Comparison of piecewise linear change point detection with traditional analytical methods for ocean and climate data

D. Banesh[1,2] · M. Petersen[2] · J. Wendelberger[2] · J. Ahrens[2] · B. Hamann[1]

Abstract

The Earth's atmosphere and oceans are largely determined by periodic patterns of solar radiation, from daily and seasonal, to orbital variations over thousands of years. Dynamical processes alter these cycles with feedbacks and delays, so that the observed climate response is a combination of cyclical features and sudden regime changes. A primary example is the shift from a glacial (ice age) state to interglacial, which is driven by a 100-thousand year orbital cycle, while the transition occurs over a period of hundreds of years. Traditional methods of statistical analysis such as Fourier and wavelet transforms are very good at describing cyclical behavior, but lack any characterization of singular events and regime changes. More recently, researchers have tested techniques in the statistical discipline of change point detection. This paper explores the unique advantages of a piecewise linear regression change point detection algorithm to identify events, regime shifts, and the direction of cyclical trends in geophysical data. It evaluates the reasons for choosing this particular change detection algorithm over other techniques by applying the technique to both observational and model data sets. A comparison of the proposed change detection algorithm to the more established statistical techniques shows the benefits and drawbacks of each method.

Keywords Change point detection · Ocean data · Fourier transform · Wavelets

Introduction

Data sets in geophysical dynamics span from minutes to millennia and meters to continents. Scientists rely on analysis tools to view dynamical processes in meaningful and insightful ways. Each tool represents time and spatial scales in a particular way, and highlights certain aspects of the data while obscuring others. A striking example is the application of the Fourier transform to paleo-climate data, where the Milankovitch orbital cycles of 100 and 41 thousand years form obvious spectral peaks (Milankovitch 1941), yet the Fourier spectrum provides no information about the exact timing, spacing, or speed of glacial to interglacial transitions. Modern climate change research has emphasized the importance of characterizing the sources of natural variability to understand anthropogenic influence on earth systems. These geophysical processes include El Niño oscillation (Oceanic Nino Index 2018), oceanic circulation, the Atlantic meridional overturning circulation (McManus et al. 2004), ice–albedo feedback (Imbrie et al. 1993), and a host of others. Traditional statistical methods for large time series data from observations and models include the Fourier and continuous wavelet transforms (Talley 2011; Wunsch 2015). The Fourier transform, or spectral analysis, of a data set allows a user to quickly identify major oscillatory components by extracting the most prominent peaks in the spectrum (von Storch and Zwiers 2002). In contrast, wavelet analysis does not average the amplitude and phase for each frequency component *over time*; it provides a localized estimate for each spectral component at a *particular point in time* (Thomson and Emery 2014). When the frequency of a data stream changes abruptly in the middle of the sequence, this positional change would be captured by wavelet analysis and would be missed by Fourier analysis.

More recently, ocean scientists have incorporated change point detection algorithms for data analysis. Change point detection, or simply change detection, is a widely used

✉ D. Banesh
  dbanesh@ucdavis.edu

1 University of California, Davis, Davis, USA

2 Los Alamos National Laboratory, New Mexico, USA

statistical approach for targeted data reduction. It includes regression-based methods, Bayesian methods, and multivariate methods, among others (Basseville et al. 1993). Time series decomposition (Hyndman and Athanasopoulos 2018) is one of the more popular change detection techniques for climate and ocean data, as it decomposes a signal into trend, seasonal/cyclical and noise components. For example, Verbesselt et al. (2010) examined satellite images showing land cover of vegetation over time to identify three types of events that might determine change: seasonal effects, gradual climate variability and abrupt change resulting from deforestation, fires or floods. Quan et al. (2016) applied time series decomposition to recorded land surface temperature data over Beijing and performed a comparison to simulated data using root mean squared (RMS) error to understand differences in observations and model predictions. Goela et al. (2016) temporally decomposed sea surface temperature data from various points off the coast of Portugal. They subtracted the seasonal component from the original data stream and applied a linear regression fit to the remainder to extract prominent change points.

These efforts have thoroughly explored applications of time series decomposition to ocean and climate data, but time series decomposition in and of itself fails to provide significant information to the scientist beyond what a Fourier analysis would. In most cases, the seasonal component of the time series decomposition is computed by identifying the major frequencies in the data, using a Fourier transform to identify these frequencies. Additionally, if all the major seasonal and cyclical components are not properly extracted, they will be reflected in the noise component, possibly corrupting this graph into misidentifying the more sporadic events in data history.

This paper introduces the use of a different technique for cyclical ocean and climate data analysis: piecewise linear regression change point detection. This algorithm identifies major *regime shifts* in cyclical climate data, providing the scientist with specific points in time where major geophysical events have occurred in history. This is a unique type of feedback that cannot be provided to the scientist with Fourier and wavelet analysis.

Piecewise linear fitting for climate applications has been previously explored in Tomé and Miranda (2004). However, there are many differences between the implementation of the technique presented in this paper and the technique presented in the related work. Firstly, the related work identifies break points by minimizing the residual sum of squares (RSS). The technique presented in the following section *uses* the RSS to define the *F*-statistic, but the primary goal of the algorithm presented is not to minimize this value but to use it to identify local behavior. Additionally, the algorithm proposed by Tomé and Miranda (2004) has several restrictions including: (1) defining a minimum distance between break points, (2) requiring that consecutive line segments reverse from increasing to decreasing or vice versa (3) and limiting the maximum number of break points to 12. None of these are limitations of the algorithm proposed in this paper. The algorithm presented does not enforce a minimum distance between break points, allowing breaks to be defined by both sudden and gradual changes. It also allows multiple consecutive trends to be both increasing or decreasing, allowing for break points to occur when the data, for example, changes from a gradual decrease to a more dramatic, immediate drop. Finally, the algorithm sets no limitations on the number of break points, rather allowing the data and algorithmic parameters to dictate the number needed to describe the variation.

The value of the change detection algorithm presented in this paper is explored through five data sets: (1) carbon dioxide records from Antarctic ice cores (Bereiter et al. 2015; (2) sedimentary oxygen-18 isotope records, a proxy of temperature, from the Benthic zone (Lisiecki and Raymo 2005); (3) the Oceanic Niño Index at the Niño 3.4 region; (4) the North Atlantic Oscillation (NAO) Index between a station in the Azores and one in Iceland (Hurrell 1995; Jones et al. 1997); and (5) ocean eddies in the Agulhas regions south of Africa (MPAS-Developers 2013; Ringler et al. 2013; Petersen et al. 2015). The paper also provides a comparison of the piecewise linear regression technique (Myers et al. 2016), to Fourier- and wavelet-based analysis to derive advantages and disadvantages of these methods. The case study for the North Atlantic Oscillation Index data also discusses the results obtained by the presented algorithm to the results shown by Tomé and Miranda (2004) to highlight the differences of the two techniques.

Change detection

A change point, in broad statistical terms, refers to a place or time such that the observed data follows one distribution up to that point and another distribution after that point (Chen and Gupta 2011). Change point detection algorithms generally serve two main purposes: (1) to decide whether there is change in the data and (2) to determine the locations where this change is present. For geophysical data, the goal in using change point detection is to extract time steps of scientific significance.

As Reeves et al. (2007) discussed in their survey paper on the application of various change point algorithms for climate data, the type of technique to use depends on the data to be analyzed. The Bayesian change point detection discussed in Ray and Tsay (2002), for example, determines change points by subdividing the data into regions with a constant mean. Other techniques (CRAN 2017) determine change by penalizing the variance of the data. However,

these techniques are not as relevant for cyclical data because ocean scientists are generally more interested in examining properties of change in the data rather than identifying uniformity. Piecewise linear regression models, such as the one presented in this section, are more appropriate, as they look for first-order changes, i.e., points at which the rate of change differs from one region to another.

Piecewise linear regression change detection

For the data discussed in this paper, a technique that captures the changes from the increasing trend to the decreasing trend, and vice versa, is vital. These *regime shifts* often signify noteworthy events in history. Therefore, a *piecewise linear regression model* based on the work presented by Myers et al. (2016) is more appropriate. Although originally designed to analyze pixel values in an image, this approach has been adapted to detect changes in other types of data. There is a variety of change point methods based on piecewise representations, including the trend filtering approach proposed by Tibshirani et al. (2014) that uses the Lasso technique (Hyun et al. 2016).

A *linear regression model* estimates the least squares line fit to a set of data points. The goal is to estimate the best linear relationship between the dependent variable on the *x*-axis, and the independent variable on the *y*-axis. However, if the relationship between the two variables is non-linear, then the linear regression model will be a poor representation of the data. This can be addressed using a *piecewise* linear regression model, where line *segments* are fit to *subsets* of the data. A line segment is used to represent the data as long as the error between the data and the fitted line segment is acceptably small. When the addition of a new data point increases the error beyond the acceptable threshold, this point is set to be the "change point", where a new line segment begins as illustrated in Fig. 1.

The piecewise linear regression model determines change points in the following way. The user first defines a buffer size, *B*, indicative of the number of data points the algorithm
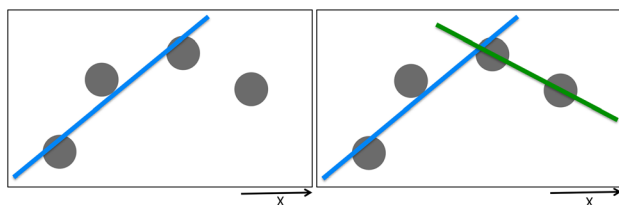
Fig. 1 Illustration of piecewise linear regression. The image on the left shows a blue line segment fit to the first three data points. The fourth point would introduce too much error to the linear regression model to be a good fit to the blue line. Therefore, a new line fit starts, as shown in green on the right, encompassing the third and fourth points. The fourth point is considered the change point

will consider to find a change point. For example, in a time-dependent data set, *B* would correspond to the number of time steps to examine. This approach ensures that even with very large data sets, calculations can still be performed efficiently by focusing on smaller regions when desired. Given a buffer size, *B*, the algorithm considers the first *B* time steps in the simulation, denoted as curr, and the subsequent *B* time steps, denoted as *buff*. It computes two residual sum-of-squares (RSS) terms for a piecewise linear fit; they are:

$$RSS_1 = RSS_{curr \cup buff}$$
$$RSS_2 = RSS_{curr} + RSS_{buff}.$$

RSS_1 determines the RSS for the combined set of curr and buff (a single line was fit to the combination of both sets), while RSS_2 determines the sum of RSS for each set curr and buff (there were separate lines fit to curr and to buff). These values are used to calculate the *F-statistic* associated with the two fits.

$$F = \frac{\left(\frac{RSS_1 - RSS_2}{p_2 - p_1}\right)}{\left(\frac{RSS_2}{T_{curr \cup buff} - p_2}\right)},$$

where $p_1 = 2$ and $p_2 = 4$, with p_1 and p_2 denoting the number of parameters in each fit; $T_{curr \cup buff}$ is the total number of time steps being considered. The *F*-statistic is used to determine whether one line or two lines would be a better representation for the selected region of data.

The user also provides a second input value, α. For any data point in *curr* and *buff*, when the data point maps to a value of the F-distribution that is larger than the given α value, this point is considered to be a change point. For certain data sets, the α criterion for change point detection can still identify a larger number of change points than desired. Therefore, a third user-defined parameter is considered, δ^2. The *F*-distribution is closely related to the variance of the two sets of data, and because in most cases, closely located data are correlated with each other, the δ^2 parameter takes this correlation into account when detecting change points. This parameter directs the algorithm to make it more difficult to select change points in the presence of auto-correlation of nearby points. For more details on this piecewise linear regression model, see Section 3 of Myers et al. (2016).

When applying this technique to ocean data sets, smaller, "more discrete" data sets resulted in more abrupt changes than the larger data examples used in Myers et al. (2016). To address this issue, a wrapper function is added to the algorithm that first searches for regions in the data, where two or more consecutive points have the same *y*-value, that is, flat regions of no change. The first data point of a flat region is automatically marked as a change point and the region of no change is marked as having no additional change points.

The piecewise linear regression algorithm is applied to every set of remaining data points between these regions of no change to determine any additional change points in the data. An additional optional parameter, n_{flat}, is introduced, ranging from two to $n + 1$, where n is the size of the entire data set. Only flat regions above that number of points are considered and marked as having change points detected, with a value of $n + 1$ indicating that a flat region, regardless of size, should not be considered. When searching large data sets of hundreds or more points, flat regions of two or three points might not necessarily indicate significant change, so this parameter allows a user greater control over the change points detected. Additionally, the first and last data points of a data set are always marked as change points.

Results and discussion

To explore the capabilities of Fourier analysis, wavelet analysis and change detection, these methods are applied to six different data sets—four observational data sets and two simulation model-derived data sets. The simulated data examples, an analyses of Cinema databases extracted from a Model for Prediction Across Scales-Ocean (MPAS-Ocean) (MPAS-Developers 2013) simulation (Petersen et al. 2013), serve as pedagogical examples to illustrate the application of the change detection algorithm to simulated ocean data. Fourier transform and continuous wavelet analysis of this data do not prove to be informative due to the small size of the data. To more directly compare Fourier and wavelet analysis to change detection, the observational data sets, data extrapolated from Antarctic ice cores (Bereiter et al. 2015), d18O readings from the Benthic zone of 57 globally distributed locations (Lisiecki and Raymo 2005), the Oceanic Niño Index from the Niño 3.4 region (Oceanic Nino Index 2018) and the North Atlantic Oscillation Index (NAO 2019) proved more valuable. These data sets range in size from many decades to millions of years, have well-studied

periodic behavior and exhibit significant geological events throughout the data. The case studies below utilize the fast Fourier transform and continuous wavelet transform functions from Matlab. The change detection model is implemented in the R statistical environment.

A pedagogical example: change detection at multiple levels of α

To demonstrate the usefulness of piecewise linear regression change detection, it is applied to data derived from an MPAS-Ocean simulation. This multi-resolution ocean simulation data set with identifiable eddies, currents and other turbulent features is commonly used in the ocean science community (Petersen et al. 2019; Golaz et al. 2019). A summary of the pipeline described in this example is shown in Fig. 2. From the MPAS-Ocean data, Cinema image databases (Ahrens et al. 2014) of surface kinetic energy are extracted. A Cinema database is a collection of images, with each image a perspective projection of the simulation data to a 2D image plane. When generating the Cinema database, the scientist must ensure that the resolution of the images is sufficient for their future analysis, similar to how they must ensure the proper resolution of their original simulation. For this MPAS-Ocean Cinema data set, the simulation is over-sampled to ensure that each component of the simulation is represented by several pixels, ensuring a high quality of input for image feature analysis. Contour features are then detected and extracted, using the method described by Banesh et al. (Banesh et al. 2017) to identify features of interest, (bright green regions in Fig. 3). The contour detection technique takes the grayscale version of a Cinema image as input. It applies a user-defined threshold value and assigns all pixels in the image above the threshold to a value of one, and all pixels below the threshold to a value of zero. Every connected set of pixels with a value of one is considered to be a derived contour. The technique described is robust enough to track slow moving features
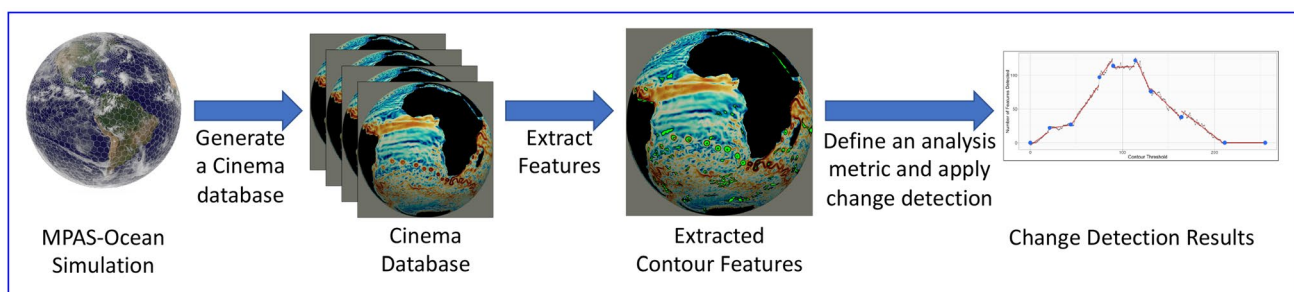
Fig. 2 A diagram of the pipeline for the application of statistical change detection to Cinema databases extracted from an MPAS-Ocean simulation. From the MPAS-Ocean simulation, temporal Cinema databases are extracted at a high image resolution. For each

image, a set of contour features are extracted at varying isovalues for analysis. Finally, for a particular parameter, e.g., isovalue or time, and for a particular metric, e.g., number of features detected, change detection is applied for statistical insight
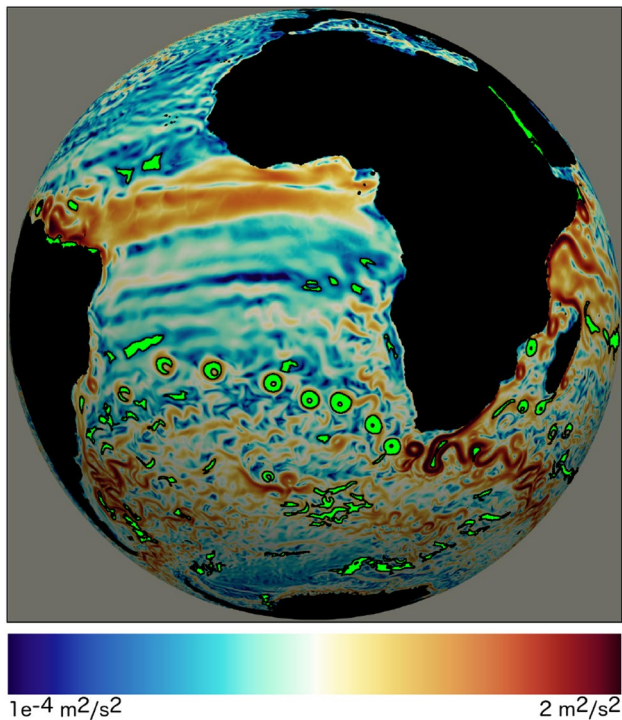
Fig. 3 MPAS-Ocean image from a Cinema database of surface kinetic energy, using a log-scale, hot–cold color map (chosen for its high discriminative power and minimal color vision issues (Turton et al. 2017; Samsel et al. 2015; Ware et al. 2018)). Contour detection was applied with a threshold value of 77. Detected regions are highlighted in bright green

over small deformations such as the curvature of the Earth's surface. Change detection analysis is applied to a metric based on these contours. There are multiple reasons why a climate scientist might opt to use Cinema databases rather than analyzing the raw MPAS-Ocean data. Feature detection on high-resolution data can be time and labor intensive and may also require high-performance computing support for analysis (Woodring et al. 2015). Furthermore, many image processing techniques that are quick to implement and apply to two-dimensional data are much more complex and may not be available for three-dimensional data sets. The generation of a Cinema data set allows for quick and accurate results (Banesh et al. 2017). The application of the presented change detection technique to model-derived data shows the capability of the overall approach where a complex, multi-dimensional model is simplified for effective statistical analysis.

Figure 4 shows the results of the change detection algorithm applied to the contour threshold parameter of the contour detection algorithm presented in Banesh et al. (2017). By holding the B and δ^2 values constant, α is varied to identify various degrees of change. In simulated data studies, such preliminary steps might be necessary to narrow the

parameter space before moving on to time-based analysis. Figure 4a detects only the highest levels of change, from zero to the maximum number of features detected. Figure 4b–d gradually detects smaller levels of change until users can determine a level that fits their needs. Each of the case studies examined in "Results and discussion" explores the capability of the change detection algorithm for a particular data set by varying the alpha parameter. This is meant to characterize the method that has been described and to show the flexibility of the algorithm through a pedagogical illustration. Though the case studies explore the usefulness of multiple alpha values, it is not expected that every user will find multiple values of alpha useful for their specific goals.

## A pedagogical example: Eddy detection in the Agulhas region

Mesoscale ocean eddies are widely studied in ocean science. They influence the ocean's biological network (Chelton et al. 2011), can contribute to heat transport over several hundred miles (Volkov et al. 2008), affect weather conditions in the ocean, and impact various other aspects of ocean dynamics (McWilliams 2008). A wide range of eddy detection and tracking techniques have been explored. Chelton et al. (2007), Williams et al. (2011) and Petersen et al. (2013) employed variations of the Okubo–Weiss criterion to identify closed regions of uniform vorticity. Chaigneau et al. (2008) and Chen et al. (2011) used versions of a parameter-based, geometric streamline clustering method, the winding-angle method, to find closed streamlines. Souza et al. (2011) compared the Okubo–Weiss and winding-angle approaches to a wavelet packet decomposition method [first introduced by Doglioli et al. (2007)] to identify where one method might perform better than another.

The goal of this study is to identify *when* eddies spin off from the Agulhas Retroflection, an important task for understanding current and mesoscale eddy behavior. To accomplish this goal, a database of MPAS-Ocean Cinema "raw data" images for 60 time steps, each time step, 5 days apart, is explored. In a raw data image, each pixel value is set to the value of the underlying MPAS-Ocean simulation it represents and is not altered by a color map or shading/lighting effects. Therefore, a contour detection algorithm that is applied to this image provides a more accurate representation of the features present in the simulation. This study focuses on the region highlighted in Fig. 5 and selects a constant contour threshold value of 13 using the contour detection algorithm discussed "A pedagogical example: change detection at multiple levels of α". Time is mapped to the *x*-axis and the number of eddies detected on the *y*-axis. The following change detection parameter values are used: $B = 3, \alpha = 1, \delta^2 = 1, n_{\text{flat}} = 2$.

Fig. 4 Contour detection (Banesh et al. 2017) applied to the MPAS-Ocean Cinema image shown in Fig. 3. The contour threshold value is varied from 0 to 255 on the x-axis and the number of features detected is plotted on the y-axis. Change detection is then applied to this graph, holding B constant at 10 and δ^2 constant at 1, while varying α. The blue dots are the change points detected, and the red lines show the piecewise linear regression fits to the data. As α decreases, the number of change points decreases and only corresponds to the higher degrees of change in the data. As α increases, change points corresponding to smaller degrees of change are included. The parameter n_{flat} is set to 257 as to have no impact on these results

(a) $B = 10$, $\alpha = 1e{-}15$, $\delta^2 = 1$
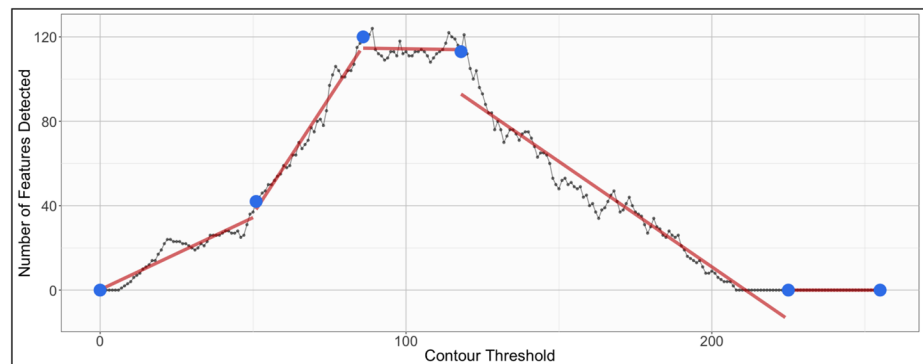
(b) $B = 10$, $\alpha = 1e{-}9$, $\delta^2 = 1$
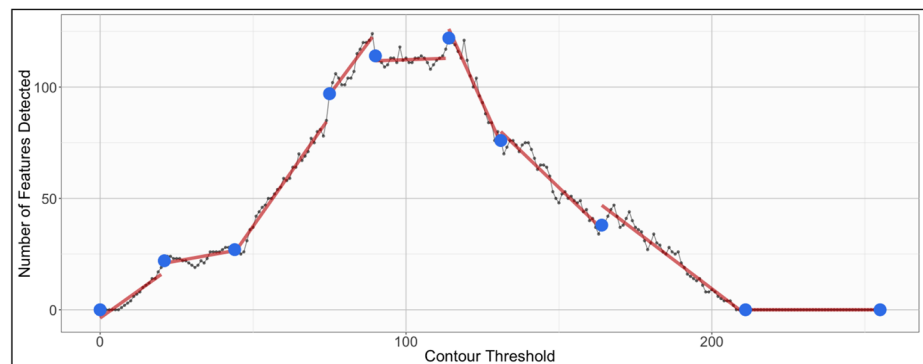
(c) $B = 10$, $\alpha = 1e{-}5$, $\delta^2 = 1$
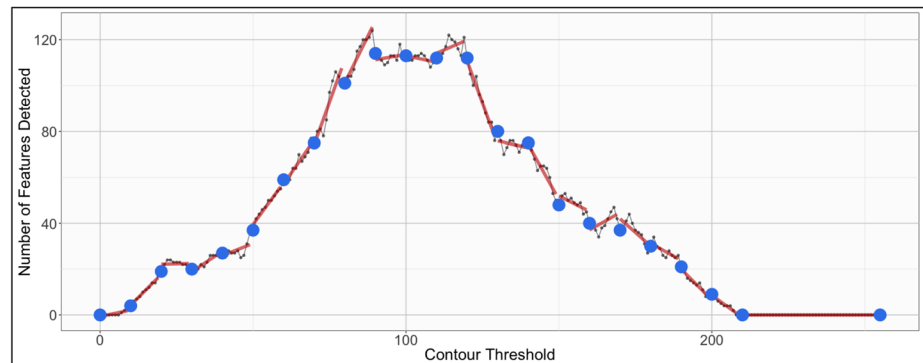
(d) $B = 10$, $\alpha = 1$, $\delta^2 = 1$
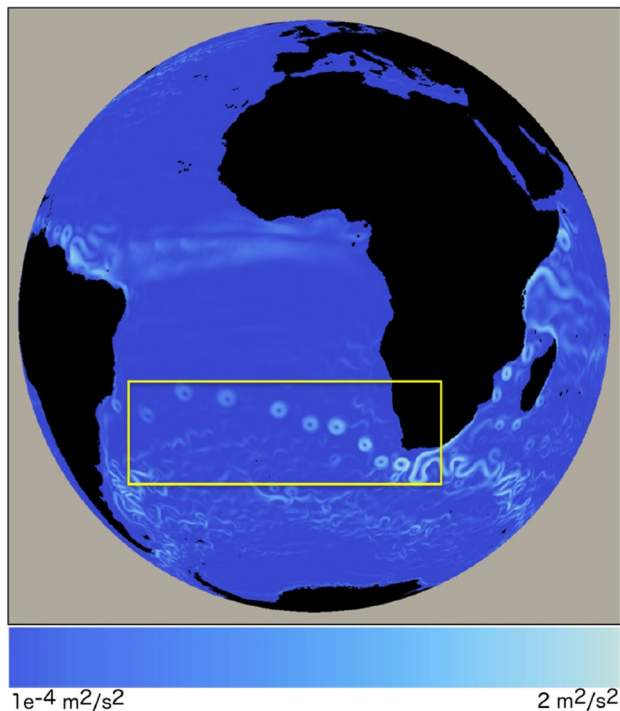
Fig. 5 MPAS-Ocean Cinema "raw data" image of kinetic surface energy. A log-scale, blue color map is used for visualization purposes only; actual analysis is conducted on the underlying data. The boxed region in yellow is the region of interest for this example

The results of change detection applied to this data are shown in Fig. 6, with change points depicted in green, orange or blue. Change points are categorized according to slope. When the slope from timeStep(change point − 1) to timeStep(change point) is positive, the change point is marked as green. When the slope is negative, it is marked as orange. All other change points are marked as blue. Of the 60 time steps analyzed, 10 are marked as change points with a positive slope. These generally indicate the start of increased activity in this region. The two main types of increased activity occur when a new eddy separates from the Agulhas Retroflection or when one eddy splits into multiple eddies during its trek across the South Atlantic. Of these ten positive-slope change points, six are time steps when a new eddy is separating from the Agulhas Retroflection; see Fig. 6b, d. A visual inspection determined that no false negatives are detected; time steps when a new eddy separates have not been missed. Effectively, this has reduced the search space in determining when new eddies emerge from the Agulhas Retroflection, from 60 time steps to 10. Though this is a smaller, representative example, for larger data sets, an automatic detection technique such as this can be invaluable.

Case study: carbon dioxide data in Antarctic ice cores

Bubbles of air trapped in yearly layers of snowfall in Antarctica and Greenland provide scientists with a nearly million-year record of the Earth's climate. Carbon dioxide (CO_2) measurements show the cycles of ice ages approximately every 100,000 years and the tight relationship between CO_2 and proxies for temperature such as oxygen-18 isotopes (Jouzel et al. 2007; Sigman et al. 2010). Kilometer-long ice cores document past climate regimes and the transitions from glacial ages (ice ages) to interglacial conditions. They are used to validate paleo-climate modeling studies and provide insights into the mechanisms of climate dynamics for guiding climate predictions (Stocker et al. 2013).

The data used in this case study, Fig. 7a is a reconstruction of the atmospheric CO_2 concentrations for the past 800ka (kilo-annum, i.e., thousands of years), extracted from ice cores from Dome C in Antarctica (Bereiter et al. 2015). The values in this data set are a composite of values from a large set of ice cores. Figure 7 shows results when performing change detection for various levels of alpha (b)–(d), and it compares these results to results obtained with continuous wavelet analysis (e) and Fourier analysis (f). Time 0 is present time, and 800ka is past time; events occurred chronologically from the right to the left of the graph.

Since CO_2 measurements portray the 100ka cyclical behavior very clearly, this is an ideal data set to explore the benefits of piecewise linear regression change point analysis. One of the advantages of this technique over other methods is its capability to provide directional and positional information that correlate the change points to events in time. Figure 7b, when examining change at the highest levels, shows that the change points (black dotted lines) correspond to the drastic and almost immediate regime transitions in the data, the shifts from a glacial to interglacial period. Fast warming is plausible due to the ice–albedo feedback effect—ice that is melting melts more ice, an effect that compounds quickly. However, cooling is much slower, as can be seen by the large regions with no change points (e.g., from about 150ka to about 10ka). Additionally, the change point algorithm provides the scientist with directionality when fitting linear segments to the data. These linear segments indicate that most of the larger regions without change (indicating more gradual change) occur when transitioning from an interglacial to glacial period, and very rarely the inverse. Change detection also gives users control over the level of change they are interested in. Similarly to determining the coastline of a continent with higher levels of resolution as one zooms in, the change in the data can be determined at finer levels of resolution by changing the α parameter, see Fig. 7b–d.

Fig. 6 Eddies are tracked in the Agulhas Retroflection region over 60 time steps. Change points are shown as a combination of blue, green and orange, where green change points indicate a positive slope from the previous step to the change point, orange indicates a negative slope, and blue indicates a zero slope. Green change points indicate the start of increased activity in the region: either a new eddy separates from the Agulhas Retroflection, as in (**b**) and (**d**), or one eddy splits into multiple eddies, as in (**c**). The new eddies are identified by the red dots in (**b**)–(**d**)
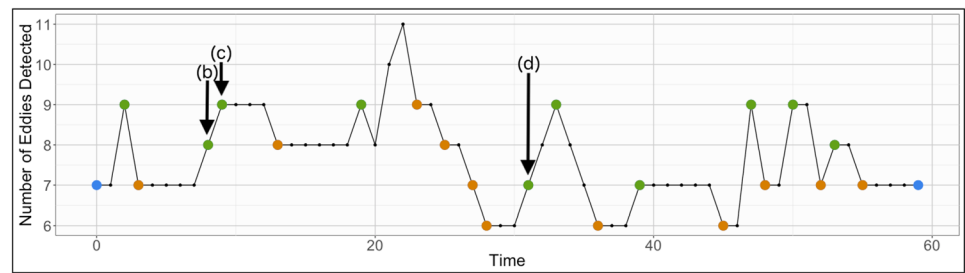
(**a**) $B = 3, \alpha = 1, \delta^2 = 1, n_{flat} = 2$
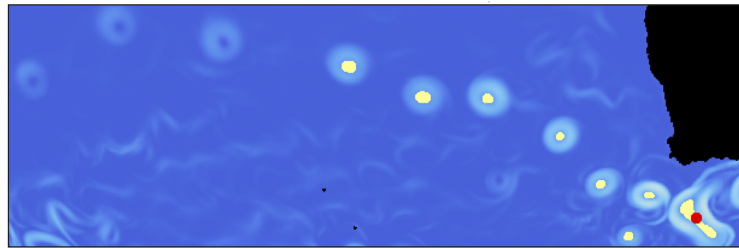
(**b**) $Time = 8, Eddies = 8$
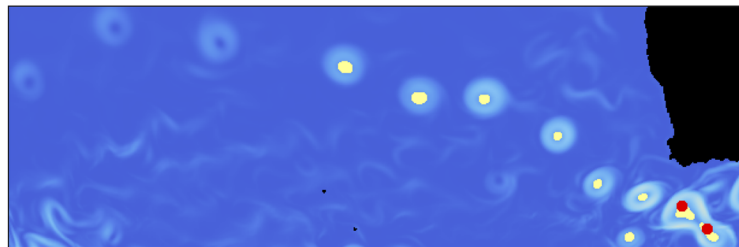
(**c**) $Time = 9, Eddies = 9$
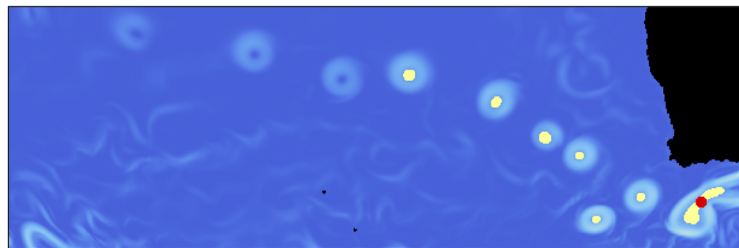
(**d**) $Time = 31, Eddies = 7$

In comparison, the continuous wavelet spectrum provides localized time-based information for frequencies in the data. Unlike the Fourier spectrum, which identifies major cycles in a data stream, but cannot correlate that information to specific time intervals, the continuous wavelet transform identifies at what time intervals certain frequencies are more or less prominent. The continuous wavelet transform can be thought of as the Fourier transform rotated by 90° counter-clockwise, so the *x* values now map to the *y*-axis, and replicated horizontally across time. In Fig. 7f, the Fourier transform highlights the Milankovitch cycles at 100 ka, 41 ka and 23 ka (red dotted lines) as some of the most prominent cyclical characteristics of this data. However, with the continuous wavelet transform, see Fig. 7e, the 100 ka cycle

is only prominent during the last 450 ka and not clearly defined for the time frame prior. This is visually evident in the difference in cyclical behavior in the first half of the data versus the second half. Though these are both informative techniques, they present a global view of the data that is difficult to relate to specific historical events.

Case study: oxygen isotopes in benthic sediment cores

This study examines a collection of oxygen-18 isotope(d18O) records gathered and synthesized from 57 deep sediment cores from around the world. The most common oxygen isotope is ^{16}O, while ^{18}O occurs in about 1 of

Fig. 7 Comparison of statistical analysis techniques for the composite CO_2 records (NCEI 2018). Given the data (**a**), change detection at multiple levels (**b**)–(**d**) is defined by varying α from $1e{-}36$ to $1e{-}10$. In these graphs, the change points are defined by the black dotted lines, and the liner regression fits by the red solid lines. The continuous wavelet transform (**e**) and Fourier transform (**f**), give the scientist a global view of the major cyclical elements
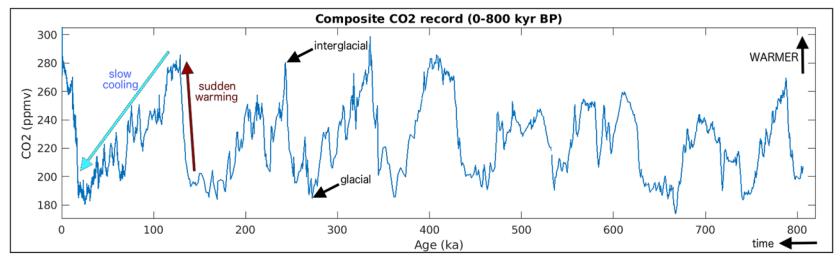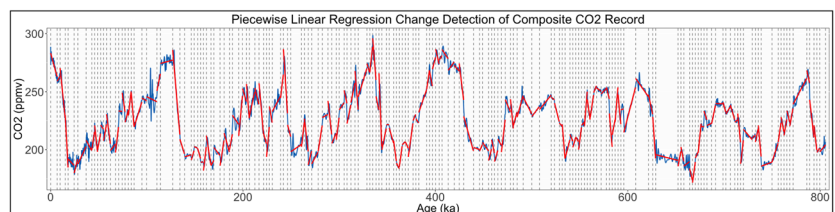

(**a**) *Composite CO_2 record data*
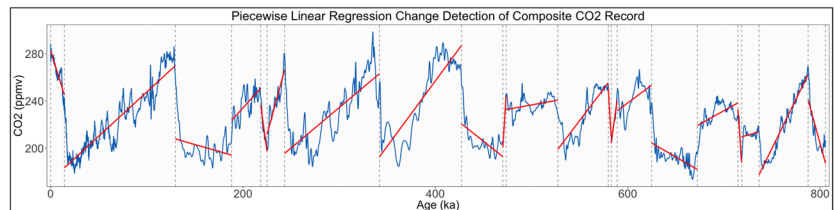
(**b**) *Change detection at: $B = 30$, $\alpha = 1e{-}36$, $\delta^2 = 1$*
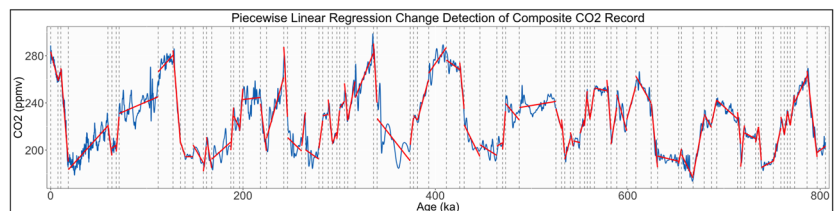
(**c**) *Change detection at: $B = 30$, $\alpha = 1e{-}21$, $\delta^2 = 1$*

(**d**) *Change detection at: $B = 30$, $\alpha = 1e{-}10$, $\delta^2 = 1$*
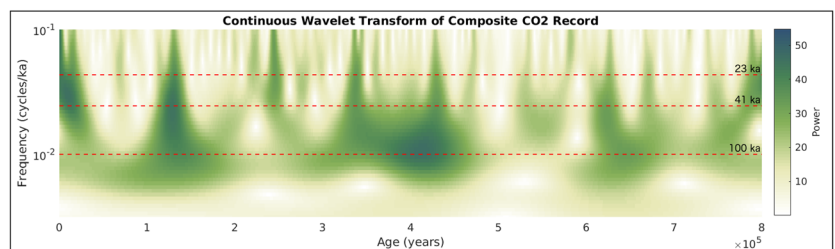
(**e**) *Continuous wavelet transform (log scale)*
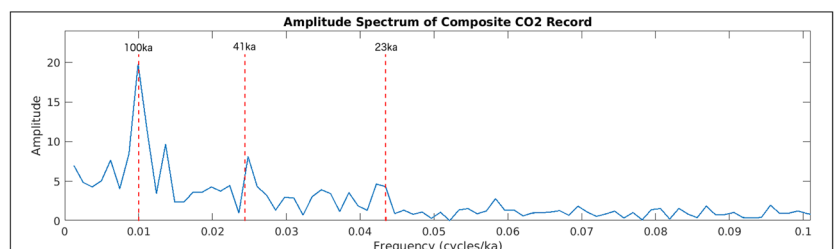
(**f**) *Fourier transform*

500 atoms. The isotope d18O serves as a proxy for temperature because "heavy water," $H_2^{18}O$, requires more energy to evaporate than "light water," $H_2^{16}O$, which has a lighter isotope of oxygen. Ocean water is enriched with d18O compared to water in rain and snowfall, and the ratio is a function of temperature. Sediment and ice core records show strong correlation between CO_2 concentrations and temperature proxies over the last million years (Jouzel et al. 2007).

Figure 8 shows the results of the various statistical algorithms applied to the Benthic d18O data set. Time 0 is present time, and 4000ka is time in the past (Lisiecki and Raymo 2005). The first 800 ka of this data set correlates strongly with the data used in the previous case study.

The change detection graphs shown in Fig. 8b–d progress from the highest levels of change to the more minute levels of change. Similar to the behavior in the previous data set, the change points (black dotted lines) in Fig. 8b capture the sudden regime shifts from the glacial to interglacial states over the past million years, and correspond to known major glacial events. The lack of change points in Fig. 8b before one million years ago signify a dramatic shift in behavior from the previous three million years to the most recent million years. This shift is behavior is reflected in the continuous wavelet transform; see Fig. 8e, where the frequencies at 100 ka are much more prominent for the first million years than for the rest of the graph. However, other than noting that there is a significant lack of cyclical behavior in data past the most recent million years, the continuous wavelet transform says little else about what is happening. In contrast, by progression through the change detection series shown in Fig. 8, the scientist can extract more information about the smaller fluctuations in the data, using break points between the rises and falls. In the process, the first few cycles over the most recent million years are also decomposed into finer components.

Case study: El-Niño-Southern Oscillation

The El-Niño-Southern Oscillation (ENSO) is an important characteristic of the coupled ocean–atmosphere system. Strictly speaking, the ENSO index is the anomaly of the monthly average sea surface temperature over a region of the equatorial Pacific, relative to the long-term mean, or climatology, for that month (Oceanic Nino Index 2018). This simple measure has far-reaching correlations with temperature and rainfall. In the El Niño state (ENSO 2.4 index greater than 0.5), the southern USA is wet and cool, while the northeastern USA is warm, while during a La Niño state (ENSO 2.4 index less than − 0.5) the opposite is true [Fig. 9a, Wang et al. (2017)].

ENSO is a classic example of oscillatory climate behavior, with periodic regime shifts from one state to another. Typical ENSO cycles last for 2–7 years, as can be seen in the Fourier transform (Fig. 9f), but the wavelet transform reveals that the exact nature of the periodicity varies from decade to decade (Fig. 9e). The correct state of ENSO is a critical factor in seasonal predictions, but the exact driving mechanisms of the timing remain poorly understood. The change point detection analysis adds to the traditional Fourier and wavelet analyses in that it highlights trends and the boundaries between them. For a low alpha value of 1e−4, change point detection reveals the fastest changes from El Nino to La Nino, and vice versa (Fig. 9b). For the mid alpha values the algorithm fits trends of 3–5 years (Fig. 9c), and for a high alpha, it fits regular trends of 2–3 years (Fig. 9d). The piecewise linear regression at varying alpha values highlights the span and direction of ENSO oscillations. As the oscillations in this data are not as dramatic as the previous two case studies, the change detection algorithm may sometimes combine smaller peaks into a singular region of relatively consistent behavior.
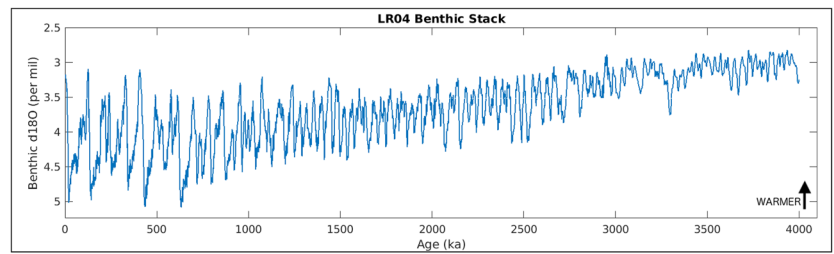
Case study: North Atlantic Oscillation

The North Atlantic Oscillation (NAO) is the most prominent pattern of atmospheric variability at northern latitudes during the cold season (November–April). It sets the position of the jet stream, and is therefore highly correlated with precipitation and surface air temperatures throughout North America and Europe. The NAO index is defined as the normalized sea level pressure difference between the Azores in the subtropics and Iceland in the North Atlantic.

The data examined in this study is a 12-month moving average of the NAO Index (NAO 2019). Figure 10 shows the results of applying different change detection parametric combinations (b)–(d), and how they compare to a continuous wavelet transform plot (e) and to Fourier analysis (f). From the Fourier and continuous wavelet analysis, it can be seen that the majority of the spectral power lies between 1- and 10-year periods that the spectrum is slightly red (stronger at lower frequencies). The NAO is not periodic at any particular frequency, because these dynamics arise from climate noise rather that oscillatory processes (Hurrell et al. 2003). The lack of periodicity can also be seen in the results of the change detection algorithm (Fig. 10b–d), where the length of the lines vary substantially. In the higher values of alpha, Fig. 10c, d, the break points correspond to more drastic regime shifts, from very high NAO Index to very low NAO Index values (or vice versa).

From these results, it is clear that the algorithm presented in this paper leads to a very different set of results when compared to Tomé and Miranda (2004) for several reasons. First and foremost, this algorithm can be tuned to capture the major shifts in the data, the anomalies that highlight the points in data history that correspond to the most drastic changes. The algorithm presented by Tomé and Miranda
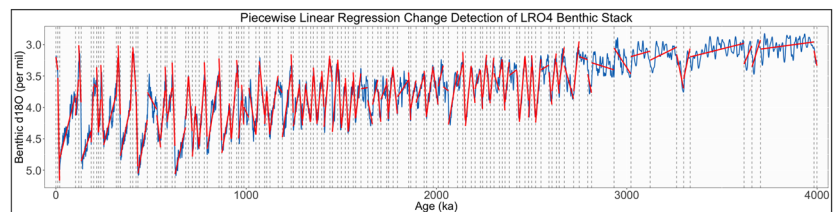
Fig. 8 Comparison of statistical analysis techniques for the d18O Benthic Stack data (Lisiecki and Raymo 2005). Given the data (**a**), change detection at multiple levels (**b**)–(**d**), is defined by varying α from 0.1 to 0.9. In these graphs, the change points are defined by the black dotted lines, and the liner regression fits, by the red solid lines. The continuous wavelet transform (**e**), and Fourier transform (**f**), give the scientist a global view of the major cyclical elements
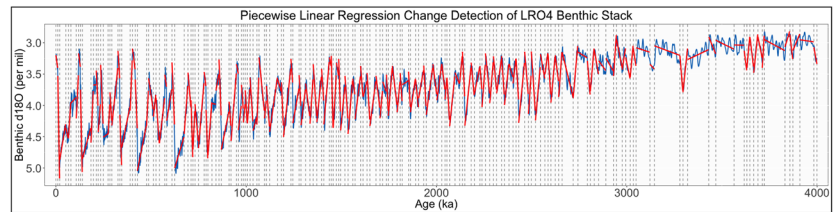

(**a**) *LR04 Benthic Stack data, proxy for temperature*
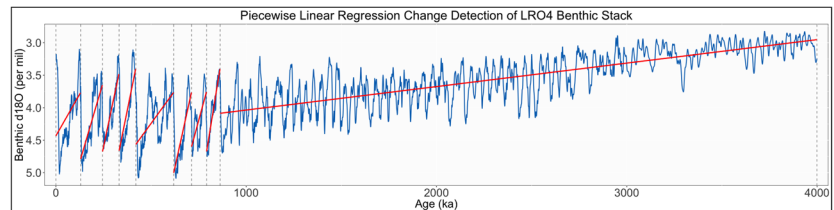
(**b**) *Change detection at: $B = 10$, $\alpha = 0.1$, $\delta^2 = 1$*

(**c**) *Change detection at: $B = 10$, $\alpha = 0.8$, $\delta^2 = 1$*

(**d**) *Change detection at: $B = 10$, $\alpha = 0.9$, $\delta^2 = 1$*
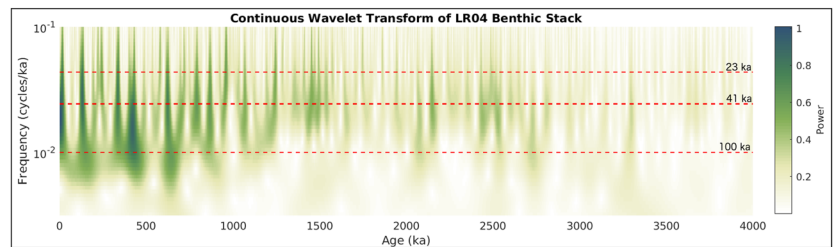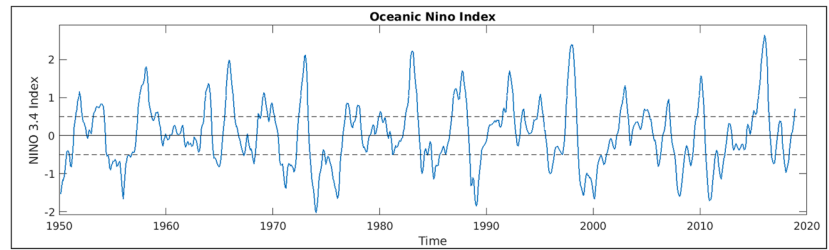

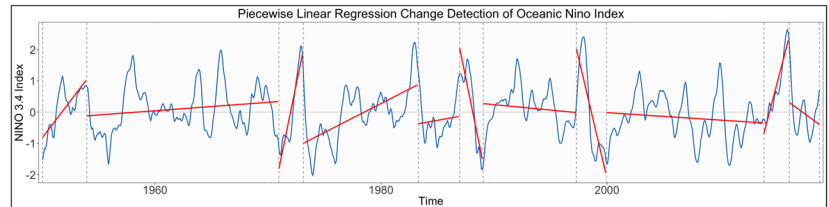(**e**) *Continuous wavelet transform (log scale)*
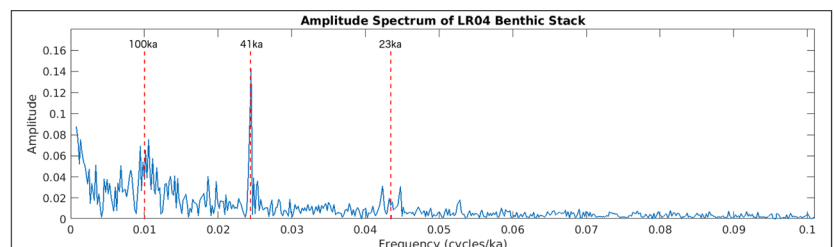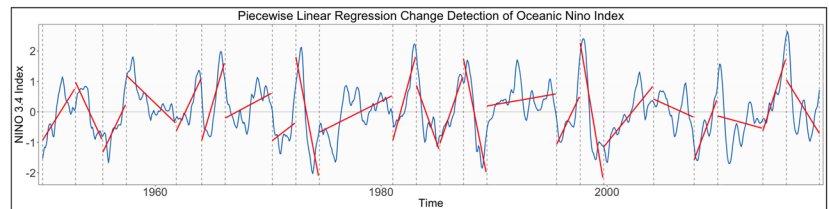

(**f**) *Fourier transform*

Fig. 9 Comparison of statistical analysis techniques for the Oceanic Niño Index (NinoData 2018). Given the data (**a**), change detection at multiple levels (**b**)–(**d**), is defined by varying α from $1e{-}4$ to 1. In these graphs, the change points are defined by the black dotted lines, and the linear regression fits by the red solid lines. The continuous wavelet transform (**e**) and Fourier transform (**f**) give the scientist a global view of the major cyclical elements
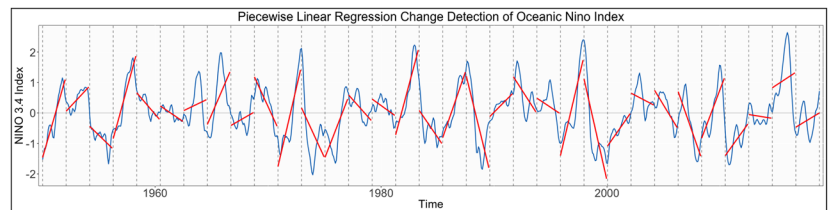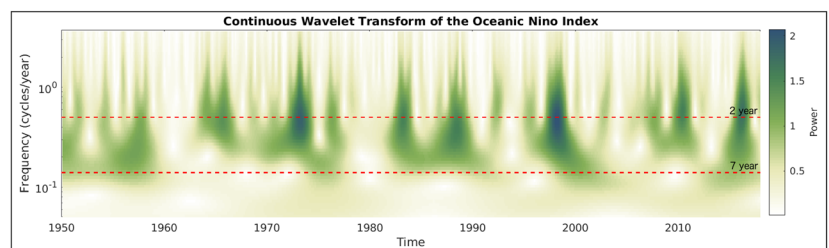

(a) *Oceanic Niño Index*

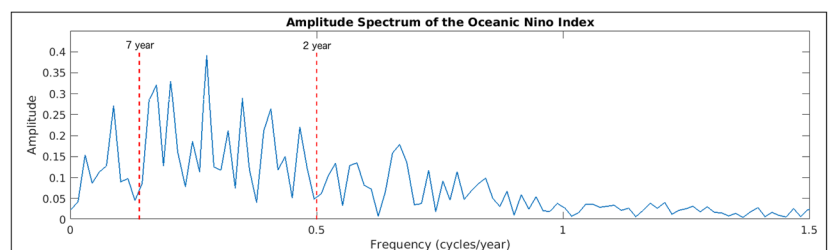**(b)** *Change detection at:* $B = 25$, $\alpha = 1e{-}4$, $\delta^2 = 1$

(c) *Change detection at:* $B = 25$, $\alpha = 1e{-}2$, $\delta^2 = 1$

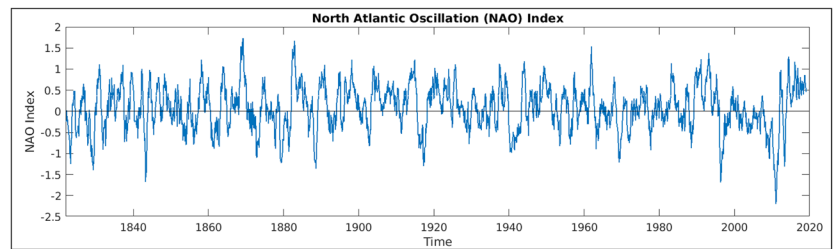**(d)** *Change detection at:* $B = 25$, $\alpha = 1$, $\delta^2 = 1$

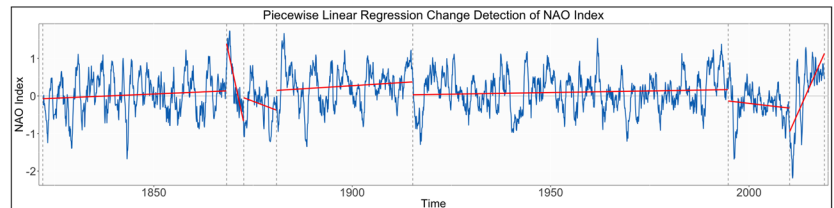**(e)** *Continuous wavelet transform (log scale)*
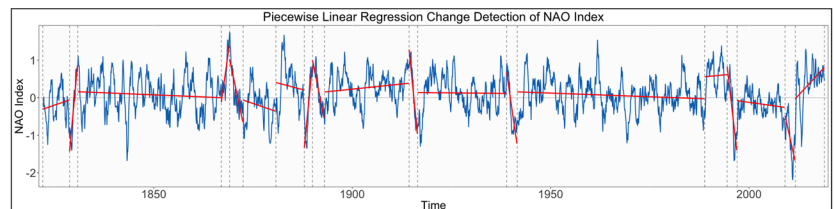

(f) *Fourier transform*

Fig. 10 Comparison of statistical analysis techniques for the 12-month moving average of the North Atlantic Oscillation (NAO 2019). Given the data (**a**), change detection at multiple levels (**b**)–(**d**) is defined by varying α from $1e-3$ to $1e-1$. In these graphs, the change points are defined by the black dotted lines, and the liner regression fits by the red solid lines. The continuous wavelet transform (**e**) and Fourier transform (**f**) give the scientist a global view of the major cyclical elements
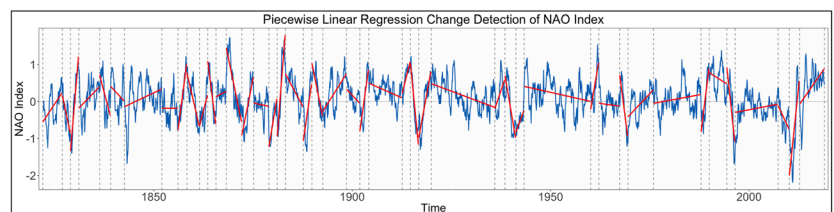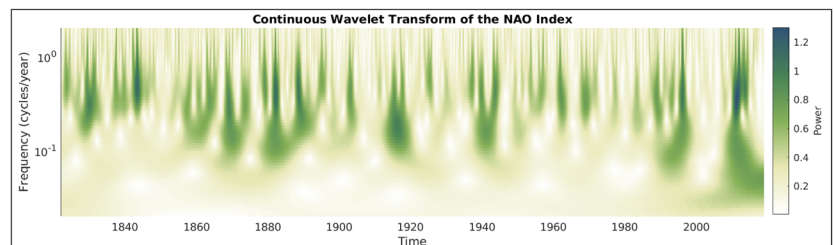
(a) *North Atlantic Oscillation Index*

(b) *Change detection at:* $B = 25, \alpha = 1e-3, \delta^2 = 1$
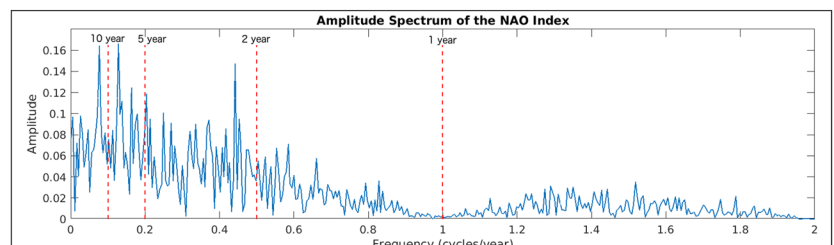
(c) *Change detection at:* $B = 25, \alpha = 1e-2, \delta^2 = 1$

(d) *Change detection at:* $B = 25, \alpha = 1e-1, \delta^2 = 1$

(e) *Continuous wavelet transform (log scale)*

(f) *Fourier transform*

Table 1 Advantages and disadvantages of statistical analysis approaches

	Advantages	Disadvantages
Piecewise linear regression change point detection	*Exact time* of events, corresponding to geologically meaningful events *Slope* of linear regression lines *provide direction* Applicable to smaller data	Not providing a global or *time-integrated view*
Fourier transform (spectral analysis)	*Cyclical components* denoted by power at each frequency *Amplitude* denoting importance	*Time* no information as to when events occur *No direction* Localized in frequency only
Continuous wavelet transform (symmetric)	*Localized time* localized in time and frequency Informing scientist when certain oscillations occur and duration of events *Amplitude/*color denoting importance	*No direction* *Exact time* of events not provided

(2004) is optimized to identify the larger trends in data rather than more localized events. Additionally, because the algorithm presented here does not require the user to define a minimum time period between each break point, as Tomé and Miranda (2004) does, it does not force a 'pseudo-frequency' on the results based on that parameter. Our results are entirely data driven, whereas defining a minimum time between break points injects a underlying periodic assumption to the data. The authors in Tomé and Miranda (2004) acknowledge in their paper that defining a minimum time between break points can lead to false results, when a break point is defined by that restriction rather than the data itself.

Discussion

Based on the presented case studies, it can be concluded that a piecewise linear regression change detection algorithm provides a unique perspective for geophysical data analysis that cannot be attained with Fourier transform or continuous wavelet analysis. This tool provides an analytic, reproducible method for identifying points of change in a mathematically meaningful way. Identifying break points and determining if and how they correlate to the Fourier and continuous wavelet transform results provided the scientists with a unique perspective that may lead to a more insightful examination of their data. Table 1 summarizes the advantages and disadvantages of the discussed techniques.

Conclusions

The four case studies and other examples presented in the paper show that piecewise linear regression change point detection allows scientists to explore cyclical geophysical data in distinctive ways when compared to Fourier or continuous wavelet transforms. Additionally, it can be inferred that linear change detection algorithms are more suited for cyclical data because it captures the behavior of regime

shifts in the data that might not be as apparent with other change detection methods. In the future, the comparative analysis presented in this paper can be expanded to directional wavelets and other statistical methods. A mapping of the benefits of one technique over another would be useful for a scientist looking to add another tool to their analysis arsenal. Future work also involves a deeper analysis of the results of the change detection algorithm. Results can be ordered to understand the largest change in data values, categorized in a histogram to understand trends in the data and compared with known historical events to gain more insight from geophysical data. This technique can also be expanded to multi-variate, n-dimensional climate and geological data for a more advanced analysis.

References

Ahrens J, Jourdain S, O'Leary P, Patchett J, Rogers DH, Petersen M (2014) An image-based approach to extreme scale in situ visualization and analysis. In: Proceedings of the international conference for high performance computing, networking, storage and analysis. IEEE Press, pp 424–434

Banesh D, Schoonover JA, Ahrens JP, Hamann B (2017) Extracting, visualizing and tracking mesoscale ocean eddies in two-dimensional image sequences using contours and moments. In: Rink K, Middel A, Zeckzer D, Bujack R (eds) Workshop on visualisation in environmental sciences (EnvirVis). The Eurographics Association. https://doi.org/10.2312/envirvis.20171103

Basseville M, Nikiforov IV et al (1993) Detection of abrupt changes: theory and application, vol 104. Prentice Hall, Englewood Cliffs

Bereiter B, Eggleston S, Schmitt J, Nehrbass-Ahles C, Stocker TF, Fischer H, Kipfstuhl S, Chappellaz J (2015) Revision of the

epica dome c co2 record from 800 to 600 kyr before present. Geophys Res Lett 42(2):542–549

Chaigneau A, Gizolme A, Grados C (2008) Mesoscale eddies off peru in altimeter records: Identification algorithms and eddy spatio-temporal patterns. Prog Oceanogr 79(2–4):106–119

Chelton DB, Schlax MG, Samelson RM, de Szoeke RA (2007) Global observations of large oceanic eddies. Geophys Res Lett 34:L15606

Chelton DB, Gaube P, Schlax MG, Early JJ, Samelson RM (2011) The influence of nonlinear mesoscale eddies on near-surface oceanic chlorophyll. Science 334(6054):328–332

Chen G, Hou Y, Chu X (2011) Mesoscale eddies in the south china sea: Mean properties, spatiotemporal variability, and impact on thermohaline structure. J Geophys Res Oceans 116:C06018

Chen J, Gupta AK (2011) Parametric statistical change point analysis: with applications to genetics, medicine, and finance. Springer Science & Business Media, Berlin

CRAN (2017) The comprehensive R archive network. https://cran.r-project.org/web/packages/ecp/index.html

Doglioli A, Blanke B, Speich S, Lapeyre G (2007) Tracking coherent structures in a regional ocean model with wavelet analysis: application to cape basin eddies. J Geophys Res Oceans 112:C05043

Goela PC, Cordeiro C, Danchenko S, Icely J, Cristina S, Newton A (2016) Time series analysis of data for sea surface temperature and upwelling components from the southwest coast of Portugal. J Mar Syst 163:12–22

Golaz JC, Caldwell PM, Van Roekel LP, Petersen MR, Tang Q, Wolfe JD, Abeshu G, Anantharaj V, Asay-Davis XS, Bader DC, Baldwin SA, Bisht G, Bogenschutz PA, Branstetter M, Brunke MA, Brus SR, Burrows SM, Cameron-Smith PJ, Donahue AS, Deakin M, Easter RC, Evans KJ, Feng Y, Flanner M, Foucar JG, Fyke JG, Griffin BM, Hannay C, Harrop BE, Hunke EC, Jacob RL, Jacobsen DW, Jeffery N, Jones PW, Keen ND, Klein SA, Larson VE, Leung LR, Li HY, Lin W, Lipscomb WH, Ma PL, Mahajan S, Maltrud ME, Mametjanov A, McClean JL, McCoy RB, Neale RB, Price SF, Qian Y, Rasch PJ, Reeves Eyre JJ, Riley WJ, Ringler TD, Roberts AF, Roesler EL, Salinger AG, Shaheen Z, Shi X, Singh B, Tang J, Taylor MA, Thornton PE, Turner AK, Veneziani M, Wan H, Wang H, Wang S, Williams DN, Wolfram PJ, Worley PH, Xie S, Yang Y, Yoon JH, Zelinka MD, Zender CS, Zeng X, Zhang C, Zhang K, Zhang Y, Zheng X, Zhou T, Zhu Q (2019) The DOE E3SM coupled model version 1: overview and evaluation at standard resolution. J Adv Model Earth Syst 11(ja):2089–2129. https://doi.org/10.1029/2018MS001603

Hurrell JW (1995) Decadal trends in the North Atlantic oscillation: regional temperatures and precipitation. Science 269(5224):676–679

Hurrell JW, Kushnir Y, Ottersen G, Visbeck M (2003) The North Atlantic oscillation: climatic significance and environmental impact. In: Washington DC American geophysical union geophysical monograph series, vol 134. https://doi.org/10.1029/GM134

Hyndman RJ, Athanasopoulos G (2018) Forecasting: principles and practice. OTexts (ISBN-13: 978-0987507112)

Hyun S, G'Sell M, Tibshirani RJ (2016) Exact post-selection inference for changepoint detection and other generalized lasso problems. arXiv preprint arXiv:160603552

Imbrie J, Berger A, Boyle E, Clemens S, Duffy A, Howard W, Kukla G, Kutzbach J, Martinson D, Mcintyre A (1993) On the structure and origin of major glaciation cycles 2. The 100,000-year cycle. Paleoceanography 8:699–735

Jones PD, Jonsson T, Wheeler D (1997) Extension to the North Atlantic oscillation using early instrumental pressure observations from gibraltar and south-west iceland. Int J Climatol 17(13):1433–1450

Jouzel J, Masson-Delmotte V et al (2007) Orbital and millennial Antarctic climate variability over the past 800,000 years. Science 317(5839):793–796. https://doi.org/10.1126/science.1141038, http://science.sciencemag.org/content/317/5839/793, http://science.sciencemag.org/content/317/5839/793.full.pdf

Lisiecki LE, Raymo ME (2005) A pliocene-pleistocene stack of 57 globally distributed benthic $\delta 18o$ records. Paleoceanography 20:PA1003

McManus JF, Francois R, Gherardi JM, Keigwin LD, Brown-Leger S (2004) Collapse and rapid resumption of Atlantic meridional circulation linked to deglacial climate changes. Nature 428(6985):834

McWilliams JC (2008) The nature and consequences of oceanic eddies. In: Ocean modeling in an Eddying regime. Washington, DC, pp 5–15 (ISBN-13: 978-0875904429)

Milankovitch M (1941) Canon of insolation and the iceage problem. In: Koniglich Serbische Akademice Beograd Special Publication, vol 132 (ISBN: 9788617066190)

MPAS-Developers (2013) MPAS. http://mpas-dev.github.io/. Accessed 20 Jun 2016

Myers K, Lawrence E, Fugate M, Bowen CM, Ticknor L, Woodring J, Wendelberger J, Ahrens J (2016) Partitioning a large simulation as it runs. Technometrics 58(3):329–340

NAO (2019) North Atlantic oscillation. https://www.esrl.noaa.gov/psd/gcos_wgsp/Timeseries/Data/nao.long.data

NCEI (2018) Antarctic ice cores revised 800kyr co2 data. https://www1.ncdc.noaa.gov/pub/data/paleo/icecore/antarctica/antarctica2015co2.xlss

NinoData (2018) Nino 3.4. https://www.esrl.noaa.gov/psd/data/correlation/nina34.data

Oceanic Nino Index (2018) Equatorial pacific sea surface temperatures. https://www.ncdc.noaa.gov/teleconnections/enso/indicators/sst/

Petersen M, Williams S, Maltrud M, Hecht M, Hamann B (2013) A three-dimensional eddy census of a high-resolution global ocean simulation. J Geophys Res (Oceans) 118:1759–1774. https://doi.org/10.1002/jgrc.20155

Petersen MR, Williams SJ, Maltrud ME, Hecht MW, Hamann B (2013) A three-dimensional eddy census of a high-resolution global ocean simulation. J Geophys Res (Oceans) 118(4):1759–1774

Petersen MR, Jacobsen DW, Ringler TD, Hecht MW, Maltrud ME (2015) Evaluation of the arbitrary Lagrangian-Eulerian vertical coordinate method in the MPAS-Ocean model. Ocean Model 86:93–113. https://doi.org/10.1016/j.ocemod.2014.12.004, http://www.sciencedirect.com/science/article/pii/S1463500314001796

Petersen MR, Asay-Davis XS, Berres AS, Chen Q, Feige N, Hoffman MJ, Jacobsen DW, Jones PW, Maltrud ME, Price SF, Ringler TD, Streletz GJ, Turner AK, Van Roekel LP, Veneziani M, Wolfe JD, Wolfram PJ, Woodring JL (2019) An evaluation of the ocean and sea ice climate of e3sm using mpas and interannual core-ii forcing. J Adv Mode Earth Syst 11(5):1438–1458. https://doi.org/10.1029/2018MS001373

Quan J, Zhan W, Chen Y, Wang M, Wang J (2016) Time series decomposition of remotely sensed land surface temperature and investigation of trends and seasonal variations in surface urban heat islands. J Geophys Res Atmos 121(6):2638–2657

Ray BK, Tsay RS (2002) Bayesian methods for change-point detection in long-range dependent processes. J Time Ser Anal 23(6):687–705

Reeves J, Chen J, Wang XL, Lund R, Lu QQ (2007) A review and comparison of changepoint detection techniques for climate data. J Appl Meteorol Climatol 46(6):900–915

Ringler T, Petersen M, Higdon RL, Jacobsen D, Jones PW, Maltrud M (2013) A multi-resolution approach to global ocean modeling. Ocean Model 69(Supplement C):211–232. https://doi.org/10.1016/j.ocemod.2013.04.010, http://www.sciencedirect.com/science/article/pii/S1463500313000760

Samsel F, Petersen M, Geld T, Abram G, Wendelberger J, Ahrens J (2015) Colormaps that improve perception of high-resolution ocean data. In: Proceedings of the 33rd annual ACM

conference extended abstracts on human factors in computing systems. ACM, New York, CHI EA '15, pp 703–710. https://doi.org/10.1145/2702613.2702975

Sigman DM, Hain MP, Haug GH (2010) The polar ocean and glacial cycles in atmospheric CO2 concentration. Nature 466:47–55

Souza JMACD, De Boyer Montegut C, Le Traon PY (2011) Comparison between three implementations of automatic identification algorithms for the quantification and characterization of mesoscale eddies in the South Atlantic ocean. Ocean Sci 7(3):317–334

Stocker TF, Qin D, Plattner GK, Tignor M, Allen SK, Boschung J, Nauels A, Xia Y, Bex V, Midgley PM (2013) Climate change 2013: The physical science basis. Tech. rep., Cambridge University, p 1535

Talley LD (2011) Descriptive physical oceanography: an introduction. Academic Press, Amsterdam, pp 167–172

Thomson RE, Emery WJ (2014) Data analysis methods in physical oceanography. Newnes, Massachusetts, USA, pp 505–546 (ISBN-13: 978-0123877826)

Tibshirani RJ et al (2014) Adaptive piecewise polynomial estimation via trend filtering. Ann Stat 42(1):285–323

Tomé A, Miranda P (2004) Piecewise linear fitting and trend changing points of climate parameters. Geophys Res Lett 31:L02207

Turton TL, Ware C, Samsel F, Rogers DH (2017) A crowdsourced approach to colormap assessment. In: Lawonn K, Smit N, Cunningham D (eds) EuroVis workshop on reproducibility, verification, and validation in visualization (EuroRV3). The Eurographics Association. https://doi.org/10.2312/eurorv3.20171106

Verbesselt J, Hyndman R, Newnham G, Culvenor D (2010) Detecting trend and seasonal changes in satellite image time series. Remote Sens Environ 114(1):106–115

Volkov DL, Lee T, Fu LL (2008) Eddy-induced meridional heat transport in the ocean. Geophys Res Lett 35:L20601

von Storch H, Zwiers FW (2002) Statistical analysis in climate research. Cambridge, United Kingdom, pp 197–199 (ISBN-13: 978-0521450713)

Wang C, Deser C, Yu JY, DiNezio P, Clement A (2017) El niño and southern oscillation (enso): a review. In: Coral reefs of the eastern tropical pacific. Springer, pp 85–106

Ware C, Turton TL, Bujack R, Samsel F, Shrivastava P, Rogers DH (2018) Measuring and modeling the feature detection threshold functions of colormaps. IIEEE Trans Vis Comput Graph 25(9):2777–2790

Williams S, Petersen M, Bremer PT, Hecht M, Pascucci V, Ahrens J, Hlawitschka M, Hamann B (2011) Adaptive extraction and quantification of geophysical vortices. IEEE Trans Vis Comput Graph 17(12):2088–2095

Woodring J, Petersen M, Schmeiβer A, Patchett J, Ahrens J, Hagen H, (2015) In situ eddy analysis in a high-resolution ocean climate model. IEEE Trans Vis Comput Graph 22(1):857–866. https://doi.org/10.1109/TVCG.2015.2467411

Wunsch C (2015) Modern observational physical oceanography: understanding the global ocean. Princeton University Press, Princeton. Appendix A2