Detecting changes in mixed-sampling rate data sequences

Lowther, A.P.[1] and Killick, R.[1] and Eckley, I.A.[1*]

July 26, 2022

Abstract

Different environmental variables are often monitored using different sampling rates; exampls include half-hourly weather station measurements, daily CO_2 data, and six-day satellite data. Further when researchers want to combine the data into a single analysis this often requires data aggregation or down-scaling. When one is seeking to identify changes within multivariate data, the aggregation and/or down-scaling processes obscure the changes we seek. In this article we propose a novel changepoint detection algorithm which can analyse multiple time series for co-occurring changepoints with potentially different sampling rates, without requiring preprocessing to a standard sampling scale. We demonstrate the algorithm on synthetic data before providing an example identifying simultaneous changes in multiple variables at a location on the Greenland ice sheet using synthetic apature radar (SAR) and weather station data.

Keywords: changepoints, segmentation, multi-frequency, multivariate.

1 Introduction

Changepoint analysis has had a substantial impact in the environmental sciences in recent years. For example, changepoint methods have been used to study animal movement (Gurarie et al., 2009; Evans et al., 2016), changes in climate variables (Beaulieu et al., 2012; Gallagher et al., 2013; Beaulieu and Killick, 2018) and vegetation phenology (Richardson et al., 2018; Xie and Wilson, 2020). Loosely speaking, changepoints are the locations within a data sequence where we observe a change in the statistical properties of the sequence, e.g. a change in mean and/or variance.

The canonical, univariate multiple-changepoint problem has been studied extensively (see e.g. Csörgő and Horváth (1997), Chen and Gupta (2012) and references therein). It involves finding both the number, M, and locations, $\boldsymbol{\tau} = [\tau_1, \ldots, \tau_M]$, of the changepoints in a single sequence. Given a sequence,

[*] Corresponding author: r.killick@lancaster.ac.uk.

[1] Department of Mathematics and Statistics, Lancaster University, Lancaster, U.K.

$\{y_t\}_{t=1}^{T}$ a popular approach used to determine the number and location of multiple univariate changepoints, solves the following optimisation problem (Jackson et al., 2005),

$$\min_{M,\boldsymbol{\tau}} \left\{ \sum_{m=1}^{M+1} \left[\mathcal{C}\left(y_{(\tau_{m-1}+1):\tau_m} \right) \right] + \beta f(M) \right\}. \tag{1}$$

Here, $\mathcal{C}$ is the cost of a segment, which can take a parametric or non-parametric form. The key is that the minimum cost, when considered across different segmentations, identifies the most appropriate segmentation for the data. Common cost functions include negative log-likelihoods, ranks, and cumulative sums for changes in mean. For a recent review of methods to detect changes in mean see Cho and Kirch (2020) and for more general changepoint methods see Truong et al. (2020). The changepoints are ordered such that $\tau_i < \tau_j \iff i < j$ and we define $\tau_0 = 0$, and $\tau_{M+1} = T$ for convenience.

Jackson et al. (2005) solve the *multiple-changepoint problem* (1) using dynamic programming (Bellman and Dreyfus, 2015). Killick et al. (2012) proposed a more computationally efficient dynamic programming algorithm with *pruning* to solve (1). The consistency of the approach proposed by Killick et al. (2012) was subsequently established by Fisch et al. (2019), who also show how pruning can be implemented when considering a minimal segment length. More recently, activity within the changepoint literature has focused on the multivariate sequence setting, either assuming changes occur contemporaneously (Wang et al., 2019; Grundy et al., 2020) or allowing changes to vary from one sequence to the next (Cho and Fryzlewicz, 2015; Wang and Samworth, 2018; Hahn et al., 2020; Tickle et al., 2020). The conditions in these settings require data to be observed across all sequences at the same time.

It is not uncommon for environmental multivariate data sequences to be observed on different time scales. This difference in sampling schemes could be due to the cost of collecting data for a number of sequences, or because the sequences originate from different sources. Consider the example shown in Figure 1. Sequences 1-3 (from top to bottom) have a daily sampling rate and are obtained from data recorded from an automatic weather station (AWS) located in the ablation area of the Greenland ice sheet (van As et al., 2011). In contrast, Sequence 4 tracks how a single pixel in images captured by the Sentenial-1 satellite varies, using synthetic aperature radar (SAR) backscatter, with a sampling rate of six days. The sequences shown in Figure 1 appear to exhibit change contemporaneously, but existing multivariate methods are only applicable if each sequence is sampled at the same rate. We could achieve the same sampling rate by aggregation or averaging the daily sequences. However, such processing of the data is not desirable for a number of reasons. Firstly, this reduces the accuracy of the changepoint location estimates due to the reduced sampling rate of the transformed sequences. In addition to this, the transformation applied may be arbitrary, and consequently might render the resulting segmentation uninterpretable. Alternatively, attempts could be made to disaggregate the sequences with the lowest sampling rates. However, this is not a trivial procedure, especially when changes are present.

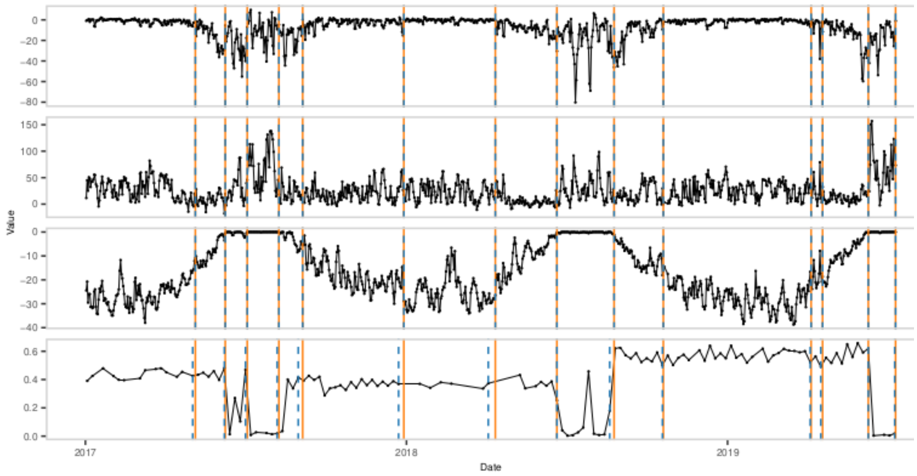


Figure 1: Sequences 1-3 (top to bottom, Latent Heat Flux (W/m^2), Sensible Heat Flux (W/m^2), Surface Temperature ($^\circ C$) show variables calculated from data collected from an automatic weather station (AWS). Sequence 4 shows the SAR evolution of a single pixel from images captured from the Sentinel-1 satellite for an area close to the AWS. The dashed (blue) vertical lines indicate the possible location of changepoints. The solid (orange) vertical lines indicate locations that can be used to segment all sequences to obtain the same segmentations derived from the changepoints.

In this article, we propose a novel changepoint approach that can detect multiple contemporaneous changes in multiple sequences where each sequence may have a different sampling rate. Our approach can be applied directly to the observed data, meaning that results are interpretable and are not affected by arbitrary transformations.

The remainder of this article is organised as follows. In Section 2 we define a novel variation of the classical changepoint problem to the mixed sampling rate (MSR) setting, and the corresponding multivariate cost function. In Section 2.1 we show that the MSR changepoint problem can be solved exactly using dynamic programming. In Section 3 we present a simulation study that demonstrates our approach can both detect the number and location of changes accurately and performs favourably in comparison to traditional approaches that might be adopted. In Section 4 we apply our approach to the data presented in Figure 1 and provide the motivation for combining these data that have different sources of origin. Finally, we conclude this article with a discussion in Section 5.

2 Changepoint detection within mixed sampling rate sequences

In this section we introduce the mixed sampling rate problem that motivates our work, and also define what we mean by a *changepoint* in such a setting. We then propose a novel cost function that incorporates the mixed sampling setting from each of the sequences. Finally, we propose an optimisation problem that can be solved to determine both the number and locations of change.

We begin by considering the collection of P sequences,

$$\left\{ \{y_{1,t}\}_{t=1}^{T_1}, \ldots, \{y_{P,t}\}_{t=1}^{T_P} \right\}, \tag{2}$$

where the number of observations for the p^{th} sequence is given by T_p. We denote the ordered time indices of sequence p as $\mathcal{T}_p = \left\{ t_{p,1}, \ldots, t_{p,T_p} \right\}$ where $|\mathcal{T}_p| = T_p$ and denote a *segment* as $y_{p,(u,v]} = \{y_{p,t} : t \in (u, v]\}$. Our work is motivated by a *collection of sequences* where each sequence may have a *unique sampling rate*, leading us to the following definition.

Definition 2.1. Let $R_1, \ldots, R_P$ denote the sampling rate of the collection of sequences given in (2). Then, a collection of sequences are said to be mixed sampling rate (MSR) sequences if

$$\exists (p_i, p_j) \in \{1, 2, \ldots, P\} \times \{1, 2, \ldots, P\} \ s.t. \ R_{p_i} \neq R_{p_j}.$$

Traditionally, a changepoint is the last observation of the previous segment. In reality, the change is somewhere between the *changepoint* and the next observation, but the location is unknown. We distinguish between *changepoints* and *MSR changepoints* as follows. The MSR changepoints appropriately segment all MSR sequences, i.e., they occur between the estimated changepoint location and the next observation in all sequences. The solid vertical (orange) lines in Figure 1 indicate these MSR changepoint locations. We denote the MSR changepoints as $\boldsymbol{\tau} = [\tau_1, \ldots, \tau_M]$. Throughout this article we assume that the changepoints, which appropriately segment each sequence, can be deduced from MSR changepoints in the following way,

$$\tau_{p,m} = \max \left\{ t \in \mathcal{T}_p : t \leq \tau_m \right\} \quad \text{for } m = 1, \ldots, M, \ p = 1, \ldots, P. \tag{3}$$

The dashed vertical (blue) lines in Figure 1 indicate the changepoint locations within each series, i.e., the last observed locations at, or prior to, the MSR changepoints. Note that, for Sequences 1-3, the changepoints and MSR changepoints coincide. However, this is not true for Sequence 4 as not all of the proposed MSR changepoint locations coincide with an observation of Sequence 4. In Section 1 we noted that, traditionally, τ_0 is the time index immediately prior to the first observation, i.e., it is 0 if the observations are at $1, 2, \ldots, T$ and $\tau_{M+1} = T$. In the MSR sequence setting, where the observed times across all sequences may not be regular, we set $\tau_0 = \min_p \{t_{p,1}\} - \epsilon = t_{\min}$ where $\epsilon > 0$ is some user-defined (small) value, and $\tau_{M+1} = \max_p \{t_{p,T_p}\} = t_{\max}$.

We next consider how to combine information from MSR sequences to determine the location of MSR changepoints. Consider the following cost function,

$$\mathcal{C}\left(\mathbf{y}_{(u,v]}\right) = \sum_{p=1}^{P} \mathcal{C}_p\left(y_{p,(u,v]}\right), \tag{4}$$

where $\mathbf{y}_{(u,v]} = \left\{\{y_{1,t}\}_{t\in(u,v]}, \dots, \{y_{P,t}\}_{t\in(u,v]}\right\}$. Equation (4) combines the information available in all sequences, giving a single measure of fit, and allows a unique cost function to be specified for each sequence. However, the cost function given in (4) may not be suitable when $y_{p,(u,v]} = \emptyset$ for some interval $(u, v]$. Further, if a cost function is based on the log-likelihood, it may be necessary to specify a minimum segment length to ensure there is enough data in a segment to estimate the likelihood parameters. More specifically, denote the minimum segment length of the p^{th} sequence L_p, then a cost function may not be well-defined when $\left|y_{p,(u,v]}\right| < L_p$ for some interval $(u, v]$, as for example we may have more parameters than data. A natural, *naive* solution to this problem may be to limit the set of *feasible* MSR changepoint locations. That is, to consider only the set of feasible changepoint vectors given by

$$\left\{[\tau_0, \tau_1, \dots, \tau_{M+1}] : \left|y_{p,(\tau_m,\tau_{m+1}]}\right| \geq L_p \text{ for } p = 1, \dots, P \text{ and } m = 0, \dots, M\right\}.$$

However, this *naive* solution could reduce the accuracy of the MSR changepoint estimates considerably, particularly when $\max_p R_p - \min_p R_p$ is large. With these considerations in mind, we propose the following cost function,

$$\tilde{\mathcal{C}}_p\left(y_{p,(u,v]}\right) = \begin{cases} \mathcal{C}_p\left(y_{p,(u,v]}\right) & \text{if } \left|y_{p,(u,v]}\right| \geq L_p \\ 0 \quad \text{otherwise} \end{cases} \qquad \text{for } p = 1, \dots, P. \tag{5}$$

Here, the cost of a segment is zero if the length of the segment is less than the minimum segment length. The set of feasible MSR changepoint vectors are now given by

$$\begin{aligned}\left\{[\tau_0, \tau_1, \dots, \tau_{M+1}] : \exists\, p \in \{1, \dots, P\} \text{ where } \left|y_{p,(\tau_m,\tau_{m+1}]}\right| \geq L_p \\ \text{for } p = 1, \dots, P \text{ and } m = 0, \dots, M\right\}.\end{aligned} \tag{6}$$

Here, we ensure that a location can be considered as a possible MSR changepoint provided that, for at least one sequence, the number of observations in the most recent segment is at least the minimum segment length of this sequence.

Thus far, we have introduced MSR changepoints and a cost function that could be used to evaluate the cost of segmenting MSR sequences. We now turn our attention to determining the number and locations of MSR changepoints. We propose using our cost function (5) in the multiple changepoint problem (1). This leads to the following problem, which we call the *Mixed Sampling Rate (MSR) Changepoint Problem*,

$$\min_{\boldsymbol{\tau}, M} \left\{ \sum_{m=1}^{M+1} \left[\tilde{\mathcal{C}}\left(\mathbf{y}_{(\tau_{m-1},\tau_m]}\right) + \beta\right] \right\}, \tag{7}$$

Here, $\tilde{\mathcal{C}}(\cdot) = \sum_{p=1}^{P} \tilde{\mathcal{C}}_p(\cdot)$ and the MSR penalty, $\beta = \sum_{p=1}^{P} \beta_p$ is given by the sum of penalties specified for each sequence. Consequently, each time an MSR changepoint is added, the objective value in (7) is penalised by a contribution from each sequence. Note, (7) allows the penalty contribution for each sequence to be unique. This is appropriate as popular penalties, e.g., SIC (Schwarz, 1978) and mBIC (Zhang and Siegmund, 2007), have penalty values that depend on the length of a sequence.

2.1 Search strategy

Now that we have clearly defined the objective function we need to optimize, we need to estimate the parameters. We first show that a dynamic program can be used to find an exact solution to the MSR changepoint problem given in (7). Our approach is, in essence, identical to optimal partitioning (OP), developed by Jackson et al. (2005). However, here we optimize our multivariate (MSR) cost function. We conclude the section by considering whether pruning might improve the computational efficiency of the procedure.

The OP approach finds an exact solution to (1) when $f(M) = M$. It works by itereratively solving increasingly complex problems, but in doing so, using information stored in previous computations, see Killick et al. (2012) for details. In order to show that the MSR changepoint problem can be solved in this way, we show that the optimal partition of data in the closed window, $[s, t]$, can be found using information in the half-open window, $[s, t)$. Let $\mathcal{P}_t = \{\boldsymbol{\tau} : t_{\min} = \tau_0 < \tau_1 \leq \ldots \leq \tau_M \leq \tau_{M+1} = t\}$ and, $F(0) = -\beta$ where $\beta = \sum_{p=1}^{P} \beta_p$. Then,

$$\begin{aligned}
F(t) &= \min_{\tau \in \mathcal{P}_t} \left\{ \sum_{m=1}^{M+1} \left[\tilde{\mathcal{C}}\left(\mathbf{y}_{(\tau_{m-1},\tau_m]}\right) + \beta \right] \right\} \\
&= \min_{\tau \in \mathcal{P}_t} \left\{ \sum_{m=1}^{M+1} \left[\sum_{p=1}^{P} \left[\tilde{\mathcal{C}}_p\left(y_{p,(\tau_{m-1},\tau_m]}\right) + \beta_p \right] \right] \right\} \\
&= \min_{s} \left\{ \min_{\tau \in \mathcal{P}_s} \left\{ \sum_{m=1}^{M} \left[\sum_{p=1}^{P} \left[\tilde{\mathcal{C}}_p\left(y_{p,(\tau_{m-1},s]}\right) + \beta_p \right] \right] \right\} + \sum_{p=1}^{P} \left[\tilde{\mathcal{C}}_p\left(y_{p,(s,t]}\right) + \beta_p \right] \right\} \\
&= \min_{s} \left\{ \min_{\tau \in \mathcal{P}_s} \left\{ \sum_{m=1}^{M} \left[\tilde{\mathcal{C}}\left(\mathbf{y}_{(\tau_{m-1},s]}\right) + \beta \right] \right\} + \tilde{\mathcal{C}}\left(\mathbf{y}_{(s,t]}\right) + \beta \right\} \\
&= \min_{s} \left\{ F(s) + \tilde{\mathcal{C}}\left(\mathbf{y}_{(s,t]}\right) + \beta \right\}.
\end{aligned} \tag{8}$$

This recursion is applied over the ordered set of observed time indices $\mathcal{T} = \bigcup_{p=1}^{P} \mathcal{T}_p$. Note, that when $P = 1$ our recursion becomes the same recursion used by Jackson et al. (2005).

Killick et al. (2012) show that for a cost function that satisfies

$$C(\mathbf{y}_{(t,s]}) + C(\mathbf{y}_{(s,T]}) + K \leq C(\mathbf{y}_{(t,T]}) \tag{9}$$

where $t < s < T$, then if

$$F(t) + C(\mathbf{y}_{(t,s]}) + K \geq F(s), \tag{10}$$

t can not be the most recent changepoint prior to T. This idea was developed further by Fisch et al. (2019) when considering segments of a minimum length. In order for the modified cost function (5) to satisfy (9), we need to consider the length of the most recent segment for each sequence. By considering s, t and T such that $|y_{p,(t,s]}| \geq L_p$ and $|y_{p,(s,T]}| \geq L_p$ for $p = 1, \ldots, P$, (9) holds and we can prove that t can be pruned following a line of proof similar to Killick et al. (2012).

3 Simulation study

We now turn to consider the performance of the MSR approach to changepoint detection in various simulated scenarios. In each scenario we compare the result of an MSR analysis against the changepoints estimated from an alternative, univariate approach, applied to each sequence in turn. We describe our general simulation framework in Section 3.1. In Section 3.2, we determine how well the MSR approach can detect a single MSR changepoint using the At Most One Change (AMOC) hypothesis test framework. Following this, Section 3.3 extends our consideration to multiple changes.

3.1 A mixed sampling rate model

Environmental datasets are often observed at different sampling rates e.g. daily, weekly, monthly and yearly. In our simulation studies, we will simulate three MSR sequences, $\boldsymbol{y}_1$, $\boldsymbol{y}_2$ and $\boldsymbol{y}_3$ with daily, monthly and yearly sampling rates respectively. These rates result in a large difference in the number of observations for a given time period. Throughout this section we will assume use of the observation window $[2000\text{-}01\text{-}01, 2020\text{-}12\text{-}17]$. This gives 7657 observations for the daily sequence, $\boldsymbol{y}_1$, with the (ordered) time indices $\mathcal{T}_1 = \{2000\text{-}01\text{-}01, \ldots, 2020\text{-}12\text{-}17\}$. We assume that the monthly sequence is observed on the last day of each month. The 251 time indices, $\mathcal{T}_2 = \{2000\text{-}01\text{-}31, 2000\text{-}02\text{-}29, \ldots, 2020\text{-}11\text{-}30\}$. Finally, we assume that the yearly sequence is observed on the final day of each year with 20 time indices, $\mathcal{T}_3 = \{2000\text{-}12\text{-}31, 2001\text{-}12\text{-}31, \ldots, 2019\text{-}12\text{-}31\}$.

We generate the observations for each sequence from a Normal distribution,

$$y_{p,t} \sim N(\mu_{p,t}, \sigma^2), \tag{11}$$

where, $\mu_{p,t} = \mu_{p,m}$ for $t \in (\tau_{m-1}, \tau_m]$, $m = 1, \ldots, M+1$, and $p = 1, 2, 3$. The MSR changepoints $\boldsymbol{\tau}$, segment means $\mu_{p,t}$, and variance σ^2, vary across simulations.

3.2 Power to detect a change

Under the AMOC framework, evidence is sought to reject the null hypothesis of no change, in favour of the alternate hypothesis of a single MSR changepoint.

The hypothesis can be stated formally as follows,

$$H_0 : \begin{cases} \mu_{p,t} = \mu_p \\ \sigma^2_{p,t} = \sigma^2_p \end{cases} \quad \forall t \in \mathcal{T}_p, \text{ for } p = 1, 2, 3, \text{ and,} \tag{12a}$$

$$H_1 : \begin{cases} \mu_{p,t_1} = \ldots = \mu_{p,t_{k_p}} \neq \mu_{p,t_{k_n+1}} = \mu_{p,t_{T_p}} \\ \sigma^2_{p,t_1} = \ldots = \sigma^2_{p,t_{k_p}} \neq \sigma^2_{p,t_{k_p+1}} = \sigma^2_{p,t_{T_p}} \end{cases} \quad \text{for } p = 1, 2, 3. \tag{12b}$$

We denote the MSR changepoint location, τ_1, such that each changepoint can be identified by

$$\tau_{p,1} = \max \{ t \in \mathcal{T}_p : t \leq \tau_1 \} \text{ for } p = 1, 2, 3.$$

We conduct the hypothesis test using the likelihood ratio test statistic

$$\Lambda = -2 \left(l_0 - l_1 \right). \tag{13}$$

Here, we use the MSR cost function to calculate the log-likelihood under the null and alternative hypothesis, given here as l_0 and l_1 respectively. We estimate the critical value by calculating the $\alpha\%$ quantile of the test statistic realisations, where data is generated without a change. We then estimate the power by calculating the proportion of times the test statistic exceeds the critical value, when data is generated with a single change. To estimate both the critical value and the power, we simulate data from the MSR model given in Section 3.1, 10000 times, with $\sigma^2 = 4$.

To estimate the critical value, all sequences are generated with mean equal to 0. To estimate the power, the MSR changepoint, $\tau_1 = 2010\text{-}06\text{-}25$ is located close to the centre of the MSR sequences and all sequences start with mean 0. Following τ_1, the daily sequence mean changes to 0.25, the monthly mean to 2 and the yearly mean to 9. We vary the size of the mean change so as to increase the difficulty in detecting the change as the length of the sequence increases. A realisation of the data used to calculate the power is given in Figure 2. The change in mean is most obvious for the yearly sequence, and arguably, the change in mean for the daily sequence is not visible by eye at this scale.

The power estimates for the MSR and univariate hypothesis tests are shown in Figure 3. The MSR test appears to have greater power than each of the univariate tests. This is because the MSR test combines information across the sequences which results in more evidence of a change. It is also less likely that spurious changes will occur simultaneously in all sequences and so false positives reduce too.

The ROC plot signifies the power we have to reject the null hypothesis of no change but does not say anything about whether the change is localized around the true value. Figure (4a) shows histograms of the changepoint locations for the tests that reject the null hypothesis when $\alpha = 0.05$. The true changepoint locations are shown by the dashed vertical (blue) lines. There are clear differences in the changepoint location estimates for the daily sequence between the MSR and univariate approaches. For the univariate approach, the distribution

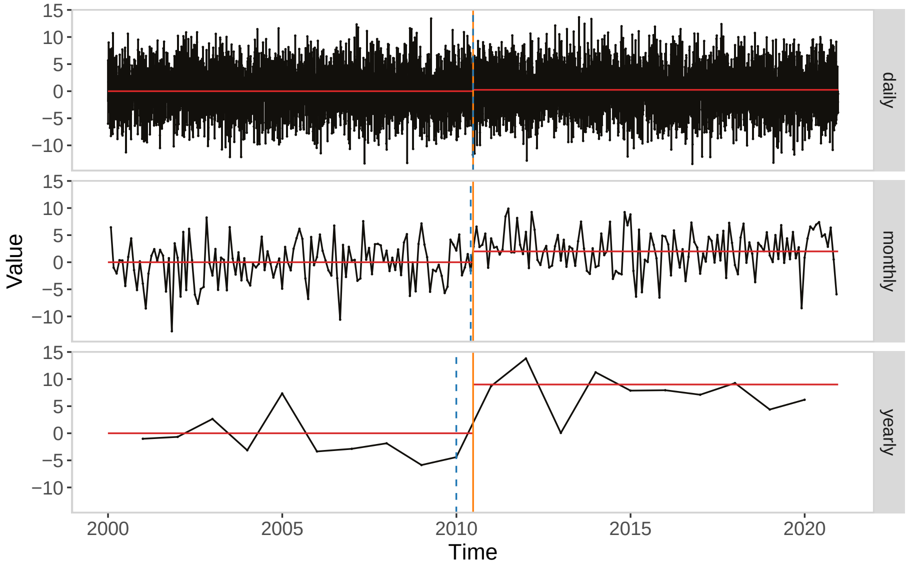


Figure 2: A realisation of data generated from the MSR model with a single change. The location of the MSR changepoint is shown by the solid vertical (orange) line. The changepoint locations are shown by the dashed vertical (blue) lines. The segment means are shown by the horizontal (red) lines.

of the location estimates appears very flat, with a very small peak centred at the true changepoint location and peaks towards the end of the observation period due to short segments. However, the estimate of the daily changepoint obtained from the MSR changepoint estimate has large peaks centred around the true location of the changepoint.

Differences in the distribution of the yearly changepoint estimate are less obvious. For both the MSR and univariate approach, the distribution is centred around the true changepoint location. The only discernible difference is the height of the peak at the true changepoint location. The height for the MSR test is approximately 8000, in comparison to 5000 for the univariate test. Figure (4b) shows the estimates of the MSR changepoint location for the tests that reject the null hypothesis. The plot contains two coloured segments located close to the centre. The widest (orange) coloured segment highlights the region up to but not including the observations either side of the changepoint for the yearly sequence. Similarly, the smaller (green) segment indicates the aforementioned region corresponding to the monthly sequence. Interestingly, the frequency of MSR location estimates falls sharply outside of the highlighted segments. It appears that the distribution of the MSR changepoint estimates is a mixture of distributions corresponding to the the changepoint estimates from the univariate approaches.

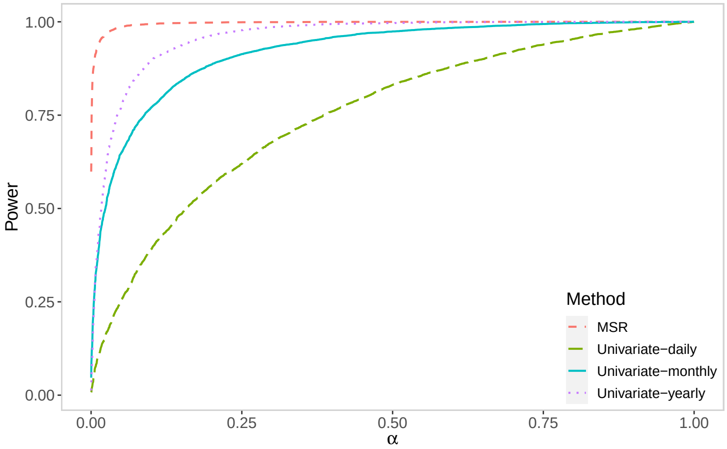
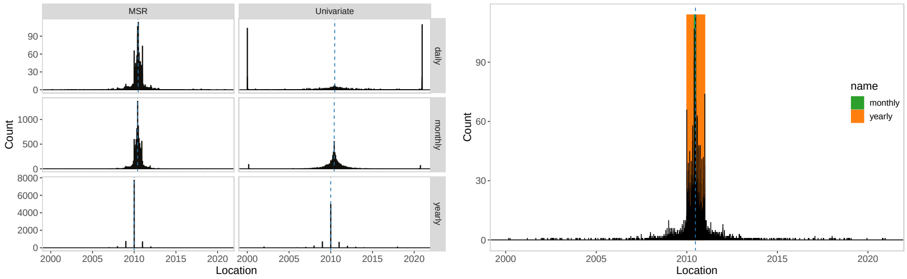


Figure 3: The Monte Carlo estimates of power in the MSR and univariate hypothesis tests.



(a) Changepoints estimates

(b) MSR changepoint estimates

Figure 4: Histograms of the location estimates corresponding to $\alpha = 0.05$.

3.3 Multiple changepoints

In this section we consider various simulated scenarios that contain multiple changes. Firstly, we aim to show that the MSR approach can accurately determine the location of the MSR changepoints. Due to the lack of existing approaches for this problem we will apply the univariate OP approach (Jackson et al., 2005) to each sequence in turn as a base comparison. For each scenario we generate 1000 repetitions from the MSR model given in Section 3.1. The cost function used for each sequence assumes that each sequence is an independent, identically distributed sample from a Normal distribution with unknown mean and unknown variance. For each sequence, we will use the BIC penalty (Schwarz, 1978), $\beta_p = 3 \log(T_p)$.

3.3.1 Changes observed in all sequences

When thinking about describing the locations of changepoints it is conventional to use an index number. However, as we have different sampling rates within the sequences, we want to use the same underlying labelling for all sequences. Thus, we use date formats in our simulations as this is a natural ordering which readers can interpret as daily/weekly/monthly/yearly observation times.

In this section, the MSR changepoint locations are, τ_1 = 2003-03-22, τ_2 = 2004-08-22, τ_3 = 2008-08-06, τ_4 = 2010-06-01, τ_5 = 2017-02-18. This ensures that there is at least one observation in each segment for each sequence. The mean alternates between 1 and -1 and the variance, $\sigma^2 = 2^2$ for each sequence. Figure 5 shows a realisation of the data. Note that segment 2 for $\boldsymbol{y}_3$ contains only one observation. Consequently, the OP approach cannot correctly identify all changepoints for y_3 as the minimum segment length of each sequence is two.

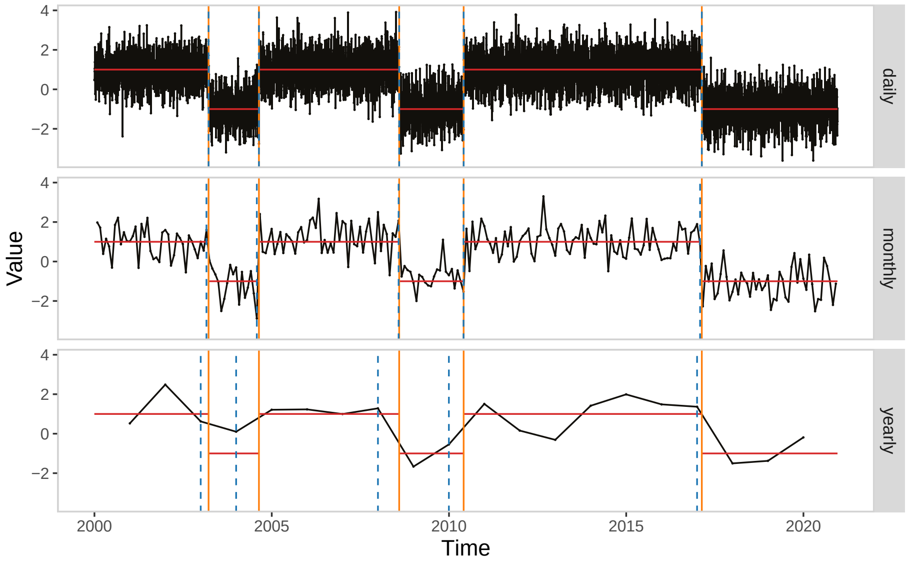


Figure 5: A realisation of the alternating change in mean example. Changes are observed in all sequences following an MSR changepoint, given by the solid vertical (orange) lines, and the changepoint locations reported in each sequence are shown by the dashed vertical (blue) lines. The solid horizontal (red) lines indicate the segment means.

Figure 6 depicts the results of applying the MSR (left) and OP (right) approaches. The columns left-to-right show the estimates for the daily, monthly and yearly sequences respectively. The histograms for the daily sequence appear similar across the two approaches, identifying changepoints at the true locations. The true changepoint locations were identified in approximatley 75%

of all simulations. Where the two approaches did not corrctly identify the true changepoint location, the changepoint estimates were very close to the true location. Due to the scale of Figure 6, it is difficult to see the side-lobes around the peaks. The MSR and OP approaches perform very differently for the monthly sequence. The MSR approach detects the true location of the changepoints in nearly all realisations of the data. In contrast, the OP-derived estimates appear to be centred around the true location of the changepoints but with higher variance. The performance of the MSR approach appears similar for both the monthly and yearly sequences. This is due to the pooling of information across series, providing more confidence in the location of the changes. However, the OP approach performs less favourably as the sequences become shorter with the performance in the yearly sequence being especially poor.

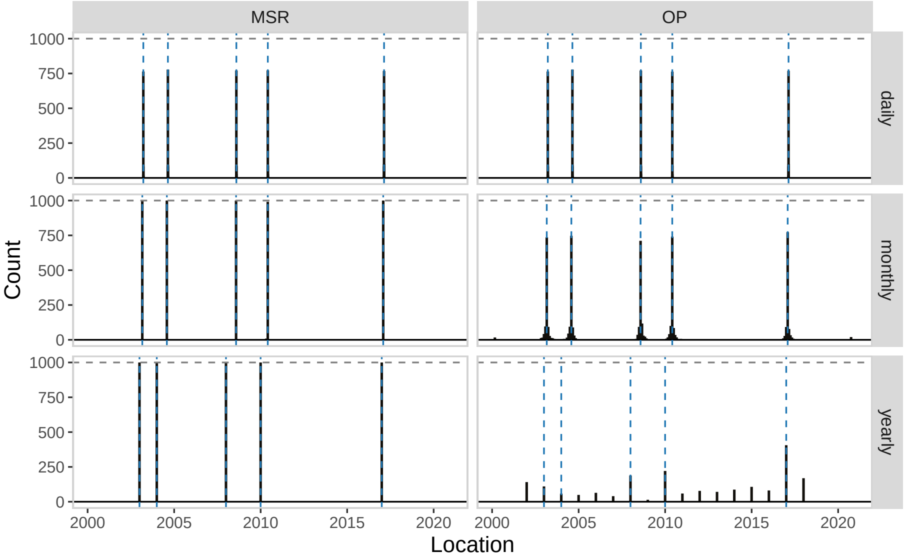


Figure 6: Histograms of the changepoint location estimates for the MSR approach (left) and the OP approach (right) for 1000 realisations of the data. Changes are observed in all sequences following an MSR changepoint and the locations are shown by the (blue) marks below the horizontal axis.

Figure 7 shows the histograms for the number of changepoints identified for each sequence for the MSR and OP approaches. The MSR approach correctly identifies the true number of changepoints for each sequence. In comparison, the OP approach correctly identifies the number of changepoints for the daily sequence in the majority of cases, but over- and then under-estimates for the monthly and yearly sequences respectively. Whilst the signal in the daily sequence may be strong enough to often detect all changes, we observe a marginal improvement in the estimate of the number of changepoints for the daily sequence using the MSR approach. Furthermore, the strength in the daily signal helps to inform accurate estimates of the changepoint locations for the monthly and yearly sequences.

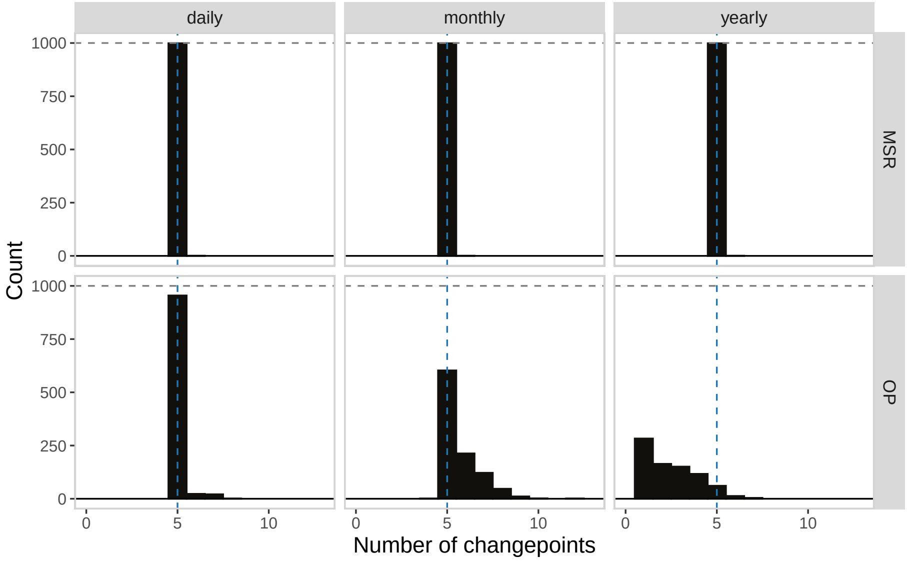


Figure 7: Histograms of the number of changepoints detected for the MSR (top) and OP approach (bottom) for 1000 realisations of the data. Changes are observed in all sequences following an MSR changepoint. The true number of changepoints are shown by the dashed vertical (blue) lines.

3.3.2 Changes not visible in some sequences

In this section we investigate the performance of the MSR approach for detecting changepoints in sequences when a change is not observed in a sequence due to the sampling rate. The MSR changepoint locations are as follows τ_1 = 2000-01-22, τ_2 = 2000-03-22, τ_3 = 2005-08-22, τ_4 = 2007-08-05, τ_5 = 2008-08-06, τ_6 = 2010-06-01, τ_7 = 2014-06-01, τ_8 = 2014-06-21 and τ_9 = 2017-02-18. For each sequence, the mean alternates between 1 and -1 and the variance, $\sigma^2 = 2^2$. A realisation of the data is given in Figure 8. Note that only $\boldsymbol{y}_1$ is observed prior to τ_1 and in the interval $(\tau_7, \tau_8]$. This means that no information from $\boldsymbol{y}_2$ or $\boldsymbol{y}_3$ can be used to detect τ_1 and τ_7 respectively.

Figure 9 depicts the histogram of the changepoint location estimates. As with the previous example, the estimate of the changepoint locations for the daily sequence appear very similar for the MSR and OP approach and the MSR changepoints are accurate for all series. In contrast, the OP estimates degrade for the monthly data and the yearly estimates show an almost random distribution.

Figure 10 shows the number of changepoints detected for each sequence.

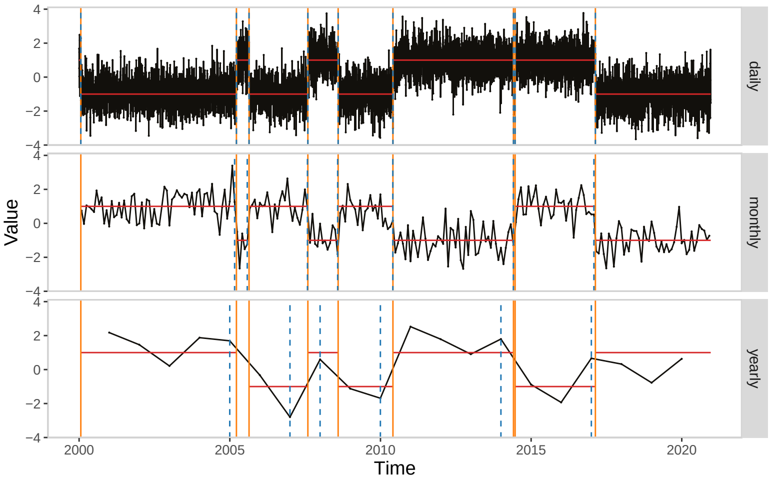


Figure 8: A realisation of the data where changes are observed in only a subset of the MSR sequences due to the sampling rate. The MSR changepoint locations are shown by the solid vertical (orange) lines. The changepoint locations are shown by the vertical dashed (blue) lines. The segment means are shown by the solid (red) horizontal lines.

Again, the MSR approach accurately determines the true number of changepoints for the daily and yearly sequence. The MSR approach appears to overestimate the number of changepoints for the monthly sequence occasionally. This may be caused by the location of MSR changepoints τ_7 and τ_8 as they are both located in June 2014. If the location estimate of τ_7 is too early, this will translate to an *extra* changepoint for the monthly sequence as it is observed on the last day of each month. Again, we observe a marginal improvement in the accuracy of the number of changepoints for the daily sequence using the MSR approach, and see strong improvements in the accuracy of the changepoint estimates for the weekly and monthly sequences.

4 Data application

We now turn to consider the utility of our proposed approach on Greenland ice sheet SAR and weather station data. The deteriorating condition of the Greenland ice sheet threatens global sea levels (Shepherd et al., 2020). As such, it is of upmost importance to closely monitor the conditions of the ice sheet. Recently, Shu et al. (2021) developed a novel classification model to distinguish between melting and non-melting states of the ice sheet using SAR backscatter. This approach uses state-space modelling to determine changes in SAR backscatter that can be attributed to the following surface types: lake, snow or ice. In this section, we apply our approach to the data presented in Figure 1 to detect

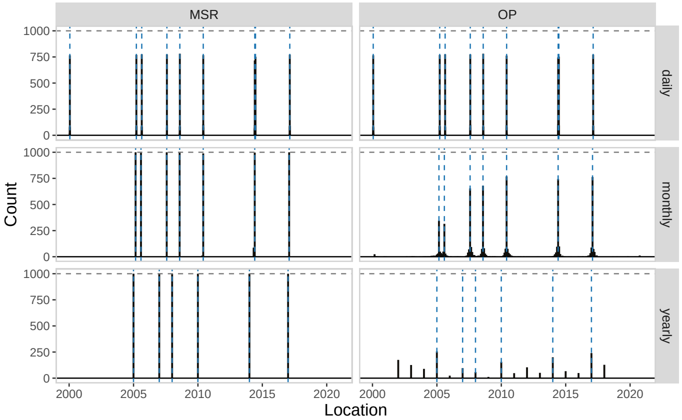


Figure 9: Histograms of the changepoint locations estimates for the MSR approach (left) and the OP approach (right) for 1000 realisation of the data. Changes are observed in only a subset of sequences following an MSR changepoint. The true changepoint locations are shown by the dashed vertical (blue) lines.

simultaneous changes in the SAR backscatter and meteorological variables. By using the meteorological variables, each having a higher sampling rate than the SAR backscatter, it may be possible to provide an earlier warning of changes to the condition of the ice sheet.

Sequences 1-3 represent three variables, *Latent Heat Flux*, *Sensible Heat Flux* and *Surface Temperature*, calculated from the Programme for Monitoring of the Greenland Ice Sheet (PROMICE) KPC_L AWS (van As et al., 2011). Sequence 4 shows the SAR backscatter data taken from a single pixel in images produced by the Sentinel-1 satellite covering the North-East sector of the Greenland ice sheet. Recall from Section 1, Sequence 4 has a sampling rate of six days, in comparison to the other sequences with a daily sampling rate.

Figure 1 shows annual variation in the AWS variables. In addition to this, the low-range surface temperature measuring device is upper-bounded by the melting point (of ice). Hence, there are periods in the summer months where the ice melts and Surface Temperature is recorded as 0^oC for prolonged periods. We detrend the sequences by applying a first difference to the AWS variables; an approach often used in the changepoint literature (Killick et al., 2013), and assume that the resulting sequences are each sampled independently from a Normal distribution with zero mean. Prior to differencing Surface Temperature, we perturb the values by adding independent noise, sampled from a $\mathrm{N}(0, 0.1^2)$ distribution, to avoid estimates of zero variance in the summer months caused by the upper bound of the measurements. We believe that the aforementioned noise added sufficiently avoids estimates of zero variance, but is subtle enough so as to not overwhelm the signal in the sequence. We assume that the SAR backscatter sequence is independently sampled from a Normal distribution with

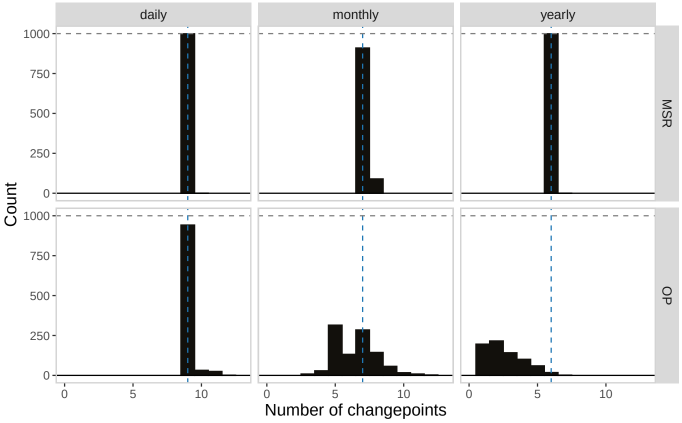


Figure 10: Histograms of the number of changepoints detected by the MSR approach (top) and OP (botttom) for 1000 realisations of the data. Changes are observed in only a subset of sequences following an MSR changepoint.

unknown mean and variance. To avoid over-fitting, we use the BIC (Schwarz, 1978) penalty for each sequence.

Figure 11 shows the analysed sequences and the resulting segmentation. Our approach detects changes in the variance of the AWS variables which coincide with changes in the mean or variance of the SAR backscatter. Few MSR changepoints coincide with observations of the SAR sequence since the MSR changepoints have daily precision and the SAR sequence has a sampling rate of six days. Consequently, the changepoints for the SAR sequence, derived from the MSR changepoints, are not co-located with the MSR changepoints; this may provide an earlier warning of imminent changes in the SAR backscatter.

5 Discussion

In this article we have proposed an optimisation based approach to detect contemporaneous changes in mean and/or variance in MSR sequences. We have demonstrated that this approach can both accurately determine the number and location of the changes. By combining information from all sequences, our approach can more accurately detect contemporaneous changes in MSR sequences in comparison to a univariate approach. In addition to this, our approach does not require the sequences to be transformed - in practise, many of which are arbitrary and the choice may vary among practitioners. In Section 3.3.2 we observed that our approach can overestimate the number of changepoints for sequences with a low sampling rate if the estimate of the MSR changepoint is incorrect. Despite this, the estimate of the number of locations for our approach appeared to be more accurate than the univariate alternative in this example.

An interesting area of future research would integrate cost functions that

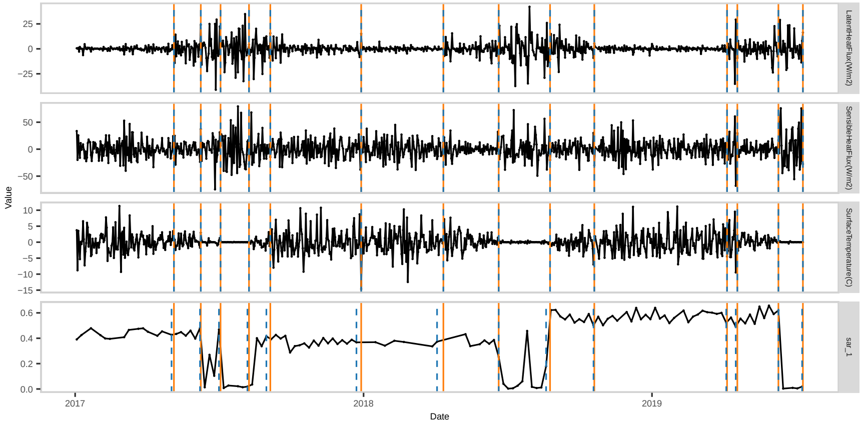


Figure 11: The first difference of the AWS variables (rows 1-3) and the SAR backscatter data. The MSR changepoints are shown by the solid vertical (orange) lines and the changepoints are shown by the dashed vertical (blue) lines.

can detect more general changes, e.g. changes in trend which will make our approach suitable to a broader audience. Our approach can also be applied to irregularly sampled sequences. The performance of the MSR approach could be studied in this setting.

6 Acknowledgement

The authors gratefully acknowledge the financial support of U.K. NERC via research grant (NE/T006102/1). We also wish to acknowledge Qingying Shu (Lancaster University) for helpful conversations regarding the Ice sheet application and Anna Hogg (Leeds University) for the original preparation of the data. Data from the Programme for Monitoring of the Greenland Ice Sheet (PROMICE) were provided by the Geological Survey of Denmark and Greenland (GEUS) at `http://www.promice.dk`.

References

Beaulieu, C., Chen, J., and Sarmiento, J. L. (2012). Change-point analysis as a tool to detect abrupt climate variations. *Phil. Trans. R. Soc. A.*, 370.

Beaulieu, C. and Killick, R. (2018). Distinguishing trends and shifts from memory in climate data. *Journal of Climate*, 31(23):9519 – 9543.

Bellman, R. E. and Dreyfus, S. E. (2015). *Applied Dynamic Programming.* Princeton University Press, Princeton.

Chen, J. and Gupta, A. K. (2012). *Parametric Statistical Change Point Analysis*. Birkh ä user Boston, second edition.

Cho, H. and Fryzlewicz, P. (2015). Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(2):475–507.

Cho, H. and Kirch, C. (2020). Data segmentation algorithms: Univariate mean change andbeyond. arXiv:2012.12814v1.

Csörgő, M. and Horváth, L. (1997). *Limit Theorems in Change-Point Analysis*. Wiley Series in Probabilty and Statistics. John Wiley & Sons.

Evans, A. L., Singh, N. J., Friebe, A., Arnemo, J. M., Laske, T. G., Fröbert, O., Swenson, J. E., and Blanc, S. (2016). Drivers of hibernation in the brown bear. *Frontiers in Zoology*, 13(7).

Fisch, A. T. M., Eckley, I. A., and Fearnhead, P. (2019). A linear time method for the detection of point and collective anomalies. arXiv:1806.01947v2.

Gallagher, C., Lund, R., and Robbins, M. (2013). Changepoint detection in climate time series with long-term trends. *Journal of Climate*, 26(14):4994 – 5006.

Grundy, T., Killick, R., and Mihaylov, G. (2020). High-dimensional changepoint detection via a geometrically inspired mapping. *Statistics and Computing*, 30:1155–1166.

Gurarie, E., Andrews, R. D., and Laidre, K. L. (2009). A novel method for identifying behavioural changes in animal movement data. *Ecology Letters*, 12(5):395–408.

Hahn, G., Fearnhead, P., and Eckley, I. A. (2020). Bayesproject: Fast computation of a projection direction for multivariate changepoint detection. *Statistics and Computing*, 30:1691–1705.

Jackson, B., Scargle, J. D., Barnes, D., Arabhi, S., Alt, A., Gioumousis, P., Gwin, E., Sangtrakulcharoen, P., Tan, L., and Tun Tao Tsai (2005). An algorithm for optimal partitioning of data on an interval. *IEEE Signal Processing Letters*, 12(2):105–108.

Killick, R., Eckley, I. A., and Jonathan, P. (2013). A wavelet-based approach for detecting changes in second order structure within nonstationary time series. *Electronic Journal of Statistics*, 7:1167 – 1183.

Killick, R., Fearnhead, P., and Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598.

Richardson, A. D., Hufkens, K., Milliman, T., Aubrecht, D. M., Chen, M., Gray, J. M., Johnston, M. R., Keenan, T. F., Klosterman, S. T., Kosmala, M., Melaas, E. K., and Friedl, Mark A. andFrolking, S. (2018). Tracking vegetation phenology across diverse north american biomes using phenocam imagery. *Scientific Data*.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464.

Shepherd, A., Ivins, E., Rignot, E., Smith, B., van den Broeke, M., Velicogna, I., Whitehouse, P., Briggs, K., Joughin, I., Krinner, G., Nowicki, S., Payne, T., Scambos, T., Schlegel, N., A, G., Agosta, C., Ahlstrøm, A., Babonis, G., Barletta, V. R., Bjørk, A. A., Blazquez, A., Bonin, J., Colgan, W., Csatho, B., Cullather, R., Engdahl, M. E., Felikson, D., Fettweis, X., Forsberg, R., Hogg, A. E., Gallee, H., Gardner, A., Gilbert, L., Gourmelen, N., Groh, A., Gunter, B., Hanna, E., Harig, C., Helm, V., Horvath, A., Horwath, M., Khan, S., Kjeldsen, K. K., Konrad, H., Langen, P. L., Lecavalier, B., Loomis, B., Luthcke, S., McMillan, M., Melini, D., Mernild, S., Mohajerani, Y., Moore, P., Mottram, R., Mouginot, J., Moyano, G., Muir, A., Nagler, T., Nield, G., Nilsson, J., Noël, B., Otosaka, I., Pattle, M. E., Peltier, W. R., Pie, N., Rietbroek, R., Rott, H., Sandberg Sørensen, L., Sasgen, I., Save, H., Scheuchl, B., Schrama, E., Schröder, L., Seo, K.-W., Simonsen, S. B., Slater, T., Spada, G., Sutterley, T., Talpe, M., Tarasov, L., van de Berg, W. J., van der Wal, W., van Wessem, M., Vishwakarma, B. D., Wiese, D., Wilton, D., Wagner, T., Wouters, B., Wuite, J., and The IMBIE Team (2020). Mass balance of the greenland ice sheet from 1992 to 2018. *Nature*, 579(7798):233–239.

Shu, Q., Killick, R., Leeson, A., Nemeth, C., Fettweis, X., Hogg, A., and Leslie, D. (2021). Characterising the ice sheet surface in north east greenland using sentinel-1 sar data. *Submitted*.

Tickle, S. O., Eckley, I. A., and Fearnhead, P. (2020). A computationally efficient, high-dimensional multiple changepoint procedure with application to global terrorism incidence. arXiv:2011.03599v1.

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

van As, D., Fausto, R. S., and PROMICE project team, . (2011). Programme for monitoring of the greenland ice sheet (promice): first temperature and ablation records. *GEUS Bulletin*, 23:73–76.

Wang, R., Volgushev, S., and Shao, X. (2019). Inference for change points in high dimensional data. Submitted to Annals of Statistics. arXiv:1905.08446v1.

Wang, T. and Samworth, R. J. (2018). High dimensional change point estimation via sparse projection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):57–83.

Xie, Y. and Wilson, A. M. (2020). Change point estimation of deciduous forest land surface phenology. *Remote Sensing of Environment*, 240:111698.

Zhang, N. R. and Siegmund, D. O. (2007). A modified bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, 63(1):22–32.