Appl. Statist. (2004)
53, *Part* 3, *pp.* 405–425

Detection and correction of artificial shifts in climate series

Henri Caussinus

Université Paul Sabatier, Toulouse, France

and Olivier Mestre

Ecole Nationale de la Météorologie, Toulouse, France

[Received May 2002. Final revision August 2003]

Summary. Many long instrumental climate records are available and might provide useful information in climate research. These series are usually affected by artificial shifts, due to changes in the conditions of measurement and various kinds of spurious data. A comparison with surrounding weather-stations by means of a suitable two-factor model allows us to check the reliability of the series. An adapted penalized log-likelihood procedure is used to detect an unknown number of breaks and outliers. An example concerning temperature series from France confirms that a systematic comparison of the series together is valuable and allows us to correct the data even when no reliable series can be taken as a reference.

Keywords: Changepoints; Climate series; Linear model; Model choice; Outliers; Penalized likelihood

1. Introduction

Many long instrumental climate records are available and might provide useful information in climate research. These data sets are essential since they are the basis of a description of the climate in the past. The reliability of these long (up to 200 years in some cases) instrumental data series has been studied for a long time by climatologists. In most cases, these series are altered by changes in the conditions of measurement, such as developments in the instrumentation, relocation of the weather-station or modification of the environment. In many cases, these changes are not mentioned in the archives, which are often incomplete. Moreover spurious observations are frequent. The induced shifts often have the same magnitude as the climate signal, such as long-term variations, trends or cycles, and might lead to wrong conclusions about the evolution of the climate. Thus the detection and the correction of these aberrations is absolutely necessary before any reliable climate study can be based on these instrumental series.

These series cannot be treated directly. Since climate signals are mostly undetermined and non-stationary, they must be removed as far as possible to reveal outliers or changes in measurement conditions. Up to now, the main principle of detection of the shifts (also called breaks) rests on the assumption that the difference between the data at the station that is tested and a reference series, usually assumed not perturbed, is fairly constant in time, up to the perturbations to be detected. It is also assumed that the distribution of the difference series is normal, and that most of the shifts are step-like changes which typically alter the average value only,

Address for correspondence: Olivier Mestre, Ecole Nationale de la Météorologie, 42 avenue Gaspard Coriolis, 31057 Toulouse Cedex 1, France.
E-mail: olivier.mestre@meteo.fr

 0035–9254/04/53405

leaving the higher moments unchanged (Alexandersson, 1986). The procedures that are in widespread use among climatologists (Potter, 1981; Alexandersson, 1986) are based on likelihood ratio tests (Hawkins, 1977; Maronna and Yohai, 1978). In these procedures the null hypothesis is tested against the presence of a single changepoint. These approaches suffer from two major limitations.

(a) They rest on the existence of a so-called reference series whose reliability cannot be proved. The different methods for creating such series (Alexandersson, 1986; Førland and Hanssen-Bauer, 1994; Easterling and Peterson, 1993) do not guarantee their quality.
(b) The number of breaks in the series tested is many and unknown. Moreover, an unknown number of outliers may spoil the data.

In the next section, we propose to remove the latter limitation by using a penalized log-likelihood approach. The general problem of determining a normal linear model with an *unknown* number of changepoints and outliers has been studied by Caussinus and Lyazrhi (1997). They formulated it as a problem of testing multiple hypotheses and provided a multidecision rule. We argue that this procedure is appropriate for our problem, and we discuss its implementation in our context.

We propose to avoid the former limitation in two ways. We show first that a direct pairwise comparison of non-reliable series is valuable for the detection of the accidents. But a better approach is obtained by defining an overall two-factor model that allows us to analyse and correct a whole set of series. However, a direct use of the penalized likelihood approach with the overall model is practically intractable for combinatorial reasons. Therefore the final proposal is a mixture of the pairwise comparison and the multidimensional approach: the former is used for preselecting a set of accidents to be considered within the framework of the latter.

In the second section, our method is applied to series of annual means of daily measured maximum and minimum temperatures provided by the Meteo-France database. The series are analysed, and the results are compared with the available archives. Finally, we briefly illustrate the usefulness of the method for assessing temperature trends during the last century.

2. The methodology proposed

The main point is to compare several possibly perturbed series instead of comparing a series with an artificial reference. For that, we introduce a two-factor linear model (time $\times$ series), with possible accidents (i.e. changepoints and/or outliers) for all series at any time, and a penalized likelihood procedure to select the best model. However, the implementation is computationally intractable without a suitable stepwise approach. The solution that is proposed is to tackle the problem in two steps. First we present a naïve technique proceeding by pairwise comparisons (incidentally, it allows us to discuss the penalization term within a fairly simple framework). Second we present the overall model and show how to use the pairwise approach to preselect the possible accidents, thus reducing the number of competing models to a tractable model.

2.1. Pairwise comparisons of the series

2.1.1. Principle

The use of a reference series whose reliability cannot be really checked is a serious limitation to the previously proposed methodologies. The procedures will detect the breaks due to the tested or to the reference series without making any distinction.

There is an easy way to circumvent the reference series. It is based on the simple statement that *between two changepoints a series is reliable* (by definition), so these sections can be used as reference series. Instead of comparing a given series with a reference series whose definition is problematic, the first idea is to compare this series with all other series within the same climatic area by making a series of differences. These difference series are then tested for discontinuities by the technique of Caussinus and Lyazrhi (1997) applied to the simple case of a normal sample (see Section 2.1.2).

At this stage, we do not know which individual series is the cause of a shift that is detected in a difference series. But, if a detected changepoint remains constant throughout the set of comparisons of a candidate station with its neighbours, it can be attributed to this candidate station. The detection of the outliers follows the same principle.

Of course we should not expect to find perfect results when we consider all the comparisons. Because of the randomness of the difference series the changepoints of weak amplitude will lead to less accurate detection and sometimes no detection at all for some comparisons (in particular in the case of simultaneous breaks). Most of the time, however, the ambiguity that is induced can be removed by considering the whole set of comparisons and using the archives of the weather-stations when available—as well as the knowledge of the climatologist.

2.1.2. Statistical technique

To detect changepoints and outliers in the difference series, we use the Caussinus and Lyazrhi (1997) procedure. We give its formulation in the case of a normal sample. We consider n normal random variables Y_i ($i = 1, \ldots, n$ is the time index) and let Y denote the column vector of the Y_is. We assume that the probability distribution of Y is n-dimensional normal, with covariance matrix I_n (identity matrix of order n) up to the unknown variance. The mean is constant between two changepoints, except for an outlier for which the mean takes any real value.

Let k be the number of changepoints and l the number of outliers. Let $\tau_1, \tau_2, \ldots, \tau_k$ be the positions of the k changepoints, and let $\delta_1, \delta_2, \ldots, \delta_l$ be the positions of the l outliers. Let $K = (\{\tau_1, \ldots, \tau_k, \}\{\delta_1, \delta_2, \ldots, \delta_l\}) \subset \{1, \ldots, n\}$ be the set of changepoints and outliers, and H_K the corresponding model. To simplify the notation, we shall set $\tau_0 = 0$, and $\tau_{k+1} = n$. Finally, let $\Delta = \{\delta_1, \delta_2, \ldots, \delta_l\}$ and $n_j = \tau_j - \tau_{j-1} - \text{card}[\{\tau_{j-1}+1, \tau_{j-1}+2, \ldots, \tau_j\} \cap \Delta]$, i.e. n_j is equal to the length of the period $[\tau_{j-1}+1, \tau_j]$ minus the number of outliers within this period.

We denote

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

and

$$\bar{Y}_j = \frac{1}{n_j} \sum_{\substack{i=\tau_{j-1}+1 \\ i \notin \Delta}}^{\tau_j} Y_i$$

for $j = 1, \ldots, k+1$.

Let $C_\emptyset(Y) = 0$ and

$$C_K(Y) = \ln \left\{ 1 - \frac{\sum_{j=1}^{k+1} n_j (\bar{Y}_j - \bar{Y})^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} \right\} + \frac{2(k+l)}{n-1} \ln(n). \tag{1}$$

The penalized log-likelihood procedure that was proposed by Caussinus and Lyazrhi (1997) is

$$\text{select } H_{K^*} \text{ such that } K^* = \arg\min_K \{C_K(Y)\} \tag{2}$$

where K is any possible set of accidents. The variance is estimated by

$$\frac{1}{n-k-l-1} \sum_{j=1}^{k+1} \sum_{\substack{i=\tau_{j-1}+1 \\ i \notin \Delta}}^{\tau_j} (Y_i - \bar{Y}_j)^2$$

where the number and positions of outliers and changepoints are those given by K^* in procedure (2).

Procedure (2) has been proved to be asymptotically Bayes invariant optimal under some specific assumptions. Let us recall the two main assumptions and verify that they are realistic for the problem that we are dealing with.

(a) Caussinus and Lyazrhi (1997) considered the simple loss function taking values 1 or 0 according to proper or wrong detection. This loss function is adequate for our problem since it favours the detection of the changes in mean at their exact place (up to one or two units) rather than in the neighbourhood of it, perhaps at the price of a larger probability of non-detection (for the discussion of this point in a similar framework, see Lyazrhi (1997)). The drawback is not very serious with the methodology that we propose since the series tested is compared with several others, thus increasing the overall probability of detection. However, our methodology requires that the accidents that are detected are as well located as possible since it rests on the conjunction of their location in the various comparisons. Moreover, a good location of the accidents detected is of considerable help when trying to explain them by means of archives (see Section 2.3.2). This property is also useful to reduce the number of candidates when implementing the final procedure of Section 2.3.

(b) It is realistic to assume that changes in the measurement conditions happen independently with probability p at each time, and that the mean number of changes np is moderate, even for large n. A similar assumption is valid for the occurrence of outliers.

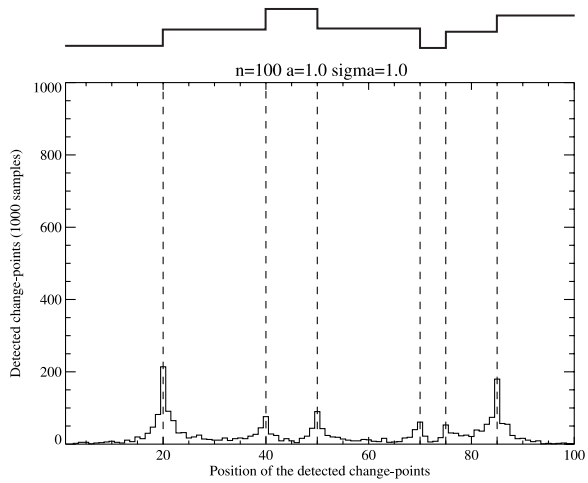
At this point, it is worth pointing out that the number of competing models increases rapidly with n. The framework is then quite different from that where more usual procedures (Akaike, 1973; Schwarz, 1978) can be expected to work as well. Actually, these have been compared with the framework that we advocate (see Section 2.1.3) and the practical results strengthen the previous discussion.

2.1.3. Simulation study

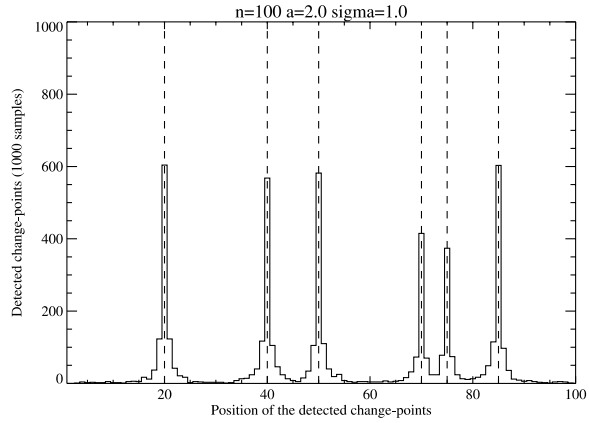
To illustrate the efficiency of the procedure proposed, we use randomly generated series which resemble the real data. Here $n = 100$ and there are six changes in mean with amplitude $\pm a$ at positions 20, 40, 50, 70, 75 and 85, as shown in Fig. 1 (broken line).

The standard deviation is constant and equal to 1. 1000 normal test series with artificial changes in mean were generated. The number k of changepoints detected and the histograms of their positions are given in Fig. 1 for $a = 1.0, 2.0, 3.0$.

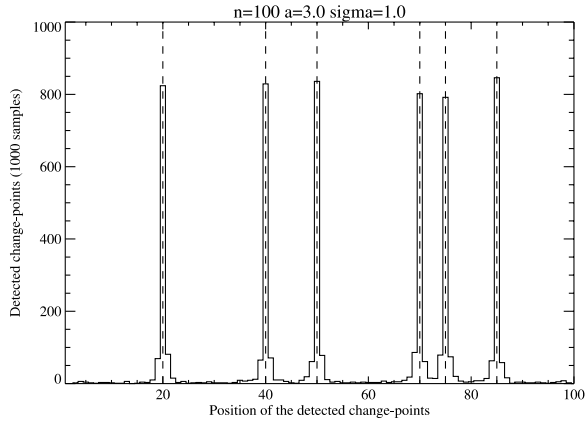
Of course the results improve when the amplitude a of the breaks increases. For low values of a, the changepoints are drowned by the residual variance and thus cannot be detected well (in number and position). For $a = 2.0$, the procedure detects four or six changepoints which are correctly placed most of the time. The detection of breaks in position 70 and 75 is always



k	$a = 1.0$
$k = 0$	0.7%
$k = 1$	25.4%
$k = 2$	19.6%
$k = 3$	38.1%
$k = 4$	12.2%
$k = 5$	3.1%
$k = 6$	0.6%
$k = 7$	0.0%
$k > 7$	0.0%



k	$a = 2.0$
$k = 0$	0.0%
$k = 1$	0.0%
$k = 2$	0.5%
$k = 3$	2.0%
$k = 4$	35.1%
$k = 5$	7.9%
$k = 6$	43.4%
$k = 7$	8.3%
$k > 7$	2.8%



k	$a = 3.0$
$k = 0$	0.0%
$k = 1$	0.0%
$k = 2$	0.0%
$k = 3$	0.0%
$k = 4$	1.2%
$k = 5$	0.5%
$k = 6$	80.9%
$k = 7$	13.1%
$k > 7$	4.3%

(a) (b)

Fig. 1. Number and position of the changepoints detected: (a) histograms of the positions of the breaks detected; (b) percentage of observed values of k

poorer, since these changepoints are very close to each other. When $a = 3.0$, the percentage of cases where six changepoints are detected reaches 80%, whereas the number of cases where $k > 6$ remains moderate. Similar results are obtained when adding outliers (see Appendix A).

Since procedure (2) rests on assumptions of normality via the penalized log-likelihood ratio (1), it is important to investigate its robustness against non-normality. A brief discussion can be found in Appendix B.

We repeated the initial experiments (normal samples with six changepoints) using Schwarz's and Akaike's penalty terms in formula (1), with respectively $k\ln(n)/(n-1)$ and $2k/(n-1)$ instead of $2k\ln(n)/(n-1)$. Some results are given in Appendix C. Akaike's criterion is clearly inappropriate. To a lesser extent, Schwarz's criterion leads also to the detection of too many changepoints.

2.2. Overall comparison and correction

We now present the overall model that is suitable for both detection and correction purposes.

2.2.1. Two-factor model

Let us consider p series belonging to the same climate area in such a way that all the series are affected by the same climatic conditions at the same time. This assumption is realistic when considering monthly or annual observations at a regional scale.

We assume that each series of observations is the sum of a climate effect, a station effect and random white noise. The station effect is constant if the series is reliable. If not, the station effect is piecewise constant between two shifts. Outliers may also occur.

Let X be a matrix of n observations X_{ij} on p series where $i = 1, \ldots, n$ is the time index and $j = 1, \ldots, p$ is the station index. Let k_j be the number of changepoints and l_j the number of outliers in series j, let $\tau_{1,j}, \tau_{2,j}, \ldots, \tau_{k_j,j}$ be the positions of these k_j changepoints and let $\delta_{1,j}, \delta_{2,j}, \ldots, \delta_{l_j,j}$ be the positions of the l_j outliers. Let $K_j = (\{\tau_{1,j}, \ldots, \tau_{k_j,j}\}, \{\delta_{1,j}, \ldots, \delta_{l_j,j}\})$ be the set of changepoints and outliers for series j. To simplify the notation, we again set $\tau_{0,j} = 0$, and $\tau_{k_j+1,j} = n$, so that K_j becomes $K_j = (\{0, \tau_{1,j}, \ldots, \tau_{k_j,j}, n\}, \{\delta_{1,j}, \ldots, \delta_{l_j,j}\})$.

In the following, *level* denotes a homogeneous subperiod between two discontinuities of a given series. For a series j with k_j breaks, let L_{jh} be the hth level ($h = 1, \ldots, k_j + 1$), i.e. L_{jh} is the interval $[\tau_{h-1,j} + 1, \tau_{h,j}]$. Note that level h for observation X_{ij} depends both on time i and on station j: when necessary it will be written $h(i, j)$.

Let μ_i be the climate effect at time i and ν_{jh} the station effect of station j for level L_{jh}. If there are no outliers, the data are described by the linear model

$$\mathbb{E}(X_{ij}) = \mu_i + \nu_{jh(i,j)},$$
$$\operatorname{var}(X) = \sigma^2 I_{np}.$$

An additional parameter is added to the mean if the data indexed by (i, j) are outliers.

The parameters are identified by $\Sigma_{i=1}^{n} \mu_i = 0$. The number of independent parameters of the model without discontinuities is $n + p - 1$.

For example, for no break in series 1,

$$\mathbb{E}(X_{i1}) = \mu_i + \nu_1$$

and, for one break at i_0 for series 2,

$$\mathbb{E}(X_{i2}) = \begin{cases} \mu_i + \nu_{21} & \text{for } i \leqslant i_0, \\ \mu_i + \nu_{22} & \text{for } i > i_0. \end{cases}$$

Remark 1. In practice, there might be several missing data. We denote their number by m.

Remark 2. The climate signal is treated as a fixed parameter so that *no assumption is made about the shape of this signal*. Moreover, conditionally on the climate signal, the disturbances may be considered independent. Finally, the local variabilities are very similar, which leads to the expression for var(X).

2.2.2. *Detection*

Assuming normality of the X_{ij}s, we now give the formulation of Caussinus and Lyazrhi's procedure in the case of the overall model.

With the previous notation we set $K = \cup_{j=1}^{p} K_j, k = \Sigma_{j=1}^{p} k_j$ and $l = \Sigma_{j=1}^{p} l_j$.

Let $\hat{\mu}_i^{\emptyset}$ and $\hat{\nu}_j^{\emptyset}$ denote the usual least squares estimates of the climate and station effects under the null hypothesis, i.e. when $K = \emptyset$, and $\hat{\mu}_i^{K}$ and $\hat{\nu}_{jh(i,j)}^{K}$ are the estimates under any given alternative hypothesis H_K.

Let $C_{\emptyset}(X) = 0$ and

$$C_K(X) = \ln\left[1 - \frac{\sum_{j=1}^{p}\sum_{i=1}^{n}\{(\hat{\mu}_i^K + \hat{\nu}_{jh(i,j)}^K)^2 - (\hat{\mu}_i^{\emptyset} + \hat{\nu}_j^{\emptyset})^2\}}{\sum_{j=1}^{p}\sum_{i=1}^{n}\{X_{ij} - (\hat{\mu}_i^{\emptyset} + \hat{\nu}_j^{\emptyset})\}^2}\right] + \frac{2(k+l)}{np - m - p - n + 1}\ln(np - m),$$

where the double sums are obviously restricted to the set of non-empty cells.

The penalized log-likelihood procedure that was proposed by Caussinus and Lyazrhi (1997) is

$$\text{select } H_{K^*} \text{ such that } K^* = \arg\min_K\{C_K(X)\}. \tag{3}$$

The variance σ^2 is estimated by

$$\frac{1}{np - m - p - n - k - l + 1}\sum_{j=1}^{p}\sum_{i=1}^{n}(X_{ij} - \hat{\mu}_i^{K^*} - \hat{\nu}_{jh(i,j)}^{K^*})^2.$$

2.2.3. *Correction*

Once detection has been achieved, final correction can be computed. Estimates $\hat{\nu}_{jh(i,j)}^{K^*}$ are used in the following way: let L_{jk_j} be the last level of series j, and $\hat{\nu}_{jk_j}^{K^*}$ the corresponding estimation of the station effect. Then, for every $X_{ij} \in L_{jh}$ $(1 \leqslant h \leqslant k_j + 1)$, the corrected X_{ij} (denoted X_{ij}^*) is given by

$$X_{ij}^* = X_{ij} - \hat{\nu}_{jh(i,j)}^{K^*} + \hat{\nu}_{j,k_j+1}^{K^*}.$$

Note that the model allows the imputation of missing data and the correction of outliers. For any missing data or outlier (i, j), the imputation is naturally given by $\hat{X}_{ij} = \hat{\mu}_i + \hat{\nu}_{jh(i,j)}$.

2.3. *Implementation*

To detect multiple accidents and changepoints in a set of series, the best approach is to use the overall model and procedure (3). Unfortunately this procedure is practically unfeasible for combinatorial reasons. To reduce the number of hypotheses to be considered, we propose to preselect the accidents by using pairwise comparisons.

2.3.1. The 'combinatorial wall'

The naïve way to implement procedures (2) and (3) is to consider all possible hypotheses, i.e. every combination of the position of the accidents (changes in means or outliers). But the number of hypotheses rises very fast with n, the length of the series, and with $k+l$, the number of accidents.

When detection is only performed for changepoints in a *normal sample*, a dynamic programming algorithm can be used (Lavielle, 1998; Hawkins, 2001). The computation time then becomes only linear in k and quadratic in n. To enable the detection of outliers as well, at a reasonable computing cost, a slightly different algorithm is used to implement procedure (2) (Mestre, 2000). This is a special stepwise algorithm where each step adds one or two more breaks or one more outlier. It is still quadratic in n, though perhaps suboptimal in a very few cases.

But the computational problems are still much heavier in the case of the overall model and procedure (3). Moreover, dynamic algorithms are not efficient in this case. In practice, this procedure is absolutely infeasible without dropping a large number of hypotheses that are not to be considered in the minimization problem. In other words, a prior selection of the possible accidents is necessary: this is the aim of the following subsection.

2.3.2. Preselection of the accidents

We propose that the first selection of possible accidents be based on pairwise comparisons of the series, as described in Section 2.1. This is efficiently achieved in an iterative way: the most striking accidents are validated. A first correction is applied to the set of series, taking into account these most probable breaks. The corrected series are again compared pairwise. Usually, this comparison displays remaining changepoints and outliers of lower amplitude, as this will be shown in the example of Section 3. The raw series are then corrected by using this extra information and compared again. Two or three steps are usually necessary to ensure a good preselection.

Let $K_{\text{Guess}\,j}$ be the set of *all possible accidents* preselected in such a way on series j and $K_{\text{Guess}} = \cup_j K_{\text{Guess}\,j}$. These are the 'candidates'. Nevertheless, another compromise is to be made to use procedure (3), since the theoretical number of hypotheses to examine is still equal to $2^{\#K_{\text{Guess}}} - 1$. A suboptimal but efficient way to proceed is the following:

$$\begin{aligned} &\text{for each } j = 1, \ldots, p \text{ select } H_{K_j^{**}} \text{ such that} \\ &K_j^{**} = \arg\min_{K_j} \left\{ C_{K_1 \cup \ldots \cup K_j \cup \ldots \cup K_p}(X) \right\}. \end{aligned} \tag{4}$$

Procedure (4) ensures the final selection of changepoints and outliers affecting the whole set of series. Note that the practical preselection should include all 'questionable' accidents, procedure (4) then being efficient to keep only the relevant ones, at the cost of a moderate extra computational effort.

Of course, at every stage of this process, we must keep in mind that the right way to analyse historical data sets is to *use both statistical inference and historical information*. Archives as unique sources of information are not sufficient since

(a) effective changes have not been necessarily registered and
(b) all changes that are found in the archives do not have a significant effect on the observations.

Therefore a statistical analysis is necessary. But archives are valuable to validate the exact dates of some shifts, providing considerable help for the selection process.

2.3.3. Correction

For practical applications, two cases may occur. When considering small regional areas, all the series are considered together. When the set of series covers larger regions, the assumption that all series have the same climatic signal no longer holds. We then correct each series by using moving neighbourhoods for both detection and correction purposes. The size and the shape of these neighbourhoods are a compromise between the knowledge of the climatologist about regional climates and the necessity to have enough data, to ensure good estimation.

Note that the most recent period remains uncorrected, and its level is taken as the reference. Therefore, once the corrected series have been inserted in the database, we may add new data each year without any correction, until a new analysis of the whole set of series is considered.

3. Application to temperature series

3.1. The temperature data set

The data set that was studied is composed of 70 annual mean series of daily minimum temperatures TN in France. Most of these series date back to the late 19th century. Their location is given in Fig. 2.

The first step is to choose a convenient scale. In the case of temperatures, for physical reasons, it is not necessary to transform the data before the analysis. This is, however, to be verified along the statistical process. It is worth noting that other kinds of data would require suitable

Fig. 2. Location map of the weather-stations that were studied

transformations, in particular the 'cumulative parameters' such as levels of precipitation or durations of sunshine.

3.2. Analysis and correction of Bourges series

3.2.1. Preselection

Let us consider the raw series of minimum temperatures for Bourges (Fig. 3). Direct observation does not reveal peculiar problems in this series. The fluctuations might be artificial or climatic in origin. Difference series are then calculated between Bourges and neighbouring raw series (Fig. 4).

The results of the pairwise comparisons with 10 stations are summarized in Fig. 5. Looking at the alignment of the changepoints in Fig. 5, we can easily detect at least three changepoints in 1910, 1967 and 1983. Owing to the large number of missing data around 1940, the Bourges series could be compared with only a few other series in this period. However, it seems there could be another break around 1943–1945. Years 1882 and 1952 are also possible dates.

Remark 3. In most of the difference series, some breaks remain isolated. For example, in the comparison Bourges *versus* Nantes, a shift is detected in 1991. This shift was provoked by the relocation of the instruments within Nantes airport, due to the construction of a new runway. Drawing the detection synthesis for Nantes reveals a good alignment of changepoints in 1991 (Fig. 6).

Coming back to Bourges, at this stage we decide to keep 1910, 1967 and 1983 as most probable shifts. The Bourges series is then corrected by using the two-factor model that was described in Section 2, K being initialized with the most probable breaks on the whole set of series. Bourges is then compared with its neighbours (Fig. 7(a)) after this first correction.

It now seems obvious that 1880 and 1882 are valid breaks. There is also evidence of problems in the series around 1945 and 1952, and the shift in 1910 does not seem to be very well corrected. The set of accidents is modified to take these possible breaks into account, and the correction

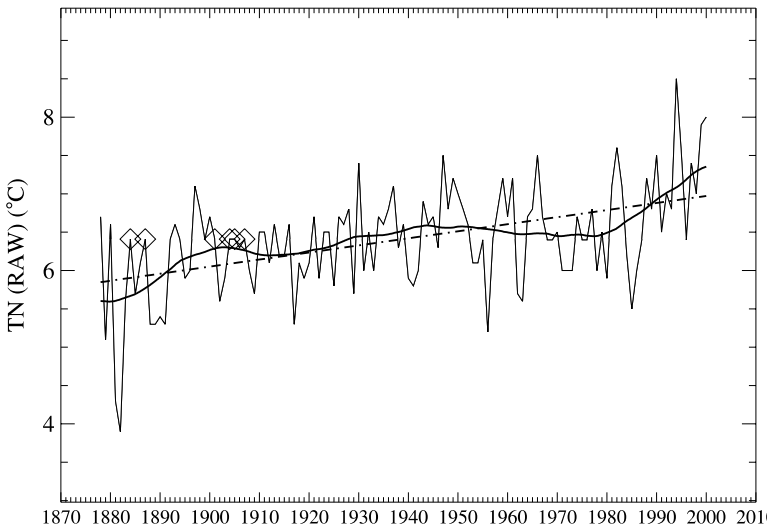


Fig. 3. Raw TN series for Bourges (mean, 6.4 °C; trend, 0.92 °C per century): ◇, missing data

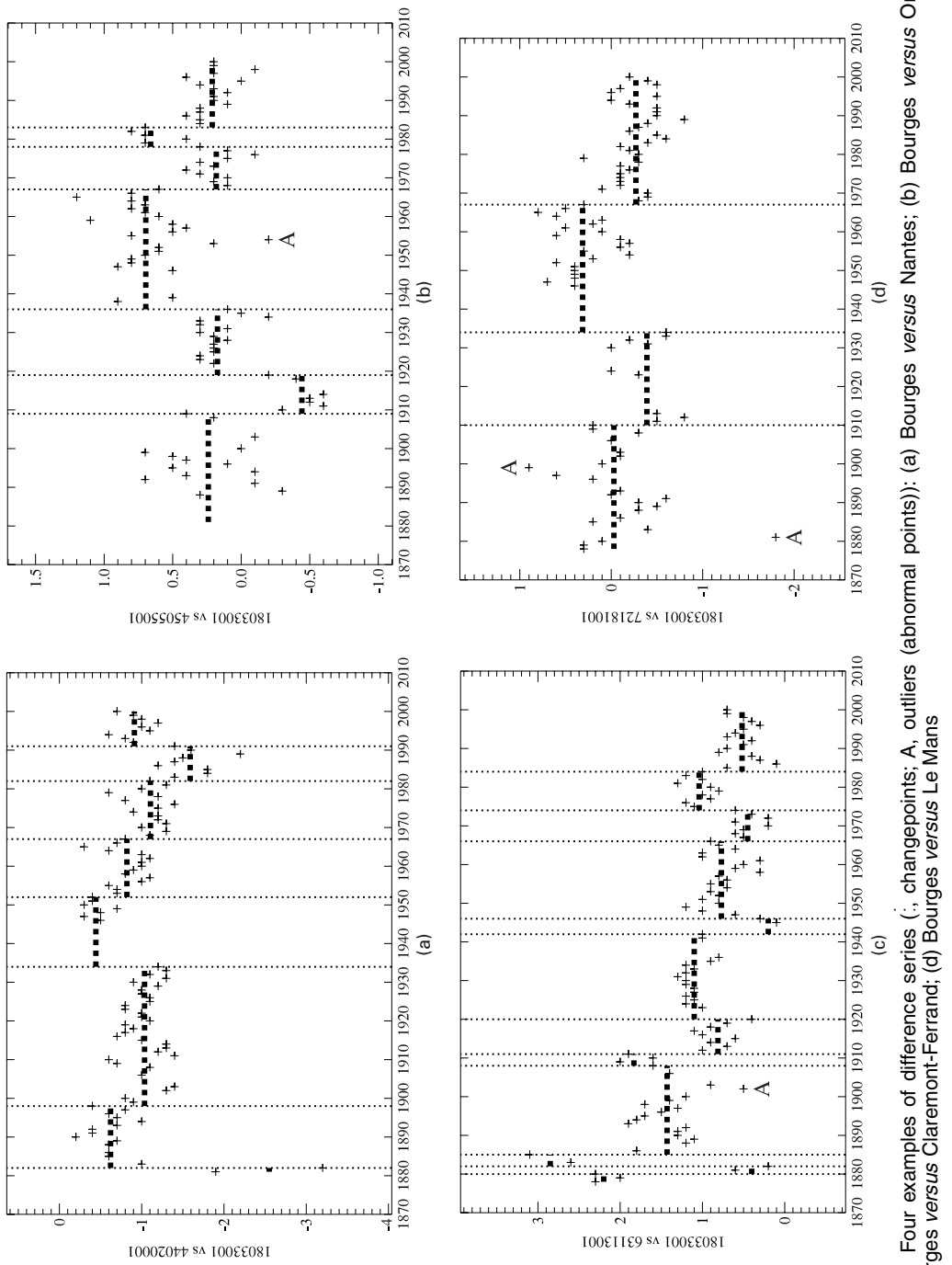


Fig. 4. Four examples of difference series (·, changepoints; A, outliers (abnormal points)): (a) Bourges *versus* Nantes; (b) Bourges *versus* Orléans; (c) Bourges *versus* Claremont-Ferrand; (d) Bourges *versus* Le Mans

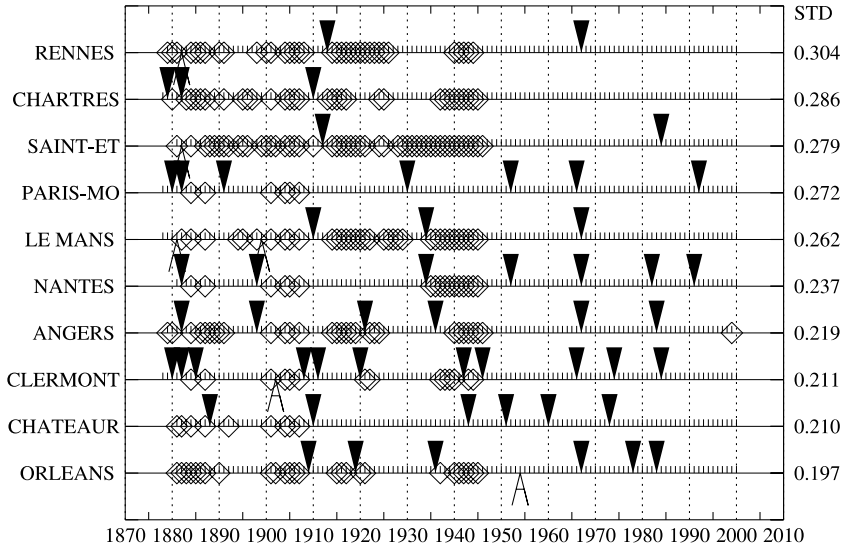


Fig. 5. Synthesis of the detected changepoints and outliers in the Bourges series (raw data): the stations are ordered from top to bottom with respect to decreasing values of the standard errors of the residuals (STD); hence, in practice, the reliability of the comparisions increases from top to bottom (▼, position of the detected changepoints in the difference series for Bourges *versus* the other stations; A, outliers; ◊, missing years in the difference series)

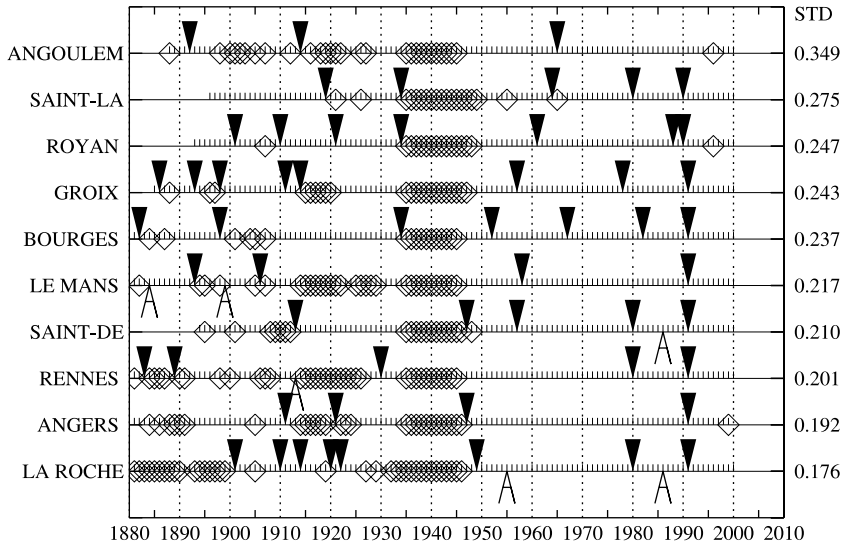


Fig. 6. Synthesis of the detected changepoints and outliers in the Nantes series (raw data)

model is computed again. The Bourges series is then compared with its neighbours once again (Fig. 7(b)).

Until 1911, observations were made in Bourges Primary School and then moved to the Observatory. From March 1945 until now, the measurements were made at Bourges airport. In 1952, the archives mention a relocation 40 m north of the original place. Again in 1967, the instruments were moved 1200 m from the previous point, facing the airport, nearby road RN151. This

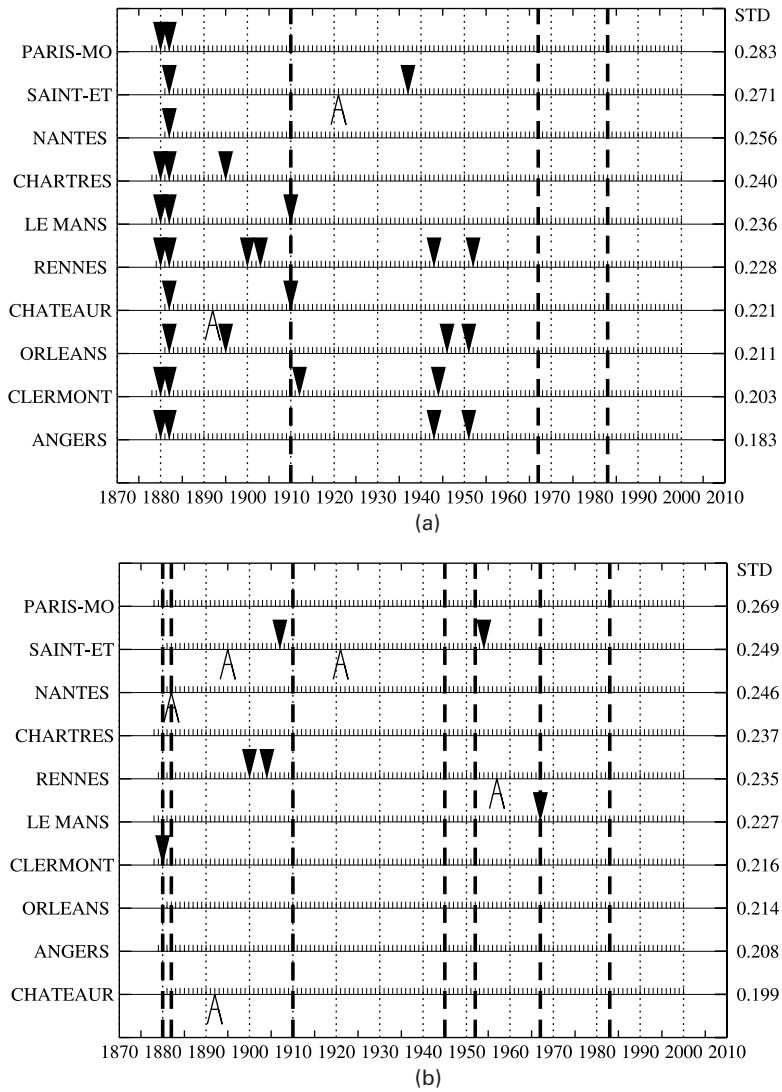


Fig. 7. Synthesis of the detected changepoints and outliers in the Bourges series (a) after first correction and (b) after second correction

change has produced the most noticeable shift in the series. The records kept do not explain the shifts in 1882 and 1983, but the coincidence with the changepoints requires that they are kept as candidates.

Remark 4. Changepoints cannot be detected well (in number and position) when a/σ is less than 1. So, indirectly, the standard error provides an estimate of the minimum amplitude of the detectable breaks which are likely to be preselected.

The very few remaining accidents in the comparison series are probably due to the first kind of errors. But, for the Bourges series, after the first correction, 1896 might have been considered a shift by a novice climatologist (Fig. 7(a)) and it is also preselected.

3.2.2. Overall procedure

To ensure the final selection, we use the overall model and procedure (4). The set of accidents K is initialized with *all* breaks and outliers previously selected on *all* series: $K = K_{\text{Bourges}} \cup K_{\text{Angers}} \cup \ldots \cup K_{\text{Paris-Montsouris}}$.

We have $K_{\text{Bourges}} = \{1880, 1882, \textit{1896}, 1910, 1945, 1952, 1967, 1983\}$. The shift in 1896 is added to K_{Bourges} for eventual validation, even if it is not probable for a trained climatologist.

Note that, in our example, the total number of guessed accidents in the set of series is greater than 75. This means more than 2^{75} different hypotheses to investigate by procedure (3), and this justifies the use of procedure (4) to lessen this number.

Let us apply this procedure on the Bourges series. In Table 1 we give the optimal subsets of shifts for given values of k_{Bourges} along with the corresponding statistic $C_K(X)$.

It is striking that the penalized log-likelihood term decreases until $k_{\text{Bourges}} = 7$ and rises again when 1896 is added to the previous breaks. For Bourges, the optimum set of shifts is then $\{1880, 1882, 1910, 1945, 1952, 1967, 1983\}$.

This leads to the estimates that are given in Table 2 for the annual values of the Bourges series.

Shift amplitudes are the differences between two consecutive level estimates. They are far from negligible. Applying these corrections to the raw data gives the corrected Bourges series in Fig. 8(a). The estimates of the μ_is (the climate effect) are given in Fig. 8(b).

To verify the underlying assumptions, the residuals are studied in Appendix D. Independence and normality assumptions can be accepted, which are consistent with the knowledge of climatologists (see, for example, Alexandersson (1986)). Temporal autocorrelation that exists in the instrumental series is well taken into account through the climate effects μ_is, in such a way that the residuals represent only local climate variability and measurement uncertainties.

Table 1. Validation of the preselected breaks

k_{Bourges}	K^{**}_{Bourges} *for the given values of* k_{Bourges}	$C_{K^{**}}(X)$
1	1967	−0.0489
2	1880 1882	−0.0983
3	1880 1882 1967	−0.1629
4	1880 1882 1910 1967	−0.1726
5	1880 1882 1910 1945 1967	−0.1786
6	1878 1882 1910 1945 1952 1967	−0.1827
7	1880 1882 1910 1945 1952 1967 1983	−0.1875
8	1880 1882 1896 1910 1945 1952 1967 1983	−0.1872

Table 2. Correction coefficients for the Bourges series

Period	*Station effect* ν (°C)	*Correction* (°C)	*Shift amplitude* (°C)
1878–1880	6.84	−0.84	−2.25 (1880)
1881–1882	4.69	1.31	1.96 (1882)
1883–1910	6.65	−0.65	−0.25 (1910)
1911–1945	6.40	−0.40	0.33 (1945)
1946–1952	6.73	−0.76	−0.25 (1952)
1953–1967	6.48	−0.48	−0.29 (1967)
1968–1983	6.19	−0.19	−0.19 (1983)
1984–2000	6.00	—	—

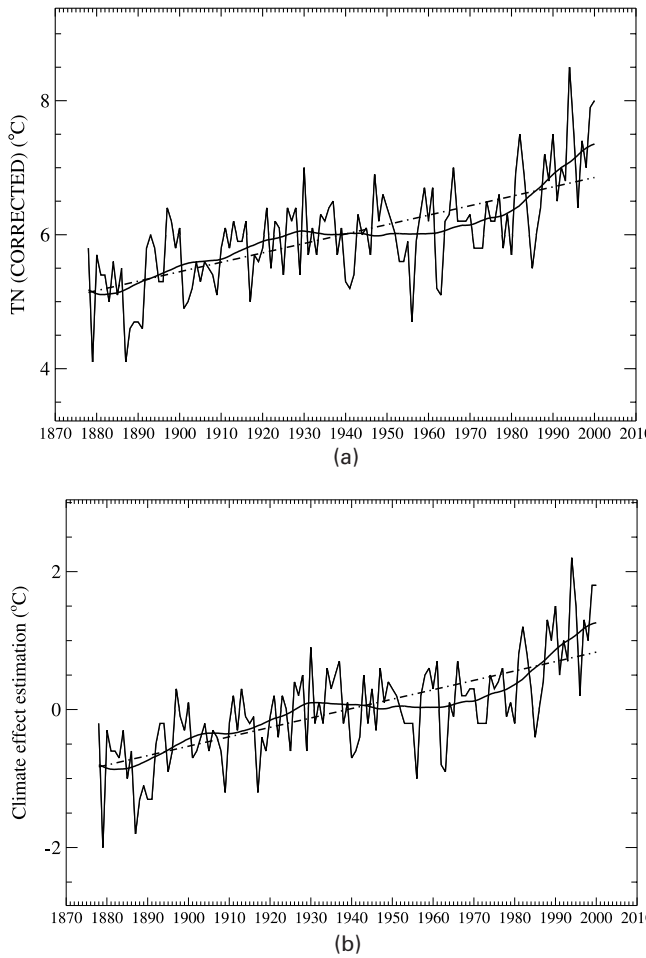


Fig. 8. (a) Corrected Bourges TN series (mean, 6.0 °C; trend, 1.41 °C per century) and (b) estimation of climate effect (mean, 0.0 °C; trend, 1.36 °C per century)

The procedure was performed for the whole data set for both minimum and maximum temperature series. It is striking that none of the original series could be considered reliable, being affected by at least three or four artificial shifts.

3.3. Temperature trends during the last century

To illustrate the usefulness of the method, we give four maps presenting the temperature trends (both minimum and maximum series) over the period 1901–2000, for raw and corrected series (Figs 9(a) and 9(c) for the raw data and 9(b) and 9(d) after correction).

Whereas the raw series show random trends, due to the poor quality of the original data, the corrected series reveal interesting features. Warming tends to be more intense in the south for maximum temperatures and in the western coasts for minimum temperatures. These patterns are consistent with some scenarios of climate change that have been investigated recently (Spagnoli *et al.*, 2002).

3.4. Discussion

The procedure that is described in this paper has been used to process a large variety of climatological series concerning rainfall, pressure, duration of sunshine, etc. In most applications, the model proposed turns out to be adequate. Some slow modifications of the environment may appear but, usually, they are negligible compared with the abrupt variations. Moreover, although the model is not perfectly adequate, it succeeds in taking them into account to a large extent. In the temperature study, a good example is given by the Paris–Montsouris series (see Fig. 9) where an urban effect has been clearly overcome. However, it might be interesting to improve

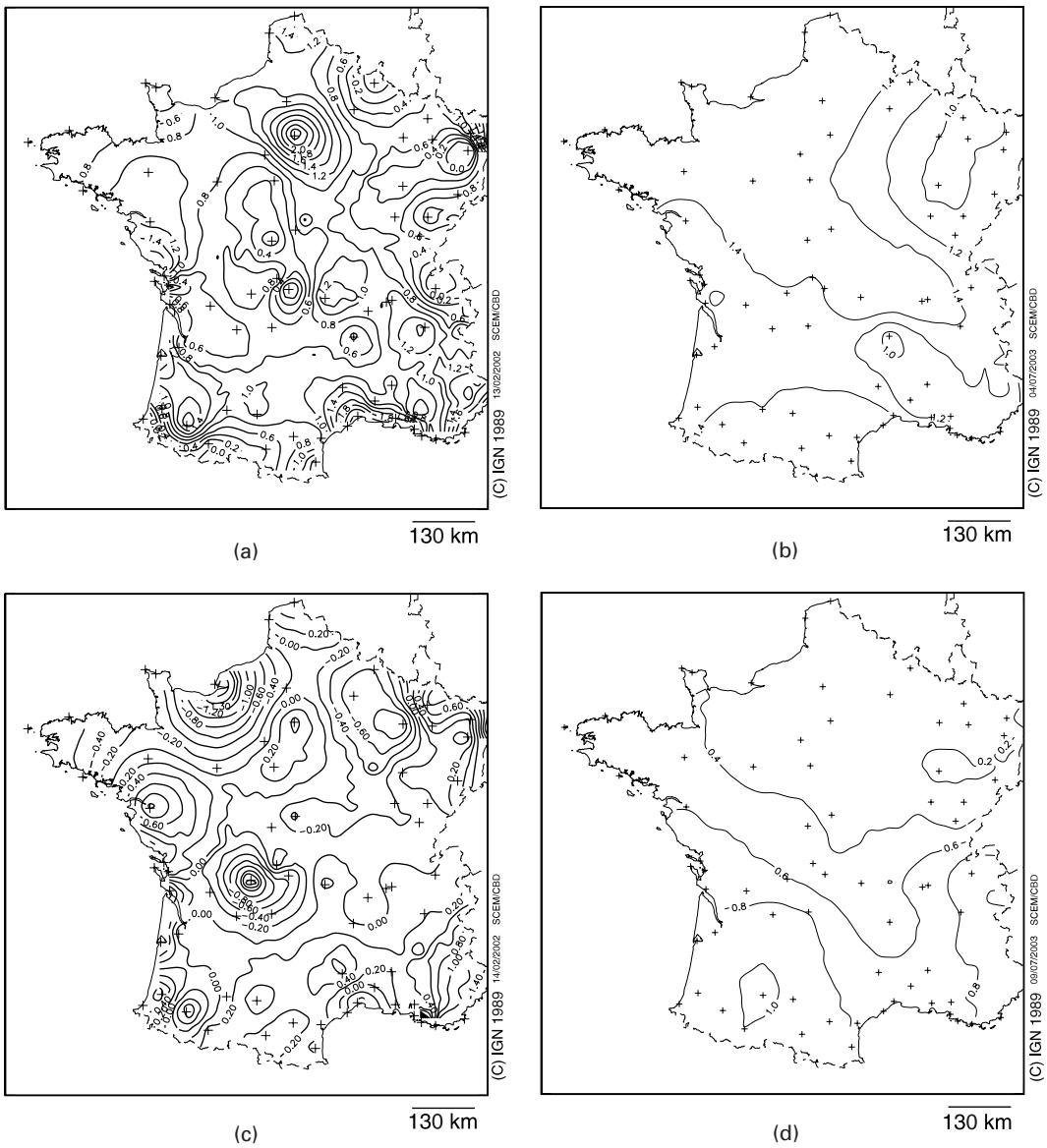


Fig. 9. Maps showing smoothed contour plots of individual linear trends for 1901–2000 estimated by ordinary least squares for each series: (a) raw minimum temperature; (b) corrected minimum temperature; (c) raw maximum temperature; (d) corrected maximum temperature

the model for better detection and correction of gradual changes together with abrupt changes.

Acknowledgements

We thank the Joint Editor and the referees for their detailed review and constructive suggestions, which led to substantial improvements.

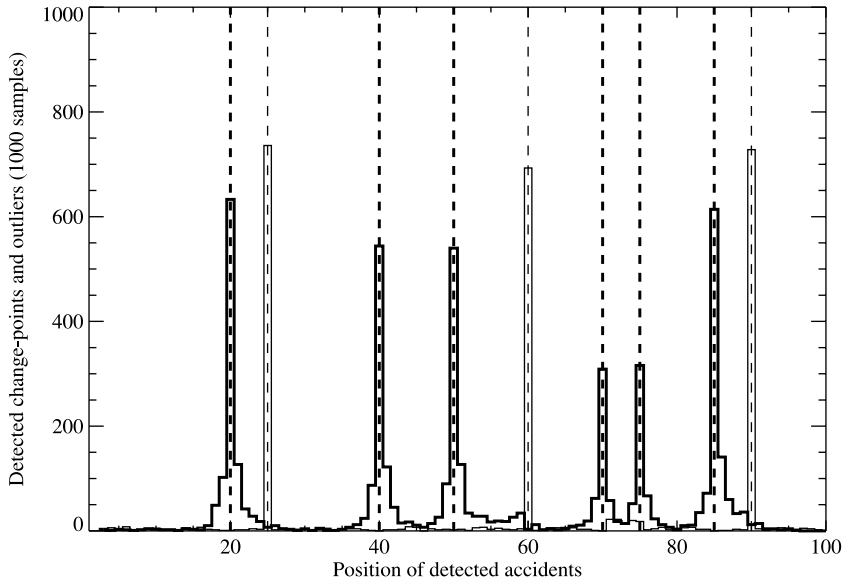


Fig. 10. Number and position of the detected changepoints and outliers (the histograms in bold correspond to changepoint detection): $n = 100$; $a = 2.0$; $b = 4.0$; $N(0, 1)$

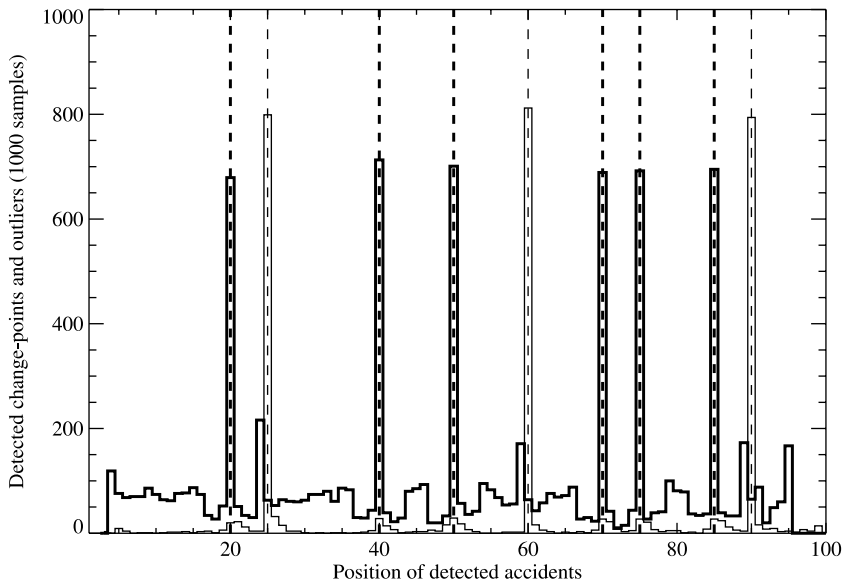


Fig. 11. Number and position of the detected changepoints and outliers in the case of a χ^2_6-distribution (the histograms in bold correspond to changepoint detection): $n = 100$; $a = 2.0$; $b = 4.0$

Appendix A: Number and position of the detected changepoints and outliers

The results that are presented in Fig. 10 are obtained from the simulated series of Section 2.1.3 with $a = 2.0$ and the addition of $\pm b = 4.0$ at positions 25 (4.0), 60 (−4.0) and 90 (−4.0) to generate outliers. It is clear that the outliers are well found and that their presence does not reduce the ability to detect the changepoints.

Appendix B: A few notes about robustness

We discuss the robustness of procedure (2) against non-normality. In most cases, 'true' changepoints (the histograms in bold) and outliers are fairly well detected but additional accidents may appear. For example, heavy-tail distributions classically generate extra outliers. In the case of a skew distribution, confusion may

Table 3. Percentage of observed values of k for various penalty terms

k	*Percentage of observed values of k for the following penalty terms†:*		
	Caussinus and Lyazrhi	*Schwarz*	*Akaike*
<4	2.5	0.0	0.0
4	35.1	0.5	0.0
5	7.9	0.6	0.0
6	43.4	8.7	0.0
7	8.3	5.5	0.0
8	2.1	10.2	0.0
9	0.7	7.8	0.0
$\geqslant 10$	0.0	66.7	100.0

†$n = 100$, $\sigma = 1.0$ and $a = 2.0$.

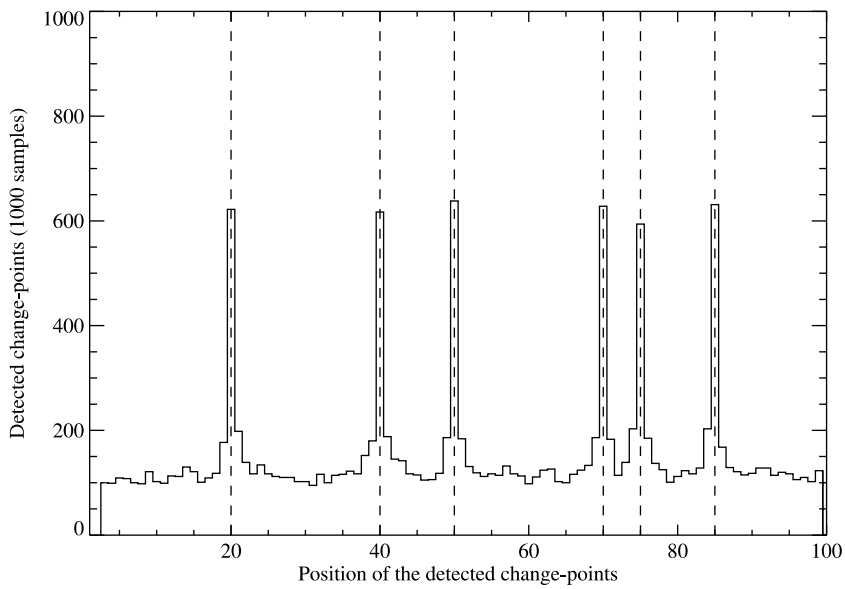


Fig. 12. Number and position of the changepoints that were detected by using Schwarz's criterion: $n = 100$; $a = 2.0$; $\sigma = 1.0$

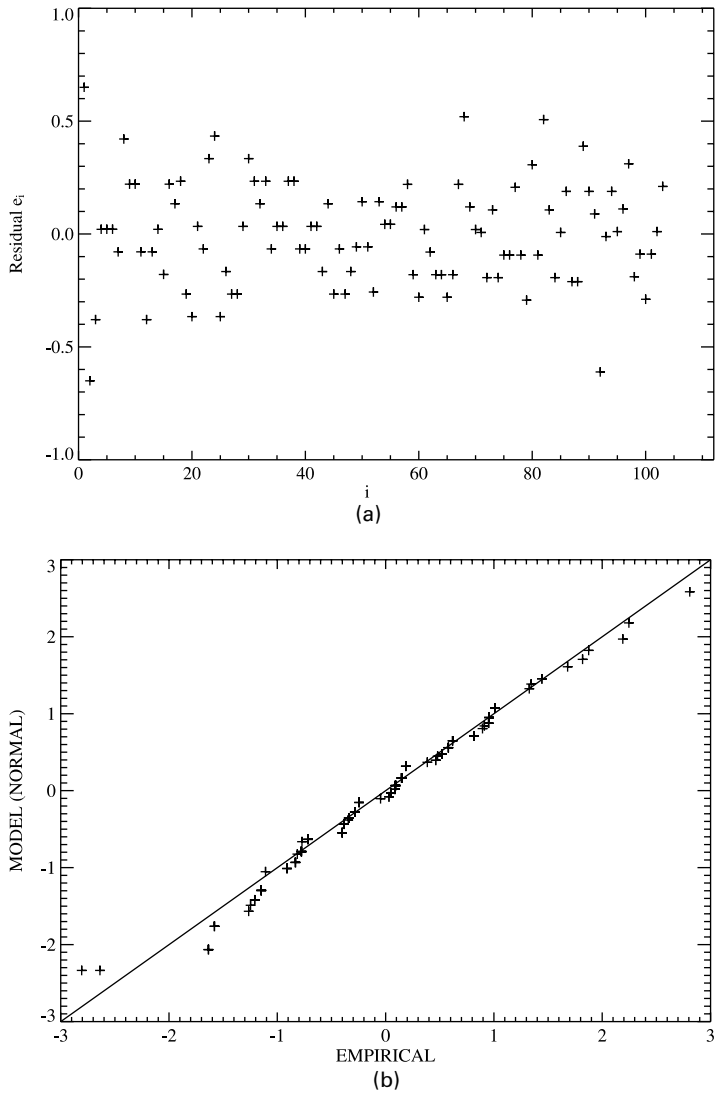


Fig. 13. (a) Time series and (b) quantile plot for the Bourges *versus* Nantes residuals (difference series)

appear between the two kinds of accident and an extra changepoint may sometimes be generated by the presence of an outlier. Fig. 11 illustrates these points for a standardized χ^2-distribution with 6 degrees of freedom.

Appendix C: Comparison of various penalty terms

Table 3 compares various penalty terms for $n = 100$ and six changepoints as in Fig. 1. Schwarz's and Akaike's criteria rest on penalty terms that are respectively $\frac{1}{2}$ and 1/4.6 of the Caussinus–Lyazrhi penalty. In the problem at hand, these weaker penalties do not sufficiently compensate for the increase in the likelihood with large values of k, so that too many changepoints are detected.

Fig. 12 gives the histogram of detected values with Schwarz's criterion. Compared with our procedure (Fig. 2, second histogram), the true changepoints are detected with slightly larger probability, but at the price of many wrong detections that are almost uniformly distributed over the period.

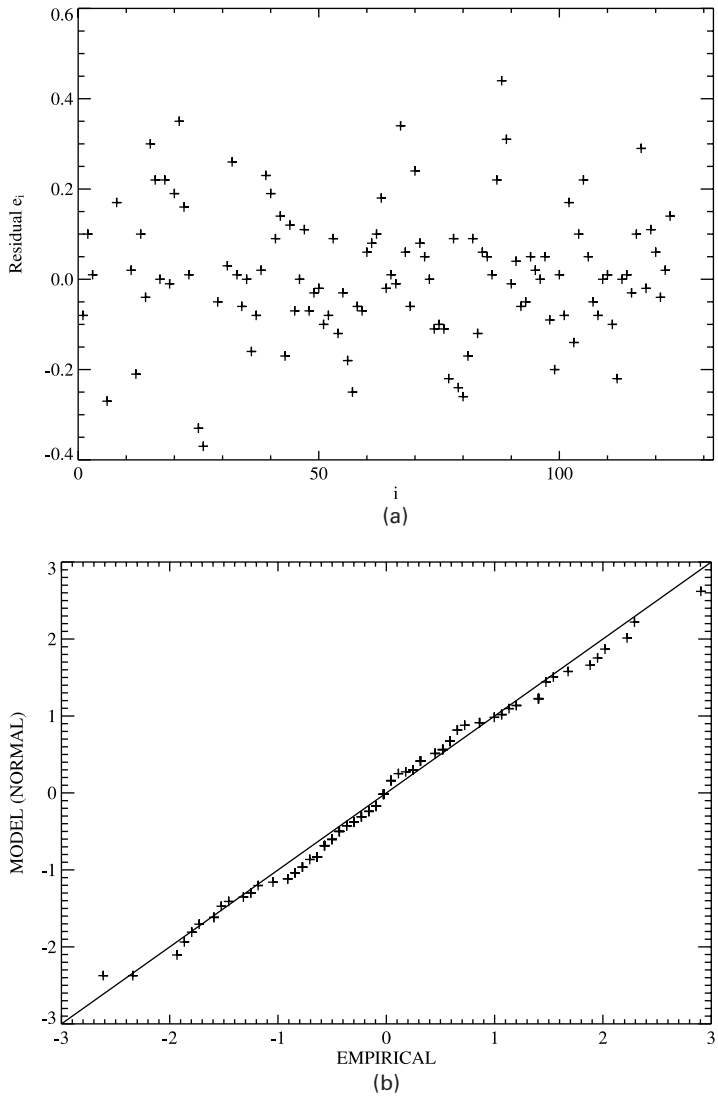


Fig. 14. (a) Time series and (b) quantile plot for the Bourges residuals (overall model)

Appendix D: Residuals

Fig. 13(a) displays the time series and Fig. 13(b) the quantile plot of the residuals for the Bourges *versus* Nantes comparison (difference series).

Fig. 14(a) gives the time series and Fig. 14(b) the quantile plot of the residuals of the Bourges series for the overall model, i.e. $X_{ij} - \hat{\mu}_i - \hat{\nu}_{jh_{(i,j)}}$ for varying i and j corresponding to the Bourges station.

In both approaches—the difference series and the overall model—the variance remains fairly constant, and the first-order autocorrelation $\gamma(1)$ takes small values (0.09 in the first example; 0.26 in the second). Quantile plots show a good consistency with the normality assumption.

References

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In *Proc. 2nd Int. Symp. Information Theory* (eds B. N. Petrov and F. Csàki), pp. 267–281. Budapest: Akadémiai Kiadó.

Alexandersson, H. (1986) A homogeneity test applied to precipitation data. *Int. J. Climatol.*, **6**, 661–675.
Caussinus, H. and Lyazrhi, F. (1997) Choosing a linear model with a random number of change-points and outliers. *Ann. Inst. Statist. Math.*, **49**, 761–775.
Easterling, D. R. and Peterson, D. C. (1993) Creation of homogeneous meteorological reference series. In *Proc. 8th Conf. Applied Climatology*, pp. 31–34. Anaheim: American Meteorological Society.
Førland, E. J. and Hanssen-Bauer, I. (1994) Homogenizing long Norwegian precipitation series. *J. Clim.*, **7**, 1001–1013.
Hawkins, D. M. (1977) Testing a sequence of observations for a shift in location. *J. Am. Statist. Ass.*, **72**, 180–186.
Hawkins, D. M. (2001) Fitting multiple change-points to data. *Comput. Statist. Data Anal.*, **37**, 323–341.
Lavielle, M. (1998) Optimal segmentation of random processes. *IEEE Trans. Signal Process.*, **46**, 1365–1373.
Lyazrhi, F. (1997) Bayesian criteria for discriminating among regression models with one possible change-point. *J. Statist. Planng Inf.*, **59**, 337–353.
Maronna, R. and Yohai, U. (1978) A bivariate test for the detection of a systematic change in means. *J. Am. Statist. Ass.*, **73**, 640–645.
Mestre, O. (2000) Méthodes statistiques pour l'homogénéisation de longues séries climatiques. *PhD Thesis*. Université Paul Sabatier, Toulouse.
Potter, K. W. (1981) Illustration of a new test for detecting a shift in mean in precipitation series. *Mnthly Weath. Rev.*, **109**, 2040–2045.
Schwarz, G. (1978) Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.
Spagnoli, B., Planton, S., Mestre, O., Déqué, M. and Moisselin, J. M. (2002) Detecting climate change at a regional scale: the case of France. *Geophys. Res. Lett.*, **29**, no. 10, 90-1–90-4.