may include a part of natural variability, due to changes in natural forcings like solar radiation or volcanic aerosols, and a part of externally forced variability, notably due to human induced greenhouse effect. Secondly, the number of available years of observation is generally limited to a hundred or less, deteriorating the estimation.

The main difficulty when estimating C using pseudo-observations is to quantify the errors and uncertainties due to model representation. The sub-regional detection case is even more problematic than the detection at continental scales due to limited available long-term simulations of unforced climate variability with RCMs. To avoid this difficulty, in this study, the use of observations to estimate C is preferred. The difficulty in distinguishing forced and internal variability can be overcome by noting that the use of real observations yields an overestimation of the internal variability from the total variability. This overestimation is conservative for the statistical tests performed. The lack of observational years remains the main weakness, and this point is precisely dealt with by the methodological development proposed here. The new "optimal fingerprint" adaptation we introduce here allows us to perform a relatively efficient detection test in the unfavorable case where the length in years of the observed series used for the estimation of C, n, is of the same order as the number of series p. This allows the undertaking of a detection study without using pseudo-observations for the estimation of the internal variability.

Another key difficulty for applying the "optimal fingerprint" method, is to compute the inverse of the matrix C. The classical adaptation of the "optimal fingerprint" computes an estimate of the covariance C at first, and then takes into account a pseudo-inverted truncation of this estimate. This is a way to decrease the errors involved in the estimation procedure. These errors may be dramatic when taking the inverse of the matrix. However, such a procedure requires the use of a truncation parameter, the value of which is difficult to choose.

The main contribution of this paper is to revisit the "optimal fingerprint" method, and to propose an adaptation of this method that yields a test procedure avoiding the use of a pseudo-inverted truncation of a C estimate (Sect. 2.3). This procedure is shown to be efficient in the sense that it is more powerful than the classical adaptation of the "optimal fingerprint", and than a relatively simple test named "guess-pattern fingerprint" also introduced by Hegerl et al (1996). This paper also discusses the choice of the parameter truncation of the classical adaptation, focusing on the efficiency of the corresponding test.

The hypothesis and methods used in this study are introduced in the next section. Section 3 presents a comparison of these different methods, especially demonstrating why the new method may be preferred. Section 3 then provides some illustrations of the new method, where the estimation of C is based on observations. We conclude in the last section.

2 Presentation of the methods

2.1 Optimal climate change detection

2.1.1 Detection framework

We start by introducing some basic notations and hypothesis, following Hasselmann (1993). The observed climate state will be represented by a p-dimensional vector ψ, each coordinate representing one observational station. Within the probabilistic framework used, ψ is considered as a random vector, taking one value each year, typically the annual or seasonal average of one climatic parameter.

The assumption is made that, in a climate change context, the observed climate vector may be decomposed such as:

$$\psi = \psi^s + \tilde{\psi}, \tag{1}$$

where ψ^s denotes the climate change signal, and $\tilde{\psi}$ denotes an internal-variability realization. It is also assumed that $\tilde{\psi}$ is centered, that is to say that:

$$\mathrm{E}(\tilde{\psi}) = 0, \tag{2}$$

denoting by E the mathematical expectation. This is virtually the case by removing the mean. Note that the first term on the right hand side of (1) is a consequence of external forcing, and it is not random. This decomposition assumes in particular that the internal variability is the same with or without climate change.

Furthermore, a second assumption is made concerning the first term of the decomposition with:

$$\psi^s = \mu g, \tag{3}$$

where μ denotes a real amplitude factor, and g is the expected p-dimensional climate change vector, taken from climate model simulations, called "guess-pattern". In practice, g corresponds to the response of the earth-system to an external forcing, which is generally, in the case of climate change detection, the anthropogenic climate change. Moreover, g is assumed to be known but is, as above mentioned, derived from an ensemble of climate model simulations.

With these notations, a detection study consists in applying a statistical test to the "null" hypothesis H_0: "$\mu = 0$" against the alternative hypothesis H_1: "$\mu > 0$". These hypothesis can be rewritten H_0: "E(ψ) = 0" and H_1: "E(ψ) = μg, with $\mu > 0$".

Note that this formulation and this type of statistical test only deal with the climate expectation, assuming that the noise structure is invariant. The study of possible changes on the variability could be investigated, but with a different approach.

As mentioned in the introduction, the matrix C, namely the covariance matrix of ψ due to internal climate variability, is a key parameter of the optimal climate change detection formalism. It will be seen that the "optimal fingerprint" method is derived while assuming that this matrix is known (Sect. 1). In practice, this is not the case, and C has to be estimated. Therefore, we will assume that n years $(\psi_i)_{1 \le i \le n}$, are available for estimating C, and another year, denoted ψ_{n+1}, will be tested for the climate change hypothesis. Note that for clarity, the theoretical part of the study will be presented with only one tested observation ψ_{n+1}, although in practice, the detection procedure can be applied to several (see Sect. 4).

Some assumptions are made about the data $(\psi_i)_{1 \le i \le n}$, that can be either observed or taken from a control run. On the one hand, they are assumed to be uncontaminated by external forcings. If these values are observed, this condition is not satisfied, but the detection will be more conservative, as stated before. On the other hand, these data are assumed to have a covariance C, that is to say the same covariance matrix than the tested observation ψ_{n+1}. This assumption can be discussed in the case of pseudo-observations taken from a climate model control run. $\tilde{\psi}$ being the random term of covariance C due to internal climate variability in Eq. (1), both assumptions can be summarized by writing:

$$\psi_i = \tilde{\psi}_i, \quad \text{for } 1 \le i \le n. \tag{4}$$

2.1.2 The optimal fingerprint

In order to introduce the optimal fingerprint, the covariance matrix C of the internal climate variability component $\tilde{\psi}$ is temporarily assumed to be known. $\tilde{\psi}$ being the only random component of ψ, C also can be seen as the covariance of ψ. Note that C being known, the data $(\psi_i)_{1 \le i \le n}$ are not used in this Sect. 2.1.2.

The fingerprint approach consists in studying the set of the linear detection variables, to determine the best one, according to the signal to noise ratio. We introduce a family of "linear detection variables" d_f with:

$$d_f = \langle \psi_{n+1}, f \rangle, \tag{5}$$

where f is an unspecified p-dimensional vector, and $\langle , \rangle$ is the symbol for the standard p-dimensional euclidean scalar product.

The use of the variable d_f naturally leads to a test T_f, whose rejection region is

$$W_f = \left\{ \psi_{n+1}, d_f = \langle \psi_{n+1}, f \rangle \ge d_f^{(\alpha)} \right\}, \tag{6}$$

considering a α-level test, and denoting by $d_f^{(\alpha)}$ the $(1-\alpha)$-quantile of d_f under H_0.

One of the main result of Hasselmann (1993) is to show that the d_f maximizing the signal-to-noise ratio is $d_{C^{-1}g}$. To use the optimal fingerprint method then leads to use the test $T_{C^{-1}g}$, that is optimal according to the signal-to-noise ratio. In the following, f_o will denote the optimal fingerprint vector:

$$f_o = C^{-1} g, \tag{7}$$

used in this test T_{f_o}.

Some others interpretations have been proposed for this result. At first, it has been shown that the "optimal fingerprint" can be seen as a regression technique (Allen and Tett 1999; Allen and Stott 2003). With this formalism, C^{-1} can be seen as an optimal metric for the regression. This point of view has also been developed in attribution studies by the same authors.

Focusing on a statistical testing theory, and using an assumption of Gaussian distribution for ψ_{n+1}, that is quite usual, other interpretations of this result can be useful. It can be shown first that, among the T_f family, T_{f_o} is the better test, in the sense that it is the most powerful. The "optimal fingerprint" detection test can also be interpreted as the likelihood ratio test (Hasselmann 1997).

2.2 Classical adaptation of the optimal fingerprint

In practice, in the case of climate studies, the covariance matrix C is not known. In such a case, and assuming that data are available for estimating C, it is much more difficult to find one optimal test. Consequently, authors have searched for an approximation of the optimal fingerprint test.

A natural idea, to adapt the Eq. (7) when C is not known might be to use the empirical covariance matrix deduced from the $(\psi_i)_{1\leq i\leq n}$ sample:

$$\hat{C} = \frac{1}{n}\sum_{i=1}^{n} \psi_i \psi_i'. \tag{8}$$

However, substituting directly the estimate $\hat{C}$ in Eq. (7) yield to a rather bad test. This test even is not defined when $p > n$, because $\hat{C}$ is not invertible. For this reason, the more frequently used adaptation of optimal fingerprints consists of applying a Moore-Penrose pseudo-inversion to a truncation of $\hat{C}$.

This adaptation can be introduced as follows. Let Σ be a $p \times p$ symmetric positive matrix. Σ is diagonalisable in an orthonormal basis, that can be written $\Sigma = P'\,\Lambda\,P$, where P denotes the changing basis matrix, and Λ the diagonal matrix of Σ-eigenvalues: (λ_1, ..., λ_k, 0, ..., 0), denoting by k the rank of Σ. The Moore-Penrose pseudo-inverse of the q-truncation of Σ, what will be denoted by Σ_q^+,can be defined, for $q \leq k$:

$$\Sigma_q^+ = P'\mathrm{diag}\left(\frac{1}{\lambda_1},\ldots,\frac{1}{\lambda_q},0,\ldots,0\right)P, \tag{9}$$

where diag(v) denotes the diagonal matrix of diagonal v. Note that Σ_k^+ is simply the Moore-Penrose pseudo-inverse of Σ, without truncation.

With this definition, using a pseudo-inverted truncation of $\hat{C}$ is equivalent to making a projection on the q first eigenvectors of $\hat{C}$ (EOFs) and optimizing the fingerprint in the reduced space of dimension q. Above all, this solution can be adopted to prevent estimation problems which can occur for the smallest eigenvalues of the empirical covariance. Indeed, the use of empirical covariance introduces a bias in the estimation of eigenvalues: the biggest eigenvalues are over-estimated, whereas the smallest ones are underestimated.

The pseudo-inverted truncation of $\hat{C}$, that will be noted $\hat{C}_q^+$, provides a possible adaptation of the optimal fingerprint test T_{f_o}, that is $T_{\hat{C}_q^+ g}$. This test is based on the detection variable $d_{\hat{C}_q^+ g}$:

$$d_{\hat{C}_q^+ g} = \langle \psi_{n+1}, \hat{C}_q^+ g \rangle. \tag{10}$$

In order to use the test $T_{\hat{C}_q^+ g}$, some methods have been developed in the literature to select q and apply a test procedure, in particular in Hegerl et al. 1996, and Allen and Tett 1999. However, the choice is generally not discussed in terms of optimality. We try to focus on this point in Sect. 3.4.

Another test, T_g, usually referred to as the guess pattern fingerprint (GPF) test, and introduced by Hegerl et al. (1996), will be used for comparison. Although it is not exactly an adaptation of T_{f_o}, it is worth using it as a basic reference. As the notation T_g suggests it, this test is based on the variable

$$d_g = \langle \psi_{n+1}, g \rangle. \tag{11}$$

It is then a quite intuitive test, because a potential change of the expectation in the g direction is searched by making a projection on g. It also can be seen as the optimal test when $C = I_p$ (I_p being the $p \times p$ identity matrix).

2.3 A new method: the regularized optimal fingerprint

We introduce here a new adaptation of the optimal fingerprints, based on the use of the Ledoit regularized estimate $\hat{C}_I$, introduced in Ledoit and Wolf (2004), and technically presented in the Appendix 1. Using $\hat{C}_I$ for estimating C, the optimal fingerprint $C^{-1}g$ can be approximated by $\hat{C}_I^{-1}g$, hereafter referred to as the ROF (regularized optimal fingerprint). A corresponding test $T_{\hat{C}_I^{-1}g}$ (ROF test) can be proposed, based on the detection variable:

$$d_{\hat{C}_I^{-1}g} = \langle \psi_{n+1}, \hat{C}_I^{-1} g \rangle. \tag{12}$$

The basic idea of this original adaptation is to use a regularization technique, by searching for a suitable covariance matrix estimate of the form:

$$\gamma\hat{C} + \rho I_p, \tag{13}$$

where I_p is the $p \times p$ identity matrix, γ and ρ being real numbers. Several arguments can justify the use of this kind of estimate.

Firstly, the use of such a regularization technique has some simple and qualitative justifications. In a "large dimension" framework, that is when n and p are close, the weakest eigenvalues of C are both strongly underestimated in $\hat{C}$, and very affected by the addition of the term ρI_p. This term, by increasing these estimated eigenvalues, decreases their weight after inversion and then in the detection algorithm. That is why this type of method is called "regularization", and provides a more stable algorithm. On the contrary, using a pseudo-inverted truncation of $\hat{C}$, by setting to 0 the dominating terms of C^{-1}, cannot provide a suitable estimate of C^{-1}. Although $\hat{C}$ or its q-truncation $\hat{C}_q$, give generally an acceptable approximation of C, the inverted or pseudo-inverted matrices C^{-1}, $\hat{C}^{-1}$, and $\hat{C}_q^+$ are usually very different, as are the directions $C^{-1}\,g$, $\hat{C}^{-1}g$ and $\hat{C}_q^+ g$, that may be used for detection.

Secondly, the use of a covariance estimate of the form (13) can be interpreted as a way to make a balance between the detection variables $d_{\hat{C}^{-1}g}$ and d_g.

Thirdly, this method can be justified in a regression framework. The links between the optimal fingerprint method and a regression of the observation vector ψ under the "guess-pattern" g has already been established (Allen and Tett 1999; Allen and Stott 2003). The basic idea is to estimate by $\hat{\mu}$ the amplitude coefficient μ such as $\mathrm{E}(\psi) = \mu g$, and to compute a confidence interval for μ. When 0 doesn't belong to this confidence interval, the hypothesis H_0: $\mu = 0$ is rejected. Within this framework, the addition of the ρI_p term to the empirical estimate exactly amounts to use a ridge-regression technique, well-known in statistics. The more classical justification of it is that it allows the decrease of the root mean square error of the estimate. The use of a ridge regression is also justified in a Bayesian framework.

The main difficulty in using this regularization technique is to find relevant estimators of the parameters γ and ρ. Various methods might be considered. The method selected in this paper is taken from Ledoit and Wolf (2004). The main concepts are reviewed in the Appendix 1, and lead to the Ledoit regularized estimate $\hat{C}_I$. Basically, the choice of this estimate can be justified by the better properties of $\hat{C}_I$ as an estimator of C. The corresponding estimate $\hat{C}_I^{-1}$ of C^{-1} has also better properties than $\hat{C}_q^+$ (see above).

3 Evaluating the methods

3.1 What kind of evaluation?

The final goal of this paper is to adapt the "optimal fingerprint" test T_{f_o}, for leading to an "efficient" test procedure when the covariance matrix C is not assumed to be known. As mentioned previously, in such a case, it is much more difficult to find one optimal test, and we will only compare the three adaptations of Sect. 2: $T_{\hat{C}_q^+ g}$, $T_{\hat{C}_I^{-1} g}$, and T_g.

Some assumptions are made for this part of the study. First, the $(\psi_i)_{1 \le i \le n}$, are assumed to be independent, centered, and normally distributed, with mean 0 and covariance C. Second, ψ_{n+1} is assumed to be independent of the $(\psi_i)_{1 \le i \le n}$, and to have a $N(\mu g, C)$ distribution, where the vector g is known, unlike the coefficient μ. As in Sect. 2, this coefficient μ is the tested parameter. Third, we assume that n and p are of the same order, that is an unfavorable case in which few data are available for estimating C (relatively to the size of C). This framework is called "large dimension" or "general asymptotics", when n and p go to the infinity together, as in Ledoit and Wolf (2004). If one wants to estimate the covariance matrix C with observations, this is a rather reasonable assumption.

While searching for an "efficient" method, we will focus first on the power of the statistical tests. Indeed, for a statistical test of a hypothesis H_0 against a defined alternative hypothesis H_1, the power is the criterion that allows to measure the efficiency of the test. Moreover, it has been mentioned that the T_{f_o}, is optimal among the T_f family in the sense that it is the most powerful. So we will search for a good approximation by searching for the most powerful adaptation.

An important difficulty, when C is assumed to be unknown, is to control the level of the proposed tests. Indeed, the level being the probability to reject H_0 whereas H_0 is true, the control of the level requires to know the distribution of the test variable d under H_0. In our case, the distribution of d under H_0 depends on C and as a consequence is difficult to compute.

This can be illustrated as follows. Using the normality assumption for the tested observation ψ_{n+1}, the distribution of the optimal fingerprint test variable d_{f_o} can be written, under H_0:

$$d_{f_o} = \langle \psi_{n+1}, f_o \rangle \sim_{H_0} \mathrm{N}(0, \sqrt{f_o' C f_o}). \tag{14}$$

The covariance C is used twice in Eq. (14): first, for computing the optimal fingerprint f_o, and second, for determining the threshold and the p-value of the test.

In our case, the distribution under H_0 has to be computed or approximated using only the observations $(\psi_i)_{1 \le i \le n}$, as well as the estimate of the optimal fingerprint vector f_o. Making errors while computing this distribution can lead to a test which hasn't a nominal level, that is to say for which the probability to wrongly reject H_0 is smaller than the expected value (the conservative test) or greater than the expected value (the permissive test). This directly impacts the power of the tests. For example, the power of a permissive test is artificially increased, the threshold being different from what it should be. Tests that do not have a nominal level cannot be compared in terms of power, and are usually dangerous to use (unless they are known to be conservative).

Finally, the ability of a test to provide a correct p-value and to have a nominal level needs to be evaluated too. Note that this feature will be called accuracy of the test.

In the following we will separate the study into three steps. Firstly, we will search for an efficient estimation of the optimal fingerprint f_o. This will be done by comparing the power of the test in the ideal case where they have a nominal level. Secondly, we will evaluate the accuracy of the ROF method. For this purpose, a bootstrap is performed in order to estimate the H_0 distribution, and then it is verified that this procedure leads to an exact test, that is to say a test having a nominal level. Thirdly, we will study specifically some aspects of the classical adaptation of the

optimal fingerprints $T_{\hat{C}_q^+ g}$, and explain why this test is not used in Sect. 4.

3.2 Comparing the efficiency

The goal of this section is to compare the efficiency of the three tests presented: $T_{\hat{C}_q^+ g}$, $T_{\hat{C}_I^{-1} g}$, and T_g.

This comparison has been mainly carried out by performing simulations, in order to compute empirically the power of these tests. As mentioned in Eq. (1), the power is the natural measure of efficiency, as long as the compared tests have a nominal level. Therefore, in those simulations, the covariance matrix C is still assumed to be known, but only for computing the distribution under H_0 (and then, the "correct" threshold and p-value). The optimal fingerprint vector, for its part, is estimated from the $(\psi_i)_{1 \leq i \leq n}$. Note that all the simulations were performed in the case $n = p$, that clearly increases the errors due to the estimation of C. The only two parameters used for those simulations are the "true" covariance matrix C and the "true" vector g. The results may depend on the chosen values.

Firstly, some simulations have been performed, using values of C and g arbitrarily chosen. This type of simulation shows that for many cases, the power of $T_{\hat{C}_q^+ g}$ is weaker than the T_g one, whatever the value of q (not shown). The power of T_g, for its part, is weaker than the one of $T_{\hat{C}_I^{-1} g}$, unless C and I_p are very close.

Secondly, in order to evaluate the tests' properties when dealing with climate, simulations have been performed using values of C and g more consistent with the real cases. Concerning g, we used the expected climate change vector g taken from a climate model, as in Sect. 4. This allows to compute the power for the hypothesis H_1 really tested. The choice of a "true" covariance matrix C close to the real one is more difficult. We then compute the simulations for three different plausible values, deduced from observed data, that are spatially centered (see Sect. 4 for details about data and spatial centering). We used the regularized covariance matrix $\hat{C}_I$ (C_1), the empirical covariance matrix $\hat{C}$ (C_2), and a spatial covariance matrix (C_3), with entries of the form:

$$\mathrm{Cov}(t_i, t_j) = \mu e^{-\lambda d(s_i, s_j)}, \tag{15}$$

t_i and t_j denoting the observation at the stations s_i and s_j, and d being the distance between the stations s_i and s_j.

Starting from either C_1, C_2 or C_3, samples $(\psi_i)_{1 \leq i \leq n+1}$, are simulated, assuming that the H_1 hypothesis is true (a positive value is chosen for the coefficient μ). The compared tests T_g, $T_{\hat{C}_I^{-1} g}$, and $T_{\hat{C}_q^+ g}$ are then applied to the sample $(\psi_i)_{1 \leq i \leq n+1}$, and the empirical powers are computed.

On the one hand, the first case C_1 must be considered as the most significant, because of two reasons : the Ledoit regularized estimate is more precise, and, in this case, the observed sample $(\psi_i)_{1 \leq i \leq n}$, is really a typical realization assuming that the Ledoit estimate is the "true" covariance. On the other hand, it is useful to study whether the results remain qualitatively the same, even if the "true" covariance matrix is weakly regular (the two other cases).

The Fig. 1 shows such a comparison for values of g and C taken from a summer daily minimum temperatures dataset covering France. It can be seen that the ROF method is more powerful in all cases; the GPF method leads sometimes to a power higher than the one of the pseudo-inverted truncation method, depending on the value of q and on the supposed "true" covariance matrix used. The results for other variables (eg daily minimum or maximum temperatures, and different seasons) are qualitatively the same (not shown). Note that equivalent simulations have been performed with a higher value of n, and gave qualitatively the same results. This allows in particular the use of the ROF method even to estimate C from pseudo-observation taken from a control run.

This result provides the main reason to prefer the ROF method to the two other ones. It is rather reinforced by a theoretical power study. Indeed, instead of computing the empirical power of $T_{\hat{C}_q^+ g}$, one can wonder what is the effect of using a pseudo-inverted truncation, by studying the power of $T_{C_q^+ g}$, and comparing it with the power of T_g. The estimation errors due to the estimation of C are then ignored. This theoretical comparison is not detailed for briefness, but it shows that even without estimation errors, $T_{C_q^+ g}$ has not necessarily better properties than T_g, unless $q = p$. It also shows that the power of $T_{C_q^+ g}$ increases with q. This result allows an interpretation of the behavior of the empirical power of $T_{\hat{C}_q^+ g}$. When q is small, the estimation errors don't matter, and the $T_{\hat{C}_q^+ g}$ power increases with q as in $T_{C_q^+ g}$. When q is large, the estimation errors, which are strongest for the smallest eigenvalues, are dominating and the power of $T_{\hat{C}_q^+ g}$ decreases with q.

3.3 Bootstrap and accuracy of the ROF method

In this Section, we describe a method for computing the threshold of the test (and more generally the p-value), starting from the sample $(\psi_i)_{1 \leq i \leq n}$,and we check the accuracy of the resulting test. Note that this task is done only for the ROF test, and we will denote by $\hat{f}_o$ the ROF:

$$\hat{f}_o = \hat{C}_I^{-1} g. \tag{16}$$

In order to compute the threshold and the p-value of the test, it is necessary to determine the distribution of the variable $d_{\hat{f}_o}$ under the hypothesis H_0. This is not a trivial problem. Indeed, only the distribution of $d_{\hat{f}_o}$ conditionally to the $(\psi_i)_{1 \leq i \leq n}$ is known:

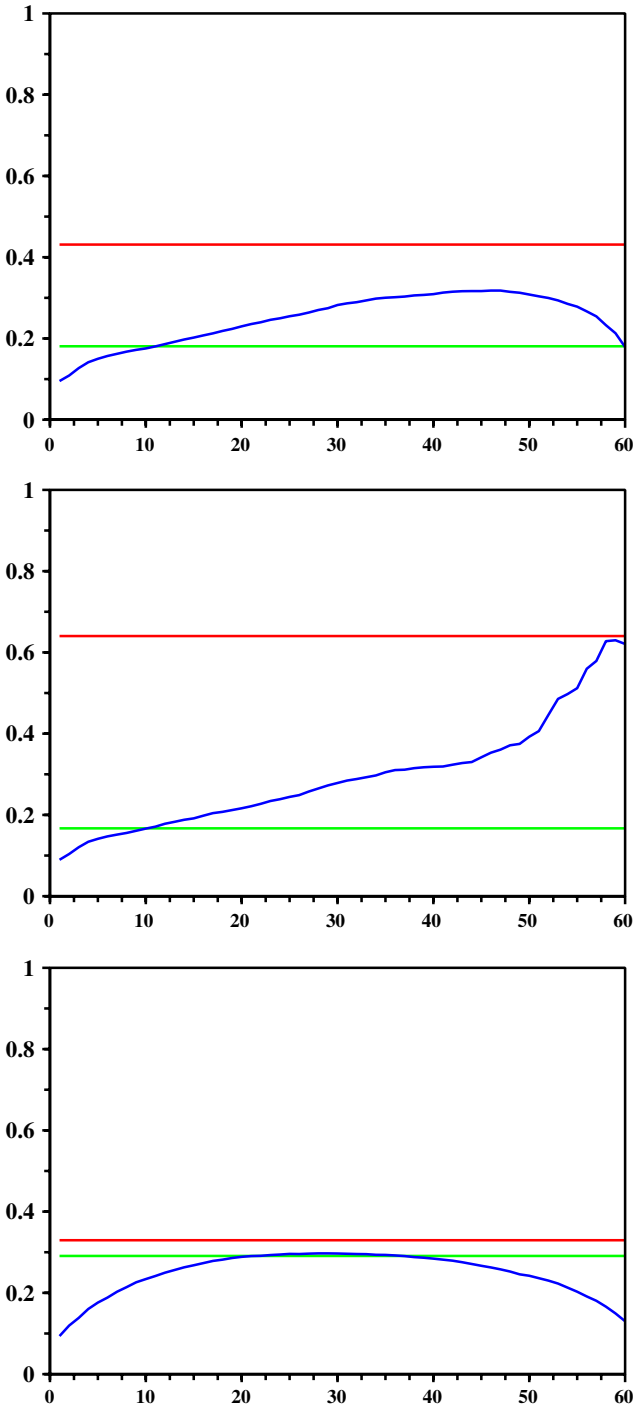

Fig. 1 Power study: Comparison of the power of the tests T_g (GPF test, *green line*), $T_{\hat{C}_I^{-1}g}$ (ROF test, *red line*), and $T_{\hat{C}_q^{+}g}$ (*blue curve*), as a function of q. Three covariance estimates are used as "true" C: the Ledoit regularized estimate $\hat{C}_I$ (*top figure*), the empirical estimate (*middle figure*), and the spatial covariance estimate (*bottom figure*); see text

$$d_{\hat{f}_o}|(\psi_i)_{1\le i\le n} \sim_{H_0} N(0,\hat{f}_o{}'C\hat{f}_o). \tag{17}$$

Moreover, the covariance C being still unknown, this formula is worthless, and it is necessary to substitute an estimate of the matrix. $\hat{C}_I$ is the main candidate, but two problems have to be taken into account. First, the use of an estimate, instead of the true value, carries some additional errors that lead to a different distribution (a classical example is the Student distribution). Second, the estimate $\hat{C}_I$ is also used in $\hat{f}_o$, and the dependence between both estimates may bias the results. This was pointed out by Allen and Tett (1999), who proposed to split the data into to two independent samples.

We here used an alternative strategy, namely a bootstrap procedure, that approximates the distribution of the normalized detection variable:

$$\delta = \frac{d_{\hat{f}_o}}{\sqrt{\hat{f}_o{}'\hat{C}_I\hat{f}_o}}, \tag{18}$$

as it can be made for a Student variable. In this way, the dependence between $\hat{C}_I$ and $\hat{f}_o$ is explicitly taken into account in order to avoid bias. Note that this (unconditioned) distribution only depends on C and g, and will be denoted by $\mathcal{D}_{C,g}$. The basic idea of the bootstrap is to estimate the distribution $\mathcal{D}_{C,g}$ by $\mathcal{D}_{\hat{C}_I,g}$, that can be simulated by Monte-Carlo technique. The details of this computation procedure are given in Appendix 2.

Let $\mathbb{T}_{\hat{C}_I^{-1}g}$ be the ROF test using this bootstrap procedure for evaluating both threshold and p-value.

A validation step is then necessary to demonstrate the accuracy of $\mathbb{T}_{\hat{C}_I^{-1}g}$ and to justify that the use of $\mathcal{D}_{\hat{C}_I,g}$ instead of $\mathcal{D}_{C,g}$ is acceptable for computing the p-value of the test. In fact, there is no absolute reason for justifying this approximation. It has been shown that $\hat{C}_I$ is a relatively good estimate of C, but this is not a conclusive argument.

This validation has been achieved by simulating the whole test procedure, starting from a pseudo sample whose covariance is known. All the details about the implementation of the validation procedure are given in Appendix 2. It can be noted that this procedure requires a starting covariance matrix C (similarly to the simulations performed in 2), and then focuses on the level of $\mathbb{T}_{\hat{C}_I^{-1}g}$.

The validation has been first applied to some simple starting matrices, and gave the required results: the level of $\mathbb{T}_{\hat{C}_I^{-1}g}$ is shown to be close to the nominal value, and its power is still greater than that of T_g. Finally, in order to validate the method for a starting covariance C next to that of the climate vector, this procedure has been applied to the matrix $\hat{C}_I$ estimated from real observations, with the same success.

3.4 Specific study of the classical approach

Although the power study presented in Sect. 3.2 could be thought as sufficient an argument to prefer the ROF method, we want here to discuss some characteristics of the

classical approach, particularly those due to the choice of the parameter q .

As mentioned while introducing the test $T_{\hat{C}_q^+ g}$, the choice of q is a first difficulty when using this approach, and some methods have been proposed to select a value. Hegerl et al. (1996) chose to study the spatial correlation between the "guess-pattern" g and the fingerprints $f_q = \tilde{C}_q^+ g$, where $\tilde{C}$ denotes an estimate of the covariance matrix C carried from a climate model simulation. The dependence of the result on q is mentioned as not changing the result for neighboring values.

Allen and Tett (1999), after having presented the optimal fingerprint formalism as regression, and some problem due to the estimation of the covariance matrix, proposed a different treatment. In particular, they proposed to use a consistency test to check whether the estimate provided by some control simulations is consistent with the observed residuals. The values of q for which this consistency hypothesis is significantly rejected are not studied. In this way, a set of values of q is used, and an answer can be given to the detection question depending on the agreement of the results of the different tests computed for each selected q.

In both methods, some values are proposed for q, but the choice is not discussed in terms of optimality. One can wonder whether the power study highlights what could potentially be a good choice of q. We have tried in Sect. 2 to focus on this point, by studying the power of the tests. Figure 1 suggests that finding the best value for q is actually a difficult task. In particular, it can be seen by comparing the three graphics that a small difference on the starting covariance matrix used for simulations should lead to a different choice.

After having chosen the value of q, which is a difficult task, the construction of a useful test using the direction $\hat{C}_q^+ g$ should require an estimation of the threshold (as in Eq. (3)). But as the empirical estimate of the covariance matrix is not well-conditioned in large dimension, it can be thought that yielding a nominal level test is very difficult. In particular, the threshold estimation would have been more difficult than the threshold estimation used in Sect. 3.3.

One can also wonder about the q-stability of the results when using the test $T_{\hat{C}_q^+ g}$. Are the results actually sensitive to the choice of q? An important dependence on q could make the interpretation much more difficult, although this wouldn't be the most conclusive argument to prefer the ROF test.

We have tried to discuss this problem theoretically, focusing first on the tests $T_{C_q^+ g}$, and then on the family $T_{\hat{C}_q^+ g}$. This part of the study is detailed in Appendix 3. It shows that the stability is generally not ensured, even without estimation problems (case $T_{C_q^+ g}$).

Another way to discuss this sensitivity may be to look at the discrepancies of the results for several values of the parameter q. We have done such a comparison for two real-case applications studied using the ROF method in Sect. 4. The results of these comparisons are represented in Fig. 2, and show that the choice of q clearly impacts the results of the test.

However, some care should be taken when interpreting Fig. 2. The p-value is actually the result provided by a statistical test, and is the criterion that allows to reject, or not, the "null" hypothesis. So the evaluation of the discrepancies between several tests should have been done by comparing their p-values. This is quite difficult in our case, because we haven't proposed a way to compute the correct p-values of the tests $T_{\hat{C}_q^+ g}$. Consequently, Fig. 2 represents the time evolution of the normalized detection variable δ introduced in Eq. (18), for several values of q, and the black dashed line shows what might have been the threshold of the tests using the basic assumption that the H_0-distribution of δ is a standard normal distribution. Although using this threshold may lead to non exact $T_{\hat{C}_q^+ g}$

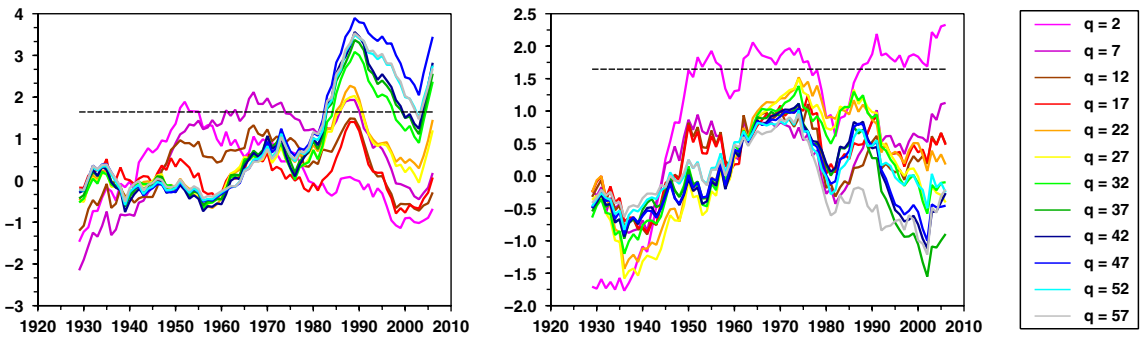

Fig. 2 q-sensitivity: Comparison of the the time evolution of the normalized detection variable δ using the classical "optimal fingerprint" adaptation $T_{\hat{C}_q^+ g}$, for different values of the truncation parameter q. The comparison is done for autumn (on the left) and summer (on the right) daily maximum temperatures, following the procedure described in Sect. 4. The *black dashed line* shows an hypothetical threshold assuming that the H_0-distribution is a standard normal distribution

tests, Fig. 2 gives a qualitative illustration of the sensitivity to the choice of q.

In the case of autumn daily maximum temperatures, it can be seen that the results at the end of the period are scattered on both side of the hypothetical threshold. In such a case, it would be difficult to conclude whether the classical approach has detected a change. It would have been even more difficult to compute a p-value, due to these discrepancies. In the case of summer daily maximum temperatures, it can be seen that the sign of the normalized detection variable may depend on the choice q. It can also be noted that some very small values of q may reject the hypothesis H_0 whereas higher values do not.

It can be noted that many detection studies, performed at global or continental scales using many values for the truncation parameter q, haven't highlighted so large discrepancies. In our case, such a sensitivity may be due to the sub-regional scale of the study (we used a dataset covering France), or to the spatial centering (see Sect. 4), that removes the main part of the signal.

Finally, the classical approach of optimal fingerprint is difficult to implement in our case: the choice of q can be debated all the more so as it really impacts the results, and the computation of a correct p-value for this test is not guaranteed. This is the reason why we have preferred to use the GPF test instead of this method as reference in the following section, devoted to the application.

4 Application

4.1 Data

In order to perform a detection study based on the presented ROF methodology, two types of data are used: observations, and climate simulations for the 21st century.

Concerning the observations, a detection study requires data of high quality, covering a period as long as possible, and with a high spatial density. Such a dataset, covering France, has been produced at Météo France using an adapted penalized log-likelihood procedure (Caussinus and Mestre 2004). This data covers the 1900–2006 period, with about 60 stations distributed over the country.

Regarding the estimate of the climate change signal ("guess pattern"), a set of eight simulations is averaged over the 2070–2099 period. The simulations are performed with the ARPEGE-Climat regional climate model with variable horizontal resolution (Gibelin and Déqué 2003). This model is forced with sea surface temperatures and ice field extensions taken from climate change simulations, corresponding to different IPCC scenarios (A2 and B2) and to different global coupled atmosphere ocean general circulation models. In order to represent the present climate, a set of three simulations performed with the same model over the 1960–1999 period, is averaged. The guess-pattern is then deduced as the difference between temperatures averaged over the 2070–2099 period, and the ones averaged over the 1960–1999 period.

Among the set of future climate simulations, two simulations are transient scenarios covering the 1960–2099 period (one A2 scenario and one B2 scenario). These two simulations will be used as an illustration of the ROF method presented here.

4.2 Implementation of the method

The precise implementation of the method is described here, starting from the two main variables used: the observations dataset Ψ, and the guess-pattern g.

The set Ψ can be seen as a sample, or a $N \times p$ matrix $(\psi_{t,s})$, where t represents the time (usually 1 year), and s the space (one station). This set has to be divided into two subsets Ψ^L and Ψ^T, that are respectively the learning sample, and the tested sample. The first one, Ψ^L, is dedicated to estimating the covariance matrix, and is usually constituted of the N_L first years of Ψ. The second one, Ψ^T groups the data that we want to test. The choice of the separating year between Ψ^T and Ψ^L is quite arbitrary, and allows to partially modulate the size of these subsets. Some care has to be taken when using the ROF algorithm when the number of years in Ψ^L is rather small. When presenting the method in Sect. 2.3, the subsets Ψ^L and Ψ^T were, respectively, the family $(\psi_i)_{1 \leq i \leq n}$, and the one-year vector (ψ_{n+1}). In the applications presented, the subset Ψ^T will usually contain more than one year, in particular for studying the time evolution of the detection variable. Note that strictly speaking, those two sets Ψ^L and Ψ^T could have been taken from different datasets, for example for testing observations, using a covariance matrix estimate deduced from a climate model run.

The vector g, taken from a climate model scenario, has to be represented on the same points as the observations, that are the p observation stations. This operation can be conducted with a basic interpolation procedure.

However, the p-dimensional space thus used for representing Ψ and g, could be inappropriate for the detection, for example due to a very irregular spatial distribution of the observation stations. In order to improve the regularity of this spatial distribution, a hierarchical clustering algorithm is applied to the observation stations (see for example Mardia et al. 1979). Such a procedure allows to reduce the spatial dimension from p to p', with $p' \leq p$. The reduced dimension p' is then chosen to provide an exact test, and maximize the power. Note that if p is close to n, or smaller, $p' = p$ is often selected. This confirms the good behavior of the test $\mathbb{T}_{\hat{C}_I^{-1}g}$ in large dimension, as mentioned in Sect. 3.

In contrast, if the spatial dimension p is higher than the size of the learning sample N_L, a reduction of the dimension could be required for ensuring the accuracy of the ROF test.

An important characteristic of this study is its focus on the regional scale and, as a consequence, on the spatial distribution of the studied variables. Indeed, detecting a change on the variable ψ can be the consequence of an uniform global change, without any regional specificity. To avoid the detection of such a global signal, each observation ψ is decomposed into the sum of a spatial mean $\overline{\psi}$ and a centered variable $\tilde{\psi}$; and it is chosen to filter out the detection variable part due to the evolution of the spatial mean $\overline{\psi}$, basing the detection only on $\tilde{\psi}$.

A preliminary temporal treatment is usually applied in order to deal with climate. Indeed, as the notations of the previous section suggest, a detection test can be applied to the climatic observations of a unique year. But rejecting the hypothesis H_0 for 1 year doesn't allow to conclude on climate change. Climate denoting the characteristics of the weather on a long time period, the detection test must be applied to a more representative variable.

The first detection studies were based on linear trends over 15, 20 or 30-year periods, with some preferences for the last one (Hegerl et al. 1996). Here the use of moving averages has been preferred, as these are less sensitive to the interannual variability. The tested vectors used are then the difference between a moving average of N_T years (N_T can be chosen), and a reference average, that is the average under the learning period (the length of which is N_L). Note that this reference average is independent of the covariance estimate $\hat{C}_I$. Thus, if the N_T years averaged are taken in Ψ^T, those tested vectors are independent of $\hat{C}_I$.

Both transformations of spatial centering and temporal treatment can be summarized by the following formulas, denoting by $(\psi_{t,s})$ the original variables, and $(\phi_{t,s})$ the transformed ones:

$$\psi_{t,.} = \frac{1}{p}\sum_{s=1}^{p}\psi_{t,s}, \tag{19}$$

$$\phi_{t,s} = \frac{1}{N_T}\sum_{h=1}^{N_T}(\psi_{t+h,s} - \psi_{t+h,.}) - \frac{1}{N_L}\sum_{h=1}^{N_L}(\psi_{h,s} - \psi_{h,.}). \tag{20}$$

Thus, the transformed vectors $\phi_t = (\phi_{t,1}, ..., \phi_{t,s}, ..., \phi_{t,p})$ are deduced from the original vectors $\psi_t = (\psi_{t,1}, ..., \psi_{t,s}, ..., \psi_{t,p})$ via a projection and an average operations.

Consequently, the covariance matrix of the vectors ϕ_t can be deduced from the original covariance matrix of the vectors ψ_t using some simple formulas. The hypothesis of independence between ψ_t and ψ_{t+1} is used here. It is important to note that this hypothesis only impacts on the threshold computation, and not on the temporal evolution of the detection variable. Furthermore, if the dependence is weak, the impact on the threshold will be weak. Finally, we compute the covariance matrix of the (ψ_t) first, then we deduce the one of the (ϕ_t), and we carry out the test.

4.3 Results: ideal case

In order to illustrate the presented ROF methodology, and to highlight its efficiency, the method is first applied to pseudo-observations derived from transient regional climate change B2 scenarios covering the 1960–2099 period. This is a kind of idealized experiment, because of two reasons. First, it allows to be placed under the hypothesis H_1, the existence of a climate change being sure. Second, it avoids the difficulties for representing the climate change vector g. Indeed, the g used can be taken from the tested simulation itself. On the contrary, to apply the algorithm to the observations with an inexact g, can impact the efficiency of the test. It has been verified that the use of a guess-pattern g taken from a different simulation performed with the same model doesn't have an impact on the results presented here, including using an A2 scenario (not shown).

The application to model data allows to verify that the power of the test proposed is higher from a quasi-experimental point of view. Indeed, a more powerful test will reject H_0 for a smaller amplitude coefficient μ, or equivalently, earlier during the 21st century.

This section is focused on one single temperature variable that is the summer daily minimum near surface temperature. Note that this variable has been chosen arbitrarily among the eight ones studied in the next section; but this choice doesn't impact the results, which are virtually the same ones for the other variables.

For this first application, the learning sample is chosen to be the first 40 years of the scenarios ($N_L = 40$), that is to say the 1960–1999 period. In a transient scenario, this period is clearly contaminated by the influence of climate change. But as mentioned previously, taking into account some climate change effect only leads to a more conservative test. The choice of 40 years for the covariance estimation is more or less arbitrary, but doesn't impact the results.

During this learning period, the covariance estimate $\hat{C}_I$ and the values taken by the tested climate vector aren't independent. As a consequence, the H_0-distribution of the detection variable is impacted, the threshold and the statistical significance have a different meaning and no conclusions can be drawn with respect to the detection issue. For this reason, the detection threshold will be represented only outside the learning period, and we take these cautions into account by representing the statistical significance with a dashed line (see Fig. 3).

In order to perform a statistical test as close as possible to the one performed in the next section, the spatial

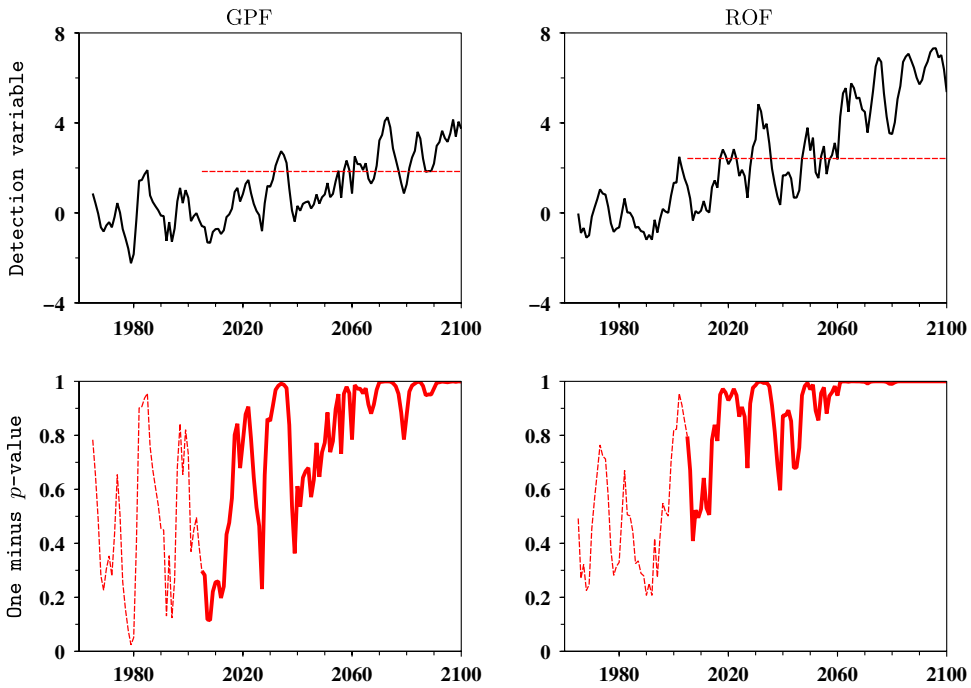

Fig. 3 Summer minimum temperatures in a transient B2 scenario: Comparison between the GPF test procedure and the ROF one for summer minimum temperatures taken from a transient B2 scenario from 1960 to 2099. The comparison is based on 5-year moving averages ($N_T = 5$), spatially centered, on 50 pseudo-stations over France. The *top figures* show the time evolution of the normalized detection variables (*black curve*), that is compared with the 95% confidence threshold (*red dashed line*), when it makes sense. The *bottom figures* represents the time evolution of the statistical significance ($1 - p$-value) of each test. The corresponding curves are dashed over the learning period

dimension of model grid, presenting $p = 220$ points over France, has been reduced to $p' = 50$ (that is approximately the value used in the application to observations). Note however that the ROF has been shown to be efficient when p and n are of the same order, but the case "p large relatively to n" is not so favorable. In particular, it is possible that the ROF test is not as accurate in such a case. Consequently, a reduction of the spatial dimension is needed in any case, due to the size of the learning sample chosen ($N_L = 40$).

It can be seen (Fig. 3) that the efficiency of the ROF technique is greater, because the detection variable exceeds the threshold more frequently, more rapidly. Note that in order to compare the efficiency of the tests, the date on which the detection variable goes beyond the threshold for the first time isn't a good criterion. In particular, the threshold can be exceeded even under the hypothesis H_0.

When the signal becomes strong enough three things happen: the statistical significance cannot go down to too small values, the detection variable "frequently" exceeds the threshold, and some values of this variable are well above the threshold. Looking at these different criteria on the case represented confirms the first impression, which is that the ROF test procedure is more efficient than the GPF one. Others variables (changing the season or studying daily maximum temperatures) give similar results.

It can be noted that for this experimental demonstration, the length N_T of the moving averages used (5 years) has been chosen arbitrarily, in order to highlight the differences between both tests. For example, the signal being strong over the 21st century, the use of 30-year moving averages over this period leads to a very significant detection, whatever the test used.

4.4 Results: observation dataset

In the case of observations, eight detection studies are performed, corresponding to the minimum and maximum near-surface temperatures, for each one of the four seasons. All studies are based on 30-year moving averages, spatially centered observations ($N_T = 30$). The learning sample consists of first 70 years, from 1900 to 1969 ($N_L = 70$). The choice of the size of the learning sample is quite arbitrary, because there is no evidence that would allow us to highlight a period less contaminated by climate change. But this choice is conservative just as in the application to the scenarios. The ROF method is applied in each case.

Some results are represented in Figs. 4 and 5. The same conventions as those of the previous section are used, for representing the 95% confidence level threshold, and the statistical significance. Over the learning period, that is prior to the year 2000 in this case, great care must be taken when interpreting the results. Indeed, the dependencies between the used $(\psi_i)_{1 \leq i \leq n}$ impact the distribution of the detection variable, and the statistical significance represented by the dashed line isn't exact.

Firstly, the strongest signal is observed on summer minimum temperatures. The p-value yields for the 1977–2006 period (that is compared with the mean over

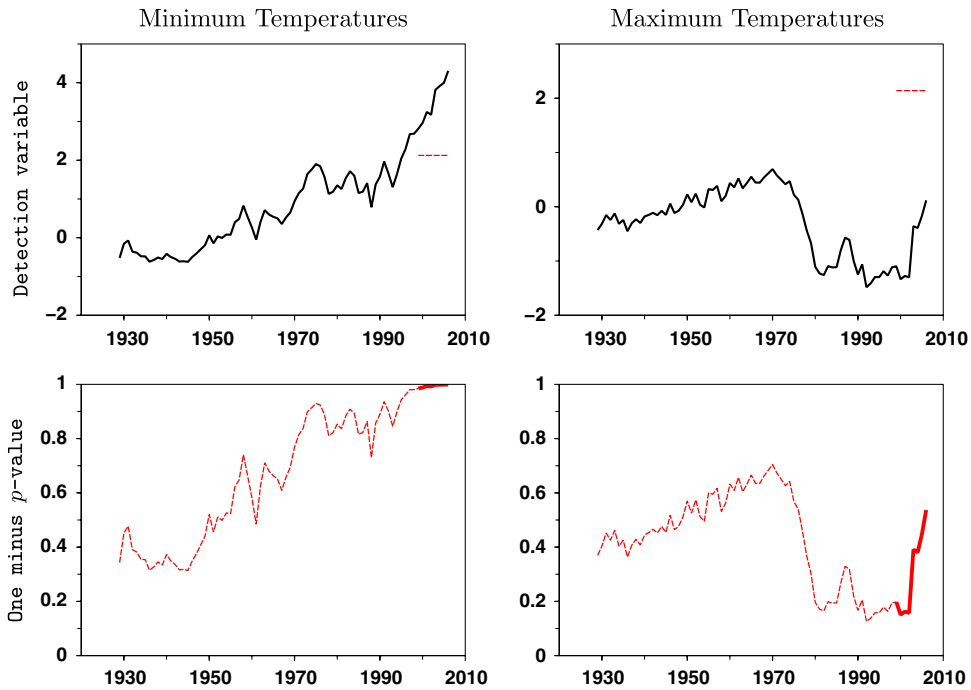

Fig. 4 Summer temperatures in observations. Results of the ROF methodology applied on observed daily minimum (*left*) and maximum (*right*) near-surface summer temperatures, from 1900 to 2006. The comparison is based on 30-year moving averages, spatially centered. The *top figures* show the time evolution of the normalized detection variables (*black curve*), that is compared with the 95% confidence threshold (*red dashed line*), over the tested sample. The *bottom figures* represent the time evolution of the statistical significance of each test. The corresponding curves are dashed over the learning period

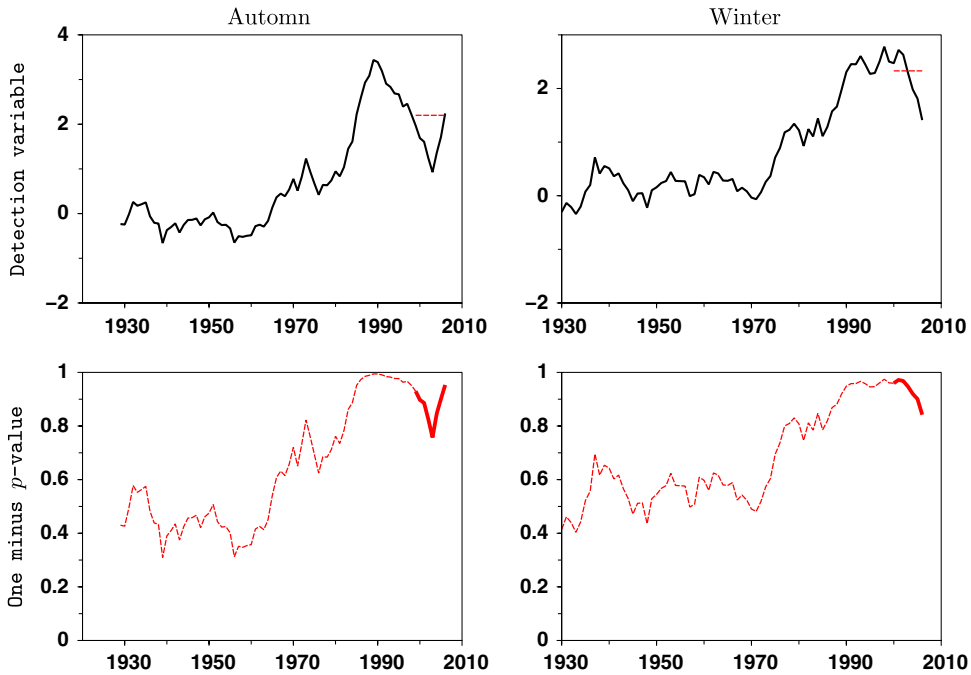

Fig. 5 Autumn and winter daily maximum temperatures in observations: results of the ROF methodology applied on observed autumn (*left*) and winter (*right*) daily maximum near-surface temperatures, from 1900 to 2006. The comparison is based on 30-year moving averages, spatially centered. The representation is the same than in Fig. 4

1900–1969) a value smaller than 10^{-3}. This result confirms and reinforces the findings of Spagnoli et al. (2002).

Secondly, two other variables present significant changes, although in a less significant way. The first one of these is the maximum autumn temperature, for which the final value of the normalized detection variable goes just beyond the threshold. Some higher values are observed, over more than one decade, for the last years of the learning period (these years being very slightly dependent on the observations used for covariance estimation). We then conclude that the detection is positive in this case. The second variable that present significant change is the winter maximum temperature. The last observed value of the associated nomrmalized detection variable is smaller than the threshold considered, but stays quite high. Moreover, significant values are observed some years before. For example, the statistical significance of the test yields to 97% in 2001 (testing the 1972–2001 average).

Table 1 Statistical significance of the ROF test: results of the ROF test applied to daily minimum/maximum temperatures and each season

Season	Min T	Max T
Summer	>0.99	0.55
Autumn	0.08	0.95
Winter	0.50	0.85
Spring	<0.01	0.78

The statistical significances ($1 - p$-values) given are obtained when testing the 1977–2006 mean, and correspond to the last value represented in Figs. 4 and 5

These two success, beyond illustrating the possibilities of the method, constitutes advances with regards to detecting climate change over France, relatively to previous studies (Spagnoli et al. 2002, Planton et al. 2005).

Thirdly, in the case of summer daily maximum temperatures, and for the other variables studied, the detection algorithm failed to detect a change. We do not represent the whole time evolution of the detection variable for each studied case, but the p-value provided by the ROF test at the end of the period are given in Table 1. It can be seen that autumn and spring daily minimum temperatures show interesting behaviour. In both cases, especially for the second one, the detection variable takes very small values, that are quite unlikely under the hypothesis H_0. The three last variables don't show a significant evolution. The explanation of such discrepancies would require further work, but it may be due to the weakness of the signal of change, or to an imperfect simulation of the pattern by the climate model, for some of the variables.

5 Conclusion

We have introduced a new adaptation of the "optimal fingerprint" statistical technique, based on the use of a well-conditioned covariance matrix estimate. This adaptation has been compared to the commonly used adaptation, in which the "optimal fingerprint" is computed in a reduced space corresponding to the first EOFs. The non-optimized "Guess Pattern Fingerprint" has also been used as a reference.

We have shown that our method is more efficient, in the sense that it yields a more powerful statistical test. Note that the comparison of the methods has been performed using a detection framework, and that the extension of the ROF methodology to the attribution problem could be a natural continuation of this work. Furthermore, to apply the "optimal fingerprint" in a reduced space requires to choose the dimension of this reduced space. Such a choice is difficult and can impact on the results. Our adaptation avoids this step, and thus is easier to implement.

Moreover, the use of regularization allows to base the covariance matrix estimation on a small-sized sample. In the case of climate change detection, the estimation of the internal climate variability can be based on observations, even if the number of available observations is reduced. Such a possibility has been used for the detection tests performed in this paper. This is an alternative to the use of an estimate based on long control simulations, and avoids the model imperfections when representing the covariance structure.

A last step is needed to achieve the implementation of a statistical detection test, that is to compute the threshold and the p-value of the test. Due to the small number of data available for estimation, we have chosen to compute this p-value and to estimate the covariance matrix on the same sample. The dependencies between these two problems are taken into account via a bootstrap procedure.

The application of the Regularized Optimal Fingerprint (ROF) method on climate data confirms the theoretical results. Firstly, the ROF has been applied on the ideal case of a climate scenario which insures that the alternative hypothesis H_1 is true. Then, it is shown from a quasi experimental point of view that the power of the ROF method is greater than the power of the GPF technique. Secondly, the ROF has been used to study a temperature dataset over France. Some previous results concerning climate change detection are reinforced, especially concerning the summer daily minimum temperatures, and some new results are highlighted. However, it is difficult to make a general conclusion about the detectability of climate change over France: among the set of eight temperature variables studied, the results show some discrepancies. Further works will be useful to answer this question, for example by using a multi-model approach.

Acknowledgments This work was supported by the European Commission Programme Energy, Environment and Sustainable Development under contract GOCE 036961 (CIRCE), and by the European Marie Curie network SEAMOCS. The authors are very grateful for the careful work done by two anonymous reviewers which has resulted in real improvements of the manuscript.

Appendix 1

Ledoit estimate $\hat{C}_I$

The main concepts of Ledoit and Wolf (2004) to lead to the Ledoit regularized estimate $\hat{C}_I$ are here reviewed.

The first question addressed is the following: given the empirical estimate of a $p \times p$ covariance matrix, $\hat{C}$, deduced from a $n \times p$ sample X, can a linear combination of the form $\tilde{C} = \gamma\hat{C} + \rho I_p$ be a more precise estimate of C

than $\hat{C}$? The answer is yes, in the sense of the mean square error:

$$\mathrm{E}\left(|\tilde{C} - C|_T^2\right), \tag{21}$$

for the classical norm:

$$|A|_T^2 = \frac{\mathrm{Tr}(\mathrm{A}'\mathrm{A})}{p}. \tag{22}$$

When C is known, the optimal values γ_o and ρ_o of γ and ρ are given by Ledoit and Wolf (2004):

$$\gamma_o = \frac{\alpha^2}{\delta^2}, \tag{23}$$

$$\rho_o = \frac{\beta^2 v}{\delta^2}, \tag{24}$$

where

$$v = \langle C, I_p \rangle_T = \frac{\mathrm{Tr}(\mathrm{C})}{p}, \tag{25}$$

$\langle .,. \rangle_T$ being the scalar product associated with the norm $|.|_T$, and:

$$\alpha^2 = |C - vI_p|_T^2, \tag{26}$$

$$\beta^2 = \mathrm{E}\left(|\hat{C} - C|_T^2\right), \tag{27}$$

$$\delta^2 = \mathrm{E}\left(|\hat{C} - vI_p|_T^2\right). \tag{28}$$

In the case when C is unknown, consistent estimators $\hat{\gamma}$ and $\hat{\rho}$ of those two optimal coefficients can be constructed, and they are shown to be convergent under general asymptotics hypothesis. Note that the "general asymptotics" framework, also called "large dimensional", is larger than the classical asymptotics framework for statistics, and it especially covers the case $n = p$, and $n,p \to \infty$. More precisely, estimates of the coefficients v,α,β and δ are first defined:

$$\hat{v} = (\hat{C}, I_p) = \frac{\mathrm{Tr}(\hat{C})}{p}, \tag{29}$$

$$\hat{\delta}^2 = |\hat{C} - \hat{v}I_p|_T^2, \tag{30}$$

$$\hat{\beta}^2 = \min\left(\hat{\delta}^2, \frac{1}{n^2}\sum_{i=1}^{n} |\psi_i\psi_i' - \hat{C}|_T^2\right), \tag{31}$$

$$\hat{\alpha}^2 = \hat{\delta}^2 - \hat{\beta}^2. \tag{32}$$

Then, the estimates of γ_o and ρ_o are deduced:

$$\hat{\gamma} = \frac{\hat{\alpha}^2}{\hat{\delta}^2}, \tag{33}$$

$$\hat{\rho} = \frac{\hat{\beta}^2\hat{v}}{\hat{\delta}^2}. \tag{34}$$

These estimates lead to a new estimate of the covariance matrix, namely the Ledoit regularized estimate:

$$\hat{C}_I = \hat{\gamma}\hat{C} + \hat{\rho}I_p. \tag{35}$$

Appendix 2

Bootstrap and validation implementation

Simulating $\mathcal{D}_{\hat{C}_I,g}$

The procedure implemented to simulate the distribution $\mathcal{D}_{\hat{C}_I,g}$ is presented here.

Step 0 From the observed $n \times p$ sample X, containing the observation vectors $(\psi_i)_{1 \le i \le n}$, the estimate $\hat{C}_I$ of C is computed.

Step 1 $\hat{C}_I$ is used to simulate a set of N samples $(X_j^*)_{1 \le j \le N}$, each one of size $n \times p$, with a Monte Carlo technique: similarly to the matrix X, the lines of a X_j^* matrix (equivalent to one of the ψ_i) are independent random vectors, the distribution of which is a $N(0, \hat{C}_I)$. The X_j^* are independent between them.

Step 2 From X_j^*, an estimate $\hat{C}_j^*$ of $\hat{C}_I$ is computed, using the estimation method presented in Appendix 1.

Furthermore, an estimate $\hat{f}_j^* = (\hat{C}_j^*)^{-1} g$ of $\hat{f}_o$ is computed.

Step 3 For each j, the conditional distribution $\mathcal{D}_{\hat{C}_I,g}$ given X_j^* is known:

$$\mathcal{D}_{\hat{C}_I,g}|X_j^* = N\left(0, \frac{\hat{f}_j^{*\prime}\hat{C}_I\hat{f}_j^*}{\hat{f}_j^{*\prime}\hat{C}_j^*\hat{f}_j^*}\right), \tag{36}$$

so that the conditional probability density function associated can be deduced.

Now, the $\mathcal{D}_{\hat{C}_I,g}$ distribution being the mixture of the conditional distributions, its density is approximated by computing the mean of the N conditional densities simulated.

Validation

The question addressed here is: in what measure the $\mathcal{D}_{\hat{C}_I,g}$ - distribution can be used instead of the $\mathcal{D}_{C,g}$ one for computing the p-value of the test?

Using the distribution $\mathcal{D}_{\hat{C}_I,g}$, we can proposed a ROF test totally defined by the sample $(\psi_i)_{1 \le i \le n}$. This test, denoted $\mathbb{T}_{\hat{C}_I^{-1}g}$ is characterized by its rejection region $\mathbb{W}_{\hat{C}_I^{-1}g}$:

$$\mathbb{W}_{\hat{C}_I^{-1}g} = \left\{ (\psi_i)_{1 \le i \le n+1}, \frac{\langle \psi_{n+1}, \hat{C}_I^{-1}g \rangle}{\sqrt{(\hat{C}_I^{-1}g)'\hat{C}_I(\hat{C}_I^{-1}g)}} \ge \mathcal{D}_{\hat{C}_I,g}^{(\alpha)} \right\}, \tag{37}$$

where $\mathcal{D}^{(\alpha)}_{\hat{C}_I,g}$ denotes the $(1-\alpha)$ quantile of the distribution $\mathcal{D}_{\hat{C}_I,g}$, estimated from the $(\psi_i)_{1\leq i\leq n}$ using the procedure presented in the previous paragraph. Note that we use here a specific notation $\mathbb{T}_{\hat{C}_I^{-1}g}$, because this test couldn't be defined by a rejection region of the form (6). So that $\mathbb{T}_{\hat{C}_I^{-1}g}$ isn't part of the T_f family.

The validation proposed here focuses on the test level. To verify that the $\mathbb{T}_{\hat{C}_I^{-1}g}$ level is nominal, it is equivalent to verify that:

$$\mathrm{P}_{H_0}(\mathbb{W}_{\hat{C}_I^{-1}g}) = \alpha. \tag{38}$$

The validation of this equality can be made similarly with the previous procedure.

Step 0 A "true" covariance matrix, C is fixed arbitrarily.

Step 1 C is used to simulate a set of N_0 samples $(X_k)_{1\leq k\leq N_0}$, similarly to the (X_j^*) of the previous paragraph: each (X_k) is of dimension $n\times p$, and its lines are independent random vectors, the distribution of which is a $N(0,C)$. Note that X_k is equivalent to the unique set of available observations X. The regularized covariance estimate $\hat{C}_I^{(k)}$ is computed, for each k.

Step 2 Following the previously presented simulation procedure, the distribution $\mathcal{D}_{\hat{C}_I^{(k)},g}$ is generated. The candidate threshold $\mathcal{D}^{(\alpha)}_{\hat{C}_I^{(k)},g}$ is deduced.

Step 3 The conditional distribution $\mathcal{D}_{C,g}$ given X_k being known, the level and the power conditionally to X_k can be deduced. Finally, the unconditioned level and power are approximated, using a relatively large number of samples N_0.

As mentioned in Sect. 3.3, this validation procedure has been successfully applied to different starting matrices C.

Appendix 3

Discussing the sensitivity to the choice of q

We first study the tests $T_{C_q^+g}$. We denote respectively by λ_i^2, g_i and ψ_i, the eigenvalues of C, and the coordinates of g and ψ in the C eigenvectors basis. The detection variable associated with $T_{C_q^+g}$ can be written:

$$d_{C_q^+g} = \langle \psi, C_q^+ g\rangle = \sum_{i=1}^{q} \frac{\psi_i g_i}{\lambda_i^2}. \tag{39}$$

Note that $\mathrm{Var}(\frac{\psi_1}{\lambda_1},\ldots,\frac{\psi_p}{\lambda_p}) = I_p$, that is to say that the $z_i = \frac{\psi_i}{\lambda_i}$ are random variables, independents and identically distributed, and, under the hypothesis H_0, with N(0,1) distribution. So,

$$d_{C_q^+g} = \sum_{i=1}^{q} z_i \frac{g_i}{\lambda_i}, \tag{40}$$

is a random walk, with weights $\frac{g_i}{\lambda_i}$.

This formula highlights the role of the family $(\frac{g_i}{\lambda_i})$, and shows that the results are generally "unstable" with respect to the choice of q, even when q is large. Indeed, some stability occurs only when the coefficients $(\frac{g_i}{\lambda_i})$ become "small", as q goes to p. In such a case, the tests $T_{C_q^+g}$ tend to $T_{C^{-1}g}$ (and they are close to each other), so the choice of q doesn't matter, as soon as q is large.

However, in the case of climate study, there isn't any evidence that this condition on the $(\frac{g_i}{\lambda_i})$ family is satisfied. Consequently, under the hypothesis H_0, the probability to reject H_0 for at least one of the considered tests can become important (clearly greater than the individual level α of each test). This illustrate the potential danger of the method and of the choice of q.

A similar study can be made for the tests $T_{\hat{C}_q^+g}$. Denoting respectively by $\hat{\lambda}_i^2$, $\hat{g}_i$ and $\hat{\psi}_i$ the eigenvalues of $\hat{C}$, and the coordinates of g and ψ in the $\hat{C}$ eigenvectors basis,

$$\hat{d}_{\hat{C}_q^+g} = \langle \psi, \hat{C}_q^+ g\rangle = \sum_{i=1}^{q} \frac{\hat{\psi}_i}{\hat{\lambda}_i}\frac{\hat{g}_i}{\hat{\lambda}_i}. \tag{41}$$

Here, the main differences with the previous case is that, first, the random variables $\frac{\hat{\psi}_i}{\hat{\lambda}_i}$ are not exactly independent and identically distributed, and, second, the $\hat{\lambda}_i$ family decreases more rapidly than the λ_i one, due to estimation errors. Consequently to this last point, the sensitivity to the choice of q is increased.

References

Allen M, Stott P (2003) Estimating signal amplitudes in optimal fingerprinting, part I: theory. Clim Dyn 21:477–491. doi: 10.1007/s00382-003-0313-9

Allen M, Tett S (1999) Cheking for model consistency in optimal fingerprinting. Clim Dyn 15(6):419–434

Barnett T, Pierce D, Schnur R (2001) Detection of anthropogenic climate change in the world's oceans. Science 292:270–274

Caussinus H, Mestre O (2004) Detection and correction of artificial shifts in climate series. J R Stat Soci: Series C (Appl Stat) 53(3):405–425

Gibelin AL, Déqué M (2003) Anthropogenic climate change over the mediterranean region simulated by a global variable resolution model. Clim Dyn 20(4):327–339

Hasselmann K (1979) On the signal-to-noise problem in atmospheric response studies: meteorology of tropical oceans, pp 251–259

Hasselmann K (1993) Optimal fingerprints for the detection of time-dependent climate change. J Clim 6(10):1957–1971

Hasselmann K (1997) Multi-pattern fingerprint method for detection and attribution of climate change. Clim Dyn 13(9):601–611

Hegerl G, Von~Storch H, Santer B, Cubash UPD Jones P (1996) Detecting greenhouse-gas-induced climate change with an optimal fingerprint method. J Clim 9(10):2281–2306

Hegerl G, Hasselmann K, Cubash U, Mitchell J, Roeckner E, Voss R, Waszkewitz J (1997) Multi-fingerprint detection and attribution analysis of greenhouse gas, greenhouse gas-plus-aerosol and solar forced climate change. Clim Dyn 13(9):613–634

IPCC (2001) Climate change 2001: the scientific basis. Contribution of Working Group I to the third assessment report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, UK and New York, NY, USA, 881 pp, [Houghton JT, Ding Y, Griggs DJ, Noguer M, van der Linden PJ, Dai X, Maskell K, Johnson CA (eds)]

Ledoit O, Wolf M (2004) A well-conditioned estimator for large-dimensional covariance matrices. J Multivar Anal 88(2):365–411

Mardia K, Kent J, Bibby J (1979) Multivariate analysis. Academic Press, London

Planton S, Déqué M, Douville H, Mestre O (2005) Impact du réchauffement climatique sur le cycle hydrologique. CR Geosci 337:193–202

Spagnoli B, Planton S, Déqué M, Mestre O, Moisselin JM (2002) Detecting climate change at the regional scale: the case of france. Geophys Res Lett 29(10):90–1, 90–4

Stott P (2003) Attribution of regional-scale temperature changes to anthropogenic and natural causes. Goephys Res Lett 30(14):2–1, 2–4. doi:10.1029/2003GL017324

Stott P, Tett S (1998) Scale-dependent detection of climate change. J Clim 11(12):3282–3294

Tett S, Jones G, Stott P, Hill D, Mitchell J, Allen M, Ingram W, Johns T, Johnson C, Jones A, Roberts D, Sexton D, Woodage M (2002) Estimation of natural and anthropogenic contributions to twentieth century temperature change. J Geophys Res 107(D16):10–1 10–24. doi:10.1029/2000JD000028

Zwiers F, Zhang X (2003) Toward regional scale climate detection. J Clim 16(5):793–797