atmosphere

MDPI

Supplementary Materials: Concurrent Changepoints in Greenland Ice Core $\delta^{18}O$ records and the North Atlantic Oscillation over the Past Millennium

István Gábor Hatvani [1], Dániel Topál [1,2], Eric Ruggieri [3] and Zoltán Kern [1,*]

S1. What is a changepoint?

Long time series are often heterogeneous in nature. As such, the most appropriate model may be one whose parameters are allowed to change through time. The point at which the statistical properties of a model change is called a "change point." Some of the earliest approaches to change point analysis include the CUSUM statistic, which monitors the cumulative sum of the residuals to see if they exceed some threshold, and binary segmentation [1], which repeatedly splits a data set and checks each remaining piece for additional change points until no further change points are detected. Examples of more recent approaches include MCMC techniques (e.g. Green [2]), which add or delete one change point at a time, and particle filters (e.g. Fearnhead and Liu [3]), where each particle represents one possible state (e.g. set of change points) for the system.

To give a concrete example, suppose that we are trying to model the mean signal in a climate system. In this scenario, the model of interest is the constant model, $Y = \mu + \epsilon$. If either the mean or variance of this model changes at any point in the time series, a change point exists [4]. Instead, if a linear (e.g. trend) model is appropriate, $Y = \beta_0 + \beta_1 x + \epsilon$, then a change in the slope (β_1), intercept (β_0), or variance of the error terms (ϵ) would indicate a change point in the data.

To give a concrete example, suppose that we have a piecewise constant function

$$Y_{1:30} = 4 + N(0, 0.08)$$

$$Y_{31:50} = 1 + N(0, 0.08)$$

where $N(0, 0.08)$ denotes a random sample from a normal distribution with mean 0 and variance 0.08, i.e. random variability around the mean (Figure S1a). In this example, data point 30 represents the "changepoint" in the model, as it is the point where the mean of the function changes from 4 to 1.

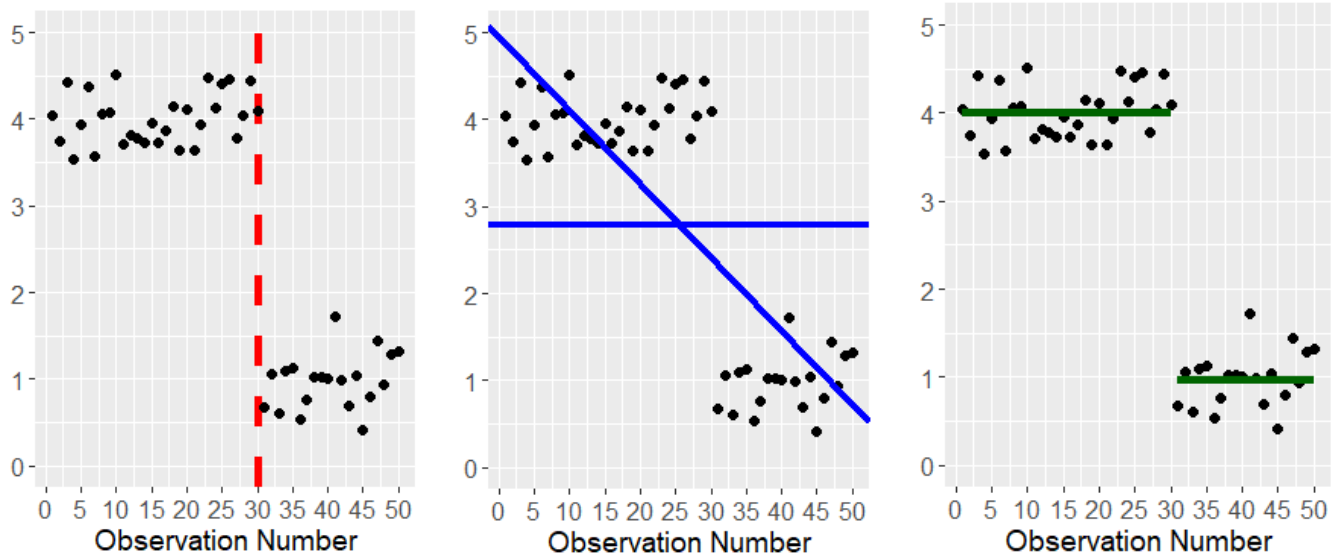


Figure S1. Simulated data (**a),** with change point indicated as a vertical dotted line (**b)** and an attempt to fit the data with a single function, either constant or linear (**c).** Fitting a piecewise function to the data.

Neither the constant nor a linear model appears to fit the simulated data well (Figure S1b), as both fail to capture the piecewise nature of the function. In this case, adding a changepoint to the model vastly improves its fit (Figure S1c). Once the location of the change point has been identified, we can estimate the parameters of the model separately for each interval. The change point in this example is obvious and easily identified by eye. However, changes in a time series that are more subtle require the aid of a change point model to be able to identify if and when they occur (Figure S2). In this example the first two change points are relatively easy to identify by eye (negative to a positive slope, change to a larger positive slope), but the third change point is less obvious. One potential model is the Bayesian Change Point algorithm of Ruggieri [5] (BCPa for short), an approach, which can generate the posterior distribution on both the number and location of changepoints in a data set.

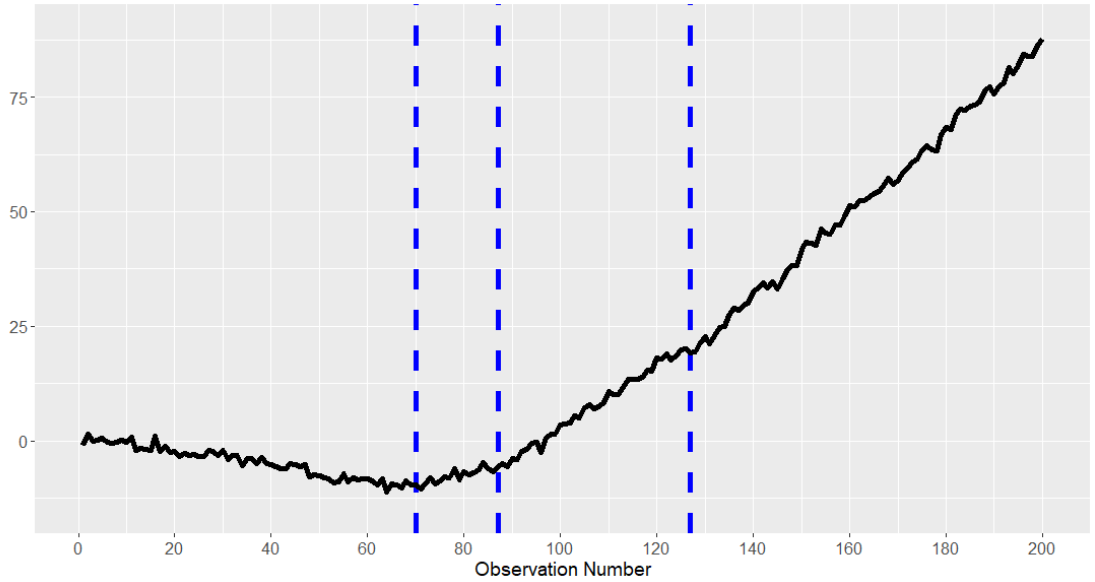


Figure S2. Simulated data with three change points indicated by dotted vertical lines at positions 70, 87, and 127. The slope of the line in each interval is -0.167, 0.261, 0.666, and 0.915, respectively.

A brute force approach to identify the "optimal" placement of k change points among N data points quickly becomes intractable as there are ${}_N C_k$ ways to place k change points among N data points. As a result, a number of change point models have been developed to work through this computational challenge [including, but not limited to CUSUM (e.g. Zeileis et al. [7]), binary segmentation (e.g. Scott and Knott [1]), MCMC (e.g. Green [2]) and particle filter approaches (e.g. Fearnhead and Liu [3])]. Here, we describe a Bayesian version of a change point model that incorporates dynamic programming recursions to piece together the different climate regimes in a computationally efficient way. BCPa assumes that the parameters of the model for any two segments of the data are independent (i.e. a product partition model [6]) and that the error terms are uncorrelated $N(0, \sigma^2)$ random variables. The model returns not only the posterior distribution on the number and location of change points in the time series (which gives us probabilistic bounds on their location), but also estimates of the parameters of the model between any two change points.

Recall that BCPa has three steps:

1) Calculate the probability of the data for all possible climate regimes in the data set

2) Recursively piece together the climate regimes identified in step (1), adding one change point at a time

3) Sample from the posterior distribution on the number and location of change points, as well as the parameters of the model in each segment.

If we use a jigsaw puzzle as an analogy, step (1) has us flip all of the pieces over so that the picture faces up and step (2) has us put the puzzle together one piece at a time. In step (3), we look back at the completed puzzle and analyze how the pieces fit together.

S2. Explanation of the parameters applied in the analysis

k_0 is a scale parameter that relates the variance of the regression parameters to the residual variance. The practical effect is to act as a "penalty" against adding change points, where a smaller value of k_0 allows for larger values of the regression parameters (relative to the error variance), but also gives a larger penalty on introducing a change point. Allowing for large values of the regression parameters is especially important for the constant term in a long time series, as its value can differ significantly from zero.

v_0 and σ_0 act as pseudo-data for estimating the value of the residual variance, $v_0$pseudo-data points of variance σ_0. Setting v_0 equal to 1 and σ_0 equal to the variance of the

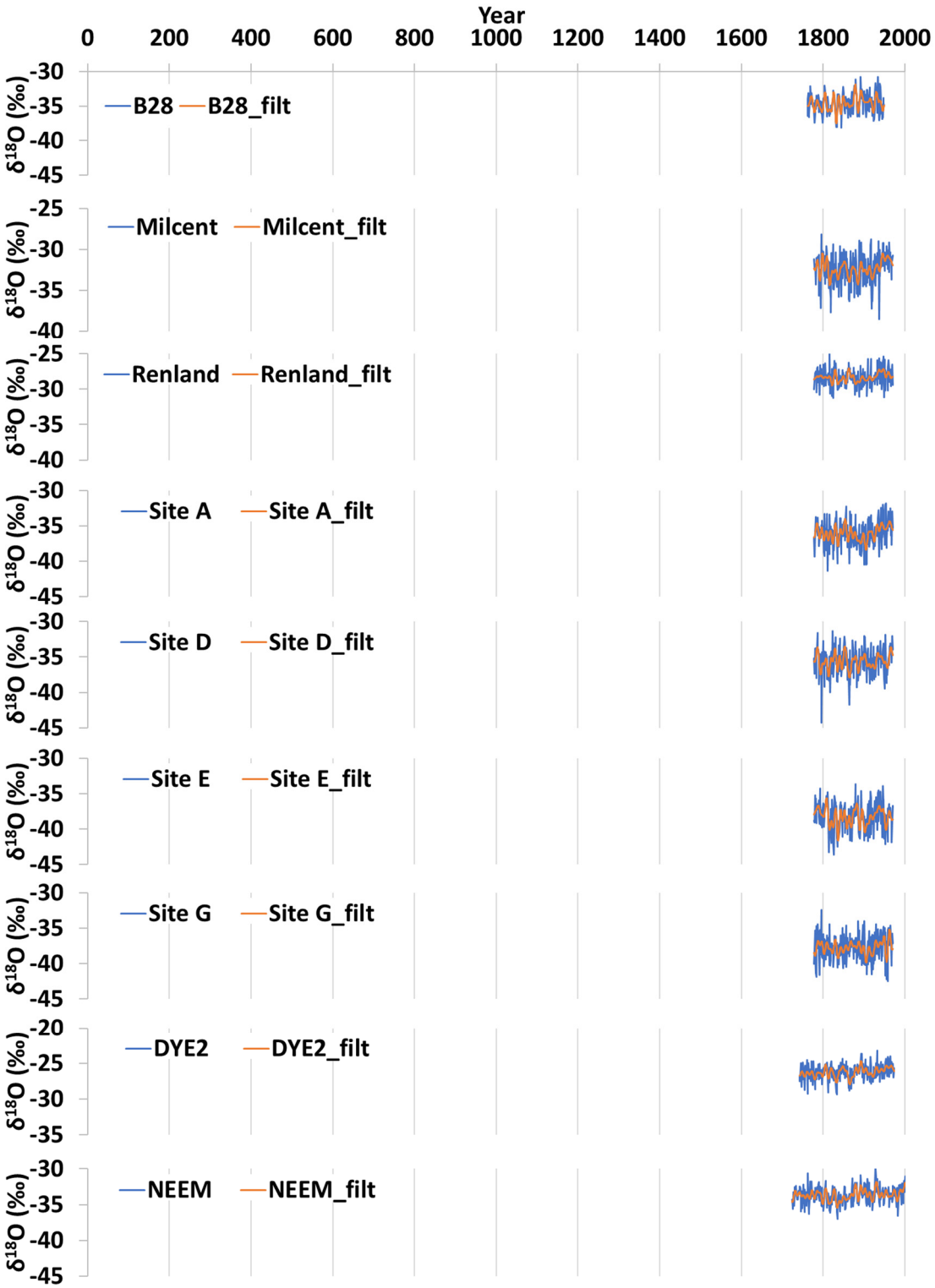
data implies that we have one prior observation of the residual error whose magnitude is equal to the variance of the data.

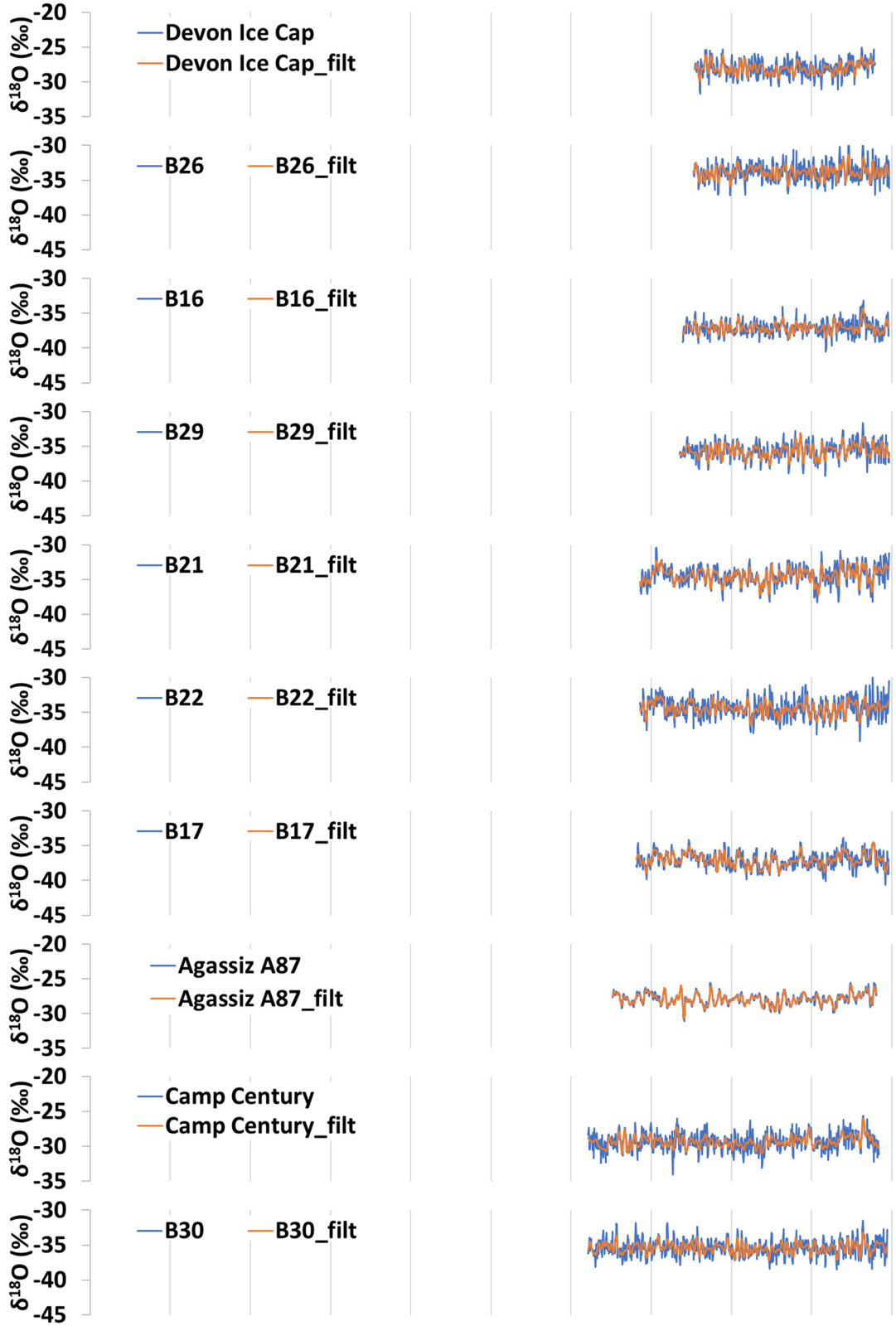
d_{min}, the minimum distance between two consecutive change points, can be set to any reasonable value for the problem of interest, but in general should be at least twice as large as the number of regression parameters that need to be estimated.

k_{max} represents the maximum number of allowed change points in the time series. The value of k_{max} should be at least as large as the expected maximum number of change points, but need not be any larger than n/d_{min}, where n is the number of observations in the data set.

num.samp is the number of sampled solutions from the posterior distribution on the number and location of change points, as well as the parameters of the regression model fit between any two change points. Larger values of *num.samp* allow for a more accurate estimate of each quantity.

As with any type of Bayesian analysis, the results can be sensitive to the choice of the prior distribution. In this situation, our prior parameters were chosen to represent a minimal amount of prior information, so as to let the actual data drive the inference. More generally, this particular model can be sensitive to the choice of the k_0, v_0, and σ_0 parameters. When this occurs, the model may display changes in the number, but not the location or distribution of the change points. In other words, if the model needs a certain amount of "evidence" to declare that a change point exists, then the choice of these parameters can alter the threshold for detection, but will not shift the location of a change point from one section of the data to another.


Year
0
200
400
600
800
1000
1200
1400
1600
1800
2000
δ18O (‰)
B28
B28_filt
Milcent
Milcent_filt
Renland
Renland_filt
Site A
Site A_filt
Site D
Site D_filt
Site E
Site E_filt
Site G
Site G_filt
DYE2
DYE2_filt
NEEM
NEEM_filt


δ18O (‰)
-20
-25
-30
-35
Devon Ice Cap
Devon Ice Cap_filt
δ18O (‰)
-30
-35
-40
-45
B26
B26_filt
δ18O (‰)
-30
-35
-40
-45
B16
B16_filt
δ18O (‰)
-30
-35
-40
-45
B29
B29_filt
δ18O (‰)
-30
-35
-40
-45
B21
B21_filt
δ18O (‰)
-30
-35
-40
-45
B22
B22_filt
δ18O (‰)
-30
-35
-40
-45
B17
B17_filt
δ18O (‰)
-20
-25
-30
-35
Agassiz A87
Agassiz A87_filt
δ18O (‰)
-20
-25
-30
-35
Camp Century
Camp Century_filt
δ18O (‰)
-30
-35
-40
-45
B30
B30_filt

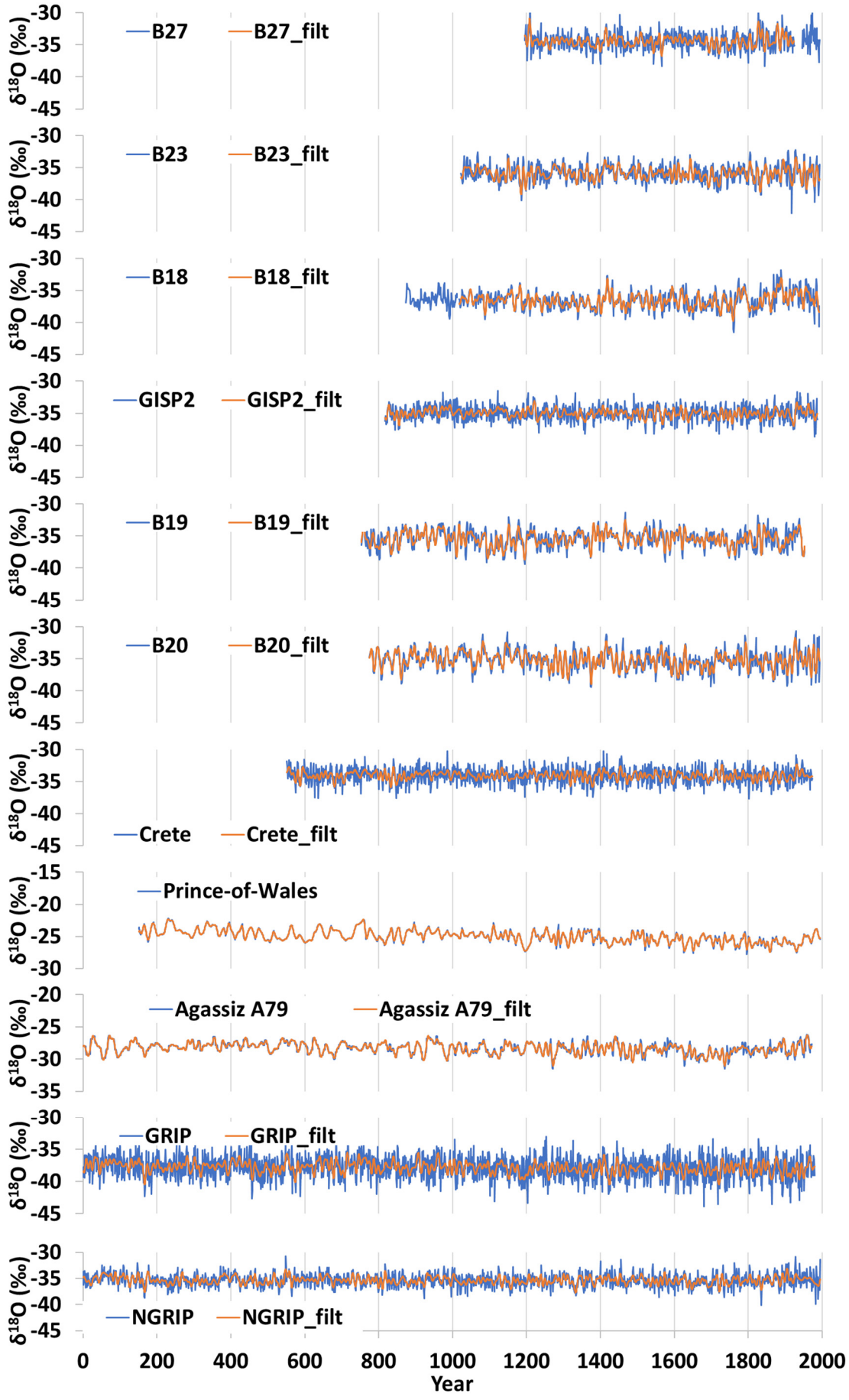

B27
B27_filt
B23
B23_filt
B18
B18_filt
GISP2
GISP2_filt
B19
B19_filt
B20
B20_filt
Crete
Crete_filt
Prince-of-Wales
Agassiz A79
Agassiz A79_filt
GRIP
GRIP_filt
NGRIP
NGRIP_filt
δ18O (‰)
Year
0
200
400
600
800
1000
1200
1400
1600
1800
2000

Figure S3. Raw and 10-yr lowpassed time series of the 30 ice core $\delta^{18}O$ records assessed in the study gathered from the Iso2k database.

S3. Software used

The research was performed in R statistical environment [8]. The Iso2k database was queried with the lipdR [9] and geoChronR [10] packages, bandpass filtering was performed with the bandpass() function of the astrochron package [11] and changepoint detection was done with the Bayesian Change Point algorithm [2]. The figures were prepared in R, MS Excel 360 and CorelDRAW 2021.

References

1. Scott, A.J.; Knott, M. A Cluster Analysis Method for Grouping Means in the Analysis of Variance. Biometrics 1974, 30, 507-512.
2. Green PJ. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika **1995,** 82(4):711–732. doi:10.1093/biomet/82.4.711
3. Fearnhead P, Liu Z. On-line inference for multiple changepoint problems. J R Stat Soc B **2007,** 69(4):589–605**.**
4. Topál, Dániel, Matyasovszky, István, Kern, Zoltán and Hatvani, István Gábor. "Detecting breakpoints in artificially modified- and real-life time series using three state-of-the-art methods". Open Geosciences, **2016**, 8(1):78-98. https://doi.org/10.1515/geo-2016-0009
5. Ruggieri, E. A Bayesian approach to detecting change points in climatic records. International Journal of Climatology **2013**, 33, 520-528.
6. Barry, D.; Hartigan, J.A. A Bayesian Analysis for Change Point Problems. Journal of the American Statistical Association **1993**, 88, 309-319.
7. Zeileis A, Leisch F, Hornik K, Kleiber C. Strucchange: an R package for testing for structural change in linear regression models. J Stat Softw **2002,** 7(2):1–38.
8. R Core Team R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing: Vienna, Austria, **2019**.
9. McKay, N.P.; Emile-Geay, J. Technical note: The Linked Paleo Data framework – a common tongue for paleoclimatology. Clim. Past **2016**, 12, 1093-1100.
10. McKay, N.P.; Emile-Geay, J.; Khider, D. geoChronR – an R package to model, analyze, and visualize age-uncertain data. Geochronology **2021**, 149-169.
11. Meyers, S.R. astrochron: An R Package for Astrochronology, **2014**.