Remote Sensing of Environment xxx (xxxx) xxxx

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

Improved change monitoring using an ensemble of time series algorithms

Eric L. Bullock*, Curtis E. Woodcock, Christopher E. Holden

Department of Earth and Environment, Boston University, 685 Commonwealth Avenue, Boston, MA 02215, USA

ARTICLE INFO

Keywords:
Change detection
Time series analysis
Landsat
Structural break detection
Land cover monitoring

ABSTRACT

An ensemble of time series algorithms improves land change monitoring. The methodology combines the Continuous Change Detection and Classification (CCDC; Zhu & Woodcock, 2014) and Cumulative Sum of Residuals (CUSUM) algorithms for break detection and the Chow Test (Chow, 1960) for removing false positives (or breaks in time series not representing land change). The algorithms included are based on fundamentally different approaches to change detection and therefore offer unique advantages. The ensemble, or the combination of the three algorithms, was applied to 3 Landsat scenes in the United States and the results were assessed based on their ability to correctly discern structural breaks from stable time periods. The CUSUM test was shown to detect significant breaks 84.18% of the time and the Chow Test correctly removed breaks in 87.4% of the breaks analyzed. The ensemble produced results with lower frequency of errors of omission and commission (Type-I and Type-II errors) than a single algorithm approach. These results indicate that using a combination of break detection algorithms can be an improvement over typical approaches that utilize only one algorithm.

1. Introduction

1.1. Remote sensing for land cover change monitoring

There is a need for spatially explicit information on land cover use and condition (Loveland et al., 1999; Rindfuss et al., 2004). Remote sensing data offers the most practical and accurate means of obtaining such information over large areas. The robust data archives of optical sensors like Landsat and MODIS have led to approaches that attempt to monitor surface properties continuously through time (Kennedy et al., 2010; Verbesselt et al., 2010; Brooks et al., 2012; Zhu and Woodcock, 2014b). These approaches have developed in response to the limitations of traditional change detection techniques that rely on comparing two images to find change (Radke et al., 2005). Many of these new methodologies have adapted classical approaches to time series analysis developed in other fields for the purpose of land change monitoring.

Time series analysis has several advantages relative to classical image processing techniques based on small numbers of images (two usually). Dense data stacks allow for an increase in input features for classification; an example being the use of indicators of seasonal variation for land cover classification (Pasquarella et al., 2017). Repeated observations also allow for the filling of data gaps due to clouds, atmospheric contaminants, or as in the case of Landsat 7, missing data due to sensor failures (Roy et al., 2008; Zhu et al., 2015b). Frequent data also allow for near-real time monitoring, with the analysis being updated with new image acquisitions (Hermosilla et al., 2017).

Time series analysis has also proven beneficial for land cover change detection (Verbesselt et al., 2010; Brooks et al., 2012; Zhu and Woodcock, 2014a; Zhu, 2017). Change monitoring methodologies have developed through years of research on structural break detection drawing on approaches developed in the fields of econometrics, biology, meteorology, and systems management (Klein, 1997). These methodologies have focused on monitoring for changes in one or multiple variables through time. The approaches are often based on regression analysis, with gradual trends being identified from the coefficients for slope or model residuals. Remote sensing applications for trend analysis include investigating gradual climatic variations (Ju and Masek, 2016; Sulla-Menashe et al., 2016), physical alterations to the land cover such as forest regrowth (Kennedy et al., 2010; DeVries et al., 2015a), changing species composition (Bullock et al., 2017), or identifying ecological responses to agricultural or urban expansion (Li et al., 2012).

In addition to trend analysis, abrupt changes can be discerned from a time series through temporal segmentation or structural break detection. These approaches often utilize a significance test derived from a change point identification algorithm. One way of categorizing break detection algorithms is whether they operate on the entirety of a given time series or successively monitor for changes as new data become available. The former is referred to as an *offline* method and the latter as *online*.

* Corresponding author.
E-mail addresses: bullocke@bu.edu (E.L. Bullock), curtis@bu.edu (C.E. Woodcock), ceholden@bu.edu (C.E. Holden).

https://doi.org/10.1016/j.rse.2019.04.018
Received 23 March 2018; Received in revised form 7 April 2019; Accepted 14 April 2019

An example of an offline algorithm for land cover monitoring that uses a time series of Landsat data is LandTrendr, which segments a time series temporally and then merges and partitions the segments to correspond to different stages of land use and condition (Kennedy et al., 2010). LandTrendr is considered an offline method because the entire time series is initially used to create the time segments and a significance test is used to determine appropriate break locations. LandTrendr has been used successfully across a wide range of forested ecosystems for change monitoring (Pflugmacher et al., 2013; Fragal et al., 2016; Shimizu et al., 2017).

Numerous approaches have also been proposed for online monitoring of change. For example, the moving sum of residuals (MOSUM) method can be used for online change monitoring by using a statistical test calculated from the moving sum of residuals from a regression model within a moving window (Zeileis et al., 2010, 2005). MOSUM has been implemented in the 'BFAST' R package for land cover monitoring, along with several other online and offline break detection tests (Verbesselt et al., 2012a, 2012b). In a different approach, Brooks et al. (2014) used an exponentially-weighted moving average (EWMA) control chart approach to find breaks in a time series. EWMA tracks the moving average of a sample observation and all previous observations with greater weight being given to closer time periods (Roberts, 1959). Since the process is iteratively performed on new data that are added to the model it is considered to be an online method.

The Continuous Change Detection and Classification (CCDC) algorithm is another online method for land cover monitoring (Zhu and Woodcock, 2014a). CCDC operates by fitting harmonic regression models to every pixel over time using all available Landsat data. New observations are compared to predicted observations based on the current model, and if the observed data deviate beyond a set threshold for all observations within a moving window period then a break is detected. The model fits are then used as inputs to a Random Forests classifier for land cover classification. CCDC is being implemented for the USGS Land Change Monitoring, Assessment, and Projection (LCMAP) initiative. LCMAP aims to produce temporally continuous and spatially explicit land cover information based on historical Landsat data (Young, 2017).

The effectiveness of an algorithm for monitoring land change depends on many factors, including: the data quality and quantity, the robustness of the algorithm to image noise or spurious changes, the metrics used as the basis for the change detection, the effectiveness of the change metric across various forms of land cover and condition change, and user-defined parameters to control the algorithm. Therefore, errors will be algorithm-specific and the land change products that they produce can be considerably different. Cohen et al. (2017) found that when comparing 7 forest disturbance algorithms, a majority of the algorithms agreed on the year and location of a disturbance in only 3% of the total mapped disturbances, with only 20.8% being agreed upon by two or more algorithms. While these discrepancies can partially be attributed to differences in input datasets, they are also due to varying approaches to change detection that lead to frequent disagreement between the algorithms.

Variation in data frequency and change metrics used by individual algorithms will contribute to differences in change monitoring results. For example, approaches that rely on a single observation a year may omit sub-annual changes or incorrectly label the date of the change due to the event occurring later in the year than the observation (Tyukavina et al., 2017). The choice of spectral bands and/or transformations will also influence the types of changes the algorithm will be able to accurately detect (Schultz et al., 2016).

The underlying change detection algorithm will also control the sensitivity of the approach to certain types of changes. For example, algorithms that characterize the trajectory or long-term trend in the time series are more suitable for the detection of gradual changes than those that do not. The timing of the event within the monitoring period can also influence the effectiveness of the algorithm. For example, CUSUM, when used offline for break detection, performs poorly at the beginning and end of the time series (Robbins et al., 2011). Methods can also vary in their sensitivity to outliers, implying that certain methodologies will be more suitable for environments with "noisy" observations such as missed clouds (Fearnhead and Rigaill, 2018). These differences imply that no one algorithm may be ideal for all change detection applications.

1.2. Ensemble algorithms

An alternative to choosing a single algorithm for land cover change analysis is utilizing multi-algorithm ensembles. The assumption motivating ensemble algorithms is that since all algorithms approach the problem differently, and can work on unique subsets or transformations of the input data, there is no single algorithm that will be optimal in all situations (Wolpert and Macready, 1997). The objective of an ensemble approach is to increase overall performance by benefiting from the unique advantages of each individual algorithm. In principle, algorithms can be combined to produce more accurate results than what would be produced from any one individually (Clemen, 1989; Dietterich, 2002).

Ensemble methods are particularly useful in problems that involve large quantities of data. In such a case a single algorithm may often not be able to properly characterize the variance or complexity of the data, and therefore will be prone to misclassification and biases. Using multiple algorithms that have unique approaches to separating the boundaries between independent groups of data can therefore be advantageous when classifying or segmenting diverse datasets (Polikar, 2009). Time series analysis of remote sensing data can utilize hundreds of images, each with millions of pixels representing a diverse combination of land covers and atmospheric effects. Additionally, different kinds of change events will have differing impacts on the data. Therefore, a single algorithm may not be the most effective way of characterizing the data in all possible situations.

One type of ensemble algorithm that has been used in the past for land cover classification utilizes Multiple Classifier Systems (MCS; Woźniak et al., 2014). In MCS all data are classified independently using different classification algorithms or derivatives of the input data and the results are combined to assign a single label. MCS for land cover applications have used different approaches to deciding among the classification results, including voting (Foody et al., 2007), rule-based decisions (Saxena et al., 2017), and stacking (Healey et al., 2018). Additionally, different approaches to choosing among classifiers have been directly compared (Kittler et al., 1998; Giacinto et al., 2000; Briem et al., 2002; Steele and Patterson, 2002; Bruzzone et al., 2004). MCS have the advantage that they can be efficiently implemented in parallel computational environments and can overcome biases of individual classifiers to achieve higher overall accuracies (Dietterich, 2002; Woźniak et al., 2014).

An alternative type of ensemble algorithm uses multiple data analysis techniques sequentially, rather than in parallel, to calculate a single output. Fig. 1 shows the difference in how a sequential ensemble is structured as compared to a parallel ensemble such as an MCS. A sequential approach differs from a parallel ensemble in that each algorithm influences the inputs used by the next algorithm. Therefore, instead of amalgamating or choosing among the results of the different algorithms, as is the case in an MCS, the calculation from a sequential ensemble approach continues to evolve while going through the series of algorithms (Dietterich, 2002). Sequential algorithm configurations are common in the field machine learning, but the general approach can be applied to any model-based algorithm (Rokach and Maimon, 2005). Sequential ensembles can potentially benefit from each of the algorithms involved without having to choose one as the final class label.

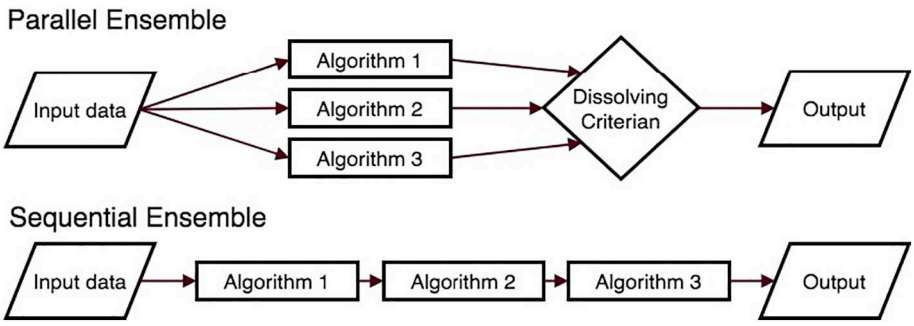

Fig. 1. Flow chart illustrating the difference between a parallel (top) and sequential (bottom) ensemble algorithm. In a parallel ensemble each algorithm performs a task using the input data independently and some method is used to choose a final answer. In the sequential ensemble the algorithms operate in order, so the input for second or third algorithms is altered by the ones before it.

1.3. Research objectives

The motivating research question is: *Given that algorithms are different and will produce varying results, will running multiple algorithms sequentially add benefit to the overall goal of time series break detection*? To explore this question we developed a sequential ensemble algorithm that combines previously developed methodologies for structural break detection and land cover monitoring. The methods considered are based on fundamentally different approaches to change detection and will therefore offer unique advantages. In contrast to previous ensemble methodologies for the purpose of land cover or change classification, we have developed ours for the sole purpose of detecting a break in a time series. Since the methodology is developed in the context of land cover monitoring the breaks should represent a change in the land surface due to a distinct change event, for example drought or logging. The proposed methodology contains two tests for changes (or breaks) in a time series followed by one test to remove unnecessary breaks. The results were analyzed based on the level of improvement the secondary algorithms provided to the overall break detection.

2. Methods

2.1. Study location and data

Three study scenes in the United States were chosen for analysis. The scenes were WRS-2 Path/Row 013/029, 033/029, and 035/032, corresponding to New England, South Dakota, and Colorado, respectively (Fig. 2). All Landsat surface reflectance data from 1984 to 2015 were utilized for the green (0.52–0.6 μm), red (0.63–0.69 μm), near-infrared (0.77–0.9 μm), and shortwave infrared (1.55–1.75 μm, 2.09–2.35 μm) bands. Clouds and cloud shadows were masked using FMask version 3.2 (Zhu et al., 2015a; Zhu and Woodcock, 2012). Processing was performed on the Boston University High Performance Shared Computer Cluster using the MATLAB and Python programming languages.

2.2. Sequential ensemble algorithm

The ensemble is composed of three algorithms for detecting structural breaks in a time series: two for finding breaks and one for removing false breaks (Type-I errors) (Fig. 3). The first algorithm, CCDC, is used to identify breaks in the time series and fit harmonic regression models for each spectral band and for every discrete time segment between model breaks. The data from the discrete time segments are then used as inputs to the second change detection algorithm, CUSUM, which is used to identify breaks that were missed by CCDC. All break points in the time series are then tested with the Chow Test to identify false, or unnecessary breaks.

2.3. First break test: CCDC

The CCDC algorithm was used first in the break detection ensemble. As the name implies, CCDC consists of two distinct components: change detection through the identification of breaks in a time series and land cover classification. Previously, the algorithm produced overall accuracies in detecting land cover change of 91% (Zhu and Woodcock, 2014a), 84% (Olofsson et al., 2016) and 87% (Zhu et al., 2016). For the ensemble algorithm only the change detection component is utilized. The accuracy of CCDC in detecting changes in a time series (as opposed to the accuracy of land cover classification) has not previously been assessed.

The break detection in CCDC is performed by fitting sinusoidal regression models to all spectral bands in a time series of Landsat observations during a training period at the start of a time series. The training model is tested for stability based on model slope and RMSE. If the model is determined to be unstable then the beginning of the training period shifts forward until a stable training period is found. The training model is used to predict the next *consecutive* observations, with *consecutive* being a parameter that in our case was set to 5. If the residuals of all 5 of the observations exceed a change threshold then a break is detected (Fig. 4). Requiring consecutive observation to exceed the change score minimizes errors of commission due to image noise.

The process repeats by moving forward in time until another break is detected or the end of the time series. In the case of training periods that need to be shifted forward due to model instability, the subsequent model is used to predict backwards in time to test for changes during the time period with no model fit. Once a break is detected a new regression model is created and the process is repeated. For a detailed explanation of the CCDC algorithm see Zhu and Woodcock (2014b).

CCDC has the advantage compared to traditional approaches to structural break detection that it was specifically designed for land cover monitoring applications based on remote sensing. For example, it includes explicit consideration for the noisy nature of remote sensing data by requiring multiple successive observations to deviate significantly from model predictions. There have also been modifications to the algorithm that attempt to reduce type-I errors, or detection of false breaks. For example, clouds and cloud shadows can cause detection of false breaks due to a sudden shift in reflectance. However, single observations that are abnormally bright in the green band or dark in the SWIR1 band are labeled as missed clouds and cloud shadows, respectively, and are removed from analysis (Zhu and Woodcock, 2014b).

Since CCDC uses an online moving-window approach there is, after initial model fitting and during periods of consistent data availability, a similar probability of breaks being detected throughout the entire time series. Furthermore, the online structure of CCDC allows for the continuous updating of change detection results in near-real time. By using a monitoring period to test for breaks based on regression residuals, CCDC is comparable to the version of MOSUM demonstrated in Verbesselt et al. (2012a, 2012b), and noticeably different than offline approaches that operate on the entirety of the data at once, such as LandTrendr or the offline CUSUM (Brown et al., 1975).

If the observations within the training window for CCDC are not representative of the entire time series then the training model will not be able to accurately predict future observations. Although CCDC tests

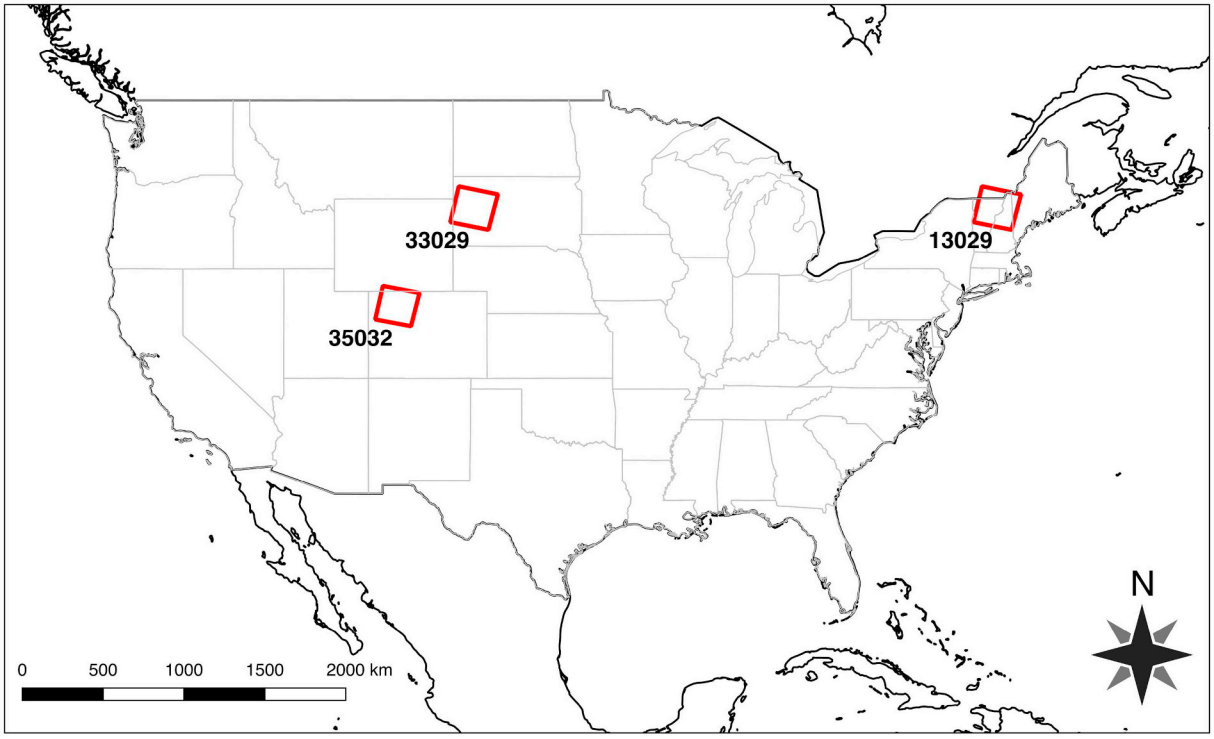

Fig. 2. Three Landsat scenes chosen as the study areas: Path/Row 35/32 (Colorado), 33/29 (South Dakota), and 13/29 (New England).

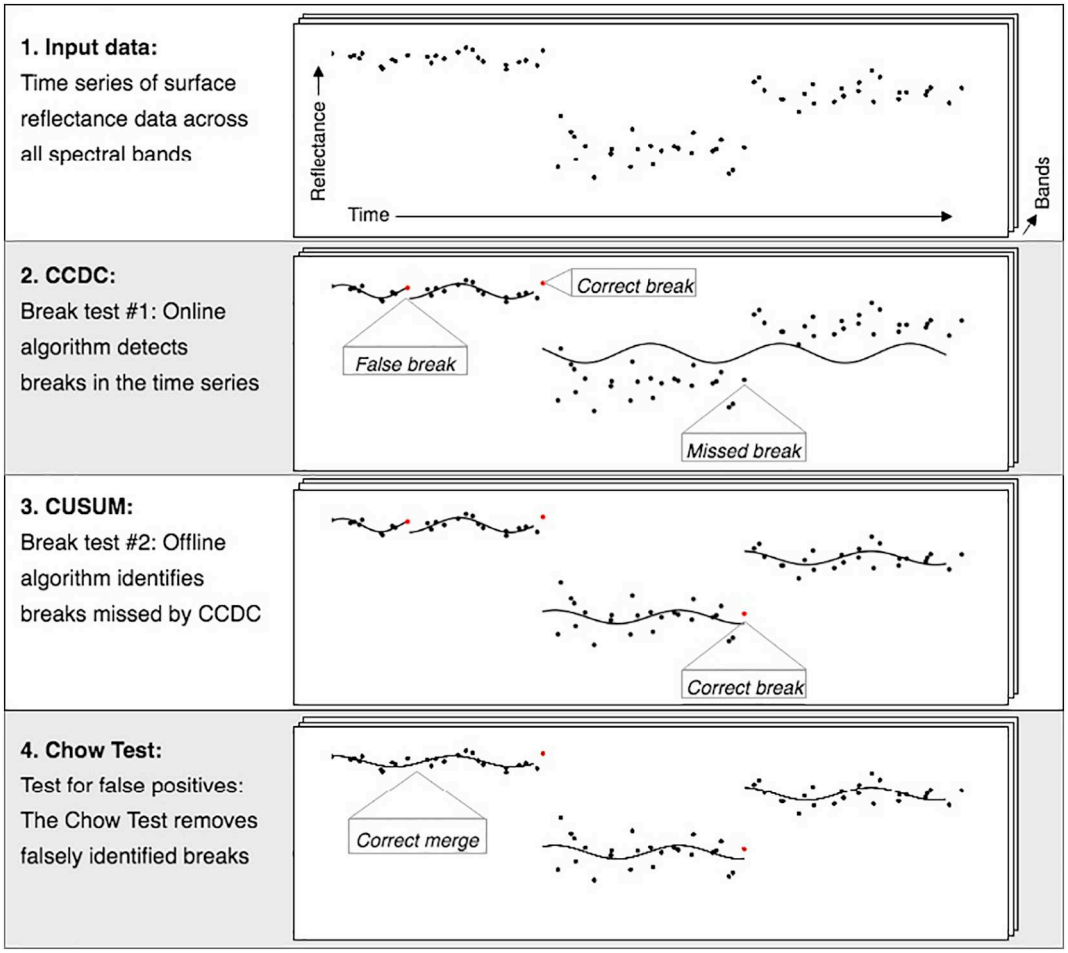

Fig. 3. The three components of the sequential ensemble algorithm. Breaks are first detected using CCDC and then CUSUM, and all break points are tested for false positives with the Chow Test.

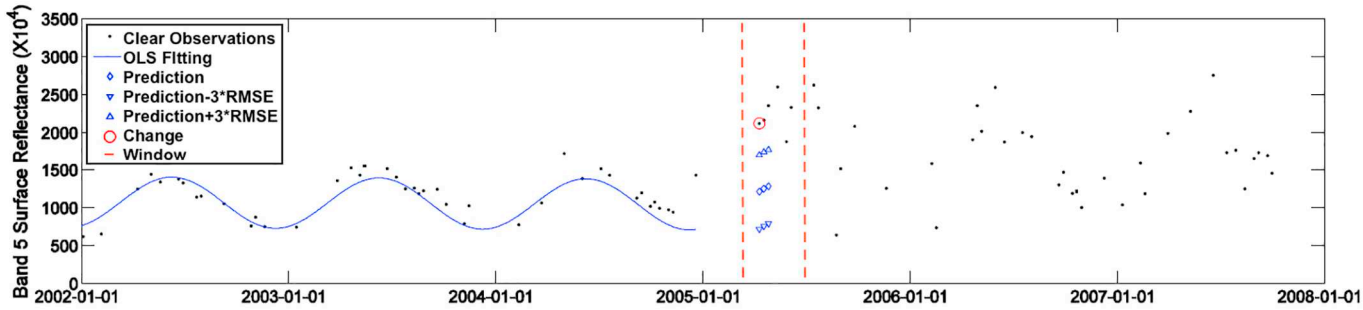

Fig. 4. An example SWIR1 time series using the CCDC break detection approach. A harmonic Ordinary Least-Squares (OLS) regression model is fit and used to predict the next sequential observation. If the predicted observation deviates from the observed for five consecutive observations then a break is detected. Modified from Zhu and Woodcock, 2014b.

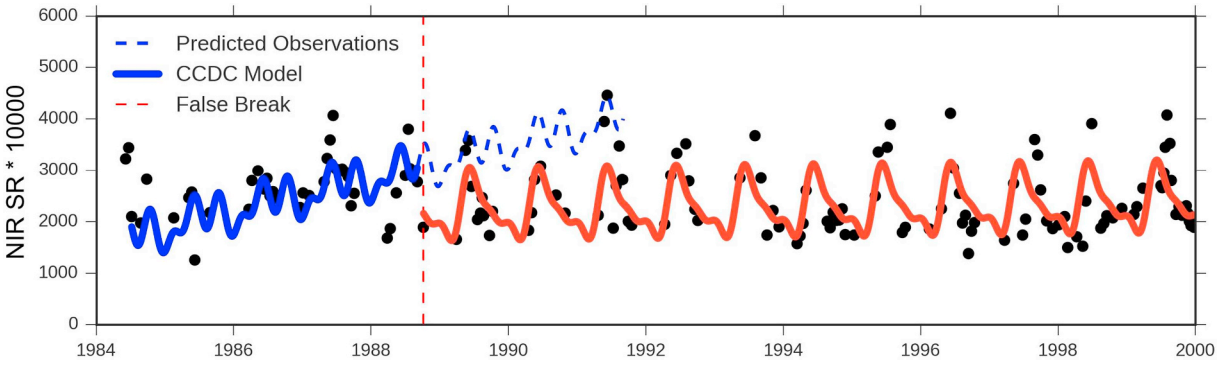

Fig. 5. Near-infrared time series in 033/029 (South Dakota) for an agricultural location. The CCDC model in the beginning of the time series contained an upward trend due to a disproportionate number of bright observations in the training period. Observations after the training period were well below the predicted, and therefore a break was detected. However, this was due to a bad model fit during the training period and not due to a real break in the time series.

for model stability in the training period it is still vulnerable to type-I errors (detection of false breaks) when the stability test fails. Fig. 5 demonstrates a situation in which CCDC will incorrectly detect a model break due to poor characterization of the data during the beginning of the time series.

2.4. The second break test: CUSUM

The second break test in the ensemble algorithm is performed after CCDC and is based on an offline CUSUM test for consistency in a linear regression model introduced by Brown et al. (1975), and extended for use with ordinary least squared (OLS) residuals in Ploberger and Kramer (1992). CUSUM is performed for each discrete time segment identified by the CCDC algorithm using the original surface reflectance data. CUSUM can be used for detecting instabilities in a model parameter, such as the mean of a time series. Since the mean of the residuals in an OLS model is required to be zero, then a stable model's residuals should fluctuate evenly and minimally around zero. If a structural break were present then the residuals would deviate consistently in one direction. Ploberger and Kramer (1992) demonstrated how the standardized OLS residuals are distributed as a Brownian Bridge, and the maximum of the absolute value of the standardized residuals can be used to test for model instability. The test statistic is used to determine the probability of observing the data assuming no structural instability. For our implementation, we used a threshold of 1% for determining a break in the time series (or a 99% probability of structural instability). Once a break is detected the time segment is split into two models (Fig. 6). The test is run in parallel across spectral bands with a break being detected when the change threshold is exceeded in a majority of the spectral bands (i.e., if CUSUM detects a change in three of the five spectral bands used for testing). Once a break is indicated, the correct break location must be determined.

There are various ways of determining the break location after CUSUM finds the existence of a break. One method is to test every possible break location to find the one that minimizes the combined sum of squared residuals (SSR) of the two models. While this approach is robust it is also computationally expensive, as every observation must be tested as a possible break point (Zeileis et al., 2003). For that reason we use a scalar-minimizing optimization routine, in which break locations are iteratively sampled until the minimum combined SSR is found using Brent's algorithm. Brent's algorithm is an efficient root-finding procedure for function minimization (Brent, 1971). The optimization approach eliminates the need to test every observation by efficiently narrowing in on an optimal break location. Brent's algorithm first attempts to determine the global minimum of the combined SSR using the fast, but not always reliable, secant and inverse quadratic interpolation methods. If convergence fails it will use the less efficient, but more robust, golden-section method. Therefore, Brent's algorithm is an attempt to be robust against converging on weak local minima while operationally efficient. Once a break location has been determined, new regression models are fit to the time segments.

The OLS residual-based CUSUM test is related to the moving sum of recursive residuals (MOSUM) test for structural break detection, with the difference being that in MOSUM the sum of residuals is calculated for a moving window of fixed width (Bauer and Hackl, 1978; Zeileis et al., 2005). MOSUM was introduced to attempt to overcome the limitation of CUSUM in detecting breaks early and late in a time series (Hinkley, 1971; Robbins et al., 2011). This limitation is due to the diminishing influence of each observation on the calculated sum of residuals. Therefore, breaks at the end of the time series will be less likely to be detected than in the middle. MOSUM, however, has been shown to detect false breaks when used for online change monitoring when the window contains multiple observations of obstructed data, such as noise from an unmasked cloud (DeVries et al., 2015a, 2015b). Since

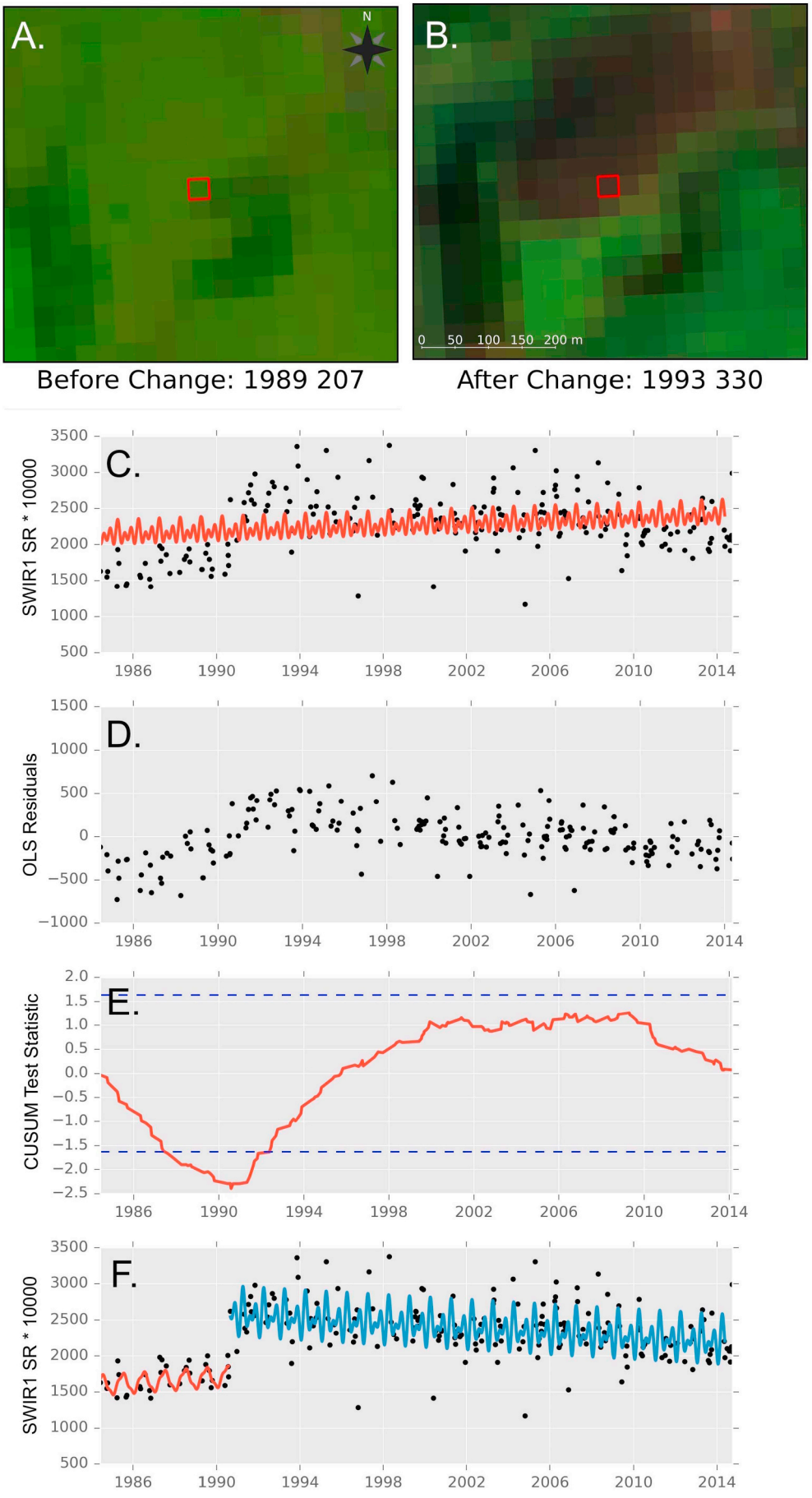

Fig. 6. A. A Landsat 5-4-3 composite before a land change. B. A Landsat 5-4-3 composite after a land cover change with a 30-m pixel highlighted in red. C. The time series of the pixel highlighted in B after the CCDC change detection step with a missed break and regression model shown in red. D. The residuals of an OLS regression model covering the time period shown in C. E. The cumulative sum of residuals with the critical region at a 5% confidence level shown in blue in which the residuals surpass in 1990. F. The time series models after CUSUM correctly identifies the change and the time series is separated into two discrete time segments separated by a model break in 1990. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

CCDC also utilizes a moving window in a monitoring period it is likely to be vulnerable to similar errors (Fig. 5). Therefore, CUSUM was chosen for the second change detection algorithm as opposed to MOSUM. It is worth noting that the combined use of the online CCDC algorithm and offline version of CUSUM is similar to 'BFASTmonitor' function in the R package 'BFAST', which by default uses MOSUM for change monitoring and a reverse-ordered CUSUM test for assessing the stability of the historical period (Verbesselt et al., 2012b, 2010).

2.5. Test for false positives: the Chow test

After the two tests to find breaks, all breaks are tested to see if they are false positives (breaks that shouldn't be included) using a modified version of the Chow Test for structural instability in a time series (Chow, 1960). The Chow Test compares the effectiveness of two separate adjacent models with one single model that spans the entire time period. The Chow Test is described as:

$$F(T) = \frac{(SSR_A - SSR_B))/k}{SSR_B(n - 2k)}$$

where,

$F(T)$ = Test statistic
SSR_A = Sum of squared residuals (SSR) of the restricted model
SSR_B = SSR of model 1 + SSR of model 2
k = # model coefficients
n = # observations in time span

The "restricted model" corresponds to the model using the pooled observations spanning the entire test period. The model is "restricted" in that the coefficients are assumed to be equal for the entirety of the time period. To test the null hypothesis that the restrictions on the model are true (and there should not be two separate groups, or in our case a model break), we calculate the Chow Test statistic. Accepting the null hypothesis therefore signifies the restrictions are valid and we merge the models into a single model spanning the entire time period.

The test statistic for the Chow Test is comparable across spectral bands, but due to correlation between bands there will likely be redundancy. Therefore, a weighted average of $F(T)$ is computed, with the weights corresponding to:

$$w_a = 1 - \frac{\sum_{i=1}^{n} r_{ai}}{n}$$

$$F(\hat{T}) = \frac{\sum_{a=1}^{n} w_a F(T)_a}{\sum_{a=1}^{n} w_a}$$

where,

w_a = Weighting coefficient for band a
n = Total # of bands
R_{ai} = Coefficient of determination (r^2) between time series of band a and i
$F(T)_a$ = Test statistic for band a
$F(\hat{T})$ = Mean test statistic across bands

Using this approach, the average is weighted by the band-to-band correlation to increase the importance of less-correlated spectral bands. If the average test statistic is below a critical value then the models are merged (Fig. 7). A weighted $F(T)$ was used after extensive testing of merging criteria for locations in each study scene that we had previously identified as errors of commission. The criteria tested included majority voting, using the maximum $F(T)$, and requiring $F(T)$ for all bands to be below the critical value. We also compared the results to the multivariate analysis of variance (MANOVA) using the same data, and found the weighted $F(T)$ to perform the best in identifying errors of commission. The distribution of weights across all the models in the 3 study scenes can be seen in Fig. 8.

The test for missed breaks is similar to the p-of-F test in the LandTrendr algorithm (Kennedy et al., 2010). In both approaches, an F-Statistic is used to determine if two models should be merged. In LandTrendr, the test is used to pick the best fit between alternative model trajectories for the entire study period. Our test does a similar procedure, except there are only two candidate trajectories: change and no change. The potential model fits and break location are instead determined by the two break tests. Additionally, LandTrendr performs the p-of-F test using yearly data, while our approach utilizes all available observations. While the LandTrendr approach reduces the computational intensity our approach allows for the possibility of detecting short duration and low-magnitude land changes by using all of the data.

Fig. 9 shows example of a time series during each stage of the ensemble algorithm. A land cover change in 1991 causes a break in the time series that is not correctly identified by CCDC. Instead, a false break is identified at the end of the time series. The two time segments identified by CCDC are then tested for missed breaks with CUSUM (Fig. 9B). The correct break is identified but the break at the end of the time series is still unnecessary. The Chow Test accurately retains the correct break that was identified by CUSUM but removes the incorrect one that was identified by CCDC (Fig. 9C). In the end there are two time segments separated by one break in the time series.

2.6. Assessment

Remote sensing studies that identify changes on a landscape often evaluate map accuracy through sample-based statistical inference (Stehman, 1999; Olofsson et al., 2013). When the area of a land cover or change process is the target population then sample units can be defined based on the pixels in a map. The sample unit will therefore represent a specific geographic location and reference labels can be assigned through some independent effort. The process of estimating map accuracy and area in such a way is well documented and a variety of approaches have been proposed, tested, and used in many studies (Stehman, 2014, 2013, 1997; Olofsson et al., 2014).

However, the objective of our assessment is not to estimate map accuracy or area, but instead to assess the performance of each algorithm in identifying breaks in a time series or, in the case of the Chow Test, identifying errors of commission. Therefore, we designed our assessment around the specific goal of evaluating the accuracy of break detection without regard to land cover classification or area estimation. The goals are to assess: (1) the accuracy of the breaks CCDC finds in a time series; (2) the accuracy of any additional breaks that CUSUM finds; and (3) the accuracy of the removal of breaks done using the Chow Test.

We defined the target population as the total changes (or breaks) in the three study scenes that represent a distinct change in land cover or condition. In order to account for errors of omission in the change detection we also included the total number of years without a break in the population. The sample units were defined as a year and pixel location, henceforth referred to as a *Pixel Year* (Fig. 10). The Pixel Years were stratified according to the effect that each algorithm had on the change detection for that year. Since the Pixel Years were selected randomly within each stratum, it was possible for the same pixel (geographic location) to have multiple sample units for different years (Table 1). The Pixel Years were stratified as follows:

1. *No Change*: There was no change identified by either CUSUM or CCDC during that year at that pixel location.
2. *CCDC Change*: A change was detected by CCDC and was not removed by the Chow Test.
3. *CUSUM Change*: A change was not detected by CCDC but was identified as change by CUSUM and not removed by the Chow Test.
4. *Chow Test Removal*: A change was detected by CCDC or CUSUM and was removed by the Chow Test.

The time series of each of the sample units were evaluated manually

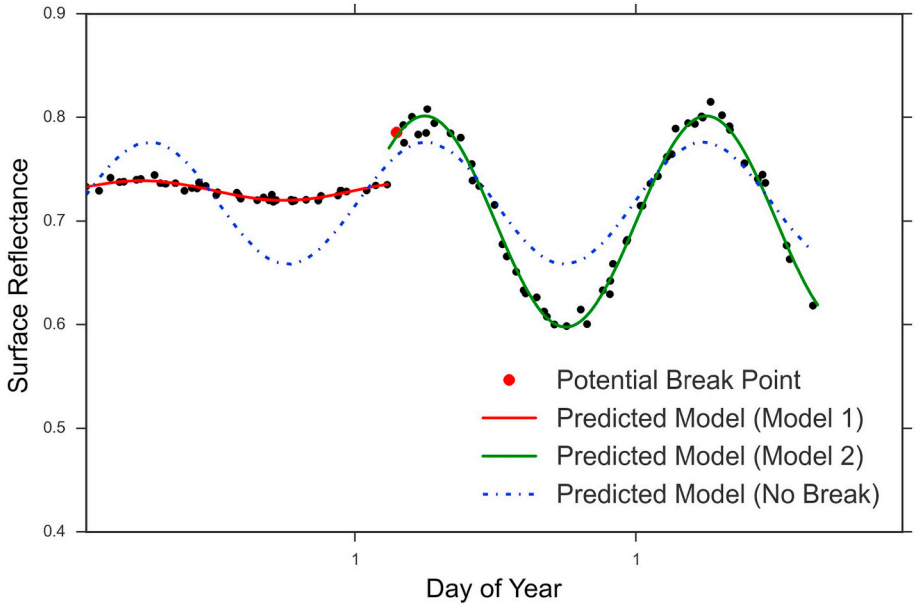

Fig. 7. The Chow Test for instability in a time series tests the null hypothesis that the coefficients from the two unrestricted models (Model 1 and Model 2) are the same as the coefficients for a single model covering the same time period. If the null hypothesis is not rejected the algorithm merges the two models, removes the breakpoint, and retains the full restricted model. In Fig. 6F, the break point is obvious and therefore the null hypothesis is rejected and the break would be retained.

to determine whether a change occurred during the Pixel Year. When possible, high-resolution imagery available on Google Earth and The National Agriculture Imagery Program (NAIP) were used to supplement the Landsat data. To account for changes that were correctly identified but with incorrect timing the response design was defined to record changes in the year after or before the pixel year.

At the time of the visual interpretation, the analyst was aware of the year being examined but did not know the stratification of the Pixel Year. The break location was analyzed for each of the spectral bands. If a change was apparent within the 3-year window of the Pixel Year then the reference label was indicated as a break. The change was required to be observable in the time series and related to a distinct change in land cover or condition. The interpretation of the breaks was performed using the AREA^2 toolbox (available from github.com/bullocke/area2) in addition to the TSTools plugin for QGIS (Holden, 2017).

Estimation of accuracy was performed in two ways: by estimating the accuracy of each algorithm in finding or removing changes relative to output of the previous algorithm in the ensemble, and by estimating the accuracy of the final results of the ensemble. For the former, the accuracies reflect different conditions for each algorithm (Table 2C). The Producer's Accuracy for CCDC indicates the proportion of changes that were accurately detected by CCDC. For CUSUM it measures the proportion of changes accurately identified by CUSUM that weren't found by CCDC, and for the Chow Test it indicates the proportion of changes incorrectly flagged by CCDC or CUSUM but correctly removed. The User's Accuracies reflect the proportion of breaks for each algorithm that were correctly added or removed.

The Pixel Years were post-stratified to simulate situations in which only CCDC was used and if CCDC and CUSUM were applied sequentially without the Chow Test (Table 2D). For each situation, a stratified

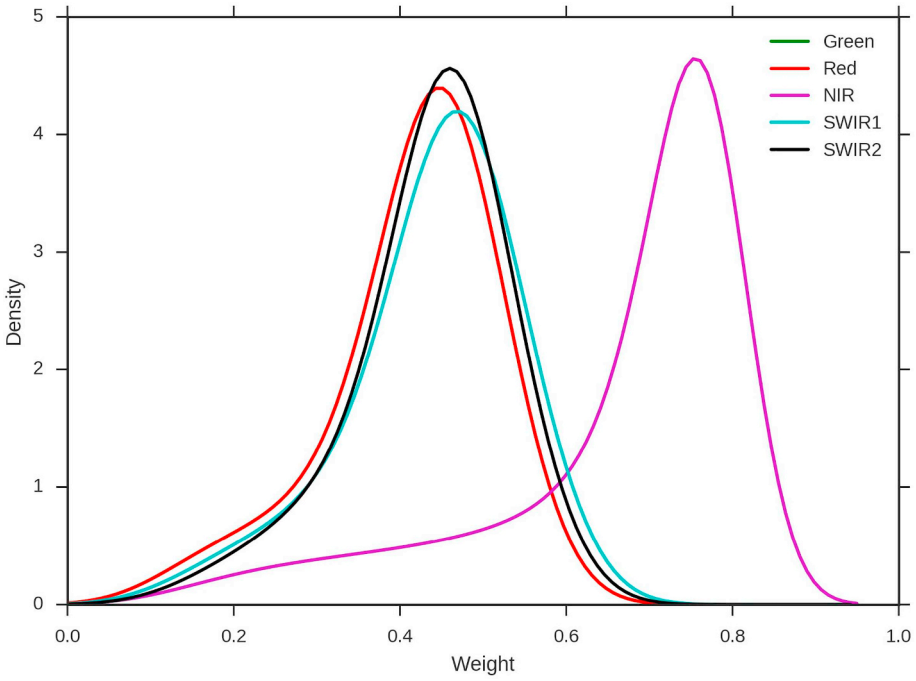

Fig. 8. Gaussian Kernel Density Estimates for the average of Chow test statistics for all models merged using The Chow Test across 3 study scenes.

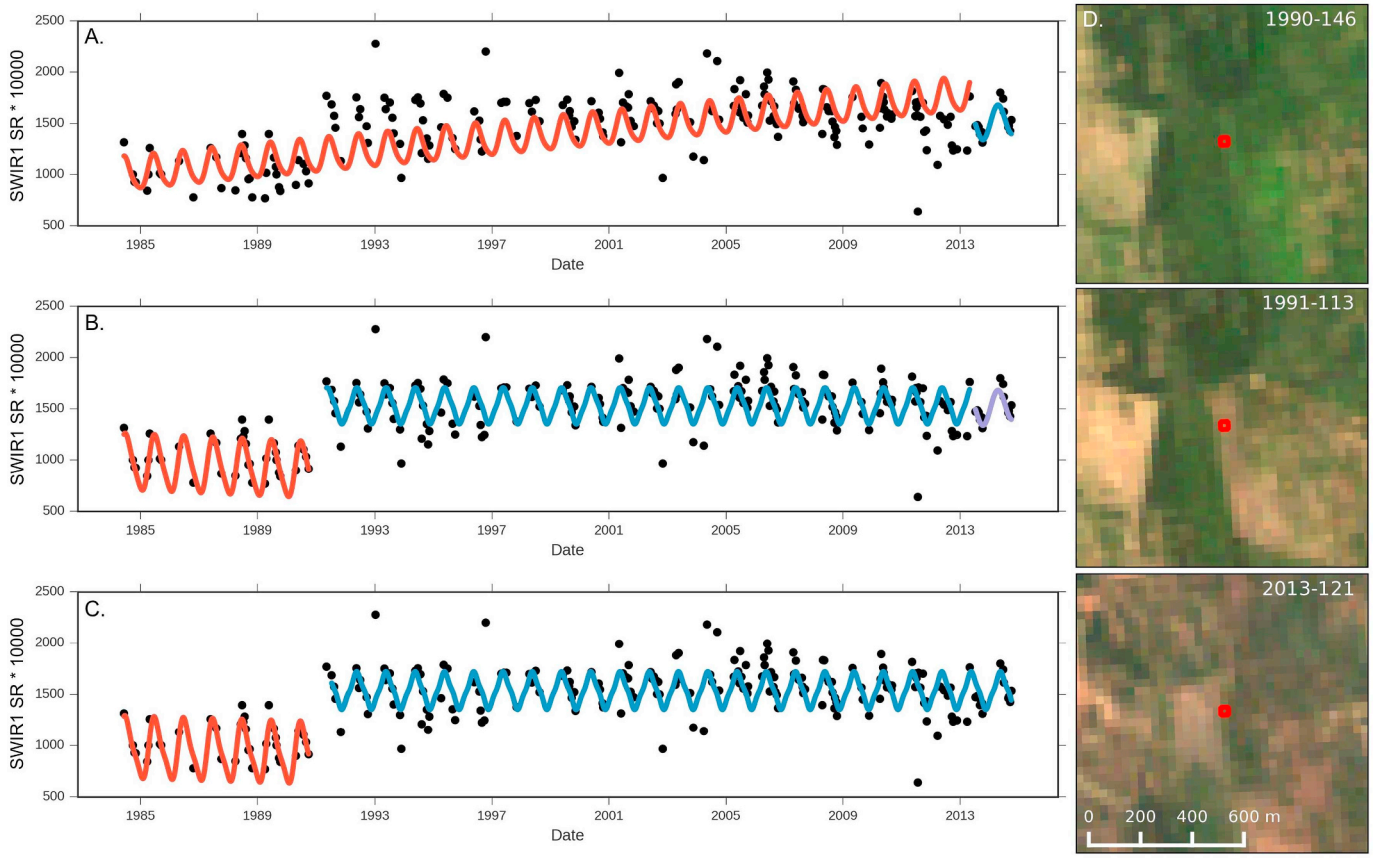

Fig. 9. A. Example SWIR1 time series after the CCDC stage with a missed break in 1991 and false break in 2013. B. The time series after the CUSUM stage with the first break correctly identified but still containing the false break in 2013. C. The time series after CUSUM and the Chow Test with the false break in 2013 correctly removed. D. Landsat 3-2-1 composites before (top) and after (middle) the correct break and at the end of the time series (bottom).

estimator was applied to the reference samples to calculate User's and Producer's Accuracies for the Change and No Change classes in addition to Overall Accuracy.

3. Results and discussion

CCDC correctly identified 71.0% of the changes in the study areas with a commission rate of 19.1% (User's Accuracy of 80.9%). Of the changes that were omitted by CCDC, only 10.1% were found by CUSUM. However, the User's Accuracy for CUSUM was 84.2%, indicating that the changes added were beneficial to the ensemble. Many of the abrupt changes due to land cover conversion were detected by CCDC, which would partially explain the high omission rate of CUSUM. In essence, CUSUM is being used to find changes that CCDC omitted, and so they are likely to be less obvious changes.

An example of CUSUM detecting a subtle and gradual change that was missed by CCDC can be seen in Fig. 11. A re-growing forest results in downward slope in the SWIR1 time series (D) that stabilizes around 2003. This change in the landscape, indicated by the break in the time series, is detected by CUSUM. CCDC, however, detected a nearby deforestation event that resulted in an abrupt and high-magnitude break

Table 1
Sample Units for the stratification of the Pixel Years. Proportions are shown for the three change strata when excluding the No Change stratum. Note that the CUSUM and Chow Test strata make up a combined 7% of the Change strata, and under 1% of the total Pixel Years.

Stratum	Samples	Proportion	Proportion (excluding no change)
1. No Change	345	0.97	N/A
2. CCDC Change	345	0.03	0.93
3. CUSUM Change	645	< 0.01	0.03
4. Chow Test Removal	645	< 0.01	0.04

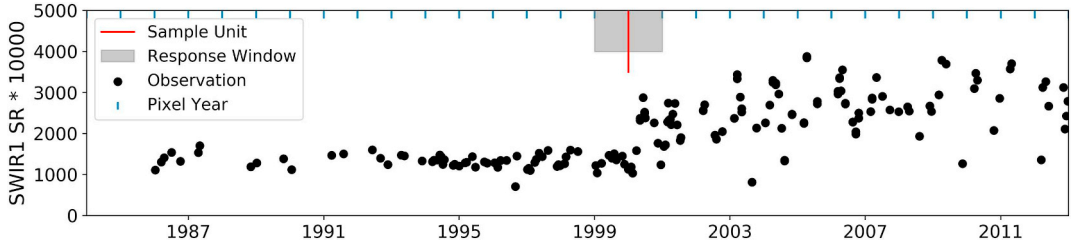

Fig. 10. An example of a SWIR1 time series showing potential sample units (Pixel Years) in addition to the three year response window. In this example, the Pixel Year corresponded to the year 2000, and the sample unit would be labeled as 'Change' if a break occurred from 1999 to 2001. In this case, there is a clear change in the time series and the sample unit would be labeled as 'Change'.

Table 2
Validation results for the second and third algorithms in the ensemble. The cells with a dotted red outline indicate the sample units that were correctly identified as change or no change by the algorithms.

	A. Reference (sample counts)		
	Change	**No change**	**Total**
No change	3	342	345
CCDC	279	66	345
CUSUM	543	102	645
Chow	81	564	645
	B. Reference (proportion)		**Total**
No change	0.9607	0.0084	0.9691
CCDC	0.0055	0.0233	0.0288
CUSUM	0.0008	0.0001	0.0009
Chow	0.0001	0.0001	0.0011
Total	0.9672	0.0328	1
	C. Algorithm specific accuracies		
	CCDC	**CUSUM**	**Chow**
User's acc. [%]	80.9	84.2	87.4
Prod. acc. [%]	71.0	10.1	13.8
	D. Result accuracies		
	CCDC Only	**CCDC + CUSUM**	**Ensemble**
User's change [%]	78.8	79.0	81.0
User's no change [%]	99.0	99.1	99.1
Prod. change [%]	71.0	74.3	74.0
Prod. no change [%]	99.0	99.0	99.4
Overall acc. [%]	98.2	98.4	98.6

in the time series (E).

Previous research comparing change detection algorithms for the purpose of monitoring forest disturbances has shown CCDC to be more susceptible to errors of omission than commission (Cohen et al., 2017). A possible explanation provided was that CCDC characterizes gradual changes as long-term trends in the regression models as opposed to attributing them to a change on a specific date. CCDC operates online with thresholds for the magnitude of change necessary to flag a change in addition to the number of consecutive observations that exceed the change threshold. Additionally, the prediction model is refit if a change is not found and, therefore, could result in a model that changes in slope as time progresses but does not cause a break in the model. Our results support this hypothesis, with CCDC having a higher User's Accuracy (80.7%) than Producer's (71.0%), i.e. more errors of omission than commission. We hypothesize that the offline version of CUSUM will be more sensitive to gradual changes as there is no refitting of the regression model and the change can be subtle as long as the residuals continuously shift in one direction.

CUSUM added 3.33% more breaks than were found with CCDC and the Chow Test removed 3.27% of the breaks found with the two break tests. The distribution of the timing of the breaks detected by CUSUM and CCDC were found to be different. While CCDC found changes at roughly the same rate for the entire time period, CUSUM was biased towards breaks in the middle of the time series (Fig. 12). These breaks were more likely to occur in models that were 10–20 years in length, while CCDC more often found breaks in models that were 1–10 years long (Fig. 13).

The breaks detected by CUSUM were found more often in the middle of the time series than the beginning or the end (Fig. 12). As previously mentioned, the inability to effectively detect breaks at the beginning and end of the time series is a commonly noted limitation of CUSUM (Robbins et al., 2011). This limitation would constrain the effectiveness of using CUSUM analysis for near-real time change monitoring. MOSUM has previously been utilized to compensate for this limitation (Verbesselt et al., 2012a). This same limitation, however, makes the CUSUM test robust against possibly false detections in short and noisy time series. In scenes 033/029 and 035/032, which are dominated by grasslands and agriculture, CCDC finds a large portion of the breaks in models that are under 3 years in duration (Fig. 13). These breaks can correspond to changes in crop management or soil moisture conditions, but are often caused from CCDC being unable to accurately characterize a noisy land cover over a short time period (Fig. 5). When bad model fits are initialized or image noise is apparent in the monitoring window, it will be susceptible to false changes. The OLS-CUSUM does not utilize a moving window, and does not perform an iterative regression that requires a training period, and thus it is less prone to these types of changes.

For the South Dakota scene, 033/029, the changes detected by CUSUM are highly concentrated in the 2000–2002 time period. We hypothesize that these changes are in response to an extreme drought that occurred in the region, one that cost the State an estimated $375–550 billion in agriculture yield (Diersen and Taylor, 2003). Over 80% (80.61%) of the breaks detected by CUSUM in the scene belonged to the Change class in the reference data, indicating they were mostly accurate. Similar results can be seen in the Colorado scene, which was experiencing the same drought. An example of a break during this time period can be seen in Fig. 14 in which CUSUM detects a break in the time series of bands 2 and 4. The event does not influence Band 5 as much as the other bands and therefore the cumulative residuals do not exceed the change threshold. However, the threshold is exceeded in a majority of the bands and therefore a break is still detected. More research would be needed to validate the drought hypothesis.

While the approach of running multiple CUSUM models in parallel across bands has proven beneficial, it does have its disadvantages. First, it does not take advantage of changes in the covariance structure between the bands. Second, there is likely to be redundancy in the tests

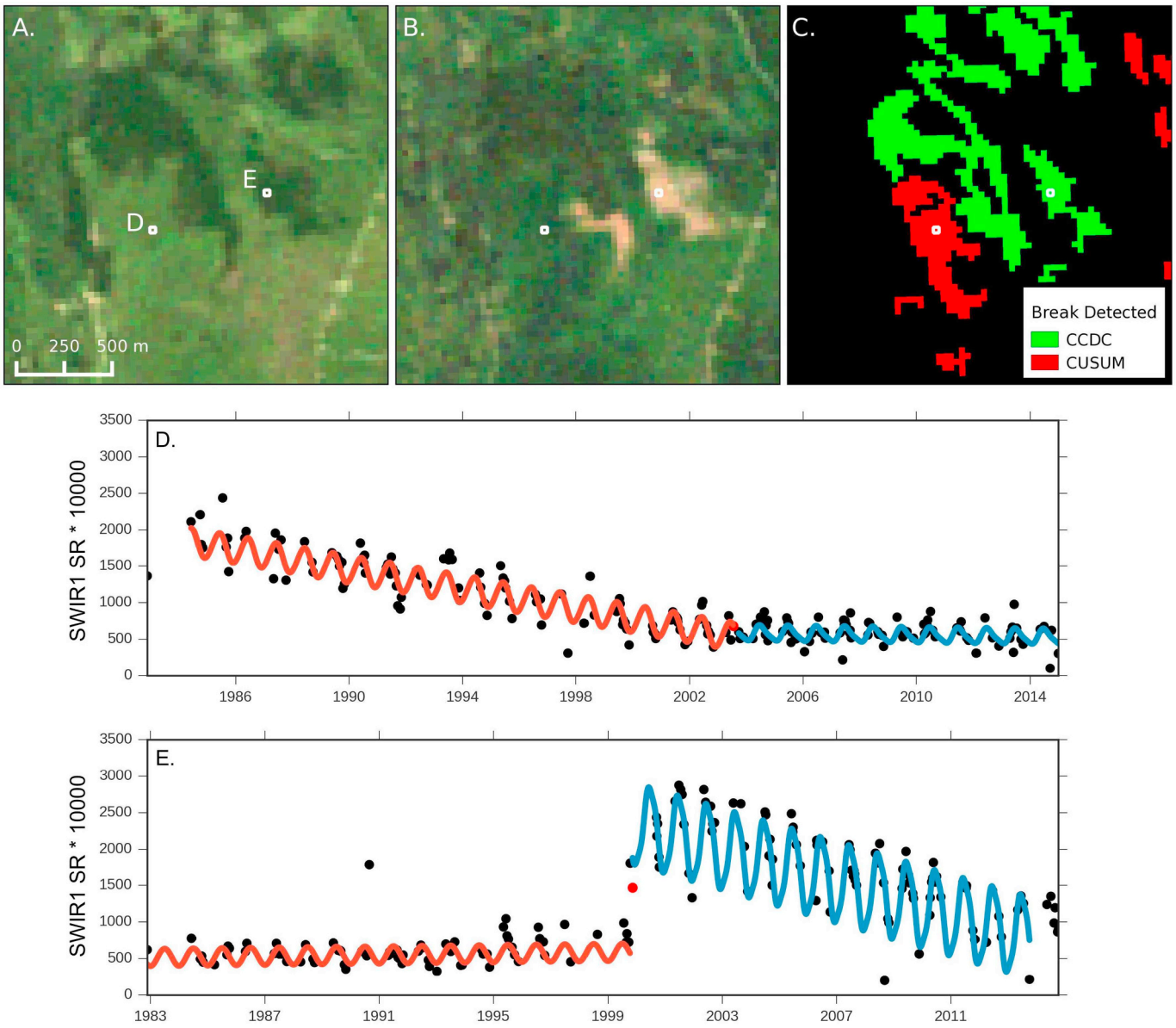

Fig. 11. A. A Landsat 5 3-2-1 image in Path/Row 013/029 (New England) for 1985-244. B. A Landsat 5 image from 2000 to 270. C. Change detection map showing areas labeled as a break by CCDC and by CUSUM. D. A SWIR1 time series for a pixel found by the CUSUM algorithm for the pixel location shown in A as "D". The pixel shows a re-growing forest until 2003 when the signal stabilizes. E. A SWIR1 time series with a change that was found in the CCDC component of the algorithm. The time series shows a deforestation event in 2000 with the pixel location being shown in A as "E". Change dates for both time series are labeled with a red dot. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

due to band-to-band correlation. Multivariate extensions of the general CUSUM test have been proposed (Woodall and Ncube, 1985; Healy, 1987; Crosier, 1988), but the parallel approach was chosen for its simplicity and tested effectiveness. Despite the shortcomings, running multiple CUSUMs in parallel across variables is a technique common in multivariate break detection (Frisén, 2011).

The commission rate, or percentage of falsely identified changes, when using both CCDC and CUSUM for change detection was 25.7%. The Chow Test had a User's Accuracy of 87.4%, indicating that a majority of the changes removed were errors of commission that were detected with CCDC or CUSUM. However, only 13.8% of the total errors of commission for the Change class were identified with the Chow Test. Therefore, similar to CUSUM, the Chow Test was found to be correctly removing (or in the case of CUSUM, adding) a majority of the changes that it modified, but was overly conservative in adding/removing changes. As a result, both algorithms had a high User's Accuracy for the Change class but low Producer's Accuracy. Testing the optimal parameters for balancing errors of omission and commission in an ensemble approach is an area for future research, but clearly less restrictive thresholds for finding change would improve overall accuracy of break detection.

The Chow Test removed breaks in models that were mostly 1–6 years in length (Fig. 15). The shortness of the models removed, in addition to the high User's Accuracy, indicates that the Chow Test is accurately identifying some of the spurious breaks that were previously incorrectly identified by CCDC. There were significantly more models merged in the scenes predominantly containing agriculture and grasslands. The fact that these land covers often have noisy time series compared to other land covers (see Zhu and Woodcock, 2014a, 2014b), indicates that CCDC is detecting too many breaks in noisy environments, but also that some of these breaks are being effectively removed by the Chow Test. Examples of the Chow Test removing unnecessary

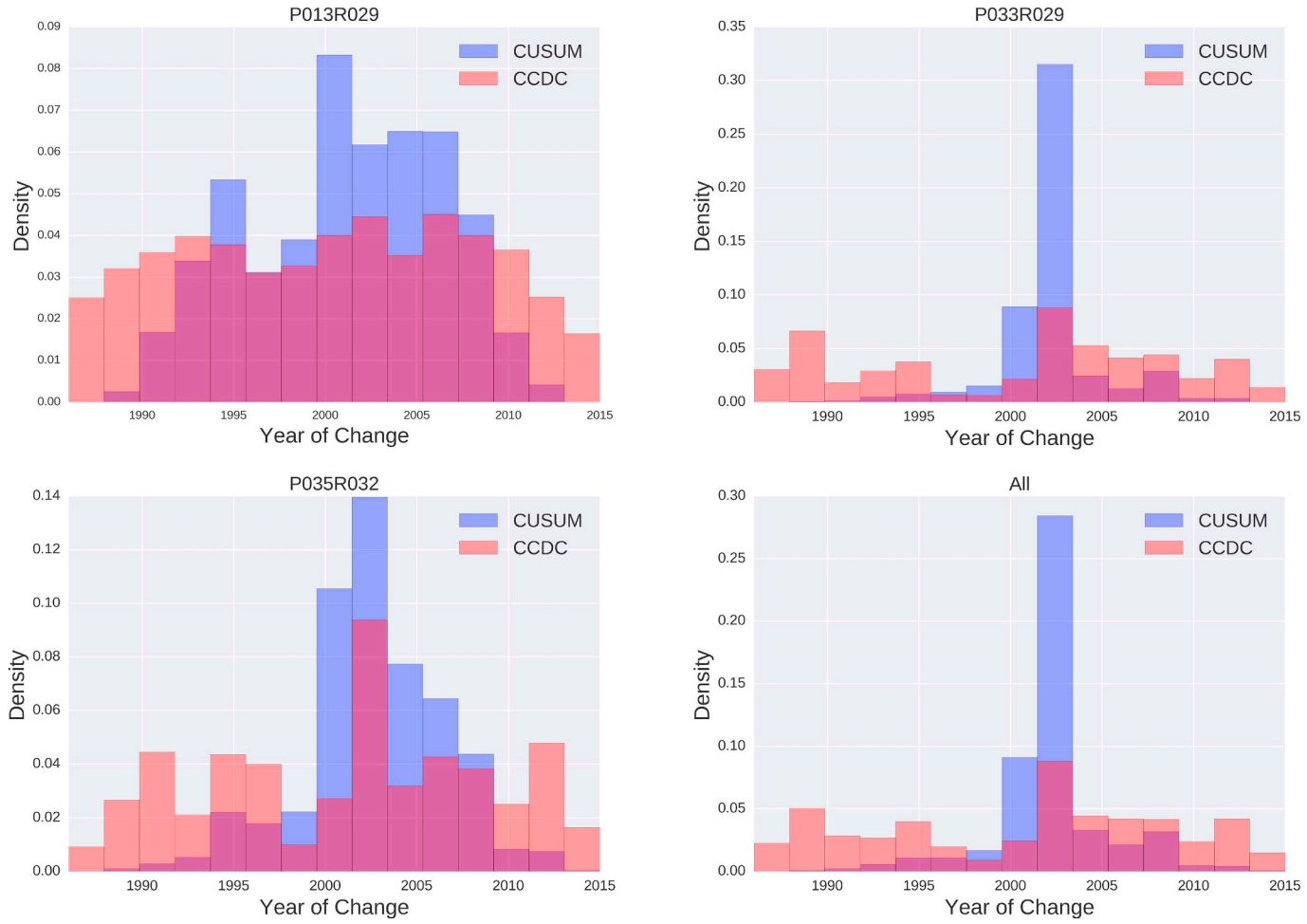

Fig. 12. Histogram of years of changes detected by CCDC and CUSUM representing the number of breaks occurring by year.

breaks near the beginning of a time segment can be seen in Fig. 16. In Fig. 16A, the segment occurs at the beginning of the time series, and in Fig. 16C it is after a new model is initiated due to a deforestation event. In these environments the Chow Test is successfully "cleaning up" the break detection results. This effect can be seen spatially in Fig. 17 in which a grassland environment undergoes a change event. Fig. 17A shows all the breaks detected by the model including those that were removed by the Chow Test. Fig. 17B shows the breaks that remain after all 3 algorithms in the ensemble. The obvious change event is still prevalent, but many small and isolated breaks are removed. The result is a "cleaner" looking map of breaks that were identified by the ensemble.

The Chow Test operates on the pixel level and without regard to spatial information or land cover. There are different approaches that could have potentially been used for the same purpose. For example, Powell and Roberts (2010) defined a set of realistic land cover conversions and removed any that did not fit the criteria, leading to a reduction in "speckle" similar to Fig. 17. In a different approach, LandTrendr uses spatial segmentation to identify disturbance "patches" and therefore minimize noise in the results. Incorporating spatial and land cover information as a means for detecting errors of commission is an area for future research.

The Overall Accuracy was not significantly different when using only CCDC, CCDC and CUSUM, and the ensemble, with each case being above 98% (Table 2D). The high Overall Accuracy was strongly influenced by the large proportion of Pixel Years that were correctly labeled as No Change (96.1%). While the Producer's Accuracy was low for CUSUM, it had the effect of raising the Producer's Accuracy of the Change class for the ensemble from 71.0% to 74.0%. Similarly, the Chow Test failed to remove an estimated 86.2% of the total errors of commission, however it improved the User's Accuracy of the Change class for the ensemble from 79.0% to 81.0%. To summarize, CUSUM and the Chow Test improved the User's and Producer's Accuracy of the ensemble even though their individual Producer's Accuracies were low. While the increase in accuracy between the ensemble and the single or double algorithm approach was minimal, the high algorithm-specific User's Accuracies (84.2% and 87.4% for CUSUM and the Chow Test, respectively) indicate that they are providing a net positive impact in the change detection. We believe that the parameterization of the algorithms to be less conservative in the change detection would have the effect of lowering the algorithm-specific User's Accuracies but also improving the ensemble results. However, we made no attempt at parameter optimization and suggest this as an area for future research.

None of the three algorithms in the ensemble were true multivariate methodologies, but instead run in parallel across bands with different approaches to dissolving the results. This was done to simplify each individual algorithm without requiring the choice of a single spectral band or transformation. Therefore, our approach can be considered a sequential ensemble, with each of the individual algorithms running in parallel across bands. Ideally, each implementation would use multivariate extensions of the univariate algorithms that we used in parallel. While true multivariate approaches for detecting change in a time series exist (Crosier, 1988; Kuncheva, 2013), most land cover monitoring applications using multivariate data for structural break detection have relied primarily on parallel univariate tests (Zhu, 2017). Regardless, we believe that running the tests in parallel across bands, and combining

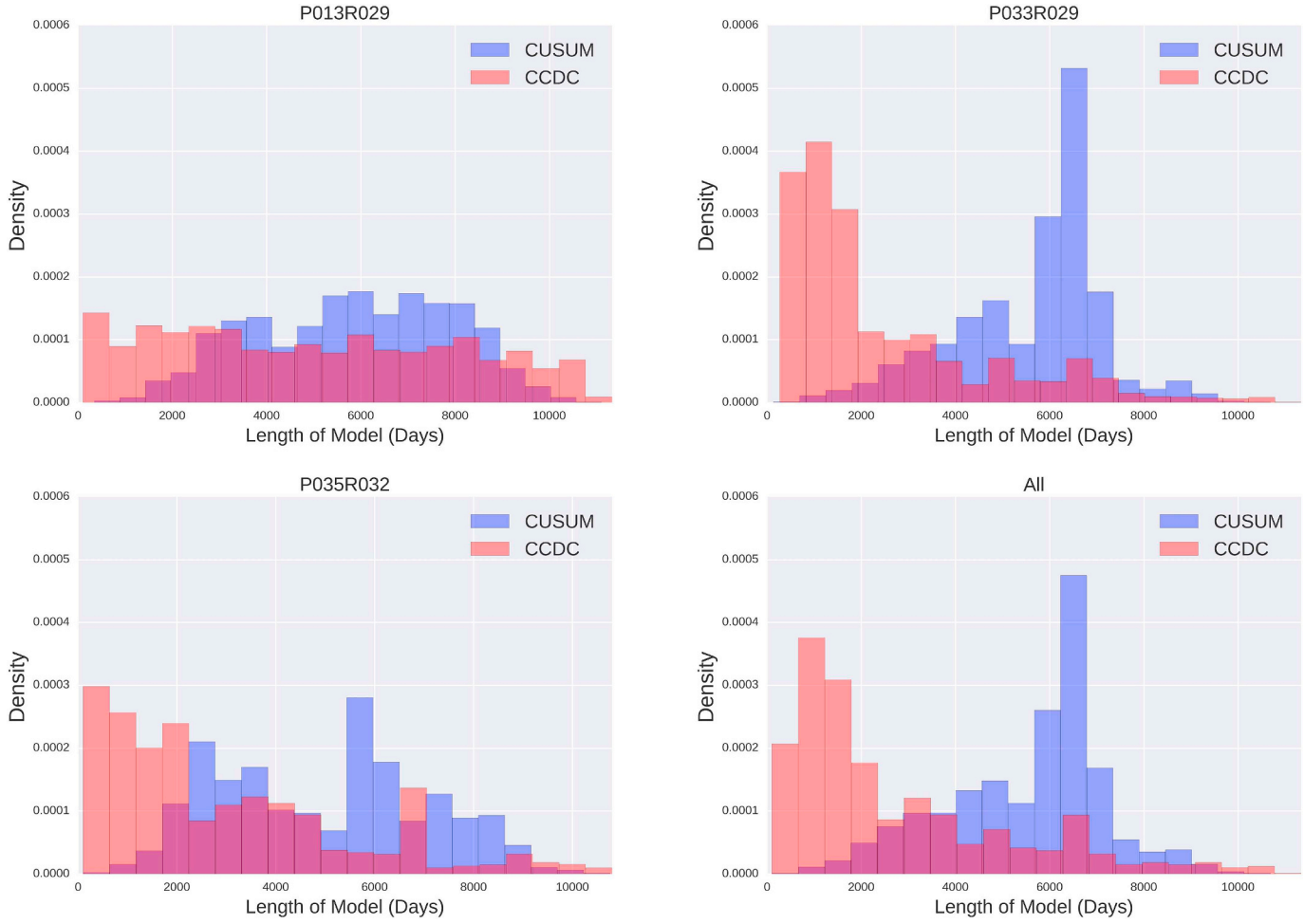

Fig. 13. Normalized density plots of lengths of models before a break is detected for CUSUM and CCDC. Note that CUSUM detected breaks more often in models that are 10–20 years long, while CCDC detected the most breaks earlier in the model period.

them sequentially, makes our approach more robust to different types of changes than approaches that rely on a single spectral band or index. This approach helps the algorithm be more useful in a variety of land change applications, such as monitoring forest regrowth (Fig. 11E), deforestation (Fig. 11E), drought (Fig. 14), and fires (Fig. 17).

It is worth noting that the increase in accuracy that can be achieved using methodologies that rely on large quantities of data may not justify the computational requirements needed to run them. Analysis was performed across approximately 200 high-speed computing nodes, with a total computation time for the three scenes of approximately 48 h (or 9600 total core hours). Given the minimal improvement in accuracy it is therefore not recommended nor is it feasible for such an approach to be run over large areas without a similar computing system. These limitations will be less constraining in the future as satellite data is

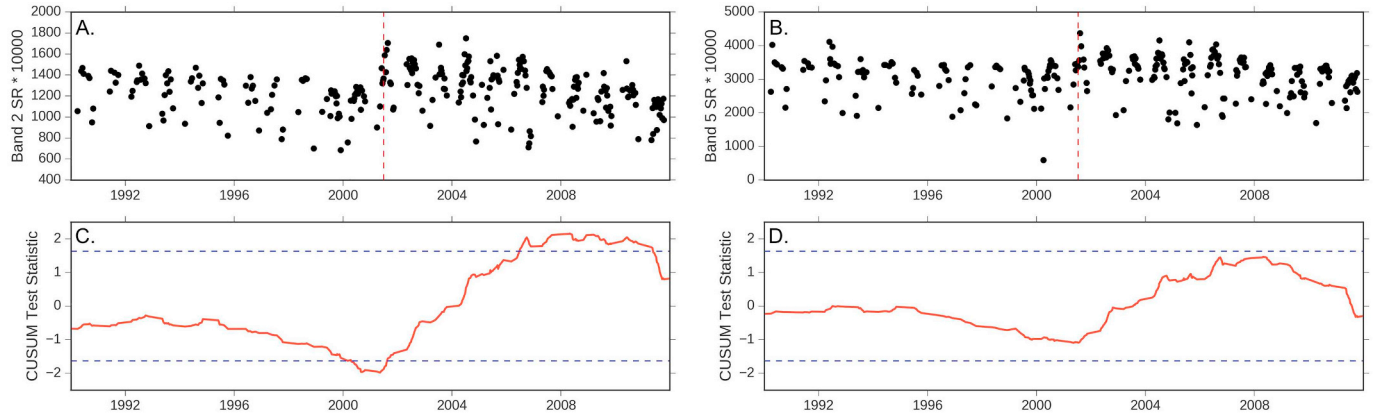

Fig. 14. The time series in a grassland in scene 035/032 (Colorado) with a model break in 2001 shown in the time series of the green (A.) and SWIR1 (B.) bands. The cumulative sum of residuals for the green time series (C.) exceeded the critical region (dashed blue), but not for the SWIR1 band (D.) CUSUM detected a large number of changes from 2000 to 2002 in the two scenes composed largely of grassland and agriculture (035/032 and 033/029). During this time period the two scenes were undergoing a severe drought. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

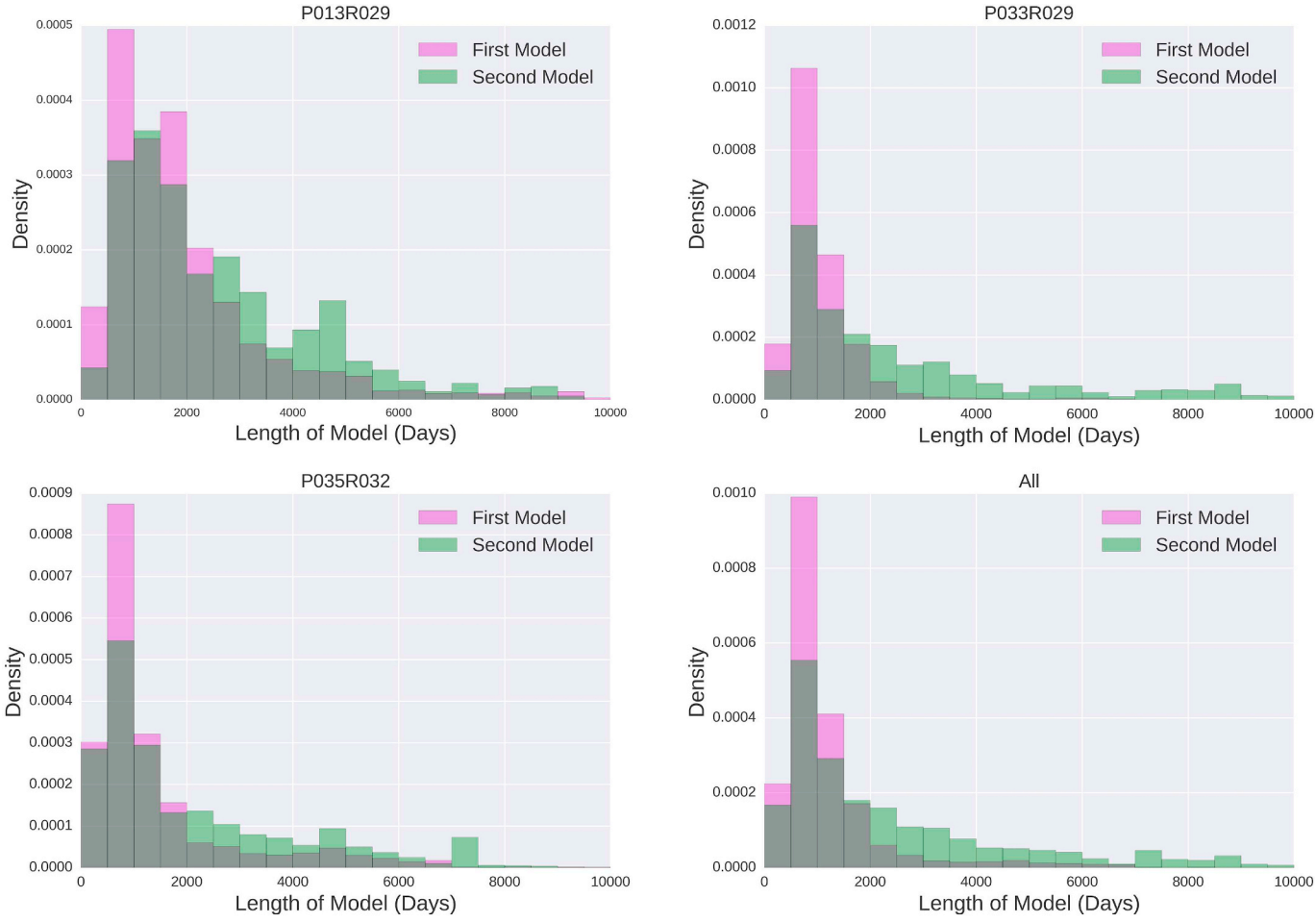

Fig. 15. Normalized histograms of model length for both regression models merged by Chow Test.

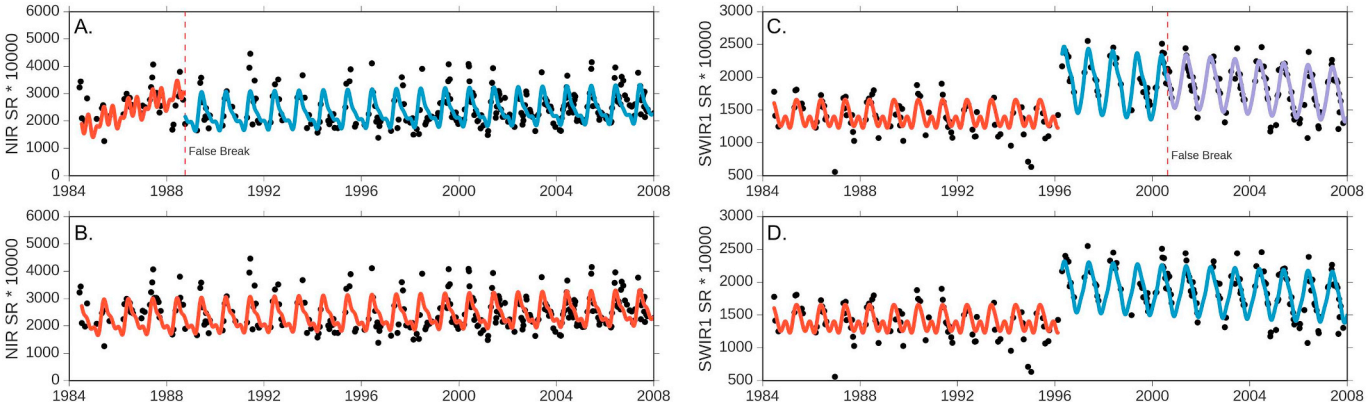

Fig. 16. Two pixels time series before and after the Chow Test. A. The near-infrared time series explained in Fig. 5 before the Chow Test and with an incorrect break in the time series; B. The model results after the Chow Test and with the break correctly removed; C. A SWIR1 time series in 033/029 undergoing a deforestation event in 1996. The break from the deforestation event is correctly labeled, however a secondary break in incorrectly labeled in 2000. D. After the Chow Test the deforestation event is still labeled as a break but the secondary break is removed.

readily becoming available on cloud computing systems such as Amazon Web Services (AWS) and the Google Earth Engine (Gorelick et al., 2017). These computing environments should allow for new approaches that utilize multiple, data intensive methodologies for detailed monitoring of Earth surface processes. Additionally, these platforms will enable easier "fine tuning" of algorithms by reducing the running time needed for a single set of parameters.

4. Conclusion

The use of ensemble algorithms has the potential to improve systems for land change monitoring. The ensemble algorithm presented here has been shown to improve the detection of breaks in time series of Landsat data. Of the breaks either added or removed by the CUSUM and Chow Tests over 84% we considered accurate and beneficial. The

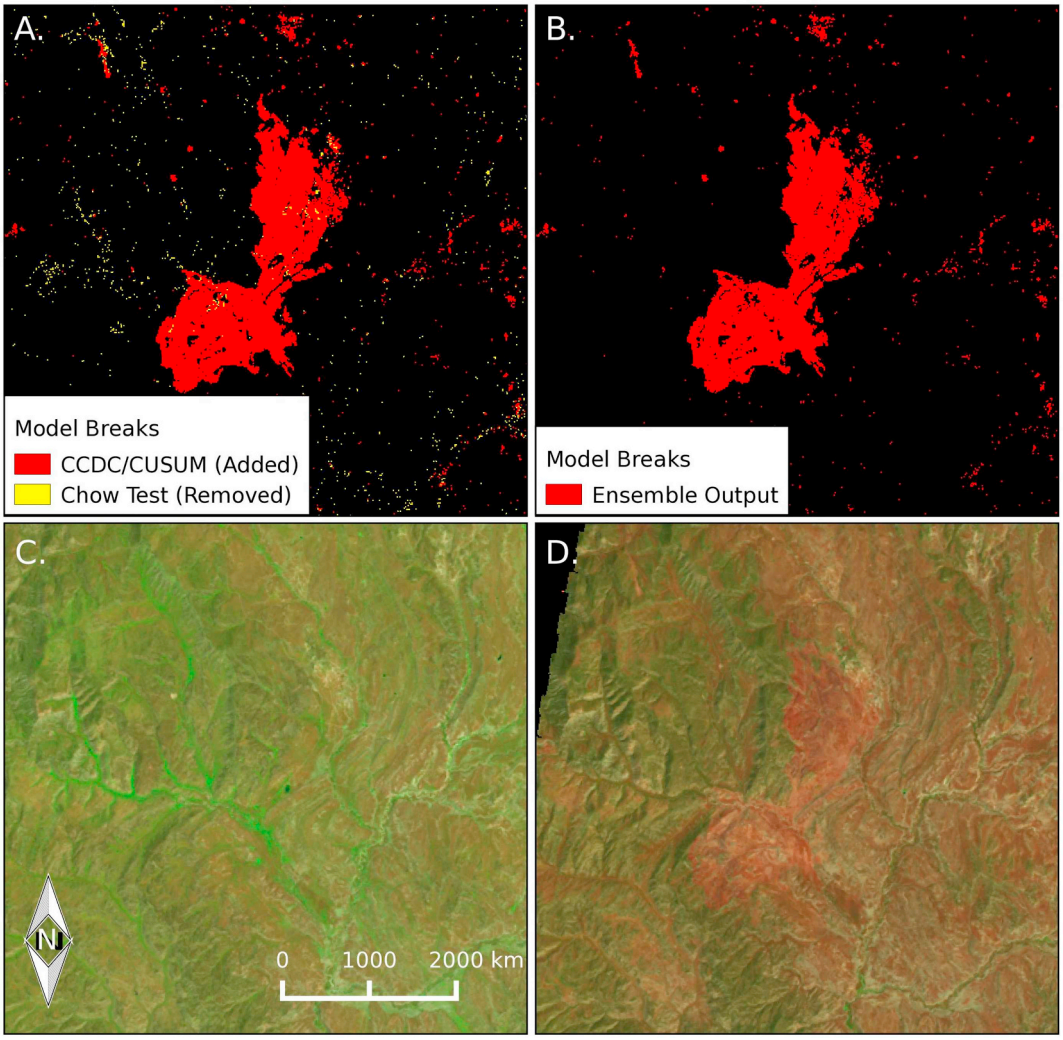

Fig. 17. Mapped break detection results in Path/Row 035/032 from 1984 to 1990. The location contains a large area on grassland that undergoes a disturbance event. The change event is likely due to a fire. A. Breaks that were added by CCDC or CUSUM and removed by the Chow Test. B. The final break detection results after all the algorithms in the ensemble. C. A Landsat 5 image from 1984 to 188 before the large change event. D. A Landsat 5 image from 1987 to 228 after the change event. Notice that the Chow Test is removing breaks in mostly small and randomly dispersed locations, leaving behind large patches that were successfully detected with CUSUM or CCDC. The Chow Test is working to visually "clean up" the break detection results.

CUSUM Test was successful in finding changes in predominantly long series time series that were not detected by CCDC. The Chow Test was effective at removing breaks due to ephemeral changes that were detected by CCDC but were deemed unnecessary. While the breaks that were added or removed were found to be mostly correct, the low Producer's Accuracies indicate that we should be less conservative in the parameterization of the break detection algorithms in order to detect more of the breaks that were previously omitted.

The inherent offline approach of both versions of the Chow and CUSUM tests utilizes the data to test for structural breaks in a different way than CCDC. By fitting regression models across the entire time period and looking for breaks based on the residuals, CUSUM has the advantage of not being prone to false breaks caused from bad training model initialization or groups of bad observations due to clouds. These short time series and corresponding breaks are common in CCDC results for "noisy" land covers, particularly agriculture, and many are being effectively removed by the Chow Test.

The presented research aimed to improve the detection of breaks in a time series that are related to changes of land cover or condition. Detecting a break is only the first step in most analysis, however. The break still needs to be attributed to a physical process such as deforestation or urbanization. It should be expected that improved break detection would lead to better land change characterization, but that was not attempted in this research. There is still much research needed on the inherent advantages and disadvantages of different break detection algorithms when used for land monitoring applications.

Acknowledgements

This research was funded by NASA through the NASA Earth Science Fellowship (16-EARTH16F-295) in addition to the USGS Landsat Science Team Program for Better Use of the Landsat Temporal Domain: Monitoring Land Cover Type, Condition and Change (grant number G12PC00070). We are also grateful for the USGS Earth Resources Observation and Science Center for the data used in the analysis. Finally, we thank the Scientific Python community for the development and distribution of the software used for our analysis.

References

Bauer, P., Hackl, P., 1978. The use of MOSUMS for quality control. Technometrics 20, 431–436.

Brent, R.P., 1971. An algorithm with guaranteed convergence for finding a zero of a function. Comput. J. https://doi.org/10.1093/comjnl/14.4.422.

Briem, G.J., Benediktsson, J.A., Sveinsson, J.R., 2002. Multiple classifiers applied to

multisource remote sensing data. IEEE Trans. Geosci. Remote Sens. 40, 2291–2299. https://doi.org/10.1109/TGRS.2002.802476.
Brooks, E.B., Thomas, V.a., Wynne, R.H., Coulston, J.W., 2012. Fitting the multitemporal curve: a Fourier series approach to the missing data problem in remote sensing analysis. IEEE Trans. Geosci. Remote Sens. 50, 3340–3353. https://doi.org/10.1109/TGRS.2012.2183137.
Brooks, E.B., Wynne, R.H., Thomas, V.A., Blinn, C.E., Coulston, J.W., 2014. On-the-fly massively multitemporal change detection using statistical quality control charts and landsat data. IEEE Trans. Geosci. Remote Sens. 52, 3316–3332. https://doi.org/10.1109/TGRS.2013.2272545.
Brown, R.L., Durbin, J., Evans, J.M., 1975. Techniques for testing the constancy of regression relationships over time. J. R. Stat. Soc. 37, 149–192.
Bruzzone, L., Cossu, R., Vernazza, G., 2004. Detection of land-cover transitions by combining multidate classifiers. Pattern Recogn. Lett. 1491–1500. https://doi.org/10.1016/j.patrec.2004.06.002.
Bullock, E.L., Fagherazzi, S., Nardin, W., Vo-Luong, P., Nguyen, P., Woodcock, C.E., 2017. Temporal patterns in species zonation in a mangrove forest in the Mekong Delta, Vietnam, using a time series of Landsat imagery. Cont. Shelf Res. https://doi.org/10.1016/j.csr.2017.07.007.
Chow, G.C., 1960. Tests of equality between sets of coefficients in two linear regressions. Econometrica 28, 591–605. https://doi.org/10.2307/1910133.
Clemen, R.T., 1989. Combining forecast: a review and annotated bibliography. Int. J. Forecast. 5, 559–583. https://doi.org/10.1016/0169-2070(89)90012-5.
Cohen, W., Healey, S., Yang, Z., Stehman, S., Brewer, C., Brooks, E., Gorelick, N., Huang, C., Hughes, M., Kennedy, R., Loveland, T., Moisen, G., Schroeder, T., Vogelmann, J., Woodcock, C., Yang, L., Zhu, Z., 2017. How similar are forest disturbance maps derived from different Landsat time series algorithms? Forests 8, 98. https://doi.org/10.3390/f8040098.
Crosier, R.B., 1988. Multivariate generalizations of cumulative sum quality control schemes. Technometrics 30, 291–303. https://doi.org/10.2307/1270083.
DeVries, B., Decuyper, M., Verbesselt, J., Zeileis, A., Herold, M., Joseph, S., 2015a. Tracking disturbance-regrowth dynamics in tropical forests using structural change detection and Landsat time series. Remote Sens. Environ. 169, 320–334. https://doi.org/10.1016/j.rse.2015.08.020.
DeVries, B., Verbesselt, J., Kooistra, L., Herold, M., 2015b. Robust monitoring of small-scale forest disturbances in a tropical montane forest using Landsat time series. Remote Sens. Environ. 161, 107–121. https://doi.org/10.1016/j.rse.2015.02.012.
Diersen, M.A., Taylor, G., 2003. Examining Economic Impact and Recovery in South Dakota From the 2002 Drought.
Dietterich, T.G., 2002. Ensemble learning. In: The Handbook of Brain Theory and Neural Networks, pp. 110–125.
Fearnhead, P., Rigaill, G., 2018. Changepoint detection in the presence of outliers. J. Am. Stat. Assoc. 1–15.
Foody, G.M., Boyd, D.S., Hernandez, C.S., 2007. Mapping a specific class with an ensemble of classifiers. Int. J. Remote Sens. 28, 1733–1746. https://doi.org/10.1080/01431160600962566.
Fragal, E.H., Silva, T.S.F., Novo, E.M.L. de M., 2016. Reconstructing historical forest cover change in the Lower Amazon floodplains using the LandTrendr algorithm. Acta Amaz 46, 13–24. https://doi.org/10.1590/1809-4392201500835.
Frisén, M., 2011. On multivariate control charts. Produção 21, 235–241. https://doi.org/10.1590/S0103-65132011005000010.
Giacinto, G., Roli, F., Bruzzone, L., 2000. Combination of neural and statistical algorithms for supervised classification of remote-sensing images. Pattern Recogn. Lett. 21, 385–397. https://doi.org/10.1016/S0167-8655(00)00006-4.
Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google Earth Engine: planetary-scale geospatial analysis for everyone. Remote Sens. Environ. https://doi.org/10.1016/j.rse.2017.06.031.
Healy, J.D., 1987. A note on multivariate CUSUM procedures. Technometrics 29 (4), 409–412.
Healey, S.P., Cohen, W.B., Yang, Z., Kenneth Brewer, C., Brooks, E.B., Gorelick, N., Hernandez, A.J., Huang, C., Joseph Hughes, M., Kennedy, R.E., Loveland, T.R., Moisen, G.G., Schroeder, T.A., Stehman, S.V., Vogelmann, J.E., Woodcock, C.E., Yang, L., Zhu, Z., 2018. Mapping forest change using stacked generalization: an ensemble approach. Remote Sens. Environ. 204, 717–728. https://doi.org/10.1016/j.rse.2017.09.029.
Hermosilla, T., Wulder, M.A., White, J.C., Coops, N.C., Hobart, G.W., 2017. Updating Landsat time series of surface-reflectance composites and forest change products with new observations. Int. J. Appl. Earth Obs. Geoinf. 63, 104–111. https://doi.org/10.1016/j.jag.2017.07.013.
Hinkley, D.K., 1971. Inference about the change-point from cumulative sum tests. Biometrika1 58, 509–523.
Holden, C.E., 2017. TSTools. https://doi.org/10.5281/zenodo.267110.
Ju, J., Masek, J.G., 2016. The vegetation greenness trend in Canada and US Alaska from 1984–2012 Landsat data. Remote Sens. Environ. 176, 1–16. https://doi.org/10.1016/j.rse.2016.01.001.
Kennedy, R.E., Yang, Z., Cohen, W.B., 2010. Detecting trends in forest disturbance and recovery using yearly Landsat time series: 1. LandTrendr - temporal segmentation algorithms. Remote Sens. Environ. 114, 2897–2910. https://doi.org/10.1016/j.rse.2010.07.008.
Kittler, J., Hatef, M., Duin, R., Matas, J., 1998. On combining classifiers. IEEE Trans. Pattern Anal. Mach. Intell. 226–239. https://doi.org/10.1109/ICPR.1996.547205.
Klein, J.L., 1997. Statistical Visions in Time: A History of Time Series Analysis. Cambridge University Press, pp. 1662–1938.
Kuncheva, L.I., 2013. Change detection in streaming multivariate data using likelihood detectors. IEEE Trans. Knowl. Data Eng. 25, 1175–1180. https://doi.org/10.1109/32.106988.
Li, Y.Y., Zhang, H., Kainz, W., 2012. Monitoring patterns of urban heat islands of the fast-growing Shanghai metropolis, China: using time-series of Landsat TM/ETM + data. Int. J. Appl. Earth Obs. Geoinf. 19, 127–138. https://doi.org/10.1016/j.jag.2012.05.001.
Loveland, T.R., Sohl, T., Sayler, K., Gallant, A., Dwyer, J., Vogelmann, J., Zylstra, G., Wade, T.G., Edmonds, C.M., 1999. Land Cover Trends: Rates, Causes, and Consequences of Late-twentieth Century US Land Cover Change. US Environmental Protection Agency, National Exposure Research Laboratory, Office of Research and Development.
Olofsson, P., Foody, G.M., Stehman, S.V., Woodcock, C.E., 2013. Making better use of accuracy data in land change studies: estimating accuracy and area and quantifying uncertainty using stratified estimation. Remote Sens. Environ. 129, 122–131. https://doi.org/10.1016/j.rse.2012.10.031.
Olofsson, P., Foody, G.M., Herold, M., Stehman, S.V., Woodcock, C.E., Wulder, M.a., 2014. Good practices for estimating area and assessing accuracy of land change. Remote Sens. Environ. 148, 42–57. https://doi.org/10.1016/j.rse.2014.02.015.
Olofsson, P., Holden, C.E., Bullock, E.L., Woodcock, C.E., 2016. Time series analysis of satellite data reveals continuous deforestation of New England since the 1980s. Environ. Res. Lett. 11, 64002. https://doi.org/10.1088/1748-9326/11/6/064002.
Pasquarella, V.J., Holden, C.E., Woodcock, C.E., 2018. Improved mapping of forest type using spectral-temporal Landsat features. Remote Sens. Environ. 210, 193–207.
Pflugmacher, D., Cohen, W.B., Kennedy, R.E., Yang, Z., 2013. Using Landsat-derived disturbance and recovery history and lidar to map forest biomass dynamics. Remote Sens. Environ. 151, 124–137. https://doi.org/10.1016/j.rse.2013.05.033.
Ploberger, W., Kramer, W., 1992. The Cusum test with Ols residuals. Econometrica 60, 271–285. https://doi.org/10.3982/ECTA9956.
Polikar, R., 2009. Ensemble learning. Scholarpedia 4, 2776. https://doi.org/10.4249/scholarpedia.2776.
Powell, Rebecca L., Roberts, Dar A., 2010. Characterizing urban land-cover change in Rondônia, Brazil: 1985 to 2000. J. Lat. Am. Geogr. 9, 183–211. https://doi.org/10.1353/lag.2010.0028.
Radke, R.J., Andra, S., Al-Kofahi, O., Roysam, B., 2005. Image change detection algorithms: a systematic survey. IEEE Trans. Image Process. 14, 294–307 (doi:10.1.1.5.291).
Rindfuss, R.R., Walsh, S.J., Turner, B.L., Fox, J., Mishra, V., 2004. Developing a science of land change: challenges and methodological issues. Proc. Natl. Acad. Sci. U. S. A. 101, 13976–13981. https://doi.org/10.1073/pnas.0401545101.
Robbins, M., Gallagher, C., Lund, R., Aue, A., 2011. Mean shift testing in correlated data. J. Time Ser. Anal. 32, 498–511. https://doi.org/10.1111/j.1467-9892.2010.00707.x.
Roberts, S.W., 1959. Control chart tests based on geometric moving averages. Technometrics 1, 239–250.
Rokach, L., Maimon, O., 2005. Data Mining and Knowledge Discovery Handbook. Springer US, New York, NY.
Roy, D.P., Ju, J., Lewis, P., Schaaf, C., Gao, F., Hansen, M., Lindquist, E., 2008. Multi-temporal MODIS-Landsat data fusion for relative radiometric normalization, gap filling, and prediction of Landsat data. Remote Sens. Environ. 112, 3112–3130. https://doi.org/10.1016/j.rse.2008.03.009.
Saxena, R., Watson, L.T., Thomas, V.A., Wynne, R.H., 2017. Scaling constituent algorithms of a trend and change detection polyalgorithm. In: Proceedings of the 25th High Performance Computing Symposium, pp. 6.
Schultz, M., Clevers, J.G.P.W., Carter, S., Verbesselt, J., Avitabile, V., Quang, H.V., Herold, M., 2016. Performance of vegetation indices from Landsat time series in deforestation monitoring. Int. J. Appl. Earth Obs. Geoinf. 52, 318–327. https://doi.org/10.1016/j.jag.2016.06.020.
Shimizu, K., Ahmed, O.S., Ponce-Hernandez, R., Ota, T., Win, Z.C., Mizoue, N., Yoshida, S., 2017. Attribution of disturbance agents to forest change using a Landsat time series in tropical seasonal forests in the Bago Mountains, Myanmar. For 8, 218 8. 218. https://doi.org/10.3390/F8060218.
Steele, B., Patterson, D., 2002. Land Cover Mapping Using Combination and Ensemble Classifiers.
Stehman, S.V., 1997. Estimating standard errors of accuracy assessment statistics under cluster sampling. Remote Sens. Environ. 60, 258–269. https://doi.org/10.1016/S0034-4257(96)00176-9.
Stehman, S.V., 1999. Basic probability sampling designs for thematic map accuracy assessment. Int. J. Remote Sens. 20, 2423–2441. https://doi.org/10.1016/S0034-4257(99)00090-5.
Stehman, S.V., 2013. Estimating area from an accuracy assessment error matrix. Remote Sens. Environ. 132, 202–211. https://doi.org/10.1016/j.rse.2013.01.016.
Stehman, S.V., 2014. Estimating area and map accuracy for stratified random sampling when the strata are different from the map classes. Int. J. Remote Sens. 35, 4923–4939. https://doi.org/10.1080/01431161.2014.930207.
Sulla-Menashe, D., Friedl, M.A., Woodcock, C.E., 2016. Sources of bias and variability in long-term Landsat time series over Canadian boreal forests. Remote Sens. Environ. 177, 206–219. https://doi.org/10.1016/j.rse.2016.02.041.
Tyukavina, A., Hansen, M.C., Potapov, P.V., Stehman, S.V., Smith-Rodriguez, K., Okpa, C., Aguilar, R., 2017. Types and rates of forest disturbance in Brazilian Legal Amazon, 2000–2013: supplementary materials. Sci. Adv. 3, e1601047. https://doi.org/10.1126/sciadv.1601047.
Verbesselt, J., Hyndman, R., Newnham, G., Culvenor, D., 2010. Detecting trend and seasonal changes in satellite image time series. Remote Sens. Environ. 114, 106–115. https://doi.org/10.1016/j.rse.2009.08.014.
Verbesselt, J., Zeileis, A., Herold, M., 2012a. Near real-time disturbance detection using satellite image time series. Remote Sens. Environ. 123, 98–108. https://doi.org/10.1016/j.rse.2012.02.022.
Verbesselt, J., Zeileis, A., Hyndman, R., 2012b. Package “bfast”.
Wolpert, D.H., Macready, W.G., 1997. No free lunch theorems for optimization. IEEE

Trans. Evol. Comput. 1, 67–82. https://doi.org/10.1109/4235.585893.
Woodall, W.H., Ncube, M.M., 1985. Multivariate CUSUM quality-control procedures. Technometrics 27 (3), 285–292.
Woźniak, M., Graña, M., Corchado, E., 2014. A survey of multiple classifier systems as hybrid systems. Inf. Fusion 16, 3–17. https://doi.org/10.1016/j.inffus.2013.04.006.
Young, S., 2017. Land change monitoring, assessment, and projection (LCMAP) revolutionizes land cover and land change research. In: United States Geological Survey Information Product. 172https://doi.org/10.3133/gip172.
Zeileis, A., Kleiber, C., Walter, K., Hornik, K., 2003. Testing and dating of structural changes in practice. Comput. Stat. Data Anal. 44, 109–123. https://doi.org/10.1016/S0167-9473(03)00030-6.
Zeileis, A., Leisch, F., Kleiber, C., Hornik, K., 2005. Monitoring structural change in dynamic econometric models. J. Appl. Econ. 20, 99–121. https://doi.org/10.1002/jae.776.
Zeileis, A., Shah, A., Patnaik, I., 2010. Testing, monitoring, and dating structural changes in exchange rate regimes. Comput. Stat. Data Anal. 54, 1696–1706. https://doi.org/10.1016/j.csda.2009.12.005.
Zhu, Z., 2017. Change detection using landsat time series: a review of frequencies, preprocessing, algorithms, and applications. ISPRS J. Photogramm. Remote Sens. 130, 370–384. https://doi.org/10.1016/j.isprsjprs.2017.06.013.
Zhu, Z., Woodcock, C., 2012. Object-based cloud and cloud shadow detection in Landsat imagery object-based cloud and cloud shadow detection in Landsat imagery. Remote Sens. Environ. 118, 83–94. https://doi.org/10.1016/j.rse.2011.10.028.
Zhu, Z., Woodcock, C.E., 2014a. Continuous change detection and classification of land cover using all available Landsat data. Remote Sens. Environ. 144, 152–171. https://doi.org/10.1016/j.rse.2014.01.011.
Zhu, Z., Woodcock, C.E., 2014b. Automated cloud, cloud shadow, and snow detection in multitemporal Landsat data: an algorithm designed specifically for monitoring land cover change. Remote Sens. Environ. 152, 217–234. https://doi.org/10.1016/j.rse.2014.06.012.
Zhu, Z., Wang, S., Woodcock, C., 2015a. Improvement and expansion of the Fmask algorithm: cloud, cloud shadow and snow detection. Remote Sens. Environ. 159, 269–277.
Zhu, Z., Woodcock, C.E., Holden, C., Yang, Z., 2015b. Generating synthetic Landsat images based on all available Landsat data: predicting Landsat surface reflectance at any given time. Remote Sens. Environ. 162, 67–83. https://doi.org/10.1016/j.rse.2015.02.009.
Zhu, Z., Fu, Y., Woodcock, C.E., Olofsson, P., Vogelmann, J.E., Holden, C., Wang, M., Dai, S., Yu, Y., 2016. Including land cover change in analysis of greenness trends using all available Landsat 5, 7, and 8 images: a case study from Guangzhou, China (2000–2014). Remote Sens. Environ. 185, 243–257.