Bayesian multiple change points and segmentation: Application to homogenization of climatic series

A. Hannart, P. Naveau

Bayesian multiple change points and segmentation: Application to homogenization of climatic series

A. Hannart[1] and P. Naveau[2]

Received 29 December 2008; revised 24 June 2009; accepted 17 July 2009; published 30 October 2009.

[1] In this paper, we describe a new multiple change point detection technique based on segmenting the time series under study into subsequences. These segments correspond to the episodes that are likely to contain a unique jump. They are found by applying Bayesian decision theory through the minimization of simple cost functions. All calculations can be performed explicitly, without falling back on Markov chain Monte Carlo methods and resulting in particularly light implementation. Through prior distributions derived from a stochastic renewal process description of jump occurrences, expert knowledge of jump amplitude and return period is also introduced in our decision process. Comparison to several multiple change point methods on simulated series lead to similar or better performance, achieved at lower computational cost.

Citation: Hannart, A., and P. Naveau (2009), Bayesian multiple change points and segmentation: Application to homogenization of climatic series, *Water Resour. Res.*, *45*, W10444, doi:10.1029/2008WR007689.

1. Introduction

[2] Long instrumental climatic records are often affected by artificial discontinuities due to changes in measurement conditions. These artificial shifts can wrongly modify the analysis of natural climate variations [*Abarca-Del-Rio and Mestre*, 2006]. The so-called change point statistical procedures have been developed to detect and remove such inhomogeneities. For a detailed review, the reader is referred to *Peterson et al.* [1998] and *Beaulieu et al.* [2007]. Classically, inhomogeneities are modelled as abrupt changes in the mean of the series, leaving its higher moments unchanged [*Alexandersson*, 1986]. Current methods simultaneously determine the number of change points and infer their positions, for instance through minimization of penalized likelihood [e.g., *Caussinus and Mestre*, 2004]. Beyond the specific context of homogenization in climatology, the change point problem is a vast, extensively treated domain of statistics, with diverse applications in econometrics, finance, biology, agronomy and hydrology among others. A general review of most common approaches can be found in work by *Reeves et al.* [2007]. In the Bayesian context, initial procedures by *Barry and Hartigan* [1992] and *Barry and Hartigan* [1993] were based on product partition models, while *Green* [1995], *Chib* [1998], and *Lavielle and Lebarbier* [2001] came up later on with different formulations. Due to the inherent complexity of the multiple change point problem, especially when the number of change points is unknown, inference of all these models have to rely heavily on Markov chain Monte Carlo (MCMC) methods. The resulting computational cost lead *Fearnhead* [2006] to propose a simplified recursive algorithm that was based on product partition models, under the assumption of independence of priors for each segment parameters. This algorithm has been further exploited by *Seidou et al.* [2007] in the particular case of a linear regression model with exact expressions for posteriors of change points positions and a straightforward simulation for posterior of the number of change points. *Fearnhead and Liu* [2007] extended the direct simulation algorithm of *Fearnhead* [2006] to online problems, and achieved a linear complexity in the number of observations through resampling from particle filters. A simpler binary segmentation procedure relying on a Bayesian criterion was also proposed by *Yang and Kuo* [2001] to deal with changes in the intensity of a Poisson process.

[3] In addition to introduce expert knowledge about the return period and amplitude of change points through prior distributions, the advantages of a Bayesian approach are twofold. In homogenization, change point locations and amplitudes posteriors are of interest to quantify the amount of uncertainty introduced by correcting inhomogeneities. It also gives an objective and coherent decision theory to decide on the number of change points and on their positions. The general motivation of this work is hence to make the most of above mentioned advantages of the Bayesian approach, while simultaneously minimizing its main disadvantage, high complexity. Our main purpose, likewise that of *Fearnhead* [2006], *Fearnhead and Liu* [2007], and *Seidou et al.* [2007], is therefore to develop an easy-to-implement Bayesian multiple change point method. Unlike these authors, we strictly constrain ourselves to exact formulae, excluding any type of simulation-based methodology. Also, differently from previous studies, we leverage Bayesian decision theory to decide on the number of change points and on their positions. More importantly, the main originality of our method consists in identifying subsequences that isolate a unique change point. To perform this task, we recall that a change point is universally defined as an abrupt shift. In our view, this characterization makes

[1] Laboratoire d'Océanographie et du Climat: Expérimentations et Approches Numériques, IPSL, CNRS, Paris, France.
[2] Laboratoire des Sciences du Climat et de l'Environnement, IPSL, CEA, CNRS, Gif-sur-Yvette, France.

0043-1397/09/2008WR007689

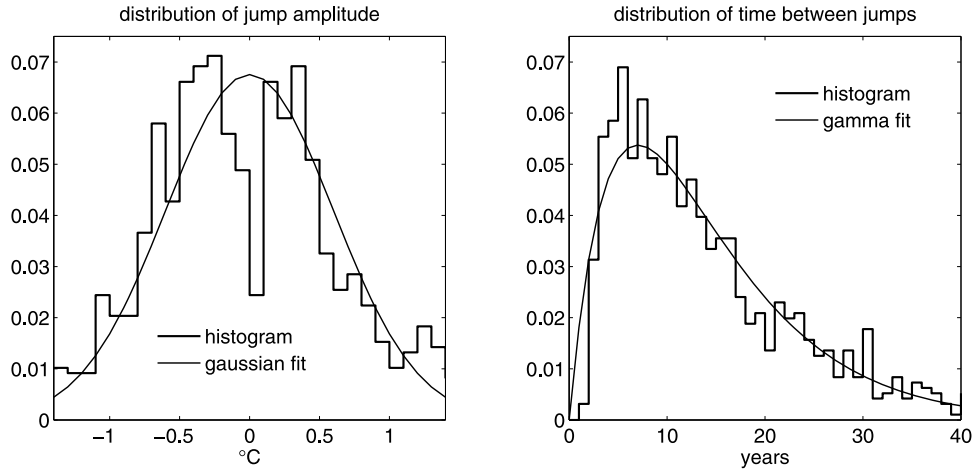

Figure 1. (left) Histogram of empirical jump amplitude and Gaussian fit ($\sigma_a = 0.6$). (right) Histogram of empirical time between jumps and gamma fit ($\lambda = 13$, $m = 2$). Empirical jump amplitude and time between jumps were obtained from homogenization results of French weather service temperature series.

change point a local concept and problem, in the sense that a change visually stands out as abrupt by comparison to its surroundings: a value that is too remote from the change point is less relevant to the detection process than a neighboring value, especially if other shifts are present between this value and the change point. In this approach, our main assumption is hence that the time series does not need to be treated globally. Rather, it could be segmented into shorter windows that capture and isolate a unique change point. If such windows could be obtained for every change points in the series, inferring their characteristics would then be straightforward using a single change point scheme. The main challenge of this approach is therefore to properly identify such subsequences, and in particular efficiently tradeoff their length. Windows should be long enough to provide enough information for accurate inference, while remaining short enough, e.g., sufficiently local, to isolate a unique change point and prevent the existence of multiple change points in the window. It is intuitive that such a tradeoff would lead to windows length typically in the order of magnitude of the average time between two consecutive jumps. If some expert knowledge about this typical jump return period is available, it would be of interest using it to perform this tradeoff. The Bayesian framework, with informative priors, is hence well suited for the proposed segmenting.

[4] This segmentation approach has a computational cost advantage, as we no longer need to work with a complex multiple change point model and a high number of interrelated parameters, leading to inextricable inference procedures, especially in the Bayesian context. Rather, two tools are required: a criterion capable of quickly quantifying the amount of evidence in favor of the existence of a single change point in a particular subsequence and a fast and single change point model to infer change point characteristics in each subsequence. This plan can be implemented because basic Bayesian single change points with explicit solutions are already available [*Lee and Heghinian*, 1977] and can be modified in a decision and cost minimization framework. Sections 2 and 3 provide these extensions of a Bayesian single change point model and combine them in a multiple change point algorithm. Sections 4 and 5 test it on simulated and real climate data, respectively. Section 6 concludes.

2. Homogenization and Single Change Point Detection

[5] Meteorologists have been measuring, correcting and interpreting records of temperatures and precipitation for many decades. For example, Figure 1 (left) displays the empirical distribution of change point amplitudes derived from temperatures homogenization results obtained by the French weather service [*Mestre*, 1998]. The bimodal shape of this empirical distribution is artificial in the sense that change points of small amplitude are practically undetectable, but they do exist as suggested by metadata [*Moberg and Alexandersson*, 1997]. Hence the solid smooth curves in Figure 1 indicates that a Gaussian fit seems adequate. The choice of a zero mean distribution reflects the lack of prior information on the sign of the jump, an assumption grounded empirically. An analysis of those data also provides information about the jump return period. An empirical estimate of the latter is about 13 years. As new time series from the same region have to be homogenized, it would be a loss to disregard the information provided by past analysis and by experts in meteorology. Integrating such prior knowledge is conceptually easy in a Bayesian framework via informative priors.

[6] To be more precise, we need to introduce a few notations for our single change point models. Denote $x = x_{1:n}$ an univariate time series of n independent real random variables $x_1, x_2, \ldots, x_n$ with overall mean μ. We assume x has zero or one change point in the mean: the binary variable $\kappa \in \{0; 1\}$ is the indicator of jump existence in x, and jump location and amplitude are called $\tau \in \{1, 2, \ldots, n\}$ and $a \in \mathbb{R}$, respectively. Concerning the distribution families, the Gaussian pdf with mean μ and variance σ^2, the zero mean normal pdf, the inverted gamma pdf with parameters α and β, the Student pdf with parameters of position α, scale β and n degrees of freedom, the Bernouilli pdf with probability ω, the uniform pdf on $\{1, \ldots, n\}$ and the dirac mass in zero are simply denoted $\mathcal{N}(. \mid \mu, \sigma^2)$, $\mathcal{N}(. \mid \sigma^2)$,

$\mathcal{IG}(.\mid \alpha, \beta)$, $\mathcal{St}(.\mid \alpha, \beta, n)$, $\mathcal{Be}(.\mid \omega)$, $\mathcal{U}(.)$ and $\delta_0(.)$, respectively. Following standard Bayesian notations, θ denotes a vector of parameters, $\hat{\theta}$ an estimator of θ, $p(.)$ a distribution and $\pi(.)$ a prior distribution. With these notations, we can introduce our first single change point Bayesian model:

$$M1:\begin{cases}\theta=(\mu,\sigma^2,\kappa,a,\tau)\in\mathbb{R}\times\mathbb{R}^+\times\{0,1\}\times\mathbb{R}\times\{1,\ldots,n\}\\ p(x|\theta)=\Pi_{i=1}^{n}\,\mathcal{N}(x_i|\mu+\delta_i(a,\tau),\sigma^2)\\ \pi(\theta)=\mathcal{N}\left(\mu|\psi_\mu,\sigma_\mu^2\right).\mathcal{IG}(\sigma^2|\alpha_\sigma,\beta_\sigma).\mathcal{U}(\tau).\mathcal{Be}(\kappa|\omega).\pi(a|\kappa,\sigma_a)\end{cases} \tag{1}$$

where $\delta_i(a,\tau)=a(\frac{\tau}{n}-\mathbf{1}_{i\le\tau})$ defines a zero mean jump sequence, κ has a Bernoulli distribution with probability ω, and $\pi(a\mid\kappa,\sigma_a)$ is the mixture defined conditionally on κ by $\mathcal{N}(a\mid\sigma_a^2)$ when $\kappa=1$ and $\delta_0(a)$ when $\kappa=0$. The definition of $M1$ is similar to the *Lee and Heghinian* [1977] and *Perreault et al.* [2000a, 2000b] models, but differs from them in several key aspects. The model now explicitly allows for the absence of a jump ($a=0$) with a nonzero probability, through the introduction of parameter κ. Metaparameter ω represents the prior probability that a jump exists and metaparameter σ_a corresponds to the typical prior jump amplitude. For the former, Figure 1 (left) provides immediate prior knowledge and for the latter the empirical knowledge about the return period λ can be converted into ω through a relationship $\omega(n,\lambda)$ on which we elaborate in section 3.

[7] Theoretical considerations can also justify our choice of prior distributions. Regarding the motivation to introduce a dirac mass in zero in $\pi(a\mid\kappa,\sigma_a)$, we emphasize that the primary purpose of model $M1$ is to decide upon the existence of a jump, a goal that can only be achieved when the event $a=0$ has a nonzero prior probability [*Robert*, 2006]. The choice of a Gaussian informative prior having a finite and informative variance σ_a^2, rather than a noninformative prior, is primarily based on the empirical evidence aforementioned. In addition, the use of noninformative priors is a well known and extensively debated difficulty in Bayesian analysis, i.e., the paradox of Jeffreys-Lindley [*Jeffreys*, 1939; *Lindley*, 1957]. The parameterization in $M1$ in terms of the mean μ and jump amplitude a is slightly different from *Lee and Heghinian* [1977] who chose (μ_1, a) and *Perreault et al.* [2000a, 2000b] who opted for (μ_1, μ_2), where μ_1 (μ_2) stands for the mean before (after) the jump. This choice has implications on the so-called reversibility of the model. We call a model "reversible" when inference is not affected by the direction in which the time series is read: this simply means that (x_1, x_2, …, x_n) and (x_n, x_{n-1}, …, x_1) should lead to the same detection of jump position and amplitude for independent data (as we assume here). In general, neither Lee's parameterization nor Perrreault's one are reversible due to the inherent asymmetry of their parameterizations but they can become reversible though, when noninformative priors on jump parameters are used. Since we explicitly chose to use prior information here, these past models are not adapted to handle reversibility.

[8] Before deciding on the existence of a change point, the posterior pdf's of jump parameters a and τ have to be derived. From the definition of M1, we can write

$$\begin{aligned}p(x|\theta)&=\prod_{i=1}^{\tau}\mathcal{N}\left(x_i|\mu-a\left(1-\frac{\tau}{n}\right),\sigma^2\right)\prod_{i=\tau+1}^{n}\mathcal{N}\left(x_i|\mu+a\frac{\tau}{n},\sigma^2\right)\\&\propto\left[\frac{1}{\sigma}\exp\left(-\frac{\left(s^2-\lambda_\tau\Delta\bar{x}_\tau^2\right)+\lambda_\tau(a-\Delta\bar{x}_\tau)^2+(\mu-\bar{x})^2}{2\sigma^2}\right)\right]^n\end{aligned}$$

where $\bar{x}$ and s^2 correspond to the classical empirical mean and variance estimators, $\Delta\bar{x}_t=\bar{x}_{t+1:n}-\bar{x}_{1:t}$ is the difference in partial means at time t, the residual variance unexplained by partitioning at time t equals $R_t=1-\frac{\lambda_t\Delta\bar{x}_t^2}{s^2}$ with the weighting factor $\lambda_t=\frac{t}{n}\left(1-\frac{t}{n}\right)$. The full posterior joint distribution is simply derived through Bayes formula

$$p(\theta|x)\propto p(x|\theta).\pi(\theta)$$

Following *Lee and Heghinian* [1977], we integrate out parameters σ and μ to get

$$p(\kappa,a,\tau|x)\propto\lambda_\tau^{-\frac{1}{2}}R_\tau^{-\frac{n-2}{2}}.\mathcal{St}\left(a|\Delta\bar{x}_\tau,\sigma_\tau^2,n-2\right).\mathcal{Be}(\kappa|\omega).\pi(a|\kappa,\sigma_a) \tag{2}$$

with $\sigma_\tau^2=\frac{R_\tau}{n\lambda_\tau}s^2$ and α_σ, β_σ, and $\sigma_\mu^{-1}\to 0$. To obtain closed forms, it is convenient to approximate the Student t distribution in (2) by a Gaussian distribution. This approximation is reasonably precise even for small values of n (we found by simulation an error of ~5% for $n=4$ and ~2% for $n=10$). Then, the equality $\mathcal{N}(.\mid\alpha,\beta).\mathcal{N}(.\mid\alpha',\beta')=\mathcal{N}(\alpha+\alpha'\mid\beta+\beta').\mathcal{N}(.\mid\frac{\alpha\beta'+\alpha'\beta}{\beta+\beta'},\frac{\beta\beta'}{\beta+\beta'})$ allows approximation of (2) by

$$\begin{aligned}p(\kappa,a,\tau\mid x)\propto\kappa\,\omega\,\lambda_\tau^{-\frac{1}{2}}R_\tau^{-\frac{n-2}{2}}&\mathcal{N}\left(\Delta\bar{x}_\tau|\sigma_\tau^2+\sigma_a^2\right)\\&\cdot\,\mathcal{N}\left(a\,|\,\Delta\bar{x}_\tau\left(1+\frac{\sigma_\tau^2}{\sigma_a^2}\right)^{-1},\sigma_\tau^2\left(1+\frac{\sigma_\tau^2}{\sigma_a^2}\right)^{-1}\right)\\&+\,(1-\kappa)\,(1-\omega)\delta_0(a)\end{aligned}$$

Integrating out a and τ, it follows that

$$p(\kappa\mid x)=\mathcal{Be}(\kappa\mid\omega^*)\omega^*=1-\left[1+\frac{\omega}{1-\omega}B\right]^{-1} \tag{3}$$

with ω^* the posterior probability of jump existence defined from the Bayes factors $B=\frac{1}{n}\sum_{\tau=1}^{n}\lambda_\tau^{-\frac{1}{2}}R_\tau^{-\frac{n-2}{2}}\mathcal{N}(\Delta\bar{x}_\tau\mid\sigma_\tau^2+\sigma_a^2)$. Denoting $p_1(.)$ a distribution conditional on $\kappa=1$, the computation of the requested marginal distributions follows

$$p(\tau\mid x)=\omega^*p_1(\tau\mid x)+(1-\omega^*)\,\mathcal{U}(\tau)\quad\text{and}$$

$$p_1(\tau\mid x)\propto\lambda_\tau^{-\frac{1}{2}}R_\tau^{-\frac{n-2}{2}}\mathcal{N}\left(\Delta\bar{x}_\tau\mid\sigma_\tau^2+\sigma_a^2\right)$$

$$p(a\mid x)=\omega^*p_1(a\mid x)+(1-\omega^*)\delta_0(a)\quad\text{and}$$

$$p_1(a\mid x)=\sum_{\tau=1}^{n}p_1(\tau\mid x).p_1(a\mid\tau,x)$$

Table 1. Performance Levels by Experiment and Method

Experiment	Description	K	n	a	Noise	BSI	PL1	PL2	MDL	SNHT
1	usual conditions	7	150	0.0	$\mathcal{N}$	0.01	0.00	0.01	0.00	0.00
2	usual conditions	7	150	1.0	$\mathcal{N}$	0.32	0.25	0.36	0.36	0.21
3	usual conditions	7	150	1.5	$\mathcal{N}$	0.57	0.55	0.61	0.63	0.39
4	usual conditions	7	150	2.0	$\mathcal{N}$	0.76	0.81	0.77	0.80	0.58
5	usual conditions	7	150	3.0	$\mathcal{N}$	0.88	0.97	0.87	0.91	0.79
6	usual conditions	15	150	1.0	$\mathcal{N}$	0.16	0.17	0.26	0.25	0.13
7	usual conditions	15	150	1.5	$\mathcal{N}$	0.37	0.32	0.49	0.49	0.23
8	usual conditions	15	150	2.0	$\mathcal{N}$	0.55	0.54	0.67	0.68	0.35
9	usual conditions	7	100	1.0	$\mathcal{N}$	0.25	0.20	0.29	0.28	0.16
10	usual conditions	7	100	1.5	$\mathcal{N}$	0.47	0.44	0.55	0.55	0.30
11	usual conditions	7	100	2.0	$\mathcal{N}$	0.66	0.69	0.70	0.72	0.42
12	usual conditions	7	150	1.5	$\chi^2(8)$	0.60	0.56	0.63	0.65	0.40
13	usual conditions	7	150	1.5	AR(1)	0.49	0.54	0.46	0.49	0.38
14	usual conditions	7	200	1.0	$\mathcal{N}$	0.35	0.28	0.39	0.38	0.24
15	usual conditions	7	200	1.5	$\mathcal{N}$	0.65	0.64	0.67	0.70	0.48
16	usual conditions	7	200	2.0	$\mathcal{N}$	0.80	0.86	0.80	0.84	0.66
17	usual conditions	11	150	1.0	$\mathcal{N}$	0.23	0.20	0.30	0.29	0.16
18	usual conditions	11	150	1.5	$\mathcal{N}$	0.47	0.41	0.57	0.58	0.30
19	usual conditions	11	150	2.0	$\mathcal{N}$	0.65	0.68	0.74	0.76	0.45
–	mean 1–19					0.49	0.48	0.53	0.54	0.35
20	very long series	250	5000	1.5	$\mathcal{N}$	0.57	0.23	0.42	0.38	0.39
21	single change point	0/1	100	1.0	$\mathcal{N}$	0.60	-	-	-	0.56
22	single change point	0/1	100	1.5	$\mathcal{N}$	0.86	-	-	-	0.82
23	single change point	0/1	100	2.0	$\mathcal{N}$	0.95	-	-	-	0.92

where $p_1(a \mid \tau, x)$ is Gaussian with mean $\Delta\bar{x}_\tau(1+\frac{\sigma_\tau^2}{\sigma_a^2})^{-1}$ and variance $\sigma_\tau^2(1+\frac{\sigma_\tau^2}{\sigma_a^2})^{-1}$, and $p_1(a \mid x)$ is a mixture of Gaussian distributions.

[9] Deciding upon the existence of a jump could be drawn from Bayes factor B using Jeffreys' absolute scale [*Jeffreys*, 1939]. Because this scale neither takes into account the prior probability ω, nor the prior amplitude σ_a, we prefer to minimize a cost function to build an estimator $\hat{\theta} = (\hat{\kappa}, \hat{a}, \hat{\tau})$. Our cost function is a basic 0-1 cost associated primarily to appropriately deciding upon the existence of the jump, and secondly, when it exists, to accurately estimating its position:

$$\mathcal{C}(\kappa,\tau;\hat{\kappa},\hat{\tau}) = \begin{cases} 0 & \text{if } [\kappa=\hat{\kappa}=0] \text{ or } [\kappa=\hat{\kappa}=1 \text{ and } |\tau-\hat{\tau}|\le h] \\ 1 & \text{otherwise} \end{cases} \tag{4}$$

To be credited by a 0 cost first requires the decision to be correct (e.g., $\kappa = \hat{\kappa}$) and second, when a jump is identified, the estimation to be accurate enough (e.g., $\mid \tau - \hat{\tau} \mid \le h$). Note that accuracy is here defined solely with respect to the estimation of jump position $\hat{\tau}$. This choice is justified by the fact that the primary goal of homogenization is to detect artificial jumps, e.g., to correctly identify break positions. The estimation of jump amplitude in the purpose of series correction is generally treated separately in an ad hoc algorithm that accounts for specificities such as seasonal differences in jump amplitude. Hence, the estimation of a is not critical in the present context. Nevertheless, the posterior average $\sum_{\tau=1}^{n} p_1(\tau \mid x)$. $\Delta\bar{x}_\tau(1+\frac{\sigma_\tau^2}{\sigma_a^2})^{-1}$ could be used as an estimator of jump amplitude, obtained by minimizing the classic quadratic cost $(a-\hat{a})^2$ under $\kappa = 1$. Note that h can be adapted depending on what is considered an acceptable level of precision for the estimation, and can be chosen consistently with the definition of decision performance. Practically, for yearly series of length in the order of 100, $h = 2$ years seems to be a fair requirement.

[10] We now minimize the average posterior cost $\rho(\hat{\kappa}, \hat{\tau} \mid x)$ defined by $\int \mathcal{C}(\kappa,\tau;\hat{\kappa},\hat{\tau}) p(\kappa,\tau \mid x)\, d\kappa d\tau$. After simplification we obtain

$$\rho(\hat{\kappa},\hat{\tau}|x) = \omega^* + \hat{\kappa}(1-\omega^*-\omega^*.\,\omega_1(\hat{\tau}\,|\,x,h))$$

where $\omega_1(\hat{\tau} \mid x, h) = \mathbb{P}\,(\mid \tau - \hat{\tau} \mid \le h \mid \kappa = 1, x) = \sum_{k=-h}^{+h} p_1(\hat{\tau}+k \mid x)$. The minimization of ρ leads to

$$\hat{\kappa} = \mathbf{1}\{\omega^*.\,(1+\omega_1(\hat{\tau}\,|\,x,h)) > 1\} \text{ and } \hat{\tau} = \text{argmax}_t\, \omega_1(t\,|\,x,h) \tag{5}$$

the estimators of jump existence and position, respectively. Equation (5) provides a complete Bayesian decision and inference scheme in the single change point context. In section 4, we find this scheme to outperform the classic procedure consisting in a decision based on the comparison of the standard normal homogeneity test (SNHT) defined by $\max_t\,(\lambda_t \Delta\bar{x}_t^2)$, to a threshold, and an estimation of jump location based on the maximum likelihood estimator defined by $\hat{\tau}_{mle} = \text{argmax}_t\,(\lambda_t \Delta\bar{x}_t^2)$. Performance levels of

both schemes are shown in Table 1 for a = 1, 1.5, 2 (experiments 21, 22, and 23).

3. Finding Subsequences

[11] We now assume that the number of change point K can be greater than one. This implies that jump return period λ also has to be integrated into model $M1$. Past knowledge can also help us to perform this extension of $M1$.

[12] A statistical analysis of aforementioned Météo France results, as shown in Figure 1 (right), indicates that a gamma distribution fits adequately the empirical distribution of time between jump arrivals with a mean λ of about 13 years and a shape parameter $m = 2$. Since the variance is equal to $\frac{\lambda^2}{m}$, parameter m can be viewed as the level of determinism associated with the jump process. When $m = 1$, the gamma pdf becomes exponential, a memoryless distribution, whereas when m goes to ∞, the gamma pdf tends to a dirac mass, i.e., a perfect memory. The short memory captured by $m = 2$ is in our view consistent with the nature of shifts in meteorological series. Station relocations, the main cause of inhomogeneities, would occur randomly with a low memory of past relocations. To model the full structure of change point occurrences, we assume jumps originate from a renewal process. Renewal processes, extensively defined and studied in probability theory [*Lefebvre*, 2005], generalize Poisson processes with an arbitrary distribution of time between events. Within this framework, the sequence of interarrival times $(T_j)_{j=2,\ldots,\infty}$ is an infinite sequence of independent and identically distributed random variables with pdf f, and T_1 the time of the first occurrence has pdf f_1. The event times $(\tau_j)_{j=1,\ldots,\infty}$ are then defined by $\tau_j = T_1 + \cdots + T_j$. Introducing $\overline{\Gamma}(x \mid m) = \frac{1}{\Gamma(m)} \int_x^{+\infty} u^{m-1} e^{-u}\, du$ the incomplete Gamma function and ν a binary variable indicating the existence of a jump at the origin of the sequence, we define the initial distribution with $f_1(t \mid m, \lambda, \nu = 1) = f(t \mid m, \lambda)$ and $f_1(t \mid m, \lambda, \nu = 0) = \lambda\, \overline{\Gamma}(\frac{t}{\lambda} \mid m)$ (section A1).

[13] To take into account the renewal process, we introduce model $M2$ which is a generalized version of $M1$, in which the prior of the change point location is now a function of f_1 and f (section A2), for example, of metaparameters m, λ and ν, instead of being uniform:

$$\pi(\tau \mid m, \lambda, \nu) \propto f_1(\tau \mid m, \lambda, \nu).f_1(n - \tau \mid m, \lambda, 0)$$

Second, the prior probability of jump existence ω in model $M1$ is no longer a metaparameter in model $M2$, but a function of metaparameters m, λ and ν. To explicit this function, we define $w_0 = \mathbb{P}(K = 0)$, $w_1 = \mathbb{P}(K = 1)$ and $\overline{\omega} = \mathbb{P}(K \leq 1) = w_0 + w_1$, and we set

$$\omega = \mathbb{P}(K = 1 \mid K \leq 1) = \frac{w_1}{w_0 + w_1}$$

Each of these quantities can be derived from m, λ and ν, explicitly or numerically (section A2). It is then immediate to show that model $M1$ is a particular case of $M2$ obtained with $m = 1$ and $\lambda = \frac{n\omega}{1-\omega}$.

[14] To identify subsequences that are most likely to contain a unique jump, a criterion characterizing each subsequence $y = x_{t_1:t_2}$, with $t_1 \in \{1, \ldots, n-1\}$ and $t_2 \in \{t_1 + 1, \ldots, n\}$, has to be introduced. We introduce the cost $\widetilde{\mathcal{C}}(y)$ of performing a single jump correction, versus performing no correction in y simply by differentiating the values of the 0–1 cost function $\mathcal{C}$ of equation (4) obtained for $\hat{\kappa} = 1$ (e.g., the cost of correcting) and $\hat{\kappa} = 0$ (e.g., the cost of not correcting):

$$\widetilde{\mathcal{C}}(y) = \mathcal{C}(y; \hat{\kappa} = 1, \hat{\tau}) - \mathcal{C}(y; \hat{\kappa} = 0, \hat{\tau}) \tag{6}$$

so that $\widetilde{\mathcal{C}}(y) = 1$ when y has zero or more than one jump; $\widetilde{\mathcal{C}}(y) = 0$ when y has one jump τ_y and $|\tau_y - \hat{\tau}_y| > h$; $\widetilde{\mathcal{C}}(y) = -1$ when y has one jump τ_y and $|\tau_y - \hat{\tau}_y| \leq h$. While other cost functions may be defined, our motivation for maintaining simple 0–1 costs is to keep our algorithm light. The mean posterior cost associated to $\widetilde{\mathcal{C}}(y)$ follows

$$\widetilde{\rho}(y|x) = 1 - \mathbb{P}(K_y = 1|x)\,.\,\left[\mathbb{P}\left(|\tau - \hat{\tau}| \leq h | K_y = 1, x\right) + 1\right]$$

To quantify $\widetilde{\rho}$, we introduce prior probabilities $\mathbb{P}(K_y \leq 1) = \overline{\omega}(n_y)$ and $\mathbb{P}(K_y = 1 \mid K_y \leq 1) = \omega(n_y)$, where n_y denotes the length of y, so that $\mathbb{P}(K_y = 1)$ a priori equals $\overline{\omega}(n_y)\,.\,\omega(n_y)$. We then update these prior probabilities based on the information provided by sequence x by applying model $M2$ to sequence y, using it to derive posterior probability of jump existence. This choice is an approximation, as it restricts the information provided by x to the information provided by y. In addition, since $M2$ holds conditionally on $K_y \leq 1$, the update is relevant for $\mathbb{P}(K_y = 1 \mid K_y \leq 1)$, but in the absence of a relevant model to update $\mathbb{P}(K_y \leq 1)$, it is left at its prior value. This approximation has the advantage of simplicity, as it avoids the use of a considerably more complex multiple change point model. Based on this choice, we obtain $\mathbb{P}(K_y \leq 1 \mid x) \simeq \overline{\omega}(n_y)$ and $\mathbb{P}(K_y = 1 \mid K_y \leq 1, x) \simeq \omega^*(n_y, y)$ where $\omega^*(n_y, y)$ stands for the posterior probability of jump existence in y resulting from the update of prior probability $\omega(n_y)$, obtained by applying model $M2$ to y. Finally, denoting $\omega_1(h, y) = \mathbb{P}(|\tau - \hat{\tau}| \leq h \mid K_y = 1, x)$, we obtain

$$\widetilde{\rho}(y \mid x) \simeq 1 - \overline{\omega}(n_y)\,.\,\omega^*(n_y, y)\,.\,(1 + \omega_1(h, y)) \tag{7}$$

[15] The problem of identifying y^* the optimal subsequence of x for a single jump correction is therefore resolved by minimizing the above quantity in y, i.e.,

$$y^* = \mathrm{argmax}_{y \in \mathcal{X}}\, \overline{\omega}(n_y)\,.\,\omega^*(n_y, y)\,.\,(1 + \omega_1(h, y)) \tag{8}$$

Once the optimal sequence y^* is identified, the decision to detect a jump is obtained by imposing a negative cost, which is equivalent to applying the detection condition (5) of model $M1$:

$$\overline{\omega}(n_{y^*})\,.\,\omega^*(n_{y^*}, y^*)\,.\,(1 + \omega_1(h, y^*)) > 1 \tag{9}$$

[16] Since $\mathrm{Card}(\mathcal{X}) \simeq \frac{1}{2} n^2$, computing the cost of all sequences y would have at least a quadratic complexity in n. However, it can be reduced easily noticing that for any y we have $2\overline{\omega}(n_y) > \overline{\omega}(n_y).\ \omega^*(n_y, y).\ (1 + \omega_1(h, y))$. It results from this inequality that whenever $\overline{\omega}(n_y) < \frac{1}{2}$, sequence y is necessarily cost positive and can be ruled out as a candidate for jump detection. Since $\overline{\omega}(n_y)$ can be expressed as a function $\phi_m(\frac{n_y}{\lambda})$ which is decreasing for any m, there exists a length $\tilde{n} = \lambda \phi_m^{-}(\frac{1}{2})$ which is proportional to λ, the

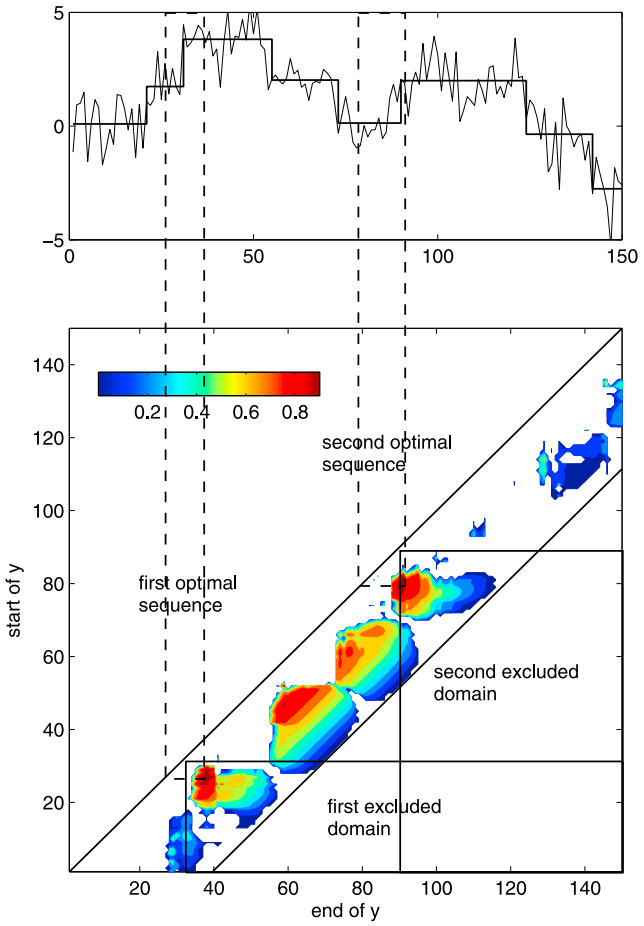

Figure 2. (top) Overall sequence x with $n = 150$, $K = 7$, and $a = 2$. (bottom) Contour plot of $\widetilde{\rho}(x_{i:j} \mid x)$ restricted to positive values, computed in the bandwidth $(0 < j - i < 1.58\,\frac{n}{K})$. Dashed rectangles highlight optimal subsequences identified at iterations 1 and 2 of the algorithm. Solid rectangles highlight the domain of value of (i, j) excluded at each iteration when x is split at the detected change point.

proportionality factor depending on m: 1.68 for $m = 1$ and 1.58 for $m = 2$ such that any sequence y having length $n_y > \tilde{n}$ can be excluded. Hence, the cost minimization can be restricted to $\widetilde{\mathcal{X}}$ the subset of $\mathcal{X}$ consisting of sequences y having length smaller than $\tilde{n} \propto \lambda$. Since $\mathrm{Card}(\widetilde{\mathcal{X}}) \simeq \frac{1}{2}\, n\tilde{n} \propto n\lambda$, the complexity of this minimization is therefore linear in n for a given prior value λ. More precisely, for a given subsequence $y = x_{i:i+l}$, computing $\widetilde{\rho}(y \mid x)$ essentially requires the computation of all the differences in partial means $\Delta\overline{y}_k = \overline{x}_{i:i+k-1} - \overline{x}_{i+k:i+l}$ for $k = 1, \ldots, l$ from which Bayes factors and posterior probabilities are derived. This has a computational cost proportional to l. Therefore, by summation for all $y \in \widetilde{\mathcal{X}}$, the total computational cost of the minimization is proportional to $\sum_{|i-j|<\tilde{n}} (i - j) \propto n\lambda^2$.

[17] Based on this subsequence identification method we now build the following recursive inference scheme.

[18] First, criterion $\widetilde{\rho}(y \mid x)$ is computed for all sequences y in $\widetilde{\mathcal{X}}$.

[19] 1. The optimal sequence $y^* = \mathrm{argmax}_{y \in x}\ \overline{\omega}(n_y)$. $\omega^*(n_y, y)$. $\omega_1(h, y)$ is identified. If $\overline{\omega}\ (n_y^*)$. $\omega^*(n_y^*, y^*)$. $(1 + \omega_1(h, y^*)) < 1$, the algorithm is stopped. Otherwise, a jump is identified in y^* at position $\hat{\tau}$.

[20] 2. The set $\mathcal{X}$ is updated by removing all sequences $y = x_{t_1:t_2}$ such that $\hat{\tau} \in [t_1, t_2]$.

[21] 3. Go to step 1 and iterate.

[22] Note that all computations are performed at step 0 of the algorithm, with complexity $\mathcal{O}(n\lambda^2)$. Iterations of steps 1 and 2 then manipulate the resulting values but no further computation is actually required. Also note that suppressions performed at step 2 basically reflect a split of the sequence at the detected jump position, before iterating the search on its remaining pieces. Figure 2 presents a cost mapping and iterations of steps 1 and 2 for a simulated sequence.

4. Simulation Results

[23] To evaluate performance and robustness of the proposed scheme, we applied it on a set of simulation-based experiments, on which we also ran various change point detection algorithms used in climate series homogenization for comparison.

4.1. Experiment Design

[24] Simulation-based experiments were designed to represent a range of situations that is diverse enough to perform sensitivity analysis, yet remaining realistic enough with respect to length, number and amplitude of jumps to mimic typical long climate records encountered in homogenization. We designed 19 experiments combining lengths n of 100, 150, 200; number of jumps K of 7, 11, 15; amplitude of jumps a of 0, 1, 1.5, 2, 3; noise of amplitude 1 having distribution Gaussian-independent, Chi2(8)-independent, Gaussian-dependent AR(1) with autocorrelation =.25. In each experiment, we simulated 1000 sequences with same length, number of jumps and amplitude of jumps, but different jump positions simulated from the renewal process described in section 3, and different noise values simulated from one of above specified distributions. On these 19 experiments, the scheme was run with prior values of amplitude an return period matching actual values, e.g., $\sigma_a = a$ and $\lambda = \frac{n}{K}$. To assess the robustness of the algorithm to the mismatch between prior and actual values, we simulated 16 additional experiments introducing different values of metaparameters (λ, σ_a) into experiment 4. To analyze performance on very long sequences, we created one experiment with $n = 5000$ and $K = 250$. Such conditions can be seen as unusual in the context of homogenization but might be representative of other contexts, such as for instance DNA sequencing. Finally, to assess performance of the detection and estimation scheme of the single change point model M1, we created three experiments with $n = 100$, $K = 0$ or 1 with probability 0.5, for $a = 1, 1.5, 2$.

4.2. Alternative Methods

[25] Other than the proposed algorithm, the Bayesian segment inference (BSI), we implemented four change point detection algorithms of increasing complexity. The first algorithm (SNHT) is a binary splitting algorithm based on the frequentist test SNHT. This simple scheme can be seen as representative of the majority of change point algorithms used for climate series homogenization. In this basic

recursive scheme, the SNHT test is applied on the entire series. It it exceeds the 95% significance level, the series is split at the location where the test statistic reaches its maximum (e.g., the maximum likelihood estimator of change point position). Then, the process is repeated recursively on the subsequences on both sides of the split, until the test statistic is below the 95% significance level.

[26] The second (PL1) and third (PL2) algorithms are more advanced and were popularized in homogenization quite recently [*Caussinus and Mestre*, 2004]. They rely on the maximization of a penalized likelihood (PL), a procedure which is applied in two steps. First, the Gaussian multiple change point model with known number of shifts K is resolved by likelihood maximization for every possible values of K. A naive maximization scheme considering every combination of shifts positions amongst the C_n^K possibilities would have complexity in n^K and be intractable. Fortunately, a dynamic optimization scheme with quadratic complexity in n is available to obtain an exact solution to this problem [*Hawkins*, 2001]: the maximum likelihood L_K^* can thus be obtained for every possible values of K with computational time in $\mathcal{O}(n^2)$. Second, the number of change points in the series is determined by maximizing a penalized log likelihood function $\log L_K^* - P_K$ where P_K increases with K and penalizes for too high dimensionality of the model. The reason for introducing such a penalty is that a direct maximization of L_K^* would invariably lead to the highest possible number of jumps, e.g., $K = n$. Several penalties have been proposed in the literature to estimate the dimension of a model. For instance, the popular Bayesian information criterion (BIC) proposed by *Schwartz* [1978] has general validity, and the corresponding penalty has a simple expression $P_K = \frac{1}{2}\frac{K}{n}\log n$. The BIC penalty can be applied directly to the present problem: algorithm PL1 corresponds to this case. Similarly, algorithm PL2 corresponds to the penalty proposed by *Caussinus and Lyazrhi* [1997] in the much more specific present context of choosing the number of change points in the mean of a Gaussian sequence. This penalty equals twice the BIC penalty, and therefore systematically leads to a lower number of shifts.

[27] The fourth method (MDL) rely on the minimization of an information criterion known as description length. Description length can be seen very generally as an alternative criterion for quantifying model fit, based on the data compression enabled by the model. The idea behind the MDL principle, as exposed by *Rissanen* [1989], is thus that the best fitting model is the one that enables maximum compression of the data. The generality and flexibility of MDL makes it relevant both for fitting a given number of model parameters and for choosing the model dimension itself. MDL has been applied to a wide range of situations (see *Saito* [1994] for a review) including the multiple change point problem. For the latter, *Davis et al.* [2006] proposed a method to detect an unknown number of change points in nonstationary AR series, with changes affecting potentially all parameters. This method enables to obtain simultaneously the number of change points and estimates of all parameters. To reduce minimization complexity, a genetic algorithm was used, leading to an approached solution. Our problem can be seen as a special case of the problem formulated by *Davis et al.* [2006], looking for changes in the mean of an AR(0). In this special case, the minimization can be implemented by adapting Hawkins algorithm instead of using Davis genetic algorithm which is more complex. This method has never been applied in the context of homogenization to our knowledge.

4.3. Metric of Detection Performance

[28] To assess performance, we compute for each experiment the average number of positives $n_{.1}$ obtained across all simulations, and split it into true positives n_{11} and false positives n_{01}. To perform the latter, a detected jump is defined to be a true (false) positive when the estimated change point falls inside (outside) an interval $[\tau - 2, \tau + 2]$ around an actual change point τ. Then, n_{11} and n_{01} can be plotted in a ROC chart that provides a mapping of detection performance. Since we need to come up with a metric for the sake of intercomparison, we use the quantity $\frac{n_{11}}{K} - \frac{n_{01}}{\frac{n}{5}-K}$. The weights $\frac{1}{K}$ and $-\frac{1}{\frac{n}{5}-K}$ associated with n_{11} and n_{01} are chosen to obtain 1 when detection is perfect (e.g., $n_{11} = K$ and $n_{01} = 0$) and 0 when detection is random. Indeed, in that case, a position τ chosen randomly has probability $\frac{5K}{n}\left(1 - \frac{5K}{n}\right)$ to be a true positive (false positive).

4.4. Comparative Results on Detection Performance

[29] Results for all experiments are shown in Table 1 and plotted in Figure 3. Averaging on all experiments, the proposed method BSI, when used with prior information matching actual situation, performs roughly at the same level than PL1, PL2 and MDL (performance = 0.51 ± 0.02) and outperforms SNHT (performance = 0.35) in every experiments. Qualitatively, two groups can be seen in the ROC chart: PL1 and SNHT characterized by low true and false positives; PL2, MDL and BSI characterized by high true and false positives. The performance of all methods is quite sensitive to jump amplitude a but with slightly different sensitivity levels, which has an incidence on their relative performance as a vary: while BSI is systematically outperformed by MDL, it outperforms PL1 for small values of a, and PL2 for large values of a.

[30] We perform a sensitivity analysis by computing the elasticity of performance to three factors that are systematically tested in the experiment set: length n, number of jumps K and amplitude a. It appears that the rank ordering of methods on their sensitivity is the same for all three factors: the sensitivity of BSI is positioned in the middle, while PL1 and SNHT are the most sensitive and PL2 and MDL are the least sensitive. The performance sensitivity of BSI to length, number of jumps and amplitude is therefore average as compared to others.

[31] We analyze robustness to non-Gaussianity ($\chi^2(8)$ noise) and to dependence (AR(1) noise) by computing the ratio of the performance obtained under such conditions to the performance obtained under a Gaussian noise. Robustness to non-Gaussianity appears to be very good for all five methods as their performance actually improve slightly instead of degrading. Differently, robustness to dependence appears to be quite contrasted: PL1 and SNHT are very robust as their performance are barely unchanged, but PL2 and SNHT robustness is weak as their performance degrade by almost 25%. Here again BSI has an in-between position as its performance decrease by 10%.

[32] Increasing length n to high values while maintaining similar jump return period (e.g., n = 5000, K = 250),

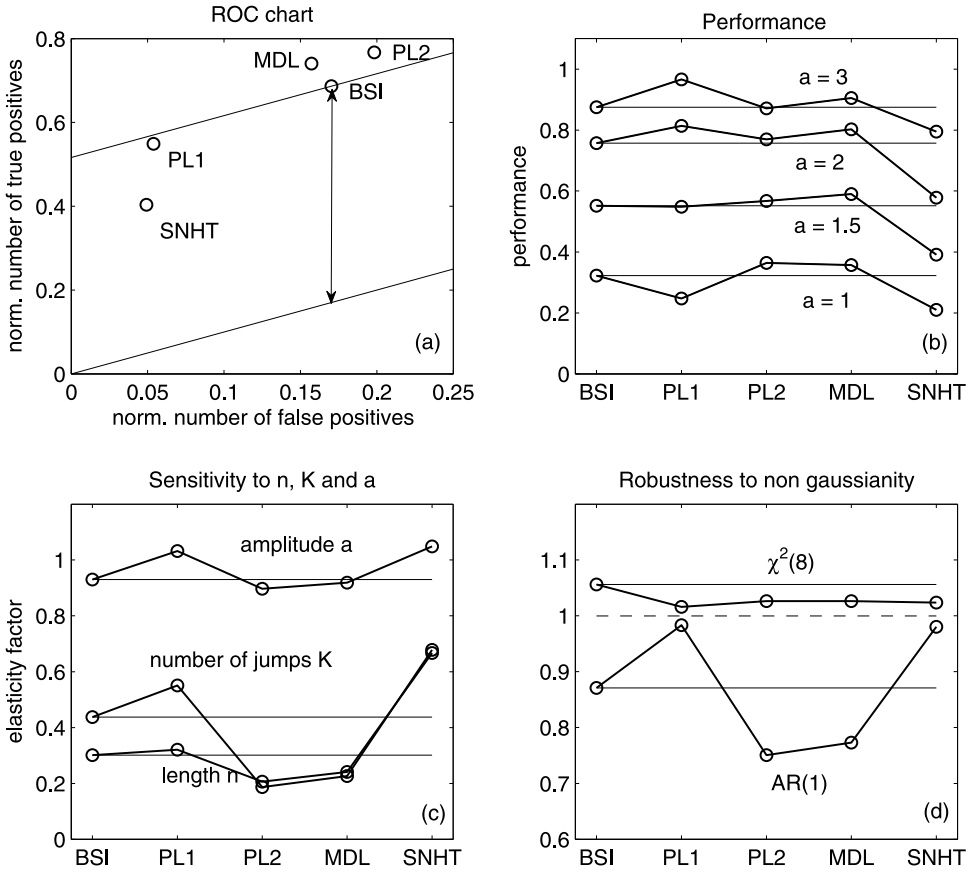

Figure 3. (a) ROC chart of percent of true positives versus percent of false positives, the difference between these two values (highlighted by the arrow) being by definition the detection performance. Each point represents values obtained by each method, averaged on all 19 "short" experiments. (b) Performance obtained with n = 150, K = 7, and a varying between 1 and 3. (c) Elasticity of performance to three factors (n, K, and a), all the rest being equal. (d) Ratio of performance obtained with $\chi^2(8)$ and $AR(1)$.

methods react quite differently (Figure 4). The performance of BSI and SNHT are maintained at the exact same level, but the performance of PL1, PL2 and MDL dramatically decrease. In this situation, BSI is comparatively better because, as is also the case of SNHT, it is insensitive to n for a fixed level of jump return period $\frac{n}{K}$. This comes from the fact that the segmenting used in BSI is purely driven by local values of x, captured through subsequences y. Therefore it is not influenced by global characteristics of x such as its total length. By contrast, PL1, PL2 and MDL are clearly

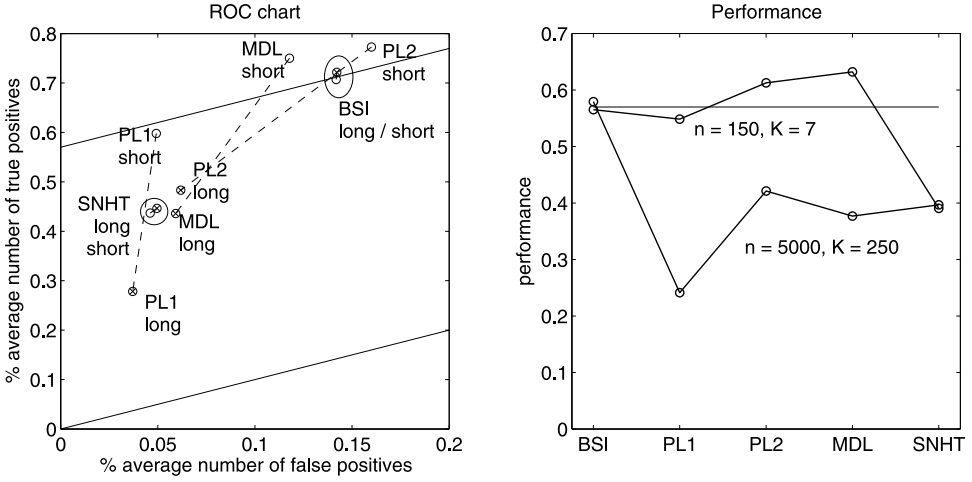

Figure 4. (left) ROC chart and (right) performance obtained for a short experiment (n = 150) and a long experiment (n = 5000) with the same value of n/K = 20 and amplitude a = 1.5.

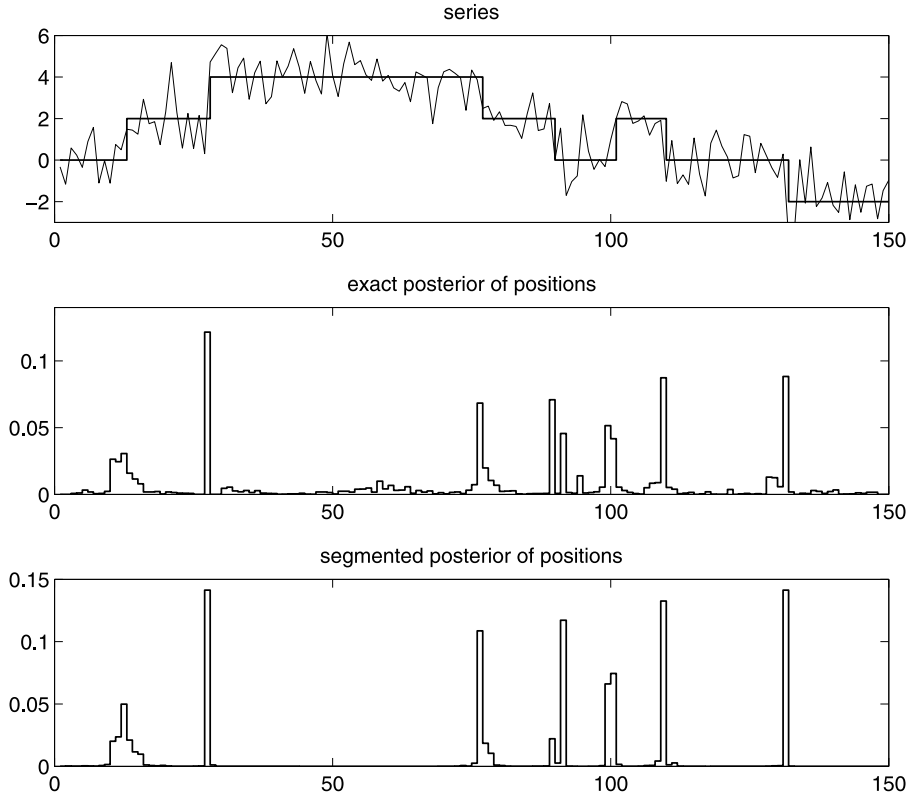

Figure 5. (top) Simulated series of length $n = 150$, $K = 7$, and $a = 2$. Posterior distributions of jump locations obtained (middle) with an advanced multiple change point scheme and (bottom) with the presented segmentation method.

not insensitive to n for a fixed jump return period $\frac{n}{K}$. This comes from the fact that the penalty terms used in PL1 and PL2 as well as the description length criterion used in MDL, are not either, since all three terms behave asymptotically for large n as $\mathcal{O}(\log(n))$. As a consequence the induced penalization becomes excessively heavy when n increases, and those methods are overly selective for high length, leading to weak performance. Therefore, for high values of n, the performance achieved by BSI is much better. Note that in addition, the comparative advantage of BSI in terms of computational time also increases with n (see below).

[33] To assess the quality of our approximation in inferring the posterior distribution of jump positions, we compare the distribution obtained with BSI to the one obtained with the advanced Bayesian scheme proposed by *Lavielle and Lebarbier* [2001]. In this method, posteriors are inferred by mean of a MCM sampler in the multiple change point framework. Although no quantified, systematic testing has been performed to gauge the match between both schemes, it has been found to be quite good in general as the example shown in Figure 5 suggests.

4.5. Comparative Results on Computational Performance

[34] We measured computational time (in elapsed CPU seconds) obtained for BSI and each four alternative methods applied to simulated series of length n increasing from 100 to 1500. We maintained the average return period $\frac{n}{K} = 20$ and used a prior return period λ matching the actual value while executing BSI. All algorithms where run on a desktop computer equipped with a 2.4 GHz Intel processor. Results are shown in Figure 6: as expected from previous considerations on complexity, computational time is found to increase with n, linearly for BSI and SNHT; quadratically for PL1, PL2 and MDL. More precisely, it approximately equals $8.8e{-}4 \cdot n$ s for BSI; $2.5e{-}5 \cdot n$ s for SNHT; $6.7e{-}6 \cdot n^2$ s for PL1, PL2 and MDL. Hence, the simplicity of SNHT scheme is clearly an advantage in terms of speed (it runs about forty times faster than BSI), although this simplicity results in a detection performance that is systematically much worse than any other tested methods, as detailed above. On the other hand, while BSI detection performance is similar to PL1, PL2 and MDL, its linear complexity makes it faster than all of these multiple change points schemes as soon as n exceeds 130; the speed ratio then keeps increasing linearly with n and quickly becomes considerable for very long series (BSI is ten times faster for $n = 1.3e{+}3$ and a hundred times faster for $n = 1.3e{+}4$). Finally, the aforementioned Bayesian scheme [*Lavielle and Lebarbier*, 2001] is found to run in approximately $15 + 6n$ s. Hence, BSI produces a reasonably good approximation of posterior distributions, roughly 10,000 times faster than an advanced MCMC multiple change point algorithm.

4.6. Robustness to Prior Information

[35] Now focusing solely on BSI, we analyze its robustness to a mismatch between prior amplitude σ_a and actual amplitude a, and between prior return period λ and actual return period $\frac{n}{K}$. We perform this analysis by computing BSI performance with $a = 2$, $n = 150$, $K = 7$ (e.g., $\frac{n}{K} = 21$), successively for σ_a varying in [0, 20] and for λ varying in [0,150]. The obtained performance pattern is similar in both

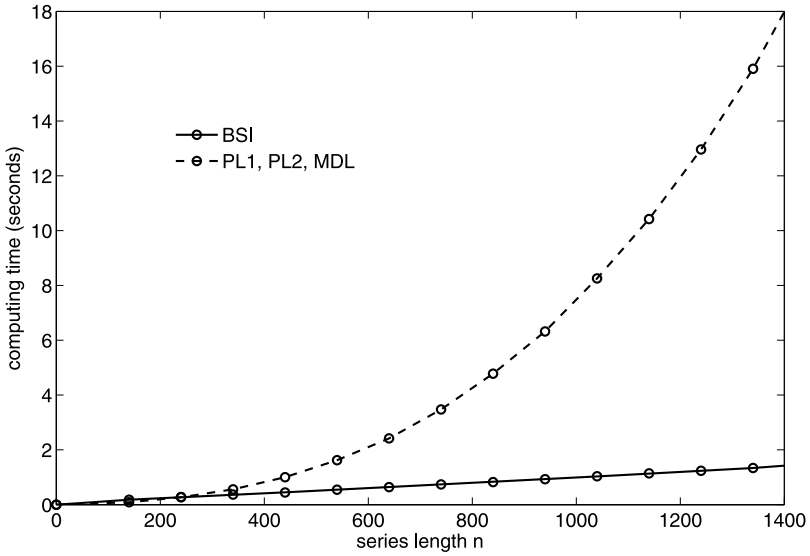

Figure 6. Computing time (in elapsed CPU seconds) as a function of series length n for BSI (dashed line) and for PL1, PL2, and MDL (solid line) implemented with the Hawkins algorithm. Computing time of SNHT is not shown, as it is visually not distinguishable from the horizontal axis.

cases (Figure 7): a sharp increase between 0 and the actual value followed by a flattening for higher values. Therefore, BSI robustness is asymmetric: very weak when values of σ_a and λ are lower than actuals as performance quickly degrades, but very strong to greater values as performance is barely unchanged. The latter robustness is striking as performance is maintained even when prior values σ_a and λ are much higher than actuals ($\times 10$). Elements of justification for such a pattern can be found in section A3. This robustness pattern suggests that a high value of σ_a and λ should be used systematically. However, while σ_a should systematically be high, using a high value of λ has numerical implications as the algorithm complexity grows in $\mathcal{O}(n\lambda^2)$. Therefore the choice of λ must be a tradeoff.

5. Application

[36] The BSI algorithm was used for homogenization of real climate data: we describe in detail the climate data and the homogenization method used, then we discuss the results obtained.

[37] The data consists of yearly average series of minimum daily temperature from sixteen stations of the French weather service. These stations are located in the southeast of France, most of them near the Mediterranean sea, in an area ranging from 42°50′N to 46°10′N in latitude and 30′W to 4°50′E in longitude. The record covers a 136 years period (1882–2007) with some data missing mostly over the late 19th century and early 20th century and during both world wars, as is common in Europe. This data has already been homogenized [*Mestre*, 1998] but since the purpose here is to test our method, we naturally used the nonhomogenized, raw data record. Further, a substantial amount of metadata on the set of selected stations is available regarding relocations and instrumental changes, so that artificial shifts are quite well documented. But again, given the purpose is to test our undocumented change point detection algorithm,

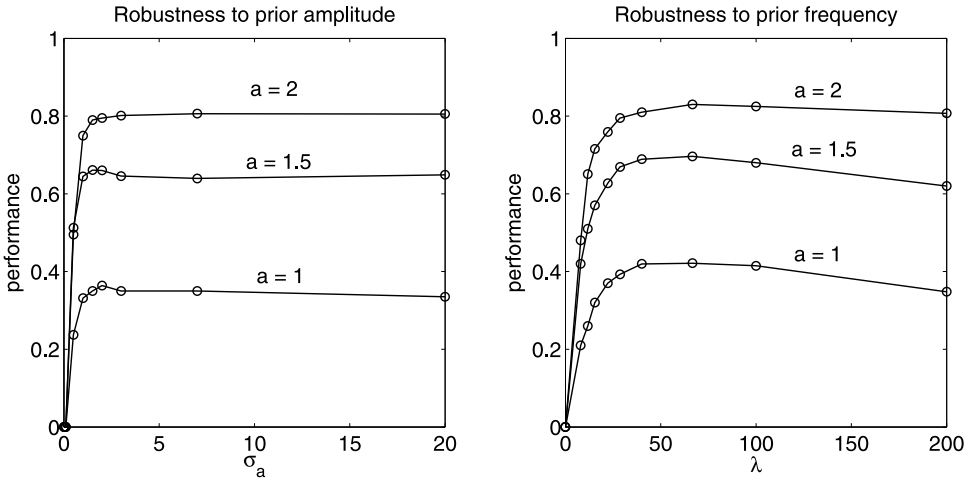

Figure 7. Performance obtained for $n = 200$; $K = 7$; and $a = 1, 1.5, 2$ when varying prior value of jump amplitude (left) σ_a and (right) λ.

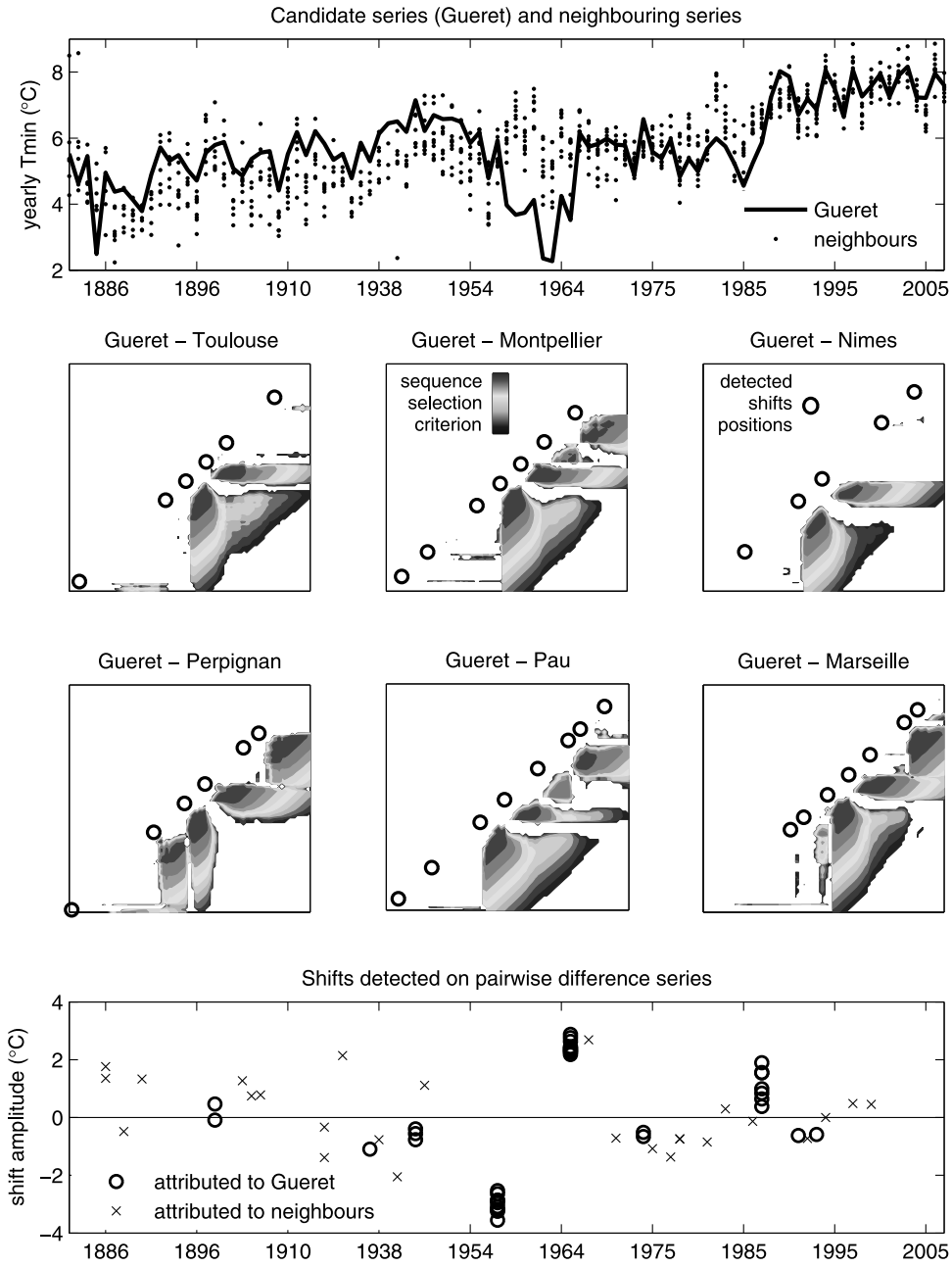

Figure 8. (top) Series of yearly averages of daily minimum temperature at Gueret station (thick line) and at neighboring stations (dots) over the period 1882–2007. (middle) Contour plots of the sequence selection criterion $\widetilde{\rho}(x_{i;j} \mid x)$ applied to six series of pairwise differences between Gueret and its most highly correlated neighboring stations ($r > 0.75$). Negative values of the criterion were excluded for clarity, and open circles on the diagonal represent the position of shifts detected in the series. (bottom) Position by amplitude chart of all shifts detected in all difference series (66 shifts) attributed either to the candidate station (open circles, 39 shifts) or to a neighbor station (crosses, 27 shifts), leading to five break detections in Gueret in 1947, 1957, 1965, 1974, and 1987.

we did not take into account this information ex ante while detecting shifts. This information was only used ex post in order to assess the ability of the algorithm to identify the shifts which existence is known from the metadata.

[38] To describe the method used, we recall the main principles at stake in homogenization; for a detailed review, the reader is referred to *Menne and Williams* [2008]. First and foremost, in a nonhomogenized climate series, instrumental shifts are mixed with the climate signal. It is widely recognized that removing the former is necessary to make the latter enough apparent for reliable detection. To do so, the relative homogeneity principle is applied: a climatological series is relatively homogeneous with respect to a synchronous series at another place if the differences of pairs of homologous averages constitute a series of random numbers (e.g., a white noise), as stated by *Conrad and Pollack* [1962]. In other words, it is assumed that inhomogeneities of a given series, referred to as the candidate

series, become apparent when processing its difference with a comparison series that has similar climatic variations. Artificial shifts are thus detected on the difference series, assuming it behaves as a white noise with jumps in the mean. In many homogenization studies, a so-called reference series obtained by averaging of sufficiently correlated nearby stations is used for comparison to the candidate series. However, since such a regional reference series is an average of potentially nonhomogeneous series, its homogeneity cannot be guaranteed: this is considered a drawback by some authors [*Caussinus and Mestre*, 2004; *Menne and Williams*, 2008]. Therefore, we prefer to implement the alternative solution used by these authors, which relies on pairwise comparisons to several neighboring station series, instead of a unique comparison to a regional reference series. Following this approach, the difference series with every sufficiently correlated neighbor (here we used $\rho > 0.70$ as a selection threshold) is derived for each candidate station. All difference series are then scanned for shifts using the BSI scheme. After this critical break detection step, the pairwise difference approach requires two additional steps, attribution and reconciliation: first, each shift detected on a paired difference series may be caused by any two series and must hence be attributed to the culprit series; second, multiple shift locations estimated on several paired difference series must be reconciled into a unique date to be used for adjusting the candidate series. Following *Caussinus and Mestre* [2004], these two steps were performed by manual review of the detected shifts. To achieve manual attribution more easily, we propose to use an original yet simple visualization tool by plotting, for each candidate, all the shifts detected on difference series in a position $\times$ amplitude chart (Figure 7, left). In such a chart, shifts that are attributable to the candidate are detected multiple times on various difference series, at a similar or identical date and with comparable amplitude, hence they tend to appear in the chart overlapping or densely grouped. Conversely, shifts that are attributable to a neighbor tend to appear isolated or more scattered. Based on visual inspection of these charts, shifts that are most obviously grouped in a given candidate chart are attributed to this candidate and subsequently removed from all its neighbors charts. Proceeding so, all shifts are iteratively attributed. Finally, reconciliation is performed by averaging estimated positions of shifts that were grouped during attribution.

[39] We applied this method to the 16 above mentioned series. Identifying correlated neighbors was easy due to the high level of correlation overall: depending on the candidate, we found between 5 and 13 neighbors satisfying to the selection criterion $\rho > 0.70$. We thus obtained 154 difference series on which BSI was run, leading to the detection of 912 shifts in total. Based on the conclusion of the robustness analysis presented in section 4, BSI was run using prior values of shifts amplitude and return period that largely and undoubtedly overestimate typical values known from past studies, by choosing $\sigma_a = 5°C$ and $\lambda = 70$ years. After visual attribution and reconciliation, we finally end up with a total of 93 shifts. In Figure 8, we illustrate each step of this process for the candidate station of Gueret. We plotted the Gueret series together with its ten selected neighbors (Figure 8, top), the contour plot of the segmentation criterion applied to difference series for its six most correlated neighbors (Figure 8, middle), and the amplitude $\times$ position chart used for visual attribution (Figure 8, bottom). The method results in detection of five shifts in Gueret in 1947, 1957, 1965, 1974 and 1987. Ex post, it appears that detected shifts match quite well with the metadata. The Gueret station has actually been relocated in 1944, 1957, 1965, 1975 and 1987: hence all five relocations were correctly detected, with an estimation error on position of 3, 0, 0, 1 and 0 years. Apart from these five relocations, no significant event susceptible to trigger a shift appear in the metadata which suggests no false negatives. Therefore, based on this particular example, the BSI detection method appears to work efficiently in practice for homogenization.

6. Conclusion

[40] We reached the objective of adapting the simple single change point framework to the multiple change point context, through the use of a recursive algorithm relying on the Bayesian decision theory and the minimization of simple cost functions. The resulting method obtain similar, if not better, performance level than three state-of-the art multiple change point methods. Yet it remains at the same time as simple and light to implement as a basic single change point iterative procedure with linear complexity in n.

[41] At a low computational cost, the method therefore benefits from the strengths of the Bayesian framework, which mainly consist in introducing expert knowledge about the return period and amplitude of change points through prior distributions, and in quantifying the uncertainty on change points characteristics through posterior distributions. In applications to homogenization, those benefits could be leveraged in two foreseeable ways. Posteriors of jumps position can be found useful to help objectivize a decision on the existence of jumps when they are detected simultaneously on multiple series of pairwise station comparison, a process which is currently performed visually. Also, joint posteriors of jumps position and amplitude can be used to derive confidence intervals on the corrected series, and to quantify the uncertainty introduced by homogenization in climatic trends further obtained.

[42] Robustness to mismatch between prior and actual values of amplitude and return period shows a pronounced asymmetry. Hence when expert information is blur, it appears to be a very safe option to boldly overestimate λ and σ_a as it will barely affect performance. The method proved particularly effective for long series with frequent jumps. In such a situation, the segmentation approach delivers its full benefit by simultaneously reducing the complexity of the problem and outperforming other tested methods. The method may therefore be of interest to other fields where such long series are found.

[43] Finally, in the context of the multiple change point problem, the stochastic description of change points occurrence used in this article may offer possibilities for Bayesian and non-Bayesian models.

Appendix A

A1. Deriving f_1 From f

[44] Parameter ν indicates the existence of a jump at the origin of the sequence. In case $\nu = 1$, τ_0 the immediate

antecedent of τ_1 is equal to zero. Since $\tau_1 - \tau_0 \sim f$, thus $\tau_1 \sim f$ and $f_1(t \mid m, \lambda, \nu = 1) = f(t \mid m, \lambda)$. In case $\nu = 0$, nothing is known regarding the position of τ_0 except that $\tau_0 \leq 0$ and $\tau_1 - \tau_0 \sim f$. Hence $f_1(t \mid m, \lambda, \nu = 1) = \mathbb{P}(\tau_1 - \tau_0 \geq t) = \int_t^{+\infty} f(u)\, du = \lambda\, \overline{\Gamma}(\lambda x \mid m)$.

A2. Deriving Priors From f and f_1

[45] We have $\mathbb{P}(K=0) = \mathbb{P}(T_1 > n)$ and $\mathbb{P}(K=1) = \int_0^n \mathbb{P}(T_1 = n-u)\, \mathbb{P}(T_2 > u)\, du$. With assumptions and definitions used in *M*4, $\mathbb{P}(T_2 > x) = 1 - \int_0^x f(u)\, du = f_1(u \mid m, \lambda, \nu = 0)$. Therefore:

$$\omega(m,\lambda,\nu) = \mathbb{P}(K=1 \mid K \leq 1) = \mathbb{P}(K=1) \,/\, \mathbb{P}(K \leq 1) = \frac{w_1}{w_0 + w_1}$$

$$w_0 = \mathbb{P}(K=0) = 1 - \int_0^n f_1(u \mid m,\lambda,\nu)\, du,$$

$$w_1 = \mathbb{P}(K=1) = \int_0^n f_1(u \mid m,\lambda,\nu)\, .\, f_1(n-u \mid m,\lambda,0)\, du,$$

To obtain $\pi(\tau)$, we write

$$\begin{aligned}\mathbb{P}(\tau = t \mid K=1) &= \mathbb{P}(\tau = t, K=1) \,/\, \mathbb{P}(K=1)\\ &= \mathbb{P}(T_1 = t, T_2 > n-t) \,/\, \mathbb{P}(K=1)\\ &= \mathbb{P}(T_1 = t)\, .\, \mathbb{P}(T_2 > n-t) \,/\, \mathbb{P}(K=1)\\ &= f_1(t \mid m,\lambda,\nu)\, .\, f_1(n-t \mid m,\lambda,\nu=0) \,/\, w_1\end{aligned}$$

since $\mathbb{P}(T_2 > u) = f_1(u \mid m, \lambda, \nu = 0)$.

A3. Robustness of the Method to a Mismatch Between Priors and Actuals

[46] The above described robustness structure comes from properties of the Bayes factor and from the nature of the cost used. To understand robustness to a λ mismatch, note that λ influences on the result via two quantities: the prior probability of jump existence $\omega(n_y)$ in the sequence y of length n_y, and the prior probability that there is at most one jump $\overline{\omega}(n_y)$. Now let us assume that the data shows strong evidence of a jump within a sequence y that is short ($n_y < \lambda$); for example, the Bayes factor associated to y is high. The Bayes factor being a very contrasted metric (hence the logarithmic scale used by Jensen to interpret it), in that case it is generally so high that even a very low prior jump probability $\omega(n_y)$ resulting from a low value of $\frac{n_y}{\lambda}$ will result in a high posterior jump probability. On the other hand, since $\frac{n_y}{\lambda}$ is low, the prior probability $\overline{\omega}$ to find 0 or 1 jump in y is high. Combining, the posterior cost $\widetilde{\rho}(y)$ will remain high. Therefore, it is very difficult for a large value of λ to discount a short sequence having strong evidence of a jump. Conversely, let us assume that the data shows strong evidence of a jump for a long sequence y, e.g., $n_y > \lambda$. Then both the Bayes factor associated to y and the prior jump probability are high, resulting in high posterior jump probability. However, since $\frac{n_y}{\lambda}$ is high, the prior probability $\overline{\omega}$ to find 0 or 1 jump in y is low. Combining, the posterior cost $\widetilde{\rho}(y)$ will be low, and even negative if $\frac{n_y}{\lambda}$ is greater than the threshold above which $\overline{\omega} < \frac{1}{2}$ (1.58 for m = 2). Therefore, it is quite easy for a small value of λ to rule out a long sequence with strong evidence of a jump. This asymmetry in the behavior of the cost when λ varies explains the asymmetry in the robustness to prior return period. The asymmetry associated to σ_a is simpler to explain: in that case, properties of the Bayes factor are only at stake. For a given sequence y, the Bayes factor is a steeply increasing function of σ_a from 0 until a maximum is reached in the neighborhood of the actual value a, and then decreases to zero. But while the increase between 0 and a is steep, the decrease is extremely slow when there is reasonable evidence of a jump in y (for $a = 1.5$ and $n = 100$, it requires $\sigma_a > 10^4$ to halve its value). Hence, it is considerably more difficult for a high value of σ_a to discount this evidence than it is for a small value.

[47] **Acknowledgments.** The authors would like to thank Olivier Mestre for interesting discussions about the statistical and practical aspects of homogenization, the European Commission COST action HOME (ES0601), Marc Lavielle for sharing his computer code, and three anonymous reviewers for very helpful comments that significantly improved this article. Alexis Hannart wishes to thank the European Commission 6th Framework programme CLARIS Project (001454) and the IPSL research action HoDem for funding the present work. We also want to thank the Centre National de la Recherche Scientifique (CNRS) and the Institut de Recherche pour le Développement (IRD) for their support in this collaboration and the Atmosphere and Ocean Department of the University of Buenos Aires for welcoming Alexis Hannart. The ANR AssimilEx project is also acknowledged by Philippe Naveau.

References

Abarca-Del-Rio, R., and O. Mestre (2006), Decadal to secular time scales variability in temperature measurements over France, *Geophys. Res. Lett.*, *33*, L13705, doi:10.1029/2006GL026019.

Alexandersson, H. (1986), A homogeneity test applied to precipitation data, *J. Climatol.*, *6*, 661–675.

Barry, D., and J. A. Hartigan (1992), Product partition models for change-point problems, *Ann. Stat.*, *20*, 260–279.

Barry, D., and J. A. Hartigan (1993), A Bayesian analysis for change-point problems, *J. Am. Stat. Assoc.*, *88*, 309–319.

Beaulieu, C., T. B. M. J. Ouarda, and O. Seidou (2007), A review of homogenization techniques for precipitation data and their applicability to precipitation series (in French), *Hydrol. Sci. J.*, *52*(1), 18–37.

Caussinus, H., and F. Lyazrhi (1997), Choosing a linear model with a random number of change-points and outliers, *Ann. Inst. Stat. Math.*, *49*, 761–775.

Caussinus, H., and O. Mestre (2004), Detection and correction of artificial shifts in climate series, *J. R. Stat. Soc., Ser. C*, *53*, 405–425.

Chib, S. (1998), Estimation and comparison of multiple change-point models, *J. Econometrics*, *86*, 221–241.

Conrad, V., and L. W. Pollack (1962), *Methods in Climatology*, Harvard Univ. Press, Cambridge, Mass.

Davis, A. R., T. C. M. Lee, and G. A. Rodriguez-Yam (2006), Structural break estimation for nonstationary time series models, *J. Am. Stat. Assoc.*, *101*, 473–495.

Fearnhead, P. (2006), Exact and efficient Bayesian inference for multiple changepoint, *Stat. Comput.*, *16*, 203–213.

Fearnhead, P., and Z. Liu (2007), Online inference for multiple changepoint problems, *J. R. Stat. Soc.*, *69*, 589–605.

Green, P. J. (1995), Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrika*, *82*(4), 711–732.

Hawkins, D. M. (2001), Fitting multiple change-points to data, *Stat. Data Anal.*, *37*, 323–341.

Jeffreys, H. (1939), *Theory of Probability*, Oxford Univ. Press, Oxford, U. K.

Lavielle, M., and E. Lebarbier (2001), An application of MCMC methods for the multiple change-points problem, *Signal Process.*, *81*, 39–53.

Lefebvre, M. (2005), *Processus Stochastiques Appliqués*, Hermann, Paris.

Lee, A. S. F., and S. M. Heghinian (1977), A shift of the mean level in a sequence of independent normal random variables—A Bayesian approach, *Technometrics*, *19*, 503–506.

Lindley, D. V. (1957), A statistical paradox, *Biometrika*, *44*, 187–192.

Menne, M. J., and C. N. Williams Jr. (2008), Homogenization of temperature series via pairwise comparisons, *J. Clim.*, *22*, 1700–1717.

Mestre, O. (1998), Homogénéité des séries du réseau climatologique d'état, 153 pp., Meteo France, Paris.

Moberg, A., and H. Alexandersson (1997), Homogenization of Swedish temperature data. Part II: Homogenized gridded air temperature compared with a subset of global gridded air temperature since 1861, *Int. J. Climatol.*, *14*, 35–54.

Perreault, L., J. Bernier, B. Bobée, and É. Parent (2000a), Bayesian change-point analysis in hydrometeorological time series. Part 1. The normal model revisited, *J. Hydrol.*, *235*, 221–241.

Perreault, L., É. Parent, J. Bernier, and B. Bobée (2000b), Retrospective multivariate Bayesian change-point analysis: A simultaneous single change in the mean of several hydrological sequences, *Stochastic Environ. Res. Risk Assess.*, *14*, 243–261.

Peterson, T. C., et al. (1998), Homogeneity adjustments of in situ atmospheric climate data: A review, *Int. J. Climatol.*, *18*, 1493–1517.

Reeves, J., J. Chen, X. L. Wang, R. Lund, and Q. Lu (2007), A review and comparison of change-point detection techniques for climate data, *J. Appl. Meteorol. Climatol.*, *46*, 900–915.

Rissanen, J. (1989), *Stochastic Complexity in Statistical Inquiry*, World Sci., Singapore.

Robert, C. (2006), *Le Choix Bayésien: Principes et Pratiques*, Springer, Paris.

Saito, N. (1994), Simultaneous noise suppression and signal compression using a library of ortho-normal bases and the minimum description length criterion, in *Wavelets in Geophysics, Wavelet Anal. Its Appl.*, vol. 4, edited by E. Foufoula-Georgiou and P. Kumar, pp. 299–324, Academic, San Diego, Calif.

Schwarz, G. (1978), Estimating the dimension of a model, *Ann. Stat.*, *6*, 461–464.

Seidou, O., J. J. Asselin, and T. B. M. J. Ouarda (2007), Bayesian multivariate linear regression with application to change point models in hydrometeorological variables, *Water Resour. Res.*, *43*, W08401, doi:10.1029/2005WR004835.

Yang, T. Y., and L. Kuo (2001), Bayesian binary segmentation procedure for a Poisson process with multiple changepoints, *J. Comput. Graphical Stat.*, *10*, 772–785.

A. Hannart, Laboratoire d'Océanographie et du Climat: Expérimentations et Approches Numériques, IPSL, CNRS, 4 place Jussieu, Paris F-75005, France. (alexis.hannart@locean-ipsl.upmc.fr)

P. Naveau, Laboratoire des Sciences du Climat et de l'Environnement, IPSL, CEA, CNRS, Orme des Merisiers, Bat. 701 C. E. Saclay, F-91191 Gif-sur-Yvette, France.