Development of new homogenisation methods for GNSS atmospheric data. Application to the analysis of climate trends and variability.

Annarosa Quarello

Sorbonne Université

ED 129 Sciences de l'Environnement

Equipe Géodésie de l'IPGP

Développement de nouvelles méthodes d'homogénéisation des données atmosphériques GNSS. Application à l'étude de la variabilité climatique.

Development of new homogenisation methods for atmospheric GNSS data. Application to the analysis of climate trends and variability.

Par Annarosa Quarello

Thèse de doctorat de l'Environnement

Dirigée par Olivier Bock et Emilie Lebarbier

Présentée et soutenue publiquement le 15/12/2020

Devant un jury composé de :

M. Guillem Rigaill	*Docteur, INRAE*
M. Enric Aguilar	*Docteur, Universitat Rovira i Virgili*
M. Olivier Mestre	*Docteur, Meteo-France*
Mme Céline Levy-Leduc	*Professeur, AgroParisTech - Univ. Paris Saclay*
Mme Laurence Picon	*Professeur, LMD*
M. Olivier Bock	*Docteur, IPGP – IGN*
Mme Emilie Lebarbier	*Professeur, Université Paris Nanterre*

Abstract

Homogenization is an important and crucial step to improve the use of observational data for climate analysis. This work is motivated by the analysis of long GNSS Integrated Water Vapor (IWV) data which have not yet been used in this context. These series are affected by inhomogeneities linked to changes in the instrumentation, in the environment, and in the data processing procedure. Because the natural variability of the IWV series is quite large, we actually work on the time series of differences, using ERA-Interim reanalysis as reference for removing the climate signal. A base assumption is that the differences contain only the signature of the abrupt changes from the GNSS series which can be detected by means of a segmentation algorithm followed by a semi-automatic attribution/validation procedure using metadata (known equipment changes).

The main contribution of this thesis is the development of a novel segmentation method in a frequentist approach, dedicated to detecting changes in the mean of the GNSS-ERA-Interim IWV difference series. We introduce a parametric model that includes a periodic bias and a heterogeneous, monthly varying, variance which fit the characteristics of the IWV difference series. The method consists of first estimating the variance using a robust estimator and then estimating the segmentation parameters (the positions of the change-points, the means of the segments) and the periodic bias model in a sequential way. The segmentation parameters and the periodic bias model are estimated iteratively for a fixed number of segments. The inference is achieved by the classical maximum likelihood procedure using a dynamic programming algorithm for the estimation of the segmentation parameters which provides the exact solution in a reasonable amount of time. The procedure is repeated for all the numbers of segments tested between 1 and a maximum of 30. Finally, the optimal number of segments is chosen using a penalized model selection strategy. Several criteria are tested.

The method is implemented in the R package GNSSseg available on CRAN. A faster version GNSSfast was also made available more recently on GitHub.

The performance of the proposed method was evaluated by numerical simulations. An application for a real dataset of 120 global GNSS stations in the global IGS network is presented for the period from January 1995 to December 2010. Inspection of the results reveals that the detected change-points contain a fraction ($\tilde{2}0\%$) of outliers which are characterized by double detections with two large offsets, generally of opposite signs, close together, e.g. a few tens of days apart. In order to detect and eliminate the outliers a screening method was developed. The final set of change-points is compared to GNSS metadata which contain information on equipment changes that occurred at the stations. The percentage of attribution remains moderate at the level of 20 %. Some of the change-points may actually be undocumented or due to the reference series (ERA-Interim).

Finally, the segmentation information (dates of the change-points) is included in a linear regression algorithm which is used to estimate the GNSS IWV trends. The estimated linear trends are tested for significance and compared to the ERA-Interim trends. Higher spatial consistency in the GNSS trends and improved consistency is found with ERA-Interim after homogenisation in regions where the reanalysis is known to perform well.

Several options are discussed to further improve the homogenisation method, such as alternative segmentation models including autocorrelation in the series and more complex bias functions. Another issue is the correct attribution of the detected change-points which may actually be due to the reference series. Validations involving dual GNSS comparisons would be useful when the network density permits. The current method already significantly improves the homogeneity of the GNSS series and may readily be used for climate trends and variability analysis.

Contents

List of Figures

List of Tables

Chapter 1

Résumé long

1.1 Introduction

Les séries longues de données sont essentielles pour l'étude, la compréhension et la modélisation des processus météorologiques et climatiques globaux. Cependant, ces séries sont souvent affectées par des inhomogénéités dues aux changements des instruments de mesure ou des observateurs, au déplacement de la station et aux changements d'environnement autour de la station (Jones *et al.* [1986]). Ces inhomogénéités se manifestent généralement par des changements abrupts dans les séries, rendant l'estimation des tendances et variabilités climatiques peu précises ou biaisées (Thorne *et al.* [2005]). L'homogénéisation de ces séries est donc une étape cruciale. Cette homogénéisation consiste à (1) détecter les changements abrupts ou ruptures ; (2) valider ces ruptures, i.e. séparer les "vraies" détections des "fausses" à l'aide de Metadata ; (3) corriger les séries de ces ruptures avant ou pendant l'estimation des tendances.

Dans cette thèse, nous sommes intéressés à un nouveau type de données : les données journalières de Contenu Intégré en Vapeur d'Eau (CIVE) mesurées par GNSS (Global Navigation Satellite Systems), appelées GNSS CIVE (IWV Integrated Water Vapor en anglais), plus précises que les mesures par radiosondes auparavant réalisées. Les études portant sur l'homogénéité de ce nouveau type de données sont récentes et peu nombreuses (Bock *et al.* [2010]; Ning *et al.* [2016]; Parracho *et al.* [2018]; Vey *et al.* [2009]). Dans ces séries, une forte variabilité naturelle a été observée rendant la détection de ces ruptures difficile. Afin de palier à ce problème et dû au fait que nous ne disposons pas de séries proches, nous avons utilisé la réanalyse ERA-Interim (Dee *et al.* [2011]) comme référence qui représente bien la variabilité atmosphérique (Parracho *et al.* [2018]) : nous considérons les séries de différence $\Delta CIVE$ entre le GNSS et les données de réanalyse ($\Delta CIVE = CIVE_{GPS} - CIVE_{ERAI}$). Dans ce travail, nous disposons d'un ensemble de données GNSS issu du réseau IGS repro1 (Bock [2017]) représenté sur la Figure 1.1). Plus précisément, nous avons les séries GNSS CIVE de 120 stations pour la période du 1er janvier 1995 au

31 décembre 2010.

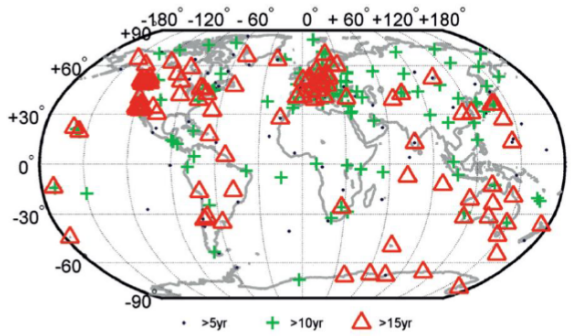
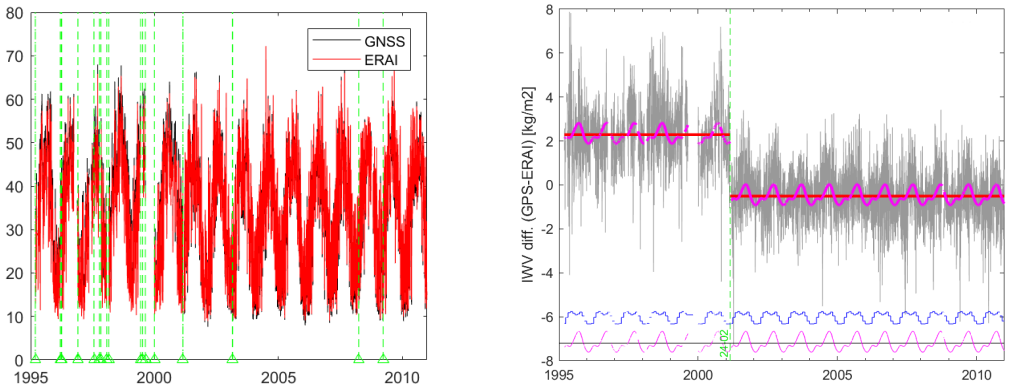


FIGURE 1.1 – Répartition des 460 stations GNSS disponibles à partir du jeu de données IGS repro1 couvrant la période du 1er janvier 1995 au 31 décembre 2010. Les différents marqueurs représentent la longueur de la série. Parmi les 460 stations, 120 sont des séries de plus de 15 ans. La source : Bock [2014].

Afin d'illustrer l'intérêt d'analyser la série de différence, regardons une série particulière qui est la série GNSS CIVE de la station CCJM (station tropicale située au nord de la mer des Philippines). La Figure 1.2 (a) présente cette série (en noir). On observe une variation saisonnière marquée avec des valeurs variants de 10 kgm^{-2} à 60 kgm^{-2} entre l'hiver et l'été, ainsi qu'une forte variabilité liée à l'évolution des conditions météorologiques. Sur cette Figure est aussi représentée la série de la réanalyse ERA-Interim associée (en rouge). Les deux jeux de données semblent en accord. Les lignes pointillées verticales donnent les changements d'équipement connus qui comprennent les changements de récepteur et d'antenne tels que trouvés dans les fichiers logs du site IGS, ainsi que deux changements de traitement en 2008 et 2009. Une inspection visuelle de la série permettant de voir si un changement connu induit une rupture dans la série n'est clairement pas évidente. La Figure 1.2 (b) présente la série de différences $\Delta CIVE$. On voit maintenant clairement apparaître une rupture le 24 février 2001 qui est associée à un changement de récepteur et d'antenne de la station CCJM. Les autres changements connus ne semblent pas produire de ruptures. En suivant la même approche que Lu & Lund [2007], nous avons ajusté un modèle de changement de moyenne à la série $\Delta CIVE$ avec un changement connu (sans y inclure de tendance puisque nous travaillons sur la différence), en rouge sur la Figure. De plus, suivant l'approche proposée par Collilieux *et al.* [2019], nous avons modélisé la présence d'un biais périodique par une série de Fourier d'ordre 4 avec une période de base de 1 an (365,25 jours) et des harmoniques de 1/2, 1/3 et 1/4 d'année (en magenta sur la Figure). Le saut dans les moyennes est estimé à 2,8 kgm^{-2}. La raison de la présence d'un biais périodique malgrè le fait que l'on travaille sur la différence s'explique par une différence de représentativité entre les deux ensembles de données. En effet, les observations GNSS peuvent capturer une certaine variabilité à petite échelle non résolue par la réanalyse (Bock & Parracho [2019]). Bien que la réanalyse soit la meilleure référence que nous puissions avoir, elle n'est pas

parfaite et ces différences doivent être prises en compte dans le modèle de segmentation afin d'éviter la sur-détection de ruptures dûe à la présence de ce biais périodique. Un deuxième point qui caractérise les différences CIVE est la variation annuelle de la variance (Bock & Parracho [2019]). Cette caractéristique est mise en évidence par la ligne bleue dans la Figure 1.2 (b) qui représente l'écart-type mensuel des résidus quotidiens après ajustement.

L'objectif principal de cette thèse était de développer un nouveau modèle de détection de ruptures ou segmentation adapté à ces deux particularités de la série temporelle de différence GNSS CIVE : un biais périodique et une variance mensuelle.



(a) (b)

FIGURE 1.2 – (*a*) Séries temporelles GNSS et ERA-Interim CIVE de la station CCJM située dans la mer des Philippines ($27N, 142E$). Les lignes vertes verticales sont les changements d'équipement documentés dans les métadonnées. (*b*) Différence CIVE (GPS - ERA-Interim) en gris. Un seul changement a été retenu a priori pour ajuster un modèle : fonction constante par morceau (superposé au signal, en rouge) + série de Fourier d'ordre 4 (en magenta en bas de graphique), le modèle ajusté est superpose en magenta sur le signal. La courbe bleue en bas de graphique montre l'écart-type des fluctuations mensuelles.

La détection de changements abrupts ou ruptures est un domaine important et très actif en statistique. L'objectif consiste à identifier des instants, positions ou dates où les propriétés statistiques des données sont différentes avant et après ces instants, typiquement la distribution. Ces ruptures délimitent ce qu'on appelle des segments. Une synthèse de l'état de l'art des méthodes de détection de ruptures, appelées aussi méthodes de segmentation, développées dans le domaine spécifique du climat est donnée dans le Tableau 2.1. La grande majorité des méthodes sont paramétriques (basées sur un modèle paramétrique par opposition à une approche non paramétrique) et fréquentiste (estimation ponctuelle par opposition à une approche bayésienne). Dans ce contexte paramétrique et fréquentiste, les deux grands types de méthodes proposées sont : (1) basées sur des tests statistiques de détection d'une rupture

(détection locale) et la méthode consiste à détecter les ruptures les unes après les autres (la solution n'est pas exacte du point de vue de la détection globale), et (2) basées sur une recherche simultanée des ruptures en utilisant des critères de vraisemblance pénalisée. Dans cette dernière approche, on peut faire une distinction selon l'exactitude ou pas de l'algorithme d'optimisation utilisé. Dans cette thèse, nous avons considéré cette dernière méthode en cherchant à utiliser un algorithme de segmentation exact (au sens du critère de vraisemblance). La Section 1.2 présente la méthode générale de segmentation : l'objectif est de montrer les difficultés tant statistique qu'algorithmique de cette approche et de présenter les solutions existantes dans la littérature.

La méthode de segmentation que nous avons développée est présentée dans la Section 1.3 dans laquelle ses performances sont étudiées à l'aide d'une étude de simulation. La Section 1.4 présente les résultats de la segmentation sur les données réelles des 120 stations. Parmi les ruptures détectées, certaines semblent plutôt correspondre à des détections sur des pics de bruit qu'on appellera des valeurs aberrantes ("outliers" en anglais). Dans cette section, nous proposons une méthode automatique pour séparer les "outliers" des "vraies" ruptures. Cette étape est appelée étape de "nettoyage" (screening en anglais). Enfin, nous utilisons l'ensemble des ruptures nettoyé des outliers pour estimer la tendance climatique dans les séries CIVE.

La Figure 1.3 donne le schéma de la procédure globale proposée. La première étape consiste à récupérer les données d'intérêts d'une part en utilisant la technique de traitement des données CIVE dérivée du GNSS et d'autre part en effectuant la différence avec la série de référence obtenue par ERAI afin d'éliminer la variabilité naturelle des séries CIVE, notées $\Delta CIVE$. Il s'ensuit l'étape de segmentatation des séries puis de "nettoyage". Les ruptures sont ensuite utilisées afin d'estimer la tendance climatique dans les séries d'origine.

1.2 Méthode générale de segmentation

Nous présentons la méthode générale de segmentation dans un cadre paramétrique, fréquentiste utilisant l'inférence par la méthode du maximum de vraisemblance (pénalisé).Les données observées $\boldsymbol{y} = \{y_t\}_{t=1,\dots,n}$ sont supposées être des réalisations de n variables aléatoires indépendantes Y_t de loi de probabilité $(P_\theta)_{\theta\in\Theta}$ où le paramètre θ est supposé affecté par $K-1$ ruptures :

$$Y_t \sim P_{\theta_k} \text{ si } t \in I_k = [\![t_{k-1}+1, t_k]\!], k = 1, \dots, K$$

où

1. K est le nombre de segments,

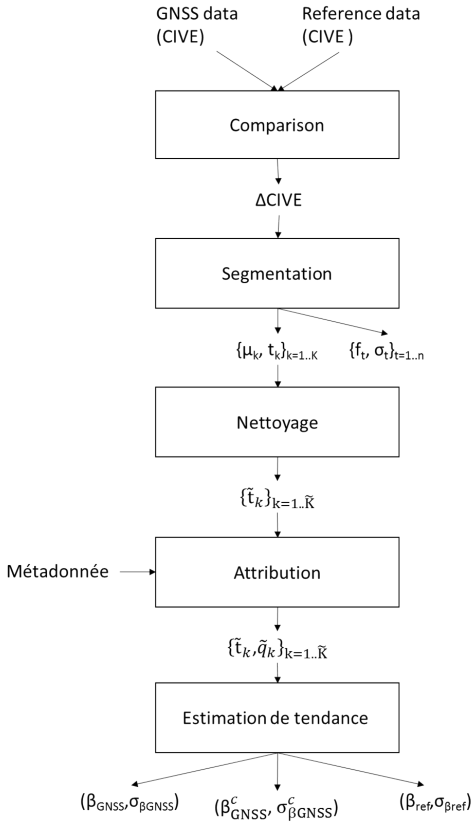


FIGURE 1.3 – Schéma de la procédure générale proposée pour l'analyse des séries journalières GNSS $\Delta CIVE$.

2. $\boldsymbol{T} = (t_1, \ldots, t_{K-1})$ le vecteur des $K-1$ ruptures qui décomposent le signal en K segments, $I_k = [\![t_{k-1}+1, t_k]\!]$, avec la convention $t_0 = 0$ et $t_K = n$.
3. $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ est le vecteur des paramètres de la distribution.

Classiquement en segmentation, l'inférence se fait en trois étapes : (1) estimation de $\boldsymbol{\theta}$ à T et K fixés, (2) estimation de T à K fixé et (3) choix de K. L'étape (1) ne présente en général pas de difficultés majeures et se fait de façon exacte. Notons $\widehat{\boldsymbol{\theta}}$ l'estimateur de $\boldsymbol{\theta}$. Les principales difficultés concernent l'étape (2) qui pose un problème algorithmique et l'étape (3) qui pose un problème statistique de sélection de modèles.

Pour l'estimation des instants de ruptures, nous cherchons à maximiser la log-vraisemblance calculée en son maximum pour $\widehat{\boldsymbol{\theta}}$

$$\widehat{\boldsymbol{T}} = \underset{\boldsymbol{T}\in\mathcal{M}_{K,n}}{\operatorname{argmax}} \log p(\boldsymbol{y}; K, \mathbf{T}, \widehat{\boldsymbol{\theta}}) = \underset{\boldsymbol{T}\in\mathcal{M}_{K,n}}{\operatorname{argmax}} \sum_{k=1}^{K} \sum_{t=t_{k-1}+1}^{t_k} \log p_{\hat{\theta}_k}(y_t),$$

où $\mathcal{M}_{K,n} = \{(t_1, \ldots, t_{K-1}) \in \mathbb{N}^{K-1}, 0 = t_0 < t_1 < \cdots < t_{K-1} < t_K = n\}$ est l'ensemble de toutes les

segmentations possibles de la grille $[\![1, n]\!]$ en K segments. Pour ce faire, nous devons explorer tout l'espace $\mathcal{M}_{K,n}$ qui est de très grande taille, $\binom{n-1}{K-1}$. Un algorithme naif ne peut être ainsi utilisé. L'algorithme maintenant bien connu qui permet d'obtenir la solution exacte en un temps algorithmique raisonnable est l'algorithme de Programmation Dynamique (DP), introduit par Bellman [1954]. La condition nécessaire pour pouvoir utiliser DP est que la quantité à maximiser soit segment-additive. En particulier, il ne sera pas possible de l'utiliser si il existe un paramètre commun aux segments. À ce stade, nous disposons d'une collection de meilleures segmentations en K segments et l'objectif est de choisir le "meilleur" K. Cette étape (3) est une question de choix de modèles qui se résout par la maximisation d'un critère de vraisemblance pénalisée :

$$\widehat{K} = \underset{K}{\operatorname{argmax}} \ \log p(\boldsymbol{y}; K, \widehat{\boldsymbol{T}}, \widehat{\boldsymbol{\theta}}) - pen(K, n).$$

Ainsi le problème se réduit au choix d'une "bonne" pénalité. Dans cette thèse, nous considérerons quatre critères : celui proposé par Lavielle [2005], noté Lav, dans lequel la pénalité est proportionnelle au nombre de segments ; celui proposé par Birgé & Massart [2001], qui mène à deux versions BM1 et BM2 car deux heursitiques pour calibrer les constantes peuvent être utilisées (voir Arlot & Massart [2009]) et, celui proposé par Zhang & Siegmund [2007], mBIC, qui est une version modifiée du critère BIC classique dédié à la segmentation dans la moyenne d'un processus gaussien. Le choix de K est un problème compliqué et délicat. Les critères étant différents, ils peuvent sélectionner un modèle différent.

1.3 Une nouvelle méthode de segmentation adaptée aux données GNSS CIVE

Nous présentons la méthode de segmentation que nous avons développée. Il s'agit d'une méthode de ruptures dans la moyenne d'un processus gaussien qui prend en compte les caractéristiques des données CIVE GNSS, à savoir un biais périodique et une variance mensuelle. Ce travail s'est basé sur un modèle proposé par Bock *et al.* [2018], qui est un modèle de détection de ruptures dans la moyenne à variance mensuelle, auquel nous avons ajoutée une composante périodique.

1.3.1 Modèle et inférence

Les données observées $\boldsymbol{y} = \{y_t\}_{t=1,\dots,n}$ sont supposées être des réalisations de n variables aléatoires indépendantes Y_t tel que

(i) la moyenne de Y_t est composée de deux termes :

- une fonction constante par morceaux $\mu_k(t)$ égale à μ_k sur l'intervalle $I_k^{\text{mean}} = [\![t_{k-1} + 1, t_k]\!]$ de longueur $n_k = t_k - t_{k-1}$ où $0 = t_0 < t_1 < \ldots < t_{K-1} < t_K = n$. Les $\mathbf{T} = \{t_k\}_{k=1,\dots,K-1}$ sont les positions des ruptures et K est le nombre d'intervalles ou segments.

– et une fonction f_t ;

(ii) la variance de Y_t dépend du mois, c'est-à-dire qu'elle est constante sur l'intervalle $I^{\text{var}}_{\text{mois}} = \{t; \text{date}(t) \in \text{mois}\}$ de longueur n_{mois} où date(t) représente la date à la position t.

Le modèle est donc le suivant

$$Y_t = \mu_k + f_t + \mathcal{E}_t, \quad \text{avec } \{\mathcal{E}_t\}_t \text{ iid} \sim \mathcal{N}(0, \sigma^2_{\text{mois}}) \text{ pour } t \in I^{\text{mean}}_k \cap I^{\text{var}}_{\text{mois}}, \tag{1.1}$$

pour $k = 1, \ldots, K$. Les intervalles $\{I^{\text{mean}}_k\}_k$ sont inconnus contrairement aux intervalles $\{I^{\text{var}}_{\text{mois}}\}_{\text{mois}}$ qui sont fixes. La composante fonctionnelle f_t décrit les variations lisses de la moyenne de la série $\Delta CIVE$.

La log-vraisemblance s'écrit

$$\log\ p(\boldsymbol{y}; K, \boldsymbol{T}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, f) \quad = \quad -\frac{n}{2}\log(2\pi) \sum_{\text{mois}} \frac{n_{\text{mois}}}{2} \log(\sigma^2_{\text{mois}}) - \frac{1}{2}\sum_{k=1}^{K}\sum_{\text{mois}}\sum_{t \in I^{\text{mean}}_k \cap I^{\text{var}}_{\text{mois}}} \frac{(y_t - \mu_k - f_t)^2}{\sigma^2_{\text{mois}}}$$

Le problème qui se pose ici est que $\boldsymbol{\sigma}$ et f sont des paramètres communs aux segments. Ainsi DP ne peut être directement utilisé. Pour palier à ce problème et conserver l'utilisation de DP, nous avons proposé la procédure d'inférence en plusieurs étapes suivante :

1. Estimation of σ^2_{mois}. Nous estimons tout d'abord la variance en utilisant un estimateur robuste proposé par Bock *et al.* [2018]. L'idée est d'appliquer l'estimateur proposé par Rousseeuw & Croux [1993] sur la séries différenciée $(Y_{t+1} - Y_t)$. Dans la mesure où l'on considère la série différenciée, la présence de la fonction f n'a pas beaucoup d'impact sur l'estimation. La variance estimée est notée $\widehat{\sigma}^2_{\text{mois}}$.

2. Estimation de f, $\boldsymbol{T}$ et $\boldsymbol{\mu}$ itérativement pour chaque valeur de K. À l'itération $[h+1]$:

(a) l'estimateur de f est l'estimateur pondéré des moindres carrés avec les poids $1/\widehat{\sigma}^2_{\text{mois}}$ obtenus sur la série $\{y_t - \mu_k^{[h]}\}_t$. De part le fait que l'effet saisonnier observé pour la station CCJM (Figure 1.2 (b)), comme par de nombreuses autres stations, est lisse, nous avons décidé de représenter f comme une série de Fourier d'ordre 4 comprenant les périodicités annuelle, semi-annuelle, ter-annuelle et trimestrielle du signal :

$$f_t = \sum_{i=1}^{4} a_i \cos(w_i t) + b_i \sin(w_i t),$$

où $w_i = 2\pi\frac{i}{L}$ est la fréquence angulaire de la période L/i et L est la durée moyenne de l'année ($L = 365.25$ jours lorsque le temps t est exprimé en jours). La fonction estimée est notée $f^{[h+1]}$.

(*b*) les paramètres de segmentation sont estimés sur la série $\{y_t - f_t^{[h+1]}\}_t$. On obtient

$$\mu_k^{[h+1]} = \frac{\sum_{\text{mois}} \sum_{t \in I_k^{\text{mean}} \cap I_{\text{mois}}^{\text{var}}} \frac{(y_t - f_t^{[h+1]})}{\widehat{\sigma}^2_{\text{mois}}}}{\sum_{\text{mois}} \sum_{t \in I_k^{\text{mean}} \cap I_{\text{mois}}^{\text{var}}} \frac{1}{\widehat{\sigma}^2_{\text{mois}}}},$$

et

$$\boldsymbol{T}^{[h+1]} = \underset{\boldsymbol{T} \in \mathcal{M}_{K,n}}{\operatorname{argmin}} \sum_{k=1}^{K} \sum_{\text{mois}} \sum_{t \in I_k^{\text{mean}} \cap I_{\text{mois}}^{\text{var}}} \frac{(y_t - f_t^{[h+1]} - \mu_k^{[h+1]})^2}{\widehat{\sigma}^2_{\text{mois}}}.$$

Cette dernière minimisation peut alors se faire via l'algorithme DP. Les estimateurs finaux sont notés $\widehat{f}$, $\widehat{\boldsymbol{T}}$ et $\widehat{\boldsymbol{\mu}}$.

3. Choix de K. Nous considérons les 4 critères donnés dans le paragraphe précécent. La variance étant estimée au préalable, le problème de segmentation peut se voir comme un problème de segmentation à variance "connue". Ainsi le contraste utilisé dans les critères pénalisé est le critère des moindres carrés.

La procédure complète de cette nouvelle méthode de segmentation est résumée dans la Figure 1.4. Différentes variantes ont été testées avant d'aboutir à la procédure proposée. Ces variantes différaient sur les aspects suivants : (1) le choix de l'initialisation de la procédure itérative entre la segmentation et l'estimation de la fonction f, avec et sans pondération ; (2) la mise à jour de l'estimation de la variance dans la procédure itérative ; (3) la sélection des composantes significatives de la série de Fourier pour l'estimation de f.

La méthode finale a été implémentée sous la forme d'un package R « GNSSseg » disponible sur le CRAN (https://cran.r-project.org/web/packages/GNSSseg/index.html). Récemment, une version plus rapide de DP a été développée par Hocking *et al.* [2018]. Le package associé est `gfpop` disponible sur GitHub. Cela nous a permis de developper une deuxième version du package plus rapide, appelée `GNSSfast`, qui peut être téléchargé depuis https://github.com/arq16/GNSSfast.git.

1.3.2 Etude de simulation

Pour évaluer la performance de la méthode, nous avons testé la procédure sur des données simulées.

1.3.2.1 Plan de l'étude et critères de qualité.

Les séries sont de longueur 400 avec 4 années de 2 mois de 50 jours chacun. Elles sont affectées par 6 ruptures aux positions $t = 55, 77, 177, 222, 300, 366$ avec des valeurs de moyennes alternant entre 0 et 1. La fonction $f(t) = 0,7\cos(2\pi t/L)$ où $L = 100$ est la durée d'un an. Nous avons fixé $\sigma_1^\star$ à 0.5 et fait varier $\sigma_2^\star$ entre 0.1 et 1.5 par pas de 0.2. Chaque configuration a été simulée 100 fois. Les vraies valeurs des paramètres sont indicés ci-dessous par "$*$". Pour évaluer la qualité des estimations obtenues, nous avons utilisé les critères suivants :

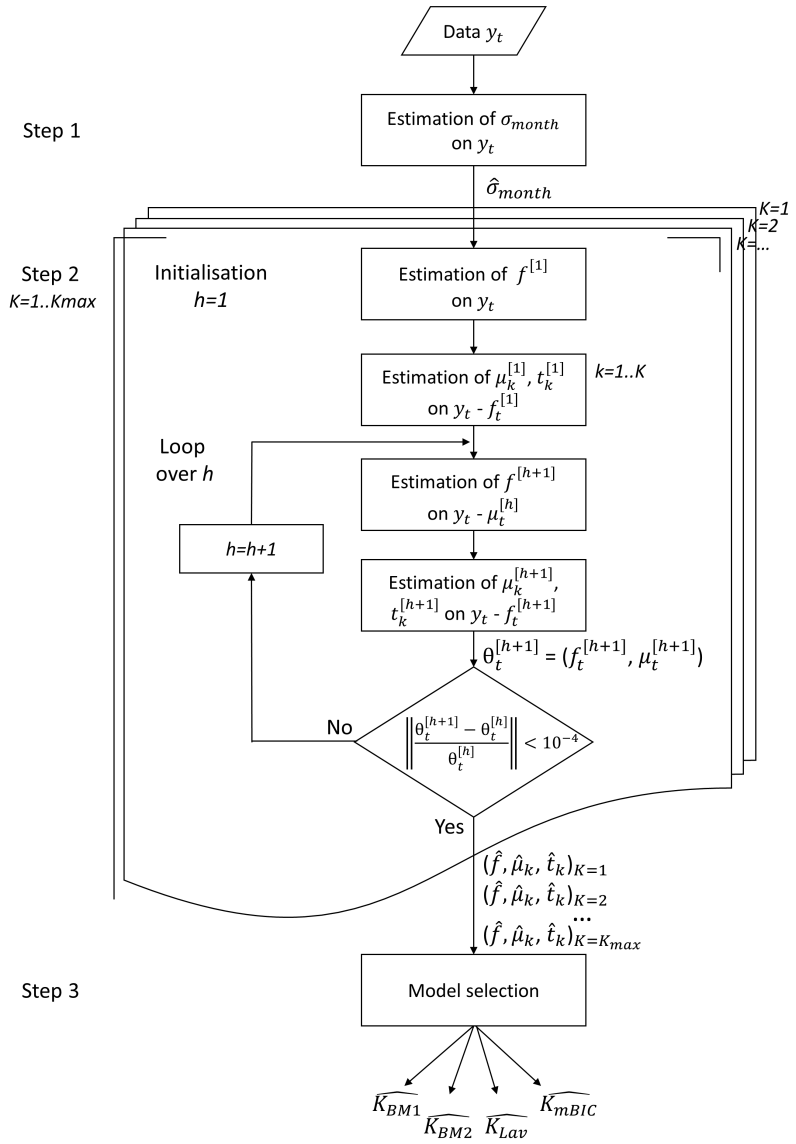


FIGURE 1.4 – Schéma de l'algorithme.

1. pour la fonction f, l'erreur quadratique moyenne (RMSE) : $\text{RMSE}(f) = \left[\frac{1}{n}\sum_{t=1}^{n}\left\{\widehat{f}_t - f_t^*\right\}^2\right]^{1/2}$;

2. pour les paramètres de segmentation, nous avons considéré plusieurs critères :

- ⋆ la différence entre le nombre de ruptures estimé et le vrai : $\widehat{K} - K^*$;
- ⋆ le RMSE des moyennes estimées ;

⋆ les deux composantes de la distance de Hausdorff, notées d_1 and d_2, et définies par :

$$d_1(\boldsymbol{t}^\star, \widehat{\boldsymbol{t}}) = \max_{\boldsymbol{t}^\star} \min_{\widehat{\boldsymbol{t}}} |\boldsymbol{t}^\star - \widehat{\boldsymbol{t}}| \quad \text{and} \quad d_2(\boldsymbol{t}^\star, \widehat{\boldsymbol{t}}) = d_1(\widehat{\boldsymbol{t}}, \boldsymbol{t}^\star).$$

Pour une segmentation parfaite, d_1 et d_2 sont nuls. Une petite valeur de d_1 signifie que les ruptures détectées sont bien positionnées et une petite valeur de d_2 qu'une grande partie des vraies ruptures sont correctement détectées.

⋆ l'histogramme de la positions des ruptures estimées.

1.3.3 Résultats

Précision des estimations des paramètres. La Figure 1.5 présente les résultats pour les 4 critères de sélection de modèle et le cas du vrai nombre de segments ($K = 7$) et les différents critères de qualité donnés ci-dessus. Pour de petites valeurs de $\sigma_2^\star$, le problème de détection est facile et tous les critères de sélection de modèle retrouvent le nombre correct de segments (Figure 1.5 (a)). Cependant pour de grandes valeurs de $\sigma_2^\star$, la détection devient plus difficile. Les différents critères de sélection se comportent légèrement différemment : Lav à tendance à retrouver le vrai nombre de segments en moyenne mais avec une grande variabilité, tandis que BM1, BM2 et mBIC ont tendance à sous-estimer le nombre de segments (plus pour mBIC). Cependant, trouver le nombre correct de segments ne signifie pas que les ruptures sont correctement positionnées. En effet, pour Lav et le cas où $K = 7$, la médiane d_1 est assez élevée (Figure 1.5 (c)). Par contre, la médiane d_2 est plus petite pour le cas où $K = 7$ par rapport aux critères testés (Figure 1.5 (d)). Enfin, le RMSE($\boldsymbol{\mu}$) est très similaire pour tous les critères (Figure 1.5 (b)), bien que Lav montre une médiane et une variabilité plus grandes lorsque $\sigma_2^\star$ est grand. La Figure 1.5 (e) donne le RMSE(f) en fonction de $\sigma_2^\star$. Les résultats ne dépendent pas beaucoup du critère de sélection, mais sont légèrement meilleurs lorsque le vrai nombre de segments est considéré et lorsque σ_2 prend des valeurs intermédiaires. Les résultats pour Lav montrent une médiane légèrement plus élevée et une plus grande variabilité.

Probabilité de détection. La Figure 1.6 donne la probabilité de détecter les positions des ruptures détectées sur les 100 simulations pour trois valeurs de $\sigma_2^\star = 0.1, 0.5$ et 1.5, et $\sigma_1^\star = 0.5$. En général, les ruptures situées dans les "mois" avec petite variance sont souvent détectées avec les trois critères et également lorsque la vraie valeur de K est considérée. Ainsi, dans le cas (a) où $\sigma_1^\star = 0.5$ et $\sigma_2^\star = 0.1$, la probabilité de détection est légèrement plus petite pour la position 222 qui est contenue dans un segment avec $\sigma_1^\star = 0.5$, et pour la position 300 pour laquelle la moyenne et la variance changent. Dans le cas (b) où $\sigma_1^\star = \sigma_2^\star = 0.5$, la probabilité de détection est plus ou moins la même pour tous les critères. Lorsque $\sigma_2^\star = 1.5$, le problème est plus difficile. Encore une fois, les ruptures situées dans les "mois" avec un bruit plus petit sont mieux détectées (positions 222 et 300) mais pour les quatre autres ruptures, les résultats

sont contrastés bien qu'ils soient tous situés dans un mois avec $\sigma_2^\star = 1.5$. Les ruptures à 55 et 77 ne sont presque jamais détectés. Pour mBIC ceci est cohérent avec le fait que la médiane $\hat{K} = 5$, c'est-à-dire qu'il manque deux ruptures en moyenne (Figure 1.5 (a)), mais les quatre autres ruptures ne sont pas si mal situés (d_1 n'est pas si grand, Figure 1.5 (c), mais d_2 est très grand, Figure 1.5 (d)). La situation est un peu similaire pour BM1. Pour Lav et la vraie valeur de K, le nombre de détections est correct (en moyenne pour Lav) mais du fait du bruit important, elles sont parfois très mal positionnés (grandes valeurs pour d_1 et d_2).

1.4 Application aux données réelles

La version finale de la nouvelle méthode de segmentation a été appliquée aux différences CIVE (GNSS moins ERA-Interim) de 120 stations. Pour valider les ruptures détectées pour les stations GNSS, nous avons utilisé les métadonnées disponibles sur le site IGS (ftp ://igs.org/pub/station/log/). En principe, les métadonnées IGS sont bien mises à jour mais il peut arriver que certaines modifications ne soient pas enregistrées ou que certaines dates soient erronées. Nous avons extrait pour chaque station les dates de changement de récepteur (R), d'antenne (A), de traitement (P) et de radôme (D).

La nouvelle méthode est étiquetée (a). Trois variantes de la méthode sont également présentées pour discuter de la sensibilité des résultats et de la performance des quatre critères de sélection. Les variantes sont : (b) seuls les termes statistiquement significatifs de la série de Fourier sont sélectionnés, (c) seule la segmentation est implémentée, c'est-à-dire que le biais périodique modélisé par la fonction f n'est pas inclus (c'est la méthode proposée par Bock *et al.* [2018]), (d) une variance homogène est considérée au lieu d'un variance mensuelle (il s'agit d'une version homoscédastique mais incluant toujours la fonction f).

Nombre de ruptures. La Figure 1.7 (a) montre le nombre des ruptures détectées pour les quatre critères. D'autres résultats sont données dans le Tableau 1.1. La répartition du nombre des ruptures détectées par station est très différente selon le critère de sélection. Plus particulièrement, mBIC détecte entre 9 et 29 ruptures par station, avec une valeur moyenne de $27,1$, c'est-à-dire que dans de nombreux cas, le plus grand nombre possible est sélectionné (29 puisque $K_{\max} = 30$). Ce comportement n'a pas été observé avec les simulations présentées dans la Section 1.2. Cela conduit à une forte sur-segmentation, ce qui n'est pas souhaité. Ce critère de pénalité n'est donc pas adapté à la nature des données analysées ici. Une raison pourrait être que l'hypothèse d'erreur gaussienne n'est pas valide avec ces séries temporelles. Une corrélation temporelle reste en effet dans les données. Par exemple pour la station CCJM (Figure 1.2) nous avons calculé le lag-1 de la fonction d'autocorrélation des résidus et trouvé une valeur de r = 0,249.

La variante (b), montre un impact marginal sur le nombre de détections et le nombre de validations

TABLE 1.1 – Comparaison des résultats de segmentation pour les quatre variantes et les quatre critères de sélection du modèle. De gauche à droite : nombre de stations avec des ruptures, nombre min / moyen / max de ruptures détectées par station, nombre total de ruptures, nombre total de valeurs aberrantes (outliers), nombre total de validations, pourcentage de validations y compris les valeurs aberrantes, pourcentage de validations sans valeurs aberrantes.

	Nsta	min	mean	max	detections	outliers	validations		
Variante (a) (segfonc)									
mBIC	120	9	27.1	29	3251	2096	267	8.2%	20.9%
Lav	114	0	4.0	28	474	129	75	15.8%	21.3%
BM1	98	0	2.8	14	335	36	70	20.9%	23.3%
BM2	107	0	3.6	18	435	64	77	17.7%	20.6%
Variante (b) (segfonc/select)									
mBIC	120	8	27.2	29	3268	2090	270	8.3%	20.7%
Lav	115	0	7.8	28	940	411	116	12.3%	20.8%
BM1	100	0	2.8	13	334	46	68	20.4%	23.4%
BM2	107	0	3.7	24	439	76	81	18.5%	22.1%
Variante (c) (segonly)									
mBIC	120	9	28.1	29	3367	1255	361	10.7%	16.4%
Lav	113	0	2.9	16	350	28	64	18.3%	19.6%
BM1	90	0	2.2	12	269	8	53	19.7%	20.2%
BM2	102	0	3.5	17	414	24	68	16.4%	17.4%
Variante (d) (seghomofonc)									
mBIC	116	0	19.0	29	2283	1637	178	7.8%	24.1%
Lav	114	0	3.5	26	415	148	56	13.5%	20.4%
BM1	92	0	2.4	19	287	40	61	21.3%	24.1%
BM2	101	0	3.2	19	387	82	68	17.6%	21.7%

pour trois critères (mBIC, BM1 et BM2). Dans les cas (c) et (d), le nombre de ruptures détectées diminue. Avec la variante (d), la variance est supposée constante ce qui a pour conséquence que la fonction estimée sera différente et les moyennes des segments également. On observe également que les écarts-types moyens sont différents ($1,19$ contre 0,84 kgm^{-2} pour la variante (a)) et moins de ruptures sont détectés. Le Tableau 1.1 montre également que le nombre de valeurs aberrantes est augmenté dans le cas (d), sauf pour mBIC qui est problématique, et le nombre de validations est diminué, mais le pourcentage de validations est pratiquement inchangé. La comparaison des quatre variantes montre que le modèle complet avec variance hétérogène et une série harmonique pour le biais périodique a les meilleures propriétés (nombre raisonnable de détections, petit nombre de valeurs aberrantes et taux élevé de validations).

Validations. Parmi les trois critères, BM1 a le plus petit nombre de valeurs aberrantes (36) et le taux de validations le plus élevé (20,9 %). Ces deux caractéristiques, ainsi que le fait que BM1 possède un nombre raisonnable de ruptures (2,8 en moyenne par station), en font le critère de sélection préféré.

Parmi les ruptures validées trouvées par BM1, il existe 53 types R, 16 A, 7 D et 13 P. Parfois plusieurs types de changements sont simultanés. Les changements de récepteurs sont le cas le plus fréquent. Il s'agit aussi du type de changement le plus fréquent dans les métadonnées. Cependant, cela contraste avec les résultats de Ning *et al.* [2016] qui n'ont pas du tout pris en compte les changements de récepteurs.

Stations particulières. La Figure 1.8 présente les résultats de segmentation des variantes (a), (c) et (d) pour quatre stations particulières.

Dans le cas de POL2, les trois variantes détectent respectivement 3, 12 et 1 rupture(s). Le signal présente une forte variation périodique qui est bien ajustée par les modèles pour les variantes (a) et (d). La variante (a) a une rupture validée (23/02/2008 pour un changement connu le 06/03/2008). La variante (c) n'a pas de validation, bien qu'elle détecte 12 ruptures. La variante (d) ne détecte qu'une seule rupture, qui est situé à 72 jours de la rupture connue la plus proche et qui coïncide avec l'une des trois ruptures trouvées par la variante (a). La détection de cette rupture est rendue difficile car elle se situe dans un mois avec un fort bruit.

Dans le cas de STJO, les variantes (a) et (d) détectent respectivement 5 et 4 ruptures, avec une valeur aberrante chacune mais pas à la même position. La variante (c) ne donne aucune détection (BM1 est souvent un critère conservatif).

Dans le cas de DUBO, les variantes (a) et (c) détectent deux ruptures à peu près à la même position, situées à proximité des changements connus, mais une seule est validée pour la variante (a). La seconde est située à 34 jours d'un changement connu pour la variante (a) et à 148 jours pour la variante (c). La la variante (a) reste la plus précise. La variante (d) a 4 détections qui consistent en fait en 2 ruptures, chacune étant associée à une valeur aberrante. Bien que le biais périodique soit ici modélisé, les deux ruptures sont assez mal localisées et donc non validées.

Enfin pour MCM4 le signal présente des inhomogénéités très marquées sous forme de plusieurs changements brusques mais aussi d'oscillations non stationnaires. Les changements brusques sont bien captés par la variante (a) qui détecte 5 ruptures parmi lesquelles 4 sont validées. Les oscillations non stationnaires ne sont que partiellement modélisées par la fonction périodique. Ce résultat suggère d'utiliser une base de fonctions plus complexes. La variante (c) fonctionne assez bien aussi et conduit à presque les mêmes détections que la variante (a). Deux changements sont validés. La variante (d) surestime en revanche le nombre des ruptures pour mieux s'adapter aux oscillations non stationnaires mais avec des détections de valeurs aberrantes. Les quatre mêmes ruptures sont validées comme avec la variante (a) mais les moyennes ajustées sont assez différentes.

Détection et suppression des erreurs aberrantes. L'inspection des résultats de la segmentation montre qu'il y a des valeurs aberrantes dues à des pics de bruit dans les séries temporelles (cf. l'exemple de STJO). Dans la Table 1.1 nous avons considéré comme "outliers" les ruptures plus proches

que 30 jours mais ce seuil avait été choisi un peu arbitrairement. Afin de déterminer le seuil de détection de manière plus rigoureuse nous avons analysé la distribution des longueurs de segments ainsi que la variation des moyennes avant et après les ruptures proches. Mais avant il est nécessaire de définir plus précisément la notion de valeurs aberrantes ou "outlier".

Formellement, soit t_i et t_{i+1} les positions de deux ruptures consécutives. Si $t_{i+1} - t_i < seuil$, alors ces ruptures t_i et t_{i+1} sont appelés "valeurs aberrantes" (outliers) et forment un "cluster" de deux valeurs aberrantes. Les différentes configurations observées des ruptures détectées sont représentées sur la Figure 1.9 :

- classe 1, appelées 'valeurs aberrantes seulement' (exemples de cas (a), (b), (c)),
- classe 2, appelées 'valeurs aberrantes et rupture' (cas (d), (e), (f)),
- classe 3, appelées 'rupture seulement' (cas (g), (h)).

Les classes 1 et 2 correspondent à un cluster de deux valeurs aberrantes ($t_{i+1} - t_i < seuil$). Dans les cas de la classe 1, la variation des moyennes avant et après le cluster n'est pas significative (selon un test statistique précisé ci-dessous), alors qu'en classe 2 elle est significative. Le but du nettoyage est donc d'éliminer les deux ruptures de classe 1, et de remplacer les ruptures de classe 2 par une seule rupture (schématisée par le point médian sur la Figure 1.9). La classe 3 est la situation normale lorsque la distance entre les deux ruptures est supérieure au seuil ($t_{i+1} - t_i \geq threshold$).

L'analyse de la distribution des longueurs des segments a montré qu'il y a un groupe de petits segments de longueurs inférieures à 50 jours séparé du reste de la distribution qui est plus étendue. Entre 50 et 100 jours il y a un minimum ce qui suggère que le seuil peut être choisi dans cet intervalle. Afin de déterminer le seuil optimal, nous avons utilisé un modele de melange et avons effectivement trouvé deux populations : une de petites longueurs et une de plus grandes longueurs (Figure 1.10). La frontiere entre les deux populations est estimée à 81 jours. Nous avons par la suite fixé le seuil de détection des valeurs aberrantes à 80 jours. Pour déterminer ensuite si une valeur aberrante est de classe 1 ou 2 nous avons fait un test d'égalité des moyennes avant et après le cluster. Nous avons utilisé un test de moyennes pondérées en prenant comme variance la variance estimée par la segmentation (elle peut être différence pour chaque point du signal).

La Figure 1.11 montre le résultat du nettoyage pour la station IISC. Avant le nettoyage, la station avait 12 ruptures, dont 8 valeurs aberrantes, regroupées en 3 clusters. Les deux premiers clusters ont 2 valeurs aberrantes chacun, tandis que le troisième cluster en a 4. La variation de la moyenne avant et après ces clusters est significative, donc tous sont classés en classe 2. Le nettoyage gardera alors le point médian de ces clusters, résultant en 7 ruptures restantes.

Nous avons également testé une autre méthode pour éviter les outliers en modifiant le paramètre 'lmin', représentant la longueur minimale des segments dans la segmentation. Il permet d'interdire à la segmentation de choisir des segments plus petits que 'lmin'. Nous avons analysé le comportement

		avant le nettoyage			après le nettoyage		
	criteria	detections	validations	% validation	detections	validations	% validation
lmin1	mBIC	3251	264	8.1	1270	146	11.50
	Lav	474	75	15.8	341	67	19.65
	BM1	335	70	20.9	292	68	23.29
	BM2	435	77	17.7	370	74	20.00
lmin10	mBIC	3056	276	9.03	1261	155	12.29
	Lav	530	84	15.85	361	70	19.39
	BM1	341	75	21.99	301	71	23.59
	BM2	491	83	16.90	413	77	18.64

TABLE 1.2 – Comparaison des résultats de segmentation pour les valeurs de 1 et 10 de la longueur de segment minimale (lmin) pour le seuil de valeur aberrante de 80. De gauche à droite : nombre des ruptures détectées, nombre total de validations, pourcentage de validations, avant et après le nettoyage.

de la segmentation pour des valeurs de lmin allant de 10 à 100. En augmentant lmin, le nombre total de segments augmente également, principalement dans les petits segments. Au-delà de lmin=10, les résultats ne sont pas très concluants. En comparant les résultats pour toutes les valeurs de lmin (de 1 à 100), suivies ou non du "nettoyage" décrit précédemment, et pour tous les critères, nous avons trouvé que le taux de validation des ruptures finales est le plus élevé pour le critère BM1 avec lmin=10 après le "nettoyage". Ce serait donc cette méthode qu'il faudrait retenir. Dans tous les cas et pour chaque critère la procédure de "nettoyage" améliore les résultats. Et dans tous les cas les meilleurs résultats sont obtenus pour BM1. La Table 1.2 montre les résultats pour lmin=1 et 10. Le pourcentage de validation pour BM1 passe de 20.90% à 23.59% entre lmin=1 avant nettoyage et lmin=10 après nettoyage.

Estimation de tendance. La tendance est estimée directement sur la série CIVE avec un modèle dans lequel sont inclus, en plus de la tendance, une fonction constante par morceaux dont les ruptures sont celles données par la segmentation (après le nettoyage) et les valeurs moyennes sont des paramètres à estimer, et une fonction harmonnique (série de Fourier d'ordre 4 dont les coefficients sont à estimer) qui représente la variation saisonnière du CIVE. Toutes les autres échelles de variabilité vont dans les résidus. L'estimation des paramètres est faite dans un premier temps par moindres carrés ordinaires (OLS). La Figure 1.12 montre les résultats pour la station ALIC qui a 5 ruptures. Le modèle est bien ajusté aux variations (lentes) du signal (graphe du haut), mais lorsque nous regardons la distribution des moyennes et la tendance nous constatons une confusion entre ces paramètres. La tendance estimée ne parait pas réaliste : $\hat{a} = 0.790 \pm 0.112 kgm^{-2}an^{-1}$, elle est trop forte et significative. En analysant les résidus on constate qu'ils sont très corrélés ($r = 0.7858$). L'hypothèse de bruit indépendant n'est pas vérifiée. L'erreur sur le paramètre n'est donc pas correcte avec l'OLS. Du coup nous avons utilisé les moindres carrés généralisés (GLS) pour tenir compte du bruit corrélé. Les estimations GLS, avec le modèle d'erreur AR(1), sont considérées comme plus réalistes. L'estimation de la tendance avec la méthode GLS est $\hat{a} = 0.820 \pm 0.307 kgm^{-2}an^{-1}$. Elle est toujours trop forte et significative bien que

l'erreur standard ait presque triplé avec le modèle de bruit AR(1). En général, nous avons testé les estimations de tendance pour toutes les stations, avec un test d'hypothèse et un seuil de significativé de $0,05$. Pour les modèles OLS et GLS qui prennent en compte les ruptures, nous avons trouvé 40 et 14 stations avec des tendances signifcatives, respectivement.

1.5 Conclusions et perspectives.

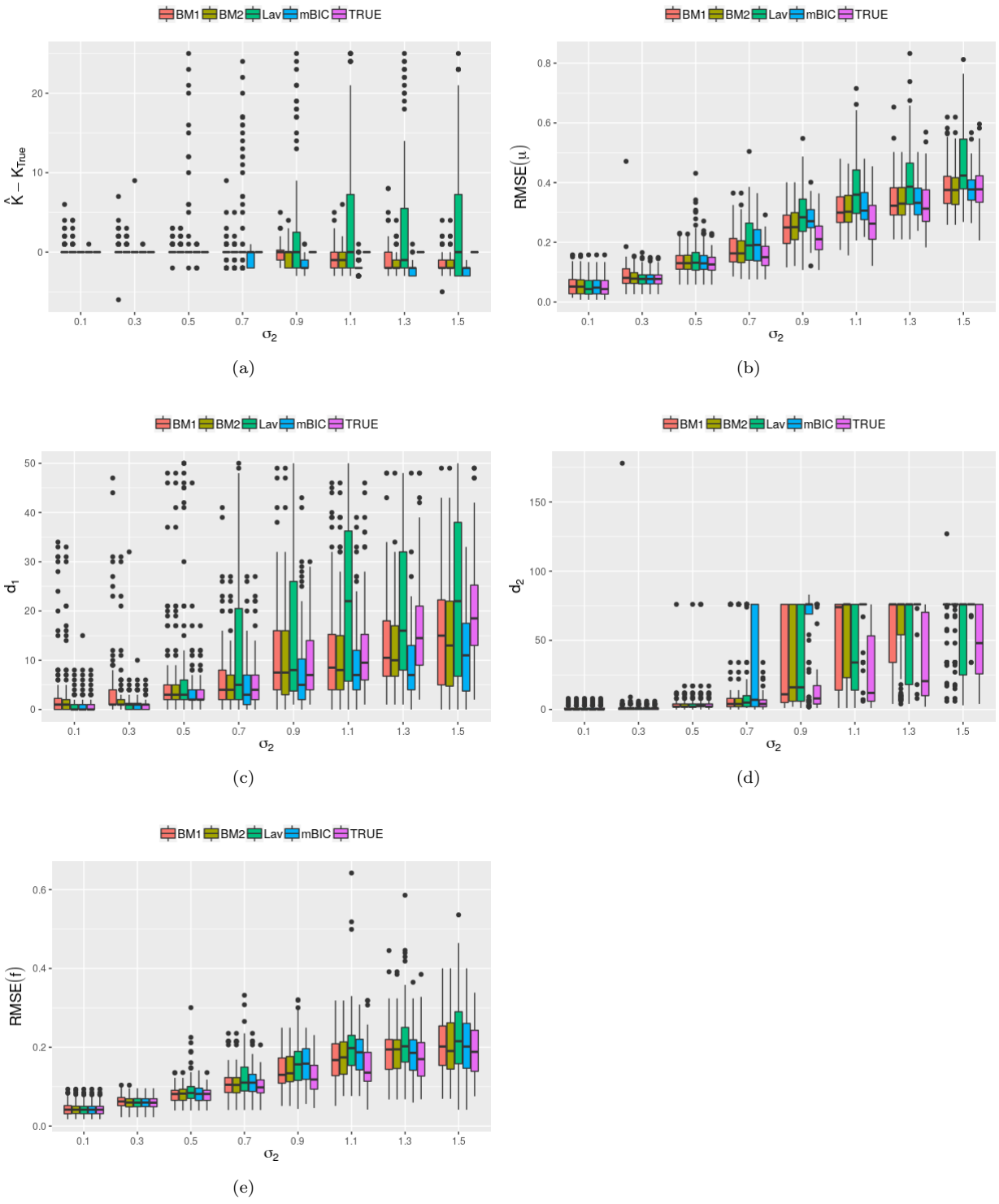
Dans cette thèse, nous avons développé une nouvelle méthode de segmentation dédiée à la détection de changements abrupt de la moyenne qui prend en compte un biais périodique et une variance hétérogène à intervalles fixes (mensuels) dans les différences CIVE entre les observations GNSS et la réanalyse ERA-Interim. La méthode a d'abord été testée et optimisée par une étude de simulation, puis appliquée aux données GNSS IWV pour 120 stations du réseau IGS mondial pour la période de janvier 1995 à décembre 2010. Nous avons vu que le comportement de la procédure dans les simulations est différent de celui des données réelles.

Segmentation. Au niveau de la phase de segmentation, nous avons vu que la procédure peut être très sensible à l'estimation de la fonction, tout comme le problème de la sélection du modèle est également délicat. Une façon d'améliorer la segmentation sur les vraies données est d'estimer le f, qui en réalité est évidemment plus complexe, avec une méthode plus flexible, non paramétrique ou semi-paramétrique. Une méthode déjà testée en segmentation est, par exemple, la méthode Lasso (Bertin *et al.* [2017]). Une autre façon d'améliorer la segmentation serait d'utiliser une perte de Hubert ou Biweight. Même si nous traitons les valeurs aberrantes a posteriori (avec la méthode de nettoyage) il pourrait être intéressant d'utiliser une sorte de telles pertes parce qu'elles permettent de supprimer les points les plus extrêmes qui pourraient avoir un fort impact à la fois sur la détection des ruptures, sur l'estimations des moyennes et sur f. De plus, l'utilisation d'une série de référence à plus haute résolution spatiale (par exemple les réanalyses ERA5 et UERRA) pour CIVE réduira les différences de représentativité.

Pour le choix du modèle final, en général, nous préférons un critère qui n'évalue pas trop de ruptures. Le critère le plus approprié pour ces données était BM1, en termes de performance sur les simulations, les quantités des ruptures estimées et les pourcentages de validation. Après segmentation, nous avons également remarqué la présence de valeurs aberrantes qui sont des ruptures proches les unes des autres. Elles sont généralement dus à des pics de bruit importants dans la série qui peuvent être supprimés avec une méthode de nettoyage. Dans le cas du critère BM1, il a détecté 20% des valeurs aberrantes et en a supprimé un tiers lorsqu'un seuil de 80 jours était utilisé. Une autre approche pour traiter le problème des valeurs aberrantes a été testée en imposant une longueur de 'lmin' dans l'algorithme de segmentation. En guise de compromis, nous avons constaté que $lmin = 10$ combiné au nettoyage produit les meilleurs résultats (taux de validation le plus élevé).

Validation. Les ruptures trouvées ont été attribuées au GNSS, puis validées à l'aide de métadonnées. Le taux de validation le plus élevé a été obtenu à partir de la combinaison de lmin égale à 10 + nettoyage (23, 59%). Les ruptures non validées peuvent être attribuées à la série de référence ou à des changements non enregistrés de métatada ou de fausses détections. Des données plus récentes couvrant une période plus longue existent (1994-2019, Bock [2019]) ainsi que des réseaux plus denses que le réseau utilisé dans cette étude. Ces données permettront de tester la significativité des ruptures détectées pour fiabiliser l'attribution des changements (mieux distinguer les ruptures dans à la série GNSS de celles dans la série de référence).

Estimation de tendance. Pour avoir une série homogène, il est possible d'estimer la tendance en intégrant les ruptures trouvées dans le modèle. Les tendances linéaires ont été estimées par les moindres carrés ordinaires et généralisés (OLS et GLS) sur la série temporelle GNSS CIVE en tenant compte des ruptures détectées par la segmentation (après le nettoyage). Bien que cette approche ait été couramment utilisée dans la communauté GNSS (Bernet *et al.* [2020]; Klos *et al.* [2018]), une confusion entre la tendance et les moyennes a été trouvée et conduira à une surestimation des sauts dans la moyenne. De plus l'estimation de tendance par GLS en supposant un processus AR (1) donne des incertitudes de tendance plus réalistes que l'OLS, mais elles sont également beaucoup plus importantes. En conséquence, seulement 12 % des stations ont une tendance significative (comparativement à 30 % avec OLS).



(a) (b) (c) (d) (e)

FIGURE 1.5 – Résultats pour les quatre critères de sélection (BM1, BM2, Lav, and mBIC) et le vrai nombre de segments (True) pour $\sigma_1^\star = 0.5$ et des valeurs différentes pour $\sigma_2^\star$. *(a)* $\hat{K} - K^\star$; *(b)* RMSE($\boldsymbol{\mu}$) ; *(c)* d_1 et *(d)* d_2 ; *(e)* RMSE de f.

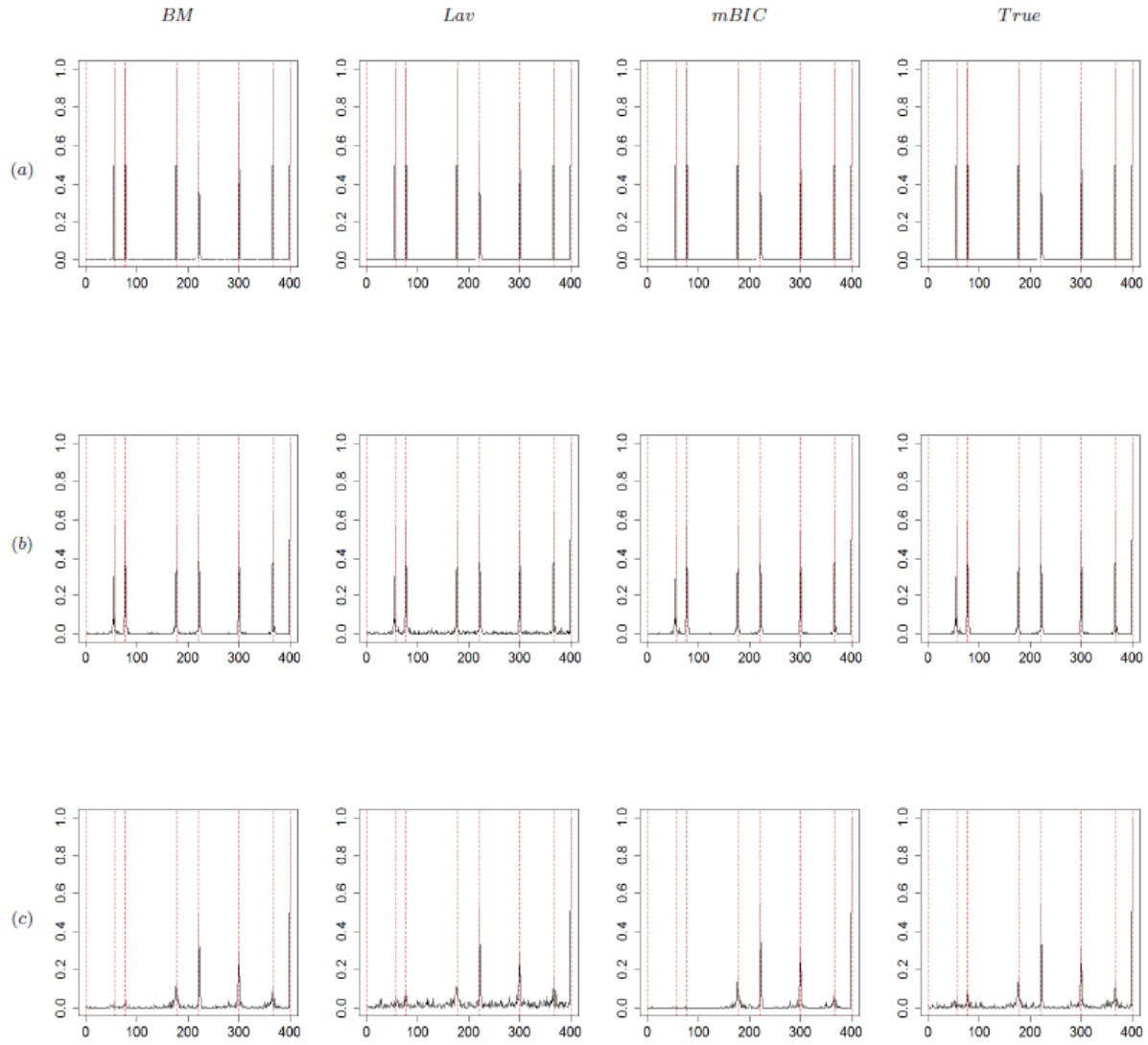


FIGURE 1.6 – Histogramme des positions des vraies ruptures avec, de gauche à droite, les critères de sélection BM, Lav et mBIC, et le vrai K (TRUE), pour $\sigma_1^\star = 0.5$ et trois valeurs différentes de $\sigma_2^\star$: (a) $\sigma_2^\star = 0.1$, (b) $\sigma_2^\star = 0.5$ et (c) $\sigma_2^\star = 1.5$. Les lignes pointillées rouges indiquent les positions des vraies ruptures.

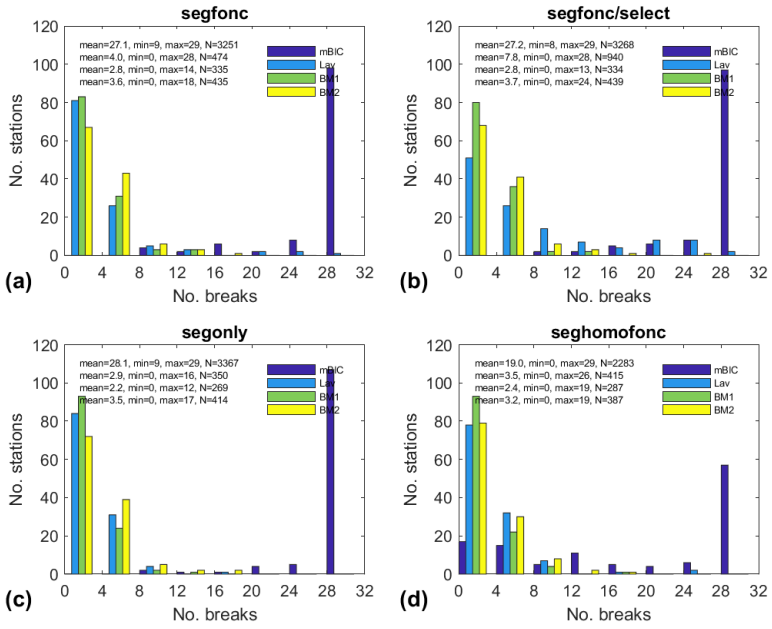


FIGURE 1.7 – Histogrammes du nombre des ruptures détectées pour quatre variantes des critères de sélection du modèle (mBIC, Lav, BM1 et BM2). Les nombres donnés dans les graphiques sont le nombre moyen, minimum et maximum des ruptures détectées par station, N est le nombre total des ruptures par méthode.

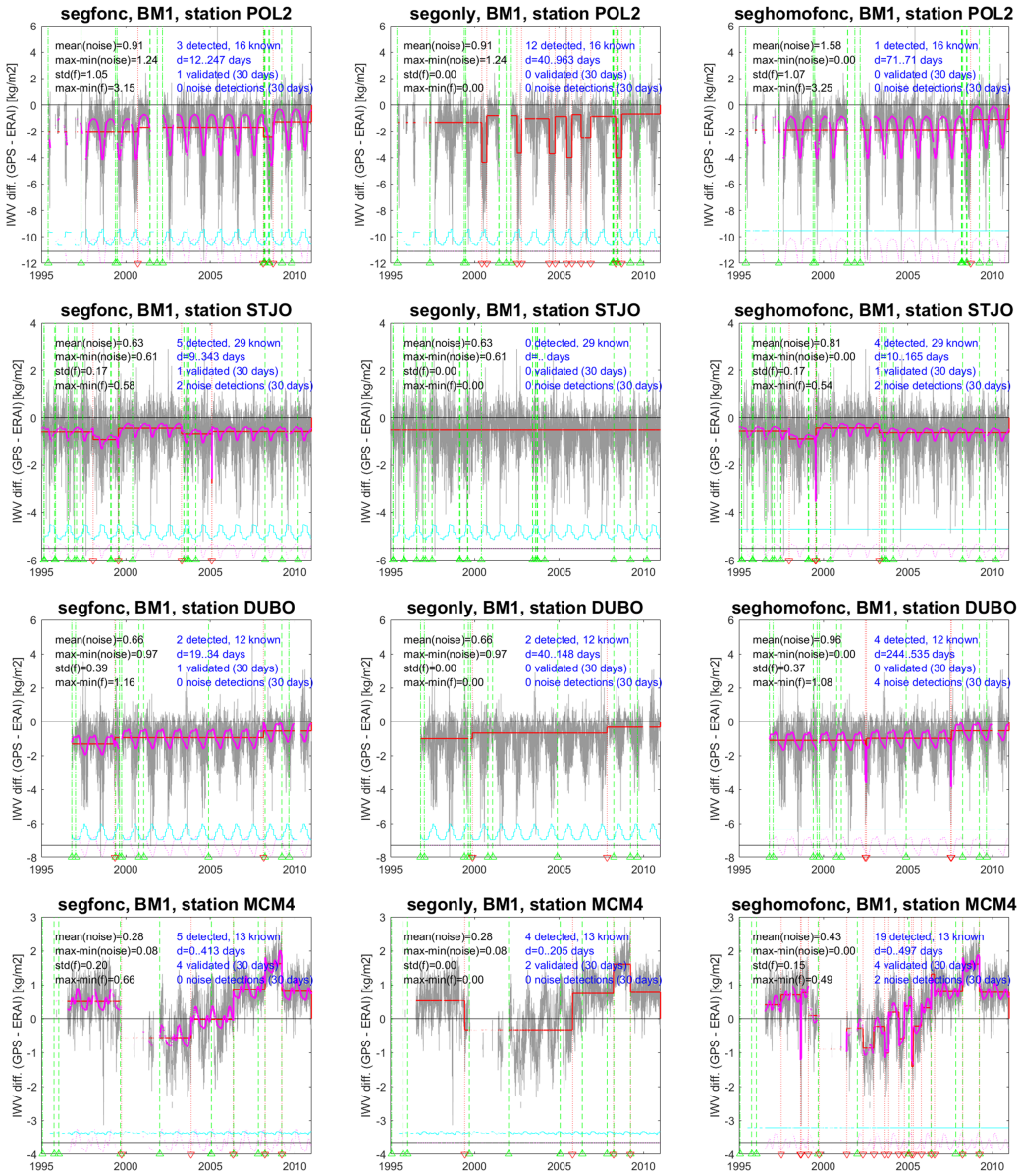


FIGURE 1.8 – Exemples de résultats obtenus avec les variantes (a), (c) et (d) du modèle de segmentation, de gauche à droite, pour quatre stations différentes : POL2, STJO, DUBO et MCM4 (de haut en bas). Le contenu des graphiques est similaire à celui de la figure 1.2 (b). Le ligne rouges verticales indiquent les ruptures détectées par la segmentation. Le texte inséré en haut à gauche des graphiques rapporte l'écart type moyen du bruit, la variation (max-min) de l'écart type du bruit, l'écart type de la fonction de biais périodique et la variation (max-min) de la fonction de biais périodique. Le texte en bleu indique le nombre total de détections et de changements connus, la distance minimale et maximale entre les ruptures détectées et les changements connus les plus proches, le nombre de détections validées et le nombre d'outliers ('noise detections') détectés avec un seuil de 30 jours.

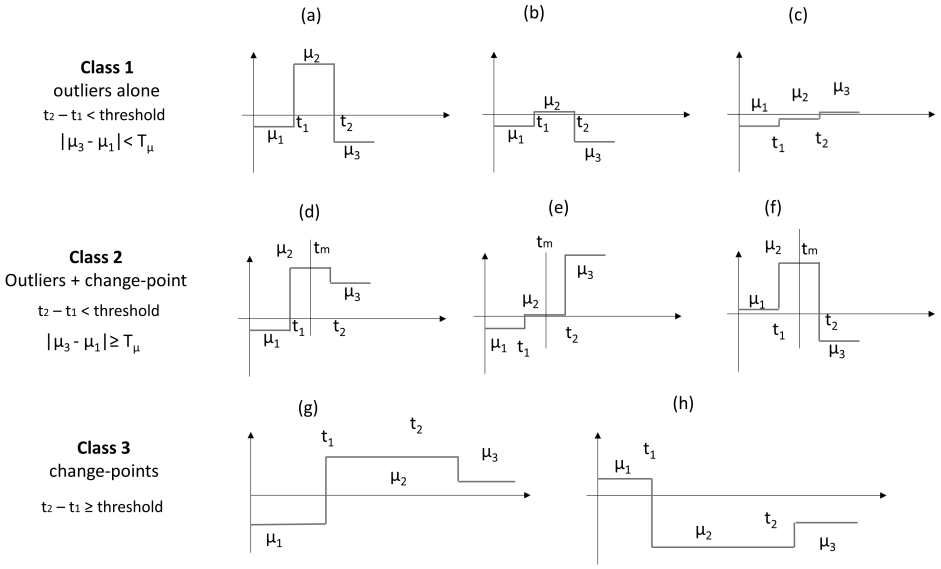


FIGURE 1.9 – Classification des ruptures détectées. Les classes 1 et 2 contiennent des détections aberrantes (outliers) définies comme telles car elles sont plus proches qu'un seuil (typ. entre 30 et 80 jours).

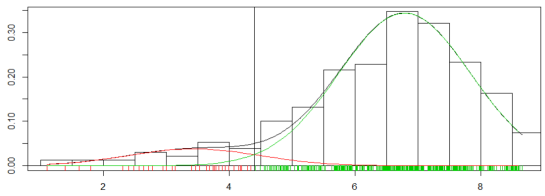


FIGURE 1.10 – La densité (logarithme de la longeur des segments pour BM1) de toutes les stations (en noir) et la densité de chacun des deux groupes (en rouge pour le premier et en vert pour le second) déterminés par un modèle de mélange. La ligne verticale noire indique la limite entre les deux groupes (81 jours) qui est optimale pour détecter les "outliers".

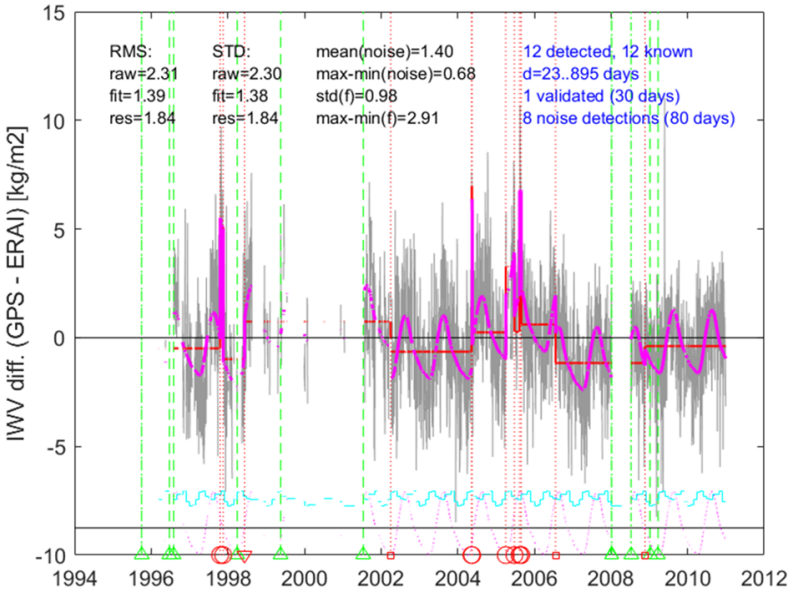


FIGURE 1.11 – Résultat de la segmentation pour la station IISC. Les lignes rouges pointillées verticales montrent les ruptures détectées et les lignes vertes pointillées verticales montrent les changements d'équipement à partir des métadonnées. Les symboles en bas indiquent le résultats de la classification des outliers : un carré rouge une rupture normale (classe 3), un cercle rouge indique une valeur aberrante (classe 1 ou 2), un triangle inversé rouge indique une rupture validée. Les valeurs aberrantes sont détectées avec un seuil de 80 jours et forment 3 clusters. Les variations de moyenne avant/après les clusters sont significatives (i.e. ils sont de classe 2).

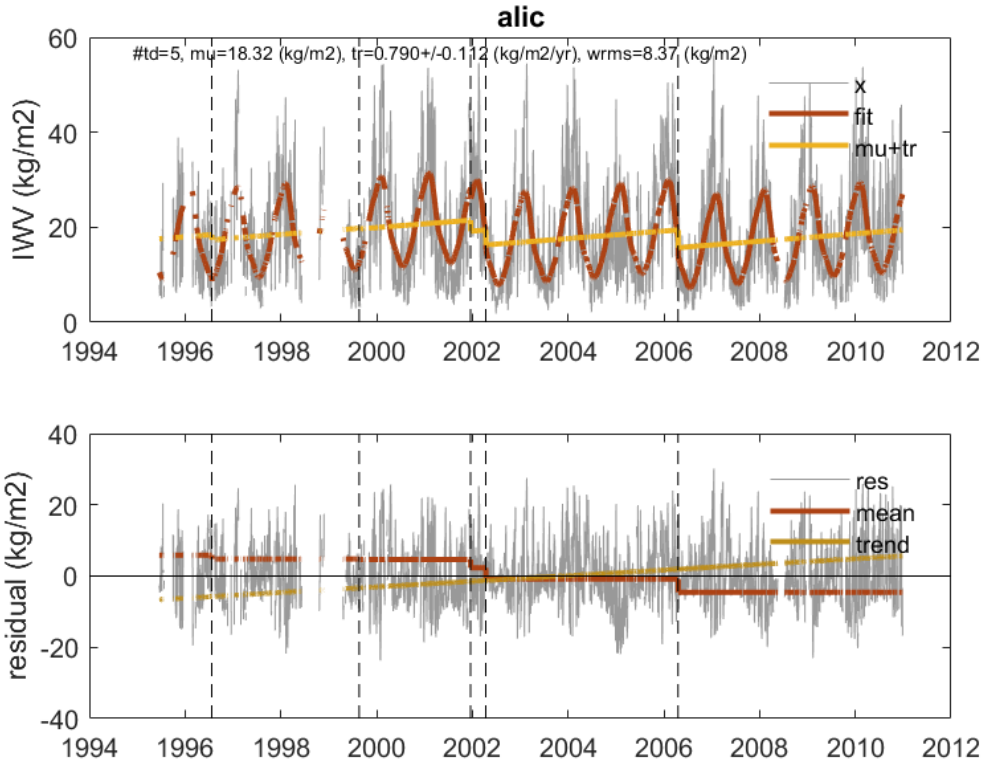


FIGURE 1.12 – Séries temporelles de GNSS IWV pour la station ALIC et modèle de tendance ajusté avec OLS : (en haut) la série est représentée en gris, la ligne rouge est le modèle ajusté et la ligne jaune est la tendance estimée + les moyennes, (en bas) la les résidus sont représentés en gris, les moyennes centrées en rouge et la tendance en jaune. Les lignes verticales noires en pointillé sont les ruptures détectées à partir de la segmentation (après le nettoyage). La valeur de tendance et son erreur standard sont données dans le graphique supérieur.

Chapter 2

Introduction

2.1 Context and problematic

2.1.1 Climate data analysis: definitions and basic concepts

Climate can be defined as the statistical description in terms of mean and variability of relevant physical quantities over a period of time ranging from months to thousands or millions of years (Planton [2013]). The quantities of interest are often those of direct impact on human life such as temperature, precipitation, and wind at the Earth's surface. Climate is thus tightly linked to weather and the distinction is made in terms of temporal periods mainly, such that climate is often considered as the long-term average of weather (typically over a period of 30 years). The state of the climate system is controlled from the interaction of five major components: the atmosphere, the hydrosphere, the cryosphere, the lithosphere and the biosphere, and the interactions between them. The system evolves in time under the influence of its own internal dynamics and because of external forcings (e.g. volcanic eruptions, solar variations, and anthropogenic forcings such as the changing composition of the atmosphere and land use change). Climate variability includes random variability (or noise) and more organised patterns or modes of variability with return periods ranging between a few years and thousands of years (Rohli & Vega [2018]). On the shorter time scales, El Niño Southern Oscillation (ENSO) is a well known pattern of tropical sea surface temperatures anomalies in the Pacific Ocean with worldwide effects and a period of the oscillation that typically varies between two and eight years (Wang [2018]). Understanding and predicting climate variability is important because of the impacts of extreme weather events (e.g. flash floods and landslides resulting from heavy precipitation, diseases and crop failures due to heat waves and droughts). Besides climate variability, one often invokes climate change which is more specifically referring to a change in the state of the climate (i.e. in the mean and/or the variability of its properties)

that persists for an extended period, typically decades or longer (Planton [2013]). Distinction should here also be made between the natural causes and those attributable to human activities, namely those altering the atmospheric composition such as green house gas (GHG) emissions responsible for the rise in average surface temperature in modern climate (ca 1850-present) known as global warming (IPCC, 2013). The effects of global warming include rising sea levels, regional changes in precipitation, more frequent extreme weather events such as heat waves, and expansion of deserts (IPCC, 2014).

The two main methods employed by climatologists are the analysis of observations and the modelling of the physical laws (processes) that determine the climate. Paleoclimatology is interested in the analysis variables such as temperature and precipitation over geological time scales (thousands to millions of years) the observations of which are provided indirectly by proxies (e.g. analysis of tree rings, or ice and sedimental cores). On the other hand, modern climatology uses direct and traceable measurements which are available for about two centuries and were primarily collected by meteorologists, both on land and on board ships. Because the initial purpose of those measurements was not for creating a long-term climate record, there is poor homogeneity in the data due to many changes in instrumentation and practice (Jones *et al.* [1986]). However, observations have long been the main source of information for understanding the physical processes determining the mean climate and its variations. Meteorological instruments have benefited the rapidly evolving technologies and gained in accuracy and spatial coverage, especially since the satellite era from the late 1970s. Modern climatology also extensively uses numerical models to study specific mechanisms in both idealized and real Earth system frameworks. Observations are crucial here as a ground truth for the validation of climate models simulations of the past. The better our climate models represent the past climate (i.e. the mean state and the variability including the extremes), the more confident we can be into their predictions of the future climate (Karl & Trenberth [2003]). Global climate models (GCMs) offer also a means to study the impact of various external climate forcings and especially separate natural and anthropogenic forcings, e.g. investigate the impact of the increase of atmospheric CO2 concentrations since the industrial revolution (Myhre *et al.* [2017]). Besides observations and climate models, atmospheric reanalyses are a hybrid modeling technique where a numerical model is run with constant assimilation of observations, hence correcting model defaults and drifts where and when they occur and serving as a physical interpolation tool where observations are lacking (Dee *et al.* [2011]). Currently, two types of reanalyses are available: modern reanalyses which assimilate surface and upper air data since the 1950s (see below) and satellite data since the late 1970s, and century reanalyses which assimilate only surface temperature, pressure, and wind data going back to the mid or late 1800s (https://reanalyses.org/).

The collection of climate observations is coordinated by several international bodies such as Global Climate Observing System (GCOS) and Detection of Atmospheric Composition Change (NDACC), see Appendix A for a description.

Long (decadal to centenial) records of observational data are essential for monitoring climate change and supporting climate research. However, long time series are often affected by inhomogeneties due to instrumentation or observer changes, station relocation and changes in the measurement conditions around the station (Jones *et al.* [1986]). These inhomogeneties manifest usually as abrupt, but also sometimes gradual, changes in the measured signals which are detrimental to estimating climate trends and variability (Thorne *et al.* [2005]). In many cases, the station history provides useful information on the causes for the changes but unfortunately not all changes are documented. Visual inspection of time series has been traditionally used to confirm documented change-points and to detect undocumented change-points. However, this task is time consuming and subjective. In recent years, a great deal of statistical methods have been developed to assist and automate the detection of change-points in climate series (see Peterson *et al.* [1998]; Reeves *et al.* [2007]; Venema *et al.* [2012]). They have been applied primarily to create high quality, homogenized, global and regional temperature based on the collection of surface and upper-air measurements as well as more recently on satellite data (Seidel *et al.* [2004]). New methods have been developed for other variables such as precipitation, surface pressure, wind speed, and humidity, and subsequent climate data sets have been released (Beaulieu *et al.* [2008]; Domonkos & Coll [2015]; Wan *et al.* [2007, 2010]; Willett *et al.* [2008]).

The homogenization of climate data consists in three main tasks: 1) the detection of change-points, i.e. the dates when the statistical properties of the studied variable undergo a significant change (e.g. in the mean or in the variability). This task is also referred to as the segmentation task because the time series is sliced into homogeneous sub-series. 2) the validation of the set of change-points issued from the previous task. This step consists in a critical analysis of the proposed segmentation result and aims at separating true detections from false detections. Metadata play an important role at this step because they can offer a rationale explanation to any observed or suspected change-point. 3) the correction of the raw data according to the set of validated change-points. This task consists in modifying the original data in order to remove any spurious non-climatic signal (e.g. by subtracting an estimate of a mean bias for a given period).

In this thesis we will be interested in a new data type, the integrated water vapour (IWV) estimates, which are derived from ground-based Global Positioning System (GPS) and more generally from Global Navigation Satellite Systems (GNSS) measurements (see Appendix B). The homogeneity of this new data type has been questioned recently in a small number of studies (Bock *et al.* [2010]; Ning *et al.* [2016]; Parracho *et al.* [2018]; Vey *et al.* [2009]). However, no specific homogenisation method has yet been developed that would account for the statistical properties of these data. The aim of this thesis is to fill this gap. Before going more into details in the next chapters, Section 2.1.2 will introduce and illustrate the homogeneity issues in surface temperature data and the traditional approach to the homogenization of this data type. Then Sections 2.1.3 and 2.1.4 will describe the specific features and characteristics of

the GNSS IWV data and explain to which extent the previous homogenization approaches need to be adapted to this new data type. Section 2.2 will introduce the general statistical framework in which this thesis will be developed, giving, in Section 2.2.1, an overview of the existing statistical methods proposed in the climate literature and, explaining, Section 2.2.2 the main difficulties faced in this context.

2.1.2 Inhomogeneities in climate data: the case of surface air temperature

Origin of inhomogeneities

Four major factors are listed by Jones *et al.* [1986] which affect meteorological station homogeneity: 1) changes in instrumentation, exposure and measurement techniques; 2) changes in station location; 3) changes in observation times and processing methods; 4) changes in the environment around the station, particularly with respect to urban growth. Instrumentation changes are desirable as new and more modern instruments are more accurate and stable, and provide often new capabilities (e.g. higher temporal sampling, automatic recalibration, etc.). However, each instrument has its own measurements biases and random errors such that an instrument change usually induces a change in the mean signal, also called a level shift (Lu & Lund [2007]), and possibly a change in the noise variance and autocorrelation (Lu *et al.* [2010]). In some special and rather uncommon cases the instruments may also show a drift which would manifest as a linear or a non-linear spurious signal and require a special signal processing procedure. Observer changes can induce inhomogeneities in time series namely when they involve a change in the time of measurement, in the instrument preparation or calibration, or type setting errors (e.g. when the observations are manually transmitted by the observer). Station location can have a direct impact on the measurements because the meteorological conditions are location dependent, e.g. temperature and pressure change rapidly with altitude, and wind speed near the surface depends on the roughness of the station environment (e.g. open meadows, small bushes, and trees around can alter differently the surface wind speed). Changes in instrumentation, location, and practice are in principle recorded in the station history, so called metadata. However, metadata records are notoriously incomplete (Li & Lund [2015]). Moreover, not every change listed in the metadata necessarily induces a shift in the time series.

Figure 2.1 shows an example of annual temperatures at Tuscaloosa, Alabama, discussed by Lu & Lund [2007]. This century-long time series shows a clear year-to-year scatter that can be attributed to inter-annual climate variability along with two change-points, in 1939 and 1957. These change-points coincide with a change of thermometer (1939) and a station relocation (1957) as recorded in the station history. Another station relocation is mentioned in 1987 but this one cannot be detected in the time series. Superposed to the time series are linear trend fits when the two change-points are ignored and taken into account, respectively. A linear trend estimate and one standard error are -0.3023 +/-

0.7755°C/Century when shifts are ignored and 3.5166 +/- 0.4186°C/Century when shifts are taken into account. It is visually evident that the level-shift adjusted trend is preferable, and hence the authors to conclude that temperature has been increasing by more than 3°C at Tuscaloosa over the 20th century. Comparatively, the unadjusted trend concludes on a temperature decrease at this site. Taking the change-points into account in the signal analysis can thus critically change the conclusions regarding the climatic trends. Taking change-points into account has also an impact on the autocorrelation of the residuals as shown in Figure 2.2. Lu & Lund [2007] conclude that the memory structure in a model ignoring the abrupt changes typically needs to be longer than for a model where the shift information is taken into account.

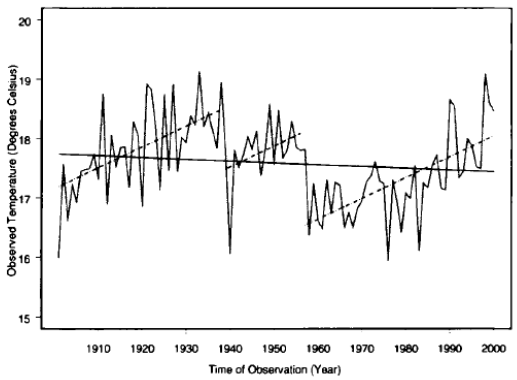


Figure 2.1 – Yearly temperatures at Tuscaloosa, Alabama, with least squares trends. Source: Lu & Lund [2007]

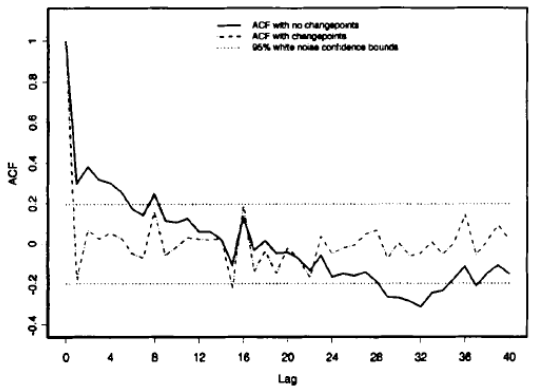


Figure 2.2 – Sample autocorrelations of ordinary least squares residuals. Source: Lu & Lund [2007]

Usage of metadata

In the example above the authors have assumed that the change-points occur at known times, which were taken from the station history record. Not all known change-points were included in the trend plus level shift model fit, however. A preselection was made based on t-tests for changes in the means and F-tests for changes in the means and trend where they concluded that the 1987 change-points was not significant and that a single overall trend slope was a better choice than a trend per segment. This simple approach cannot be applied in most cases, however, because the metadata are often incomplete. The more general problem is thus one in which the number of change-points and their times are both unknown. Many different statistical methods have been developed over the past years to tackle this problem (Costa & Soares [2009]; Peterson *et al.* [1998]; Reeves *et al.* [2007]; Venema *et al.* [2012]). Section 2.2.1 below gives a short review of the methods. Metadata remain nevertheless a useful source of information which are used either to validate the detected change-points or to formulate a prior distribution in a Bayesian approach (Li & Lund [2015]). WMO guidelines on metadata and homogenization are provided in Aguilar *et al.* [2003].

Absolute and relative segmentation approaches

Figure 2.1 showed an example of an annual temperature time series in which a linear trend could easily be detected and interpreted as a climate signal (i.e. a warming of the air close to the surface over the 20th century). To properly estimate this trend the authors took both effects into account simultaneously (trend and offsets) in the homogenization method. Similarly, when monthly time series are analysed the seasonal variations, which are a dominant component in the signal at that temporal scale, need to be taken into account. Figure 2.3(a) shows the monthly temperatures at Tuscaloosa, Alabama, discussed by Lu *et al.* [2010] in which the seasonal variations are the main visible feature. Detecting the change-points directly in this time series would be very difficult. To overcome this difficulty, two main approaches are commonly used. 1) When the data from the target station are considered in standalone the approach is said to be absolute. The seasonal variations can here be handled in two ways: i) mean seasonal variation is removed by subtracting to each month the mean value for that month. The resulting monthly anomaly signal is then analyzed. ii) the mean seasonal variation is represented by either a parametric or a non-parametric model and its parameters are fitted simultaneously with the trends and offsets. 2) The second approach consists in using a reference series from one or several nearby stations which are exposed to the same climate signal (i.e. including the seasonal but also the trend component). This segmentation approach is referred to as the relative approach. Figure 2.3(b) shows the monthly temperature adjusted for the seasonal mean for Tuscaloosa, Alabama, along with the fitted trend and change-points. Two change-points are detected in April 1939 and July 1957, respectively, and the trend slope estimate is 0.00258 °C/month (3.10 °C/century) +/- 0.00039 °C/month (0.47 °C/century). In that case the authors

used a statistical segmentation method which estimates the change-point number, locations, and the time series regression parameters (Lu *et al.* [2010]). The method will be described in more details in Section 1.2. The two change-points detected by the method are validated with the station metadata and the estimated trend slope is consistent with the estimate found from the annual data discussed above.

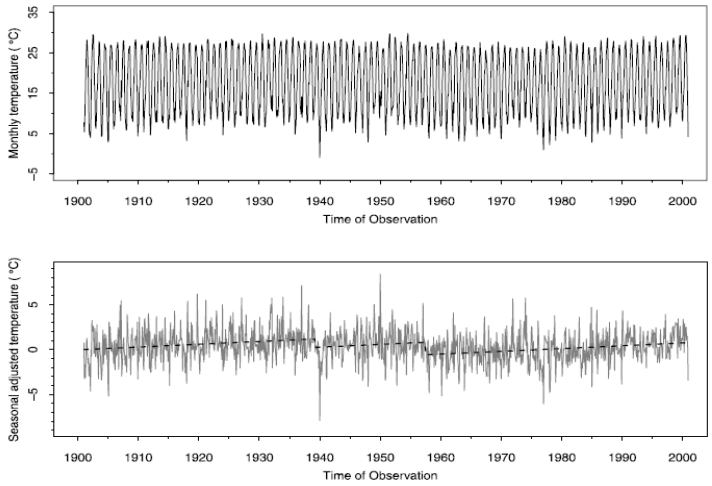


Figure 2.3 – The Tuscaloosa data with change-point structure imposed. Source: Lu *et al.* [2010].

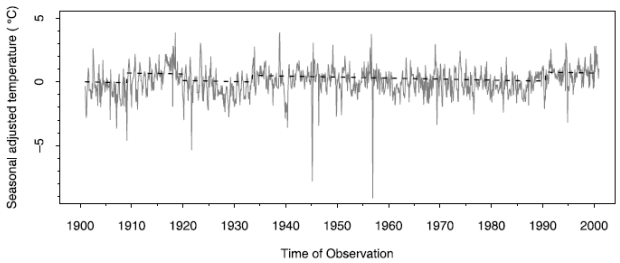


Figure 2.4 – The Tuscaloosa minus the reference data with change-point structure imposed. Source: Lu *et al.* [2010]

Figure 2.4 shows the results when the relative approach is used, i.e. when the segmentation is applied to Tuscaloosa minus a reference series that averages three neighboring stations (Lu *et al.* [2010]). Here four change-points are detected. None of them is found within the same year as the metadata information (station changes are occured in 1921, 1939, 1957 and 1987). However, three of them are within 3 or 6 years and the authors to conclude that the relative analysis is superior to the absolute one. The extra change-point in 1909 is possibly attributed by the authors to inhomogeneities in the reference series at two of the neighboring stations. Though there is consensus that relative methods are

generally superior to absolute methods (see e.g. Venema, 2012), this example points nevertheless to the difficulty of handling inhomogneities in the reference series. Much effort has actually been devoted to the development of multiple change-point methods which tackle this issue (e.g. Caussinus & Mestre [2004]; Menne & Williams [2009]).

2.1.3 The role of water vapour in climate

Water vapor is a key component of the global hydrologic cycle and plays a major role in many atmospheric processes contributing to the weather and climate. It is essential for the development of disturbed weather, influences the planetary radiative balance, and influences surface fluxes and soil moisture (Sherwood *et al.* [2010]). The latent heat released when atmospheric water vapor condenses and the cooling of air through evaporation or sublimation of condensate affect strongly atmospheric circulations making water vapour an active player in dynamic processes that shape the global circulation of the atmosphere (Schneider & Levine [2010]). Through its abundance (about 0.2% of the total mass of atmospheric gases), it is also the dominant greenhouse gas in the atmosphere. However, compared to the long-lived greenhouse gases (CO2, CH4, N2O...) which act as the drivers of the greenhouse effect, water vapour acts as fast feedback variable. The water vapor feedback is the process whereby an initial warming of the planet, caused, for example, by an increase in atmospheric carbon dioxide, leads to an increase in the humidity of the atmosphere. Because water vapor is itself a greenhouse gas, this increase in humidity causes enhanced warming by a factor of 2 to 3 of the initial warming (Held & Soden [2000]). The rate at which water vapour increases per 1 K of temperature increase is controlled by the Clausius–Clapeyron (CC) relation (see e.g. Held & Soden [2006]). Under the assumption of constant relative humidity, this rate is about 7% K^{-1}. At global scale, observational and modelling studies have suggested that the relative humidity is maintained and that water vapour in the atmosphere closely follows the temperature in agreement with the C-C equation (Held & Soden [2006]; Semenov & Bengtsson [2002]). However, at a regional scale, deviations from C-C law are observed and the strength of the feedback can vary largely (O'Gorman & Muller [2010]). In addition, the short residence time of water vapour in the atmosphere and its small scale variability make its representation in global weather and climate models extremely challenging (Sherwood *et al.* [2010]). Ground-based observational networks and satellite missions are thus important sources of moisture information for monitoring the change in atmospheric composition in the context of global warming, constraining atmospheric reanalyses and validating climate model simulations.

Global and regional trends in temperature and water vapour have been shown to vary considerably among the various climate models due to differences in numerical approximation and physical parameterizations (IPCC, 2014). The differences among the IPCC AR5 models in the tropics lie in a range of factors up to 1:5 (see Flato *et al.* [2013] P774, Fig. 9.9). Accurate and homogeneous water vapour

observations are expected to play a central role in quantifying the uncertainty and identifying the errors sources in the models. In this respect, Parracho *et al.* [2018] showed that a major issue in the global climate models may be linked with atmospheric circulation that controls decadal moisture trends and interannual variability. Water vapour trends are also a good diagnostic of the homogeneity of reanalyses and satellite data. For example, Schröder *et al.* [2016], compared several reanalyses and satellite data sets and identified change-points in the reanalyses that coincided with changes in the observing system (e.g. start and end of assimilation of satellite data in the reanalyses).

Humidity measurements in the troposphere have been made since the 1950s using radiosonde balloons equipped with pressure, temperature, and humidity sensors. But the global radiosonde data record covers mostly the Northern Hemisphere, and its usefulness for climate monitoring is limited by errors and biases associated with the instruments and by discontinuities due to changes in sensors and procedures over time (Dai *et al.* [2011]; Ross & Elliott [2001]; Wang & Carlson [2001]). Total column water vapor (TCWV) and profiles of tropospheric humidity have been measured by satellites since the end of the 1970s, but long-term trends are difficult to derive from satellite data because of intercalibration limitations.

2.1.4 Inhomogeneities in GNSS IWV data

The ground-based GNSS receivers are presently one of the most reliable techniques for sensing TCWV, also referred to as IWV and precipitable water (PW), with accuracy at the level of 1–2 kg m^{-2} *or* 5% *in all weather conditions* (*Bocket al.* [2007]; *Wang et al.* [2007]). *GNSS IWV has been extensively used for detecting humidity biase*

Sources of inhomogeneity in GNSS IWV data

It is obvious that changes in instrumentation occurring either in the space segment (e.g. replacement of older generation satellites by new satellites) or, more often, at the ground level are susceptible to break homogeneity of the raw measurements. Fortunately, GNSS ground stations are rather robust devices which don't need to be changed often except e.g. to adapt to the new emerging satellite constellations. Hence, all tracking stations switched progressively from GPS only to GPS, GLONASS, Galileo, and Beidou complient instrumentation over the past years.

The replacement of an antenna is likely to introduce a change in the antenna's phase center offset (PCO) and phase center variation (PCV) map (the variation of phase as a function of the radio wave incidence angle). Phase center offsets induce an uncertainty in the position of the antenna reference point and thus alter the stability of the station coordinates but have only marginal impact on the ZTD stability. A change in PCV on the other hand can induce a small bias both in station coordinates (especially in the vertical component) and in ZTDs. In order to minimize this effect, the IGS introduced antenna calibration models for all tracking stations used in the IGS network. Nevertheless, it may happen that for older antenna types the PCO/PCV models are not as accurate as for recent types and the antenna

change may induce a small offset in the ZTD time series (see the example of CCJM below).

Many permanent GNSS stations are equipped with antenna radomes as a means of protection against general wear, to prevent the buildup of debris and snow, and to discourage people and animals from disturbing the antenna (https://kb.unavco.org/kb/article/unavco-resources-radomes-520.html). Antenna radomes affect the signal propagation thereby altering the antenna's absolute PCO and PCV. When the appropriate PCO and PCV models for each antenna and radome combination are applied during the data processing, this effect should only produce a small bias. Vey et al., 2009, illustrated the effect of a combined change of antenna and radome which produced an offset in IWV of -1.3 kg m-2 at station HOFN (Iceland). Smaller offsets were also reported for other stations by these authors.

GNSS receivers are high-tech electronic devices which benefit from continous improvement and evolutions. In a receiver lifetime, a large number of firmware updates are usually applied: ca 70% of the receiver changes reported in the IGS site logs are firmware updates while 30%are hardware changes. Hardware changes most often occur after a system failure which can be detected from the drop in the number of daily measurements and the increase in the noise, drift and/or other spurious signal in the coordinates and ZTDs (Parracho *et al.* [2018]; Vey *et al.* [2009]). Periods of receiver malfunctioning should be removed from the analysed IWV time series. Fortunately, these cases are rare. In regular operations, the main mechanism through which a receiver change can induce a change in the mean ZTD estimate is via the satellite geometry, e.g. when the tracking capabilities including more or less low elevation satellites. Changing the elevation cutoff angle in the receiver software can have a similar effect. The reason is that PCV errors, mapping function errors and satellite geometry determined by the cutoff angle act together to determine the ZTD and station height bias.

The impact of multipath is difficult to predict, partly because the electric and magnetic properties of the stations' environments are not well determined and partly because the coupling effects with metallic objects nearby is complex and highly variable. Multipath changes can be natural such as growing/declining of vegetation, or due to human interventions, e.g. cutting of vegetation and construction of buildings nearby. They would induce either gradual or abrupt changes in ZTD estimates, respectively.

IGS elaborated recommendations for installing permanent stations, including the siting and monumentation rules to protect from excessive multipath and guarantee long term positioning stability (https://kb.igs.org/hc/en-us/articles/202011433-Current-IGS-Site-Guidelines). Station history information is recorded in the so-called IGS site logs which are intended to report all instrumentation changes (receiver, antenna, radome, external clock, meteorological sensor, etc.), also including firmware updates and elevation cutoff changes. However, the description of the station environment in terms of elevation mask is optional. Besides the site logs, information on the data quality is produced daily by IGS data centres using TEQC software (Estey & Meertens [1999]) as well as by the IGS analysis centres as a by-product of the data processing; the latter are software-dependent however. The analysis of both information types could be used to create additional metadata, e.g. to detect a change in the receiver

tracking performance, in satellite health, in station multipath, etc.

Homogenization of GNSS IWV series

First, it should be recalled that the detection of discontinuities in GNSS station coordinate solutions has been practiced for a long time in the geodetic community, either for computing terrestrial reference frames (Collilieux *et al.* [2011]) or analysing station velocities due to tectonic motions (Williams [2003]). Until now, change-points were usually detected by visual inspection of time series. Statistical segmentation methods have also been tested in the framework of Detection of Offsets in GPS time series Experiment (DOGEx) but the study concluded that manual methods almost always gave better results than automated or semi-automated methods (Gazeaux *et al.* [2015]). More recently, new statistical approaches have been developed which show more promising results; they are the semi-parametric segmentation of multiple series (Bertin *et al.* [2017]) and a factor model approach for the joint segmentation with between-series correlation (Collilieux *et al.* [2019]). Both approaches have been tailored to fit specific characteristics of a global network of GNSS station coordinate solutions. They include namely common biases represented by a functional part or between-station correlations. In all cases the change-points parameters remain station specific (i.e. independent from one station to another) and the segmentation model is based on the absolute approach (see Section 2.1.2). Following the same methodology, this thesis aims at adapting existing segmentation methods and developing new methods that fit the statistical properties of IWV time series. GNSS IWV time series are structurally very different from GNSS coordinate time series (they include flicker noise and only small seasonal signals), and the hypothesis of common time-dependent biases or spatial correlations are weak. For these reasons the above-mentioned methods don't apply.

In the GNSS-derived IWV daily series the signal is stronger due to the natural variability of the IWV variable, making more difficult to find the abrupt changes. It is thus mandatory to use a relative homogenization technique.

In this work, we analysed GNSS data from the IGS network shown in Figure 2.5. Because this network is quite sparse, the construction of references series from neighboring stations is hard. Therefore we used the ERA-Interim reanalysis Dee *et al.* [2011] as a reference and analyzed the IWV differences between the GNSS and the reanalysis data: $\Delta IWV = IWV_{GPS} - IWV_{ERAI}$. It was shown in a previous study that the ERA-Interim reanalysis represents well the signal of the atmospheric variability Parracho *et al.* [2018].

Characteristics of GNSS IWV series

Figure 2.6(a) shows a time series of daily GNSS IWV data from the IGS repro1 data set Bock [2017]. The IWV time series at this tropical station (CCJM, Japan, located north of the Philippines Sea) exhibits a

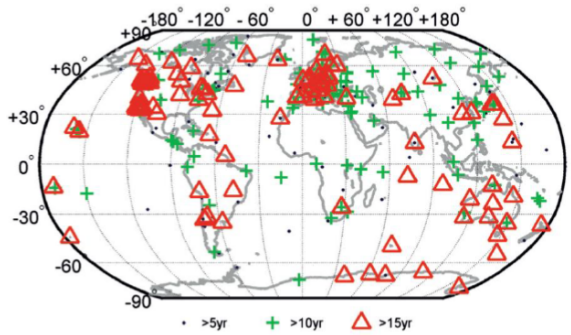
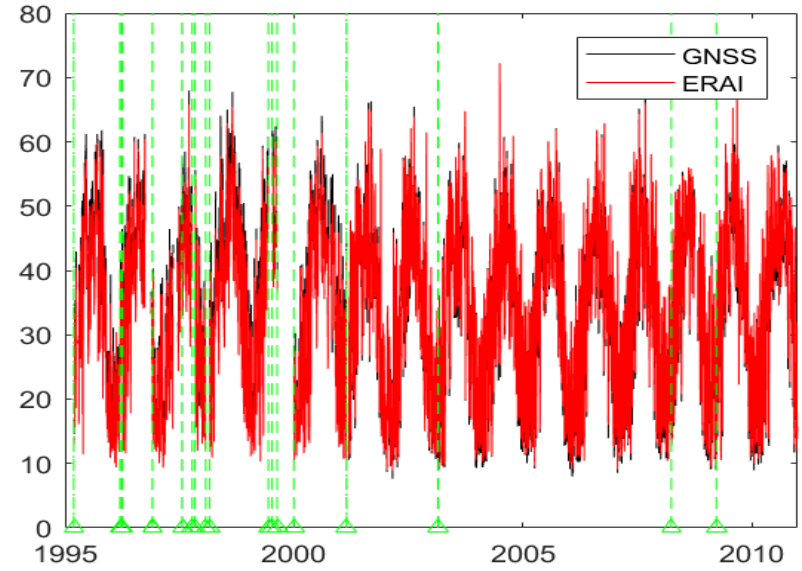


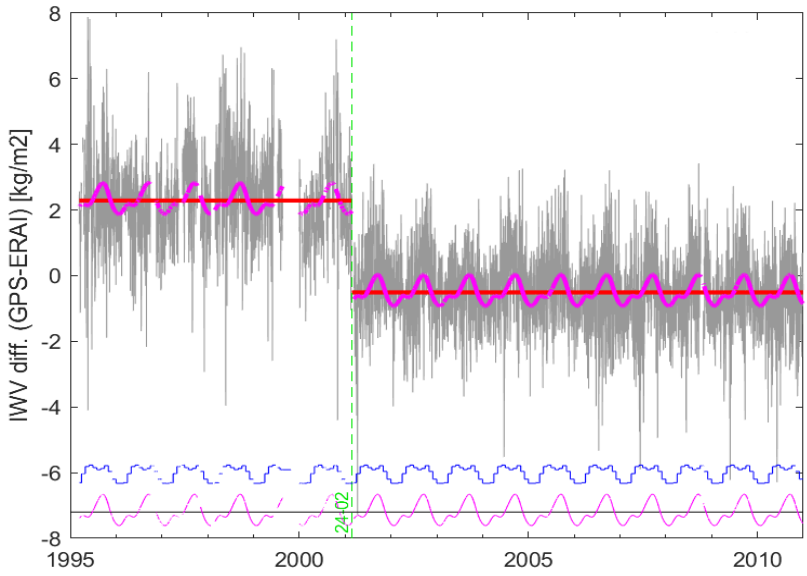
Figure 2.5 – Distribution of 460 GNSS stations available from the IGS repro1 dataset covering the period from 1 January 1995 to 31 December 2010. The different markers represent the length of the time series. Among the 460 stations, 120 have time series longer than 15 years. Source: Bock [2014].

marked seasonal variation, with values varying from 10 kgm^{-2} to 60 kgm^{-2} between winter and summer, as well as a strong day-to-day variability linked to the changing weather situations. The figure also shows the IWV time series from the ERA-Interim reanalysis (Dee *et al.* [2011]). It can be seen that the two data sets are in good agreement. They represent the seasonal and daily variations with high similarity. The vertical dotted lines show the known equipment changes for the GNSS station. They include the dates of receiver and antenna changes as found in the IGS site log files, and also two processing changes in 2008 and 2009. It is not obvious from the inspection of the time series if any change induces a break in the GNSS time series.

Figure 2.6(b) shows the time series of IWV differences, $IWV_{GPS} - IWV_{ERAI}$. This plot reveals one obvious change-point on 24 Feb 2001 which coincides with a change of receiver and antenna at the station. The other known changes don't seem to produce inhomogeneities (at least no change in the mean difference can be detected visually). Following the same approach as Lu & Lund [2007] we fitted a least-squares model to the ΔIWV data with a change in the mean signal at a known time. Our model did not include a trend because we analyse a difference series (target minus reference). On the other hand, following the approach of Collilieux *et al.* [2019], we model the presence of a periodic bias with a Fourier series of order 4 with a base period of 1 year (365.25 days) and harmonics of 1/2, 1/3 and 1/4 of a year. The jump in the means is estimated to 2.8 kgm^{-2}. Among the 120 stations that we examined, this offset is actually the largest that was encountered. The reason why this offset is that large is because the particular antenna and radome models that were used during the processing of the former period were incorrect (they were actually not known and alternate models were used while instead the data should not have been processed). During the later period the proper model was available and used, and the bias with respect to ERA-Interim disappeared. Another feature present in the Figure is the periodic bias represented by the magenta line. This bias is explained by small differences in representativeness



(a)



(b)

Figure 2.6 – (*a*) GNSS and ERA-Interim IWV time series at station CCJM located in the sea of the Philippines (27.096°N, 142.185°E).(*b*) IWV difference (GPS - ERA-Interim) in grey shading. The vertical green lines show the equipment changes documented in the metadata.

between the two data sets. Indeed, the GNSS observations can capture some small-scale variability not resolved by the reanalysis (Bock & Parracho [2019]). Although the reanalysis is the best available reference data set we can have, it is not perfect and this differences must be taken into account in the segmentation model in order to avoid the over-detection of change-points in place of variations in the bias. Finally, a third feature shown in this example is the annual variation in the variance show by the blue line which represents the monthly standard deviation of the daily post-fit residuals. Few existing segmentation models take this feature into account. Finally, we also computed the lag-1 of the autocorrelation function of the residuals and found a value of r = 0.249. This value is relatively small and allows for neglecting the serial dependence in the $IWV_{GPS} - IWV_{ERAI}$ differences in a first step.

2.2 Statistical framework for change-point detection

Change-point detection analysis constitutes an important and active current area of statistics. The proposed statistical methods dedicated to the specific climate field were naturally inspired by the many statistical methods developed in more general settings. The purpose consists in identifying instants, positions or dates where the statistical properties of the data before and after these instants are different, typically in the distribution. These change-points delimit what are called segments. This change-point detection problem is an old subject in statistics, dating back to 1954 Page [1954] with testing a potential single change-point. Over the decades, change-point methods have developed rapidly and intensively with multiple change-points, different types of data and other assumptions. A big distinction between them concerns the objective itself of the detection: *on-line* and *off-line* detection. The former consists in detecting change-points as soon as they occur in the time series, whereas the *off-line* approach consists in detecting all the change-points once the whole series is observed. In this thesis, we consider the *off-line* detection of multiple change-points. We can refer to Ardia *et al.* [2019]; Jandhyala *et al.* [2013]; Truong *et al.* [2020] for a review, admittedly not exhaustive, of recent numerous methods.

Section 2.2.1 presents a state-of-art of the statistical methods proposed in climate field literature. In this thesis, we focus on parametric methods in a frequentist framework using penalized criteria. In Section 2.2.2, we precise the main difficulties that arise in this approach which are both statistical (for the choice of the number of change-points) and algorithmic (for the location of change-points) and present the current solutions proposed in the literature.

2.2.1 Overview of change-point detection methods in climate

This section aims at presenting the main homogenization methods proposed in the climate literature in the sense that they are the most used and/or adapted by the community. In particular, they are listed in various review papers (see Easterling & Peterson [1995], Peterson *et al.* [1998], Aguilar *et al.* [2003],

Reeves *et al.* [2007], Ducré-Robitaille *et al.* [2003], Ribeiro *et al.* [2016]) and have been also assessed in the COST action HOME Venema *et al.* [2012] for monthly temperature and precipitation observations and, more recently, in COST Action GNSS4SWEC for GNSS IWV Van Malderen *et al.* [2020]. These methods are detailed in Table 2.1 with different types of information. The two first columns contain the name and the reference of the method, following by the characteristics of the data to which they are applied (the time resolution and the type of series in terms of relative or absolute). The last two columns list the papers in which the method is used or adapted and the review papers in which it appears respectively. For some of these methods, free software is available. Table 2.2 gives the list of the software by specifying the name of the latest version, its computer language and availability. The methods are based on different statistical approaches and tools. Among them, we have the classical distinctions between parametric versus non-parametric methods and frequentist versus Bayesian approaches, and the use of classical inference procedures based on tests or maximum penalized likelihood procedures. Each of these points is precised for the different methods in columns from 5 to 8 of Table 2.1 providing details as for example about the used or developed tests, with an algorithmic point of view in column 9. Information about specific features taken into account in the modeling is given in column 10. Based on these statistical and algorithmical considerations, we built a possible classification between the methods in the form of a tree given in Figure 2.7 allowing to have a quick and more readable vision of the difference between them. The numbers appearing in this tree correspond to the numbers of the methods listed in Table 2.1. In the following paragraphs, we explain and detail each of the criteria used to build this classification tree (columns from 5 to 10 of Table 2.1 or levels of the tree reported on the left of Figure 2.7).

Method			Data		Statistical approach					References	
num	name	reference	time step	comp	model	approach	inference	search	specific features	users	reviews
1[a]	PRODIGE	Caussinus & Mestre [2004]	y, m	relative (pairwise)	parametric	frequentist	penalized likelihood (CL)	DP (optimal)		Venema *et al.* [2012]	Aguilar *et al.* [2003]; Ribeiro *et al.* [2016]
2[a]	ACMANT (Adapted Caussinus-Mestre Algorithm for Networks of Temperature Series)	Domonkos & Coll [2015, 2017]; Domonkos *et al.* [2011]	y, m, d	relative	parametric	frequentist	penalized likelihood (CL)	DP (optimal)		Van Malderen *et al.* [2020]; Venema *et al.* [2012]	Ribeiro *et al.* [2016]
3[a]	HOMER (HOMogenizaton softwarE in R)	Mestre *et al.* [2013]	y, m	relative (pairwise)	parametric	frequentist	penalized likelihood (mBIC)	DP (optimal)			
4	MFixedHetero	Bock *et al.* [2018]	d	relative	parametric	frequentist	penalized likelihood (BM, Lavielle, mBIC)	DP (optimal)	monthly variance		
5		Lu *et al.* [2010]	y, m, d	relative, absolute	parametric	frequentist	penalized likelihood (MDL)	GA (sub-optimal)	AR(p), trend, periodic (mean, variance, correlation)		
6		Li & Lund [2012]	y	relative, absolute	parametric	frequentist	penalized likelihood (MDL)	GA (sub-optimal)	non gaussian, AR(1)		
7	SNHT(Standard Normal Homogeneity test)	Alexandersson [1986]	y	relative	parametric	frequentist	test (MLR)	BS (sub-optimal)		Alexandersson & Moberg [1997]; Van Malderen *et al.* [2020]; Venema *et al.* [2012]	Reeves *et al.* [2007]; Ribeiro *et al.* [2016]
8[a]	MASH (Multiple Analysis of Series for Homogenisation)	Rimoczi-Paal *et al.* [1999]; Szentimrey [2007, 2008]	y, m, d	relative	parametric	frequentist	test (MLR)	Multiple Comparison (sub-optimal)		Auer *et al.* [2005]; Venema *et al.* [2012]	Aguilar *et al.* [2003]
9		Lund & Reeves [2002]	y	relative, absolute	parametric	frequentist	test (F test)	successive detections and corrections (sub-optimal)	change in mean, trend	Wang [2003]	
10[a]	RHtest	Wang [2008a,b]; Wang *et al.* [2007]	y, m	relative, absolute	parametric	frequentist	test (PMT, PMF, PMTred, PMFred)	BS (sub-optimal)	AR(1) in PMTred, PMFred	Venema *et al.* [2012] (PMF), Ning *et al.* [2016] (PMTred)	Ribeiro *et al.* [2016]
11		Lu & Lund [2007]	m, d, h	relative, absolute	parametric	frequentist	test (F-test)	one change-point	AR(p), trend		Reeves *et al.* [2007]
12		Menne *et al.* [2009]	m	relative	parametric	frequentist	test (MLR)	BS (sub-optimal)		Venema *et al.* [2012]	Ribeiro *et al.* [2016]
13[a]	AnClim	Stepanek *et al.* [2009]	y, m, d	relative	parametric	frequentist	test (MLR)	BS (sub-optimal)		Venema *et al.* [2012]	Ribeiro *et al.* [2016]
14[a]	RHtests_dlyPrcp	Wang *et al.* [2010]	d	absolute	parametric	frequentist	test (PMFred)	BS (sub-optimal)	non gaussian, AR(1)		
15[a]	Climatol	Guijarro [2011, 2013, 2018]	m, d	relative	parametric	frequentist	test (t-test, SRMD, SNHT)	BS, moving windows (sub-optimal)		Van Malderen *et al.* [2020]; Venema *et al.* [2012]	Ribeiro *et al.* [2016]
16		Vincent [1998]	y	relative	parametric	frequentist	test	multiple nested model tests (sub-optimal)	change in mean and trend	Bonsal *et al.* [2001]; Vincent & Gullett [1999]	Reeves *et al.* [2007]
17		Seidou *et al.* [2007]	y	relative, absolute	parametric	Bayesian	likelihood	one change-point			Beaulieu *et al.* [2008]; Ribeiro *et al.* [2016]
18		Seidou & Ouarda [2007]	y	relative, absolute	parametric	Bayesian	likelihood	optimal			Beaulieu *et al.* [2008]; Ribeiro *et al.* [2016]
19		Li & Lund [2015]	y, m	relative, absolute	parametric	empirical Bayes	likelihood (Bayesian MDL)	MCMC (sub-optimal)	metadata, AR(1)	Li *et al.* [2019]	
20		Hewaarachchi *et al.* [2017]	d	relative, absolute	parametric	empirical Bayes	likelihood (Bayesian MDL)	GA (sub-optimal)	metadata, AR(1), periodic (mean, variance, correlation)		
21		Karl & Williams [1987]	y	relative, absolute	non-parametric	frequentist	test (Mann-Whitney)	BS (sub-optimal)		Ducré-Robitaille *et al.* [2003]; LANZANTE [1996]; Van Malderen *et al.* [2020]	Reeves *et al.* [2007]
22		Craddock [1979]	m	relative	/	/	visual	visual		Brunetti [2009]; Venema *et al.* [2012]	Ribeiro *et al.* [2016]

Table 2.1 – Review of statistical methods in climate field. The abbreviations are defined in Table 2.3

[a] existing software

num	latest version	language	url
1	PRODIGE		
2	ACMANT3	Windows executables	http://www.c3.urv.cat/softdata.php
3	HOMER	R package	http://www.c3.urv.cat/softdata.php
8	MASH v3	Windows executables	https://www.met.hu/en/omsz/rendezvenyek/homogenization_and_interpolation/software/
10	RHtestV4	R package, FORTRAN	https://github.com/ECCC-CDAS/RHtests
13	AnClim	Windows executables	http://www.climahom.eu/software-solution/anclim
14	RHtests_dlyPrcp	R package	https://github.com/ECCC-CDAS/RHtests
15	CLIMATOL3.1.1	R Package	https://cran.r-project.org/web/packages/climatol/index.html

Table 2.2 – Free software implementing some of the methods listed in Table 2.1.

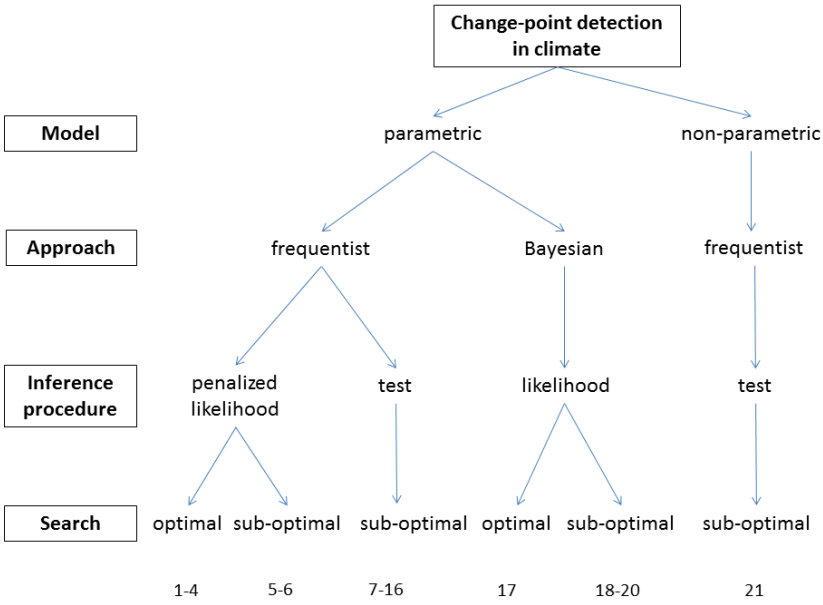


Figure 2.7 – Statistical classification of change-point methods in climate. The levels of the leafs are reported on the left of the tree and numbers at the end of each branch refer to change-point methods detailed in the Table 2.1.

abbreviation	definition
y, m, d, h	year, month, day, hour
CL	Caussinus & Lyazrhi
BM	Birgé & Massart
mBIC	Modified Bayesian information criterion
MDL	Minimum Description Length
DP	Dynamic Programming
GA	Genetic Algorithm
MLR	Maximum Likelihood Ratio
AR	autoregressive
BS	Binary Segmentation
MCMC	Markov Chain Monte Carlo
PMF	Penalized Maximal F test
PMT	Penalized Maximal t test
SRMD	Squared Relative Mean Difference

Table 2.3 – Abbreviations used in Table 2.1.

Models. A first distinction that can be made between the methods is based on parametric or non-parametric models. The latter assumes the data to be distribution-free contrary to the former for which the data are supposed to be drawn from a particular probability distribution. As we can observe in Table 2.1 or in Figure 2.2, all the proposed methods in climate are parametric, except the ones proposed by Ducré-Robitaille *et al.* [2003]; Karl & Williams [1987]; LANZANTE [1996]; Reeves *et al.* [2007]; Van Malderen *et al.* [2020] which are based on the classical non-parametric Mann-Whitney test. The parametric approach in segmentation consists in assuming that the series is modeled by a P_θ distribution, and some or all of the parameters θ that are subject to changes. This approach requires a precise knowledge of the signal under study for first determining the distribution and after choosing which are the parameters affected by the abrupt changes. In climate, the distribution is largely supposed Gaussian but, when it is not suitable, for example for precipitation data, another distribution are considered and methods should be redeveloped according to, as in Li & Lund [2012]; Wang *et al.* [2010]. Concerning the affected parameters, in the vast majority changes are assumed to affect the signal mean, but some authors also considered changes in trend as Lund & Reeves [2002]; Vincent [1998].

Approaches. Among the parametric methods, one can distinguish the frequentist and Bayesian approaches. The frequentist approach considers the parameter as fixed and proposes a point estimation by maximizing a well-suitable contrast function (e.g. the likelihood $p(\mathbf{Y}|\theta)$ and $\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \log p(\mathbf{Y}|\theta)$) whereas the Bayesian approach treats the parameter as a random variable drawn from a distribution, called the prior distribution, and aims at retrieving its posterior distribution, i.e. its distribution given

the data $p(\theta|\mathbf{Y})$. In climate, this approach has been considered by Hewaarachchi *et al.* [2017]; Li & Lund [2015]; Seidou & Ouarda [2007]; Seidou *et al.* [2007]. Compared to the frequentist approach, its advantage is that it makes it possible to take into account information on the change-point locations using the metadata through the prior distribution, as in Hewaarachchi *et al.* [2017]; Li & Lund [2015]. If Seidou *et al.* [2007] and Seidou & Ouarda [2007] adopt a pure Bayesian approach, Hewaarachchi *et al.* [2017] and Li & Lund [2015] considered an empirical Bayes-type approach that can be seen as a mixture between frequentist and Bayesian since some parameters are point estimated and some are integrated. As Table 2.1 shows, most of the proposed methods in climate are addressed in a frequentist statistical framework.

Inference procedure. The most commonly detection techniques proposed in climate in a parametric framework are based on two classical stastistical approaches: hypothesis tests and maximum likelihood inference. In the latter, the change-point detection problem is thus formulated as an estimation problem solved via the maximization of the (log-)likelihood criterion raising both algorithmical and model selection issues. In the former, which constitutes the huge part of this literature, it is formulated as a statistical hypothesis test for testing the presence of a single change-point leading to sequential detection procedures for the multiple change-point detection problem. Formally, given a time series $\boldsymbol{y} = \{y_t\}_{t=1,\dots,n}$, if a change-point exists at a time $t = \varphi$, then the behavior of series $\{y_t\}_{t=1,\dots,\varphi}$ differs from the behavior of series $\{y_t\}_{t=\varphi+1,\dots,n}$ in some sense. In a parametric framework, the observed data are modeled by a random process $\boldsymbol{Y} = \{Y_t\}_{t=1,\dots,n}$ with a probability distribution P_θ and this translate as a change in P_θ in terms of the parameter θ, i.e. the parameter θ of the two sub-series (the data on segments $[\![1, \varphi]\!]$ and $[\![\varphi+1, n]\!]$) is different. Naturally neither the existence or the location of the change-point φ are known in practice. Using the test-based approach, the purpose consists thus in first testing the presence of a change-point and then estimating its position. The first problem can be formulated in terms of the two following hypothesis:

$$\begin{cases} H_0 : & Y_1, \dots, Y_n \quad \text{i.i.d} \sim P_\theta \\ \textit{against} & \\ H_1 : \exists\, t \in (1, \dots, n) & \begin{cases} Y_1, \dots, Y_\varphi \quad \text{i.i.d} \sim P_{\theta_1} \\ Y_{\varphi+1}, \dots, Y_n \quad \text{i.i.d} \sim P_{\theta_2} \\ \theta_1 \neq \theta_2 \end{cases} \end{cases}$$

The standard tests are most used in climate are the t-test, F-test, and the likelihood ratio test (see methods lines from 7 to 16 in Table 2.1 and reference papers listed in column 'users'). Whatever the

choice of the statistical test, the general procedure is the following: each position is tested with an alternative hypothesis saying that a change occurs at this position. If the maximal value of the test statistic among the $n-1$ possible positions is larger than a critical value then the null hypothesis is rejected and the change-point is estimated as the position for which the test statistic attains a local maximum. In order to build the test decision, Monte Carlo simulations are usually used to approximate the distribution of the maximal test statistic since it is quite complex. Improvements of the t-test and the F-test have been proposed by Wang [2008a,b]; Wang *et al.* [2007, 2010] in view of an equally detection power of each position-test leading to penalized versions of these two tests (PMT and PMF). This procedure allows to detect possibly one single change-point. When multiple detection is wanted, sub-segmentation techniques can be used. This search consists in recursively applying the test until no significant change-point is found. Several algorithms have been proposed with application to climate data and are presented in the next paragraph.

The second approach, based on penalized likelihood, was considered by some authors (see lines from 1 to 6 in Table 2.1). Several methods were proposed which differ in the formulation of the penalty function and the search algorithm. In this thesis we also adopted the penalized likelihood approach as is detailed in Section 2.2.2. As we have seen in the previous paragraph, a Bayesian approach has been considered by some authors for single detection or multiple detection (see methods lines from 17 to 20 in Table 2.1). As in the parametric framework, the adopted criterion is based on the likelihood function.

Search. This paragraph presents an algorithmic point-of-view of the methods. More precisely, the question arising here concerns the fact that the multiple detection inference is performed in an exact manner or not. The exact version means that the solution is optimal according to the considered inference criterion and the non-exact to a sub-optimal one. The multiple change-point detection methods using recursive tests will leads necessarily to sub-optimal solutions since all the possible segmentation solutions are not considered. In this approach, among the proposed algorithms in climate, the most widely used search is the Binary Segmentation (BS) since it is conceptually simple and easy to implement: if a change-point is detected, the time series is partitioned in two series. The test is then applied on each of the two series separately until there is no more significant change-point (see Wang *et al.* [2007], Wang [2008a], Menne & Williams [2009], Stepanek *et al.* [2009]). A modification is proposed by Guijarro [2011] using moving windows. Another iterative procedure proposed by Lund & Reeves [2002], consists at each iteration in a detecting/correcting procedure: a change-point is detected on the whole series using a statistical test and the mean shift associated with the latter is removed. Although these algorithms lead to sub-optimal solutions, they have an advantage in the case of long series of being faster than exact

algorithms with a complexity linear in the length of the series. When all the change-points are detected simultaneously, whatever the frequentist or Bayesian approach, an algorithmical problem appears. This problem is well known and is due to the need of visiting the whole segmentation space which can be huge. In climate, in a frequentist framework, two well known algorithms are used: the dynamic programming (DP) by Bock *et al.* [2018]; Caussinus & Mestre [2004]; Domonkos *et al.* [2011] and a genetic-type algorithm (GA), that uses principles of genetic selection and mutation, by Li & Lund [2012]; Lu *et al.* [2010]. With the former the solution is optimal whereas with the latter it is sub-optimal. This point is discussed in the next section. In a Bayesian framework, there is no analytic expression of the posterior distribution of the change-points. A classical way of solving the problem is to carry out Monte Carlo simulations. In climate, Li & Lund [2015]; Li *et al.* [2019] used the classical Markov Chain Monte Carlo (MCMC) method. This approximation leads thus to a sub-optimal solution. The only algorithm that gives the optimal solution is described by Fearnhead [2006] which was used by Seidou & Ouarda [2007].

Specific features. In addition to the abrupt changes, climate series present different characteristics which, if not taken into consideration, can influence the segmentation process. In particular, effects that are not included in the segmentation model will be captured by the segmentation solution leading to many false detection. First working on the series of differences using a relative approach (see Section 2.1.2), the climate trend is, in theory, removed. However, in the case of an absolute approach, it is fundamental to integrate this trend in the segmentation model as proposed by Lu *et al.* [2010]; Lund & Reeves [2002]; Lund *et al.* [2007]; Vincent [1998]. Other characteristics can be taken into account, such as temporal dependence and periodic features which can affect the mean or the covariance. Dependencies such as periodic and autoregressive Gaussian process of order p have been considered (see lines 5, 6, 10, 11, 14, 19, and 20 of Table 2.1). Some authors also modelled the variance of the process (Bock *et al.* [2018]; Hewaarachchi *et al.* [2017]; Lu *et al.* [2010]).

2.2.2 The maximum penalized-likelihood approach used in this work. Two main issues.

The inference approach adopted in this thesis sets in a parametric and frequentist framework. A vast majority of methods setting on this approach use a two-step strategy: (i) estimate the change-point locations as well as the corresponding distribution parameters θ for a fixed number of change-points and (ii) choose the number of change-points. In most cases, the estimation of θ is not a problem because the exact solution can be derived. The major difficulties are: 1) to estimate the change-point positions, which leads to an algorithmic issue, and 2) to estimate the number of change-points, which is a model

selection problem.

Estimating the change-point positions. The classical maximum likelihood tools cannot be applied due to the discrete nature of the change-point parameters. This therefore implies the exploration of the whole segmentation space that is huge. Such an exploration is prohibitive in terms of computational time when performed in a naive manner (with a complexity in $\mathcal{O}(n^K)$ if K is the number of segments or $K-1$ is the number of change-points and n is the length of the series). If the search of the optimal segmentation according to the likelihood criterion can be based on a sequential principle as the binary segmentation (see for example Olshen *et al.* [2004] and Fryzlewicz *et al.* [2014]) leading to (acceptable) sub-optimal solutions, there exists an efficient algorithm that enables to recover the exact solution in a fast manner (with a complexity in $\mathcal{O}(Kn^2)$): the dynamic programming (DP) algorithm. This algorithm was introduced by Bellman [1954] and used for the first time in segmentation by Auger & Lawrence [1989] under the name "segment neighborhood". For the past ten years, many authors have proposed pruned versions of this algorithm that still remain exact: Killick *et al.* [2012]; Maidstone *et al.* [2017]; Rigaill [2015] with a linear complexity in n. DP is still widely used today, also in climate (Bock *et al.* [2018]; Caussinus & Mestre [2004]; Domonkos & Coll [2015]; Mestre *et al.* [2013]). All these exact algorithms can be used as long as the criterion to be optimized is additive on the segments. This condition is not always satisfied, typically when others effects that are not affected by the changes are added in the segmentation model. In this case, some authors suggest the use of a genetic algorithm (GA) providing an approximate solution to the segmentation (Li & Lund [2012]; Lu *et al.* [2010]).

Choosing the number of change-points. In most applications, this number is unknown and needs to be estimated. This model selection problem can be solved using penalized criteria: a penalty term is added to the inference criterion (here the log-likelihood) in order to take into account the complexity of the model. The problem is thus reduced to the choice of an appropriate penalty function. It is well known that the classical criteria such as Akaike information criterion (AIC) proposed by Akaike [1973] and Bayesian information criterion (BIC) proposed by Schwarz [1978] are not theoretically adapted in the segmentation context for different reasons (see Birgé & Massart [2007] and Zhang & Siegmund [2007], respectively) and tend to overestimate the number of segments in practice. Modified versions of the AIC and BIC dedicated to the segmentation framework have been proposed by Lebarbier [2005] based on the works of Birgé & Massart [2001] and Zhang & Siegmund [2007]. The latter is called modified Bayesian information criterion (mBIC). In the specific climate context, other penalties have been proposed. Caussinus & Mestre [2004] proposed a BIC-like penalty considering both change-points and

outliers. Some authors proposed penalties based on the MDL (Minimum description length) criterion proposed by Rissanen [1978] (see for example Li & Lund [2012]; Lu *et al.* [2010]).

2.3 Outline of this work

The main goal of this thesis was to develop a new segmentation model adapted to the special features of the GNSS IWV difference time series highlighted in Section 2.1.3 and Figure 2.6. As explained in Section 2.2.2 we decided to consider a penalized-likelihood approach using DP which allows to find the exact solution to the associated segmentation problem. Knowing the position of the change-points will allow to correct the original GNSS IWV time series for inhomogeneities (abrupt changes in the mean) or alternatively implement a trend estimation method that takes the inhomogneneities into account. The structure of the following chapters describes the development and the application of this method to the global GNSS data set introduced in Section 2.1.3 and Figure 2.5.

Chapter 3 provides a general segmentation framework using a penalized likelihood inference method. Two classic and widely used models are presented, with particular emphasis on the algorithmic part: the homoscedastic model which considers changes in the mean of an independent Gaussian process and the heteroscedastic model which considers changes both in the mean and in the variance. A third model is also presented which was proposed by Bock *et al.* [2018] to account for the changing monthly variance observed in the GNSS IWV differences (see Figure 2.5). However, it is shown that these models fail when the data contain an additional signal such as a periodic bias highlighted in the GNSS IWV differenes (Figure 2.5).

Chapter 4 introduces a new model developed in this thesis. This model is an extension of the previous model (Bock *et al.* [2018]) and accounts for the periodic bias. This chapter describes the model, inference, and algorithmic aspects which allowed to use DP. The method is tuned and its performance is assessed by extensive numerical simulations. The method has been released as an R package available to the scientific community.

Chapter 5 shows the results of our method on real data. It also presents our approach to deal with the outliers. A trend estimation for the data is proposed.

Finally, Chapter 6 discusses and concludes on the findings of the research presented in this thesis, as well as its limitations, lessons learned during the conducting process, and some guidelines for future work.

Chapter 3

Segmentation methods

A general segmentation method is presented in this chapter. The considered approach is parametric, frequentist and is based on a penalized likelihood inference. Section 3.1 introduces the general segmentation model. Sections 3.2 and 3.3 present the classical maximum likelihood method used to estimate the parameters and the DP algorithm used to solve the optimization problem. Section 3.4 introduces several penalties that will be used for the choice of the number of changes or segments. In Sections 3.5 and 3.6 the method is applied on three specific models. The two firsts are classic and very widely used models called the homoscedastic and the heteroscedastic gaussian models. They consider changes in the mean and changes both in the mean and in the variance respectively. Then, we present a model that is "between the two latter": the changes affected only the mean and the variance is heterogeneous but on fixed and known intervals. This model has been proposed by Bock *et al.* [2018] for a first analysis of the GNSS data. With one simulated series, we illustrate that these models fail when the signal includes a periodic bias as is evidenced on the real GNSS IWV differenced data (see Figure 2.6 (b)).
The model proposed by Bock *et al.* [2018] will be the basis of the new model developed in this work and described in Chapter 4.

3.1 General Model

We observe $\boldsymbol{y} = \{y_t\}_{t=1,\dots,n}$ a finite sequence of observed data which are realizations of n independent random variables Y_t that are supposed to be drawn from a probability distribution $(P_\theta)_{\theta\in\Theta}$ such that

the parameter θ is affected by $K-1$ changes. The model is thus:

$$Y_t \sim \begin{cases} P_{\theta_1}, & t = 1, \dots, t_1 \\ P_{\theta_2}, & t = t_1 + 1, \dots, t_2 \\ \dots, & \\ P_{\theta_k}, & t = t_{k-1} + 1, \dots, t_k \\ \dots, & \\ P_{\theta_K} & t = t_{K-1} + 1, \dots, n, \end{cases}$$

where $\theta_k \neq \theta_{k+1}, \forall k \in \{1, \dots, K-1\}$. Modeling the data is thus the first issue. The difficulty is to choose both the distribution P_θ and which parameter is affected by the changes (all can be affected or some of them only).

The parameters of the model to be estimated are the following ones:

1. K the number of the segments,
2. $\boldsymbol{T} = (t_1, \dots, t_{K-1})$ the $K-1$ change-points, that decompose the signal in K segments, $I_k = [\![t_{k-1} + 1, t_k]\!]$, with the convention $t_0 = 0$ and $t_K = n$.
3. $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)$ the parameters of the distribution.

Once the model has been specified, the issue regards the inference of all the parameters given below.

3.2 Inference

The classical maximum likelihood method is used to estimate the parameters. If we suppose that the probability distribution P_θ admits a density f_θ and since the Y_t are independent, the likelihood is written as follows

$$p(\boldsymbol{y}; K, \boldsymbol{T}, \boldsymbol{\theta}) = \prod_{t=1}^{n} p(y_t; \boldsymbol{\theta}, \boldsymbol{T}) = \prod_{k=1}^{K} \prod_{t=t_{k-1}+1}^{t_k} f_{\theta_k}(y_t)$$

and the log-likelihood as

$$\log p(\boldsymbol{y}; K, \boldsymbol{T}, \boldsymbol{\theta}) = \sum_{k=1}^{K} \sum_{t=t_{k-1}+1}^{t_k} \log f_{\theta_k}(y_t) \tag{3.1}$$

The inference is conventionally done in three steps:

(i) Estimating the distribution parameters $\boldsymbol{\theta}$, $\boldsymbol{T}$ and K being fixed. We get:

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \Theta^K}{\operatorname{argmax}} \ \log p(\boldsymbol{y}; K, \boldsymbol{T}, \boldsymbol{\theta}) = \underset{\boldsymbol{\theta} \in \Theta^K}{\operatorname{argmax}} \sum_{k=1}^{K} \sum_{t=t_{k-1}+1}^{t_k} \log f_{\theta_k}(y_t).$$

Thus

$$\widehat{\theta}_k = \underset{\theta_k \in \Theta}{\operatorname{argmax}} \sum_{t=t_{k-1}+1}^{t_k} \log f_{\theta_k}(y_t).$$

(ii) Finding the change-point locations $\mathbf{T}$ for a fixed number of segments K (or $K-1$ change-points). To this aim, we have now to maximize the log-likelihood calculated at its maximum:

$$\widehat{\boldsymbol{T}} = \underset{\boldsymbol{T} \in \mathcal{M}_{K,n}}{\operatorname{argmax}} \log p(\boldsymbol{y}; K, \mathbf{T}, \widehat{\boldsymbol{\theta}}) = \underset{\boldsymbol{T} \in \mathcal{M}_{K,n}}{\operatorname{argmax}} \sum_{k=1}^{K} \sum_{t=t_{k-1}+1}^{t_k} \log f_{\widehat{\theta}_k}(y_t),$$

where $\mathcal{M}_{K,n} = \{(t_1, \ldots, t_{K-1}) \in \mathbb{N}^{K-1}, 0 = t_0 < t_1 < \cdots < t_{K-1} < t_K = n\}$ is the set of all possible segmentations of the grid $[\![1, n]\!]$ in K segments. While the distribution parameters are continuous, the change-point parameters are discrete. This point constitutes the major difficulty in segmentation. Indeed, in this case the likelihood is not differentiable with respect to these parameters and the classical maximum likelihood tools can not be used. We have thus to explore the whole space $\mathcal{M}_{K,n}$. The size of this space is huge since it is $\binom{n-1}{K-1}$ causing an algorithmic problem. Indeed, from a computational point of view, an exhaustive exploration would have a complexity in $\mathcal{O}(n^K)$. The only way (until 2011) to reduce this complexity and obtained the exact solution is to use the dynamic programming (DP) introduced by Bellman [1954]. Its complexity is linear in K and quadratic in the length of the time series ($\mathcal{O}(Kn^2)$). More details on this algorithm are given in Section 3.3.

At this point, we have a collection of the best segmentation of the data in $k = 1, \ldots, K$ segments and the purpose is to choose the "best" one.

(iii) Choosing the number of segments K. The fit to the data, given by the likelihood, will always increase with the number of segments. In order to choose an appropriate segmentation, a common strategy consists in adding a penalty term depending of the number of segments:

$$\widehat{K} = \underset{K}{\operatorname{argmax}} \log p(\boldsymbol{y}; K, \widehat{\boldsymbol{T}}, \widehat{\boldsymbol{\theta}}) - pen(K, n).$$

Thus the problem is reduce to the choice of a "good" penalty. A more detailed discussion and presentation of different penalties adapted in segmentation framework will be given in the Section 3.4.

3.3 Dynamic Programming

The dynamic programming algorithm is a recursive algorithm, based on the Bellman optimality principle "Sub-paths of the optimal path are themselves optimal" (Bellman [1954], and it was introduced for the first time in the context of segmentation by Fisher [1958].

Recall that the optimisation problem for estimating the change-points is the following

$$\min_{\boldsymbol{T}\in\mathcal{M}_{K,n}} \sum_{k=1}^{K} \sum_{t=t_{k-1}+1}^{t_k} -\log f_{\hat{\theta}_k}(y_t). \tag{3.2}$$

This optimization problem is actually a shortest path problem, that can be solved thanks to DP in an exact manner. Let define

- $\mathcal{C}(i,j) = \sum_{t=i}^{j} -\log f_{\hat{\theta}_{ij}}(y_t)$ is the cost of the path connecting i to j directly, e.g. the cost of the segment $I_{i->j}$ with $\hat{\theta}_{ij}$ is the estimation of θ using the data $\{y_i, \ldots, y_j\}$,
- $\mathcal{C}_k(i,j)$ the cost of the best path connecting i to j in k sub-paths, e.g. the best segmentation from i to j in k segments.

Using these notations, the optimization problem of interest (3.2) is written

$$\mathcal{C}_K(1,n) = \min_{1\leq t_1<,\ldots,<t_{K-1}} \sum_{k=1}^{K} \mathcal{C}(t_{k-1}+1, t_k).$$

This cost can be computed using a recursive formulation.

Recursive formulation. Considering first the case of $K = 2$ segments, i.e. one change-point, we have

$$\begin{aligned}
\mathcal{C}_2(1,n) &= \min_{1\leq t_1<n} \sum_{k=1}^{2} \mathcal{C}(t_{k-1}+1, t_k) = \min_{1\leq t_1<n} \Big\{ \mathcal{C}(1,t_1) + \mathcal{C}(t_1+1,n) \Big\} \\
&= \min_{1\leq h<n} \Big\{ \mathcal{C}_1(1,h) + \mathcal{C}(h+1,n) \Big\}
\end{aligned}$$

For any K,

$$\begin{aligned}
\mathcal{C}_K(1,n) &= \min_{1\leq t_1<\cdots<t_{K-1}<n} \sum_{k=1}^{K} \mathcal{C}(t_{k-1}+1,t_k) \\
&= \min_{1\leq t_1<\cdots<t_{K-1}<n} \left\{ \sum_{k=1}^{K-1} \mathcal{C}(t_{k-1}+1,t_k) + \mathcal{C}(t_{K-1}+1,n) \right\} \\
&= \min_{K-1\leq t_{K-1}<n} \left\{ \min_{1\leq t_1<\cdots<t_{K-2}<t_{K-1}} \sum_{k=1}^{K-1} \mathcal{C}(t_{k-1}+1,t_k) + \mathcal{C}(t_{K-1}+1,n) \right\} \\
&= \min_{K-1\leq h<n} \left\{ \mathcal{C}_{K-1}(1,h) + \mathcal{C}(h+1,n) \right\}
\end{aligned}$$

Description of DP. This algorithm requires the calculation of the cost of all segments for $1 \leq i \leq j \leq n$ stored in an upper diagonal matrix $\mathcal{D}$ $n \times n$, called cost matrix.

$$mat\mathcal{D} = \begin{cases} \mathcal{C}(i,j) & \text{if } i \leq j \\ +\infty & \text{otherwise} \end{cases}$$

Step k: for $2 \leq k \leq K$, $k \leq j \leq n$,

do

$$\mathcal{C}_k(1,j) = \min_{k-1\leq h<j} \{\mathcal{C}_{k-1}(1,h) + \mathcal{C}_1(h+1,j)\}$$

DP allows to calculate the cost of the best segmentation in K segments, but does not have as output the optimal segmentation itself. The algorithm is completed by a backtracking procedure, which allows to rebuild the associated optimal change-points. The DP algorithm has a complexity of the order of $O(Kn^2)$ if the calculation of the cost matrix is of the same order.

Sufficient condition for using DP. DP can be applied if and only if the quantity to be optimized is segment-additive, i.e. $-\log p(\boldsymbol{y}; K, \boldsymbol{T}, \widehat{\boldsymbol{\theta}}) = \sum_{k=1}^{K} C_k$ where C_k denotes a function of k (see for example Bai & Perron [2003]; Lavielle [2005]; Picard *et al.* [2005]). A sufficient condition to have this segment-additivity is the independence between the segments in terms of observations (the Y_t of different segments are independent) and in terms of parameters (the parameters are all segment-specific). Thus the presence of global (or common) parameters among the segments hampers the use of DP. Note that if there exists some dependence between the Y_t, DP can be applied only if the dependence parameter is (also) affected by the abrupt changes. For more details, see Section 2.6.1 in Chakar *et al.* [2017]. Some climate models, as as seen in 2.2.1, take into account some additional effects, such as temporal dependence and periodic features (see for example Hewaarachchi *et al.* [2017]; Lu *et al.* [2010] and Bock *et al.* [2018]). In this case, DP cannot be applied directly and alternatives have been proposed leading to approximate solutions: a

two-step inference procedure using again DP as in Bock *et al.* [2018] (see Section 3.6) or the use of an other algorithm, the GA algorithm as in Lu *et al.* [2010].

Recently, faster (linear or quasi-linear in n in n in many cases) but still exact versions of DP have been proposed: PELT (Killick *et al.* [2012]) and PDPA (Rigaill [2015]) and the mixture of them (Maidstone *et al.* [2017]). For PDPA, the necessary condition is there that is one single parameter (thus affected by abrupt changes). The PELT algorithm integrates a penalty proportional to K (so it simultaneously estimates the change-point locations and their number). However, it is known that such AIC-type penalty selects too many segments (Cleynen *et al.* [2014]).

3.4 Model Selection

The procedure described in Section 3.2 provides the best segmentation, according to the likelihood, in a fixed number of segments K, which is unknown in practice. The choice of K can be seen as a model selection problem. To this purpose, the most common strategy is to use penalized contrast criteria. Most often, the contrast denoted c here is the least-squares criterion or minus the log-likelihood, and the number of segments K or change-points $K-1$ is chosen by minimizing $c(\boldsymbol{y}; K, \widehat{\boldsymbol{T}}, \widehat{\boldsymbol{\theta}}) + pen(K, n)$ for some suitable penalty function $pen(K, n)$. As already mentioned in Section 2.2.2, classical penalties such as AIC (Akaike [1973]) and BIC (Schwarz [1978]) are not adapted to the segmentation framework (Birgé & Massart [2007] and Zhang & Siegmund [2007]) and tend to overestimate the number of segments. A huge literature has been devoted to propose penalties dedicated to the segmentation framework (see Lavielle [2005]; Lebarbier [2005]; Yao & Au [1989]; Zhang & Siegmund [2007]) and in the specific climate context (see Caussinus & Mestre [2004]; Li & Lund [2012]; Lu *et al.* [2010]). In this thesis, we will consider three of them:

⋆ **the one proposed by Lavielle [2005]; Yao & Au [1989], denoted Lav** in which the penalty is proportional to the number of segments:

$$\widehat{K} = \underset{K}{\operatorname{argmin}}\ c(\boldsymbol{y}; K, \widehat{\boldsymbol{T}}, \widehat{\boldsymbol{\theta}}) + \beta K, \tag{3.3}$$

where β is the penalty constant chosen using an adaptive method proposed by the author.

⋆ **the one proposed by Birgé & Massart [2001], denoted BM** and calibrated by Lebarbier [2005] for segmentation in the mean of Gaussian process:

$$\widehat{K} = \underset{K}{\operatorname{argmin}}\ c(\boldsymbol{y}; K, \widehat{\boldsymbol{T}}, \widehat{\boldsymbol{\theta}}) + \alpha K \left[5 + 2\log\left(\frac{n}{K}\right)\right], \tag{3.4}$$

where the penalty constant α can be calibrated using the slope heuristic proposed by Arlot & Massart [2009].

⋆ **the one proposed by Zhang & Siegmund [2007]** called the modified BIC (mBIC) which is a modified version of the classical BIC criterion dedicated to the segmentation in the mean of a Gaussian process $y_t \sim \mathcal{N}(\mu_k, \sigma^2)$: two versions exist, one with known variance and one with unknown variance. The former is :

$$\widehat{K} = \underset{K}{\operatorname{argmin}} \frac{1}{2}\sum_{k=1}^{K}\sum_{t=t_{k-1}+1}^{t_k}(y_t - \widehat{\mu}_k)^2 + \frac{1}{2}\sum_{k=1}^{K}\log(\widehat{t}_k - \widehat{t}_{k-1}) - \left(\frac{3}{2} - K\right)\log(n). \tag{3.5}$$

Ardia *et al.* [2019] show that the MDL criterion, used by Li & Lund [2012] and Lu *et al.* [2010], can be seen as a Bayesian criterion with appropriate prior distributions for change-point models. As a consequence, the obtained based-MDL penalties looks like the mBIC. In particular, both penalties integrate a term depending on the segment lengths of the segmentation.

The choice of K is a complicated and delicate problem. Moreover, because their formulations are different, each criterion may select a different model.

3.5 Classical Gaussian segmentation models

In this section we will restrict the study to the cases of an independent Gaussian process

$$\forall t \in \{1, \dots, n\},\ Y_t \sim \mathcal{N}(\mu(t), \sigma(t)^2)$$

in order to illustrate the inference procedure. Two models can be considered: the homoscedastic model called (M_1) and heteroscedastic model called (M_2). In (M_1), the mean is the only parameter affected by the abrupt changes, and the variance is considered constant during the time. The distribution parameters are $\boldsymbol{\theta} = (\boldsymbol{\mu}, \sigma^2)$, where $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_K)$, and the model is written as follows

$$(M_1) \qquad Y_t = \mu_k + \mathcal{E}_t,\ \mathcal{E}_t\ i.i.d. \sim \mathcal{N}(0, \sigma^2),$$

for $t \in I_k = [\![t_{k-1} + 1, t_k]\!]$ with $k \in [\![1, K]\!]$.

In the heteroscedastic model (M_2), the abrupt changes affect simultaneously the mean and the variance. The distribution parameters are then $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$, where $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_K)$ and $\boldsymbol{\sigma}^2 = (\sigma_1^2, \sigma_2^2, \dots, \sigma_K^2)$. And the model is written as follows

$$(M_2) \qquad Y_t = \mu_k + \mathcal{E}_t,\ \mathcal{E}_t\ ind. \sim \mathcal{N}(0, \sigma_k^2),$$

for $t \in I_k = [\![t_{k-1}+1, t_k]\!]$ with $k \in [\![1, K]\!]$.

The inference procedure is described in the Section 3.2. We take up the structure of the aforementioned paragraph, introducing the log-likelihood and focus on the two first steps of the inference (the estimation of the distribution parameters and the change-point locations when K is fixed).

Log-likelihood. The log-likelihood for model (M_1) is

$$\begin{aligned}\log p(\boldsymbol{y}; K, \boldsymbol{T}, \boldsymbol{\mu}, \sigma^2) &= \sum_{k=1}^{K} \sum_{t=t_{k-1}+1}^{t_k} \log f_{(\mu_k, \sigma^2)}(y_t), \\ &= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{k=1}^{K}\sum_{t=t_{k-1}+1}^{t_k}(y_t - \mu_k)^2\end{aligned}$$

and for model (M_2) is

$$\begin{aligned}\log p(\boldsymbol{y}; K, \boldsymbol{T}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) &= \sum_{k=1}^{K} \sum_{t=t_{k-1}+1}^{t_k} \log f_{(\mu_k, \sigma_k^2)}(y_t) \\ &= \sum_{k=1}^{K} -\frac{(t_k - t_{k-1})}{2}\log(2\pi\sigma_k^2) - \sum_{k=1}^{K}\left(\frac{1}{2\sigma_k^2}\sum_{t=t_{k-1}+1}^{t_k}(y_t - \mu_k)^2\right).\end{aligned}$$

j

Estimation of $\boldsymbol{\theta}$, T being fixed (step (i)). The estimators of the mean and the variance are the classical maximum likelihood estimators:

$$\widehat{\mu}_k = \frac{1}{(t_k - t_{k-1})}\sum_{t=t_{k-1}+1}^{t_k} Y_t \qquad \text{for } (M_1) \text{ and } (M_2),$$

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{K}\sum_{t=t_{k-1}+1}^{t_k}(Y_t - \widehat{\mu}_k)^2 \qquad \text{for } (M_1),$$

$$\widehat{\sigma}_k^2 = \frac{1}{(t_k - t_{k-1})}\sum_{t=t_{k-1}+1}^{t_k}(Y_t - \widehat{\mu}_k)^2 \qquad \text{for } (M_2)$$

Finding the change-point locations T (step (ii)). Recall that $\widehat{\mathbf{T}}$ is obtained by maximizing the log-likelihood calculated at its maximum for $\boldsymbol{\theta}$:

$$\widehat{\boldsymbol{T}} = \underset{\boldsymbol{T} \in \mathcal{M}_{K,n}}{\operatorname{argmax}} \ \log p(\boldsymbol{y}; K, \boldsymbol{T}, \widehat{\boldsymbol{\theta}}).$$

For model (M_1), we get

$$\begin{aligned}
\widehat{\mathbf{T}} &= \underset{\mathbf{T}\in\mathcal{M}_{K,n}}{\operatorname{argmax}} \log p(\mathbf{y}; K, \mathbf{T}, \widehat{\boldsymbol{\mu}}, \widehat{\sigma}^2) \\
&= \underset{\mathbf{T}}{\operatorname{argmax}} \left(-\frac{n}{2}\big(\log(2\pi) + \log(\widehat{\sigma}^2)\big) - \frac{1}{2\widehat{\sigma}^2} n\widehat{\sigma}^2 \right) \\
&= \underset{\mathbf{T}}{\operatorname{argmax}} \; -\frac{n}{2}\left(\log(2\pi) + 1 + \log\left(\frac{1}{n} \sum_{k=1}^{K} \sum_{t=t_{k-1}+1}^{t_k} (y_t - \widehat{\mu}_k)^2 \right) \right) \\
&= \underset{\mathbf{T}}{\operatorname{argmin}} \sum_{k=1}^{K} \sum_{t=t_{k-1}+1}^{t_k} (y_t - \widehat{\mu}_k)^2
\end{aligned}$$

Thus DP applied by considering as the cost of the segment $[\![i, j]\!]$,

$$\mathcal{C}(i, j) = \sum_{t=i}^{j} (y_t - \widehat{\mu}_{i,j})^2$$

where $\widehat{\mu}_{ij} = \overline{y}_{ij} = \frac{1}{n_{ij}} \sum_{t=i}^{j} y_{ij}$ and $n_{ij} = j - i + 1$. The model (M_1) has a common parameter: the variance σ^2. We have seen that a sufficient condition for using DP is that there are no common parameters. It turns out that the maximization in distribution parameters $\boldsymbol{\theta}$ comes down to a problem of optimizing a segment-additive quantity. This model is the only exception, to our knowledge, with a common parameter where it still works.

For model (M_2), we get

$$\begin{aligned}
\widehat{\mathbf{T}} &= \underset{\boldsymbol{T}\in\mathcal{M}_{K,n}}{\operatorname{argmax}} \log p(\boldsymbol{y}; K, \boldsymbol{T}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\sigma}}^2) \\
&= \underset{\mathbf{T}}{\operatorname{argmax}} \left(\sum_{k=1}^{K} -\frac{(t_k - t_{k-1})}{2} (\log(2\pi) + \log(\widehat{\sigma}_k^2)) - \sum_{k=1}^{K} \frac{1}{2\widehat{\sigma}_k^2} (t_k - t_{k-1}) \widehat{\sigma}_k^2 \right) \\
&= \underset{\mathbf{T}}{\operatorname{argmax}} \left(\sum_{k=1}^{K} -\frac{(t_k - t_{k-1})}{2} \left(\log(2\pi) + \log\left(\frac{1}{t_k - t_{k-1}} \sum_{t=t_{k-1}+1}^{t_k} (y_t - \widehat{\mu}_k)^2 \right) \right) - \frac{n}{2} \right) \\
&= \underset{\mathbf{T}}{\operatorname{argmin}} \sum_{k=1}^{K} (t_k - t_{k-1}) \log\left(\frac{1}{t_k - t_{k-1}} \sum_{t=t_{k-1}+1}^{t_k} (y_t - \widehat{\mu}_k)^2 \right)
\end{aligned}$$

Thus DP applied by considering as the cost of the segment $[\![i, j]\!]$,

$$\mathcal{C}(i, j) = \sum_{t=i}^{j} n_{ij} \log\left(\sum_{t=i}^{j} \frac{(y_t - \widehat{\mu}_{ij})^2}{n_{ij}} \right).$$

The model (M_2) satisfies the necessary and sufficient conditions to use DP: y_t are independent and all parameters are segment-specific.

3.6 Segmentation in the mean with heterogeneous variance on fixed time-intervals

Recall that in this thesis, we are motivated by the homogenization of the GNSS ΔIWV series presented in Section 2.1.4. As we have seen, these series present some particular characteristics in addition to the abrupt changes: a monthly change in variance (with a period of one year) and a smoothly varying bias as illustrated in Figure 2.6 (b). First let us consider the latter characteristic (and thus forgetting the smoothly varying bias). The two classical models presented in the above section will not be adapted. In order to illustrate this point, we applied them on a simulated time series. We consider a series of length of $n = 400$ with 4 "years" of 2 "months" of 50 "days" each and with standard deviations changing every month. A total of 6 change-points are considered at positions $t = 55, 77, 177, 222, 300, 366$ and the mean within each segment alternates between 0 and 1. The standard deviations of the two months are $\sigma_1 = 0.2$ and $\sigma_2 = 1.2$. The simulated series is represented in Figure 3.1. Note that the change-points located at the positions $t = 55, 77, 177, 366$ are more difficult to detect because they are located in segments with a high standard deviation (σ_2), the change-point located at $t = 300$ corresponds to both a change in the mean and in the variance and the one at $t = 222$ belongs to a segment with a small standard deviation (σ_1) thus easier to detect. The segmentation solutions obtained with the homoscedastic model (M_1) and the heteroscedastic model (M_2) using the BM model selection criterion are given in Figure 3.2 (a) and (b), respectively. The heteroscedastic model finds 8 change-points: only one change in the mean is detected and all the changes in the variance are detected, as requested by this model and so expected. The homoscedastic model retrieves 4 change-points among the 6, the two missing change-points being positioned in a segment with a high standard deviation.

Recently, a segmentation model has been proposed by Bock *et al.* [2018] including a variance changing on given and fixed time intervals. An evenly-spaced time interval of one month was chosen with a period of one year, as this corresponds with the dominant mode of variability seen in the data series. The variance can also change somehow from year to year, but this variation is neglected. The model writes as:

$$Y_t = \mu_k + \mathcal{E}_t, \ \mathcal{E}_t \ iid \sim \mathcal{N}(0, \sigma^2_{month}), \quad \forall t \in I_k \cap I_{\text{month}}$$

for $t \in I_k = [\![t_{k-1} + 1, t_k]\!]$ with $k \in [\![1, K]\!]$, and if date(t) is the date at the position t, $I_{\text{month}} =$

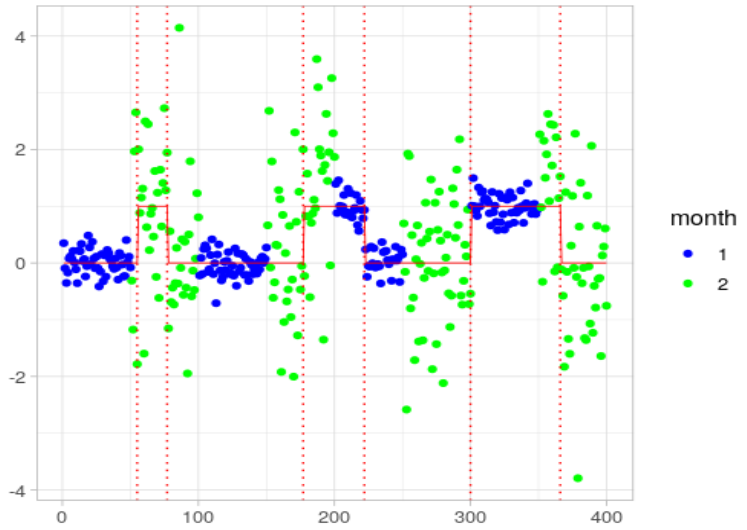


Figure 3.1 – A simulated time series of length $n = 400$ with 6 change-points (vertical dotted red lines) with standard deviation $\sigma_1 = 0.2$ in blue and $\sigma_2 = 1.2$ in green. The red line corresponds to the mean of the signal.

$\{t; \text{date}(t) \in \text{month}\}$. The log-likelihood is

$$\log p(\boldsymbol{y}; K, \boldsymbol{T}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) = \sum_{k=1}^{K} \sum_{month} \sum_{t \in I_k \cap I_{month}} -\frac{1}{2} \log(2\pi\sigma^2_{month}) - \sum_{k=1}^{K} \sum_{month} \sum_{t \in I_k \cap I_{month}} \frac{(y_t - \mu_k)^2}{2\sigma^2_{month}}.$$

where $\boldsymbol{\sigma}^2 = (\sigma^2_{month})_{month}$. The maximum likelihood estimators of μ_k and σ_{month} are respectively

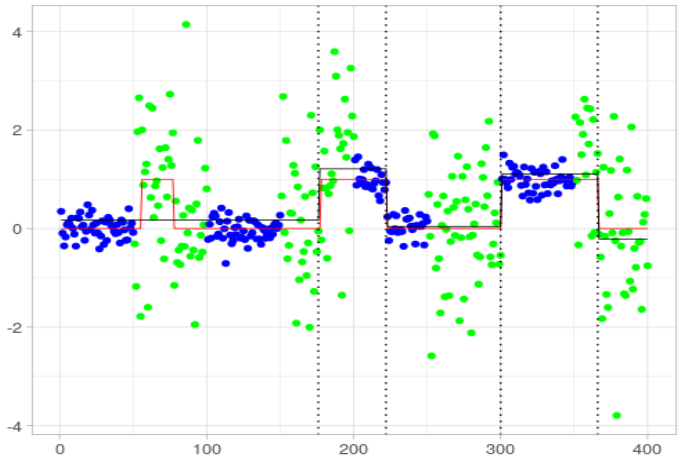
$$\widehat{\mu}_k = \frac{\sum_{month} \sum_{t \in I_k \cap I_{month}} \frac{Y_t}{\widehat{\sigma}^2_{month}}}{\sum_{month} \sum_{t \in I_k \cap I_{month}} \frac{1}{\widehat{\sigma}^2_{month}}}, \tag{3.6}$$

and

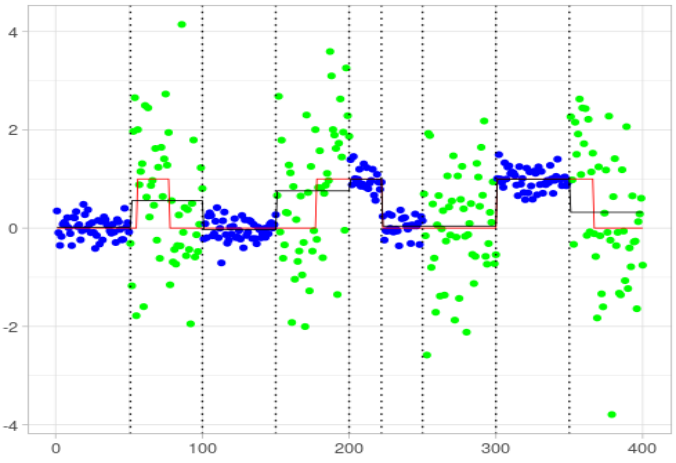
$$\widehat{\sigma}^2_{month} = \frac{1}{n_{month}} \sum_{t \in I_k \cap I_{month}} (Y_t - \widehat{\mu}_k)^2.$$

As we can see, the estimators are inter-dependent. The inference step (i) therefore already poses difficulties. Moreover, another problem arises: σ_{month} links some segments together, so DP cannot be used to estimate the change-points as the segment-additivity condition is not met (see Section 3.3). In order to keep possible the use of DP, the solution proposed by Bock *et al.* [2018], following the strategy proposed by Chakar *et al.* [2017], is to first estimate the variances and then use the classical inference with 'known' variances:

- Estimating the variances σ^2_{month}. The main problem is to estimate the variance in the presence of



(a) Homoscedastic model



(b) Heteroscedastic model

Figure 3.2 – Obtained segmentation with the homoscedastic (a) and the heteroscedastic (b) models on the simulated time series plotted in Figure 3.1. The vertical dotted black lines correspond to the estimated change-points, the black lines to the estimated mean and the red line to the true mean.

change-points. Since the classical estimator would fail, Bock *et al.* [2018] proposed to use a robust estimator proposed by Rousseeuw & Croux [1993] and apply it to the differenced time series, $Y_t - Y_{t-1}$. This series is centered except at the change-point positions (i.e. only $K-1$ ($K \ll n$) differences are non-centered) which can be seen as outliers. The scale estimator of Rousseeuw & Croux [1993] is robust with respect to a small number of outliers. The estimated monthly standard deviation write finally:

$$\widetilde{\sigma}_{month} = \frac{Q_{CR,n}((Y_{t+1} - Y_t)_t)}{\sqrt{2}},\ t \in month \tag{3.7}$$

where for a process $\mathbf{X}$

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

with

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

where Φ denotes the cumulative distribution function of a standard Gaussian random variable. Note that the $Q_{CR,n}(\mathbf{X})$ estimator is proportional to the first quartile of the absolute differences. The $\widetilde{\sigma}_{month}$ is computed over all years for the month considered.

- Classical inference with 'known' variances. The estimators of μ_k are given by equation (3.6) in which $\widehat{\sigma}^2_{month}$ is replaced by $\widetilde{\sigma}^2_{month}$ resulting in a classical weighted least-squares estimator with weights $1/\widetilde{\sigma}^2_{month}$. Then to estimate the change-points, the optimization problem is

$$\begin{aligned}\widehat{\mathbf{T}} &= \underset{\mathbf{T}\in\mathcal{M}_{K,n}}{\operatorname{argmax}}\ \log p(\mathbf{y}; K, \mathbf{T}, \widehat{\boldsymbol{\mu}}, \widetilde{\sigma}^2) \\ &= \underset{\mathbf{T}}{\operatorname{argmin}} \sum_{k=1}^{K} \sum_{month} \sum_{t\in I_k \cap I_{month}} \frac{(y_t - \widehat{\mu}_k)^2}{\widetilde{\sigma}^2_{month}}.\end{aligned}$$

Thus DP applied by considering as the cost of the segment $[\![i, j]\!]$,

$$\mathcal{C}(i,j) = \sum_{t=i}^{j} \frac{(y_t - \widehat{\mu}_{ij})^2}{\widetilde{\sigma}^2_t}.$$

We applied this method on the same simulated time series as in the previous paragraph plotted in Figure 3.1. The obtained result is given in Figure 3.4: we can observe that all the true change-points are retrieved. More simulation results, including comparisons to the homoscedastic and heteroscedastic models, as well as application to real data with this model are presented in Bock *et al.* [2018].

Now let's get back to the smoothly varying bias seen in the GNSS ΔIWV data (Figure 2.6 (b)). As

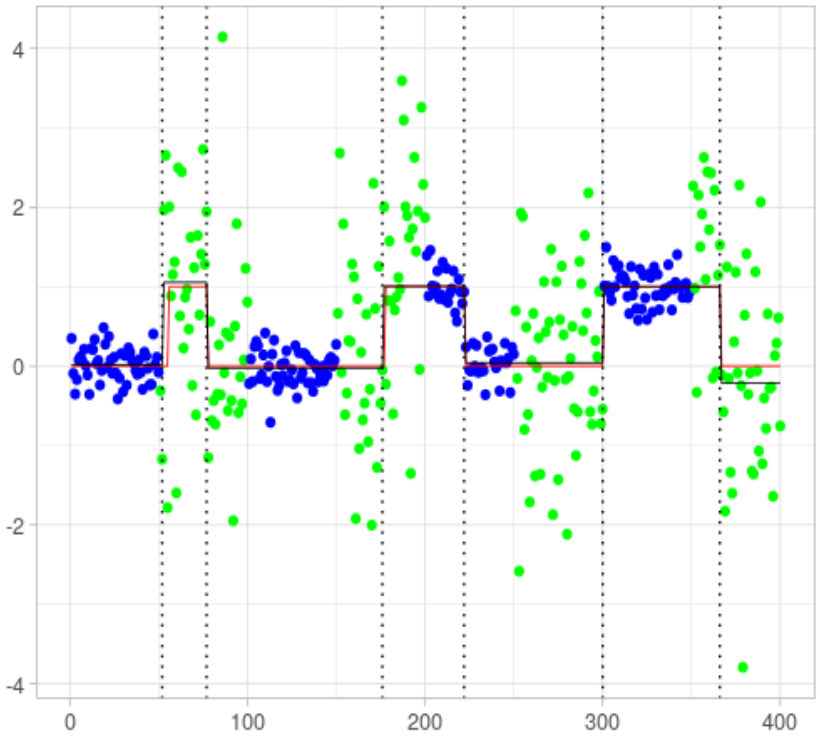


Figure 3.3 – Segmentation obtained with the model proposed by Bock *et al.* [2018] on the simulated time series plotted in Figure 3.1. The vertical dotted black lines correspond to the estimated change-points, the black lines to the estimated mean and the red line to the true mean.

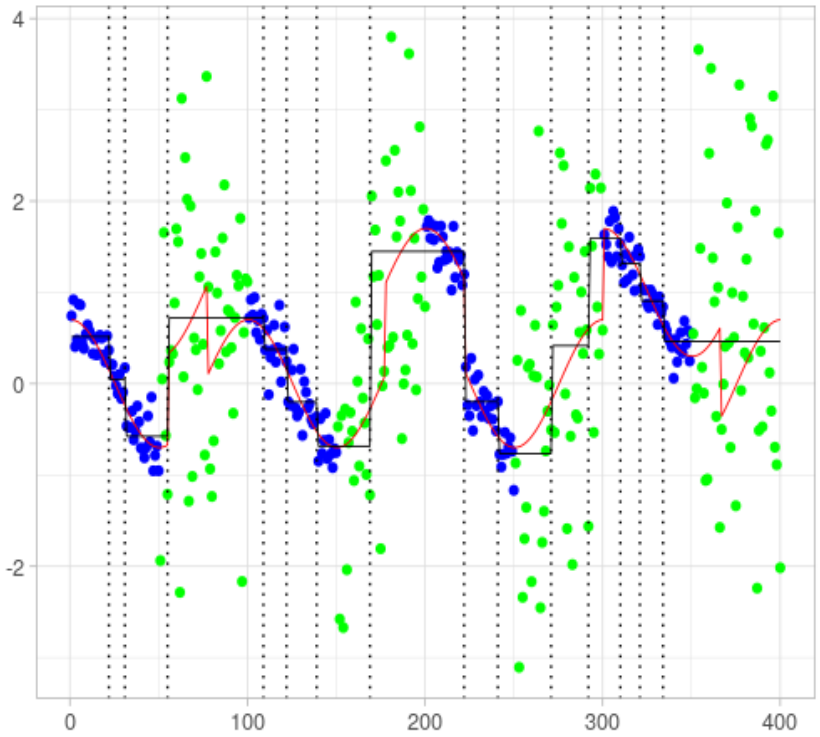


Figure 3.4 – Segmentation obtained with the model proposed by Bock *et al.* [2018] on the simulated time series in which a periodic function has been added. The vertical black lines correspond to the estimated change-points, the black line corresponds to the estimated mean, the global mean (the mean and the function) is in red.

explained in Section 2.1.4 this bias is due to representativenenes differences between the GNSS point observation and the reanalysis. In order to evaluate the impact of this feature, we added a periodic function: $f_t = 0.7\cos(2\pi t/L)$, where $L = 100$ is the length of a year, to the previously simulated series. The resulting series is plotted in Figure 3.4 with its true average (the mean plus the function) in red. The segmentation obtained with the model proposed by Bock *et al.* [2018] is also given in this figure. We can observe that the segmentation captures the functional with an overestimation of the number of change-points. This phenomenon is also observed on real GNSS ΔIWV series as will be illustrated in Chapter 5. To solve this issue, a new model is developed Chapter 4 to take into account the possible presence of a smoothly varying bias which is modelled as an additive functional part.

Chapter 4

A new segmentation method adapted to GNSS IWV difference data

In this chapter, we present the new segmentation model we developed in order to better fit the characteristics of GNSS ΔIWV data, namely a monthly variance and a smoothly varying bias (see Section 2.1.4). Segmentation models which do not include these features were shown to fail (see Chapter 3). The main idea for the new model is to add a functional part to the model proposed by Bock *et al.* [2018] which already modeled the monthly variance (see Section 3.6). The proposed model is described in Section 4.1 and the inference procedure in Section 4.2. As classical in segmentation (see Chapter 3), the inference is performed in two steps: 1) estimation of the variance, the functional part and the segmentation parameters for fixed a number of segments, and 2) selection of the number of segments. An additional algorithmic difficulty for this new model compared to the model of Bock *et al.* [2018] is that the function is a global parameter and again hampers to use DP. To circumvent this problem, we propose to estimate iteratively the functional part and the segmentation parameters. Section 4.3 presents a numerical simulation study to assess the performance of the proposed method. Note that the associated algorithm results from tests of different possible variants (e.g. updating the monthly variance during the iterative procedure instead of estimating it once at the beginning, or interchanging the order of the estimation of the functional and the segmentation at the initialization). Section 4.4 presents the results of these variants. Finally, Section 4.5 presents the R packages GNSSseg and GNSSfast which are

two releases of the method that have been made available to the community.

4.1 Model

Let be $\boldsymbol{y} = \{y_t\}_{1,\dots,n}$ the observed series with length n that is supposed to be modeled by a Gaussian independent random process $\boldsymbol{Y} = \{Y_t\}_{t=1,\dots,n}$ such that

(i) the mean of $\boldsymbol{Y}$ is composed of two terms:

- a piece-wise constant function $\mu_k(t)$ equal to μ_k on the interval $I_k^{\text{mean}} = [\![t_{k-1}+1, t_k]\!]$ with length $n_k = t_k - t_{k-1}$ where $0 = t_0 < t_1 < \ldots < t_{K-1} < t_K = n$. The $\mathbf{T} = \{t_k\}_{k=1,\dots,K-1}$ are the times of the abrupt changes or change-points and K is the number of intervals or segments.
- and a function f_t;

(ii) the variance of $\boldsymbol{Y}$ is month-dependent, i.e. it is constant within the interval $I_{\text{month}}^{\text{var}} = \{t; \text{date}(t) \in \text{month}\}$ with length n_{month} where $\text{date}(t)$ stands for the date at the position t.

The resulting model is thus the following

$$Y_t = \mu_k + f_t + \mathcal{E}_t, \quad \forall t \in I_k^{\text{mean}} \cap I_{\text{month}}^{\text{var}}, \tag{4.1}$$

where the errors $\{\mathcal{E}_t\}_t$ are centered independent Gaussian with heterogeneous variance, i.e. $\{\mathcal{E}_t\}_t$ i.i.d. $\sim \mathcal{N}(0, \sigma^2_{\text{month}})$ if $t \in I_{\text{month}}^{\text{var}}$ and for $k = 1, \dots, K$. The intervals $\{I_k^{\text{mean}}\}_k$ are unknown contrary to the intervals $\{I_{\text{month}}^{\text{var}}\}_{\text{month}}$ that are fixed. The functional component f_t describes the smooth variations of mean of the series ΔIWV.

4.2 Inference

As usual in segmentation based on the maximum likelihood inference procedure framework (see Chapter 3), the inference is performed in two steps that are here:

Step 1 Estimate $\boldsymbol{T} = (t_1, \dots, t_{K-1})$ the $K-1$ change-points, $\boldsymbol{\mu} = (\mu_k)_k$ the K means, $\boldsymbol{\sigma}^2 = (\sigma^2_{\text{month}})_{\text{month}}$ the monthly variances, and f the function, with K being fixed.

Step 2 Choose the number of segments K.

The log-likelihood of the model defined by Eq. (4.1) is equal to

$$\begin{aligned} \log\ p(\boldsymbol{y}; K, \boldsymbol{T}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, f) \quad &= \quad -\frac{n}{2}\log(2\pi) \sum_{\text{month}} \frac{n_{\text{month}}}{2} \log(\sigma^2_{\text{month}}) \\ &\quad -\frac{1}{2}\sum_{k=1}^{K} \sum_{\text{month}} \sum_{t \in I_k^{\text{mean}} \cap I^{\text{var}}_{\text{month}}} \frac{(y_t - \mu_k - f_t)^2}{\sigma^2_{\text{month}}} \end{aligned} \tag{4.2}$$

4.2.1 Step 1: Inference of $\boldsymbol{T}$, $\boldsymbol{\mu}$, $\boldsymbol{\sigma}^2$ and f, with K being fixed

As we have seen in Chapter 3, the use of the DP algorithm is now classical to estimate the change-point positions. This is the only algorithm that enables to retrieve the exact maximum likelihood solution. However, DP can be applied if and only if the quantity to be optimized is additive with respect to the segments. Here the presence of the 'global' parameters σ^2_{month} and f will link the segments and the required condition will not be satisfied. In order to circumvent this algorithmical problem and keep the use of DP, we propose to proceed in two steps: (1) we estimate the variances using a robust estimator as in Chakar *et al.* [2017] and Bock *et al.* [2018] and (2) we estimate iteratively f and the segmentation parameters (i.e. the change-points and the means) using DP as in Gazeaux *et al.* [2015] and Bertin *et al.* [2017].

The proposed algorithm is thus the following:

(1) Estimation of σ^2_{month} . As seen in Section 3.6, Bock *et al.* [2018] proposed a consistent estimator for the variance parameter based on the robust one proposed by Rousseeuw & Croux [1993]. This estimator is defined by Eq. (3.7). We use here this estimator even in the presence of the function f because the latter does not have much impact on the resulting estimation (in the application, this smoothly varying bias is almost completely cancelled out in the differentiated series). The estimated variance is noted $\widehat{\sigma}^2_{\text{month}}$.

(2) Estimation of f and both $\boldsymbol{T}$ and $\boldsymbol{\mu}$ iteratively . The procedure consists in minimizing the minus log-likelihood given in Eq. (4.2) iteratively. At iteration $[h+1]$:

(*a*) the estimator of f is the weighted least-squares estimator with weights $1/\widehat{\sigma}^2_{\text{month}}$ applied to $\{y_t - \mu_k^{[h]}\}_t$. Based on the seasonal character of the smoothly varying bias observed for station CCJM (Figure 2.6 (b)) and shared by many other stations, we decided to represent f as a Fourier series of order 4 accounting for annual, semi-annual, terannual, and quarterly periodicities in the signal:

$$f_t = \sum_{i=1}^{4} a_i \cos(w_i t) + b_i \sin(w_i t),$$

where $w_i = 2\pi \frac{i}{L}$ is the angular frequency of period L/i and L is the mean length of the year ($L = 365.25$ days when time t is expressed in days). The estimated function is denoted $f^{[h+1]}$.

(b) the segmentation parameters are estimated based on $\{y_t - f_t^{[h+1]}\}_t$. We get

$$\mu_k^{[h+1]} = \frac{\sum_{\text{month}} \sum_{t \in I_k^{\text{mean}} \cap I_{\text{month}}^{\text{var}}} \frac{(y_t - f_t^{[h+1]})}{\widehat{\sigma}^2_{\text{month}}}}{\sum_{\text{month}} \sum_{t \in I_k^{\text{mean}} \cap I_{\text{month}}^{\text{var}}} \frac{1}{\widehat{\sigma}^2_{\text{month}}}}, \tag{4.3}$$

and

$$\boldsymbol{T}^{[h+1]} = \underset{\boldsymbol{T} \in \mathcal{M}_{K,n}}{\operatorname{argmin}} \sum_{k=1}^{K} \sum_{\text{month}} \sum_{t \in I_k^{\text{mean}} \cap I_{\text{month}}^{\text{var}}} \frac{(y_t - f_t^{[h+1]} - \mu_k^{[h+1]})^2}{\widehat{\sigma}^2_{\text{month}}},$$

where we recall that $\mathcal{M}_{K,n} = \{(t_1, \ldots, t_{K-1}) \in \mathbb{N}^{K-1}, 0 = t_0 < t_1 < \ldots, t_{K-1} < t_K = n\}$ is the set of all the possible partitions of the grid $[\![1, n]\!]$ in K segments. This minimization can now be obtained using DP where the cost of the segment $[\![i, j]\!]$ is

$$\mathcal{C}(i, j) = \sum_{t=i}^{j} \frac{(y_t - f_t^{[h+1]} - \widehat{\mu}_{i,j}^{[h+1]})^2}{\widehat{\sigma}_t^2},$$

where $\widehat{\mu}_{i,j}^{[h+1]}$ is the estimated mean in the considering segment given by Eq. (4.3) and $\widehat{\sigma}_t^2$ is the value of the variance at position t.

The final estimators are denoted $\widehat{f}$, $\widehat{\boldsymbol{T}}$ and $\widehat{\boldsymbol{\mu}}$.

4.2.2 Choice of K

The best K is selected again using three model selection criteria presented in Section 3.4: the ones proposed by Lavielle [2005], Birgé & Massart [2001] and Zhang & Siegmund [2007] denoted respectively Lav, BM and mBIC. Since in our estimation procedure the variances are estimated first, our segmentation problem can be seen as one in which the variance is 'known'. We thus propose to use the least-squares based criterion defined as follows:

$$c(\boldsymbol{y}; K, \widehat{\boldsymbol{T}}, \widehat{\boldsymbol{\theta}}) = \sum_{k=1}^{K} \sum_{\text{month}} \sum_{t \in \widehat{I}_k^{\text{mean}} \cap I_{\text{month}}^{\text{var}}} \frac{(y_t - \widehat{f}_t - \widehat{\mu}_k)^2}{\widehat{\sigma}^2_{month}}. \tag{4.4}$$

Some remarks about the two first criteria which involves penalty constants to be calibrated:

- for the Lav's criterion, β is the penalty constant chosen using an adaptive method . The method involves a threshold S which is fixed to $S = 0.75$, as suggested by Lavielle [2005].
- for the BM's criterion, the penalty constant α can be calibrated using the slope heuristic proposed

by Arlot & Massart [2009]. Two methods are proposed actually: the "dimension jump" and the "data-driven slope estimation" which are referred to as BM1 and BM2, respectively, hereafter.

4.2.3 In practice and different choices

In practice, the iterative procedure in Step 1 of the inference (Section 4.2.1) is performed for $K = 1, \ldots, K_{\max}$ where $K_{\max}$ should be 2 or 3 times larger than the expected number of change-points. For both the simulations (hereafter) and the applications (in Chapter 5), we used $K_{\max} = 30$.

The iterative procedure needs a proper initialization procedure and a stopping rule. For the initialization, the function f is estimated first, using a unweighted least-squares criterion. For the stopping rule the change of f_t and μ_k between two successive iterations is checked against a fixed threshold. The convergence of the iterative procedure is accelerated using the stopping test proposed by Varadhan & Roland [2008].

The final algorithm was derived after testing several different options discussed in Section 4.4. It is summarized in Figure 4.1.

4.3 Simulations

4.3.1 Simulation Design and Quality Criteria.

Simulation Design. The same simulation design as in Section 3.6 was used here. The time series have a length of $n = 400$ with 4 "years" of 2 "months" of 50 "days" each and with standard deviations changing every month, and with 6 change-points at positions $t = 55, 77, 177, 222, 300, 366$ and mean values alternating between 0 and 1. The periodic function was again modelled by $f(t) = 0.7\cos(2\pi t/L)$ where $L = 100$ is the length of one year. Since we consider here only two months, the standard deviations is alternating between two values, σ_1 and σ_2, for which several batches of simulations were generated with different values: $\sigma_1 = 0.1$, 0.5, or 0.9 and $\sigma_2 = 0.1$ to 1.5 by step of 0.2. Each batch contained 100 time series. Figure 4.2 shows an example of one such series.

Quality Criteria. The quality of the results will be quantified by analyzing the differences between the estimates and their corresponding true values. In the following, the estimates will be denoted with a hat $\widehat{x}$ and the true values with a star superscript $x^\star$.

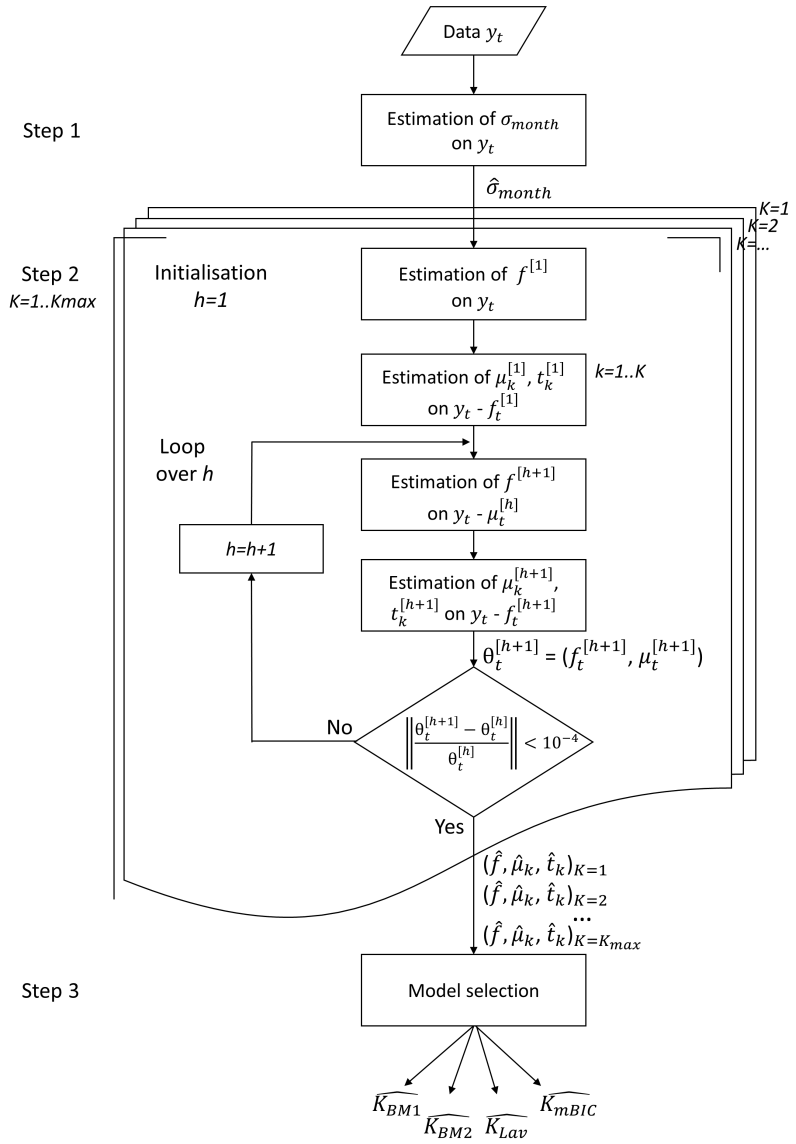


Figure 4.1 – Schematic of the algorithm.

1. for the estimated variance or the standard deviation parameters, we analyze the difference with respect to the true values for each of the two months;

2. for the function f, we compute the root mean square error (RMSE) of the estimated function: $\text{RMSE}(f) = \left[\frac{1}{n} \sum_{t=1}^{n} \left\{ \widehat{f}_t - f_t^* \right\}^2 \right]^{1/2}$;

3. for the segmentation parameters, the following criteria are considered:

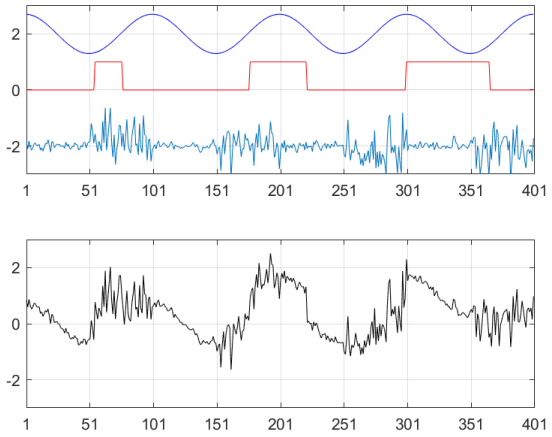


Figure 4.2 – Example of a simulated time series (black solid line in lower panel) of length $n = 400$ with $K = 7$ segments (red solid line), function $f(t) = 0.7\cos(2\pi t/L)$ (blue solid line), noise (cyan solid line) with standard deviation $\sigma_1 = 0.1$ and $\sigma_2 = 0.5$ (changing every $L/2 = 50$ points, starting with σ_1).

- ⋆ the difference between the estimated number of segments and the true one $\widehat{K} - K^*$;
- ⋆ the RMSE of the estimated mean parameter $\widehat{\boldsymbol{\mu}}$:
  $\text{RMSE}(\boldsymbol{\mu}) = \left[\frac{1}{n}\sum_{t=1}^{n}\{\widehat{\mu}_t - \mu_t^*\}^2\right]^{1/2}$;
- ⋆ the distance between the estimated positions of the change-points $\widehat{\boldsymbol{t}}$ and the true ones $\boldsymbol{t}^\star$; this distance is measured with the help of the two components of the Hausdorff distance, d_1 and d_2, defined as:

$$d_1(a,b) = \max_b \min_a |a-b| \quad \text{and} \quad d_2(a,b) = d_1(b,a).$$

In our case, $d_1(\boldsymbol{t}^\star, \widehat{\boldsymbol{t}})$ quantifies the largest distance between an estimated change-point and the true ones. However, it does not say if some change-points are missing. This complementary information is given by $d_2(\boldsymbol{t}^\star, \widehat{\boldsymbol{t}})$ which quantifies how close the true change-points are to the detected ones. A perfect segmentation results in both null d_1 and d_2. A small d_1 means that the detected change-points are well positioned and a small d_2 that a large part of the true change-points are correctly detected. A common situation found in practice is the one where the number of change-points is under-estimated, with a small d_1 and a large d_2. In that case, some change-points are undetected but the detected ones are correctly located. This situation is satisfying here since in our applications one prefer to miss some (small) change-points rather than having too many false detections.

⋆ the histogram of the change-point locations provide a measure of the probability of the position of the change-points.

4.3.2 Simulation Results

Only the results for $\sigma_1^\star = 0.5$ are illustrated hereafter. The results for the others values of $\sigma_1^\star$ are discussed at the end of the Section.

Accuracy of the variance estimates. Figure 4.3 presents the estimation errors of $\hat{\sigma}_1$ and $\hat{\sigma}_2$ for different values of $\sigma_2^\star$. It is seen that the variance estimator works well and the estimated standard deviations are retrieved with the same accuracy as in Bock *et al.* [2018] despite the presence of the periodic bias. The dispersion increases when $\sigma_2^\star$ is increasing as one can expect.

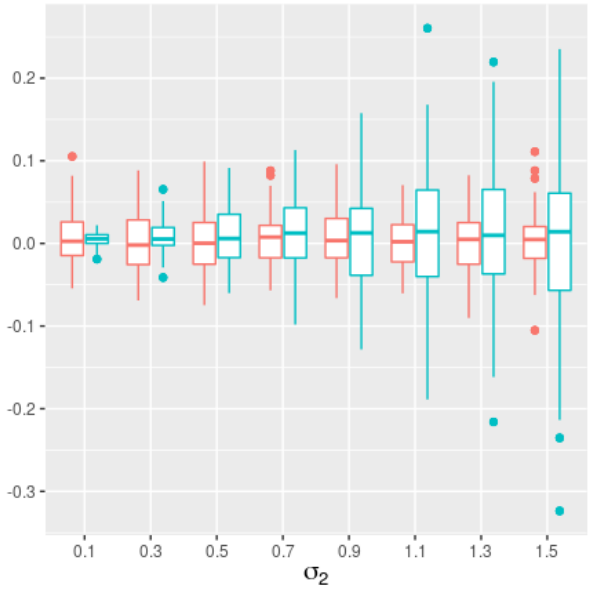


Figure 4.3 – Boxplots of standard deviation estimation errors: $\hat{\sigma}_1 - \sigma_1^\star$ in red and $\hat{\sigma}_2 - \sigma_2^\star$ in blue, with $\sigma_1^\star$=0.5 and $\sigma_2^\star = 0.1, \ldots, 1.5$. Each case includes 100 simulations.

Accuracy of segmentation parameter estimates. Figure 4.4 shows the results for the four model selection criteria and the special case where the number of segments K is fixed to the true value ($K = 7$). For small values of $\sigma_2^\star$, the detection problem is easy and all the model selection criteria retrieve the correct number of segments (Figure 4.4(a)). However for large values of $\sigma_2^\star$, the detection becomes difficult, and the errors increase. The different criteria behave slightly differently. Lav tends to

	Nsta	detect	valid	valid
mBIC	120	3251	267	8.2%
Lav	114	474	75	15.8%
BM1	98	335	70	20.9%
BM2	107	435	77	17.7%

Table 4.1

give the true number of segments in median, but with a large dispersion, while BM1, BM2, and mBIC tend to underestimate the number of segments (more for mBIC). However, finding the correct number of segments does not mean that the change-points are properly positioned. Indeed, for Lav and the case when $K = 7$, the median d_1 is still quite large (Figure 4.4(c)). On the other hand, the median d_2 is smaller for the case when $K = 7$ compared to the tested criteria (Figure 4.4(d)). Finally, RMSE($\boldsymbol{\mu}$) is very similar for all the criteria (Figure 4.4(b)), though Lav shows a larger median and dispersion when $\sigma_2^\star$ is large. When $\sigma_2^\star$ takes intermediate values the case when $K = 7$ yields slightly improved results.

4.1

Probability of detection. Figure 4.5 shows the percentage of the change-point detections for three values of $\sigma_2^\star = 0.1, 0.5$ and 1.5, and $\sigma_1^\star = 0.5$. In general, the change-points located in the "months" with smaller variance are more often recovered with all three criteria, and also when the true K is used. Hence, in the case (a) when $\sigma_1^\star = 0.5$ and $\sigma_2^\star = 0.1$, the probability of detection is slightly smaller for the position 222, which is contained in a segment with $\sigma_1^\star = 0.5$, and for the position 300 where both the mean and the variance change. In the case (b) when $\sigma_1^\star = \sigma_2^\star = 0.5$, the probability of detection is more or less the same for all the change-points and all the criteria. When $\sigma_2^\star = 1.5$, the problem is more complicated. Again the change-points located in the "months" with smaller noise are better detected (positions 222 and 300) but for the other four change-points the results are contrasted although they are all located in months with $\sigma_2^\star = 1.5$. The change-points at 55 and 77 are almost never detected. For mBIC this is consistent with the fact that the median $\hat{K}$=5, i.e. two change-points are missing, on average (Figure 4.4(a)), but the other four change-points are not so badly located (d_1 is not that large, Figure 4.4(c), but d_2 is very large, Figure 4.4(d)). The situation is a bit similar for BM1. On the other hand, for Lav and the true K, the number of detections is correct (on average for Lav) but due to the large noise they are sometimes very badly positioned (large d_1 and d_2).

Accuracy of the function estimate. Figure 4.6 shows RMSE(f) as a function of $\sigma_2^\star$. As expected, the errors increase when $\sigma_2^\star$ increases. The results do not much depend on the selection criterion, but the

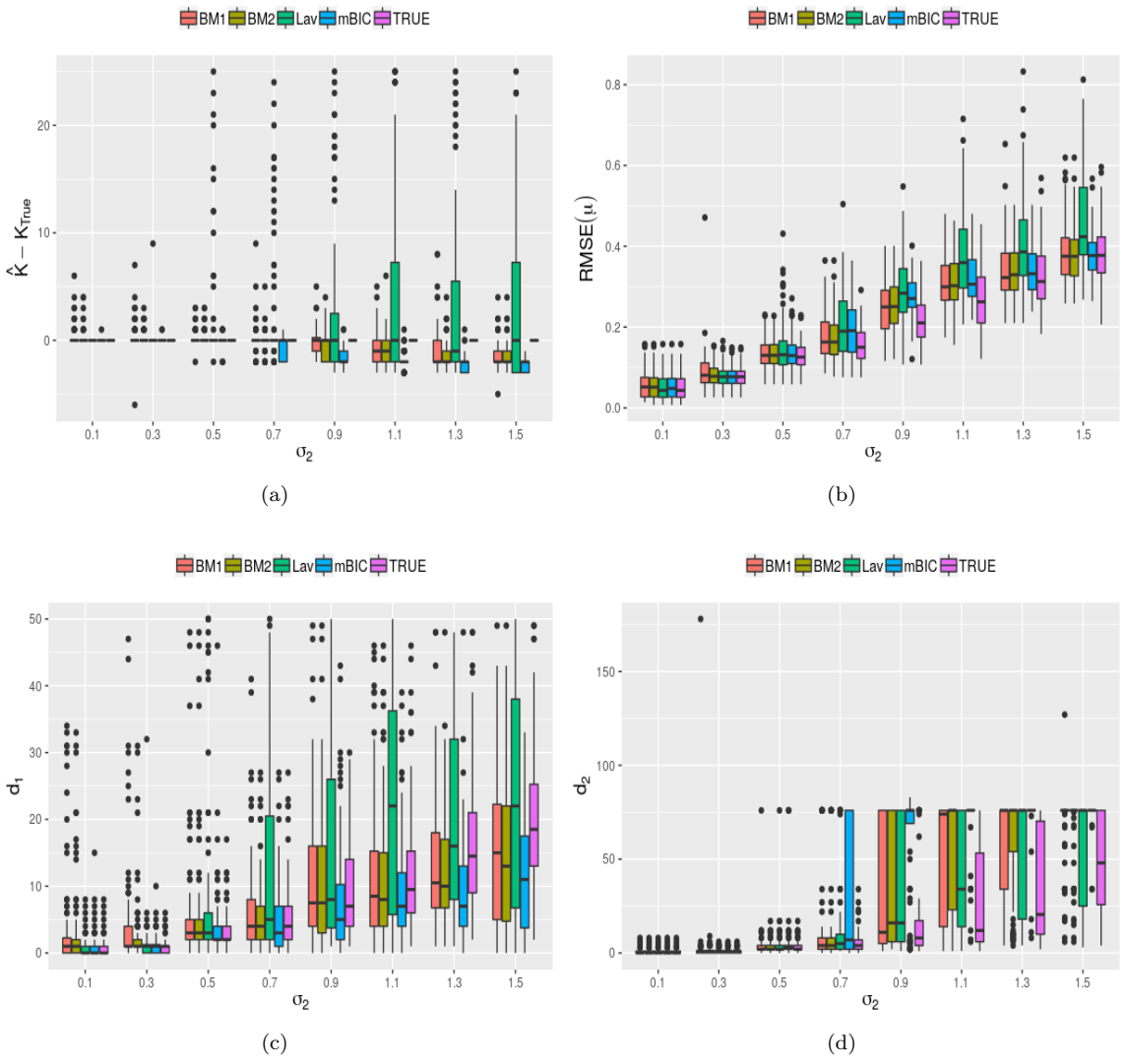


Figure 4.4 – Results with the four selection criteria (BM1, BM2, Lav, and mBIC) and with the true number of segments (True), for $\sigma_1^\star = 0.5$ and different values of $\sigma_2^\star$. *(a)* $\hat{K} - K^\star$; *(b)* RMSE($\boldsymbol{\mu}$); *(c)* first Hausdorff distance d_1 and *(d)* second Hausdorff distance d_2.

results are slightly better when the true number of segments is known and when $\sigma_2^\star$ takes intermediate values. The results for Lav show a slightly larger median and larger dispersion.

The results for other values of $\sigma_1^\star$ are very similar for BM1, BM2, mBIC, and the case when the true K is used. The results are slightly improved for $\sigma_1^\star = 0.1$ and slightly degraded for $\sigma_1^\star = 0.9$, as expected, see Figures 4.7 and 4.8, respectively. The results for Lav are more chaotic, with either

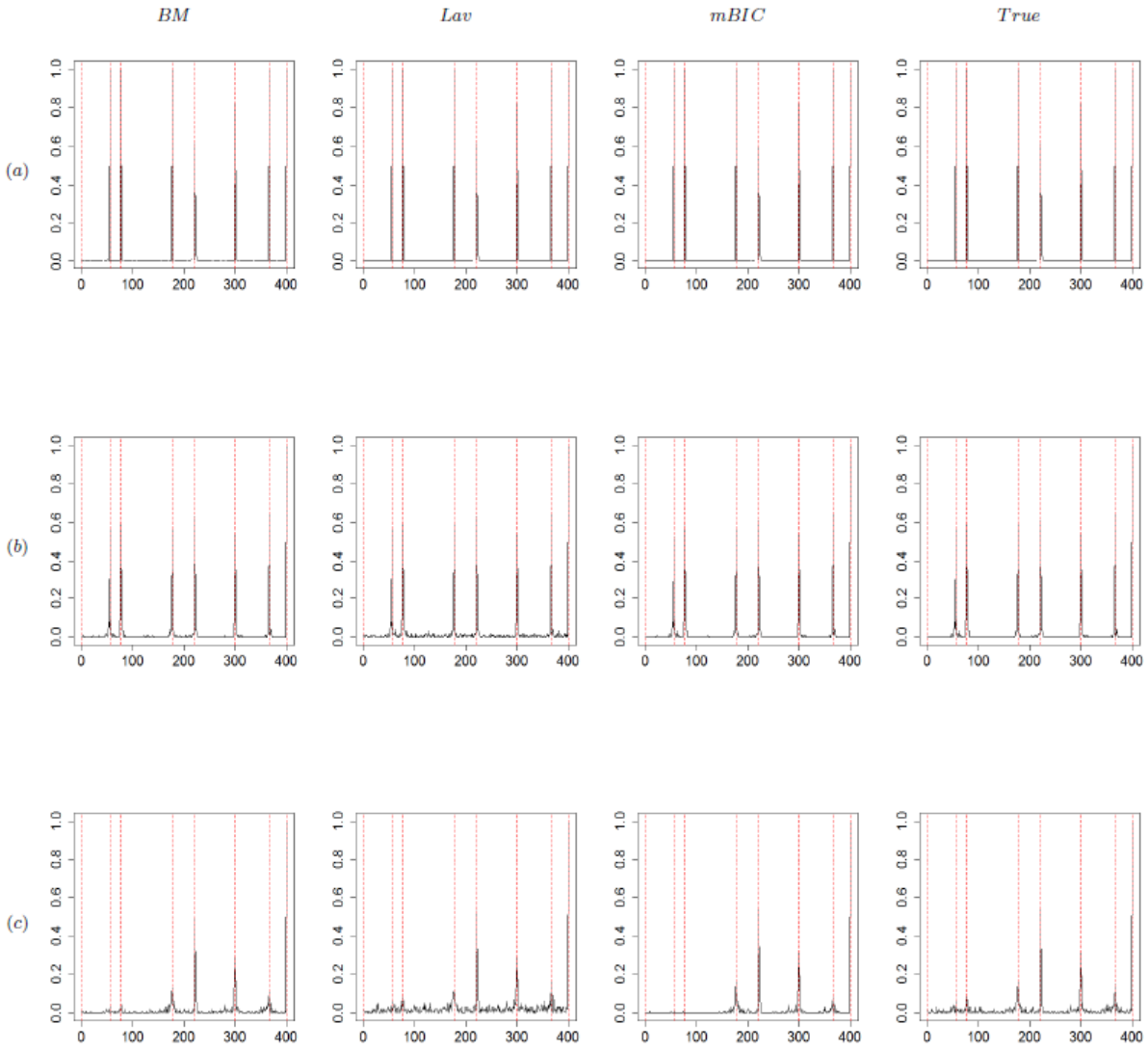


Figure 4.5 – Histogram of change-point detections with, from left to right, the BM, Lav, and mBIC selection criteria, and the case when the true number of segments is used (TRUE), for $\sigma_1^\star = 0.5$ and three different values for $\sigma_2^\star$: (a) $\sigma_2^\star = 0.1$, (b) $\sigma_2^\star = 0.5$ and (c) $\sigma_2^\star = 1.5$. The red dotted lines indicate the positions of the true change-points.

large under-estimation of K for the smaller $\sigma_1^\star$ and over-estimation of K for the larger $\sigma_1^\star$, with large subsequent degradation of the other quality criteria. In general, under-estimating K leads to an increase of RMSE(μ), while over-estimating K leads to an increase of d_1.

The main conclusions from the simulation study are the following:

- The proposed method works well but the results are sensitive to the choice of the function form

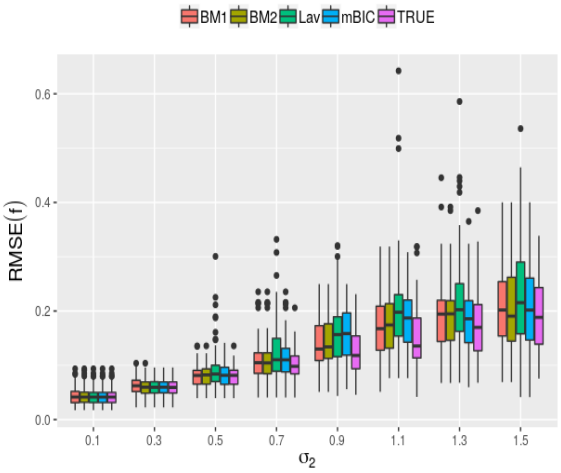


Figure 4.6 – RMSE of the estimated function f for $\sigma_1^\star = 0.5$ and different values for $\sigma_2^\star$.

due to its possible confusion with the change-points. Performing a selection of the statistically significant parameters of the function appears as a good way to reduce this problem and improves slightly the change-point detection with our simulated data (see Section 4.4).

- Concerning the model selection criteria, BM1, BM2, and mBIC, provide very similar results. They behave well and detect correctly the number and position of change-points when the noise is not too large. When the noise is heavy some change-points are missed but this is a counterpart of the limited number of false detections. The Lav criterion shows much larger dispersion in the number of change-points and, though the estimated number is close to the truth in median, some change-points are not properly located (larger d_1 and d_2) with an impact on the estimated $\boldsymbol{\mu}$ and f.

4.4 Tested Alternatives

In this section, we discuss several other implementations of the algorithm that were tested before selecting the final form described in Section 4.2.

(1) Updating the variance: instead of estimating the standard deviation vector $\boldsymbol{\sigma}$ once at the beginning only (step 1), we tested a version of the algorithm where $\boldsymbol{\sigma}$ was updated at each iteration in the loop (step 2). Figure 4.9 shows the results when the variance vector is updated during the iterative procedure: a positive impact of this procedure was to provide slightly more accurate estimates for the variance with also a positive impact on the estimated function, f_t, and the seg-

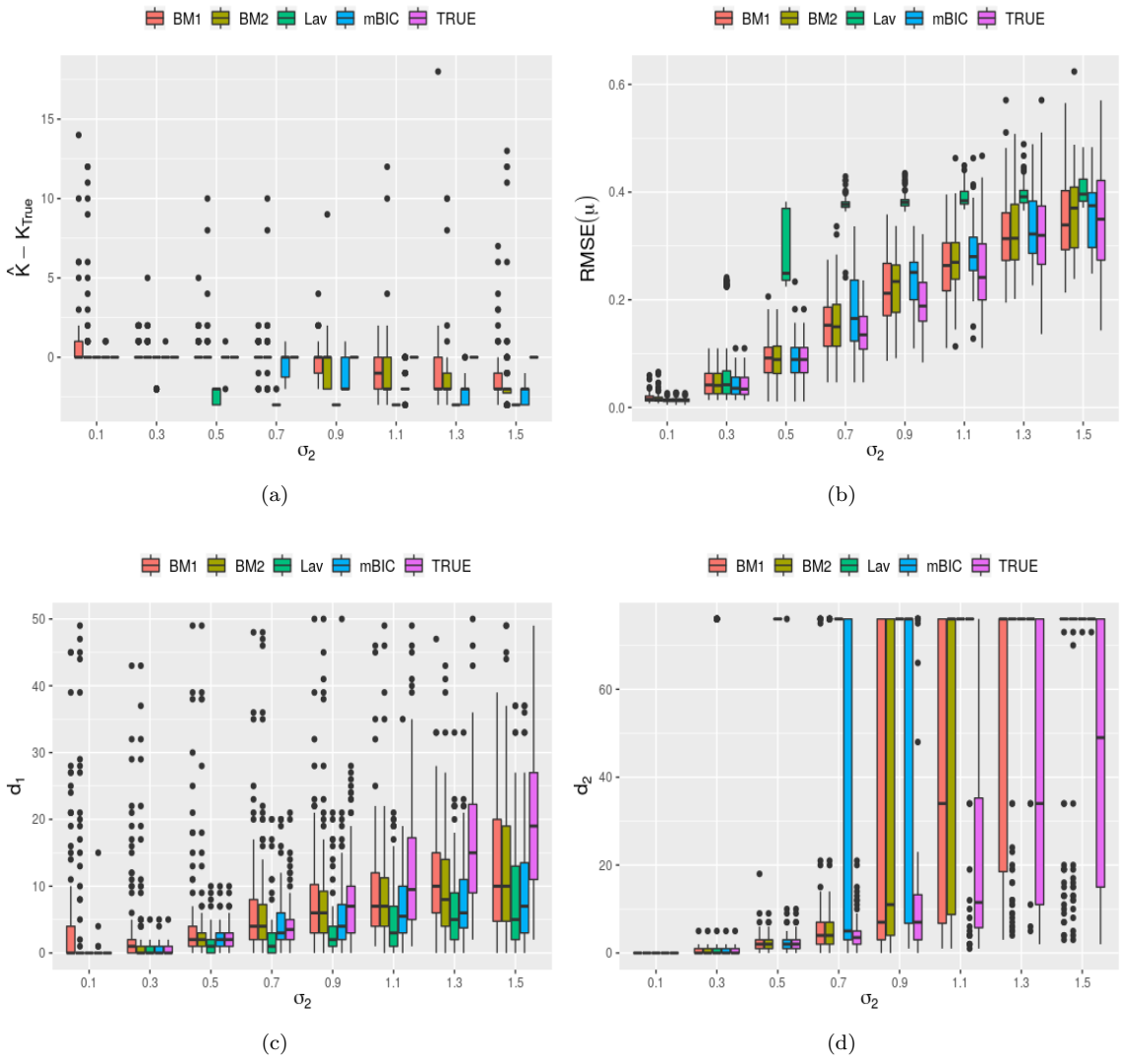


Figure 4.7 – Results with the four selection criteria (BM1, BM2, Lav, and mBIC) and with the true number of segments (True), for $\sigma_1^\star = 0.1$ and different values of $\sigma_2^\star$. *(a)* $\hat{K} - K^\star$; *(b)* RMSE($\boldsymbol{\mu}$); *(c)* first Hausdorff distance d_1 and *(d)* second Hausdorff distance d_2.

mentation parameters ($\boldsymbol{\mu}$ and $\boldsymbol{T}$). However, the small changes in variance at each iteration slowed down the convergence of the algorithm with actually only small improvement of the accuracy of the estimated function and parameters in the end. For this reason, we did not select this procedure. This test also showed that our method is not very sensitive to the accuracy of the variance.

(2) Variants of the initialization. In the standard initialization procedure described in Section 4.2, f is estimated first using an unweighted regression and then the segmentation is performed on

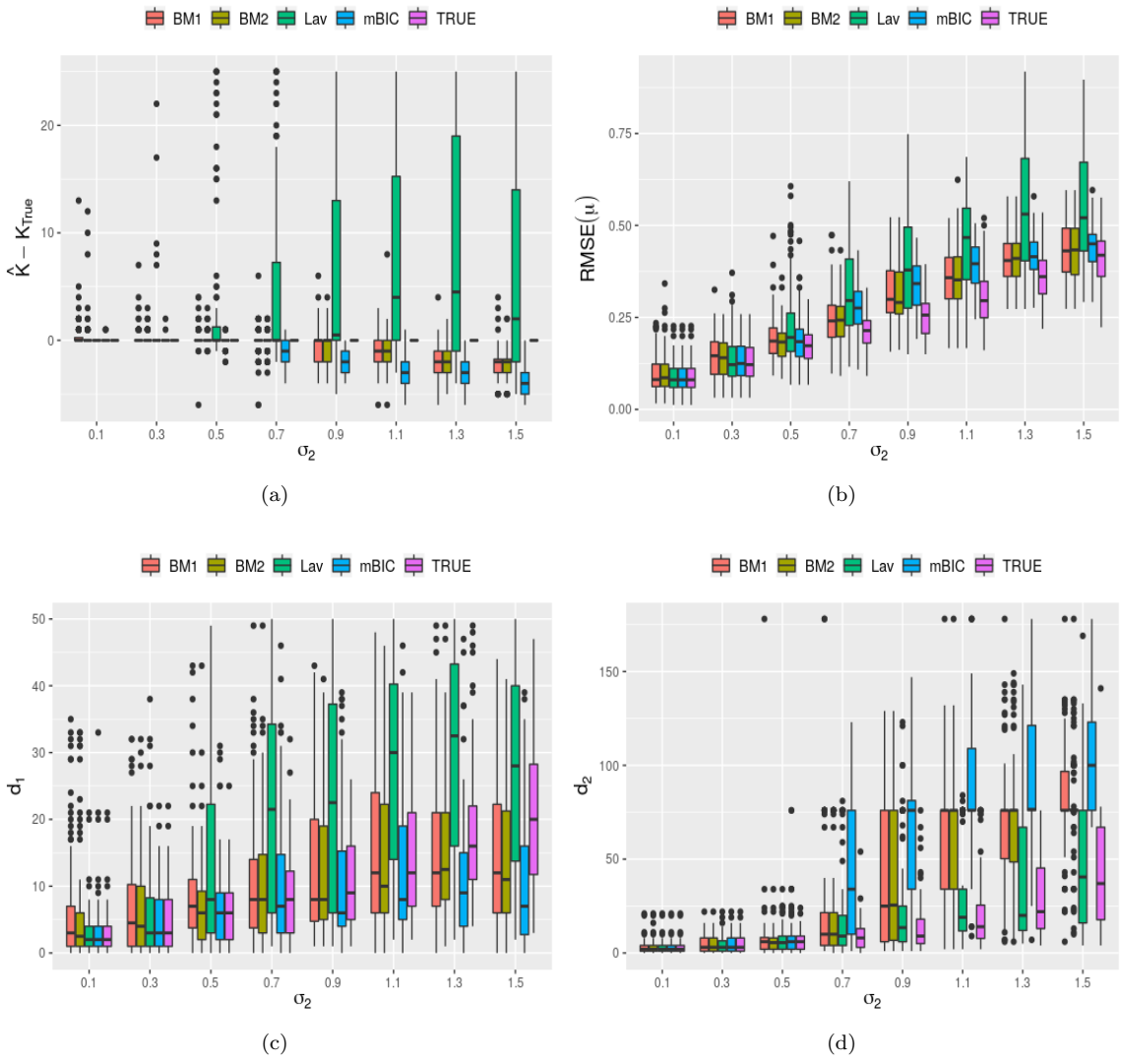


Figure 4.8 – Results with the four selection criteria (BM1, BM2, Lav, and mBIC) and with the true number of segments (True), for $\sigma_1^\star = 0.9$ and different values of $\sigma_2^\star$. *(a)* $\hat{K} - K^\star$; *(b)* RMSE($\boldsymbol{\mu}$); *(c)* first Hausdorff distance d_1 and *(d)* second Hausdorff distance d_2.

$y_t - f_t$. Here we show the results for three variants: (a) the segmentation is performed first and then f is estimated on $y_t - \mu_t$ using a weighted regression (b) f is estimated first using a weighted regression (as in the loop) and then the segmentation is performed on $y_t - f_t$; (c) f is estimated first using a weighted regression (as in the loop) but on on $y_t - \bar{y}$ (recentred signal).

Figure 4.10 shows the results for case (a). Compared to Figure 4.4, the results are significantly degraded for all values of σ_2. Especially, the larger d_1 indicates that change-points are badly

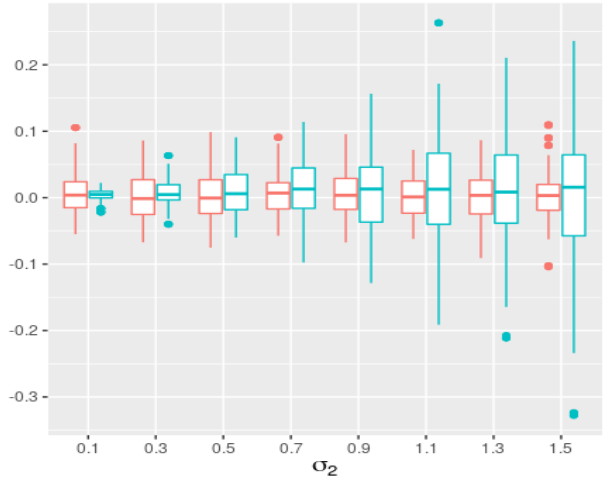


Figure 4.9 – Boxplots of standard deviation estimation errors for the alternative (1) in 4.4, when the variance vector is updated during the iterative procedure: $\hat{\sigma}_1 - \sigma_1^\star$ in red and $\hat{\sigma}_2 - \sigma_2^\star$ in blue, with $\sigma_1^\star$=0.5 and $\sigma_2^\star = 0.1, \ldots, 1.5$. Each case includes 100 simulations.

located. The reason is that in the initialization, the segmentation catches the variations in the signal due to the periodic function following by a wrong estimation of f. Then the iterative procedure does not change this effect and leads naturally to an over-segmentation in addition of the bad estimation of f. This particularly marked for small values of the noise σ_2 and for the Lav's criterion whatever σ_2.

Figure 4.11 shows the results for case (b). The results are degraded as well but less than previously and mainly for larger σ_2. The reason of this effect can be explained by the fact that the change-points belonging to small variance periods are absorbed by f degrading thus its estimation at this initialization step. And as for the case (a), the iterative procedure does not change too much this effect.

The results for case (c) (not shown here) are very similar to those obtained with our initialization procedure. This alternative is equivalent to include a constant term in the linear regression to estimate f. Its estimation is less degraded compared to case (b) and the loop corrects it.

Our choice of estimating first the function f using an unweighted regression is more flexible in the sense that it does not capture the all segmentation effect at the initialization step allowing thus the iterative procedure to correctly separate the function and the segmentation terms.

(3) Selection of the function model. The sensitivity of the procedure to the initialization step dis-

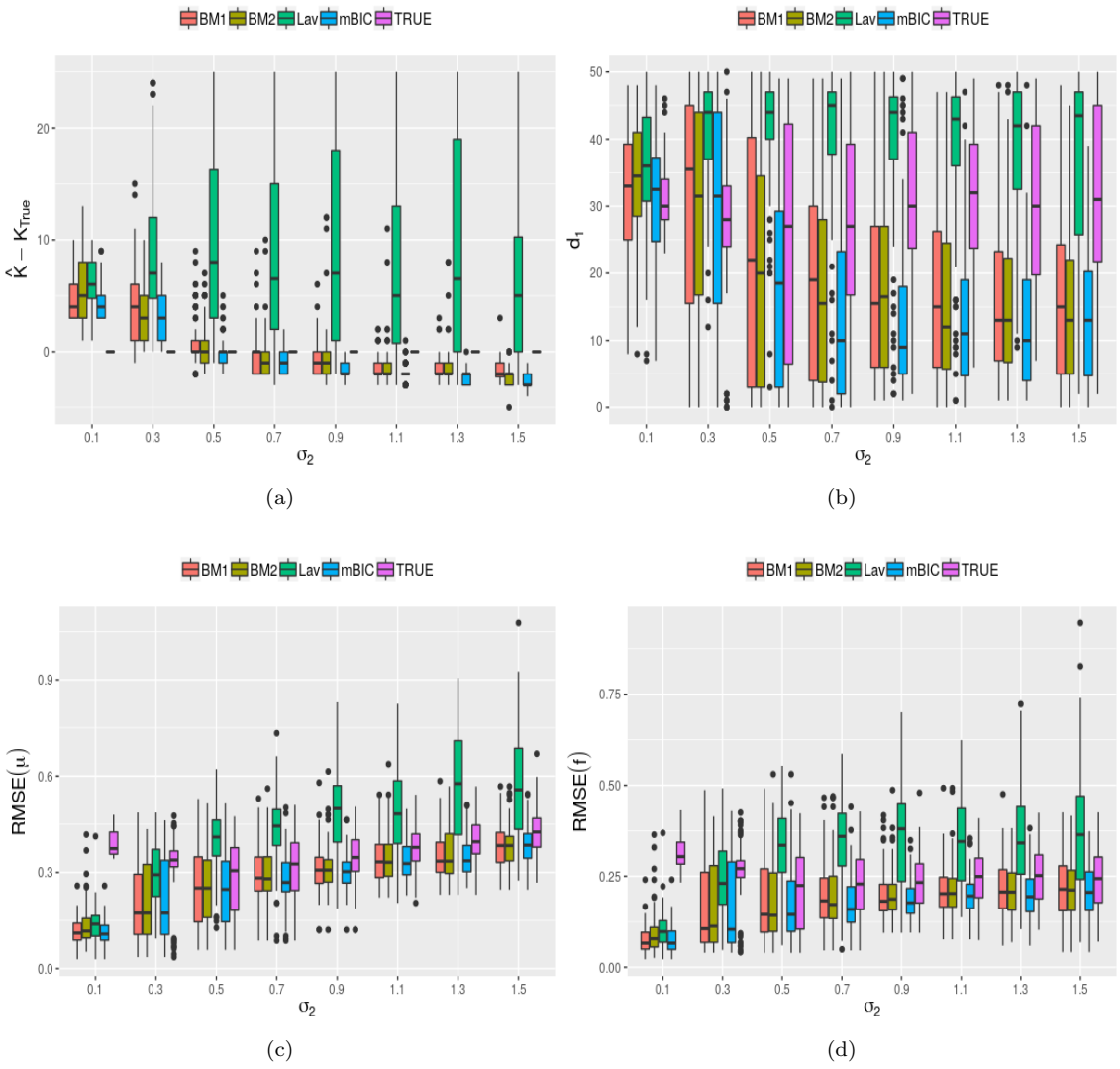


Figure 4.10 – Simulation results when the segmentation is performed first in the initialization step. *(a)* $\hat{K} - K^\star$; *(b)* first Hausdorff distance d_1; *(c)* RMSE($\boldsymbol{\mu}$); *(d)* RMSE(f).

cussed above highlights the possible confusion between the function and segmentation. This sensitivity can be further explored by testing different models for f. The idea behind is that simpler models might be less confused with the segmentation making the procedure more accurate in terms of change-point locations. We tested two alternatives: (a) the shape of f is known up to a scaling factor, i.e. $f_t = a_1 \cos(2\pi t/L)$; (b) the full Fourier series of order 4 is used in the linear regression, then only the statistically significant terms are selected based on their p-values according to a threshold of 0.001. Figures 4.12 and 4.13 show that the results for these two cases are

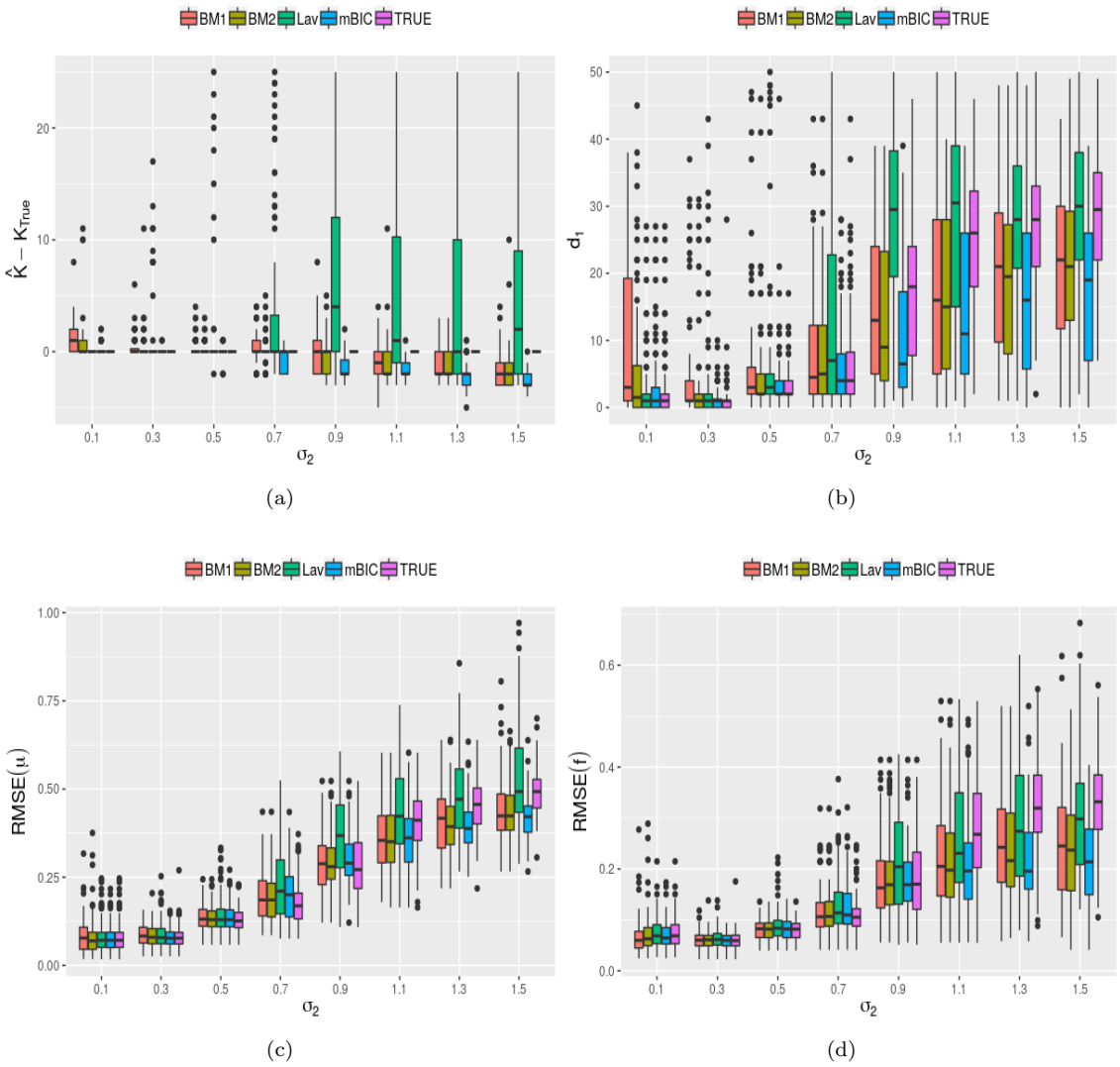


Figure 4.11 – f is estimated using a weighted regression in the initialization step, as a function of σ_2.

both consistent and improve the segmentation results compared to our method (Figures 4.4 and 4.6). Especially, the overall RMSE of the fitted function is strongly reduced. The impact on the positions and amplitudes of the change-points is rather small, however, and the impact in the case of real data is negligible (see Section 5).

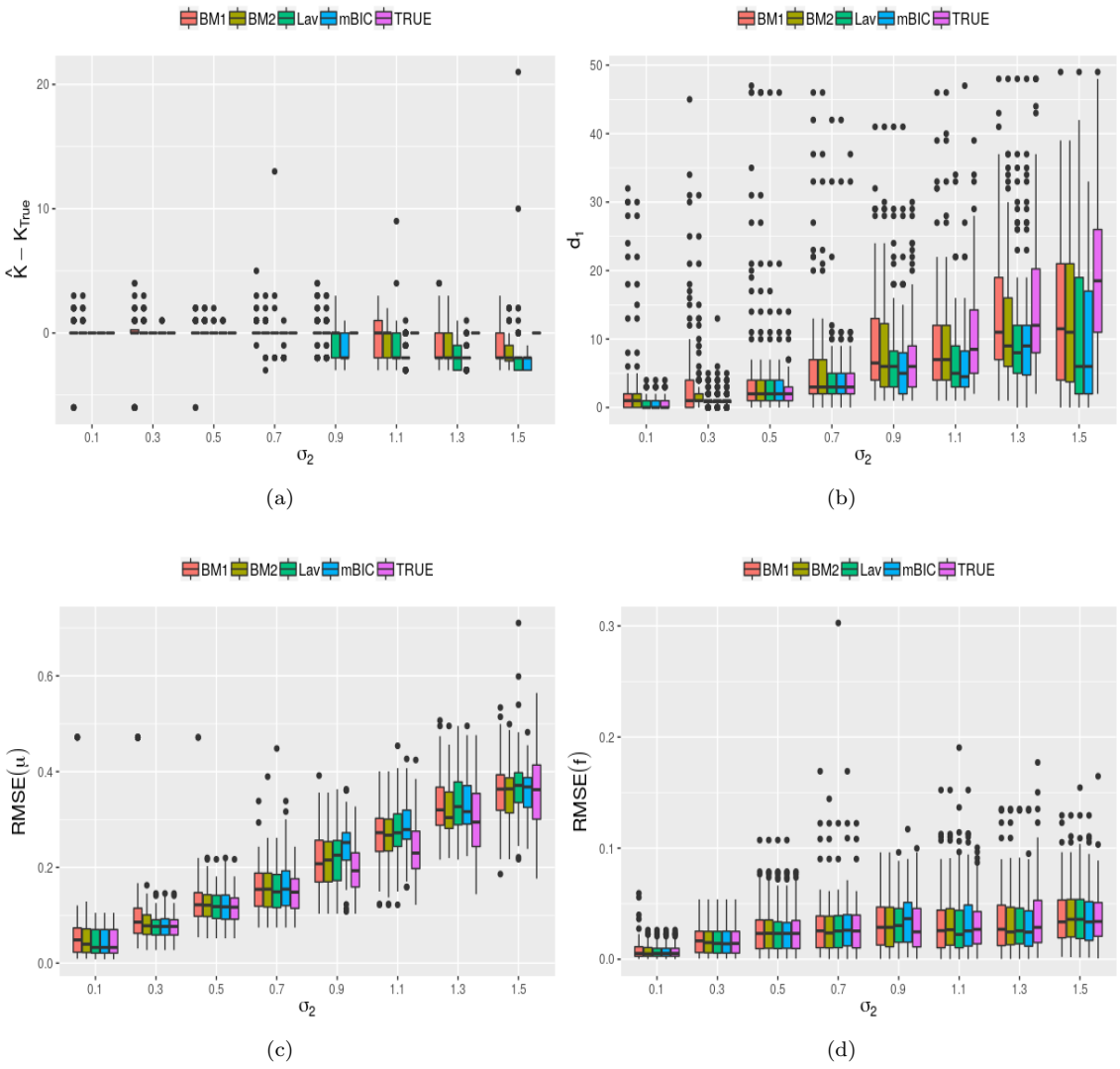


Figure 4.12 – f is estimated using the true shape, as a function of σ_2.

4.5 R packages

4.5.1 Presentation of the package GNSSseg

The procedure, proposed in 4.2, and with its alternatives, is been developed as a R package named `GNSSseg`, avaible on the CRAN (https://cran.r-project.org/web/packages/GNSSseg/index.html). The main function is called GNSSseg with the following arguments:

$$\text{GNSSseg}(Data, lyear, lmin, K_{\max}, selection.K, S, f, selection.f, threshold, tol)$$

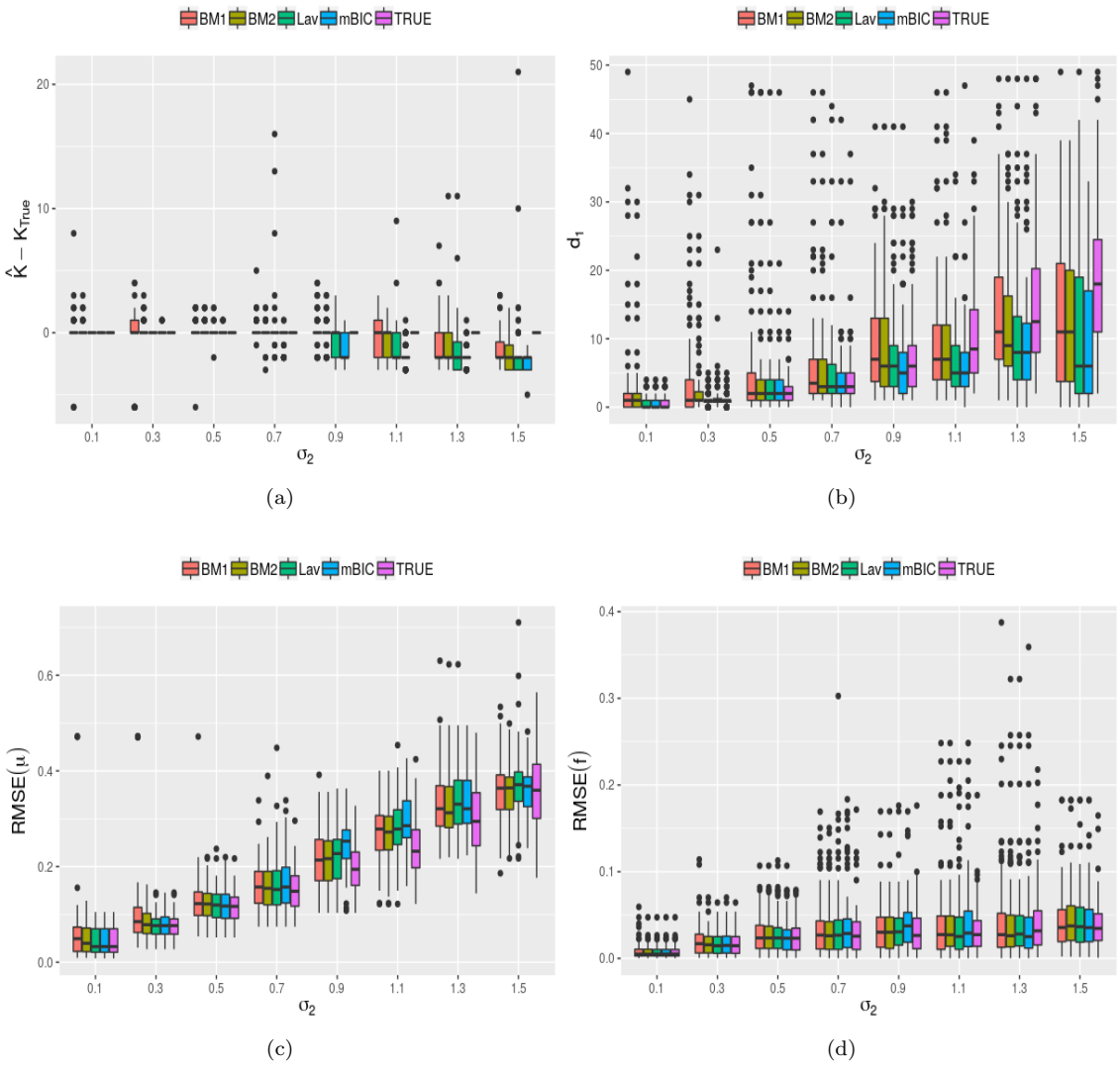


Figure 4.13 – The statistically significant parameters of f are selected (p-values¡0.001).

where:

- *Data* is a data frame containing two fields: *signal* for the signal and *date* for the time information associated to the signal, both fields are vectors of length n. The *date* field should be of class POSIXct (i.e. an implementation of GMT). The time does not need to be continuously sampled (i.e. the data can contain gaps and/or the signal can contain NA values). The temporal resolution of the signal should be daily;
- *lyear* is the length of the year in the signal. Default is 365.25;

- *lmin* is the minimum length of the segments. Default is 1;
- $K_{\max}$ is the maximal number of segments. Default is 30;
- *selection.K* can be used to choose either all (*selection.K*="All") or only one specific selection criterion (possible values are "Lav", "mBIC", "BM_slope" for BM1, and "BM_BJ"for BM2) and also to run the algorithm for one specific value of K only (*selection.K*="none"). In the latter case, the value of K used is the one specified in the variable $K_{\max}$;
- S is the threshold used in the Lav criterion. Default is 0.75
- f allows to skip the estimation of the function and perform only the segmentation (using option f=FALSE). The default value for f is TRUE.
- *selection.f* allows to apply the selection based on a significance test on the parameters of the function, possible values are TRUE and FALSE. The default value is *selection.f*=FALSE (no selection applied).
- *threshold* in the case when *selection.f* is TRUE, this parameter can be used to the threshold on the p-values. The default value is $threshold = 0.05$. Note that a threshold equal to 1 is equivalent to a selection based on AIC only.
- *tol* is the threshold used in stopping rule of the iterative procedure (step 2) of the general algorithm. The default value of *tol* (10^{-4}) was chosen empirically and can be modified. Smaller values will need more iterations for the algorithm to converge;

The main function returns the final estimates of $\widehat{K}, \widehat{t}_k, \widehat{\mu}_k, \widetilde{\sigma}^2_{month}, \widehat{f}_t$ for each of the model selection criteria selected with *selection.K*.

4.5.2 GNSSfast: improvement of execution time

The `GNSSseg` package takes time on long series in practice due to the segmentation applied iteratively. Recently, a faster version of DP has been developed by Hocking *et al.* [2018]. The associated package is `gfpop` and is available on the git repository https://github.com/vrunge/gfpop.git. It reduces the computational time to $\mathcal{O}(Kn \log n)$ (compared to $\mathcal{O}(n^2)$ for DP). We integrated it in `GNSSseg` resulting in a new package called `GNSSfast`. This new package can be downloaded from https://github.com/arq16/GNSSfast.git.

We evaluated empirically the improvement of our algorithm on ten time series from the data described in Section 5.1 on a machine Ubuntu 18.04.2 LTS; the length of the series n is between 5000 and 6000. In average, the segmentation takes 41 minutes (2463 seconds) with `GNSSseg` (DP) against one

minute and half (79 seconds) with `GNSSfast`.

Note that although `GNSSfast` significantly improves the speed, it is not yet possible to change the minimum length of the segments, which is therefore set to 1.

Chapter 5

Application to real data

5.1 Dataset, metadata, and validation procedure

Outline of the data analysis procedure. Figure 5.1 represents the schematic of the homogenisation procedure applied to the real data. The first step consists in forming the daily IWV differences which will served as the input to the homogenisation. The IWV data from the GNSS data set and the reference data set may not be on the same time grid and some resampling, interpolating, and averaging may be necessary. This computation is done by the "comparison" software. In this work we used the reprocessed GNSS IWV data from 120 global GNSS stations (Figure 1.1) and ERA-Interim reanalysis as the reference data set. The reprocessed data set is limited to the period from 1 January 1995 to 31 December 2010. The daily GNSS IWV, ERA-Interim IWV, and the differences are publicly available from Bock [2017].

The second, and main step, is the "segmentation". We applied the new method described in Chapter 4 as well as three variants. The variants implement simplified models we use to investigate the impact of including the monthly variance and smoothly varying bias in the new method. The results are presented in the Section 5.2.

The segmentation method detects sometimes a couple or more change-points located close together which we call outliers. They are usually due to spikes in the noise and shows thus large variations in the mean. Such detections are unwanted and the next step in the processing called "screening" aims at removing them. A basic outlier detection scheme is used in Section 5.2 which compares the position difference between successive change-points to a predetermined threshold of 30 days. A more elaborate screening method is described in later Section 5.3. The screening result is a reduced set of change-points

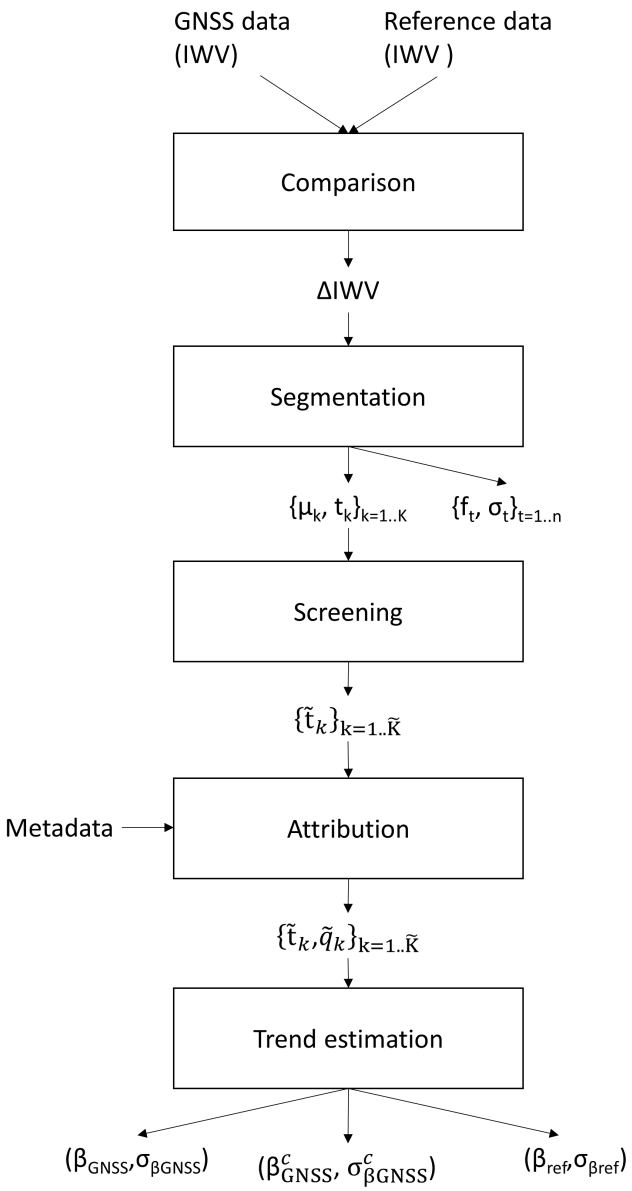


Figure 5.1 – Schema of the general homogenisation procedure on GNSS ΔIWV daily series.

$\{\tilde{t}_k\}$.

The next step is the "attribution" where the detected change-points will be attributed a flag $\{\tilde{q}_k\}$ to say whether their origin can be attributed to "GNSS" with some confidence or not. Different methods of attribution are discussed in Section 5.4. A very straightforward method consists in comparing every detected change-point to the equipment changes known from the metadata. When a coincidence is found

within some predetermined range, the change-point is considered as attributed to GNSS origin. When the range is small (e.g. 30 days), the probability that the matching is due to chance is quite small (on the order of 10^{-4}). In 5.2 we will use this method to "validate" the segmentation results. More details about the usage of the GNSS metadata is given in the next paragraph.

The final step in our analysis is the computation of the trends taking the segmentation results into account. The estimation method and results are presented in 5.5.

GNSS metadata and validation procedure. For validating the detected change-points for the GNSS stations we used the available metadata from the IGS site-logs (ftp://igs.org/pub/station/log/). They contain for each station the dates of changes of receiver (R), antenna (A), and radome (D). Experience shows that equipment changes do not produce systematically a break in the GNSS IWV time series. The most important changes are those affecting the antenna and its electromagnetic environment, the satellite visibility, and the number of observations, Vey *et al.* [2009]. For instance, Ning *et al.* [2016] considered only antenna and radome changes, as well as addition/removal of microwave absorbing material which was known by the authors for one specific station (ONSA). However, there is some evidence that changes in the receiver settings also induce inhomogeneities, e.g. when the elevation cutoff angle is changed (Vey *et al.* [2009]). So we decided to include receiver changes as well. We also included the dates of processing changes (P) which occurred at a few stations in 2008 and 2009, this issue is discussed in Parracho *et al.* [2018]. In principle, the IGS metadata are well maintained but it may happen that some changes are not recorded or that some dates are wrong. Undocumented changes might occur due to changes in the environment, e.g. cutting of vegetation and construction of buildings nearby the antenna as well as seasonal changes in multipath due to growing/declining vegetation may also impact the measurements and produce either abrupt or gradual changes. As a consequence, though metadata represent a valuable source of validation, a full matching between detected change-points and metadata is not to be expected.

Because of noise in the signal, the detected changes may also not coincide perfectly with the known changes and we must allow some flexibility in the validation procedure. A window of 30 days before or after a documented change was used for the automatic validation of the detected change-points. A visual inspection was also performed to check if the invalidated change-points make sense.

5.2 Segmentation Results

5.2.1 General results

The final version of the new segmentation method was applied to IWV differences (GNSS minus ERA-Interim) from 120 stations. Figure 5.2 shows the number of detected change-points for the four criteria. The new method is labelled (a). Three variants of the method are also presented to discuss the sensitivity of the results and the performance of the four selection criteria. The variants are: (b) only the statistically significant terms of the Fourier series are selected (this optional test was introduced in Section 4.4 to test if reducing the number of degrees of freedom in the function leads to better results), (c) only the segmentation is implemented, i.e. the periodic bias modelled by the function f is not included (this is the method proposed by Bock *et al.* [2018]), (d) a homogeneous variance is considered instead of a monthly variance (this is a homoscedastic version but still including the functional). Statistics on the number of detected change-points are included in Figure 5.2. More statistics including the number of validations and outliers are given in Table 5.1. In this sub-Section, change-points are flagged as outliers when their position difference is smaller than 30 days. This choice is consistent with the validation window of 30 days.

Let us first discuss the results for the final version of the new method shown in Figure 5.2(a). Over the 120 GNSS stations, mBIC, Lav, BM1, and BM2 detect a total of 3251, 474, 335, and 435 change-points, respectively. The distribution of the number of change-points per station is very different depending on the selection criterion. Most notably, mBIC detects between 9 and 29 change-points per station, with a mean value of 27.1, i.e. in many cases the largest possible number is selected (29 since $K_{\max} = 30$). This behaviour was not observed with the simulations shown in Section 4. From Table 5.1 we see that mBIC actually has many outliers (2096 out of 3251 detections). Comparison of contrast values reveals that mBIC selects solutions with smaller SSR values than the other criteria, i.e. the model selected by mBIC generally explains better the observed signal. However, this is at the expense of strong over-segmentation, which is not wanted. This penalty criterion is thus not well adapted to the nature of the data analyzed here. One of the reasons might be that the hypothesis of Gaussian errors is not valid with these time series, due to serial correlation in the data mentioned in the 2 (r = 0.249 for station CCJM, Figure 2.6) and spikes in the noise. The three other selection criteria provide much more consistent results, with mean number of change-points of 2.8, 3.6 and 4.0, respectively, for BM1, BM2, and Lav. Among the three criteria, we see from Table 5.1 that BM1 has the smallest number of outliers (36) and the highest rate of validations (20.9%). These two features, and also the fact that BM1 has a reasonable number of change-points (2.8 on average per station), make this selection criterion the preferred one.

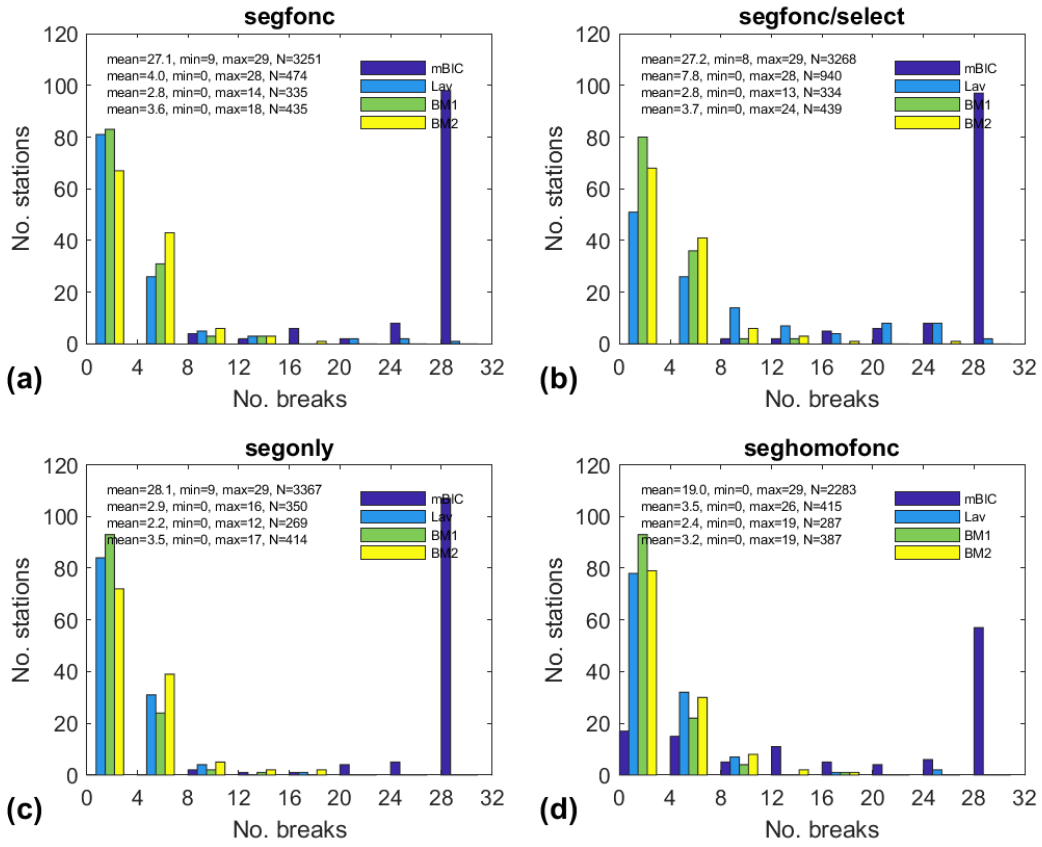


Figure 5.2 – Histograms of the number of change-points detected for four variants of the model selecting criteria (mBIC, Lav, BM1, and BM2). The numbers given in the plots are the mean, min, and max number of change-points detected per station, N is the total number of change-points per method.

These general results are more or less the same for the 3 variants, with some slight differences. Variant (b), with selection of the Fourier series coefficients, shows marginal impact on the number of detections and the number of validations for three criteria (mBIC, BM1, and BM2). Only for Lav do the mean and total number of detections increase (by nearly a factor of 2). We do not have a precise explanation for this, but it reveals some instability in the model selection with this criterion. Instability could also be guessed from the maximal number of detections of 28 already seen in variant (a). It means that in some cases, Lav selects a number of segments very close to the maximum ($K_{\max} = 30$). BM1 and BM2 have also more outliers with this variant, though the total number of detections is almost unchanged. So, contrary to the simulation results, there is no benefit of selecting the functional model with the real data.

Table 5.1 – Comparison of segmentation results for the four variants and the four model selection criteria. From left to right: Number of stations with change-points, min/mean/max number of detected change-points per station, total number of change-points, total number of outliers, total number of validations, percentage of validations including outliers, percentage of validations without outliers.

	Nsta	min	mean	max	detect	outliers	valid	valid	valid without outliers
Variant (a) (segfonc)									
mBIC	120	9	27.1	29	3251	2096	267	8.2%	20.9%
Lav	114	0	4.0	28	474	129	75	15.8%	21.3%
BM1	98	0	2.8	14	335	36	70	20.9%	23.3%
BM2	107	0	3.6	18	435	64	77	17.7%	20.6%
Variant (b) (segfonc/select)									
mBIC	120	8	27.2	29	3268	2090	270	8.3%	20.7%
Lav	115	0	7.8	28	940	411	116	12.3%	20.8%
BM1	100	0	2.8	13	334	46	68	20.4%	23.4%
BM2	107	0	3.7	24	439	76	81	18.5%	22.1%
Variant (c) (segonly)									
mBIC	120	9	28.1	29	3367	1255	361	10.7%	16.4%
Lav	113	0	2.9	16	350	28	64	18.3%	19.6%
BM1	90	0	2.2	12	269	8	53	19.7%	20.2%
BM2	102	0	3.5	17	414	24	68	16.4%	17.4%
Variant (d) (seghomofonc)									
mBIC	116	0	19.0	29	2283	1637	178	7.8%	24.1%
Lav	114	0	3.5	26	415	148	56	13.5%	20.4%
BM1	92	0	2.4	19	287	40	61	21.3%	24.1%
BM2	101	0	3.2	19	387	82	68	17.6%	21.7%

In variant (c), the result for mBIC is slightly worse (more detections) but with fewer outliers. For the three other criteria the number of detections decreases significantly. The latter behaviour was actually not expected. Our interpretation is that when the periodic bias is not modelled, the segmentation algorithm has two options: (i) either put additional change-points to better fit the periodic variations in the signal, but this would lead to many more change-points (4 per year, i.e. a total of 64 per station for a 16-year time series) and cost too much to the selection criteria because of the penalty term, (ii) or select only those change-points with a large amplitude that are not confounded with the periodic bias. The observed result (Figure 5.2(c) and Table 5.1) suggest that BM1, BM2, and Lav prefer the second, more conservative option. As a consequence fewer change-points are detected and these may still include some outliers. Stated in other words, variant (a) including the periodic bias is actually capable of detecting smaller offsets, which makes it more efficient for the homogenization purpose. Note that with variant (c),

	before screening			after screening		
criteria	detections	validations	% validation	detections	validations	% validation
mBIC	3251	264	8.1	1270	146	11.50
Lav	474	75	15.8	341	67	19.65
BM1	335	70	20.9	292	68	23.29
BM2	435	77	17.7	370	74	20.00

Table 5.2

	Nsta	min	mean	max	detect	valid	valid
mBIC	120	9	27.1	29	3251	267	8.2%
Lav	114	0	4.0	28	474	75	15.8%
BM1	98	0	2.8	14	335	70	20.9%
BM2	107	0	3.6	18	435	77	17.7%

Table 5.3

the situation described by option (i) occurs nevertheless in some cases as will be illustrated in the next sub-section, and though the number of outliers and validations both decrease for BM1, BM2, and Lav, the percentage of validations remains nearly the same (Table 5.1). So, variant (a) clearly works better than variant (c) in the sense it detects more change-points but with the drawback of more outliers. This point is further discussed in the last section.

In variant (d) the variance is assumed to be constant. This has two consequences: (i) the function is fitted with uniform weights which in general leads to an estimated function $\hat{f}$ and an estimated mean $\hat{\mu}$ of different shapes, (ii) the estimated variance is larger than the mean variance of the variant (a) (the average mean standard deviations amount to 1.19 vs. $0.84 kgm^{-2}$, respectively) and fewer change-points are detected. Table 5.1 confirms that with this method fewer change-points are detected than with variant (a), however the number of outliers is increased (except for mBIC which is again a special case). The number of validations is also decreased, but the percentage of validations is almost unchanged.

The comparison of our four model variants showed that the complete model with heterogeneous variance and a full functional model for the periodic bias has the best properties (reasonable number of detections, small number of outliers, and high rate of validations). Among the four model selection criteria, BM1 and BM2 behave better than Lav and mBIC, with a small advantage for BM1 (higher validation rate). The last row of 5.1 not discussed so far computes the rate of validation slightly differently by excluding the outliers that are not validated from the total number of change-points. The assumption

here is that we can achieve a proper screening by removing all unnecessary (here invalidated) outliers. The numbers rise as the denominator of the ratio becomes smaller. The conclusions are unchanged as BM1 gets still be best score and reaches 23-24%. This rather optimistic outlier-corrected validation rate will be revised in Section 5.3. Figure 5.3(a) shows that the yearly-mean standard deviation of the noise ranges between 0 and 2 kgm^{-2}, with a mean value over the 120 stations of $0.84kgm^{-2}$. The seasonal excursion is of $0.63kgm^{-2}$ on average, which reflects the importance of modelling the heterogeneous variance. Figure 5.3(b) presents a measure of the magnitude of the periodic bias for BM1. With an average value of 0.33 kgm^{-2} it is clear that the periodic bias is not negligible and modelling it improves the segmentation results as shown by comparing the results of variant (d) and (a). Figure 5.3(c) shows that the distribution of offsets (changes in mean) is nearly symmetrical. The mean absolute value of $1.27kgm^{-2}$ is relatively large. The dip centred on zero reflects the fact that the smaller offsets are more difficult to detect. The most frequently detected offsets are found around +/- $0.5kgm^{-2}$. The larger offsets (up to +/- $10kgm^{-2}$) are outliers. The distribution of signal-to-noise ratio (SNR, computed as the absolute value of offset divided by standard deviation of noise) is peaking at 0.6 and the larger values (up to 10) correspond again to outliers. The mean SNR of 1.55 reflects the fact that offsets have in general an amplitude comparable to the noise.

5.2.2 Examples of special cases

We compared the results of the different variants station by station. For most stations, the results were identical whatever the method, however some special cases are worth mentioning. Here we will only discuss the results obtained with criterion BM1. With variant (c) there are actually 66 stations which have the same number of detections than variant (a). Although in general the change-points are located at the same position in the time series, this is not always the case. For 18 stations variant (c) detects more change-points and in 36 cases, it detects fewer. Station POL2 is an example for the former situation and station STJO an example of the latter. DUBO is an example where the same number is detected but the change-points are not located at the same position. With variant (d), the number of stations with equal, more, and fewer number of detections is: 57, 24, and 39, respectively. Examples are: EBRE, MCM4, and POL2.

The results for a selection of four stations are given in Figure 5.4:

- In the case of POL2, variants (a), (c) and (d) detect 3, 12, and 1 change-point(s), respectively. The signal shows a strong periodic variation which is well fitted by the models of variant (a) and (d) but

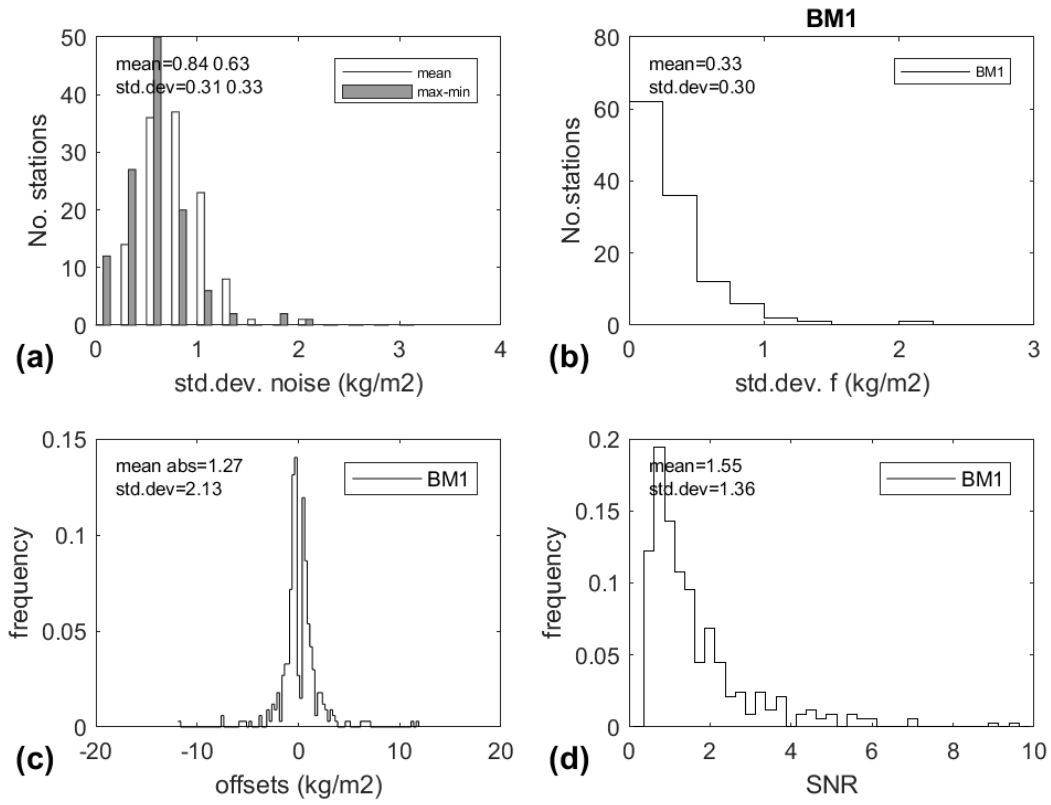


Figure 5.3 – Histograms of segmentation results for the final method with selection criterion BM1: (a) Number of stations with respect to the estimated standard deviation of the noise (mean and max-min of the 12 monthly values); (b) Number of stations with respect to the standard deviation of the estimated function; (c) Distribution of offsets of detected change-points; (d) Distribution of SNR of detected change-points..

is erroneously captured by the segmentation in variant (c). Variant (a) has one validated change-point (detected date: 2008-02-23, known change: 2008-03-06, type of change: P). Variant (c) has no validation, although it detects 12 change-points. Variant (d) detects only one change-point, which is located 72 days from the nearest known change-point and is thus not validated, but it coincides with one of the three detections found by variant (a). The detection of this change-point is made difficult because it is located in a month with heavy noise.

- In the case of STJO, variants (a) and (d) detect 5 and 4 change-points, respectively, with one outlier each but not at the same position. Among the detected change-points, one is exactly the same (detected: 2003-04-18, known: 2003-06-08, type: R) but is not validated, and one is close (detected by variant (a): 1999-07-20, by variant (d): 1999-07-19, known: 1999-07-29, type: R)

and is validated. Variant (c) gives no detection, the conservative option (ii), discussed above at page 104, is selected by BM1.

- In the case of DUBO, variants (a) and (c) detect two change-points at almost the same position but not exactly. Both are located close to known changes but only one is validated for variant (a) (detected: 1999-05-07, known: 1999-05-26, type: R). The second one is located 34 days from a known change for variant (a) and 148 days for variant (c). Though variant (c) works not bad, it is not as accurate as variant (a) because the periodic bias is neglected. Variant (d) has 4 detections which actually consist in 2 change-points, each being associated with an outlier. Although the periodic bias is modelled here, both change-points are quite badly located and thus not validated.
- Finally for MCM4 the signal has very marked inhomogeneities in the form of several abrupt changes but also non-stationary oscillations. The abrupt changes are well captured by variant (a) who detects 5 change-points among which 4 are validated (types are in chronological order: R, R, P, P). The non-stationary oscillations are only partly modelled by the periodic function. This is a special case where even the model used in variant (a) is not well adapted to such oscillations. This result advocates for an improvement of the functional basis. In that case, variant (c) works quite well too and leads to almost the same detections as variant (a), but only the two P changes are validated. Variant (d) on the other hand over-estimates the number of change-points to better fit the non-stationary oscillations but with detections of outliers. The four same change-points are validated as with variant (a) but the fitted means are quite different.

Among the 70 validated change-points found with variant (a) by BM1 (see Table 5.1), there are 53 R, 16 A, 7 D, and 13 P types (note that these numbers don't sum up to 70 because in many cases the changes involve several types). We find here that receiver changes are the most frequent explanation for inhomogeneities. This is not surprising since they are the most frequent change-type occurring at GNSS stations (among the 1731 known changes, 1142 are of R type, 389 A, 70 D, and 425 P). However, this is in contrast with Ning *et al.* [2016]'s results who did not consider receiver changes at all. Most of the receiver changes documented in the IGS sitelogs actually refer to operating software updates which don't have much impact on the observations as long as they don't involve a change in the minimum elevation cutoff angle. Hardware changes on the other hand are more prone to have an impact. We performed a quality control based on the observation files with TEQC software Estey & Meertens [1999] and found that in many cases hardware updates lead to changes in the multipath diagnostic parameters and in some occasions in the percentage of observations. Examples are the receiver changes at STJO on 1999-08-06 (from ROGUE_SNR_8000 to AOA_SNR_12_ACT) and at MCM4 on

2002-01-03 (from ROGUE_SNR_8000 to AOA_SNR_12_ACT) and 2006-05-19 (from AOA_SNR_12_ACT to ASHTECH_ZXII3), discussed above. At MCM4, strong oscillations are visible in the multipath diagnostics (mp1 and mp2) during the AOA_SNR_12_ACT period, similar to those seen in the IWV differences (Figure 5.4). This reveals a malfunctioning of the GNSS equipment during that period also associated with a jump in the mean signal at the beginning and at the end of the period.

5.2.3 Comparison with Ning *et al.* [2016]

Similar to this study, Ning *et al.* [2016] analyzed the homogeneity of GNSS-ERAI IWV differences for a global network of 101 GNSS sites with a least 15 years of observations. Their series were used with monthly sampling whereas here we used daily sampling. They used the PMTred test Wang [2008a] to detect abrupt changes in the mean IWV difference but this model does not include a periodic bias. They detected a total of 62 change-points, at 47 stations, among which 45 detections were attributed to the GNSS series, 16 to ERAI, and 1 was undetermined. Their attribution method was based on the comparison of the GNSS candidate series to two or three references series (ERAI, another nearby GNSS series, and/or a nearby VLBI series). Consistency between the two or three detected offsets was used to attribute the change-points to GNSS and disagreement to ERAI (by default). They also used the GNSS metadata and validated 13 detections, but they included only antenna, radome, and known microwave absorbing material changes. Their validation window was +/- 6-month wide, i.e. much larger than our +/- 30-day window. We reanalyzed their results based on the metadata we had for 42 IGS sites of their GNSS network. We found that 10 out of 12 unvalidated GNSS detections could actually be explained with receiver changes and 2 with receiver+antenna changes using the same 6-month window. Six of the detections agreed with receiver changes within less than 2 months. Regarding the change-points they attributed to ERAI, we found that 5 out of 15 coincide with GNSS changes (2 receiver changes and 3 antenna changes). However, inspection of the GNSS-ERAI IWV difference time series suggests that in many cases these detections might be due to outliers and gaps in the time series. This suggests that the PMTred test which they used is quite sensitive to fluctuations in the noise similar to variant (d) discussed in the previous sub-section.

The comparison of our results for variant (a) with Ning *et al.* [2016]'s results for 31 common stations which have change-points leads to the following conclusions: (i) our method detects nearly twice more change-points than PMTred (107 vs. 43), (ii) among 32 PMTred detections attributed to GNSS, about 1/3rd coincide with ours within +/- 2 months, 1/3rd within 2-6 months and 1/3rd within more than 6 months, (iii) among 11 PMTred detections attributed to ERAI, 4 change-points coincide with ours within +/- 1 month (the others being about 6 months or more apart) and none of them can be explained

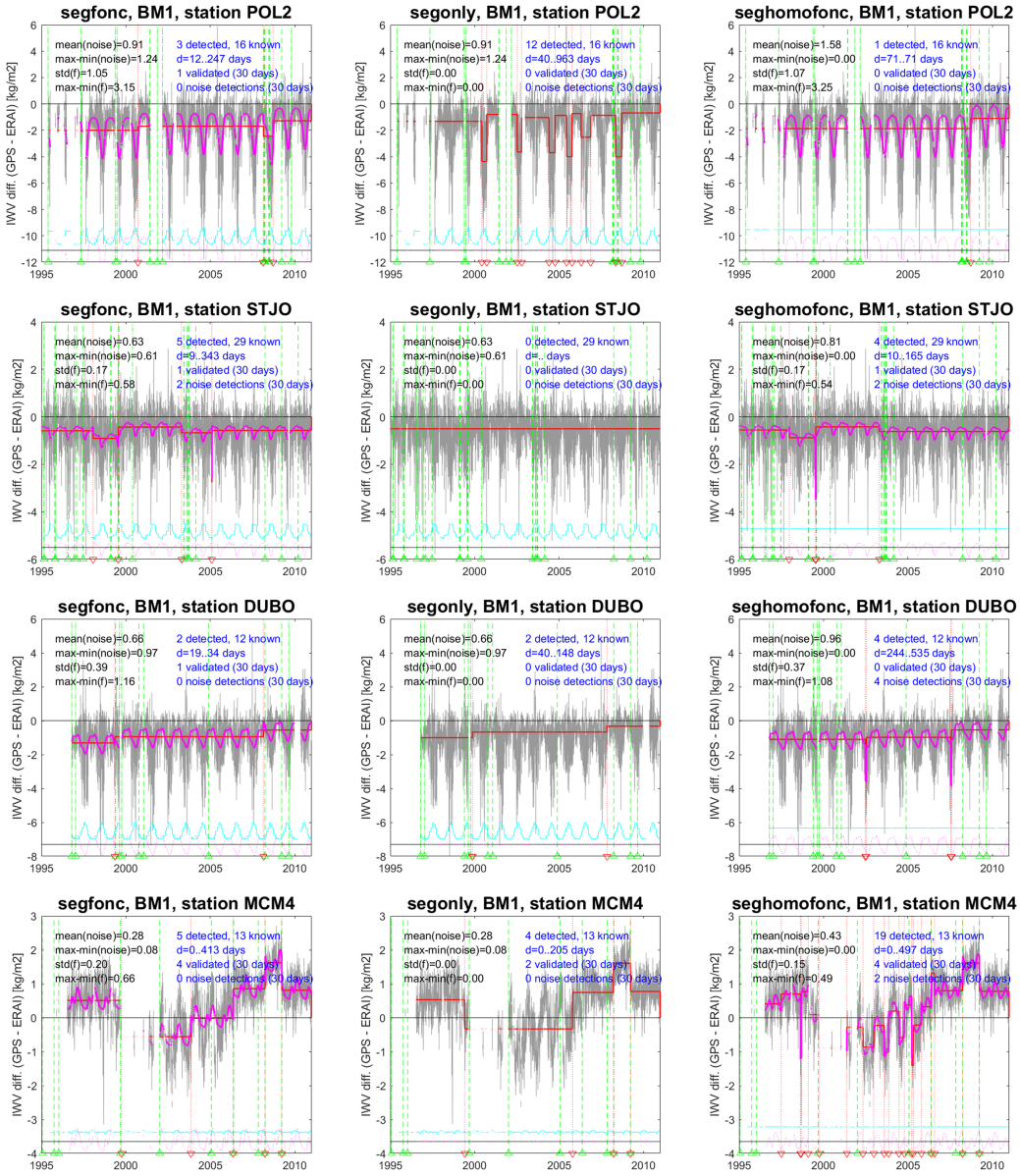


Figure 5.4 – Examples of results obtained with variants (a), (c), and (d) from left to right, for four different stations: POL2, STJO, DUBO, and MCM4 (from top to bottom). The content of the plots is similar to Fig. 2.6(b). The text inserted at the top left of the plots reports the mean standard deviation of the noise, the variation (max-min) of the standard deviation of the noise, the standard deviation of the periodic bias function, and the variation (max-min) of the periodic bias function. The text in blue reports the total number of detections and of known changes, the minimum and maximum distance between detected change-points and the nearest known changes, the number of validated detections, and the number of noise detections.

by GNSS metadata, even when including receiver changes. Inspection of the IWV differences and the TEQC diagnostics led us also to the conclusion that 4 change-points attributed to ERAI may indeed be due to inhomogeneities in ERAI. The stations and dates involved are: GODE (1998-08-06), HOB2 (2006-06-10), and WUHN (1999-02-14 and 2006-09-27). For station WUHN, Parracho *et al.* [2018] also detected an anomaly in the ERAI IWV series around 2006-09 which coincides with a change in radiosonde type from the station at the city of Wuhan, China. Since the radiosonde data are assimilated into ERAI, any abrupt change in these data may be transferred to the reanalysis.

5.3 Screening outliers

In this Section we will re-examine the outlier detection method introduced in the previous sub-section. Formally, let be t_i and t_{i+1} the positions of two consecutive change-points. If $t_{i+1} - t_i < threshold$, then these change-points t_i and t_{i+1} are called "outliers" and they form a "cluster" of two outliers. Outliers are most of the time due to spikes in the noise (see the example of STJO in Figure 5.4 in the case of segfonc model, the 4th and 5th change-points). However, a couple of outliers may still be associated with a "true" change-point (STJO in Figure 5.4 in the case of seghomofonc model, 3rd and 4th change-points). A "true" change-point can be characterized by a significant change in the mean before and after the cluster of outliers. Figure 5.5 explains further the three possible situations that may arise with two change-points:

- class 1, called 'outliers alone' (example cases (a), (b), (c)),
- class 2, called 'outliers and change-point' (cases (d), (e), (f)),
- class 3, called 'change-point only' (cases (g), (h)).

Class 1 and 2 correspond to the detection of a cluster of two outliers ($t_{i+1} - t_i < threshold$). In the cases of class 1, the variation of the means before and after the cluster is not significant (according to a statistical test), whereas in class 2 it is significant. The aim of the screening is thus to eliminate both change-points (outliers) of class 1, and to replace the couple of outliers of class 2 by a single change-point (schematized by the mid-point on Figure 5.5). Class 3 is the regular situation when the distance between the two change-points is larger than the "outlier-detection" threshold: $t_{i+1} - t_i \geq threshold$. There can be more than two consecutive change-points which are pairwise closer than the outlier-detection threshold. All these change-points will be grouped in one cluster and the test on the means will be performed between the mean before and after the cluster. The screening may in that case replace or remove more than two change-points.

In Section 5.3.1, we will re-examine the choice of the outlier-detection threshold (fixed to 30 in the previous sub-section). Therefore we will analyse the distribution of segment lengths and search for a class of small segments that may be identified as the outliers according to our definition. Then we will describe the test for equal means used to distinguish between class 1 and 2. The outlier detection method will be applied on the segmentation results obtained with the new segmentation method. Section 5.3.2 presents the results. In Section 5.3.3 we discuss an alternative approach to eliminate the short segments with the help of the minimal length parameter (denoted lmin) introduced in Section 4.5.

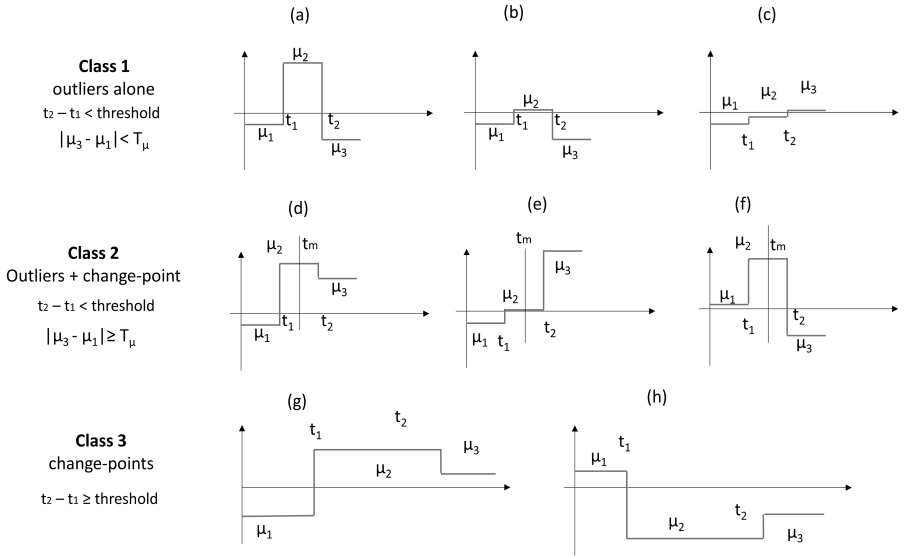


Figure 5.5 – The different observed configurations of the detected change-points.

5.3.1 Threshold and proposed test

Analysis of segment length. Figure 5.6 presents the histogram of the segment lengths obtained with the model selection criterion BM1 for all the stations. The distribution is roughly exponential: small segment lengths are the most frequent and long segments the least, except a small peak around 5000 days which corresponds to stations with no change-point. Figure 5.7 is a zoomed version for lengths smaller than 240 days. We see that there is a peak for the bin $[10, 20]$ followed by a dip for the bin $[20, 30]$. The initial choice of an outlier-detection threshold of 30 days was not bad but a more extended dip is seen between 50 and 100. A better choice may thus be in this interval instead. For choosing in a more objective way this outlier-detection threshold, we will use a mixture model below.

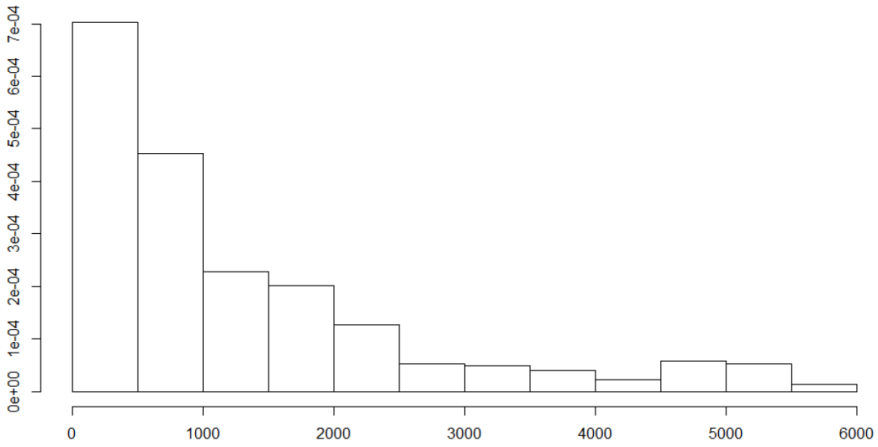


Figure 5.6 – Histograms of the segment lengths, for BM1 with a bin size of 500 days.

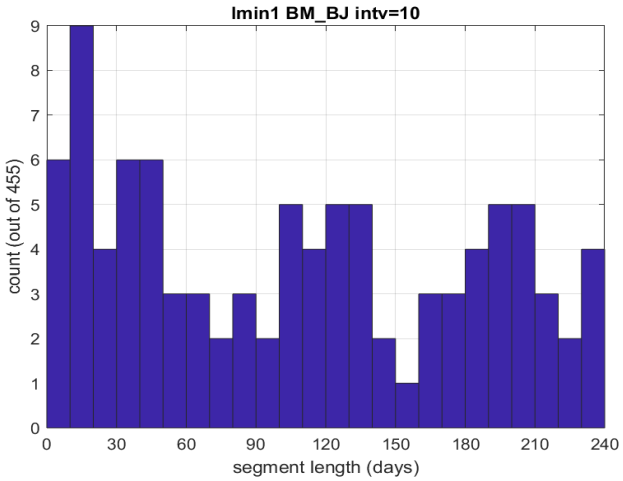


Figure 5.7 – Histograms of the segment lengths, for BM1, zoomed for lengths from 1 to 240, with a bin size of 10 days.

Mixture models. We propose to classify the lengths of the segments using a Gaussian mixture model with common variance (Biernacki *et al.* [2000]; McLachlan & Peel [2004]; Titterington *et al.* [1985]). We use the classical Integrated Completed Likelihood (ICL) criterion proposed by Biernacki *et al.* [2000] to choose the number of groups. Indeed this criterion is well suitable compared to the BIC one in this clustering objective context. Figure 5.6 showed that the distribution of the segment lengths is rather exponential. To apply a Gaussian mixture model we changed the scale to the logarithm one, see Figure 5.8. The ICL criterion selects 2 groups. Figure 5.9 shows the estimated density for each of them. The boundary between the two groups is at 81. This results is consistent with the subjective analysis done

in the previous paragraph, where we noted a dip in the distribution of segment lengths between 50 and 100.

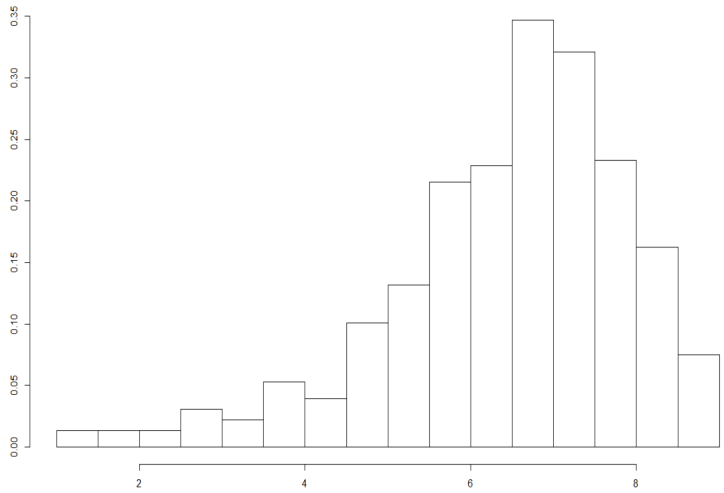


Figure 5.8 – The logarithmic distribution of the length of the segments.

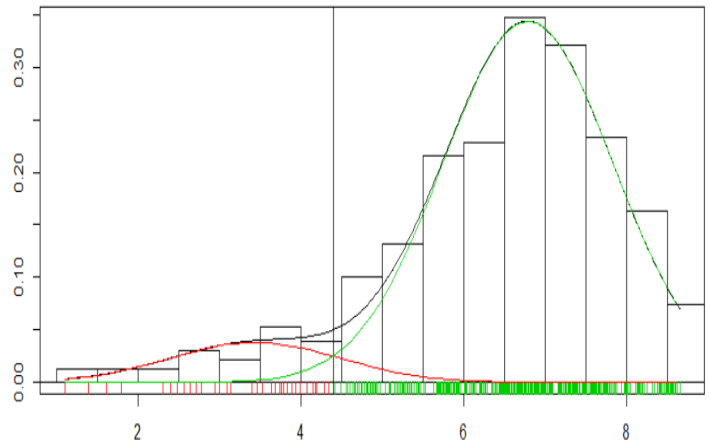


Figure 5.9 – The density of all the data (in black) and the density of each of the two groups (in red for the first and in green for the second). The black vertical line indicates the boundary between the two groups.

Test Once the outlier-threshold is chosen (once the outliers are isolated from the "true" change-points, i.e. classes 1 and 2 from the class 3), we have to decide if the associated clusters are in class 1 or in class 2. As explained in the introduction of this Section, the distinction is based on the difference between the means before and after the cluster. To test if the difference is significant (class 2) or not (class 1), we

build a test of equality of means. According to the model Eq. 4.1 of Chapter 4, $\{Y_t\}_t$ is an independent Gaussian process with a mean equals to $\mu_k + f_t$ if t belongs to segment k and a variance that is monthly different. More formally, let be t_i and t_j the positions of the first and last outliers of the cluster, then the means we want to compare are μ_i and μ_{j+1}. Denote by I_i and I_{j+1} the associated segments. If we note $\tilde{Y}_t = Y_t - f_t$, we have that

$$\tilde{Y}_t \sim \mathcal{N}(\mu_i, {\sigma_t}^2) \quad \text{if } t \in I_i$$

where σ_t^2 is the variance at position t. Here we assume that the function f, the intervals I_i and the variances are known. In practice, we used their estimates (see Chapter 4). The test hypotheses are

$$\begin{cases} H_0 : & \mu_i - \mu_{j+1} = 0 \\ against & \\ H_1 : & \mu_i - \mu_{j+1} \neq 0 \end{cases}$$

Recall that the maximum likelihood estimators of the means μ_i and μ_{j+1} are $\widehat{\mu}_i = \frac{\sum_{t\in I_i} \frac{\tilde{Y}_t}{\sigma_t^2}}{\sum_{t\in I_i} \frac{1}{\sigma_t^2}}$ and $\widehat{\mu}_{j+1} = \frac{\sum_{t\in I_{j+1}} \frac{\tilde{Y}_t}{\sigma_t^2}}{\sum_{t\in I_{j+1}} \frac{1}{\sigma_t^2}}$ respectively (see Chapter 4). We thus consider the following test statistic

$$T = \frac{\widehat{\mu}_i - \widehat{\mu}_{j+1}}{\sqrt{\frac{1}{\sum_{t\in I_i} \frac{1}{\sigma_t^2}} + \frac{1}{\sum_{t\in I_{j+1}} \frac{1}{\sigma_t^2}}}},$$

and its law under H_0 is $\mathcal{N}(0,1)$. Indeed, we have that

$$\frac{\tilde{Y}_t}{{\sigma_t}^2} \sim \mathcal{N}\Big(\frac{\mu_i}{\sigma_t^2}, \frac{1}{\sigma_t^2}\Big) \implies \sum_{t\in I_i} \frac{\tilde{Y}_t}{\sigma_t^2} \sim \mathcal{N}\Big(\mu_i \sum_{t\in I_i} \frac{1}{\sigma_t^2}, \sum_{t\in I_i} \frac{1}{\sigma_t^2}\Big) \implies \widehat{\mu}_i \sim \mathcal{N}\Big(\mu_i, \frac{1}{\sum_{t\in I_i} \frac{1}{\sigma_t^2}}\Big)$$

We reject H_0 with a significance level of 5% if $|t_{\text{obs}}| > 1.96$ or equivalently if p-value$= 2\Phi(-|t_{\text{obs}}|) < 0.05$ where t_{obs} is the observed test statistic and Φ is the cumulative distribution function of $\mathcal{N}(0,1)$.

5.3.2 Outliers detection

In this Section, we applied the proposed screening method on the segmentation results of Section 5.2. For the outlier-detection threshold we considered the initial value of 30 days that we used in Section 5.2 and the optimal value of 80 days determined with the mixture model. We restrict the analysis to the segmentation results obtained with the BM1 selection criterion.

The total number of detected change-points with BM1 is 335. With a threshold of 30 days, 36 change-points are detected as outliers which are grouped in 18 clusters all of which are composed of 2

outliers only. Applying the test for equality in the means, 17 clusters are classified in class 2 and only 1 in class 1. With a threshold of 80 days, more outliers are detected: 70 outliers grouped in 34 clusters, 33 clusters of 2 outliers and 1 cluster of 4 outliers (station IISC shown below). Among them, 27 clusters are classified in class 2 and 7 in class 1. The screening eliminates the outliers belonging to class 1 and keeps the midpoint between the last and first outliers of the clusters of class 2. As a result, 315 change-points remain after the screening with the threshold of 30 days and 292 change-points with the threshold of 80 days.

Figures 5.10 and 5.11 show the case of station IISC for the threshold of 30 days and 80 days, respectively. With the threshold of 30 days, before the screening, the station has 12 change-points (dotted red lines) of which 6 outliers (marked with a red circle) are grouped in 3 clusters of 2 outliers each. The first two clusters are classified in class 2 while the third cluster is in class 1 (marked with an "x" on the bottom). The screening will thus keep the mid-point for the first two clusters and remove the two outliers for the third cluster. After the screening, 8 change-points remain. With the threshold of 80 days, there are again three clusters. The first two clusters are the same but the third cluster now also includes the two change-points that are located earlier in the time series (all four change-points of this cluster are in the year 2005). The change in mean before and after this cluster is significant (although they appear visually close). All three clusters are classified in class 2. After the screening, 7 change-points remain.

In order to judge the benefit of the screening we inspected the number and percentage of validations with respect to the metadata, before and after screening. The number of validations before the screening was 70. After the screening is decreased to 69 and 68 with the threshold of 30 days and 80 days, respectively. The lost validations were outliers which belonged all to class 2 (outlier + change-point). The reason why they were lost is because the initial dates were replaced with the midpoints and they were beyond the validation window of 30 days. More importantly, the percentage of validations is slightly increasing with both detection thresholds. It goes from 20.89% before the screening to 21.90% after the screening with the threshold of 30 days and 23.3% with the threshold of 80 days.
Results for the other criteria are presented in Tables 5.4 and 5.5 below for the threshold of 80 days (see the top parts of these Tables with lmin= 1). An improvement in the percentage of validations is observed for all four criteria.

5.3.3 Test of different minimum segment size

In this sub-section we discuss an alternative approach to eliminate the short segments with the help of the minimal length parameter (denoted lmin) introduced in Section 4.5. This parameter will force

the segmentation algorithm to separate successive change-points by at least "lmin" days. It is thus a mean to avoid small segment lengths. First we will investigate the impact of increasing lmin on the distribution of the detected segment lengths. Then we will analyse the impact of one specific station, IISC which was already analysed in Section 5.3.2. Finally, we will study the impact on the percentage of validation of various lmin values combined with the screening method described in the Section 5.3.2 for the different model selection criteria.

Impact of changing lmin on the distribution of segment lengths Figures 5.12 shows the histograms of the detected segment lengths for lmin values between 1 and 100. Two general features are evidenced. First, as lmin increases, the total number of segments increases (this number is reported on the y-axis of each plot). Second, the increase concerns mainly the small segment lengths, with a strong peak in the bin of the smallest segments (from lmin to lmin+10). A similar behaviour is found with the other criteria (BM2 and Lav, but not with mBIC which shows the opposite tendency but this criterion has in all cases too many detections and is not reliable in our application). Table 5.4 reports the results for all the criteria. We can suspect that this increasing is due to the fact that the segmentation need to compensate the forbidden detections (imposed by lmin).

This behaviour is actually not very satisfying if one expected to use only lmin to get rid of the outliers. Indeed, if we would use e.g. lmin= 80 to be consistent with the mixed model results discussed above, we would still have many small segments with lengths = 80..90. So a better solution would be to combine a value of lmin> 1 but not too big and the screening method. This option is discussed below. But first, we will analyse in more detail the impact of changing lmin on the segmentation results for one specific station.

Impact of changing lmin on the segmentation results for IISC Figure 5.13 shows the segmentation results for station IISC, criterion BM1, when lmin takes values between 1 and 100. The case lmin=1 was already discussed above. The segmentation detects 12 change-points among which 8 are outliers grouped in 3 clusters (the outlier detection threshold is 80 days). The first cluster has two outliers at dates 1997-10-21 and 1997-11-18 and its segment length is 23 (not counting the days with no observations). The second cluster has starting on 2004-05-12 also two outliers which are 5 days apart. The last cluster includes 4 outliers located between 2005-04-04 and 2005-08-29. It is composed of two pairs of outliers detected on two noise peaks. When lmin is increased to a value of 10, the second cluster disappears but there is still one change-point which is associated to a visible change in mean on

2004-05-02. The first cluster on the other hand is present and the dates are unchanged, and the two change-points disappeared from the third cluster while the remaining two are still outliers are located at the same dates. When lmin is further increased, the change-points which were in the first and third clusters with lmin= 1 are still detected with increasing segment lengths. In the case of the first cluster, the two change-points remain close and are detected as outliers until lmin= 60 (outliers are not detected for lmin= 80 or 100 because the outlier-threshold is 80 days). In the case of the third cluster the behavior is more complicated with two peaks for lmin= 20, 50 and 60, and even more for lmin= 80 and 100. These peaks seem to catch the quasi-periodic behavior of the signal which cannot be fitted by the purely period function. This supposition is reinforced by the decreasing of the amplitude of the estimated function with respect to the increasing of lmin: 2.93 for lmin= 10, 2.84 for lmin= 60 and 2.23 for lmin= 80.

For the purpose of the screening, one could conclude that the segmentation solution for lmin= 10 is good enough: it allowed to remove one cluster with a small segment clearly associated with an isolated noise peak. The other clusters are more difficult to remove because they are associated with a burst of noise of some width. In terms of number of change-points, the results for station IISC follow the general tendency which is an increase when lmin is increased, with some fluctuation however. Especially, there is one situation (lmin= 40) where the number of change-points is zero. Inspection of other stations reveals that it is not uncommon that for certain values of lmin the number of change-points can drop to zero. However, for many stations the number is steadily increasing, which means that new change-points appears, which are in some cases outliers (separated by less than 80 days). The case of no detected change-points can be explained by the marginal behavior of the BM1's heuristic. Indeed, this heuristic aims at calibrating the penalty constant based on the biggest jump of dimension with respect to this constant. In some cases, the higher jump is reached by several values of the constant. By default, the lower dimension is considered and can often lead to zero detection whereas the others jumps allow some detections. This is the case of the station IISC with lmin=40: with the first jump we have no detection whereas with the second one, 10 change-points are detected.

Performance of combining lmin and the screening method Here we evaluate the performance of the screening method combined with lmin. Table 5.4 and Table 5.5 summarize the results before and after the screening. Among the different criteria, the highest percentage of validation is always obtained with BM1, independently of lmin, both before and after the screening. The overall best values for BM1 are found for lmin=10, which are 21.99% and 23.59%, before and after the screening, respectively. The screening improves the results for all lmin values and all criteria. In Table 5.5 are presented the same results as Table 5.4 after the screening. As we have seen before in 5.3.1, the result for

lmin= 1 is improved after the screening, passing by a percentage of validation, the best in both cases for BM1 of 20.90% to 23.29%. For larger lmin, the best without the screening was lmin= 10 and criterion BM1 with a percentage of validation of 21.99%, that was still less good than the results for lmin= 1 and BM1 after the screening. After applying the screening, the best is for the lmin= 10 and BM1 with a percentage of validation of 23.59%. This is also, in absolute the best score obtained. In general, better scores are obtained after having applied the screening (in particular, for the criteria Lav and BM2 with lmin= 1). This comparison proves that actually the screening proposed in 5.3.1 improves the results of the segmentation, eliminating the outliers in the detected change-points.

We have seen that increasing lmin also increases the number of outliers, after the screening the percentage of validation improved with respect at the results without applying the screening. Another criterion for an indication of whether by increasing lmin we have a better segmentation, in general, is to look at the percentage of clusters validated after screening, in Table 5.5. In general, the percentage of clusters considered significant is higher for BM2 and lmin = 1. Until lmin= 40, BM1 and BM2 percentage are higher than 75%. For lav and mBIC is generally always lower than other criteria. This percentage indicates the quantity of the change-points detected for each lmin that are kept as "good". In this case lmin= 1 seems to show the best segmentation.

5.4 Attribution

The goal of the attribution is to determine the cause of the change-points detected by the segmentation method. Since the method is operating on IWV differences, there is ambiguity whether the cause is a change in the mean of the target series (GNSS) or the reference series (ERAI in our case). The most widely used method is to compare the detected change-points to metadata (Venema *et al.* [2012]). In Section 5.2 we compared our segmentation results to the IGS metadata available for our GNSS reprocessed data set and found a validation rate about 21% for BM1. After the outlier screening, this number slightly increased to 23%. The remaining change-points can be due to undocumented GNSS equipment changes and changes in the station's environment, inhomogeneity in the ERAI reference series, and/or false detections due to noise spikes not detected as outliers and undmodelled effects (e.g. autocorrelation in the difference series). We assume that the first two explanations are the most likely.

One way to further disentangle the GNSS and ERAI causes is to test the significance of every detected change-point by comparing the mean of IWV differences over the homogeneous parts of the series between the target GNSS series and another reference series different from ERAI. Past studies used dual GNSS comparisons, as well as comparisons involving DORIS and VLBI data (Bock *et al.* [2010]; Ning *et al.*

	criteria	detections	validations	% validation	outliers	cluster	class 2	% class 2
lmin1	mBIC	3251	264	8.1	2714	1027	733	71.37
	Lav	474	75	15.8	194	85	61	71.76
	BM1	335	70	20.90	70	34	27	79.41
	BM2	435	77	17.7	113	55	48	87.27
lmin10	mBIC	3056	276	9.03	2432	936	637	68.05
	Lav	530	84	15.85	231	101	62	61.38
	BM1	341	75	21.99	64	32	24	75
	BM2	491	83	16.90	128	63	50	79.36
lmin20	mBIC	2883	296	10.27	2158	838	595	71
	Lav	556	87	15.65	234	101	70	69.30
	BM1	392	72	18.37	73	36	27	75
	BM2	570	91	15.96	171	82	64	78.04
lmin30	mBIC	2689	306	11.38	1857	744	511	68.68
	Lav	783	120	15.33	379	159	100	62.89
	BM1	439	82	18.68	92	45	36	80
	BM2	676	105	15.53	209	99	83	83.83
lmin40	mBIC	2541	318	12.51	1592	665	470	70.67
	Lav	917	127	13.85	426	186	127	68.27
	BM1	453	85	18.76	93	45	32	71.11
	BM2	727	109	14.99	222	109	82	75.22
lmin50	mBIC	2376	302	12.71	1259	544	382	70.22
	Lav	1145	150	13.10	540	233	160	68.66
	BM1	512	92	17.97	96	47	31	65.95
	BM2	774	110	14.21	205	100	75	75
lmin60	mBIC	2262	288	12.73	969	431	299	69.37
	Lav	1225	165	13.47	430	194	118	60.82
	BM1	555	92	16.58	79	39	24	61.53
	BM2	901	132	14.65	226	108	69	63.88
lmin80	mBIC	2077	261	12.56	0	0	0	0
	Lav	1453	194	13.35	0	0	0	0
	BM1	614	106	17.26	0	0	0	0
	BM2	911	136	14.92	0	0	0	0
lmin100	mBIC	1873	232	12.38	0	0	0	0
	Lav	1627	195	11.98	0	0	0	0
	BM1	688	115	16.71	0	0	0	0
	BM2	1158	169	14.59	0	0	0	0

Table 5.4 – Comparison of segmentation results for different values of the minimum segment length (lmin) for the outlier-threshold 80 before the screening. From left to right: number of detected change-points, total number of validations, percentage of validations, total number of outliers, number of clusters, number of clusters in class 2 and the percentage of cluster in class 2.

	criteria	detections	validations	% validation
lmin1	mBIC	1270	146	11.50
	Lav	341	67	19.65
	BM1	292	68	23.29
	BM2	370	74	20.00
lmin10	mBIC	1261	155	12.29
	Lav	361	70	19.39
	BM1	301	71	23.59
	BM2	413	77	18.64
lmin20	mBIC	1320	175	13.18
	Lav	392	74	18.87
	BM1	346	70	20.23
	BM2	463	84	18.14
lmin30	mBIC	1343	174	12.81
	Lav	504	88	17.26
	BM1	383	79	20.37
	BM2	550	97	17.45
lmin40	mBIC	1419	197	13.81
	Lav	618	102	16.50
	BM1	392	77	19.64
	BM2	587	95	16.18
lmin50	mBIC	1499	206	13.68
	Lav	765	111	14.51
	BM1	447	82	18.34
	BM2	644	100	15.53
lmin60		1592	225	14.07
		913	139	15.22
		500	86	17.20
		744	119	15.99

Table 5.5 – Comparison of segmentation results after screening for different values of the minimum segment length (lmin) for the outlier-threshold 80. From left to right: number of detected change-points, total number of validations, percentage of validations.

[2016]). Other reanalyses such as MERRA-2 might also be useful (Parracho *et al.* [2018]). Using this approach, the detected change-points can be confirmed (if the difference in mean is significant) or not. If it is confirmed by one or several additional comparisons, the cause is attributed to GNSS. If not, the ERAI reference series might be tested as well to attribute possibly the cause to ERAI. A limitation with this method is that, not having a large availability of GNSS stations or other observing techniques in our global network, we cannot carry out an exhaustive check of all target stations. However, this approach might be useful in dense regional or national networks.

5.5 Trends estimation

The homogenisation method developed in Chapter 4 combined with the treatment of outliers proposed above allow to obtain the abrupt changes appearing in the GNSS IWV series. This information can then be used to estimate correctly the climate trend in the series. To this aim, one solution would be to correct the raw series IWV from these change-points and then estimate the trend on the corrected series. However, as we have observed on the particular station CCJM (see Figure 2.6 (a) in the introduction), the IWV series exhibits also a marked seasonal variation. We thus proposed a global model on IWV in which are integrated, in addition to the trend, a piece-wise constant mean on fixed intervals (the intervals obtained by the segmentation) and a function h that aims to model the seasonal variation. After presenting this model, we show a preliminary analysis on a particular series.

Model and estimation Denote by Y_t the signal IWV at position t, the model is the following:

$$Y_t = m_k + a \cdot x_t + h_t + \mathcal{E}_t, \qquad t \in \hat{I}_k = [\![\hat{t}_{k-1} + 1, \hat{t}_k]\!] \text{ with } k \in [\![1, \hat{K}]\!], \tag{5.1}$$

where the vector of errors $\mathcal{E} = \{\mathcal{E}_t\}$ are centered and $\hat{t}_k$ are the segments obtained by the proposed procedure (combining the segmentation method with the screening) on the IWV differences (using the ERAI reference). As classically, we estimate the parameters using the least-square method. For the estimation of the function h, we follow the same approach as for the estimation of f in 4, i.e. using a Fourier series of order 4:

$$h_t = \sum_{i=1}^{4} a_i \cos(w_i t) + b_i \sin(w_i t),$$

where $w_i = 2\pi \frac{i}{L}$ is the angular frequency of period L/i and L is the mean length of the year ($L = 365.25$ days when time t is expressed in days).

Let consider the matricial formulation of this model:

$$\boldsymbol{Y} = \mathbf{X}\beta + \mathbf{E}, \tag{5.2}$$

where $\boldsymbol{Y} = (Y_1\ Y_2 \dots\ Y_n)^T$ is a n-dimensional column vector, $\mathbf{X}$ the design matrix with dimension $[n \times (\hat{K} + 1 + 2 \times 4)]$ ($\hat{K}$ for the segments, 1 for the trend and 2×4 for the coefficients of the Fourier decomposition), β the $\hat{K} + 9$-dimensional column vector of the unknown coefficient and $\mathbf{E}$ the column vector of errors with size n. We estimate these parameters using the classical least-square criterion:

- if the errors are assumed to be independent, i.e. $Cov(E) = \sigma_2 \mathbf{I}_n$ where $\mathbf{I}_n$ is the identity matrix with size n, the adapted criterion is the classical Ordinary Least-Squares criterion (OLS) leading to the classical OLS estimator:

$$\widehat{\beta}^{\text{OLS}} = Arg \min_{\beta} \|\boldsymbol{Y} - \mathbf{X}\beta\|^2 = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{Y}, \tag{5.3}$$

- if the errors are assumed to be correlated, i.e. $Cov(E) = \Omega$ that is not diagonal, the adapted criterion is the classical Generalized Least-Squares criterion (GLS) leading to the classical GLS estimator:

$$\widehat{\beta}^{\text{GLS}} = Arg \min_{\beta} \|\boldsymbol{Y} - \mathbf{X}\beta\|^2_{\Omega^{-1}} = (\mathbf{X}^T\Omega^{-1}\mathbf{X})^{-1}\mathbf{X}^T\Omega^{-1}\boldsymbol{Y}. \tag{5.4}$$

  In practice, Ω is typically unknown. In this case, an estimator of Ω denoted $\widehat{\Omega}$ is substituted in (5.4). Ω is classically estimated using the OLS estimated residuals $\widehat{\mathcal{E}}_t = Y_t - \mathbf{X}\widehat{\beta}^{\text{OLS}}_t$.

Preliminary analysis on a case study Let us first examine the OLS results for one particular station (ALIC). This station has 5 change-points after the screening. Figure 5.14 shows the GNSS IWV time series, the fitted model, and the residuals. The estimated trend value is $\hat{a} = 0.790\ kgm^{-2}yr^{-1}$ and its standard error is $\sigma_a = 0.112 kgm^{-2}yr^{-1}$. The trend value is remarkably large and significant with respect to the standard error. The trend value without change-points estimated is $\hat{a} = -0.089\ kgm^{-2}yr^{-1}$ which is indeed much smaller. The lower plot shows that the centred time series of the piece-wise constant function representing the mean values and the trend go in opposite directions. There is confusion between the trend and the mean parameters. This problem is actually reflected in the covariance matrix of the parameters as strong correlation between parameters m_k and a ($r = 0.83$ with between m_1 and a and $r = -0.92$ between m_6 and a). The standard deviation of the parameters is also increased ($\sigma_{m_1} = 1.04 kgm^{-2}$ and $\sigma_{m_6} = 0.57 kgm^{-2}$) compared to the case when no change-points are estimated ($\sigma_{m_1} = 0.12 kgm^{-2}$ and $\sigma_a = 0.030 kgm^{-2}yr^{-1}$). Including change-points increases significantly the error in the parameters. The comparison of the trends estimated given by both OLS solutions (with

and without change-points included) shows a huge difference in the values: they are of opposite signs and differ in magnitude by factor of ~ 8.9. Although the trend estimated without the change-points is not reliable (because we know the time series is inhomogeneous), it is still more reasonable compared to published trends. For example, Parracho *et al.* [2018] found trends for the same station of -0.117, -0.070, and -0.081 $kgm^{-2}yr^{-1}$ for monthly IWV data from GNSS (inhomogeneous), ERA-Interim, and MERRA-2 reanalysis, respectively. It should be noted that Parracho *et al.* [2018] computed trend using the Theil-Sen method (Sen [1968]; Theil [1992]) which is known to be more robust than OLS but cannot include the change-points.

Another feature seen in Figure 5.14 (lower plot) is the temporal variation of the residual's magnitude. It is quite common that IWV variability is larger in the summer period than in the winter period (Bock & Parracho [2019]). This feature would suggest to use a model for the variance of the errors that is time-dependent, i.e. $Cov(E) = \Omega$ would be diagonal but with different diagonal elements. The OLS solution is in that case not optimal and the GLS solution is required (which in that case is called the weighted least-squares, WLS, solution). Further insight into the residual's properties is given by 5.15 which shows that there is also strong serial correlation in the time series. This is actually expected since atmospheric variability can be quite large and our model only accounts for a smooth seasonal variation modelled by a fourth order Fourier Series.

To account for these features we tested both the WLS and GLS solution. In the WLS we determined the error variance by computing the empirical variance on a moving window of size 30 days on the residuals from the first OLS. In the GLS we specified an autoregressive model of order 1 (AR(1)) with a lag-1 correlation coefficient of 0.7858 (estimated from the residuals from the first OLS). The estimated parameters (in the case 5 change-points are included) changed only slightly, but the standard errors changed significantly. In the case of WLS, the errors decreased. In the case of the GLS they increased. Because the autocorrelation in the data is real, the WLS error estimates are biased as the OLS estimates were. The GLS estimates with the AR(1) error model is thought to be more realistic, although it does not account for the heteroscedasticity of the errors. The trend estimate with the GLS method was $\hat{a} = 0.820$ $kgm^{-2}yr^{-1}$ with a standard error of $\sigma_a = 0.307 kgm^{-2}yr^{-1}$. The error on the trend increased by a factor of 2.72 (actually the error increased on all parameters by almost the same amount). Comparing the trend and its error, it can still be concluded that the trend is significant at the level of 0.05.

If the noise structure is more complex than an AR(1), it can be difficult to identify and to represent it correctly with the GLS approach. An interesting alternative is to use the moving block bootstrap (MBB) technique (Mudelsee [2019]) which takes into account implicitly the distribution and the correlation of the noise, without modeling them explicitly.

General results for all stations We tested if the estimated trend was different from zero to a given value (0.05) significance level. We have therefore carried out a hypothesis test where the null hypothesis H0: $a = 0$, against the alternative hypothesis, H1: $a \neq 0$. If the null hypothesis is true, the t-statistic is $t = \hat{a}/\sigma_a$ and its law under H_0 is $\mathcal{N}(0, 1)$. We reject H_0 with a significance level of 5% if p-value$= \Phi(-|t_{\text{obs}}|) < 0.05$ where t_{obs} is the observed test statistic and Φ is the cumulative distribution function of $\mathcal{N}(0, 1)$. If we consider the OLS and GLS models without taking into account the change points, respectively 67 and 29 trend estimates are significant at the 0.05 level. However, we know that these trends estimates are not reliable because even a single change-point is enough to bias the trend estimate. If the model includes the change points detected with the segmentation method described in Section 5.2, we have 40 and 14 stations with significant trend estimates at the 0.05 level. Figures 5.16 shows the trends of the GLS solution for both cases (with and without including the change-points). Figures 5.17 shows the same results with error bars. In these Figures we see that ALIC is the worst case where the trend is increased. There are other cases where the trend is actually decreased, and cases where it is unchanged (the latter include stations with no change-points).

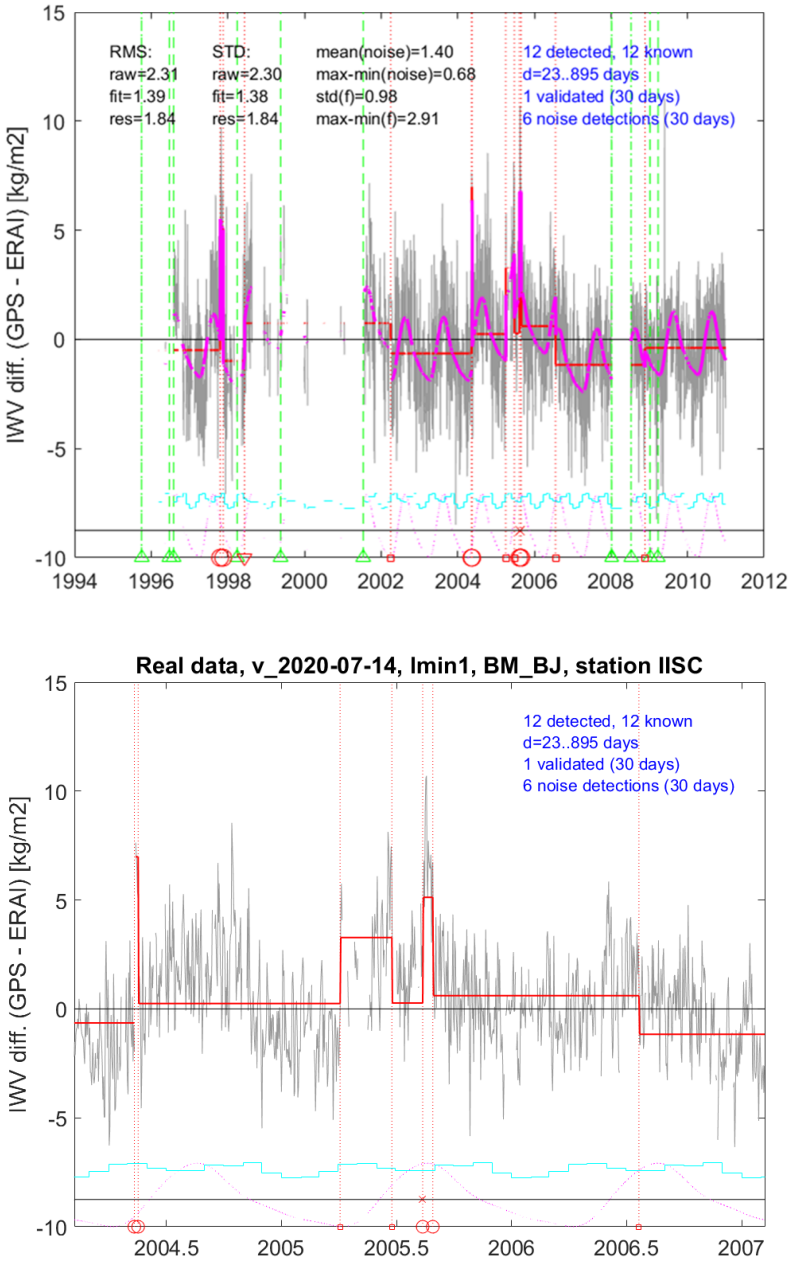


Figure 5.10 – Outlier detection and classification for the case of station IISC, with a threshold of 30 days. Upper: full time series. The vertical dotted red lines show the detected change-points and the vertical dashed green lines show the equipment changes from metadata. Symbols on the bottom: a red circle indicates an outlier, a red square a regular change-point, a red inverted triangle a validated change-point. On the black horizontal line at -9, the red symbol "x" atop the first outlier of a clusted indicates that the change in mean is not significant (class 1) and the screening will remove both outliers. For the other clusters (class 2), the screening with replace the change-points by the mid-point. The lower plot shows a zoom on the class 1 outliers.

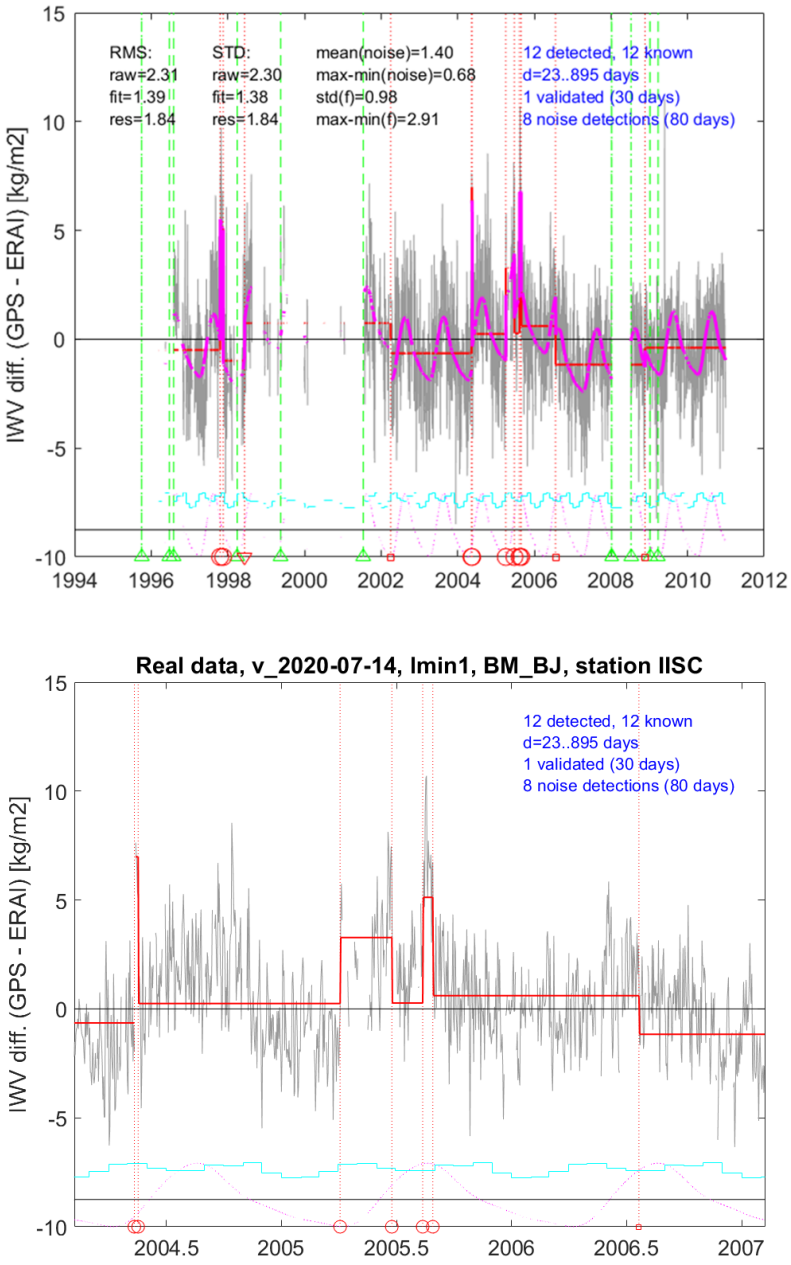


Figure 5.11 – Similar to Figure 5.10 but with the threshold of 80 days. Note that the four change-points of year 2005 are all outliers and belong to the same cluster of class 2.

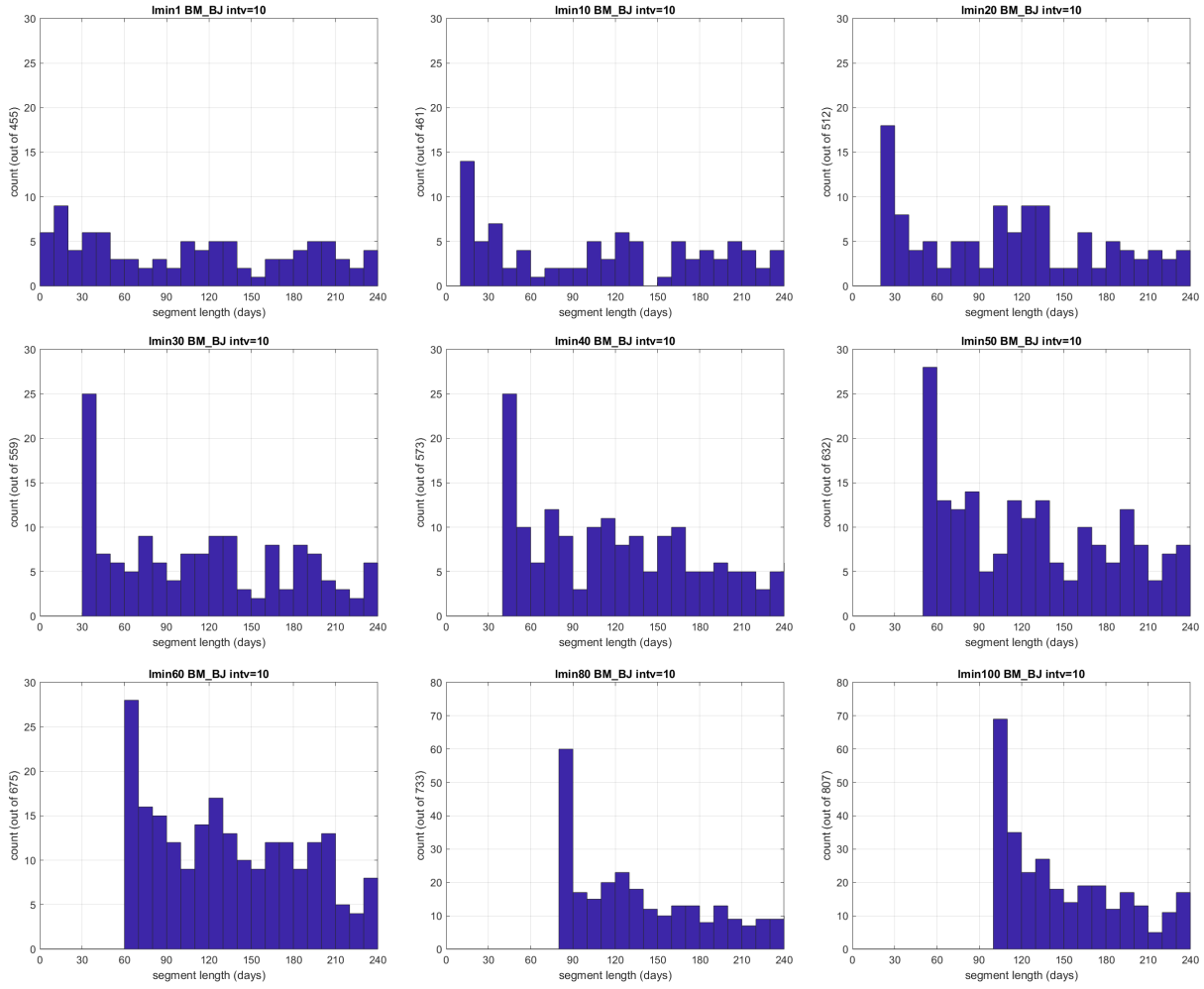


Figure 5.12 – Histograms of the detected segment lengths, for BM1, for lmin varying from 1 to 100. Note the change in vertical axis for the latter two plots.

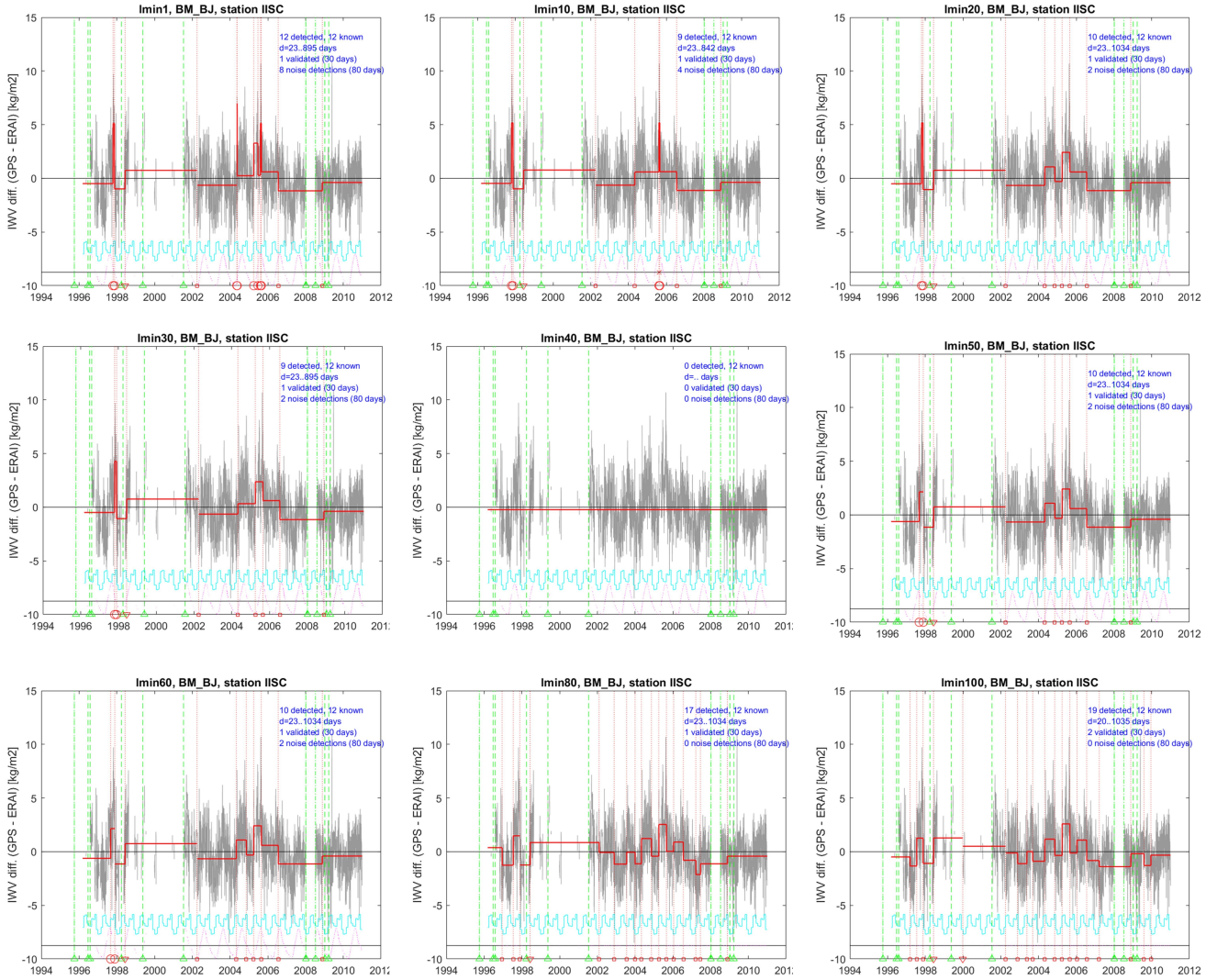


Figure 5.13 – Similar to Figure 5.10 but for different lmin values (1 to 100, see figure titles) and an outlier detection threshold of 80 days. The estimated periodic function is not added to the means for clarity.

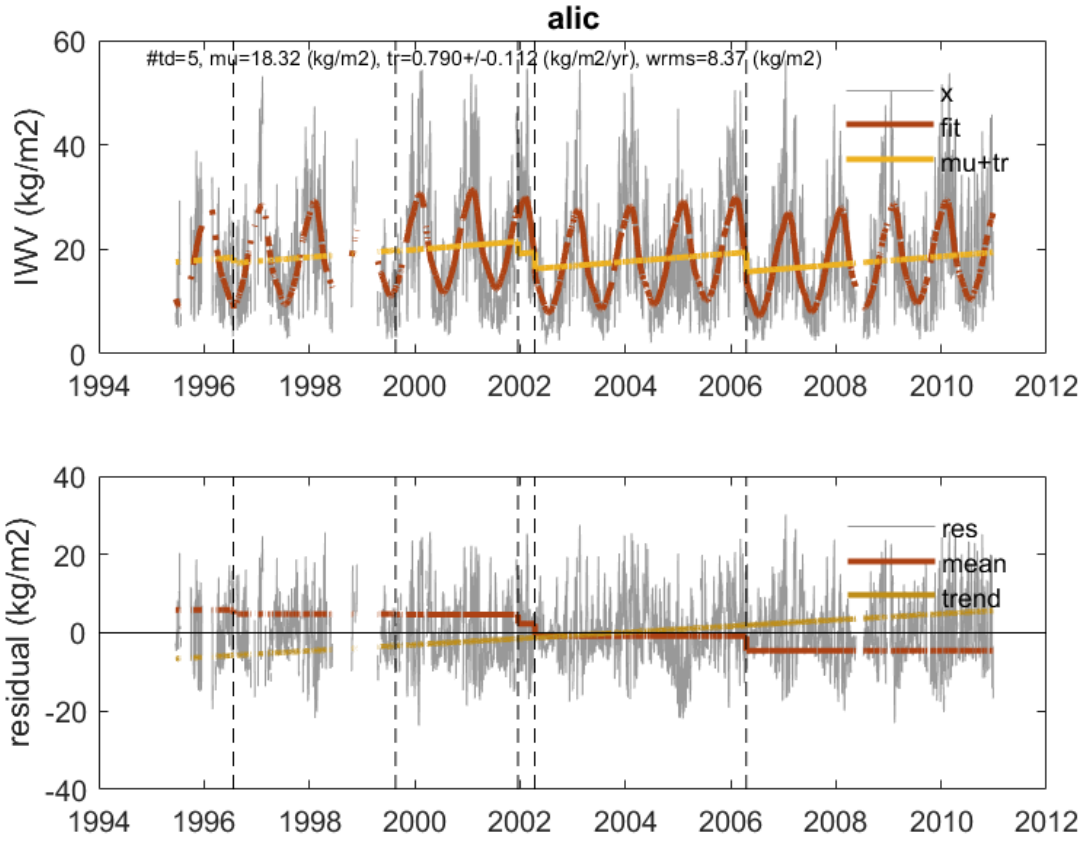


Figure 5.14 – Time series of GNSS IWV for the station ALIC and fitted trend model with OLS: (top) the time series is plotted in gray, the red line is the fitted model, and the yellow line is the estimated trend + means, (bottom) the residuals are plotted in gray, centred means in red, and the trend in yellow. The vertical black dashed lines are the detected change-points from the segmentation (after the screening). The trend value and its standard error are given in the upper plot.

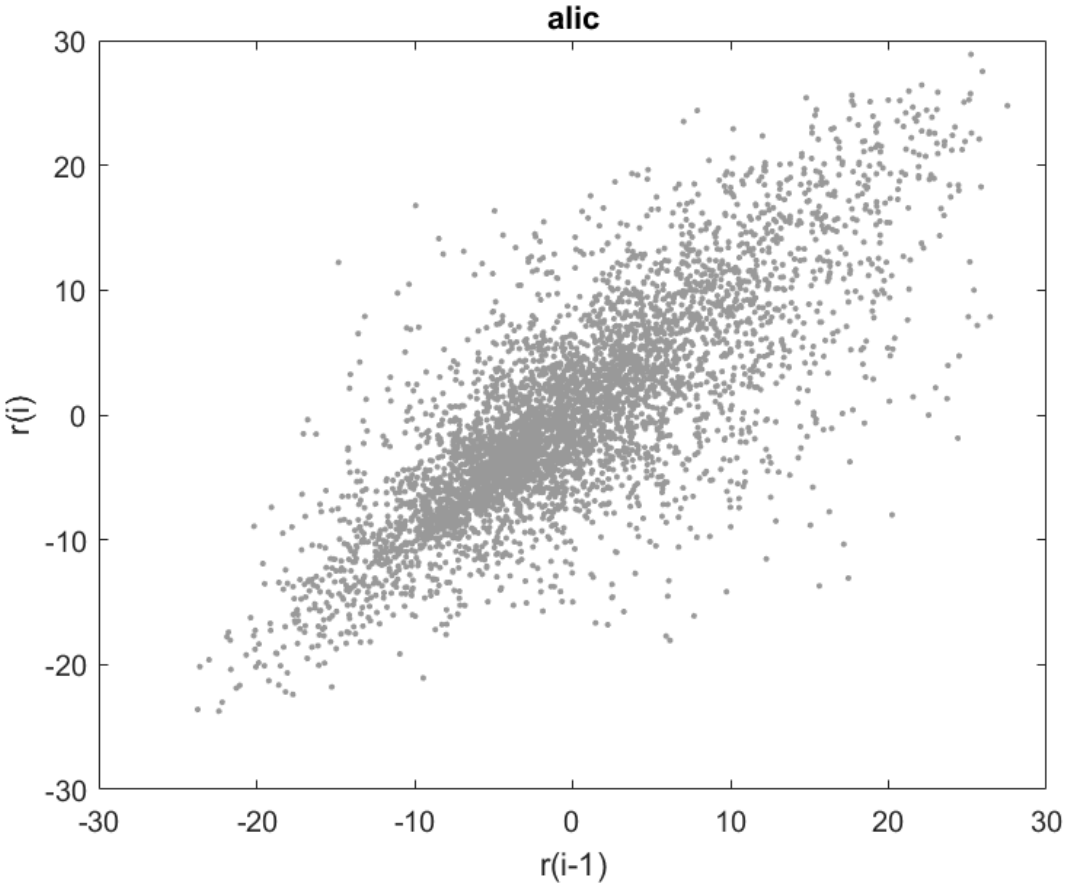


Figure 5.15 – OLS regression residuals from 5.14

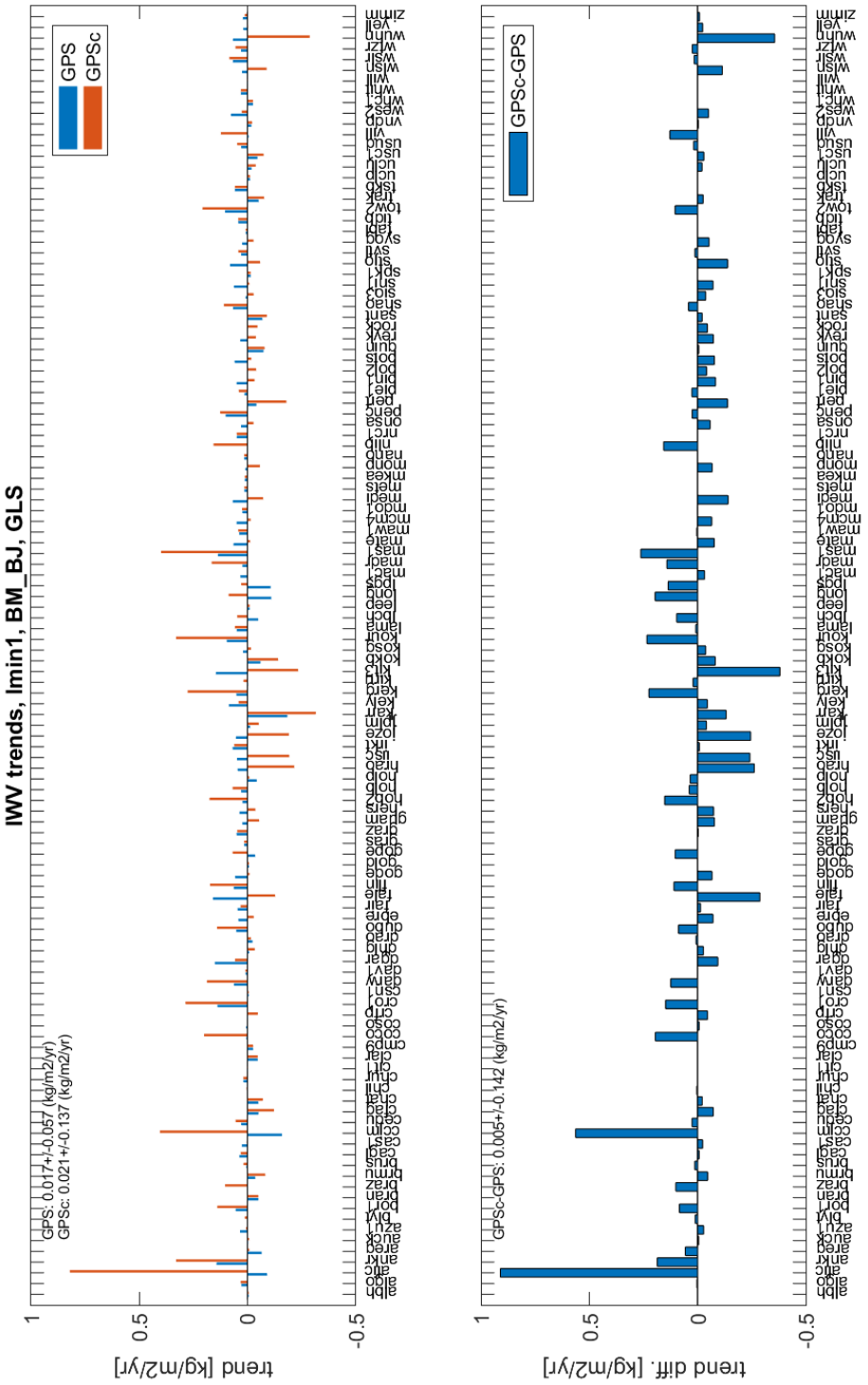


Figure 5.16 – Trend estimates (on the top) and difference (on the bottom) between the GLS estimate for the trend without considering the change-points (GPS) and GLS estimate integrating change-points in the model (GPSc.)

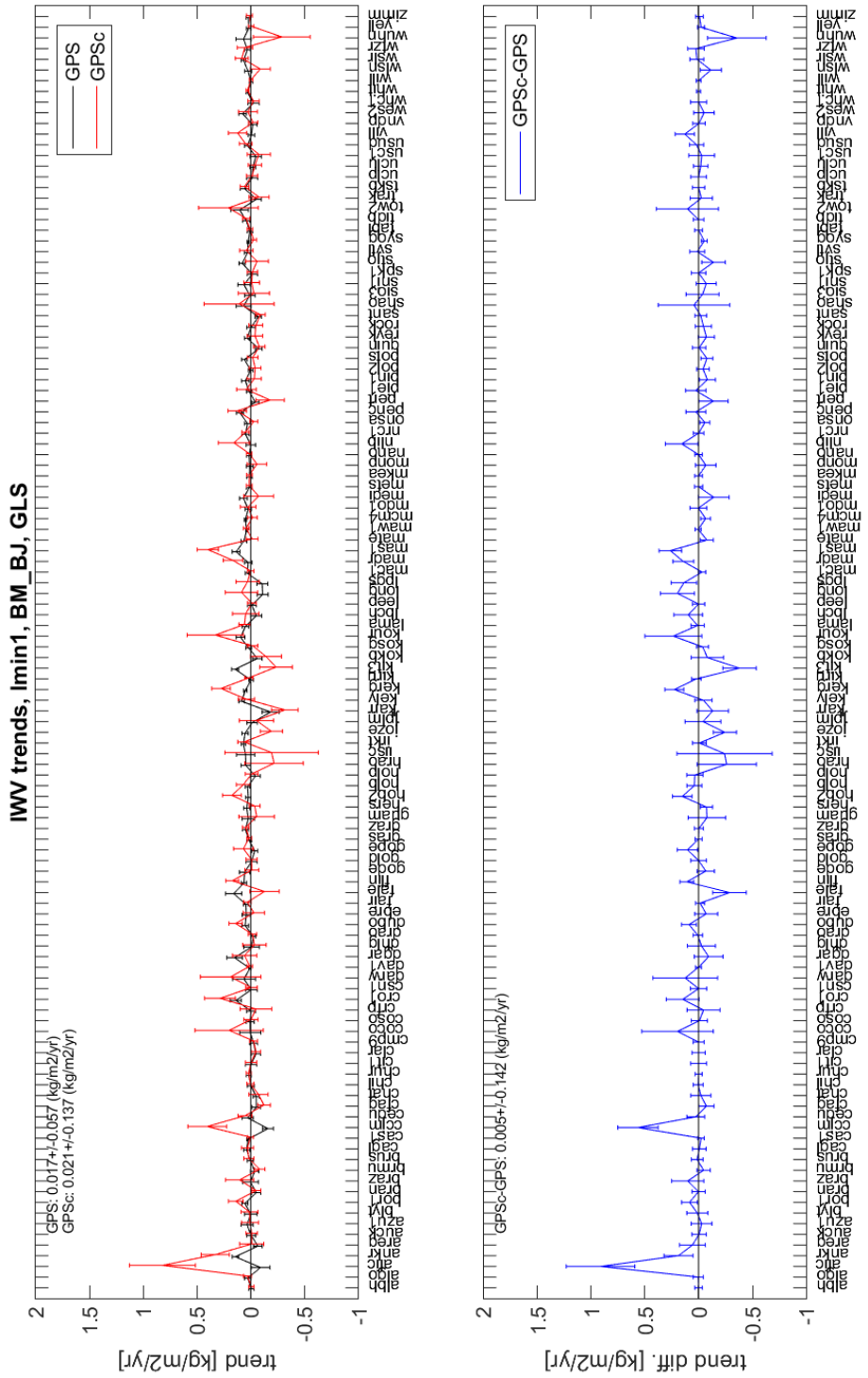


Figure 5.17 – Trend estimates (on the top) and difference (on the bottom) between the GLS estimate for the trend without considering the change-points (GPS) and GLS estimate integrating change-points in the model (GPSc.), with error bars.

Chapter 6

Conclusions and perspectives

6.1 Discussion and conclusions

In this thesis we have developed a new segmentation method devoted to the detection of abrupt changes in the mean which takes into account a periodic bias and a heterogeneous variance on fixed intervals (monthly) in the IWV differences between GNSS observations and ERA-Interim reanalysis. The method was tested and optimized first through a simulation study and then applied to IWV GNSS data for 120 stations of the global IGS network for the period from January 1995 to December 2010. The method works well and was published on the CRAN. Below we discuss a few limitations and issues that were noticed and possible ways to solve them.

Segmentation: sensitivity of the results to the bias function f. The simulation study revealed a sensitivity of the segmentation results to the way the periodic bias f is initialized and estimated. Indeed, there is possible confusion between the segmentation parameters and the period bias. The selection of significant parameters of the function model was shown to be able to stabilize the problem in the case of the simulations. With the real data, on the other hand, this option did not have a significant impact on the segmentation results. Indeed, in the real data, the smoothly varying bias which is modeled so far by a low order periodic function may be more complex. Also the noise in the real data is not Gaussian iid but has some autocorrelation (although it is believed to be short term) which may contribute to the bias variations. Using a more complex function basis for the estimation of f or a non-parametric approach may help to solve this issue. It is also expected that the use of a reference IWV data set with reduced representativeness errors (e.g. ERA5 reanalysis) would decrease the smoothly varying bias.

Segmentation: model selection. The model selection is a delicate and difficult problem. The tested criteria give different results both with simulations and with the real GNSS data. In the simulations, Lav has an unstable behavior, while BM1, BM2, and mBIC give very similar results. The best results are found with BM1 which estimates the lowest number of change-points and outliers, and achieves the highest validation rate. In real data, mBIC strongly over-segments, Lav remains unstable but approaches the behavior of BM1 and BM2, but the best criterion is again BM1. In general, we prefer a criterion which does not estimate too many change-points. We also noticed the presence of outliers which are couples of change-points located close together. They are typically due to large noise spikes in the series which can be removed by a screening method.

Screening. The detection and removal of outliers was treated with a screening method which consists in testing the variations in mean of the segment before and after a group of "close" outliers (called a cluster). The method proved efficient. In the case of BM1 criterion, it detected 20% of change-points as outliers and removed one third of them when a threshold of 80 days was used to detect outliers ("close" change-points) and a significance level of 0.05 for the change of means. Another approach to handle the outlier issue was tested by imposing a minimum segment length in the segmentation algorithm. This approach had one main drawback which is that although segments with length smaller than the threshold $lmin$ are avoided, the total number of change-points increased as $lmin$ was increased. Again this tendency reflects bad behavior of the model selection criteria (all except mBIC which still largely over-segments). As a compromise, we found that $lmin = 10$ combined with the screening yields the best results (highest validation rate). Other strategies could consist of better filtering the initial data (screening of GNSS IWV), using a reference with smaller representativeness differences (e.g. ERA5 reanalysis), using a different criterion to be optimized in the segmentation that is less sensitive to noise spikes than the least-squares or the Gaussian PDF assumption in the log-likelihood (Eq. 4.2). In Section 6.2 below we show some preliminary results with a Biweight loss function.

Attribution. Attribution of the change-points to the GNSS using metadata produced a rate of 23%. This rate remains low. Possible causes can be undocumented GNSS equipment changes, changes in the station's environment, inhomogeneity in the ERAI reference series, and/or false detections due to unmodelled effects in the signal. Another approach to confirm the GNSS origin of the detected change-points is to test the significance of the changes in the mean when the target GNSS series is compared to another, nearby, GNSS series. Such a method was implemented by Ning *et al.* [2016] who had obtained a similar $\tilde{2}0\%$ of validations from metadata. They cross-compared 59 out of 62 GNSS stations of their

global network which probably included baselines of several 100 kilometers (this information is not given in the publication). Using denser GNSS networks and data from other observing systems (DORIS, VLBI, radiosondes) might help to implement this approach in the future.

Estimation of linear trends. The linear trends were estimated by ordinary and generalized least-squares (OLS and GLS) on the GNSS IWV time series including change-points detected by the segmentation (after screening). Although this approach has been commonly used in the GNSS community (Bernet *et al.* [2020]; Klos *et al.* [2018]) a confusion between the trend and mean parameters was found. This issue gets worse when the number of change-points increases. Another issue is due to autocorrelation of the "noise" in the IWV series which actually represents the day to day atmospheric variability. The lag-1 correlation coefficients for the 120 stations range between 0.22 and 0.81, with a median of 0.64. The trend estimation by GLS assuming an AR(1) process gives more realistic trend uncertainties than OLS but they are also much larger. As a results only 12% of the stations have a significant trend (compared to 30% with OLS) but due to the confusion effect, these trends are probably mis-represented. This problem still needs to be improved.

Correction of the IWV time series. To produce a corrected (homogenized) series, it is necessary to remove the offsets in the mean between segments. This can be can done in two ways: either subtract the variations in the means estimated during the segmentation or subtract the means estimated during the trend estimation. Both approaches are unsatisfying so far. In this first case the offsets might be due to a change-point in the reference series. As long as the change-points cannot efficiently be attributed to GNSS this approach will introduces spurious offsets in the corrected signal. In the second case, the confusion between the estimated trends and means will over-estimate the offsets and produce similar errors. These issues need to be solved first.

Applications on Benchmark data of the COST GNSS4SWEC. Our new segmentation method was used as part of the COST GNSS4SWEC Benchmark on three synthetic data sets based on the analysis of real data Van Malderen *et al.* [2020]. All three data sets included abrupt changes in the mean that were randomly distributed, seasonal signals (annual, semi-annual, 3 and 4 months), and different noise processes. The "easy" data set simulated Gaussian white noise; the "moderate" data set simulated white noise plus AR(1) noise; and the "complex" data set moreover included gaps and local trends. Our method obtained the best scores among all the segmentation methods tested exaequo with the ACMANT method Domonkos & Coll [2017] also based on a penalised maximum likelihood approach. However, with "complex" data set, all methods have trouble because none of the methods includes the

autocorrelation in the model. There are some methods that include a dependence in the series (see 2.1), however these methods are sub-optimal. One evolution of our method which is optimal would consist in taking the autocorrelation into account. This point is further discussed below.

6.2 Perspectives

The limitations of the present version of the segmentation method and the general homogenization and trend estimation procedures have been highlighted and discussed in the previous sub-section. Here we discuss some additional ideas of improvements of the method that would require new developments (new models and inference approaches).

Improving the estimation of the function f. Using a non-parametric or semi-parametric approach could bring more flexibility and improve the estimation of f and consequently the segmentation (estimation of the change-points). One could consider the semi-parametric approach proposed by Bertin *et al.* [2017]. They considered a segmentation model including a functional part for which a dictionary approach is used to estimate it. A Lasso procedure is used to select the relevant functions of the dictionary. This approach allows for example the estimation of functions that are smooth and also show some irregularities.

Improving the segmentation using Hubert or Biweight losses. A way to avoid the detection of outliers could be to consider an adapted loss function as the Hubert or the Biweight loss (instead of the log-likelihood). Indeed, in this case the mean and the functional estimates would not take into account the extreme data points and thus avoid the detection of outliers. Both the segmentation results and functional estimates would be more robust. We did a preliminary test of this method using the gfpop R package (Hocking *et al.* [2018]) and we applied the segmentation with a weighted Biweight loss on the particular station IISC (corrected from the functional estimated by our current method). Figure 6.1 compares this method to the results obtained with our current method shown earlier (Figure 5.11). Recall that the current procedure detects 12 change-points of which 8 outliers and, after screening 7 remained. Now, the robust segmentation detects only 5 change-points: 4 change-points are common and one (on 2005-01-31) replaces a cluster of outliers detected with the screening (see Chapter 5.3.1).

Improving the segmentation by including dependence. On the station CCJM, we computed the lag-1 of the autocorrelation function of the residuals and found a value of $r = 0.249$. The value slightly decreases to 0.223 when it is computed from the residuals of the GLS estimation including the screened change-points. The r values for all the stations range between 0.204 and 0.655 with a median

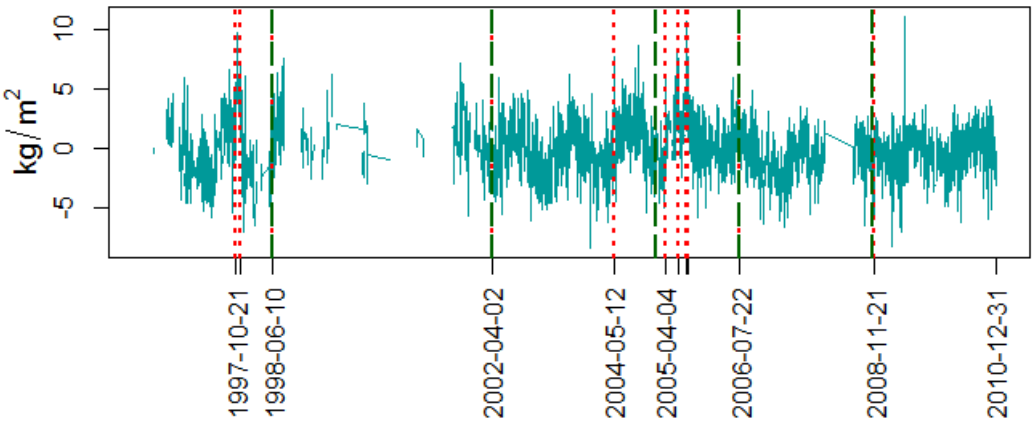


Figure 6.1 – Segmentation result obtained with a robust method (Biweight loss) for the station IISC (dashed green lines) compared to the current method (same as in Figure 5.11) represented as dotted red lines.

value of 0.386. In order to take this dependence into account in the model used for the segmentation, we could follow the work of Chakar *et al.* [2017] who proposed a segmentation model with an autoregressive noise (not affected by the change-points).

Application to other data sets. New GNSS and reanalysis data sets are now available, in particular global GNSS reprocessed series spanning a longer period (1994-2019, Bock [2019]), denser GNSS regional networks, e.g. in Europe (Pacione *et al.* [2017]), and higher-resolution reanalyses (e.g. ERA5 and UERRA). It would be interesting to compare the segmentation results and the trend estimates obtained with these new data sets to those found in this work. Longer time series are also of special interest to investigate more extensively the climate trends and variability.

Appendix A

GCOS and NDACC networks

The Climate Observing System (GCOS) program was established in 1992 and stimulates, coordinates, and facilitates the taking of needed observations by national or international organizations to support their own requirements as well as common goals
(https://gcos.wmo.int). It provides an operational framework for integrating and enhancing the observational systems of participating countries and organizations into a comprehensive system focused on the requirements for climate issues. It includes in-situ observations from atmospheric, ocean, and terrestrial instruments, as well as satellite data. The atmospheric component comprises a baseline network of about 1000 ground stations providing temperature and precipitation observations and a network of radiosonde stations providing upper-air measurements of temperature, pressure, wind, and humidity from balloons. The GCOS surface and upper air data archives contain high quality continuous observations dating back to the early 1900s and 1950s, respectively.

The Network for the Detection of Atmospheric Composition Change (NDACC) is composed of more than 70 globally distributed, ground-based, remote-sensing research stations with more than 160 currently active instruments providing high quality, consistent, standardized, long-term measurements of atmospheric temperatures and trace gases, particles, spectral UV radiation reaching the Earth's surface, and physical parameters for detection of trends in overall atmospheric composition, understanding their impacts on the stratosphere, troposphere, and mesosphere, establishing links between climate change and atmospheric composition, testing and validating atmospheric measurements from satellites, supporting process-focused scientific field campaigns, and testing and improving theoretical models of the atmosphere http://www.ndaccdemo.org. The NDACC began network operations as The Network for Detection of Stratospheric Change (NDSC) in January 1991 and includes data back to the 1960s.

Appendix B

Principles of the GNSS IWV technique.

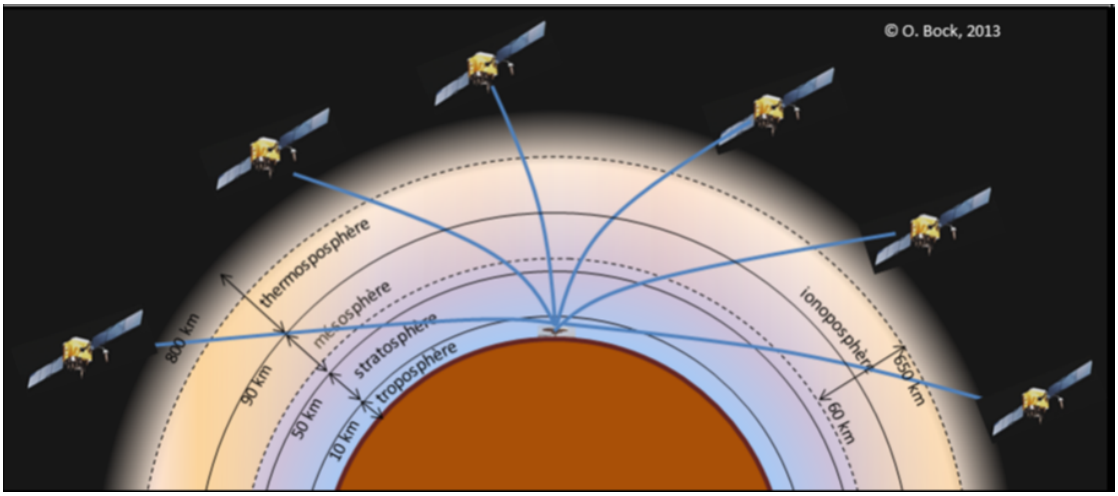


Figure B.1 – Propagation of GPS signals. Source Bock, 2013.

The GNSS IWV technique is based on the estimation of the propagation delay of radio waves transmitted from a constellation of GNSS satellites that are measured by ground-based receiving stations Bevis *et al.* [1992], as shown in Figure B.1. It is a remote sensing technique that involves complex signal processing and auxiliary data. The technique involves a constellation of satellites (nominally 24 satellites in the case of GPS) for which the orbits are determined from the analysis of the measurements collected by a ground-based tracking network. In order to link the satellite positions and the Earth-fixed receiver positions, Earth Rotation Parameters (ERPs) as well as other system parameters are also required (e.g. satellite clock offsets, inter-system biases when the measurements from several systems, e.g. GPS and

GLONASS, are analysed together). These auxiliary data are produced by the system operators and broadcasted as a message contained in the GNSS satellite signals to the general user for navigation purposes (Hofmann-Wellenhof *et al.* [1993]). For scientific applications such as geodesy, geophysics, weather forecasting and climate monitoring, the accuracy requirements on these auxiliary data are much more stringent. High quality products are thus provided by the International GNSS Service (IGS) who coordinates the analysis of a global reference network since 1994 (http://www.igs.org/network). Initially composed of about 50 stations, the IGS network grew rapidly to about 300 stations and has been stabilized around this number since 2003 (Dow *et al.* [2008]). Figure 2.5 gives a view of the station network used in this work (120 stations with time series of 15 years). In addition to the IGS network, many national and regional networks have been established, based on public and private initiatives for both scientific and commercial activities. Part of the observations collected by these networks is publicly available and many of these data are adequate for being reprocessed and used for scientific applications (Jones *et al.* [2019]).

The GNSS data processing technique for scientific applications consists typically in the analysis of daily batches of code and phase measurements of one or several stations. For each station, one set of 3D coordinates is estimated daily along with a group of zenith tropospheric delays (ZTDs) with a time interval between five minutes and two hours, as well as a number of other parameters (phase ambiguities, receiver clock offsets, tropopsheric gradients), depending on the software and processing options (see Jones *et al.* [2019], for more details on processing options and https://www.unavco.org/software/data-processing/postprocessing/postprocessing.html for a description of processing software).

The IWV estimates are derived from the ZTDs (Bevis *et al.* [1992]; Bock [2014]), for more detailed see Chapter 2 of Bernardes Parracho [2017]. The difference ZWD = ZTD - ZHD corresponds to the zenith wet delay; ZHD is the zenith hydrostatic delay, it is computed from the barometric surface pressure at the level of the GNSS antenna; Tm is the weighted mean temperature in the atmospheric column above the antenna.

The uncertainty in the GNSS IWV estimates is dependent on the uncertainty in each of the three variables: ZTD (determined by the quality of the phase measurements and the data processing procedure), K(Tm) (determined by the accuracy of the Tm data and the empirically determined refractivity constants K2 and k3) and ZHD (dependent on the accuracy of the surface pressure data and refractivity constant k1). The quality of the phase measurements depends on the quality of the instrumentation (GNSS receiver and antenna, including the stability of the receiver clock) but also of the environment. Signal reflection and scattering on the environment (ground surface, vegetation, buildings) that is detected by the receiving antenna is known to interfere with the direct signal coming from the satellites

and generate signal fading and phase errors which a have detrimental impact on the quality of the station coordinates and ZTDs (Elósegui *et al.* [1995]). The accuracy of the estimated parameters is also depending on the tropospheric propagation modelling approach, especially the mapping functions used to relate the delays in the direction of the satellites to the delay at the zenith under the assumption of a perfectly layered atmosphere (Boehm & Schuh [2013]). A detailed discussion and evaluation of each of the error sources is given in Bock & Parracho [2019]; Ning *et al.* [2016]; Parracho *et al.* [2018].

In order to guarantee a high accuracy and homogeneity of the ZTD and IWV estimates for climate trend analysis, it is crucial to adopt a frozen processing procedure and use consistent and homogeneous auxiliary data, namely reprocessed satellite orbits, clocks, and ERPs (Ostini [2012]; Steigenberger [2006]).

References

Aguilar, E., Auer, I., Brunet, M., Peterson, T. & Wieringa, J. (2003). Guidelines on climate metadata and homogenization. *World Meteorological Organization (WMO)TD No. 1186/World Climate Data Monitoring Program (WCDMP) No. 53, Geneva; 52.*. 45, 53, 55

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. *In B.N. Petrov et F. Csaki (Eds.), Second International Symposium on Information Theory*, 267–281. 61, 68

Alexandersson, H. (1986). A homogeneity test applied to precipitation data. *Journal of Climatology*, **6**, 661–675. 55

Alexandersson, H. & Moberg, A. (1997). Homogenization of swedish temperature data. part i: Homogeneity test for linear trends. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, **17**, 25–34. 55

Ardia, D., Dufays, A. & Criado, C.O. (2019). Frequentist and bayesian change-point models: A missing link. *Available at SSRN 3499824*. 53, 69

Arlot, S. & Massart, P. (2009). Data-driven calibration of penalties for least-squares regression. *J. Mach. Learn. Res.*, **10**, 245–279. 21, 69, 82

Auer, I., Böhm, R., Jurković, A., Orlik, A., Potzmann, R., Schöner, W., Ungersböck, M., Brunetti, M., Nanni, T., Maugeri, M. *et al.* (2005). A new instrumental precipitation dataset for the greater alpine region for the period 1800–2002. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, **25**, 139–166. 55

Auger, I.E. & Lawrence, C.E. (1989). Algorithms for the optimal identification of segment neighborhoods. *Bulletin of Mathematical Biology*, **51**, 39–54. 61

Bai, J. & Perron, P. (2003). Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, **18**, 1–22. 67

Beaulieu, C., Seidou, O., Ouarda, T., Zhang, X., Boulet, G. & Yagouti, A. (2008). Intercomparison of homogenization techniques for precipitation data. *Water Resources Research*, **440**. 42, 55

Bellman, R. (1954). The theory of dynamic programming. *Bulletin of the American Mathematical Society*, **60**, 503–515. 21, 61, 65, 66

Bernardes Parracho, A.C. (2017). *Study of trends and variability of atmospheric water vapour with climate models and observations from global gnss network*. Theses, Université Pierre et Marie Curie - Paris VI. 141

Bernet, L., Brockmann, E., von Clarmann, T., Kämpfer, N., Mahieu, E., Mätzler, C., Stober, G. & Hocke, K. (2020). Trends of atmospheric water vapour in switzerland from ground-based radiometry, ftir and gnss data. *Atmospheric Chemistry and Physics*, **20**, 11223–11244. 32, 136

Bertin, K., Collilieux, X., Lebarbier, E. & Meza, C. (2017). Semi-parametric segmentation of multiple series using a dp-lasso strategy. *Journal of Statistical Computation and Simulation*, **87**, 1255–1268. 31, 50, 80, 137

Bevis, M., Businger, S., Herring, T.A., Rocken, C., Anthes, R.A. & Ware, R.H. (1992). Gps meteorology: Remote sensing of atmospheric water vapor using the global positioning system. *JOURNAL OF GEOPHYSICAL RESEARCH*, **97**, 787–801. 140, 141

Biernacki, C., Celeux, G. & Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**, 719–725. 113

Birgé, L. & Massart, P. (2001). Gaussian model selection. *Journal of the European Mathematical Society*, **3**, 203–268. 21, 61, 68, 81

Birgé, L. & Massart, P. (2007). Minimal penalties for gaussian model selection. *Probability Theory and Related Fields*, **138**, 33–73. 61, 68

Bock, O. (2014). Les systèmes de positionnement et de navigation par satellite : Application à la météorologie et à la climatologie. *La Météorologie*, **8**, 38. 7, 9, 17, 51, 141

Bock, O. (2017). Global gps integrated water vapor - igs repro1 - v1. 16, 50, 99

Bock, O. (2019). Global gnss iwv data at 436 stations over the 1994-2018 period. 32, 138

Bock, O. & Nuret, M. (2009). Verification of nwp model analyses and radiosonde humidity data with gps precipitable water vapor estimates during amma. *Weather and Forecasting*, **24**, 1085–1101. 48

Bock, O. & Parracho, A. (2019). Consistency and representativeness of integrated water vapour from ground-based gps observations and era-interim reanalysis. *Atmos. Chem. Phys.*, **19**, 9453–9468. 17, 18, 53, 124, 142

Bock, O., Bouin, M.N., Walpersdorf, A., Lafore, J.P., Janicot, S., Guichard, F. & Agusti-Panareda, A. (2007). Comparison of ground-based gps precipitable water vapour to independent observations and nwp model reanalyses over africa. *Quarterly Journal of the Royal Meteorological Society*, **133**, 2011–2027. 48

Bock, O., Bouin, M.N., Doerflinger, E., Collard, P., Masson, F., Meynadier, R., Nahmani, S., Koité, M., Gaptia Lawan Balawan, K., Didé, F., Ouedraogo, D., Pokperlaar, S., Ngamini, J.B., Lafore, J.P., Janicot, S., Guichard, F. & Nuret, M. (2008). West african monsoon observed with ground-based gps receivers during african monsoon multidisciplinary analysis (amma). *Journal of Geophysical Research: Atmospheres*, **113**. 48

Bock, O., Willis, P., Lacarra, M. & Bosser, P. (2010). An inter-comparison of zenith tropospheric delays derived from doris and gps data. *Adv. Space Res.*, **46**, 1408–1447. 16, 42, 48, 119

Bock, O., Collilieux, X., Guillamon, F., Lebarbier, E. & Pascal, C. (2018). A breakpoint detection in the mean model with heterogeneous variance on fixed time-intervals. *Statistics and Computing*, **63**, 22–32. 10, 21, 22, 26, 55, 60, 61, 62, 63, 67, 68, 72, 73, 75, 76, 77, 78, 80, 85, 102

Boehm, J. & Schuh, H. (2013). *Atmospheric Effects in Space Geodesy*. Springer Atmospheric Sciences, Springer Berlin Heidelberg. 142

Bonsal, B.R., Zhang, X., Vincent, L.A. & Hogg, W.D. (2001). Characteristics of daily and extreme temperatures over canada. *Journal of Climate*, **14**, 1959–1976. 55

Brunetti, M.a. (2009). Estimating local records for northern and central italy from a sparse secular temperature network and from 1961-1990 climatologies. *Advances in Science and Research*, **3**, 63. 55

Caussinus, H. & Mestre, O. (2004). Detection and correction of artificial shifts in climate series. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **53**, 405–425. 47, 55, 60, 61, 68

Chakar, S., Lebarbier, E., Lévy-Leduc, C., Robin, S. *et al.* (2017). A robust approach for estimating change-points in the mean of an AR(1) process. *Bernoulli*, **23**, 1408–1447. 67, 73, 80, 138

Cleynen, A., Dudoit, S. & Robin, S. (2014). Comparing segmentation methods for genome annotation based on rna-seq data. *Journal of Agricultural, Biological, and Environmental Statistics*, **19**, 101–118. 68

Collilieux, X., Métivier, L., Altamimi, Z., Van Dam, T. & Ray, J. (2011). Quality assessment of gps reprocessed terrestrial reference frame gps solut. *GPS Solutions*, **15**, 219–231. 50

Collilieux, X., Lebarbier, E. & Robin, S. (2019). A factor model approach for the joint segmentation with between-series correlation. *Scandinavian Journal of Statistics*, **46**, 686–705. 17, 50, 51

Costa, A.C. & Soares, A. (2009). Homogenization of climate data: Review and new perspectives using geostatistics. *Mathematical Geosciences*, **41**, 291–305. 45

Craddock, J.M. (1979). Methods of comparing annual rainfall records for climatic purposes. *Weather*, **34**, 332–346. 55

Dai, A., Wang, J., Thorne, D.E., P. W.and Parker, Haimberger, L. & Wang, X.L. (2011). A new approach to homogenize daily radiosonde humidity data. *J. Clim.*, **24**, 965–991. 48

Dee, D.P., Uppala, S., Simmons, A., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M., Balsamo, G., Bauer, d.P. *et al.* (2011). The era-interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the royal meteorological society*, **137**, 553–597. 16, 41, 50, 51

Domonkos, P. & Coll, J. (2015). Homogenization of precipitation time series with acmant. *Theor Appl Climatol*, **122**, 303–314. 42, 55, 61

Domonkos, P. & Coll, J. (2017). Homogenisation of temperature and precipitation time series with acmant3: method description and efficiency tests. *International Journal of Climatology*, **37**, 1910–1921. 55, 136

Domonkos, P., Sigró, J. & Poza, R. (2011). Adapted caussinus-mestre algorithm for networks of temperature series (acmant). *International Journal of Geosciences*, **02**. 55, 60

Dow, J., Neilan, R. & Rizos, C. (2008). The international gnss service in a changing landscape of global navigation satellite systems. *Journal of Geodesy*, **83**, 191–198. 141

Ducré-Robitaille, J.F., Vincent, L.A. & Boulet, G. (2003). Comparison of techniques for detection of discontinuities in temperature series. *International Journal of Climatology*, **23**, 1087–1101. 54, 55, 57

Dunn, R.J.H., Stanitski, D.M., Gobron, N., Willett, K.M., Ades, M., Adler, R., Allan, R., Allan, R.P., Anderson, J., Argüez, A., Arosio, C., Augustine, J.A., Azorin-Molina, C., Barichivich, J., Barnes, J., Beck, H.E., Becker, A., Bellouin, N., Benedetti, A., Berry, D.I., Blenkinsop, S., Bock, O., Bosilovich, M.G., Boucher, O., Buehler, S.A., Carrea, L., Christiansen, H.H., Chouza, F., Christy, J.R., Chung, E.S., Coldewey-Egbers, M., Compo, G.P., Cooper, O.R., Covey, C., Crotwell, A., Davis, S.M., de Eyto, E., de Jeu, R.A.M., VanderSat, B., DeGasperi, C.L., Degenstein, D., Girolamo, L.D., Dokulil, M.T., Donat, M.G., Dorigo, W.A., Durre, I., Dutton, G.S., Duveiller, G., Elkins, J.W., Fioletov, V.E., Flemming, J., Foster, M.J., Frey, R.A., Frith, S.M., Froidevaux, L., Garforth, J., Gupta, S.K., Haimberger, L., Hall, B.D., Harris, I., Heidinger, A.K., Hemming, D.L., peng (Ben) Ho, S., Hubert, D., Hurst, D.F., Hüser, I., Inness, A., Isaksen, K., John, V., Jones, P.D., Kaiser, J.W., Kelly, S., Khaykin, S., Kidd, R., Kim, H., Kipling, Z., Kraemer, B.M., Kratz, D.P., Fuente, R.S.L., Lan, X., Lantz, K.O., Leblanc, T., Li, B., Loeb, N.G., Long, C.S., Loyola, D., Marszelewski, W., Martens, B., May, L., Mayer, M., McCabe, M.F., McVicar, T.R., Mears, C.A., Menzel, W.P., Merchant, C.J., Miller, B.R., Miralles, D.G., Montzka, S.A., Morice, C., Muhle, J., Myneni, R., Nicolas, J.P., Noetzli, J., Osborn, T.J., Park, T., Pasik, A., Paterson, A.M., Pelto, M.S., Perkins-Kirkpatrick, S., Pétron, G., Phillips, C., Pinty, B., Po-Chedley, S., Polvani, L., Preimesberger, W., Pulkkanen, M., Randel, W.J., Rémy, S., Ricciardulli, L., Richardson, A.D., Rieger, L., Robinson, D.A., Rodell, M., Rosenlof, K.H., Roth, C., Rozanov, A., Rusak, J.A., Rusanovskaya, O., Rutishäuser, T., Sánchez-Lugo, A., Sawaengphokhai, P., Scanlon, T., Schenzinger, V., Schladow, S.G., Schlegel, R.W., Eawag Schmid, M., Selkirk, H.B., Sharma, S., Shi, L., Shimaraeva, S.V., Silow, E.A., Simmons, A.J., Smith, C.A., Smith, S.L., Soden, B.J., Sofieva, V., Sparks, T.H., Stackhouse, P.W., Steinbrecht, W., Streletskiy, D.A., Taha, G., Telg, H., Thackeray, S.J., Timofeyev, M.A., Tourpali, K., Tye, M.R., van der A, R.J., Robin, V.B.v.d.S., van der SchrierW. Paul, G., van der Werf, G.R., Verburg, P., Vernier, J.P., Vömel, H., Vose, R.S., Wang, R., Watanabe, S.G., Weber, M., Weyhenmeyer, G.A., Wiese, D., Wilber, A.C., Wild, J.D., Wong, T., Woolway, R.I., Yin, X., Zhao, L., Zhao, G., Zhou, X., Ziemke, J.R. & Ziese, M. (01 Aug. 2020). Global climate. *Bulletin of the American Meteorological Society*, **101**, S9 – S128.

Easterling, D.R. & Peterson, T.C. (1995). A new method for detecting undocumented discontinuities in climatological time series. *International Journal of Climatology*, **15**, 369–377. 53

Elósegui, P., Davis, J.L., Jaldehag, R.T.K., Johansson, J.M., Niell, A.E. & Shapiro, I.I. (1995). Geodesy using the global positioning system: The effects of signal scattering on estimates of site position. *Journal of Geophysical Research: Solid Earth*, **100**, 9921–9934. 142

Estey, L. & Meertens, C. (1999). Teqc: The multi-purpose toolkit for gps/glonass data. *GPS Solutions*, **3**, 42–49. 49, 108

Fearnhead, P. (2006). Exact and efficient bayesian inference for multiple changepoint problems. *Stat Comput*, **116**, 203–213. 60

Fisher, W.D. (1958). On grouping for maximum homogeneity. *Journal of the American Statistical Association*, **53**, 789–798. 66

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S.C., Collins, W., Cox, P., Driouech, F., Emori, S., Eyring, V., Forest, C., Gleckler, P., Guilyardi, E., Jakob, C., Kattsov, V., Reason, C. & Rummukainen, M. (2013). *Evaluation of climate models*, 741–882. Cambridge University Press, Cambridge, UK. 47

Fryzlewicz, P. *et al.* (2014). Wild binary segmentation for multiple change-point detection. *The Annals of Statistics*, **42**, 2243–2281. 61

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

Gazeaux, J., Lebarbier, E., Collilieux, X. & Métivier, L. (2015). Joint segmentation of multiple gps coordinate series. *Journal de la Société Française de Statistique*, **156**, 163–179. 50, 80

Guijarro, J. (2011). Climatological series shift test comparison on running windows. *Idojaras*, **117**, 35–45. 55, 59

Guijarro, J. (2013). Climatological series shift test comparison on running windows. *Idojaras*, **117**, 35–45. 55

Guijarro, J. (2018). Homogenization of climatic series with climatol. 55

Held, I.M. & Soden, B.J. (2000). Water vapor feedback and global warming. *Annual Review of Energy and the Environment*, **25**, 445–475. 47

Held, I.M. & Soden, B.J. (2006). Robust responses of the hydrological cycle to global warming. *Journal of Climate*, **19**, 5686–5699. 47

Hewaarachchi, A.P., Li, Y., Lund, R. & Rennie, J. (2017). Homogenization of daily temperature data. *Journal of Climate*, **30**, 985–999. 55, 58, 60, 67

Hocking, T.D., Rigaill, G., Fearnhead, P. & Bourque, G. (2018). Generalized functional pruning optimal partitioning (gfpop) for constrained changepoint detection in genomic data. 23, 97, 137

Hofmann-Wellenhof, B., Lichtenegger, H. & J., C. (1993). *Global Positioning System. Theory and practice.*. 141

Jandhyala, V., Fotopoulos, S., MacNeill, I. & Liu, P. (2013). Inference for single and multiple change-points in time series. *Journal of Time Series Analysis*, **34**, 423–446. 53

Jones, J., Guerova, G., Douša, J., Dick, G., de Haan, S., Pottiaux, E., Bock, O., Pacione, R. & van Malderen, R. (2019). *Advanced GNSS Tropospheric Products for Monitoring Severe Weather Events and Climate: COST Action ES1206 Final Action Dissemination Report*. Springer International Publishing. 141

Jones, P.D., Raper, S.C.B., Bradley, R.S., Diaz, H.F., Kellyo, P.M. & Wigley, T.M.L. (1986). Northern hemisphere surface air temperature variations: 1851–1984. *Journal of Climate and Applied Meteorology*, **25**, 161–179. 16, 41, 42, 43

Karl, T.R. & Trenberth, K.E. (2003). Modern global climate change. *Science*, **302**, 1719–1723. 41

Karl, T.R. & Williams, C.N. (1987). An approach to adjusting climatological time series for discontinuous inhomogeneities. *Journal of Climate and Applied Meteorology*, **26**, 1744–1763. 55, 57

Killick, R., Fearnhead, P. & Eckley, I.A. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, **107**, 1590–1598. 61, 68

Klos, A., Hunegnaw, A., Teferle, F.N., Abraha, K.E., Ahmed, F. & Bogusz, J. (2018). Statistical significance of trends in zenith wet delay from re-processed gps solutions. *GPS Solutions*, **22**, 51. 32, 136

LANZANTE, J.R. (1996). Resistant, robust and non-parametric techniques for the analysis of climate data: Theory and examples, including applications to historical radiosonde station data. *International Journal of Climatology*, **16**, 1197–1226. 55, 57

Lavielle, M. (2005). Using penalized contrasts for the change-point problem. *Signal Processing*, **85**, 1501–1510. 21, 67, 68, 81

Lebarbier, E. (2005). Detecting multiple change-points in the mean of Gaussian process by model selection. *Signal Processing*, **85**, 717–736. 61, 68

Li, S. & Lund, R. (2012). Multiple changepoint detection via genetic algorithms. *Journal of Climate*, **25**, 674–686. 55, 57, 60, 61, 62, 68, 69

Li, Y. & Lund, R. (2015). Multiple changepoint detection using metadata. *Journal of Climate*, **28**, 4199–4216. 43, 45, 55, 58, 60

Li, Y., Lund, R. & Hewaarachchi, A. (2019). Multiple changepoint detection with partial information on changepoint times. *Electron. J. Statist.*, **13**, 2462–2520. 55, 60

Lu, Q. & Lund, R.B. (2007). Simple linear regression with multiple level shifts. *Canadian Journal of Statistics*, **35**, 447–458. 9, 17, 43, 44, 51, 55

Lu, Q., Lund, R. & Lee, T.C.M. (2010). An mdl approach to the climate segmentation problem. *The Annals of Applied Statistics*, **4**, 299–319. 9, 43, 45, 46, 55, 60, 61, 62, 67, 68, 69

Lund, R. & Reeves, J. (2002). Detection of undocumented changepoints: A revision of the two-phase regression model. *Journal of Climate*, **15**, 2547–2554. 55, 57, 59, 60

Lund, R., Wang, X.L., Lu, Q.Q., Reeves, J., Gallagher, C. & Feng, Y. (2007). Changepoint detection in periodic and autocorrelated time series. *Journal of Climate*, **20**, 5178–5190. 60

Maidstone, R., Hocking, T., Rigaill, G. & Fearnhead, P. (2017). On optimal multiple changepoint algorithms for large data. *Stat. Comput.*, **27**, 519–533. 61, 68

McLachlan, G.J. & Peel, D. (2004). *Finite mixture models*. John Wiley & Sons. 113

Menne, M.J. & Williams, C.N. (2009). Homogenization of temperature series via pairwise comparisons. *Journal of Climate*, **22**, 1700–1717. 47, 59

Menne, M.J., Williams, C.N. & Vose, R.S. (2009). The u.s. historical climatology network monthly temperature data, version 2. *Bulletin of the American Meteorological Society*, **90**, 993–1008. 55

Mestre, O., Domonkos, P., Picard, F., Auer, I., Robin, S., Lebarbier, É., Boehm, R., Aguilar, E., Guijarro, J., Vertachnik, G., Klancar, M., Dubuisson, B. & Stepanek, P. (2013).

HOMER : a homogenization software - methods and applications. *IDOJARAS*, **117**, 47 – 67. 55, 61

Mudelsee, M. (2019). Trend analysis of climate time series: A review of methods. *Earth-Science Reviews*, **190**, 310 – 322. 124

Myhre, G., Myhre, C. & Forster, P.a.L. (2017). Halfway to doubling of co2 radiative forcing. *Nature Geosci 10*, **10**, 710–711. 41

Nilsson, T. & Elgered, G. (2008). Long-term trends in the atmospheric water vapor content estimated from ground-based gps data. *Journal of Geophysical Research*, **113**. 48

Ning, T. & Elgered, G. (2012). Trends in the atmospheric water vapor content from ground-based gps: The impact of the elevation cutoff angle. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing - IEEE J SEL TOP APPL EARTH OBS*, **5**, 744–751. 48

Ning, T., Wickert, J., Deng, Z., Heise, S., Dick, G., Vey, S. & Schöne, T. (2016). Homogenized time series of the atmospheric water vapor content obtained from the gnss reprocessed data. *Journal of Climate*, **29**, 2443–2456. 5, 16, 28, 42, 48, 55, 101, 108, 109, 119, 135, 142

O'Gorman, P. & Muller, C. (2010). How closely do changes in surface and column water vapor follow clausius–clapeyron scaling in climate change simulations? *Environ. Res. Lett.*, **5**, 025207. 47

Olshen, A.B., Venkatraman, E.S., Lucito, R. & Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**, 557–572. 61

Ostini, L. (2012). *Analysis and Quality Assessment of GNSS-derived Parameter Time Series*. Verlag nicht ermittelbar. 142

Pacione, R., Araszkiewicz, A., Brockmann, E. & Dousa, J. (2017). EPN-repro2: A reference GNSS tropospheric data set over europe. *Atmospheric Measurement Techniques*, **10**, 1689–1705. 138

Page, E.S. (1954). CONTINUOUS INSPECTION SCHEMES. *Biometrika*, **41**, 100–115. 53

Parracho, A.C., Bock, O. & Bastin, S. (2018). Global iwv trends and variability in atmospheric reanalyses and gps observations. *Atmospheric Chemistry and Physics*, **18**, 16213–16237. 16, 42, 48, 49, 50, 101, 111, 122, 124, 142

Peterson, T.C., Easterling, D.R., Karl, T.R., Groisman, P., Nicholls, N., Plummer, N., Torok, S., Auer, I., Boehm, R., Gullett, D., Vincent, L., Heino, R., Tuomenvirta, H.,

Mestre, O., Szentimrey, T., Salinger, J., Førland, E.J., Hanssen-Bauer, I., Alexandersson, H., Jones, P. & Parker, D. (1998). Homogeneity adjustments of in situ atmospheric climate data: a review. *International Journal of Climatology*, **18**, 1493–1517. 42, 45, 53

Picard, F., Robin, S., Lavielle, M., Vaisse, C. & Daudin, J.J. (2005). A statistical approach for array cgh data analysis. *BMC Bioinformatics*, **6**, 27. 67

Planton, S. (2013). Ipcc, 2013: Annex iii: Glossary [planton, s. (ed.)]. in: Climate change 2013: The physical science basis. contribution of working group i to the fifth assessment report of the intergovernmental panel on climate change [stocker, t.f., d. qin, g.-k. plattner, m. tignor, s.k. allen, j. boschung, a. nauels, y. xia, v. bex and p.m. midgley (eds.)]. cambridge university press, cambridge, united kingdom and new york, ny, usa. https://www.ipcc.ch/site/assets/uploads/2018/02/AR5_SYR_FINAL_Annexes.pdf. 40, 41

Reeves, J., Chen, J., Wang, X.L., Lund, R. & Lu, Q.Q. (2007). A review and comparison of changepoint detection techniques for climate data. *Journal of Applied Meteorology and Climatology*, **46**, 900–915. 42, 45, 54, 55, 57

Ribeiro, S., Caineta, J. & Costa, A. (2016). Review and discussion of homogenisation methods for climate data. *Physics and Chemistry of the Earth, Parts A/B/C*, **94**, 167 – 179, 3rd International Conference on Ecohydrology, Soil and Climate Change, EcoHCC'14. 54, 55

Rigaill, G. (2015). A pruned dynamic programming algorithm to recover the best segmentations with 1 to k_max change-points. *Journal de la Société Française de Statistique*, **156**, 180–205. 61, 68

Rimoczi-Paal, A., Kerenyi, J., Mika, J., Randriamampianina, R., Dobi, I., Imecs, Z. & Szentimrey, T. (1999). Mapping daily and monthly radiation components using meteosat data. *Advances in Space Research*, **24**, 967–970. 55

Rissanen, J. (1978). Modelling by the shortest data description. *Automatica*, **14**, 465–471. 62

Rohli, R.V. & Vega, A.J. (2018). *Climatology (fourth ed.)*.. Jones Bartlett Learning.ISBN 9781284126563. 40

Ross, R.J. & Elliott, W.P. (2001). Radiosonde-based northern hemisphere tropospheric water vapor trends. *Journal of Climate*, **14**, 1602–1612. 48

Rousseeuw, P.J. & Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, **88**, 1273–1283. 22, 75, 80

Schneider, O.P.A., T. & Levine, X.J. (2010). Water vapor and the dynamics of climate changes. *Reviews of Geophysics*, **48**. 47

Schröder, M., Lockhoff, M., Forsythe, J.M., Cronk, H.Q., Vonder Haar, T.H. & Bennartz, R. (2016). The gewex water vapor assessment: Results from intercomparison, trend, and homogeneity analysis of total column water vapor. *Journal of Applied Meteorology and Climatology*, **55**, 1633–1649. 48

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464. 61, 68

Seidel, D.J., Angell, J.K., Christy, J., Free, M., Klein, S.A., Lanzante, J.R., Mears, C., Parker, D., Schabel, M., Spencer, R., Sterin, A., Thorne, P. & Wentz, F. (2004). Uncertainty in Signals of Large-Scale Climate Variations in Radiosonde and Satellite Upper-Air Temperature Datasets. *Journal of Climate*, **17**, 2225–2240. 42

Seidou, O. & Ouarda, T. (2007). Recursion-based multiple changepoint detection in multivariate linear regression and application to river streamflows. *Water Resources Research*, **43**. 55, 58, 60

Seidou, O., Asselin, J.J. & Ouarda, T.B.M.J. (2007). Bayesian multivariate linear regression with application to change point models in hydrometeorological variables. *Water Resources Research*, **43**. 55, 58

Semenov, V. & Bengtsson, L. (2002). Secular trends in daily precipitation characteristics: greenhouse gas simulation with a coupled aogcm. *Climate Dynamics*, **19**, 123–140. 47

Sen, P.K. (1968). Estimates of the regression coefficient based on kendall's tau. *Journal of the American Statistical Association*, **63**, 1379–1389. 124

Sherwood, S.C., Roca, R., Weckwerth, T.M. & Andronova, N.G. (2010). Tropospheric water vapor, convection, and climate. *Reviews of Geophysics*, **48**. 47

Steigenberger, P. (2006). Reprocessing of a global gps network. 142

Stepanek, P., Zahradníček, P. & Skalák, P. (2009). Data quality control and homogenization of air temperature and precipitation series in the area of the czech republic in the period 1961–2007. *Advances in Science and Research*, **3**, 23–26. 55, 59

Szentimrey, T. (2007). Manual of homogenization software mashv3.02. *Hungarian Meteorological Service*, 65. 55

Szentimrey, T. (2008). Development of mash homogenization procedure for daily data. proceedings of the fifth seminar for homogenization and quality control in climatological databases. *WCDMP-No. 71*, 123–130. 55

Theil, H. (1992). A rank-invariant method of linear and polynomial regression analysis. 345–381. 124

Thorne, P.W., Parker, D.E., Christy, J.R. & Mears, C.A. (2005). Uncertainties in climate trends: Lessons from upper-air temperature records. *Bulletin of the American Meteorological Society*, **86**, 1437–1442. 16, 42

Titterington, D.M., Smith, A.F.M. & Makov, U.E. (1985). *Statistical analysis of finite mixture distributions*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, John Wiley & Sons, Ltd., Chichester. 113

Truong, C., Oudre, L. & Vayatis, N. (2020). Selective review of offline change point detection methods. *Signal Processing*, **167**, 107299. 53

Van Malderen, R., Pottiaux, E., Klos, A., Domonkos, P., Elias, M., Ning, T., Bock, O., Guijarro, J., Alshawaf, F., Hoseini, M., Quarello, A., Lebarbier, E., Chimani, B., Tornatore, V., Zengin Kazanci, S. & Bogusz, J. (2020). Homogenizing gps integrated water vapor time series: Benchmarking break detection methods on synthetic data sets. *Earth and Space Science*, **7**, e2020EA001121, e2020EA001121 2020EA001121. 54, 55, 57, 136

Varadhan, R. & Roland, C. (2008). Simple and globally convergent methods for accelerating the convergence of any em algorithm. *Scandinavian Journal of Statistics*, **35**, 335–353. 82

Venema, V.K.C., Mestre, O., Aguilar, E., Auer, I., Guijarro, J.A., Domonkos, P., Vertacnik, G., Szentimrey, T., Stepanek, P., Zahradnicek, P., Viarre, J., Müller-Westermeier, G., Lakatos, M., Williams, C.N., Menne, M.J., Lindau, R., Rasol, D., Rustemeier, E., Kolokythas, K., Marinova, T., Andresen, L., Acquaotta, F., Fratianni, S., Cheval, S., Klancar, M., Brunetti, M., Gruber, C., Prohom Duran, M., Likso, T., Esteban, P. & Brandsma, T. (2012). Benchmarking homogenization algorithms for monthly data. *Climate of the Past*, **8**, 89–115. 42, 45, 54, 55, 119

Vey, S., Dietrich, R., Fritsche, M., Rülke, A., Steigenberger, P. & Rothacher, M. (2009). On the homogeneity and interpretation of precipitable water time series derived from global gps observations. *Journal of Geophysical Research: Atmospheres*, **114**. 16, 42, 48, 49, 101

Vincent, L. & Gullett, D. (1999). Canadian historical and homogeneous temperature datasets for climate change analyses. *International Journal of Climatology*, **19**, 1375 – 1388. 55

Vincent, L.A. (1998). A technique for the identification of inhomogeneities in canadian temperature series. *Journal of Climate*, **11**, 1094–1104. 55, 57, 60

Wan, H., Wang, X. & Swail, V. (2007). A quality assurance system for canadian hourly pressure data. *Journal of Applied Meteorology and Climatology*, **46**, 1804–1817. 42

Wan, H., Wang, X.L. & Swail, V.R. (2010). Homogenization and trend analysis of canadian near-surface wind speeds. *Journal of Climate*, **23**, 1209–1225. 42

Wang, C. (2018). A review of ENSO theories. *National Science Review*, **5**, 813–825. 40

Wang, C.H.L., J. & Carlson, D.J. (2001). Water vapor variability in the tropical western pacific from 20-year radiosonde data. *Journal of Climate - J CLIMATE*, **18**, 752–766. 48

Wang, J. & Zhang, L. (1998). Climate applications of a global, 2-hourly atmospheric precipitable water dataset derived from igs tropospheric products. *J. Geod.*, **83**, 209–217. 48

Wang, J., Zhang, L., Dai, A., Immler, F., Sommer, M. & Vömel, H. (2012). Radiation dry bias correction of vaisala rs92 humidity data and its impacts on historical radiosonde data. *Journal of Atmospheric and Oceanic Technology*, **30**, 197–214. 48

Wang, X. (2003). Comments on "detection of undocumented changepoints: A revision of the two-phase regression model". *Journal of Climate - J CLIMATE*, **16**. 55

Wang, X.L. (2008a). Accounting for autocorrelation in detecting mean shifts in climate data series using the penalized maximal t or f test. *Journal of Applied Meteorology and Climatology*, **47**, 2423–2444. 55, 59, 109

Wang, X.L. (2008b). Penalized maximal f test for detecting undocumented mean shift without trend change. *Journal of Atmospheric and Oceanic Technology*, **25**, 368–384. 55, 59

Wang, X.L., Wen, Q.H. & Wu, Y. (2007). Penalized maximal t test for detecting undocumented mean change in climate data series. *Journal of Applied Meteorology and Climatology*, **46**, 916–931. 48, 55, 59

Wang, X.L., Chen, H., Wu, Y., Feng, Y. & Pu, Q. (2010). New techniques for the detection and adjustment of shifts in daily precipitation data series. *Journal of Applied Meteorology and Climatology*, **49**, 2416–2436. 55, 57, 59

Willett, K.M., Jones, P.D., Gillett, N.P. & Thorne, P.W. (2008). Recent changes in surface humidity: Development of the hadcruh dataset. *Journal of Climate*, **21**, 5364–5383. 42

Williams, S.D.P. (2003). Offsets in global positioning system time series. *Journal of Geophysical Research: Solid Earth*, **108**. 50

Yao, Y.C. & Au, S.T. (1989). Least-squares estimation of a step function. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, **51**, 370–381. 68

Zhang, N.R. & Siegmund, D.O. (2007). A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, **63**, 22–32. 21, 61, 68, 69, 81