A breakpoint detection in the mean model with heterogeneous variance on fixed time intervals

Olivier Bock, Xavier Collilieux, François Guillamon, Emilie Lebarbier, Claire Pascal

A BREAKPOINT DETECTION IN THE MEAN MODEL WITH HETEROGENEOUS VARIANCE ON FIXED TIME-INTERVALS

O. BOCK, X. COLLILIEUX, F. GUILLAMON, E. LEBARBIER AND C. PASCAL

Abstract. This work is motivated by an application for the homogeneization of GNSS-derived IWV (Integrated Water Vapour) series. Indeed, these GPS series are affected by abrupt changes due to equipment changes or environemental effects. The detection and correction of the series from these changes is a crucial step before any use for climate studies. In addition to these abrupt changes, it has been observed in the series a non-stationary of the variability. We propose in this paper a new segmentation model that is a breakpoint detection in the mean model of a Gaussian process with heterogeneous variance on known time-intervals. In this segmentation case, the dynamic programming (DP) algorithm used classically to infer the breakpoints can not be applied anymore. We propose a procedure in two steps: we first estimate robustly the variances and then apply the classical inference by plugging these estimators. The performance of our proposed procedure is assessed through simulation experiments. An application to real GNSS data is presented.

1. Introduction

Breakpoint detection aims at detecting abrupt changes, called breakpoints, in the distribution of a signal. Such problems arise in many fields, such as genomics [1, 2, 3, 4], medical [5], econometrics [6, 7, 8], geodesy [9, 10] or climate [11, 12, 13]. This massive number of applications results in an abundant literature on this subject. The motivation of our work comes from the analysis of GNSS-derived Integrated Water Vapour (IWV) series. The IWV plays a significant role in climate studies. However, these series have been affected by abrupt changes due to equipment changes, changes in processing procedure and/or changes in electromagnetic properties of the environment at the measurement site [14, 15]. A change in the mean in the signal therefore marks the presence of such an abrupt change. The statistical purpose consists thus in detecting the instants at which the mean changes in the process, that is continuous here. Many approaches have been proposed in the literature about this problem. Among them, we focus on segmentation methods. More precisely, the model of interest will be the following: the signal is supposed to be a realization of an independent Gaussian process whose parameters are affected by an unknown number of changes at unknown times. Two models can be considered, according to the characteristics of the signal that are affected by the changes: it can be either the mean of the signal only (usually called the homoscedastic model) or both the mean and the variance (usually called the heteroscedastic model), as proposed by [2] in a genomic application field or by [10] in a geodesic application for the analysis of GPS coordinates series. However, in the GNSS-IWV series, it has been observed a non-stationary of the variance due to increased variability of IWV in summer. Inspection of the annual variation of the series shows that a monthly sampling of the variance will be adequate. Consequently, the two above models will fail.
The model we propose in this work is thus a segmentation in the mean of a Gaussian process model with heterogeneous variances in the sense that the stationarity time-intervals of the variance are fixed (the months for the application).

It is now well known in segmentation framework that segmentation raises algorithmic issues due to the discrete nature of the breakpoint parameters. Indeed, the inference of these parameters requires to visit the whole segmentation space, which is prohibitive in terms of computational time when the visit is performed in a naive way. The Dynamic Programming (DP) algorithm (introduced by [16] and used for the first time in segmentation by [17]) and, recently its pruned versions [18, 19, 20], is the

Date: June 26, 2018.
2010 *Mathematics Subject Classification.* 62G05, 62M10,62P12.
Key words and phrases. Breakpoint detection; Robust estimation; GNSS time-series.

only efficient algorithm that retrieves the exact solution (i.e. the optimal segmentation according to the log-likelihood or least-square contrasts for example) in a faster way. However this algorithm can only be used if the quantity to be optimized is segment-additive (see for example [8] or [5]). In other words, a sufficient condition to satisfy this assumption is the fact that the segments are not linked both in terms of observations (i.e. independence) and parameters (i.e. no common parameters). In our case, the both stationary time-intervals of the means and the variances do not coincide. Two problems will appear: first the estimators of these two parameters will be linked and then we have no hope that DP can be applied. In order to circumvent this problem and retain the use of DP, we consider the same inference strategy as in [21] or [22] which consists in a two-step procedure: we first estimate the 'nuisance' parameters (here the variances) and then we apply the classical inference procedure by plugging these estimators.

The problem is thus reduced to the estimation of the variance parameter in a series with changes in the mean. Due to the presence of breakpoints in the series, the classical estimators for the variance will fail. Here, we follow the same idea as in [21] who proposed a robust estimator of the autocorrelation parameter for estimating breakpoints in the mean of an AR(1) process. Briefly speaking, instead of using the raw series, the idea is to work with the differenciated series that is then a zero-mean Gaussian process except at the position of the breakpoints. These points can be then seen as outliers and a robust approach can be used to obtain a good estimator of the scale parameter, as [23] proposed. We adapt in particular this estimator to our case for which, using the results of [24], we obtain asymptotic properties.
For the second step of the inference, if DP can be applied to obtain the best segmentation of the series in a given number of segments, the question arises of the choice of this number. This question has been widely investigated. In this paper, we propose to adapt the criteria proposed by [5], [25] and [26].

This paper is organized as follows: Section 2 presents the proposed segmentation model, describes the algorithmic issue for the inference and gives the outline of the proposed inference strategy. The details of this strategy are given in Section 3. More precisely, the robust estimator of the variance and the different model selection criteria for choosing the number of segments are given. A simulation study is performed in Section 4 and Section 5 is dedicated to an application of our method on GNSS-derived IWV series.

2. Model and inference issue

2.1. **Model.** We observe a series $y = \{y_t\}_{t=1,\dots,n}$ modeled by a Gaussian independent random process $Y = \{Y_t\}_{t=1\dots,n}$ such that

- $\star$: the mean of Y is affected by $K-1$ abrupt changes at some unknown instants, called breakpoints, $0 = t_0 < t_1 < \ldots < t_{K-1} < t_K = n$ and is constant between two breakpoints or within the interval $I_k^{\text{mean}} = [\![t_{k-1}+1, t_k]\!]$, denoted segment, and
- $\star$: the variance of Y is also subject to known $J-1$ changes, i.e. the variance is constant within each interval I_j^{var} and different from one to another.

The model is thus the following:

$$Y_t \sim \mathcal{N}(\mu_k, \sigma_{\mathrm{j}}^2) \quad \forall t \in I_k^{\text{mean}} \cap I_j^{\text{var}}, \tag{1}$$

for $k = 1, \dots, K$ with K is the number of segments or intervals I_k^{mean} and for $j = 1, \dots, J$ with J is the number of intervals I_j^{var}. Contrary to the heteroscedastic model, the intervals I_j^{var} and I_k^{mean} are not assumed to be the same.

2.2. **Segmentation inference: an algorithmic issue.** Parameter inference in model (1) amounts to estimating the number of segments K, the breakpoints $\boldsymbol{t} = (t_k)_{k=1,\dots,K-1}$ and the distribution parameters, i.e. the means $\boldsymbol{\mu} = (\mu_k)_{k=1,\dots,K}$ and the variances $\sigma^2 = (\sigma_j^2)_{j=1,\dots,J}$. To this end, we use a (penalized) maximum-likelihood framework and proceed as classically in segmentation inference in three steps: (i) estimate the distribution parameters, the breakpoints and their number being fixed, (ii) estimate the breakpoints for a fixed K and (iii) choose the number of segments K.

The log-likelihood of model (1) is

$$\log p(y;\boldsymbol{t},\boldsymbol{\mu},\sigma) = -\frac{n}{2}\log(2\pi) - \sum_{j=1}^{J}\frac{n_j}{2}\log(\sigma_j^2) - \frac{1}{2}SSwg_K(\boldsymbol{t},\boldsymbol{\mu},\sigma^2), \tag{2}$$

where

$$SSwg_K(\boldsymbol{t},\boldsymbol{\mu},\sigma^2) = \sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{t\in I_k^{\text{mean}}\cap I_j^{\text{var}}}\frac{(y_t-\mu_k)^2}{\sigma_j^2}, \tag{3}$$

and n_j is the length of interval I_j^{var}. Recall that in the segmentation framework, it is now well known that the step (ii) leads to a discret optimization problem and that the only efficient algorithm that retrieves the solution (exact solution in a fast way) is the Dynamic Programming algorithm (DP). This algorithm can be applied under the constraint that the quantity to be optimized is additive with respect to the segments or intervals I_k^{mean} (see for example [8], [2] or [5]). Here the optimization problem for breakpoint estimation is

$$\min_{\boldsymbol{t}\in\mathcal{M}_{K,n}}\min_{\boldsymbol{\mu}\in\mathbb{R}^K}\min_{\sigma\in\mathbb{R}^{+\mathbf{J}}} -\log p(y;\boldsymbol{t},\boldsymbol{\mu},\sigma) = \min_{\boldsymbol{t}\in\mathcal{M}_{K,n}} -\log p(y;\boldsymbol{t},\widehat{\boldsymbol{\mu}},\widehat{\sigma}),$$

where $\log p(y;\boldsymbol{t},\boldsymbol{\mu},\sigma)$ is given in (2) and $\mathcal{M}_{K,n} = \{(t_1,\ldots,t_{K-1})\in\mathbb{N}^{K-1}, 0=t_0<t_1<\ldots,t_{K-1}<t_K=n\}$ is the set of all possible partitions in K segments of the grid $[\![1,n]\!]$. The carriers of the mean parameters and the variance parameters being not the same, I_k^{mean} for μ_k and I_j^{var} for σ_j^2, two problems appear: first the estimators of these two parameters are linked, as we observe on their expressions:

$$\widehat{\mu}_k = \frac{\sum_{j=1}^{J}\sum_{t\in I_k^{\text{mean}}\cap I_j^{\text{var}}}\frac{Y_t}{\widehat{\sigma}_j^2}}{\sum_{j=1}^{J}\sum_{t\in I_k^{\text{mean}}\cap I_j^{\text{var}}}\frac{1}{\widehat{\sigma}_j^2}} \quad , \quad \widehat{\sigma}_j^2 = \frac{1}{n_j}\sum_{k=1}^{K}\sum_{t\in I_k^{\text{mean}}\cap I_j^{\text{var}}}(Y_t-\widehat{\mu}_k)^2. \tag{4}$$

Then we have no hope that $-\log p(y;\boldsymbol{t},\widehat{\boldsymbol{\mu}},\widehat{\sigma})$ will be segment-additive so DP can not be used to estimate the breakpoints.

3. Inference procedure

In order to keep possible the use of DP, we consider the same strategy as proposed by [21] or [22] which consists in

- **(1):** estimating the variance parameters (see Section 3.1), the estimators are denoted $\tilde{\sigma}_j^2$,
- **(2):** using the classical inference with 'known' variances. In this case,
  - ⋆: the mean estimators are the same as (4) where $\widehat{\sigma}_j^2$ is replaced by $\tilde{\sigma}_j^2$,
  - ⋆: the optimization problem for breakpoint estimation is

$$\begin{aligned}\min_{\boldsymbol{t}\in\mathcal{M}_{K,n}} -\log p(y;\boldsymbol{t},\widehat{\boldsymbol{\mu}},\widehat{\sigma}) &= \min_{\boldsymbol{t}\in\mathcal{M}_{K,n}}\sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{t\in I_k^{\text{mean}}\cap I_j^{\text{var}}}\frac{(y_t-\widehat{\mu}_k)^2}{\tilde{\sigma}_j^2}\\ &= \min_{\boldsymbol{t}\in\mathcal{M}_{K,n}} SSwg_K(\boldsymbol{t},\widehat{\boldsymbol{\mu}},\tilde{\sigma}^2)\\ &= SSwg_K(\widehat{\boldsymbol{t}},\widehat{\boldsymbol{\mu}},\tilde{\sigma}^2),\end{aligned} \tag{5}$$

    and DP applied.
  - ⋆: the number of segments K is chosen according to a model selection strategy which consists in maximizing a penalized log-likelihood (see Section 3.2).

3.1. **A robust estimator of the scale parameters in presence of breakpoints.** For a sake of simplicity, let us consider that the variance of the process Y is homogeneous, i.e. in model (1) we have $J=1$, $\sigma_j^2=\sigma^2$ whatever j (i.e. $I_j^{\text{var}}=[\![1,n]\!]$) and the purpose is to estimate σ. Since we need to estimate it in a series with breakpoints, classical estimators failed. The objective is to provide a robust (faced to the presence of breakpoints) estimator of σ. Following [21], the idea is to work on the differentiated series $X_t=(Y_{t+1}-Y_t)_t$ since the means of this novel series is equal to 0 except at the breakpoint positions (only $K-1$ ($K\ll n$) differences are non-centered). These latter breakpoints can then be seen as outliers and robust approaches can be used to estimate σ. [23] proposed a robust estimator of the scale parameter

of an independent Gaussian stationary process $\boldsymbol{X}$ that is proportional to the first quartile of the n^2 differences $\{|X_i - X_j|\,;\ 1 \leq i < j \leq n\}$, i.e.

$$Q_{CR,n}(\boldsymbol{X}) = c_Q \left\{|X_i - X_j|\,;\ 1 \leq i < j \leq n\right\}_{\left(\lceil \frac{1}{4} C_n^2 \rceil\right)}, \tag{6}$$

with

$$c_Q = \frac{1}{\sqrt{2}\Phi^{-1}\left(\frac{5}{8}\right)} \approx 2.2191, \tag{7}$$

to ensure the consistency of the estimator, and where Φ denotes the cumulative distribution function of a standard Gaussian random variable. The asymptotic properties of this estimator have been studied by [24] for Gaussian stationary processes with either short-range or long-range dependence.
Using this estimator, the robust estimator σ we proposed in our context is given in Proposition 3.1 for which asymptotic properties are obtained.

Proposition 3.1. *Let $(Y_t)_t$ and $(E_t)_t$ such that $Y_t = \mu_k + E_t$ if $t \in I_k = [\![t_{k-1}+1, t_k]\!]$ for $k = 1, \ldots, K$ where $(E_t)_t$ are i.i.d centered Gaussian with variance σ^2 and let further assume that $Y_0 \sim \mathcal{N}(\mu_1, \sigma^2)$. Let denote $\boldsymbol{X} = (X_t)_{t=0,\ldots,n-1} = (Y_{t+1} - Y_t)_{t=0,\ldots,n-1}$ and $(\nu_t)_{t=0,\ldots,n-1} = (E_{t+1} - E_t)_{t=0,\ldots,n-1}$. Let*

$$\tilde{\sigma}_n = Q_n(\boldsymbol{X}) = c_Q\ \frac{Q_{CR,n}(\boldsymbol{X})}{\sqrt{2}}, \tag{8}$$

where $Q_{CR,n}(\boldsymbol{X})$ and c_Q are given in (6) and (7) respectively. Then, Q_n satisfies the following Central Limit Theorem

$$\sqrt{n}(Q_n(\boldsymbol{X}) - \sigma) \xrightarrow{d} \mathcal{N}(0, \sigma'^2)\ ,\ \text{as } n \to \infty\ ,$$

where

$$\begin{aligned} \sigma'^2 &= \sigma\mathbb{E}[\mathrm{IF}^2\left(\nu_0/\sqrt{2}\sigma\right)] + 2\sigma \sum_{h\geq 1} \mathbb{E}\left[\mathrm{IF}(\nu_0/\sqrt{2}\sigma)\mathrm{IF}(\nu_h/\sqrt{2}\sigma)\right]\ , \\ \mathrm{IF}(x) &= c_Q \left(\frac{1/4 - \Phi(x + 1/c_Q) + \Phi(x - 1/c_Q)}{\int_{\mathbb{R}} \phi(y)\phi(y + 1/c_{Q,\Phi})\mathrm{d}y}\right), \end{aligned}$$

and where Φ and ϕ denote the cumulative distribution function and the probability distribution function of a standard Gaussian random variable, respectively.

Proof: since the proposed estimator is proportional to the CR's one (6), the asymptotic result is simply obtained using the results obtained by [24]: Theorem 2 is applied on $(\nu_t)_t$ with $\gamma_\nu(0) = 2\sigma^2$ and since $\sum_{h\geq 1} |\gamma_\nu(h)| < \infty$.

Note that by working on the differenciated series $(X_t)_{t=0,\ldots,n-1} = (Y_{t+1} - Y_t)_{t=0,\ldots,n-1}$, the dependence is lost but it is of short-range and remark that X_t is a Gaussian process with variance $2\sigma^2$ (that explained the normalization by $\sqrt{2}$ in (8)).

Come back to our segmentation model (1) and using Proposition 3.1, the proposed estimator of σ_j is

$$\tilde{\sigma}_{j,n} = Q_n(\boldsymbol{X}^{(j)}), \tag{9}$$

where $\boldsymbol{X}^{(j)} = (X_t^{(j)})_t = (Y_{t+1} - Y_t)_{t \in I_j^{\mathrm{var}}}$. We note $\tilde{\sigma}_n = (\tilde{\sigma}_{j,n})_j$.

3.2. **Selecting the number of segments.** In order to select the number of segments K, we consider three criteria proposed by [5], [25] and [26]. We use these criteria forgetting the fact that σ_j^2 has been estimated in a first step. The two first criteria are penalized contrast criteria which differ from the form of the penalty and depend on constants to be calibrated contrary to the last one. The penalty proposed by [5], denoted Lav, depends on the number of parameters in a model with dimension K (i.e. a segmentation with K segments) denoted D_K. It is defined as follows:

$$\mathrm{Lav}(K) = SSwg_K(\widehat{\boldsymbol{t}}, \widehat{\boldsymbol{\mu}}, \tilde{\sigma}^2) + \beta D_K, \tag{10}$$

where $SSwg_K(\boldsymbol{t}, \boldsymbol{\mu}, \sigma^2)$ is the sum of squares given in (3). $D_K = K$, the K means. The constant β is chosen using an adaptative method which involves a threshold s, taken in the simulation study and the applications to $s = 0.7$ as suggested by [5].
Applying the works of [27] in the segmentation context, [25] proposed a more complex penalty in which,

in addition to D_K, the number of possible segmentations with K segments (that is $\binom{n-1}{K-1}$) is taken into account for. This criterion is denoted BM and is defined as follows:

$$\text{BM}(K) = SSwg_K(\widehat{\boldsymbol{t}}, \widehat{\boldsymbol{\mu}}, \tilde{\sigma}^2) + \alpha \left[5D_K + 2K \log\left(\frac{n}{K}\right)\right]. \tag{11}$$

This penalty also depends on a constant α which can be calibrated in practice using the slope heuristic method proposed in [28]. More pecisely, there exists two algorithms based on this heuristic: the dimension jump algorithm and the data-driven slope estimation algorithm. We use for the simulations and the application the package R `capushe` [30] and denote BM1 and BM2 the criteria BM where the constant is calibrated using these two algorithms respectively. Note that these criteria have to be minimized.
The last criterion is a modified version of the classical BIC criterion [29] adapted by [26] to the segmentation in the mean with homogeneous variance framework, and so-called mBIC. Two versions are derived depending on the knowledge or not of the variance. Here, we considered the one for which the variance is supposed to be known, denoted mBIC,

$$\text{mBIC}(K) = -\frac{1}{2} SSwg_K(\widehat{\boldsymbol{t}}, \widehat{\boldsymbol{\mu}}, \tilde{\sigma}^2) - \frac{1}{2}\sum_{k=1}^{K} \log\left(\widehat{n}_k\right) + \left(\frac{3}{2} - K\right) \log\left(n\right), \tag{12}$$

where $\widehat{n}_k = \widehat{t}_k - \widehat{t}_{k-1}$ is the length of the kth segment of the best segmentation with K segments (i.e. of $\widehat{\boldsymbol{t}}$). Note that this criterion has to be maximized.

4. Simulation study

In order to assess the performance of our procedure, we conduct the simulation study described below. Note that we indiced the true parameters by *.

4.1. **Simulation design and quality criteria.** We use a similar design as in [21] for the segmentation parameters (breakpoint locations and means) and mimic our motivation application in the sense that the series include several years and the variance time-intervals correspond to the months. We consider series of length $n \in \{200, 800\}$ with 4 years of $n/4$ points each and 2 months by year with standard deviation $\sigma_1^\star$ and $\sigma_2^\star$ respectively. $\sigma_1^\star$ is fixed to 0.5 and $\sigma_2^\star$ varies from 0.1 to 1.5 by step of 0.2. The series are affected by 6 breakpoints ($K^\star = 7$, the true number of segments) located at positions $\boldsymbol{t}^\star = (27, 38, 88, 111, 150, 183)$ for $n = 200$ and $4 \times \boldsymbol{t}^\star$ for $n = 800$. The mean within each segment alternates between 0 and 1, starting with $\mu_1 = 0$. Each configuration is simulated 100 times.

Moreover, in order to exhibit the need of a new segmentation model for our motivated application, we compare our segmentation method, called MFixedHetero, with the two more classical segmentation models (see for example [2]) with

- $\star$: changes in the mean with homogeneous variance, called MHomo: Y_t *ind.* $\sim \mathcal{N}(\mu_k, \sigma^2)$ if $t \in I_k = [\![t_{k-1} + 1, t_k]\!]$,
- $\star$: changes in both the mean and the variance, called MHetero: Y_t *ind.* $\sim \mathcal{N}(\mu_k, \sigma_k^2)$ if $t \in I_k = [\![t_{k-1} + 1, t_k]\!]$.

In order to evaluate the performance of our proposed method, we use the following criteria:

- $\star$: the difference between the estimated standard deviation and the true one, $\tilde{\sigma}_{\bullet,n} - \sigma_\bullet^\star$,
- $\star$: the difference between the estimated number of segments and the true one, $\hat{K} - K^\star$,
- $\star$: the two components of the Hausdorff distance $d_1(\boldsymbol{t}^\star, \hat{\boldsymbol{t}})$ and $d_2(\boldsymbol{t}^\star, \hat{\boldsymbol{t}})$ where

$$d_1(a, b) = \max_b \min_a |a - b|,$$

and $d_2(a, b) = d_1(b, a)$, in order to study the quality of the estimation of the breakpoint locations. A perfect segmentation results in both null d_1 and d_2. Under-segmentation results in a small d_1 and a large d_2, provided that the estimated breakpoints are correctly located.

4.2. **Results.**

Estimation of $\sigma_1^\star$ and $\sigma_2^\star$. Figure 1 presents the proposed estimator for the two variances $\sigma_1^\star$ and $\sigma_2^\star$. We observe that it performs well to estimate the variances resulting in a similar performance in terms of segmentation estimation (see Figures 2(a) and 3 for the selection of K, and Figure 4 for the locations of the breakpoints when the variances are estimated or fixed to the true values). We can also note that the accuracy of the variance estimations increases with the length of the series n.

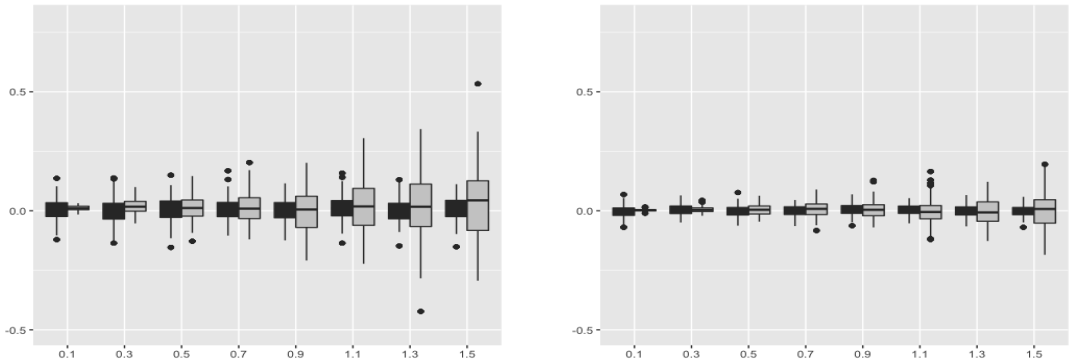

FIGURE 1. Boxplots of $\tilde{\sigma}_{1,n} - \sigma_1^\star$ in black and $\tilde{\sigma}_{2,n} - \sigma_2^\star$ in grey for different values of $\sigma_2^\star$ with $n = 200$ (left) and $n = 800$ (right).

Segmentation estimation for MFixedHetero. Only the results for $n = 200$ are presented here, the results for $n = 800$ leading to the same conclusions.

Figure 2(a) compares the estimated number of segments obtained with the considered model selection criteria for different noise levels of $\sigma_2^\star$. The two components of the Haussdorff distance (d_1 and d_2) calculated on the obtained segmentations are plotted in Figures 2(b) and 2(c) respectively. These distances are also computed for the optimal segmentations with the true number of segments (on the same figures). In addition, the histograms of breakpoint locations are given in Figure 4 for three values of $\sigma_2^\star$ when the number of segments is selected using mBIC or fixed to the true value and when the variances are estimated or not (the other criteria giving similar results).

First recall that in a segmentation in the mean context, it has been observed that when the noise is small, the detection problem is easy and the procedure detects the true breakpoints. However, when the problem gets difficult (large variance), the procedure tends to underestimate the number of segments in order to avoid the detection of false breakpoints (see for example [21]). In our simulation design, among the six breakpoints, four belong to an interval with variance $\sigma_2^{2\star}$, the fourth one, $t_4^\star$, belongs to an interval with variance $\sigma_1^{2\star}$ and the fifth one $t_5^\star$ corresponds to both a change in the mean and the variance. We thus observe that our procedure performs as expected and whatever the model selection criteria. First, the variance $\sigma_2^{2\star}$ does not alter the detection of the breakpoint $t_4^\star$. When $\sigma_2^{2\star}$ is small, all the true breakpoints are recovered with a less of accuracy for $t_4^\star$ and $t_5^\star$, and when $\sigma_2^{2\star}$ becomes large, the procedure tends to underestimate the number of segments with estimated breakpoints that are close to the true ones (d_1 smaller compared to the segmentations with the true number of segments). For a very high value of $\sigma_2^{2\star}$, almost only $t_4^\star$ is detected. We can also observe that our method performs as well as when the variances are known. Moreover, even if the different criteria for selecting the number of segments show a global same behaviour, there exist some slight differences: BM1 fails when the detection problem is very easy due to the calibration heuristic, BM2 tends to detect a little more number of segments compared to the other criteria when the variance is large.

Comparison with models MHomo and MHetero. Figure 5 displays the boxplots of the number of segments selected by Lav, BM and mBIC for models MHomo and MHetero and Figures 6 and 7 give the histograms of the breakpoint locations for the different model selection criteria and three values of $\sigma_2^\star$,

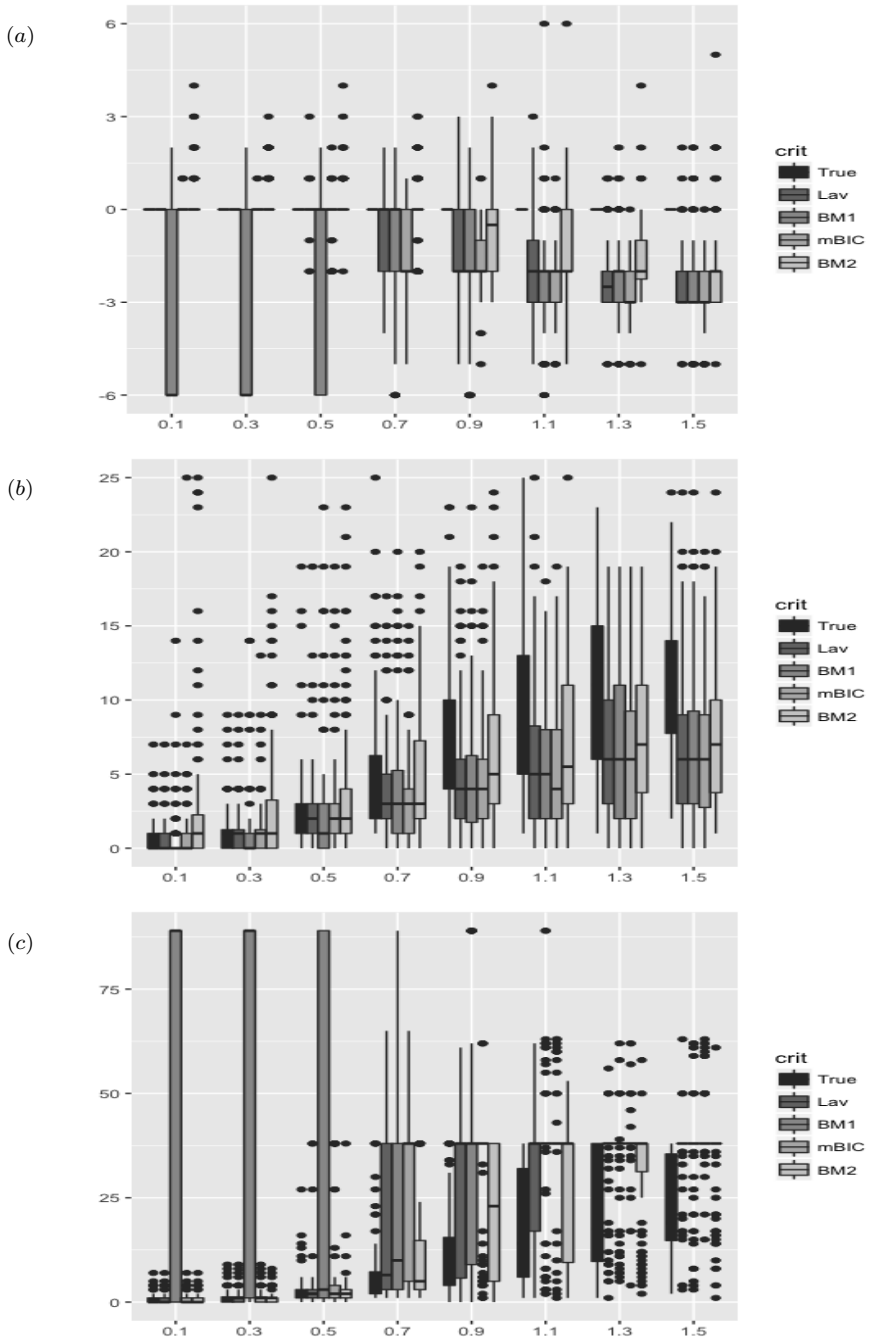

FIGURE 2. Boxplots of (a) $\hat{K} - K^\star$, (b) the first component of the Hausdorff distance (d_1) and (c) the second component of the Hausdorff distance (d_2) for $n = 200$ and for different values of $\sigma_2^\star$, obtained for MFixedHetero and the different model selection criteria.

obtained with MHomo and MHetero respectively. Note that the imposed changes of variance are located at the positions $25, 50, 75, 100, 125, 150, 175$ and the true breakpoints at $27, 38, 88, 111, 150, 183$. Logically MHomo, MHetero and MFixedHetero lead to close performances in terms of segmentation when the two variances are close, even if for MHetero we can observe an overestimation by mBIC (Figure 7 (b-left) or Figure 5 (right)) and a less of accuracy with Lav and BM2. With model MHetero, as expected,

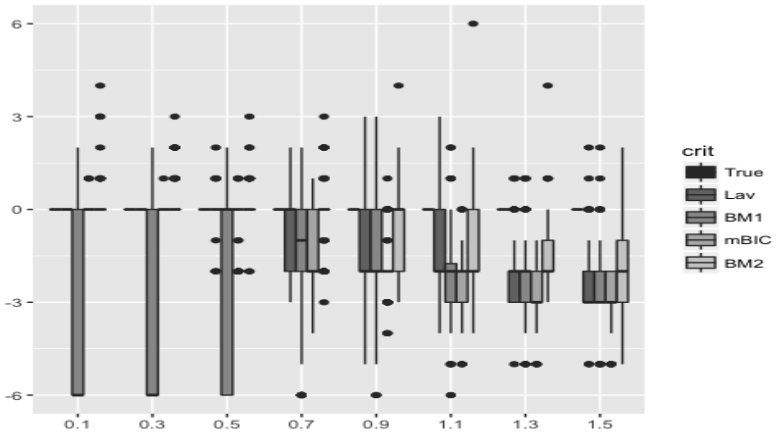

FIGURE 3. Boxplots of $\hat{K} - K^\star$ for $n = 200$ and for different values of $\sigma_2^\star$, obtained for MFixedHetero with the true values of $\sigma_1^\star$ and $\sigma_2^\star$ and the different model selection criteria.

the changes in the variance are also detected, with more difficulty compared to the detection of the changes in the mean. This explained the overestimation of the estimated number of segments. This is more marked with mBIC. Model MHomo behaves similarly as model MFixedHetero, except when the variance is too large (Figure 6 (c)). In this latter case, MFixedHetero can be able to detect the fourth breakpoint $t_4^\star$ contrary to MHomo for which the estimated standard deviation is larger than 0.5 in the corresponding interval (1.27 in average).

5. Application to GNSS-derived Integrated Water Vapour series

Context and data description. The GNSS-derived IWV series are used to study and verify climate model predictions of atmospheric water vapour trends and variability connected to climate change (global warming) [31]. The biases induced by the abrupt changes are small and thus difficult to distinguish from the natural climate variation of the measured IWV signal [15]. The most commonly used approach is the relative abrupt change detection which compares the candidate series to one or several reference series (e.g. from nearby stations) which are assumed to contain nearly the same climate signal [32, 11]. In the case of our application, the stations in the global GNSS network are usually too far from each other to remove completely the climate signal in the differences. Instead, we extract the reference time series for each station from a gridded global atmospheric model reanalysis. In this work, we use the ECMWF reanalysis, ERA-Interim (ERAI) [33]. The considered series are thus the differences between the GNSS-IWV signal and the ERA-Interim one, resulting in so called GNSS-ERAI series.

In this application we consider the two stations SYOG (Syowa, Antarctica) and ONSA (Onsala, Sweden) contributing to the International GNSS Service (IGS) network of continuously operating reference stations (www.igs.org). The IWV data retrieved from these GPS measurements are described in [31]. In the present work, the IWV data series are used with daily time sampling. The equipment changes are available from the so-called IGS sitelogs and are given in Table 1.

Model (1) **for this application.** For these series, the variance time-intervals correspond to the different months, i.e.

- $\star$: $j =$ month, $J = 12$ and each interval $I^{\text{var}}_{\text{month}}$ is the union of several intervals among the considered years, $I^{\text{var}}_{\text{month}} = \bigcup_{\text{year}} I^{\text{var}}_{\text{year,month}}$ where $I^{\text{var}}_{\text{year,month}}$ is the time-interval of the month 'month' of the year 'year'.
- $\star$: σ_j^2 is estimated by $Q_n(\boldsymbol{x}_{\text{month}})$ with $\boldsymbol{x}_{\text{month}} = ((y_{t+1} - y_t)^{\text{year}}_{date(t) \text{ and } date(t+1) \in \text{month}})_{\text{year}}$, i.e. the differentiated series of the considered month of all the years, and where Q_n is defined by (8).

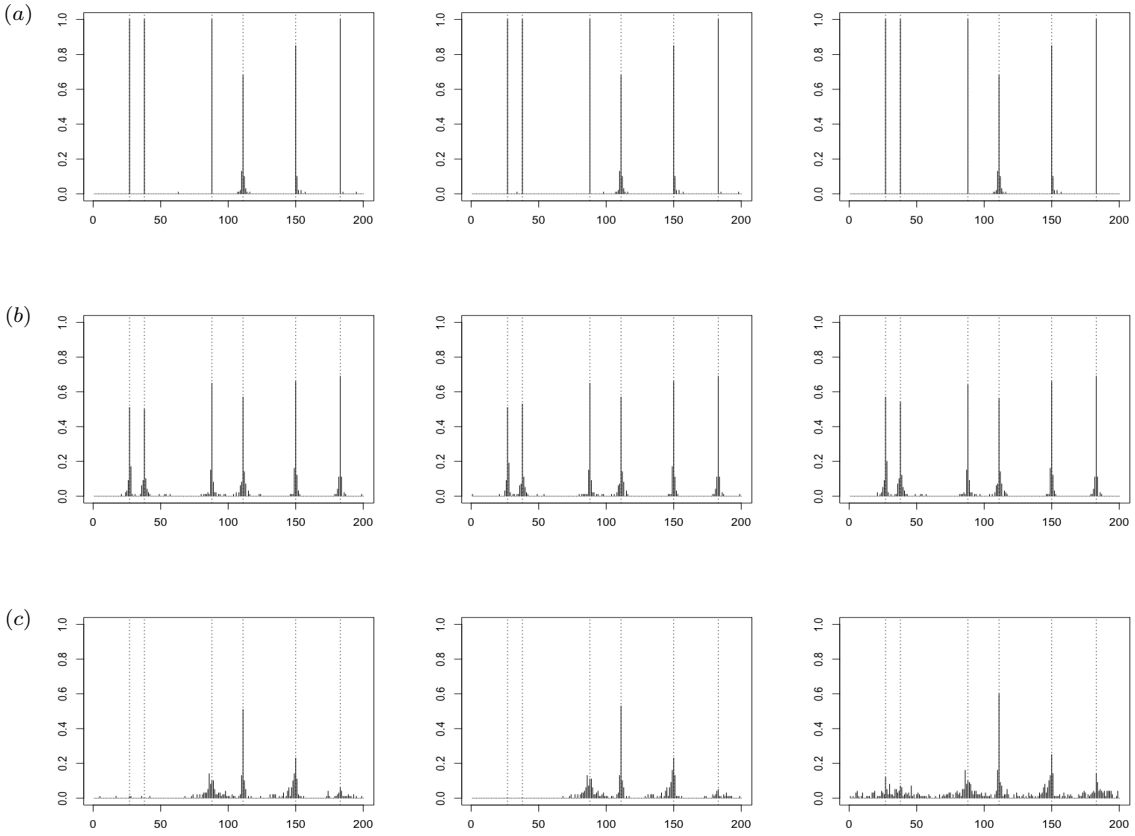

FIGURE 4. Frequencies of each possible breakpoint for MFixedHetero when the number of segments is selected with mBIC and the variances are estimated (left), when the number of segments is selected with mBIC and the variances are the true values (middle) and when the number of segments is true ($K = 7$) and the variances are estimated (right), with $n = 200$. The value of $\sigma_2^\star$ is fixed to 0.1 (a), 0.5 (b) and 1.5 (c). The dotted lines correspond to the true breakpoint locations.

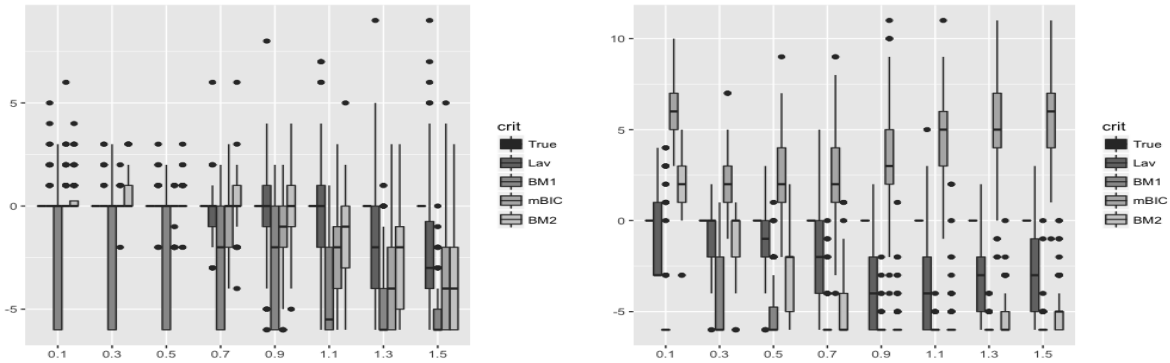

FIGURE 5. Boxplots of $\hat{K} - K^\star$ for $n = 200$ and different values of $\sigma_2^\star$ with MHomo (left) and MHetero (right) using the different model selection criteria.

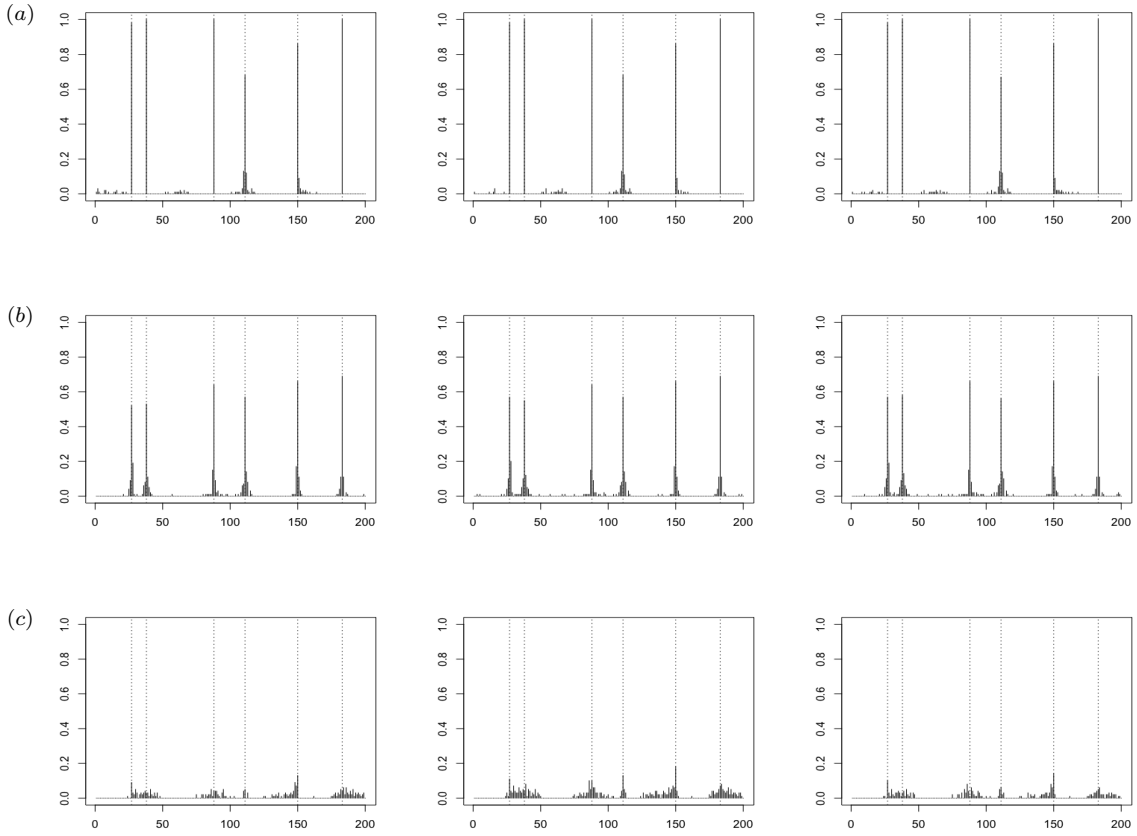

FIGURE 6. Frequencies of each possible breakpoint for MHomo when the number of segments is selected with the criteria mBIC (left), Lav (middle) and BM2 (left), with $n = 200$. The value of $\sigma_2^\star$ is fixed to 0.1 (a), 0.5 (b) and 1.5 (c). The dotted lines correspond to the true breakpoint locations.

Results. For the SYOG series, all the criteria select four breakpoins, except for mBIC that selects 81 ones. The results are plotted in Figure 8. All the four breakpoints correspond (exactly for dates 2008-03-31 and 2009-03-26 and are close for dates 1999-12-16 and 2007-02-15) to known equipment changes (the dashed lines (in black)). This segmentation is also obtained by both the models MHomo and MHetero with BM2. This can be explained by the fact that the monthly variances are quite similar (see the estimated standard deviation of each month Figure 8 (middle)).

The results for the series ONSA are given in Figure 9. The criteria select different number of segments: $\hat{K} = 2$ for Lav and BM2, $\hat{K} = 15$ for BM1 and $\hat{K} = 76$ for mBIC. The big abrupt change at date 1999-02-04 is always detected and is associated to a change in receiver, antenna and radome. When $\hat{K} = 15$, only one break corresponds to a known change and two others are close. Contrary to the previous series, the estimated monthly variances are different (higher in summer) resulting in a different segmentation for models MHomo and MHetero (see Figure 9 (c) where the criterion Lav is considered). The breakpoint at date 1999-02-04 is detected with the both. However we observe an overestimation of the number of breakpoints and the estimated breakpoints are clearly not linked to known equipment changes. Note that all equipment changes do not impact the time series [15].

For both series, we observe an overestimation of the number of segments when using the mBIC criterion. By looking to the estimated means (Figure 8 (middle) and Figure 9 b-left), this overestimation links to

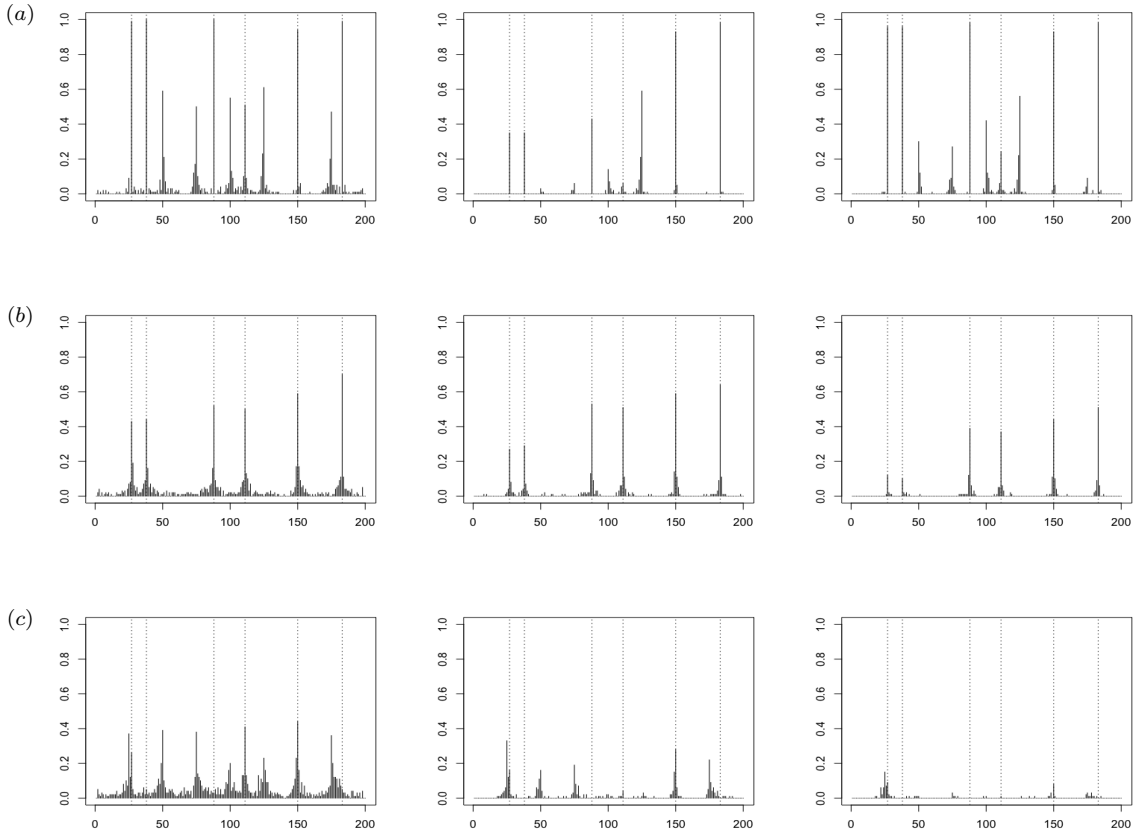

FIGURE 7. Frequencies of each possible breakpoint for MHetero when the number of segments is selected with the criteria mBIC (left), Lav (middle) and BM2 (left), with $n = 200$. The value of $\sigma_2^\star$ is fixed to 0.1 (a), 0.5 (b) and 1.5 (c). The dotted lines correspond to the true breakpoint locations and the changes of variances are fixed at locations $25, 50, 75, 100, 125, 150, 175$.

the detection of outliers and seems to capture a periodic signal. This latter point can be due to the fact that a periodic tendency remains despite the correction by the ERAI model [31].

REFERENCES

[1] Braun JV, Braun R, Müller HG: **Multiple changepoint fitting via quasilikelihood, with application to DNA sequence segmentation**. *Biometrika* 2000, **87**(2):301–314.

[2] Picard F, Robin S, Lavielle M, Vaisse C, Daudin JJ: **A statistical approach for CGH microarray data analysis**. *BMC Bioinformatics* 2005, **6**:27.

[3] Cleynen A, Dudoit S, Robin S: **Comparing segmentation methods for genome annotation based on rna-seq data**. *Journal of Agricultural, Biological, and Environmental Statistics* 2014, **19**:101–118.

[4] Lévy-Leduc C, Delattre M, Mary-Huard T, Robin S: **Two-dimensional segmentation for analyzing Hi-C data**. *Bioinformatics* 2014, **30**(17):i386–i392.

[5] Lavielle M: **Using penalized contrasts for the change-point problem**. *Signal Processing* 2005, **85**(8):1501–1510.

[6] Lai TL, Liu H, Xing H: **Autoregressive models with piecewise constant volatility and regression parameters**. *Statistica Sinica* 2005, **15**:279–301.

[7] Lavielle M: **Detection of multiple changes in a sequence of dependent variables**. *Stochastic Processes and their Applications* 1999, **83**:79–102.

[8] Bai J, Perron P: **Computation and analysis of multiple structural change models**. *J. Appl. Econ.* 2003, **18**:1–22.

Series	known changes
SYOG	1995-03-15 (RA)
	1996-01-17 (R)
	1999-12-24 (R)
	2000-02-03 (R)
	2002-01-26 (R)
	2007-01-25 (R)
	2008-03-31 (P)
	2009-03-26 (P)
ONSA	1999-02-01 (RAD)
	1999-07-03 (R)
	2003-08-19 (R)
	2004-03-10 (R)
	2007-11-01 (R)
	2008-03-31 (P)
	2008-05-15 (R)
	2009-03-26 (P)

TABLE 1. Known changes in the two considered series. All changes corresponds to a change of receiver (R), antenna (A), radome (D), or processing (P). RA and RAD indicate combined changes.

[9] Gazeaux J, Williams S, King M, Bos M, Dach R, Deo M, Moore AW, Ostini L, Petrie E, Roggero M, Teferle FN, Olivares G, Webb FH: **Detecting offsets in GPS time series: First results from the detection of offsets in GPS experiment**. *Journal of Geophysical Research (Solid Earth)* 2013, **118**:2397–2407.

[10] Gazeaux J, Lebarbier E, Collilieux X, Métivier L: **Joint segmentation of multiple GPS coordinate series**. *Journal de la Société Française de Statistique* 2015, **156**(4):163–179.

[11] Caussinus H, Mestre O: **Detection and correction of artificial shifts in climate series**. *Applied Statistics* 2004, **53**:405–425.

[12] Mestre O, Domonkos P, Picard F, Auer I, Robin S, Lebarbier E, Böhm R, Aguilar E, Guijarro JA, Vertacnik G, et al.: **HOMER: a homogenization software–methods and applications** 2013.

[13] Lu Q, Lund R, Lee T: **An MDL approach to the climate segmentation problem**. *The Annals of Applied Statistics* 2010, **4**:299–319.

[14] Vey S, Dietrich R, Fritsche M, Rlke A, Steigenberger P, Rothacher M: **On the homogeneity and interpretation of precipitable water time series derived from global GPS observations**. *Journal of Geophysical Research: Atmospheres* 2009, **114**(D10).

[15] Ning T, Wang J, Elgered G, Dick G, Wickert J, Bradke M, Sommer M, Querel R, Smale D: **The uncertainty of the atmospheric integrated water vapour estimated from GNSS observations**. *Atmospheric Measurement Techniques* 2016, **9**:79.

[16] Bellman R: **The theory of dynamic programming**. *Bulletin of the American Mathematical Society* 1954, **60**(6):503515.

[17] Auger I, Lawrence C: **Algorithms for the optimal identification of segments neighborhoods**. *Bull Math Biol* 1989, **51**:3954.

[18] Killick R, Fearnhead P, Eckley IA: **Optimal detection of changepoints with a linear computational cost**. *Journal of The American Statistical Association* 2012, **107**(500):1590–1598.

[19] Rigaill G: **A pruned dynamic programming algorithm to recover the best segmentations with** 1 **to** K_{max} **change-points**. *Journal de la Socit Franaise de Statistique* 2015, **156**(4):180–205.

[20] Maidstone R, Hocking T, Rigaill G, Fearnhead P: **On Optimal Multiple Changepoint Algorithms for Large Data**. *arXiv eprint 1409.1842* 2014.

[21] Chakar S, Lebarbier E, Levy-Leduc C, Robin S: **A robust approach for estimating change-points in the mean of an AR(1) process**. *To appear in Bernoulli (arXiv:1403.1958)* 2015.

[22] Cleynen A, Robin S: **Comparing change-point locations of independent profiles**. *Statistics and Computing* 2014.

[23] Rousseeuw PJ, Croux C: **Alternatives to the Median Absolute Deviation**. *Journal of the American Statistical Association* 1993, **88**(424):1273–1283, [http://www.tandfonline.com/doi/abs/10.1080/01621459.1993.10476408].

[24] Lévy-Leduc C, Boistard H, Moulines E, Taqqu MS, Reisen VA: **Robust estimation of the scale and of the autocovariance function of Gaussian short-and long-range dependent processes**. *Journal of Time Series Analysis* 2011, **32**(2):135–156.

[25] Lebarbier E: **Detecting Multiple Change-Points in the Mean of Gaussian Process by Model Selection**. *Signal Processing* 2005, **85**:717–736.

[26] Zhang NR, Siegmund DO: **A Modified Bayes Information Criterion with Applications to the Analysis of Comparative Genomic Hybridization Data**. *Biometrics* 2007, **63**:22–32.

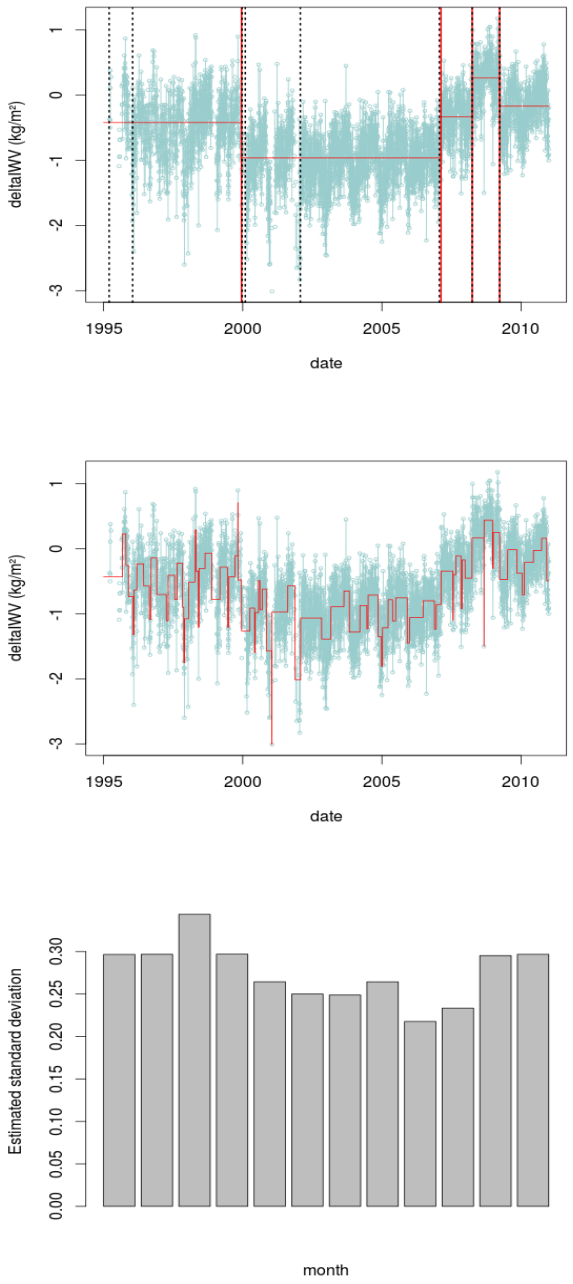

FIGURE 8. Results for the series SYOG: the estimated breakpoint with $\hat{K} = 5$ (top), the series with the estimated mean with $\hat{K} = 82$ (middle) and the estimated standard deviation for each month (bottom). Solid lines (in red): the estimated breakpoints and the fitted expectation. Dashed lines (in black): known equipment changes (see Table 1).

[27] Birg L, Massart P: **Gaussian model selection**. *Journal of the European Mathematical Society* 2001, **3**:203–268.
[28] Arlot S, Massart P: **Data-driven calibration of penalties for least-squares regression**. *J. Mach. Learn. Res.* 2009, **10**:245–279 (electronic), [http://www.jmlr.org/papers/volume10/arlot09a/arlot09a.pdf[pdf]].
[29] Schwarz G: **Estimating the dimension of a model**. *Ann. Statist.* 1978, **6**:461464.

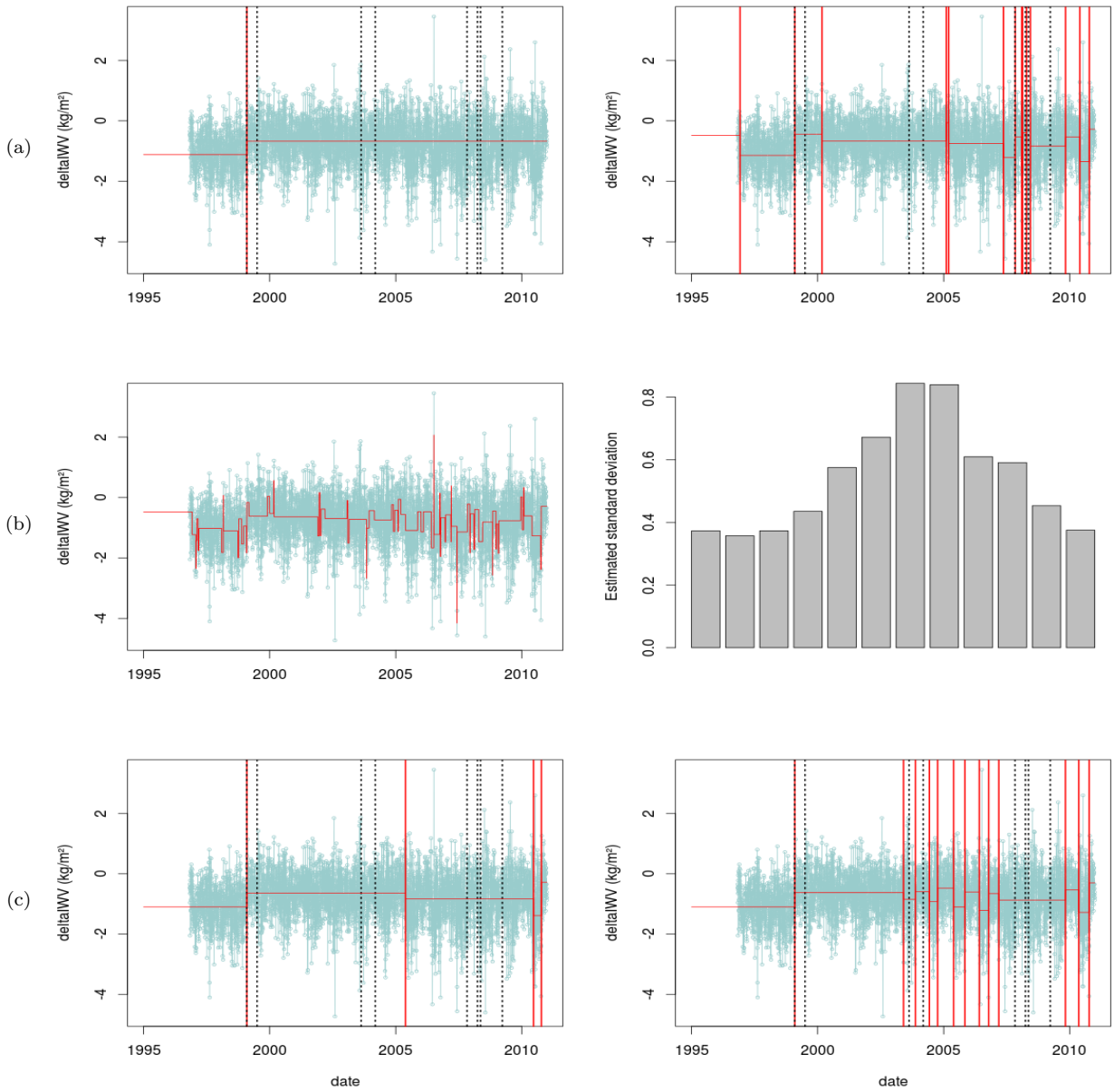

FIGURE 9. Results for the series ONSA. (a) the estimated breakpoints with $\hat{K} = 2$ (left: Lav and BM2) and $\hat{K} = 15$ (right: BM1). (b) the estimated mean with $\hat{K} = 74$ (left: mBIC) and the estimated standard deviation for each month (right). (c) the estimated breakpoints obtained with model MHomo (left: $\hat{K} = 5$ with Lav) and with model MHetero (left: $\hat{K} = 14$ with Lav). Solid lines (in red): the estimated breakpoints and the fitted expectation. Dashed lines (in black): known equipment changes (see Table 1).

[30] Baudry JP, Maugis C, Michel B: **Slope heuristics: overview and implementation**. *Statistics and Computing* 2011, **22**(2):455470.

[31] Parracho AC, Bock O, Bastin S: **Global IWV trends and variability in atmospheric reanalyses and GPS observations**. *Atmospheric Chemistry and Physics Discussions* 2018, :Under–Review.

[32] Lindau R, Venema V: **On the multiple breakpoint problem and the number of significant breaks in homogenization of climate records**. *Idojaras, QJ Hung. Meteorol. Serv* 2013, **117**:1–34.

[33] Dee DP, Uppala S, Simmons A, Berrisford P, Poli P, Kobayashi S, Andrae U, Balmaseda M, Balsamo G, Bauer dP, et al.: **The ERA-Interim reanalysis: Configuration and performance of the data assimilation system**. *Quarterly Journal of the royal meteorological society* 2011, **137**(656):553–597.

E-mail address: olivier.bock@ign.fr,Xavier.Collilieux@ign.fr, francois.guillamon@agroparistech.fr,
E-mail address: emilie.lebarbier@agroparistech.fr, claire.pascal@agroparistech.fr