Environmental Modelling and Software 139 (2021) 104993

Contents lists available at ScienceDirect

Environmental Modelling and Software

journal homepage: http://www.elsevier.com/locate/envsoft

Evaluating the ability of numerical models to capture important shifts in environmental time series: A fuzzy change point approach

M.J. Hollaway [a,*], P.A. Henrys [a], R. Killick [b], A. Leeson [c,d], J. Watkins [a]

[a] UK Centre for Ecology and Hydrology, Lancaster Environment Centre, Bailrigg, Lancaster, UK
[b] Department of Maths and Statistics, Lancaster University, Lancaster, UK
[c] Lancaster Environment Centre, Lancaster University, Bailrigg, Lancaster, UK
[d] Data Science Institute, Lancaster University, Bailrigg, Lancaster, UK

ARTICLE INFO

Keywords:
Changepoints
Fuzzy-logic
Data science
Uncertainty
Evaluation framework

ABSTRACT

Numerical models are essential tools for understanding the complex and dynamic nature of the natural environment. The ability to evaluate how well these models represent reality is critical in their use and future development. This study presents a combination of changepoint analysis and fuzzy logic to assess the ability of numerical models to capture local scale temporal events seen in observations. The fuzzy union based metric factors in uncertainty of the changepoint location to calculate individual similarity scores between the numerical model and reality for each changepoint in the observed record. The application of the method is demonstrated through a case study on a high resolution model dataset which was able to pick up observed changepoints in temperature records over Greenland to varying degrees of success. The case study is presented using the DataLabs framework, a cloud-based collaborative platform which simplifies access to complex statistical methods for environmental science applications.

1. Introduction

The natural environment is a complex system that evolves through time in response to drivers such as climate change, economic change and social change (IPCC, 2018; Schröter et al., 2005). To understand the wide range of feedbacks and interactions involved in the earth system, numerical models (of varying complexities and computational requirements) are becoming increasingly relied upon. Recent advances in high powered computing have resulted in models that are capable of running at finer spatial and temporal resolutions and/or include more processes, and thus better represent the dynamic natural environment (Collins et al., 2011; Gutjahr et al., 2019; Hu et al., 2018; Savage et al., 2013; Swart et al., 2019). These developments are particularly important as many environmental processes are local in nature and exhibit high spatial variability, e.g. air pollution episodes, localised heavy rainfall, or ice sheet melt. Therefore, in theory, finer resolution models should be able to better capture this variability than their coarser scale counterparts. Furthermore, finer scale models are able to provide high resolution predictions of future environmental change, under a warming climate. However, with this enhanced capability comes increased scrutiny of uncertainty in the model structure, parameters and outputs (Beven, 2006) and how these uncertainties are communicated to model users, developers and ultimately decision makers.

This increasing need to quantify uncertainty in outputs, along with the rapid rise in the volume and variety of 'big data' in environmental science, has resulted in increasingly complex datasets from which scientists wish to answer key questions. The field of data science provides potential solutions to extract information from ever growing complex environmental datasets (Tso et al., 2020) along with providing the ability to drive new scientific insight and better constrain uncertainties (Hollaway et al., 2018). However, utilisation of such techniques often requires experts from different domains to work together in an open and transparent way. Therefore there is a need to facilitate such collaborative efforts in order to use complex statistical methods to answer environmental science challenges, an example of which is the evaluation of complex numerical models.

Typically, numerical models are evaluated against observations from a variety of different sources (E.g. Sensor networks or satellite data) with global metrics often employed to assess how well the model captures the overall behaviour of the system (Gleckler et al., 2008; Pincus et al., 2008). These integrated quantities often include the coefficient of determination (R^2) which assesses how well a numerical model

[*] Corresponding author.
E-mail address: mhollaway@ceh.ac.uk (M.J. Hollaway).

https://doi.org/10.1016/j.envsoft.2021.104993

Available online 19 February 2021

represents the overall variance seen in the observations or the root mean square error (RMSE) which evaluates the overall magnitude of the forecast errors by the numerical model. Whilst providing a good overall summary of model performance, these metrics do have limitations in their use. For example, if the numerical model consistently over/under predicts the observations it can still return a high R^2 (Krause et al., 2005). Therefore it is often used in conjunction with other metrics such as RMSE, however this itself is scale dependant and therefore cannot be used to compare different model outputs. Furthermore, many processes in the natural environment are local in both space and time, and therefore a good numerical model performance using integrated metrics does not necessarily translate to the model being a good predictor at the local scale. Examples of such local scale events can be seen in long term atmospheric records over varying temporal scales. These can range from abrupt shifts in the long term statistical properties of air temperature (e.g. the rapid rise in global mean temperature in the late 1990s (IPCC, 2018), through to shorter term shifts associated with seasonal variability. Therefore, if a numerical model is to be classed as suitably representing the reality of the natural environment it should be able to capture variability at a range of spatial and temporal scales with an acceptable degree of accuracy.

In recent years there has been a move towards incorporating advanced statistical techniques into climate model evaluation, for example, the analysis of extreme events (Leeson et al., 2018). To date however, no previous studies have focussed on the degree of accuracy to which models capture the timing of changepoints in the long term temperature record.

This study presents a new approach to numerical model evaluation that utilises changepoint analysis (detection of shifts in the statistical properties of time series data) to assess the ability of a model derived time series to capture different modes of temporal variability seen in the observed record. Changepoints are first identified in the modelled and observed time series, then their locations – together with an estimate of uncertainty calculated through bootstrap samples - are used in combination with fuzzy logic to develop a metric which captures the degree to which the location of modelled changepoints agrees with those identified in the observations. The method is then demonstrated using a case study of its application to a high resolution model reanalysis dataset designed to simulate the climate of the Greenland Ice Sheet.

2. Materials and methods

2.1. Changepoint detection in discrete time series

Changepoint detection is essentially a statistical method that is used to estimate the point (or points) in a time series where there is an abrupt shift in its statistical properties such as the mean, variance or both, conditional on an assumed model (Eckley et al., 2011). For a discrete time series of ordered data, $y_{1:n} = (y_1, y_2, \ldots\ldots y_n)$, the optimal location and number of changepoints (m) are identified based on a chosen cost function and a penalty to avoid over fitting. The number and locations of the changepoints in this study are identified using the pruned exact linear time (PELT) algorithm (Killick et al., 2012). More detail on PELT can be found in Killick et al. (2012), but in short, the algorithm performs an exact search of the time series and considers all possible combinations for any number of changepoints (up to a maximum, specified using a minimum segment length). Here, the modified Bayesian information criterion (MBIC; Zhang and Siegmund (2007)) is used in combination with PELT to detect the optimal number of changepoints in the time series. This helps reduce the identification of short segments as the MBIC penalty balances the overall fit against the length of each segment.

2.2. Estimating confidence intervals on changepoints

In order estimate the uncertainty in the locations of the changepoints identified using PELT, confidence intervals (CIs) are constructed. These are calculated as follows:

1. For changepoint location, k_i, isolate the segments to its left ($y_{k_{i-1}+1}$, …, y_{k_i}) and right (y_{k_i+1}, …, $y_{k_{i+1}}$) and generate a bootstrap sample of each segment separately giving ($\widetilde{y}_{k_{i-1}+1}$, …, $\widetilde{y}_{k_i}$) and ($\widetilde{y}_{k_i+1}$, …, $\widetilde{y}_{k_{i+1}}$) respectively.
2. Combine the output from step 1 and treat as a time series with a single changepoint. This is the series defined by $\widetilde{y}_{k_{i-1}+1}, \ldots, \widetilde{y}_{k_{i+1}}$, with length l_i.
3. Estimate the location of the changepoint for this sample using the same approach used to calculate the original m changepoints.
4. Repeat steps 1 to 3 N times.
5. Calculate the 2.5 and 97.5 percentiles of the N bootstrap samples giving the lower and upper confidence interval for the changepoint location.
6. Repeat steps 1–5 for each of the m changepoints in the time series.

This will provide 95% confidence intervals for each changepoint identified by the PELT algorithm. If another confidence interval range (E.g. 90% or 99%) is required, the percentiles set in step 5 can be changed accordingly.

2.3. Comparing changepoints between 2 time series

In order to compare the timing of changepoints between two time series and take account of the uncertainty in the estimation of the changepoint locations (represented by the constructed CIs) a new metric is proposed that is based around fuzzy logic (Matthé et al., 2006; Meyer and Hornik, 2009; Zadeh, 1965).

Here, the aim is to evaluate whether the numerical models are able to reproduce changepoints in the observational record and thus demonstrate the model's ability to capture key processes in the environmental system in their outputs. Note therefore, that the changepoints (and associated CIs) in the observed record are assumed to be the 'truth' and are used as the benchmark with which to evaluate the model.

The observed and modelled changepoints are converted into triangular fuzzy numbers centred on the changepoint location with the corresponding upper and lower confidence intervals as boundaries. A normalised similarity score is computed between each observed and climate model fuzzy pairs with a score of 0 indicating no similarity and a score of 1 indicating perfect similarity (Fig. 1). In this case the measure is the Jaccard similarity score which is the ratio between the fuzzy intersection (i.e. the area containing membership to both fuzzy numbers) and the fuzzy union (i.e. the area containing membership of either of the fuzzy numbers being compared) of each pair. The performance of the climate model at each observed changepoint is recorded as the model changepoint that returns the highest similarity or 0 if no points show similarity. If more than one modelled changepoint shows similarity with an observed, the one returning the highest score will be associated with that observed changepoint. If there are more modelled changepoints than observed, some will not be included in this pairwise comparison. Conversely, if there are more observed changepoints than model changepoints, not every observed will have a match. This approach ensures that each model changepoint is only associated with one observed.

The similarity scores are then summed across the total number of observed changepoints to give an overall score for the given climate model at that site and normalised to unity by dividing by the total number of observed changepoints, thus allowing comparison of performance across different sites. The total score thus ranges from 0 (no observed changepoints captured) to 1 (all observed changepoints captured perfectly). It should be noted that if the timing of the changepoint is captured perfectly by the climate model but the CIs differ, the normalised score will be lower than 1 as different CIs indicate that the statistical representations of the two time series are different.

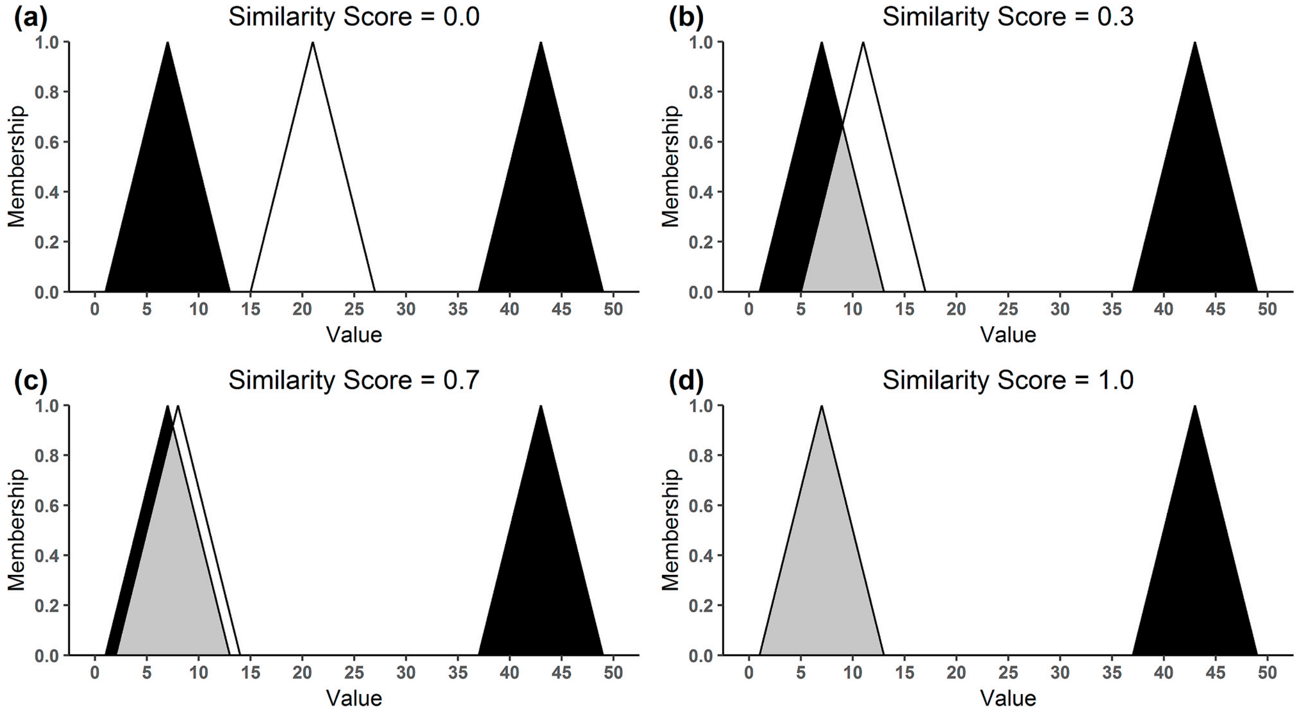

Fig. 1. Schematic to show similarity score calculation for changepoints from 2 different time series. The solid black triangles represent the changepoint and associated confidence intervals for the first time series (assumed to be the truth). The solid white triangles represent the changepoint and associated confidence intervals from the second time series (the time series being evaluated). The grey triangle represents the Jaccard index/similarity score. A-D represent increasing value of similarity between changepoints from none (0) to perfect (1). Two black triangles are shown to represent the situation that not all changepoints will have a match.

In addition, the proportion of climate model changepoints that are observed changepoints is calculated as well as the proportion of observed changepoints that are captured (using the above criteria). Here an observed changepoint is classed as captured if there is intersection with the CI of a climate model changepoint. These summary metrics provide an overall view of how well the climate model captures the statistical properties of the observed dataset in terms of its marginal behaviour in time, with the individual similarity scores highlighting which local (in time) events are particularly well, or poorly, captured by the model.

Missing data in the observed time series are treated as "missing at random" and therefore are ignored in the analysis. For consistency, where there are missing data in the observed time series, the corresponding data is removed from the numerical model time series. In order to ensure a sufficient sample size for processing for changepoints, a minimum threshold of 100 complete timesteps of data was assigned, with any segment not meeting this threshold discarded from the analysis.

2.4. *Case study: Application to the evaluation of a high resolution model reanalysis product over the Greenland ice sheet*

The ERA5 dataset is a state-of-the-art global reanalysis product developed by the European Centre for Medium Range Weather Forecasting (ECMWF) which provides a detailed record of global atmospheric conditions from 1979 through to present day (Hersbach et al., 2020). The dataset is based around the Integrated Forecasting System (IFS) and utilises data assimilation techniques using observations and satellite data to produce the final product representing the best state of global meteorological conditions at hourly intervals and ~31 km horizontal resolution. The ERA5 dataset supersedes the ERA-Interim reanalysis dataset (Dee et al., 2011) which operates at lower temporal (6 hourly) and spatial (~79 km) resolutions and has been commonly used to force regional climate models, including over Greenland (Fettweis et al., 2013, 2017). Given the high temporal and spatial resolution of ERA5, along with the use of data assimilation of observations, it is reasonable to expect that ERA5 should perform well, in terms of accurately representing local scale temporal variability when compared to the observed record.

The changepoint evaluation method described above is applied to ERA5 air temperature time series data, in order to evaluate its performance against observations from automatic weather stations (AWSs) from the Greenland Climate Network (GC-Net; Steffen et al., 1996). This network of 18 AWSs provides, amongst other key meteorological variables, long term hourly temperature records from the mid-1990s through to the present day. For this work, hourly data is aggregated to daily mean temperature for the 14 AWSs with the most complete records for the period 2000 through to 2017. These time series are then analysed using the approach presented above and the locations of changepoints established. As the changepoint detection algorithm is conditional on the assumed statistical model used to represent the temperature time series, if the underlying data structure is poorly understood, erroneous results can be produced (Beaulieu and Killick, 2018). The daily temperature time series in this study exhibit seasonality and auto-correlation between days, which must be accounted for in the fitting of the changepoint algorithm. In this case, pre-screening of the data revealed a first order auto-regressive model (AR1) fitted to the data enabled the best detection of changepoints in the time series. Therefore an AR1 model is fitted to the data prior to the changepoint analysis. PELT is then used on the residuals of the AR1 time series to identify changepoint locations based on a change in variance over time. The same process is then applied to the model time series and the ability to capture the observed changepoint evaluated. As the focus of this study is to develop a new approach to evaluating the ability of climate models to pick up local scale changes in the statistical properties of an observed temperature time series, a marginal approach in space is taken and the changepoint analysis is applied to each of the 14 GC-Net AWS sites independently.

For this case study, all analysis were conducted using R version 3.5.3 (R Core Team, 2019) and executed within a Jupyter notebook. Version

2.3.1 of the 'changepoint' R package (Killick and Eckley, 2014; Killick et al., 2019) was used to detect changepoint locations using PELT (Killick et al., 2012) and version 1.0–18 of the 'sets' package (Meyer and Hornik, 2009) was used to calculate the fuzzy based evaluation scores.

2.5. Implementation of the method into the DataLabs framework

In order to champion open science, collaboration and ease access to complex statistical methods for environmental science applications, the method presented here is implemented into the DataLabs framework (Hollaway et al., 2020). These tools sit in a cloud based computational environment that can scale in resources depending on the complexity and volume of data that is required to be processed. Furthermore, DataLabs can provide access to analytical methods at different levels of abstraction ranging from raw code to a graphical user interface that drives the workflow. This can foster collaboration between scientists of different areas of expertise in an open and transparent environment, seen as a key advancement in the field of data science.

3. Results

3.1. Identification of changepoints in the observed time series

A summary of the changepoint locations estimated using PELT between 2000 and 2017 inclusive is shown in Fig. 2. In general, most of the stations show 2 changepoints in each calendar year with the first typically falling during the spring (March to May) and the second falling in late summer/early autumn (August to October). From the PELT fit (not shown), it is clear that the time series tends to be more variable during the winter months and less so during summer. This allows physical inference to be made from the timing of the changepoints, which potentially correspond to the onset and end of the ice melt season in a given year. Overall, where there is data available (periods of missing data are highlighted in grey in Fig. 2) this pattern of changepoint timing holds for most stations. The exception is Summit which returns no estimated changepoints for the period 2009–2011. During the estimation of the changepoint locations (using the bootstrap approach described above), the mean temperatures of each segment are also calculated. The resulting warmest segments at each site (red shading in Fig. 2) indicate that the 2012 'summer season' (as inferred from the changepoint locations) is warmest at 4 of the 14 measurements stations. It is known that 2012 was particularly high melt year on Greenland (Nghiem et al., 2012) and thus this is a potential key event to focus on for the evaluation of the numerical model.

3.2. Evaluation of ERA5 in terms of capturing observed changepoints

The changepoint analysis is repeated on the ERA5 temperature time series (for the grid cells corresponding to each AWS site) to identify changepoint locations (and associated confidence intervals). As per the observations, each location is treated independently in space. The new fuzzy logic based metric is then used to evaluate the model's ability to capture the observed changepoints at each station (Fig. 3).

Overall, ERA5 captures the timing of the changepoints in the observations with varying degrees of accuracy returning overall similarity scores that range from 0.17 at JAR2 to a best performance of 0.51 at Summit. This indicates that there are fairly significant differences between the timing of the changepoints seen in observations and corresponding estimated changepoints in the ERA5 simulated time series. Furthermore, lower scores tend to occur at sites where the percentage of observed changepoints captured by ERA5 is low (i.e. no intersection at all between the modelled and observed changepoint confidence intervals) or the percentage of model changepoints that are true changepoints is also low (Fig. 3).

A closer look at the individual changepoint evaluation scores (corresponding to each changepoint in the observations time series) at the 2 best performing (NASA-U and Summit) and the 2 worst performing sites (NASA-SE and JAR2) provides further information to inform interpretation of the overall performance metric (Fig. 4).

At NASA-U and Summit, ERA5 tends to produce high scores for capturing each individual changepoint, particularly in the latter half of the record (2012 onwards) where similarity scores of 0.65 or higher are

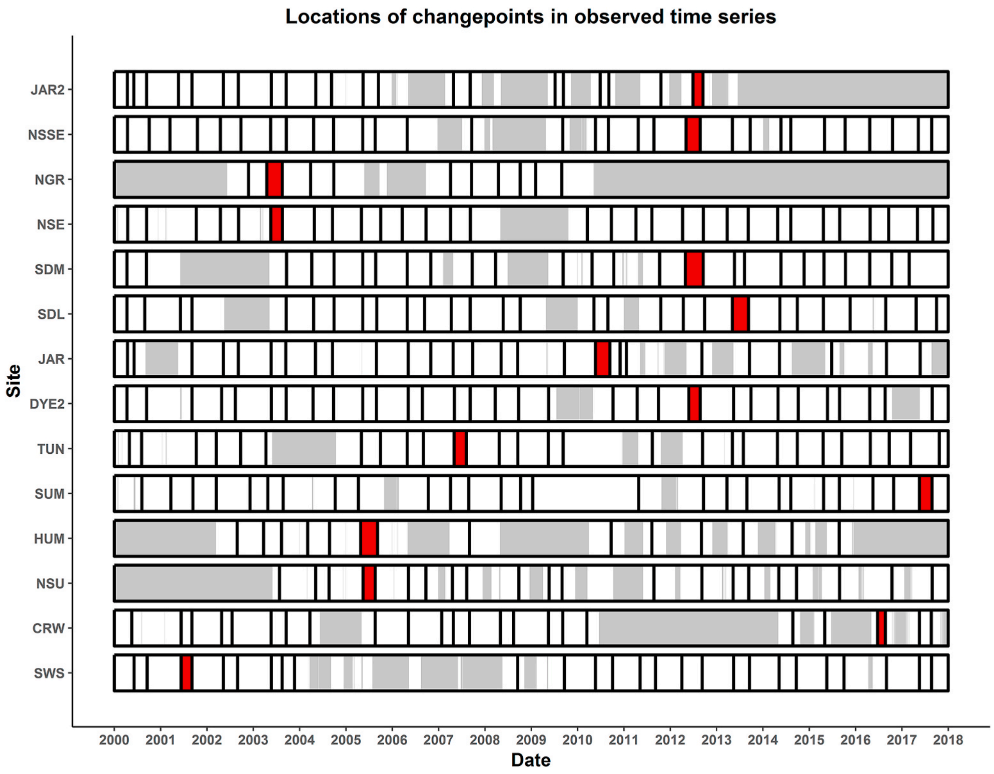

Fig. 2. Location of changepoints in the observed temperature time series at the 14 GC-Net stations for the time period 1/1/2000 to 31/12/2017. The y-axis shows the station name and the x-axis shows time. Each bar represents a single station with the vertical black lines indicating a changepoint and the grey shading indicating missing data from the record. The red shading indicates the segment (based on the changepoint locations) with the highest mean temperature across all years for each station.

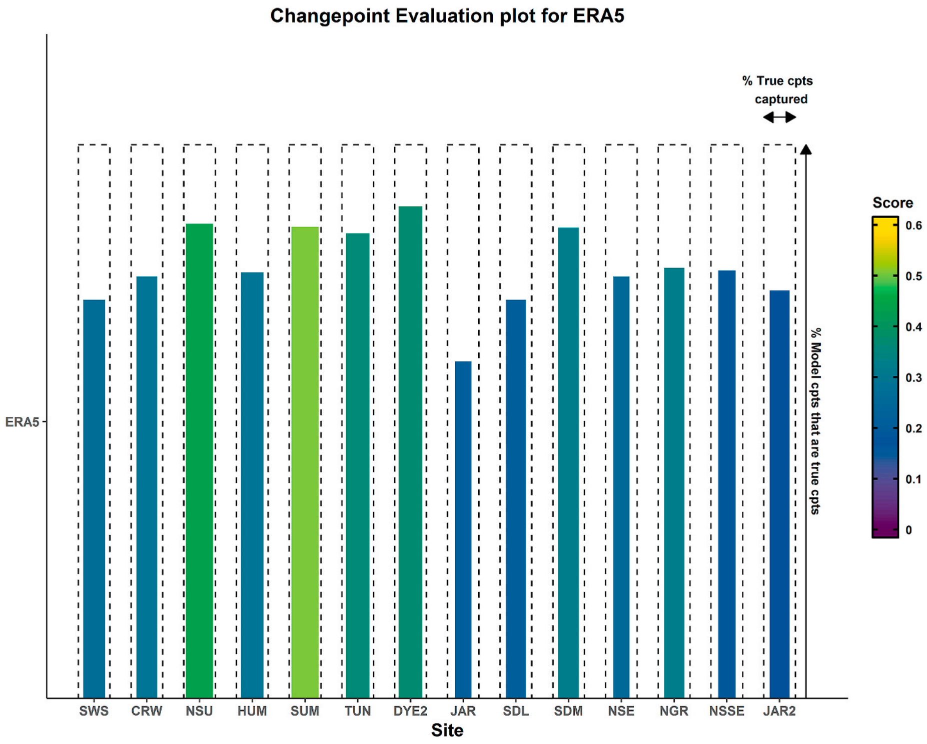

Fig. 3. ERA5 performance at each GC-Net AWS, defined by the three summary metrics outlined in Section 2. The coloured box width shows the percentage of true changepoints captured and the coloured box height shows the percentage of model changepoints that are true changepoints. The dashed boxes represent a perfect score of 1 on each of these measures. The colour of each box represents the overall normalised score for the changepoint evaluation metric. Note that the scale for the changepoint evaluation metric has been capped at 0.6.

frequently returned. At Summit, from 2012 to 2017, the timing of the observed changepoint is captured perfectly on 4 occasions (out of 11) and within 2 days on 4 other occasions, resulting in similarity scores of 0.65–1.0. Here the range in scores is driven by differences in the confidence intervals of the observed and climate model changepoints. At NASU-U, for the same period, the timing of observed changepoints is captured perfectly on 3 occasions (out of 7). For the remaining 4 occasions in this period, the model misses the timing by 2–10 days, including failing to capture the observed changepoint in 2017 at all (i.e. returns a similarity score of 0). Again, the range in similarity scores at NASA-U for 2012–2017 (0.18–0.82 for non-zero similarity scores) is also driven by differences in the confidence intervals. ERA5 tends to capture a similar proportion of the observed changepoints at each site (Fig. 3), resulting in the higher overall metrics when normalised across all changepoints (0.51 at Summit and 0.44 at NASA-U).

At NASA-SE and JAR2, ERA5 captures far fewer of the observed changepoints, and returns lower similarity scores when they are captured (Fig. 4). At NASA-SE, similarity scores of 0.45–0.82 are returned for 2015–2017 inclusive where the climate model captures the timing of the observed changepoints either exactly or within one day. Despite this, similarity scores of less than 1.0 are due to differences in the confidence intervals overlap. However, from 2003 to 2006, 7 changepoints are estimated in the observed time series, with ERA5 failing to estimate any at all. Conversely at JAR2, with the exception of 2010, ERA5 and observed changepoint pairs occur in most years (Fig. 4), however the model tends to return lower similarity scores (0.03–0.75) with ERA5 failing to capture the timing of the observed changepoints by between 2 and 16 days. Overall, despite capturing a smaller proportion of observed changepoints (17 out of 31) at NASA-SE than at JAR2 (14 out of 22), the generally higher similarity scores at NASE-SE leads to a slightly better overall performance (0.19 compared to 0.17 at JAR2).

3.3. *Focus on capturing key events – the 2012 warm year*

As highlighted in Section 3.1, the changepoint analysis on the 14 AWS temperature time series identified the summer months of 2012 as the warmest on average at 4 of the stations. Furthermore, previous studies have highlighted that 2012 was an unusually warm year and one of extreme melt over the Greenland Ice Sheet (Hanna et al., 2014; Nghiem et al., 2012). As such, this provides an ideal localised (in time) event to evaluate how well ERA5 captures the timings of these changepoints. This can be done by critiquing the similarity scores at sites where the changepoint locations in the observations dataset could be interpreted as the start and end of the summer months (Table 1). The timing of the summer months is captured best at Southdome (Fig. 5d) with the model capturing the start of the season (evaluation score of 0.712) better than the end (0.365). Here, the start of the season is captured perfectly with the non-perfect similarity score being driven by the difference in confidence intervals (lower panel Fig. 5d). ERA5 estimates the end of the season as being 15 days later than the observations, however given the relatively large uncertainty in the changepoint location, there is a degree of overlap in the confidence intervals resulting in the similarity score of 0.365.

The timing of the summer season is captured fully (i.e. ERA5 returns non-zero evaluation scores for both the start and the end) at 4 of the remaining 5 sites, Swisscamp (Fig. 5a), DYE2 (Fig. 5b), Saddle (Fig. 5c), and NASA-SE (Fig. 5e). Overall, the timing of the start of the summer season is captured better than the end at DYE2 (0.528/0.045 for the start/end), Saddle (0.170/0.098) and NASA-SE (0.101/0.003) with only DYE2 capturing the timing perfectly. The reverse is seen at Swisscamp where the end of the season is captured very well (0.719) with the timing captured perfectly and the non-prefect similarity score again driven by the overlap in the confidence intervals. The start of the season is only just captured (returns a non-zero score of 0.016), estimating it to be about a month later than the observations (Fig. 5a). At the remaining sites, ERA5 performs poorly, either not estimating any changepoints at all for 2012 at NASA-E (not shown in Fig. 5) or failing to capture the start of the summer at all at JAR2 (the model estimating the changepoint around a month earlier than the observations with no overlap of the confidence intervals). The end of the summer at JAR2 is also captured poorly with ERA5 estimating the changepoint around 10 days earlier than the observations, returning a low similarity score (0.106). The uncertainty in the changepoint locations (as signified by the confidence intervals), indicates that the underlying statistical representation of the

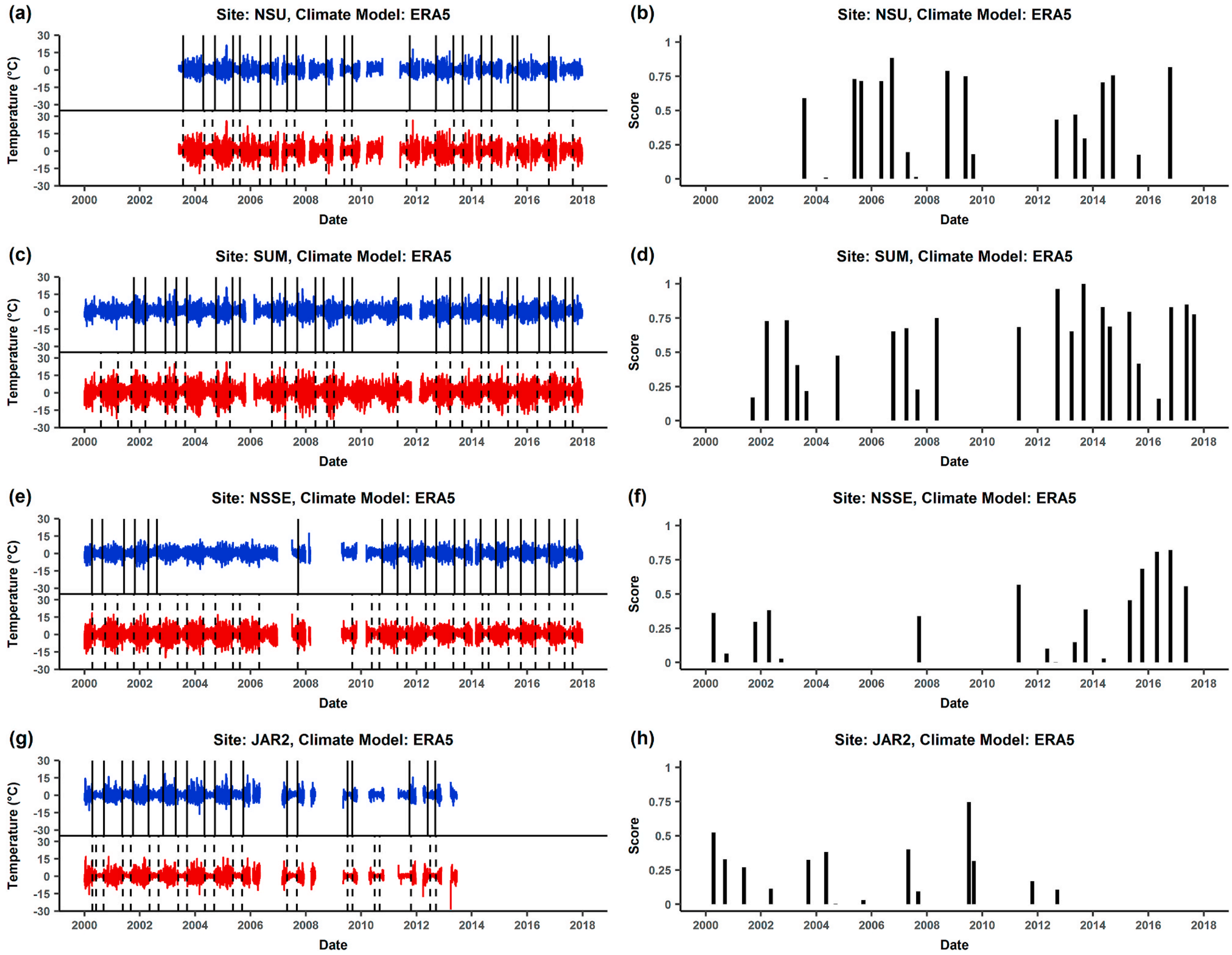

Fig. 4. Evaluation of ERA5 at the 2 best performing sites (NSU and SUM) and the 2 worst performing sites (NSSE and JAR2) using the changepoint evaluation metric. The left column (a, c, e, g) shows the residuals of the climate model time series (after the AR1 model fit) and estimated changepoint locations (blue lines and vertical solid black lines respectively) and the corresponding observed residual time series and estimated changepoint locations (red lines and vertical dashed black line respectively). The right hand column (b, d, f, h) shows the individual similarity scores for each corresponding site on a normalised scale.

time series from ERA5 differs to that of the observations at these sites, leading to the discrepancies in the timings of the changepoints. For example, at JAR2, at the start of summer the model displays much larger uncertainty (Fig. 5f) in the changepoint location indicating that the variance of the time series differs greatly to that of the observations. This could indicate the methods used to produce the reanalysis time series potentially fail to capture the variability seen in reality on this particular event and location. Furthermore, this also suggests that the summer season is offset a month earlier in the ERA5 model and could have implications for ice melt should the reanalysis product be used to drive other regional models simulating ice dynamics.

During the calculations of the confidence intervals for the changepoint locations, the bootstrap samples were also used to estimate the corresponding segment means (and associated confidence intervals), given the uncertainty in the changepoint locations. These were used to compute similarity scores using a similar fuzzy union approach to the changepoint evaluation (Table 1). Of the sites that capture the summer season, only 2 return non-zero similarity scores (for the segment means); Swisscamp and Southdome. At Southdome, despite capturing the timing of the season well, ERA5 does not estimate the mean temperature well; only returning a score of 0.112. In contrast at Swisscamp, the similarity score for the mean is slightly higher (0.125) however ERA5 fails to capture the start of the season well.

4. Discussion

The method presented in this study provides a new approach for assessing numerical models by evaluating how well they capture observed local scale temporal events (i.e. changepoints). The fuzzy logic based metric also factors in the uncertainty in the changepoint locations (represented by bootstrapping confidence intervals) in both datasets into the evaluation. The application of the method is demonstrated in the evaluation of the ERA5 reanalysis dataset using temperature data from the GC-Net monitoring network in Greenland.

4.1. Understanding numerical model performance at capturing local scale temporal events

Using the normalised summary metric, ERA5 returns overall similarity scores ranging from 0.17 to 0.51 (Fig. 3), when averaged across all observed changepoints at a given site. As similarity scores are also calculated for each individual changepoint, these can give an indication of potential reasons for the model performing poorly. At sites where ERA5 performs the best (Summit (0.51) and NASA-U (0.44)) the model tends to score 0.65 or higher for many changepoints in the latter half of the record. This indicates that ERA5 not only captures the timing of the changepoints well but there is also strong overlap in the confidence intervals between the changepoints of each time series. Strong overlap

Table 1
Timing of observed changepoints in 2012 at a subset of GC-Net sites along with similarity scores for corresponding changepoints (if identified) from the ERA5 model. The evaluation score for ERA5's ability to capture the segment mean is also presented. Here a score of zero indicates either the model does not capture the mean or the model does not estimate a corresponding full summer season. Note, only the sites where changepoints are identified in the observation time series for the start and the end of the summer season are shown. Similarity scores are presented to 3 decimal places due to very low scores returned for some sites (E.g. NSSE).

Site	Start Date	End Date	ERA5 similarity score (start)	ERA5 similarity score (end)	ERA5 similarity score (mean)
SWS	02/04/2012	09/09/2012	0.016	0.719	0.125
DYE2	27/05/2012	25/08/2012	0.528	0.045	0.000
SDL	13/04/2012	29/09/2012	0.170	0.098	0.000
SDM	30/04/2012	17/09/2012	0.712	0.365	0.112
NSE	07/04/2012	18/09/2012	0.000	0.000	0.000
NSSE	06/05/2012	25/08/2012	0.101	0.003	0.000
JAR2	28/06/2012	15/09/2012	0.000	0.106	0.000

indicates that, accounting for uncertainty in the changepoint location, there is little difference in the statistical properties of the time series (in this case variance) suggesting ERA5 is representing reality well for that particular event. At sites where ERA5 performs poorly (e.g. NASA-SE) the model, despite returning reasonably high scores when it does capture changepoints, fails to capture the majority of changepoints at the start of the record leading to an overall low similarity score. Interestingly, when compared across all sites the model tends to perform better at sites located in the dry snow zone of the ice sheet, where there is very little melting, and poorly at the ablation zone sites (Swisscamp, JAR and JAR2) where there tends to be the most melt (Leeson et al., 2018), which plays an important role in the surface mass balance of the ice sheet. Therefore, as reanalysis datasets are often used to drive detailed regional climate models that include detailed representations of melt processes (E.g. MAR (Fettweis et al., 2017)), the failure to capture local scale events by datasets such as ERA5 could propagate through the model chain. Therefore consistency across sites in either the geography or timing of poor performance can be used to aid development efforts as to potential processes and feedbacks in the ERA5 model that require further investigation. This offers potential advantages over using global metrics to evaluate model performance.

A comparison with evaluation using the traditional integrated metrics yields another advantage of utilising this local event based method as part of the model evaluation workflow. Table 2 shows the associated R^2 for each site indicating that ERA5 generally captures the general temperature trends well (R^2 of 0.65–0.97). Overall, the sites where the R^2 indicates good performance, tend to also perform better at capturing the local scale events. However, there are some notable exceptions at JAR (0.88 R^2) and JAR2 (0.72 R^2) that return good performance on the global metrics but produce some of the lowest scores using the changepoint metric (0.22 and 0.17 respectively). Therefore, the good performance across the record does not translate to the ability to capture local scale events well. This agrees with previous work where fidelity at global scales does not always translate to finer scale events when using complex, computationally heavy models (Medley et al., 2013).

The individual similarity scores can also evaluate how well the model captures key events that are known to be important. In this case, using the changepoint locations, 2012 was identified as a potential anomalous warm year with ERA5 able to capture this with mixed results.

The summer season is captured at 5 of the 7 sites (that recorded changepoints that could be inferred as the start and end of the summer season) with either missed changepoints or poor capturing of timing leading to poor performance at the other 2 sites. Further to this, as the overall similarity scores are normalised across the record, if there is a known event that is of critical importance for the model to capture, greater weightings can be applied when the overall score is calculated. As this method enables focus on localised events in time, it could be used to flag other potential events (in this case summer seasons) that could also be anomalous, and are critical for the climate model to capture.

The changepoint analysis used in this study suggests the 2012 summer season is warmest (over the 20 year record of available data) across a large number of sites which corresponds to one of the largest melt years in history over Greenland (Hanna et al., 2014; Nghiem et al., 2012). Therefore, the method could be run on longer term data records to highlight other extreme summer years in the past and provide further constraints on model evaluation. The method utilised here, can use changing patterns in the statistical properties of the data to detect the onset of particular events (e.g. in this case summer season). This can provide a more robust constraint on how well the model captures local scale temporal changes rather than simply looking at annual time series or arbitrary definitions of a season. In this study, the method was applied to detect changepoints on a seasonal scale, however with the availability of long term records it could be adapted to critique the data for longer term changes. This could be combined with using the technique to evaluate multiple climate models of different complexities and resolutions to assess whether the incorporation of more processes leads to better model fidelity at capturing local scale temporal events and would constitute a natural follow on study to this work.

4.2. *Issues to consider*

Despite the advantages this new model evaluation method offers, there are some issues that need to be considered in its application, particularly when using highly seasonal environmental data. Typically, datasets from the environmental domain exhibit high levels of seasonality, are non-stationary and can exhibit high levels of auto-correlation. Changepoint detection algorithms can be sensitive to all of these things and lead to the identification of spurious changepoints in the time series (Beaulieu et al., 2012; Beaulieu and Killick, 2018) and overfitting of the algorithm. This can often make it difficult to make any inference as to the physical cause of the changepoint (e.g. onset of a particular season or a change in instrument calibration or location). Therefore, due to the changepoint algorithm being conditional on the underlying model that is specified for the data (AR1 in this study), if the underlying data structure is not adequately understood, the method could produce erroneous results as discussed in Beaulieu and Killick (2018).

The GC-Net temperature data and corresponding model data in this work were investigated for seasonality and autocorrelation prior to fitting of PELT and it was concluded that an AR1 model was the best model to fit to the data. However, not all environmental datasets would be suitable for fitting an AR1 model and therefore some prior exploration of the data is required. Finally, the amount of missing data in the time series can also impact the ability to detect changepoints and result in issues of overfitting of PELT. The method presented here allows specification of the minimum length of continuous data for the time series to have for PELT to be applied to detect changepoints. In this case, it was set at 100 days to focus on seasonal changes, however this setting can be varied dependent on the temporal scale of features which want to be investigated.

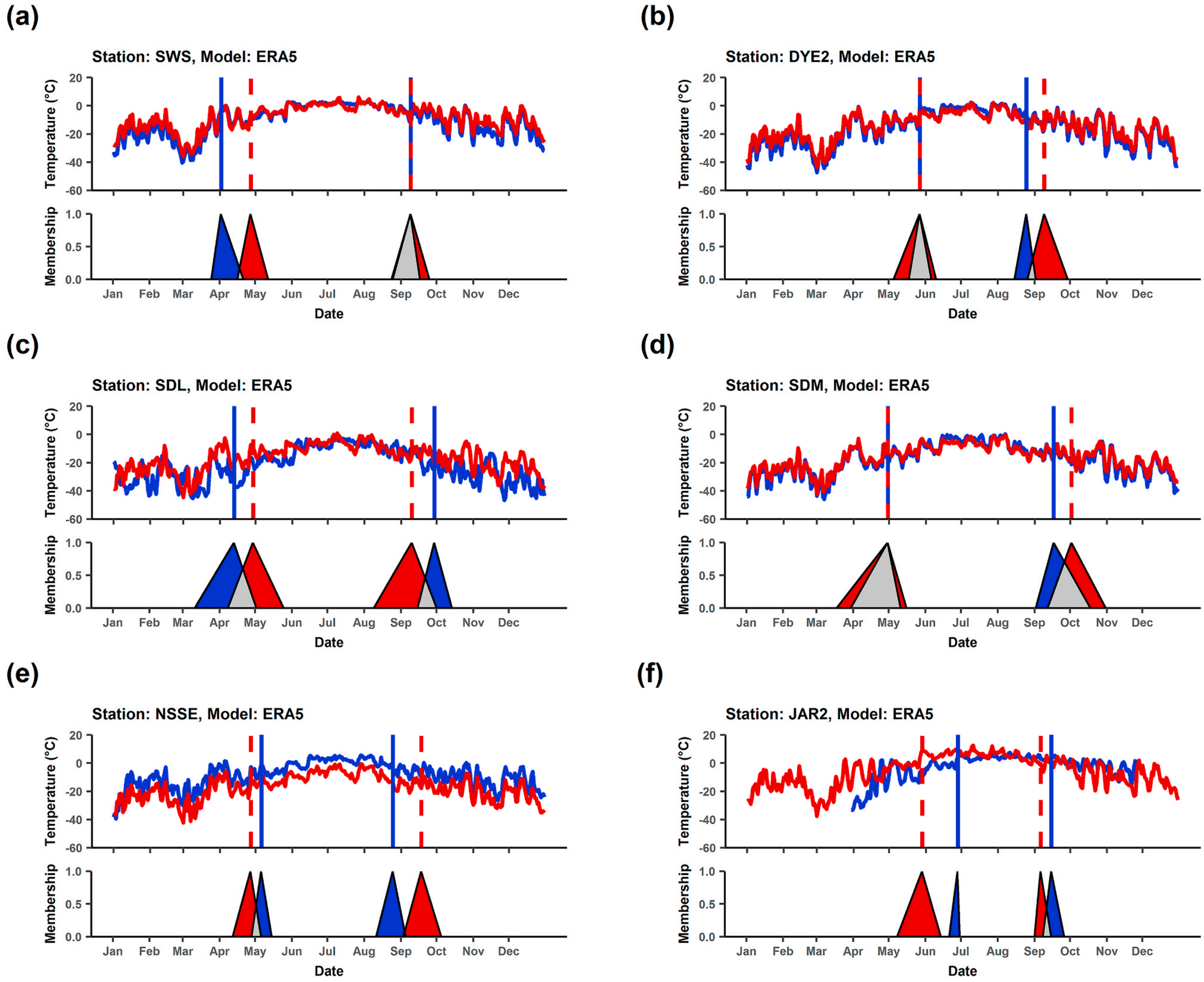

Fig. 5. Evaluation of ERA5 for 2012 at subset of sites. The observed temperature time series is shown by the blue line and the ERA5 time series is shown by the red line. The blue solid vertical lines show the changepoints in the observations and the solid vertical red lines show the estimated changepoints from ERA5. The blue triangles represent the confidence intervals around the observed changepoint (triangle point) and the red triangles represent the confidence intervals around the modelled changepoint (triangle point). The grey shaded areas highlight regions of overlap between the confidence intervals of observed and ERA5 changepoints (I.e. returns an evaluation score). Note data for site NSE is not shown in the plot despite the changepoints for the start and end of the summer season being identified in the observations (Table 1). This is because the ERA5 did not estimate any changepoints for the same period.

Table 2
ERA5 evaluation at the GC-Net sites showing R^2 and overall similarity score using the changepoint metric.

Site	ERA5 changepoint metric	ERA5 R^2
SWS	0.27	0.94
CRW	0.29	0.88
NSU	0.44	0.96
HUM	0.29	0.91
SUM	0.51	0.96
TUN	0.36	0.97
DYE2	0.37	0.96
JAR	0.22	0.88
SDL	0.22	0.65
SDM	0.32	0.96
NSE	0.26	0.91
NGR	0.33	0.94
NSSE	0.19	0.72
JAR2	0.17	0.72

4.3. Contribution to data science solutions to environmental science challenges

To facilitate access to the complex workflow of integrating statistical (I.e. Changepoint analysis) and process based models (i.e. ERA5), for a range of users with different expertise, the analytical workflow presented here is implemented into the DataLabs framework. These cloud based tools provide a consistent and coherent environment for scientists from different backgrounds (e.g. environmental scientists, statisticians and computer scientists) to come together and collaborate in the development of novel methods. Furthermore, through visualisation dashboards such as RShiny (Chang et al., 2018), users can run complicated analytical methods without having to access complex code (Fawcett, 2018; Slater et al., 2019). This enables the dissemination of results to a wide range of user abstractions. The method presented above sits in a modular series of R Markdown (Allaire et al., 2018) notebooks that perform the data extraction, the changepoint analysis itself and calculation of the fuzzy based evaluation metric. Finally, an RShiny application sits over the R code in the notebooks which enables the user to explore the performance of the model at each site (an example of the application is available with this manuscript). The open and transparent nature of the DataLabs enables the method to meet FAIR (Findable, accessible, interoperable and reusable (Wilkinson et al., 2016)) standard recommended for scientific data. This allows users to understand the assumptions made in the execution of the workflow and enable reproducibility of the method and adaptability to datasets from other domains. The user is also able to tailor the lab to bring in a different changepoint algorithm or indeed combine the analysis with other approaches that critique model performance. E.g. the lab could be updated to also evaluate the climate models using extreme value theory, as has been done previously by Leeson et al. (2018). This can serve as a key tool in facilitating the use of data science methods to tackle some of environmental sciences grand challenges (Blair et al., 2019).

5. Conclusions

A new approach to numerical model evaluation has been developed by utilising a combination of changepoint analysis (using the PELT algorithm developed by Killick et al. (2012)) and fuzzy logic to assess the ability of climate models to capture key events seen in the observed record. Uncertainty in the changepoint locations are used in combination with a fuzzy union based metric to assign individual similarity scores to each changepoint in the observations time series to measure how well the numerical model captures that particular changepoint. This allows focus of the model evaluation to be placed on local scale temporal events and quantify whether strong performance using global integrated quantities translates to the local scale. In addition, the method can be used to identify common events that indicate good or poor performance highlighting potential areas to focus further model development on. This was demonstrated through a case study using a regional climate model which was able to pick up observed changepoints in temperature records over Greenland to varying degrees of success.

In order to facilitate access to data science and statistical approaches for environmental scientists, the method has also been incorporated into the DataLabs framework. This allows users to interact in a collaborative way utilising the method standalone, porting it to other datasets or combining it with other approaches (e.g. extreme value theory (Leeson et al., 2018; Toulemonde et al., 2015)) for a more robust model evaluation exercise. This helps provide a collaborative platform to tackle environmental data sciences' grand challenges (Blair et al., 2019).

Code and data availability

Software Name: Fuzzy changepoint application to evaluate numerical model ability to capture important shifts in environmental time series. Hardware Requirements: PC, System requirements: Windows, Linux, Program language: R, Program size: 60 KB, Licence: OGL v3, Available at the NERC Environmental Information Data Centre (EIDC): https://doi.org/10.5285/49d04d55-90a7-4106-b8fe-2e75aba228e4 (Hollaway, 2021). The R code to run the Fuzzy Changepoint based analysis case study presented in this paper is available as either a Jupyter or R Markdown notebook and is available on GitHub: https://github.com/mjhollaway/Fuzzy_cpt_eval. The accompanying R Shiny application is available at the following URL: https://dsne-fuzzycpteval.datalabs.ceh.ac.uk/. The Gc-Net weather station data are publically available for download from http://cires1.colorado.edu/steffen/gcnet/. The ERA5 reanalysis dataset is available for download from the Copernicus Climate Change Service (C3S) Climate Data Store at https://cds.climate.copernicus.eu/#!/search?text=ERA5&type=dataset.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was funded by the UK Engineering and Physical Sciences Research Council (EPSRC) Data Science for the Natural Environment (DSNE) project (Grant No. EP/R01860X/1). This work was also supported by the UK Status, Change and Projections of the Environment project, a National Capability award funded by the UK Natural Environmental Research Council (NERC: NE/R016429/1). Additional funding for the DataLabs came through a NERC Capital bid as part of the Environmental Data Services. The authors would like to thank Iain Walmsley and Oladimeji Awe for assistance in deploying the Shiny Application in DataLabs.

References

Allaire, J.J., Horner, J., Xie, Y., Marti, V., Porte, N., 2018. Markdown: 'Markdown' rendering for R. https://CRAN.R-project.org/package=markdown. (Accessed 17 April 2020).

Beaulieu, C., Chen, J., Sarmiento, J.L., 2012. Change-point analysis as a tool to detect abrupt climate variations. Phil. Trans. Math. Phys. Eng. Sci. 370 (1962), 1228–1249. https://doi.org/10.1098/rsta.2011.0383.

Beaulieu, C., Killick, R., 2018. Distinguishing trends and shifts from memory in climate data. J. Clim. 31 (23), 9519–9543. https://doi.org/10.1175/jcli-d-17-0863.1.

Beven, K., 2006. A manifesto for the equifinality thesis. J. Hydrol. 320 (1), 18–36. https://doi.org/10.1016/j.jhydrol.2005.07.007.

Blair, G.S., Henrys, P., Leeson, A., Watkins, J., Eastoe, E., Jarvis, S., Young, P.J., 2019. Data science of the natural environment: a Research roadmap. Frontiers in Environmental Science 7 (121). https://doi.org/10.3389/fenvs.2019.00121.

Chang, W., Cheng, J., Allaire, J.J., Xie, Y., McPherson, J., 2018. Shiny: web application framework for R. https://CRAN.R-project.org/package=shiny. (Accessed 17 April 2020).

Collins, W.J., Bellouin, N., Doutriaux-Boucher, M., Gedney, N., Halloran, P., Hinton, T., Hughes, J., Jones, C.D., Joshi, M., Liddicoat, S., Martin, G., O'Connor, F., Rae, J., Senior, C., Sitch, S., Totterdell, I., Wiltshire, A., Woodward, S., 2011. Development and evaluation of an Earth-System model –. HadGEM2. Geosci. Model Dev. 4 (4), 1051–1075. https://doi.org/10.5194/gmd-4-1051-2011.

Dee, D.P., Uppala, S.M., Simmons, A.J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M.A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A.C.M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A.J., Haimberger, L., Healy, S.B., Hersbach, H., Hólm, E.V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A.P., Monge-Sanz, B.M., Morcrette, J.-J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N., Vitart, F., 2011. The ERA-Interim reanalysis: configuration and performance of the data assimilation system. Q. J. R. Meteorol. Soc. 137 (656), 553–597. https://doi.org/10.1002/qj.828.

Eckley, I.A., Fearnhead, P., Killick, R., 2011. Analysis of changepoint models. In: Cemgil, A.T., Barber, D., Chiappa, S. (Eds.), Bayesian Time Series Models. Cambridge University Press, Cambridge, pp. 205–224.

Fawcett, L., 2018. Using interactive shiny applications to facilitate research-informed learning and teaching. J. Stat. Educ. 26 (1), 2–16. https://doi.org/10.1080/10691898.2018.1436999.

Fettweis, X., Box, J.E., Agosta, C., Amory, C., Kittel, C., Lang, C., van As, D., Machguth, H., Gallée, H., 2017. Reconstructions of the 1900–2015 Greenland ice sheet surface mass balance using the regional climate MAR model. Cryosphere 11 (2), 1015–1033. https://doi.org/10.5194/tc-11-1015-2017.

Fettweis, X., Franco, B., Tedesco, M., van Angelen, J.H., Lenaerts, J.T.M., van den Broeke, M.R., Gallée, H., 2013. Estimating the Greenland ice sheet surface mass balance contribution to future sea level rise using the regional atmospheric climate model MAR. The Cryosphere 7 (2), 469–489. https://doi.org/10.5194/tc-7-469-2013.

Gleckler, P.J., Taylor, K.E., Doutriaux, C., 2008. Performance metrics for climate models. J. Geophys. Res.: Atmosphere 113 (D6). https://doi.org/10.1029/2007jd008972.

Gutjahr, O., Putrasahan, D., Lohmann, K., Jungclaus, J.H., von Storch, J.S., Brüggemann, N., Haak, H., Stössel, A., 2019. Max Planck institute earth system model (MPI-ESM1.2) for the high-resolution model intercomparison project (HighResMIP). Geosci. Model Dev. (GMD) 12 (7), 3241–3281. https://doi.org/10.5194/gmd-12-3241-2019.

Hanna, E., Fettweis, X., Mernild, S.H., Cappelen, J., Ribergaard, M.H., Shuman, C.A., Steffen, K., Wood, L., Mote, T.L., 2014. Atmospheric and oceanic climate forcing of the exceptional Greenland ice sheet surface melt in summer 2012. Int. J. Climatol. 34 (4), 1022–1037. https://doi.org/10.1002/joc.3743.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R.J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., Thépaut, J.-N., 2020. The ERA5 global reanalysis. Q. J. R. Meteorol. Soc. 146 (730), 1999–2049. https://doi.org/10.1002/qj.3803.

Hollaway, M.J., Beven, K.J., Benskin, C.M.W.H., Collins, A.L., Evans, R., Falloon, P.D., Forber, K.J., Hiscock, K.M., Kahana, R., Macleod, C.J.A., Ockenden, M.C., Villamizar, M.L., Wearing, C., Withers, P.J.A., Zhou, J.G., Barber, N.J., Haygarth, P. M., 2018. A method for uncertainty constraint of catchment discharge and phosphorus load estimates. Hydrol. Process. 32 (17), 2779–2787. https://doi.org/10.1002/hyp.13217.

Hollaway, M.J., Dean, G., Blair, G.S., Brown, M., Henrys, P.A., Watkins, J., 2020. Tackling the Challenges of 21st-Century Open Science and beyond. A Data Science Lab Approach, Patterns. https://doi.org/10.1016/j.patter.2020.100103.

Hollaway, M.J., 2021. Fuzzy changepoint application to evaluate numerical model ability to capture important shifts in environmental time series. NERC Environ. NERC Environ. Inf. Data Centre. (Model). https://doi.org/10.5285/49d04d55-90a7-4106-b8fe-2e75aba228e4.

Hu, L., Keller, C.A., Long, M.S., Sherwen, T., Auer, B., Da Silva, A., Nielsen, J.E., Pawson, S., Thompson, M.A., Trayanov, A.L., Travis, K.R., Grange, S.K., Evans, M.J., Jacob, D.J., 2018. Global simulation of tropospheric chemistry at 12.5 km resolution: performance and evaluation of the GEOS-Chem chemical module (v10-1) within the NASA GEOS Earth system model (GEOS-5 ESM). Geosci. Model Dev. (GMD) 11 (11), 4603–4620. https://doi.org/10.5194/gmd-11-4603-2018.

IPCC, 2018. Climate Change 2013: the Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Killick, R., Eckley, I.A., 2014. changepoint: an R Package for Changepoint Analysis. J. Stat. Software 58 (3), 19. https://doi.org/10.18637/jss.v058.i03.

Killick, R., Fearnhead, P., Eckley, I.A., 2012. Optimal detection of changepoints with a linear computational cost. J. Am. Stat. Assoc. 107 (500), 1590–1598. https://doi.org/10.1080/01621459.2012.737745.

Killick, R., Haynes, K., Eckley, I.A., 2019. changepoint: an R Package for changepoint analysis. , R package version 2.3.1. https://CRAN.R-project.org/package=changepoint.

Krause, P., Boyle, D.P., Bäse, F., 2005. Comparison of different efficiency criteria for hydrological model assessment. Adv. Geosci. 5, 89–97. https://doi.org/10.5194/adgeo-5-89-2005.

Leeson, A.A., Eastoe, E., Fettweis, X., 2018. Extreme temperature events on Greenland in observations and the MAR regional climate model. Cryosphere 12 (3), 1091–1102. https://doi.org/10.5194/tc-12-1091-2018.

Matthé, T., De Caluwe, R., De Tré, G., Hallez, A., Verstraete, J., Leman, M., Cornelis, O., Moelants, D., Gansemans, J., 2006. Similarity between Multi-Valued Thesaurus Attributes: Theory and Application in Multimedia Systems. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 331–342.

Medley, B., Joughin, I., Das, S.B., Steig, E.J., Conway, H., Gogineni, S., Criscitiello, A.S., McConnell, J.R., Smith, B.E., van den Broeke, M.R., Lenaerts, J.T.M., Bromwich, D. H., Nicolas, J.P., 2013. Airborne-radar and ice-core observations of annual snow accumulation over Thwaites Glacier, West Antarctica confirm the spatiotemporal variability of global and regional atmospheric models. Geophys. Res. Lett. 40 (14), 3649–3654. https://doi.org/10.1002/grl.50706.

Meyer, D., Hornik, K., 2009. Generalized and Customizable Sets in R. 2009, vol. 31, p. 27. https://doi.org/10.18637/jss.v031.i02, 2.

Nghiem, S.V., Hall, D.K., Mote, T.L., Tedesco, M., Albert, M.R., Keegan, K., Shuman, C.A., DiGirolamo, N.E., Neumann, G., 2012. The extreme melt across the Greenland ice sheet in 2012. Geophys. Res. Lett. 39 (20) https://doi.org/10.1029/2012gl053611.

Pincus, R., Batstone, C.P., Hofmann, R.J.P., Taylor, K.E., Glecker, P.J., 2008. Evaluating the present-day simulation of clouds, precipitation, and radiation in climate models. J. Geophys. Res.: Atmosphere 113 (D14). https://doi.org/10.1029/2007jd009334.

R Core Team, 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Coputing, Vienna, Austria. https://www.R-project.org/.

Savage, N.H., Agnew, P., Davis, L.S., Ordóñez, C., Thorpe, R., Johnson, C.E., O'Connor, F. M., Dalvi, M., 2013. Air quality modelling using the Met Office Unified Model (AQUM OS24-26): model description and initial evaluation. Geosci. Model Dev. (GMD) 6 (2), 353–372. https://doi.org/10.5194/gmd-6-353-2013.

Schröter, D., Cramer, W., Leemans, R., Prentice, I.C., Araújo, M.B., Arnell, N.W., Bondeau, A., Bugmann, H., Carter, T.R., Gracia, C.A., de la Vega-Leinert, A.C., Erhard, M., Ewert, F., Glendining, M., House, J.I., Kankaanpää, S., Klein, R.J.T., Lavorel, S., Lindner, M., Metzger, M.J., Meyer, J., Mitchell, T.D., Reginster, I., Rounsevell, M., Sabaté, S., Sitch, S., Smith, B., Smith, J., Smith, P., Sykes, M.T., Thonicke, K., Thuiller, W., Tuck, G., Zaehle, S., Zierl, B., 2005. Ecosystem Service supply and vulnerability to global change in Europe. Science 310 (5752), 1333. https://doi.org/10.1126/science.1115233.

Slater, L.J., Thirel, G., Harrigan, S., Delaigue, O., Hurley, A., Khouakhi, A., Prosdocimi, I., Vitolo, C., Smith, K., 2019. Using R in hydrology: a review of recent developments and future directions. Hydrol. Earth Syst. Sci. 23 (7), 2939–2963. https://doi.org/10.5194/hess-23-2939-2019.

Steffen, K., Box, J.E., Abdalati, W., 1996. Greenland climate network: GC-Net. In: Colbeck, S.C., Crrel (Eds.), Special Report on Glaciers, Ice Sheets and Volcanoes, Trib. To M. Meier.

Swart, N.C., Cole, J.N.S., Kharin, V.V., Lazare, M., Scinocca, J.F., Gillett, N.P., Anstey, J., Arora, V., Christian, J.R., Hanna, S., Jiao, Y., Lee, W.G., Majaess, F., Saenko, O.A., Seiler, C., Seinen, C., Shao, A., Sigmond, M., Solheim, L., von Salzen, K., Yang, D., Winter, B., 2019. The Canadian earth system model version 5 (CanESM5.0.3). Geosci. Model Dev. (GMD) 12 (11), 4823–4873. https://doi.org/10.5194/gmd-12-4823-2019.

Toulemonde, G., Ribereau, P., Naveau, P., 2015. Applications of extreme value theory to environmental data analysis. In: Chavez, M., Ghil, M., Urrutia-Fucugauchi, J. (Eds.), Extreme Events: Observations, Modeling, and Economics. American Geophysical Union Washington DC, pp. 7–21.

Tso, C.-H.M., Henrys, P., Rennie, S., Watkins, J., 2020. State tagging for improved earth and environmental data quality assurance. Frontiers in Environmental Science 8 (46). https://doi.org/10.3389/fenvs.2020.00046.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., 't Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B., 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3 (1), 160018. https://doi.org/10.1038/sdata.2016.18.

Zadeh, L.A., 1965. Fuzzy sets. Inf. Control 8 (3), 338–353. https://doi.org/10.1016/S0019-9958(65)90241-X.

Zhang, N.R., Siegmund, D.O., 2007. A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. Biometrics 63 (1), 22–32. https://doi.org/10.1111/j.1541-0420.2006.00662.x.