

LETTER • OPEN ACCESS

## A global time series dataset to facilitate forest greenhouse gas reporting

To cite this article: Noel Gorelick *et al* 2023 *Environ. Res. Lett.* **18** 084001

View the [article online](#) for updates and enhancements.

You may also like

- [Coronene-based charge-transfer complexes](#)  
Yukihiro Yoshida, Kazuhide Isomura, Yoshihide Kumagai et al.
- [Nanorectenna spectrally-selective plasmonic hot electron response to visible-light lasers](#)  
R M Osgood, M Kang, K-B Kim et al.
- [Analysis of adaptability estimation algorithms in dynamic systems](#)  
J N Zuleta and R A L. Franco

ENVIRONMENTAL RESEARCH  
LETTERS

## LETTER

## OPEN ACCESS

## RECEIVED

12 December 2022

## REVISED

26 June 2023

## ACCEPTED FOR PUBLICATION

29 June 2023

## PUBLISHED

11 July 2023

Original content from  
this work may be used  
under the terms of the  
[Creative Commons  
Attribution 4.0 licence](#).

Any further distribution  
of this work must  
maintain attribution to  
the author(s) and the title  
of the work, journal  
citation and DOI.



## A global time series dataset to facilitate forest greenhouse gas reporting

Noel Gorelick<sup>1,2,\*</sup> , Zhiqiang Yang<sup>3</sup> , Paulo Arévalo<sup>4</sup> , Eric L Bullock<sup>3</sup> , Katherin Patricia Insfrán<sup>5</sup> and Sean P Healey<sup>3</sup> <sup>1</sup> Remote Sensing Laboratories, University of Zürich, Zürich, Switzerland<sup>2</sup> Google Switzerland, Zurich, Switzerland<sup>3</sup> US Forest Service, Rocky Mountain Research Station, Ogden, UT, United States of America<sup>4</sup> Department of Earth and Environment, Boston University, Boston, MA, United States of America<sup>5</sup> Técnicos del Instituto Forestal Nacional (INFONA), San Lorenzo, Paraguay

\* Author to whom any correspondence should be addressed.

E-mail: [gorelick@gmail.com](mailto:gorelick@gmail.com)**Keywords:** time-series, global dataset, greenhouse gas, earth engine, CCDC**Abstract**

We have developed a version of the Continuous Change Detection and Classification algorithm within the Google Earth Engine environment. It has been used with 20 years of Landsat data (1999–2019) to produce a new, publicly available global dataset of pre-computed time series break points and harmonic coefficients. We present results from regional use cases demonstrating classification and change detection with this new dataset and compare them to other temporal compositing techniques. Our results demonstrate that gains in overall accuracy using CCDC may be small on a yearly basis, but they are consistent, and improvements in temporal coherence—correctly detecting land use transitions and temporal trends—can be significant. These improvements can translate into better estimates of land use change activity and reduce the uncertainty in the greenhouse gas emissions estimates in REDD+ reporting.

**1. Introduction**

Accounting mechanisms tracking the climate impact of land use change generally involve two components: activity data, which reflect how much land area changed from one land use to another, and emission and removal factors, which quantify mean carbon flux due to those changes (GFOI 2020). Nationally Determined Contributions (NDCs) to the Paris Agreement and other international climate change mitigation frameworks generally involve the reduction of land use transitions that represent net carbon emission. Monitoring such transitions in the form of activity data is vital in estimating emissions from land use changes such as deforestation.

There is great potential to derive land use change information needed in this context from time series of satellite imagery, particularly from platforms such as Landsat with a measurement history spanning almost 40 years (Gómez *et al* 2016, Wulder *et al* 2022),

and the more contemporary Sentinel-2 missions (Drusch *et al* 2012). Annual remotely sensed maps of land use change can be used to generate activity data either directly (Houghton *et al* 2012) or, preferably, as stratification to support the efficient collection of high-quality sample data (Arévalo *et al* 2020, Espejo *et al* 2020). In the latter case, more accurate land use change maps provide more efficient stratification, resulting in more precise estimates of the area of land use and change (Czaplewski and Patterson 2003, Olofsson *et al* 2020).

Under the United Nations Framework Convention on Climate Change (UNFCCC) program for reducing emissions from deforestation and degradation (REDD+), national organizations have reported on their forest reference emission levels (FRELs). REDD+ reporting countries derive their FREL activity data from maps based on the classification of satellite imagery, and while some still use a manual classification-based visual interpretation

to develop their reports (Thailand 2022, Dominica 2022, Indonesia 2022), many are evolving towards supervised classification. In these methods, compositing of pixel values acquired at different dates is commonly performed before classification to counteract the effects of cloud and shadow contamination and to reduce radiometric and phenological noise (Van Doninck and Tuomisto 2017). Compositing also helps mitigate the effects of data unevenly distributed in time or space and standardizes the inputs for classification.

A variety of temporal compositing strategies have been proposed, ranging from simple annual or seasonal statistical reductions such as medians (Roberts *et al* 2017), medoids (Flood 2013), or percentiles (Hansen *et al* 2014), to complex ranking and selection mechanisms designed to select the most representative pixel (White *et al* 2014) or highlight changes during different parts of the year (Potapov *et al* 2020). Countries reporting their greenhouse gas inventories to the UNFCCC select their path through this myriad of options based on their budget in terms of time, expense, technical capacity, and specific geographic challenges. For instance, Paraguay reported using a direct change classification method based on a stack of three best-pixel composites (Paraguay 2015), while Ecuador used annual medoid composites in a post-classification change detection scheme (Ecuador 2020). Peru used direct change classification on a set of annual percentiles, interval means, and linear trends between two years (Peru 2016), and Chile used two-date greenest pixel composites with a series of multitemporal band indices following a modification of the MIICA protocol (Jin *et al* 2013, Chile 2021).

In cases when a lack of data leaves gaps in intra-annual composites, gaps may be filled using data from earlier and later years (Potapov *et al* 2020). Time series analysis methods extend this idea by fitting a function to pixel values through time to produce a continuous and terse representation of spectral change within and across years, simplifying information from hundreds or thousands of images into a few parameters that can be efficiently used in prediction models. Harmonic functions are often used for this purpose to incorporate cyclical reflectance change due to phenology into the models (Brooks *et al* 2013, Wilson *et al* 2018, Bullock *et al* 2020). Classification of each pixel to a land use class then proceeds using a small number of time series harmonic coefficients per radiometric band.

The Continuous Change Detection and Classification algorithm (CCDC, Zhu and Woodcock 2014) is one prominent example of the family of harmonic fitting models. The CCDC algorithm identifies structural breakpoints in a time series by iteratively fitting harmonic curves to the data and

declaring a breakpoint when the difference between predicted and observed reflectance values exceed a dynamic threshold. CCDC outputs include harmonic parameters describing reflectance patterns between each detected breakpoint pair. These parameters may be used directly as independent variables supporting classification, or they can be used to create a synthetic cloud-free image for any historical date (Zhu *et al* 2015). CCDC has been successfully utilized to map forest disturbances (Tang *et al* 2019), urban impervious surfaces (Deng and Zhu 2020), boreal changes (Wang *et al* 2020), and other instances of land cover and land use change.

Using CCDC is computationally expensive, requiring highly iterative model fitting over deep stacks of time series imagery (Pasquarella *et al* 2022). These costs have limited the use of CCDC in practical applications, particularly in developing world settings where improved forest greenhouse gas reporting is urgently needed. We have consequently implemented a version of the CCDC algorithm in the Google Earth Engine environment (Gorelick *et al* 2017) and report the processing of 20 years of the global Landsat archive. This dataset is available as a public asset on Google Earth Engine ([https://developers.google.com/earth-engine/datasets/catalog/GOOGLE\\_GLOBAL\\_CCDC\\_V1](https://developers.google.com/earth-engine/datasets/catalog/GOOGLE_GLOBAL_CCDC_V1)) under a standard CC-BY license.

Given the importance of mapping land use transition in the context of international carbon accounting and considering the new availability of analysis-ready Landsat time series data (Dwyer *et al* 2018), the goal of this paper is to introduce this dataset and explore advantages it may offer in land cover, land use, and biomass mapping, both in terms of point-in-time accuracy and the coherence of predictions made for the same location over time. Only 45 countries, representing 61% of the world's forests, report carbon stocks to the UN Food and Agriculture Organization using designed national forest inventories or high-quality remote sensing methods (Nesha *et al* 2021). This paper evaluates, in settings with ample independent monitoring data, the effects of Landsat preprocessing methods on maps of deforestation and biomass density, both keystone variables in forest carbon accounting.

## 2. Methods

We ported version 12.30 of the MATLAB implementation of CCDC into Google Earth Engine. This port makes it easy to run CCDC on the entire history of Landsat reflectance imagery. However, the Landsat time series at many locations can contain more than 1000 observations, and running CCDC on an image stack that deep can consume upwards of 1000 CPU hours per square degree of land surface area. In order

to facilitate affordable and easy access to this tool, we pre-generated a global dataset of Landsat-based CCDC-detected segments and coefficients and made this product available as an Earth Engine dataset. The dataset was created from the Landsat 5, 7, and 8 Collection-1, Tier-1, surface reflectance time series, using all daytime images between 1999 and 2019. Each image was preprocessed to mask pixels identified as cloud, shadow, or snow, according to the 'pixel\_qa' band, saturated pixels, and pixels with an atmospheric opacity >300 as identified by the 'sr\_atmos\_opacity' and 'sr\_aerosol' bands. Pixels repeated in north/south scene overlap were deduplicated. The results were processed using CCDC and output at 30 m/pixel in 2° tiles for all landmasses between -60° and +85° latitude.

The output contains, on a per-pixel basis, the start, end, and breakpoint times of each detected temporal segment, the harmonic coefficients modeling the segment, the root mean square error (RMSE) indicating how well the model fits, the magnitude of the change that caused the breakpoint to be detected, and an estimate of the probability of the change being real. Each of these outputs is variable-length by nature since the number of detected segments can vary from pixel to pixel. To accommodate the variable number of segments, the CCDC algorithm produces a multi-band image where each pixel in each band contains a 1D or 2D variable-length array, whose sizes are based on the number of segments detected. All of the output bands are described in table 1.

Using this dataset, we performed experiments to investigate its efficacy compared to several other compositing methods in two case studies related to forest carbon monitoring: land use and change mapping for the continental United States and above-ground biomass mapping in Paraguay.

### 2.1. Implementation validation

To validate the Earth Engine implementation of CCDC, we processed a collection of time-series samples using the original MATLAB implementation and the Earth Engine version and compared the results. At each of 57 500 random points, a time series of surface reflectance values were extracted from the combined Landsat 5, 7, and 8, Collection-1, Tier-1 datasets spanning 1986–2020, and preprocessed to remove clouds, cloud shadow, snow, saturated and duplicate pixels.

The MATLAB implementation of CCDC uses LASSO regression in the model fitting process to reduce overfitting of the time series. However, because MATLAB's LASSO implementation is not open source and could not be examined, the results from the Earth Engine implementation are expected to differ slightly. To mitigate this effect as much as

possible, the validation samples were processed twice with each version; once using ordinary least squares (OLS) regression, under which an exact comparison of the outputs could be made, and again using LASSO regression where only the number of segments, the dates and the RMSE of the model predictions were compared.

Our validation showed that using OLS regression, the number of segments, segment start and end times, and the number of observations per fit segment were identical for all samples. The harmonic coefficients and RMSE values matched to within 0.0005% relative error. Using LASSO regression, the segment endpoints only differed in 13 of the 51 011 samples, with most of those disagreeing by only one observation in length. For the 107 491 detected segments, 99.93% of the RMSE values varied by less than 0.1%, with only 117 values varying by 1% or more.

### 2.2. Estimating land use and change over time

As described earlier, activity data comprising information about land use transition is a key component of systems that account for greenhouse gas implications of land use change. Frameworks for tracking deforestation trends often use maps of forest loss (often approximated by more directly measurable loss of tree cover) together with high-quality land use observations collected at sample locations to create statistical estimates of deforestation area. In our land use classification and change experiment, we created and compared a time series of annual land use classifications computed using feature vectors generated from the global CCDC dataset and several other statistical compositing methods (table 2).

This experiment used expert time series photo-interpretations covering 25 000 CONUS-based plots from the LCMAP & LCMS projects (Pengra *et al* 2020, Housman *et al* 2022, Xian *et al* 2022) as reference data. This dataset has both land cover and land use interpretations spanning 1986–2020, and this experiment used the dataset's six land use classes: Forest, Developed, Agriculture, Wetland, Rangeland, and Other. For each plot, we extracted annual data for six feature vectors using three statistics-based compositing methods: median, geometric median, and medoid, at annual and three-month seasonal timesteps from the same Landsat-based collections used to generate the global CCDC dataset. We also added a statistics-based feature vector approximating the percentiles method used by the GLAD project (Potapov *et al* 2020). This feature vector contains per-band quartile means, two half means, min, max, mean, and median, and the magnitude differences between the min and max, half means, and the first and fourth quartiles.

**Table 1.** CCDC dataset layers.

| Band name           | Per-pixel array Size | Description                                                           |
|---------------------|----------------------|-----------------------------------------------------------------------|
| tStart              | # Segments           | Start time of each segment                                            |
| tEnd                | # Segments           | End time of each segment                                              |
| tBreak              | # Segments           | Break time of each segment                                            |
| numObs              | # Segments           | Number of observations in each segment                                |
| changeProb          | # Segments           | Probability of change for each segment                                |
| inputband_coefs     | # Segments x 8       | Harmonic coefficients. 8 per segment.<br>Repeated for each input band |
| inputband_rmse      | # Segments           | RMSE of the fitted model per segment.<br>Repeated for each input band |
| inputband_magnitude | # Segments           | Magnitude of change<br>Repeated for each input band                   |

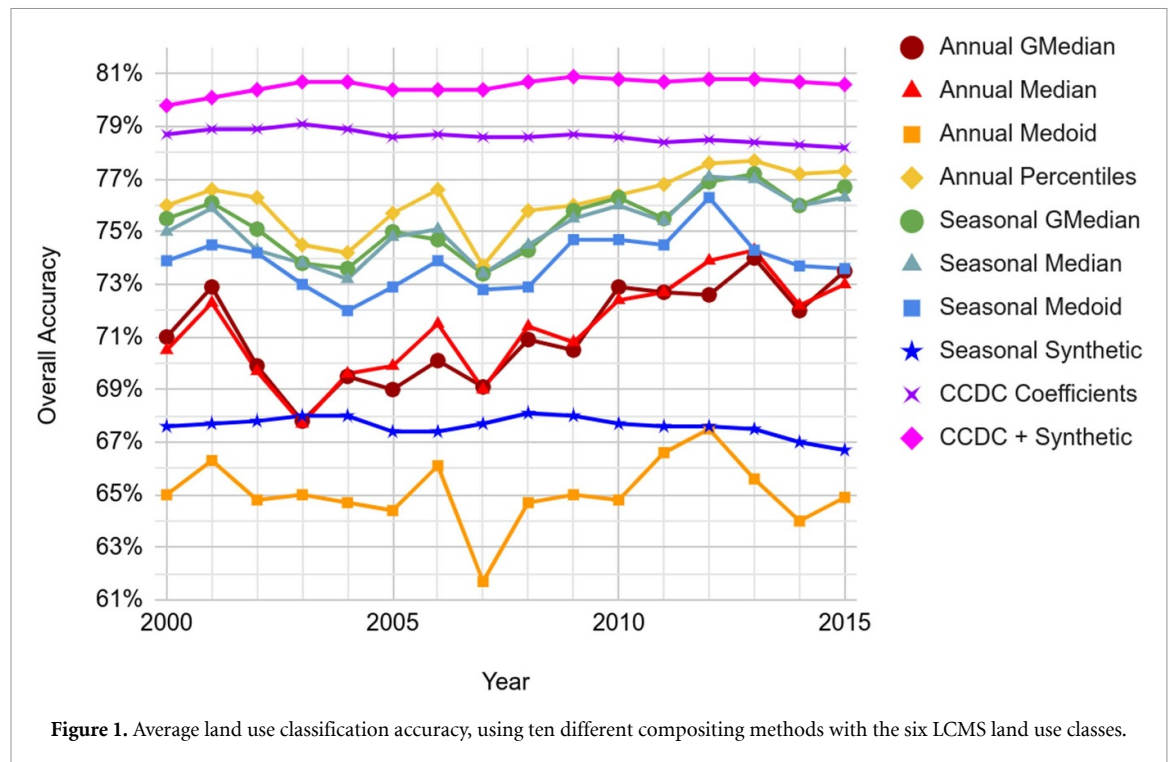
**Table 2.** Compositing methods used in this study. Annual composites include data between 1/1 and 12/31. Seasonal composites include data in three-month ranges.

| Compositing method | Description                                                                                                                     | # Covariates annual/seasonal | Reference                    |
|--------------------|---------------------------------------------------------------------------------------------------------------------------------|------------------------------|------------------------------|
| Median             | Per-band medians                                                                                                                | 7/21                         |                              |
| Geometric median   | Geometric (L1) median, the point minimizing the sum of the Euclidean distances to the inputs                                    | 7/21                         | Haldane (1948), Small (1990) |
| Medoid             | Whole observation closest to the per-band medians                                                                               | 7/21                         | Flood 2013                   |
| Percentiles        | Per-band min, max, mean, median, quartile means (Q1, Q2, Q3, Q4), half means (H1, H2) and differences: max-min, H2-H1 and Q4-Q1 | 91                           | Potapov <i>et al</i> 2020    |
| Seasonal synthetic | Synthetic images generated at 3 month intervals (Apr. Jul. Oct)                                                                 | 21                           | Zhu <i>et al</i> 2015        |
| CCDC coefficients  | Raw CCDC harmonic coefficients                                                                                                  | 56                           | Zhu and Woodcock 2014        |
| CCDC + Synthetic   | CCDC coefficients and synthetic images                                                                                          | 77                           |                              |

Additionally, we included feature vectors generated from the global CCDC dataset, influenced by the methodology of the USGS LCMAP project (Xian *et al* 2022). Since CCDC parameters remain the same across years until the algorithm finds a breakpoint, there is no chance of predicting an inter-segment land use class change if the only predictor variables are the segments' harmonic coefficients. While this leads to an expectation of reduced false positive errors for predicted land use transitions, this can potentially increase false negative error rates if breakpoints omitted by CCDC systematically prevent the identification of real transitions or slow, successional changes that might occur without discrete change in land use. To mitigate this effect, LCMAP uses a feature vector that combines an average annual CCDC-predicted synthetic image with the raw CCDC

harmonics to allow better classification of slower changes that might not trigger CCDC's break-finding function. With this in mind, we tested three feature vectors generated from the global CCDC dataset; one using the raw CCDC coefficients extracted at the center of the year (e.g. fractional year 2016.5), one using three synthetic images generated at three-month intervals over the North American growing season (e.g. 1 April 2016, 1 July 2016, 1 October 2016), and a third combining the coefficients and synthetic images.

A stratified random sampling of 1000 points from each of the 6 LCMAP classes was taken from the reference data to train a random forest classifier, which was then applied to the remaining data yearly from 2000 to 2015. This process was repeated 16 times, and the results were averaged. Additionally, to test



the stability of each predictor over time, we computed the average number of segments detected per plot and the percentage of plots that agreed with the reference data in all years. We also prepared a post-classification change estimate of the forest vs. non-forest land use classes between the years 2005 and 2015 and computed the estimated areas and  $\pm 95\%$  confidence intervals using the recommended best practices (Olofsson *et al* 2014)

### 2.3. Estimating above-ground biomass over time in Paraguay

Our second experiment demonstrates the potential for directly predicting a time series of above-ground biomass density (AGBD). In this experiment, we used field plot biomass estimates from 286 Paraguayan NFI plots collected 2014–2016 (Sato *et al* 2015) as reference data and predicted total AGBD estimates for the forested regions of Paraguay. We used the same ten compositing methods from the previous experiment; however, the seasonal composites were created for all four seasons instead of the three previously used due to minimal snow cover in Paraguayan winter. Composites were created for each year between 2000 to 2019 and a random forest regression was used to generate the estimates. The results were masked with an annual forest/non-forest mask (Hansen *et al* 2013) and summarized by year.

## 3. Results

### 3.1. Land use mapping and change

Our land use and change results show that the two feature vectors containing the CCDC coefficients produced consistently higher accuracy than the other feature vectors (figure 1). Moreover, the CCDC-based feature vectors produced substantially more stable classifications over time. The average numbers of segments (where one segment indicates no change over the whole period) in the CCDC-based input datasets (1.55, 1.43, and 2.25; table 3) were much closer to the observed average number of segments from the reference data (1.12) than any of the other options (the lowest of which was 3.59 from ‘Annual Percentiles’).

Adding synthetic images to the CCDC parameter feature vector increased average accuracy, but as anticipated, it also led to more false positive segments resulting in more segments detected (table 3).

In the forest change test, in which the six use classes were simplified to forest and non-forest, the CCDC-based feature vectors outperformed the other methods, with better overall, user’s, and producer’s accuracies as well as smaller population-level uncertainties of deforestation when combined with reference data using a post-stratified estimator (Olofsson *et al* 2013, 2014).

The annual compositing methods underperformed similarly in both the land use and forest

**Table 3.** Summarized results from the land use classification and forest change experiments on ten types of compositing methods. Columns 2–5 are for predictions using all six land use classes, and columns 6–10 are for a post-classification forest change estimation between 2005 and 2015. For the land use classification, the overall accuracy (OA) averaged over all classifiers, the average number of detected breaks, and the percentage of plots in which the predictions match the reference data in all years are shown. For the forest change estimate, the overall accuracy and the user's (UA) and producer's accuracies (PA) for the forest loss class are shown, along with the mapped area of loss, the error-adjusted estimate of loss, and the  $\pm 95\%$  confidence interval as a percentage of the error-adjusted area of loss. The best-performing method is highlighted in green in each column.

| Compositing method | Annual 6-class land use classifications |             |               | 2005–2015 forest change |         |         |                        |                      |               |
|--------------------|-----------------------------------------|-------------|---------------|-------------------------|---------|---------|------------------------|----------------------|---------------|
|                    | OA                                      | Avg. # seg. | Exact matches | OA                      | Loss UA | Loss PA | Mapped loss area (Mha) | Est. loss area (Mha) | $\pm 95\%$ CI |
| Annual GMedian     | 71.2%                                   | 4.55        | 33.2%         | 85.7%                   | 6.0%    | 31.3%   | 15.42                  | 2.95                 | 11.2%         |
| Annual median      | 71.3%                                   | 4.44        | 33.4%         | 85.3%                   | 5.6%    | 33.8%   | 18.00                  | 3.00                 | 10.5%         |
| Annual medoid      | 65.1%                                   | 6.23        | 16.0%         | 79.2%                   | 2.5%    | 46.3%   | 53.71                  | 2.86                 | 11.2%         |
| Percentiles        | 76.2%                                   | 3.59        | 46.9%         | 89.8%                   | 10.1%   | 30.0%   | 9.04                   | 2.90                 | 11.3%         |
| Seas. GMedian      | 75.4%                                   | 3.81        | 44.2%         | 86.9%                   | 8.1%    | 33.8%   | 11.61                  | 2.77                 | 10.7%         |
| Seas. median       | 75.2%                                   | 3.78        | 44.2%         | 87.0%                   | 8.1%    | 35.0%   | 11.43                  | 2.77                 | 10.9%         |
| Seas. medoid       | 73.9%                                   | 4.42        | 34.4%         | 87.1%                   | 8.3%    | 37.5%   | 12.01                  | 2.86                 | 10.7%         |
| Seas. synth        | 67.6%                                   | 2.25        | 59.0%         | 86.2%                   | 5.9%    | 40.0%   | 19.70                  | 2.93                 | 10.6%         |
| CCDC coeff.        | 78.6%                                   | 1.43        | 79.7%         | 87.8%                   | 10.5%   | 28.8%   | 8.24                   | 3.01                 | 12.2%         |
| CCDC + Synth       | 80.6%                                   | 1.55        | 76.9%         | 92.3%                   | 15.2%   | 53.8%   | 10.83                  | 3.01                 | 9.9%          |
| Reference          |                                         | 1.03        |               |                         |         |         |                        |                      |               |

**Table 4.** Normalized root mean square error of prediction vs. the NFI field-collected reference data for ten compositing methods predicting above-ground biomass in Paraguayan forests. These were computed using 20-fold cross-validation with a 5% holdout and no replacement.

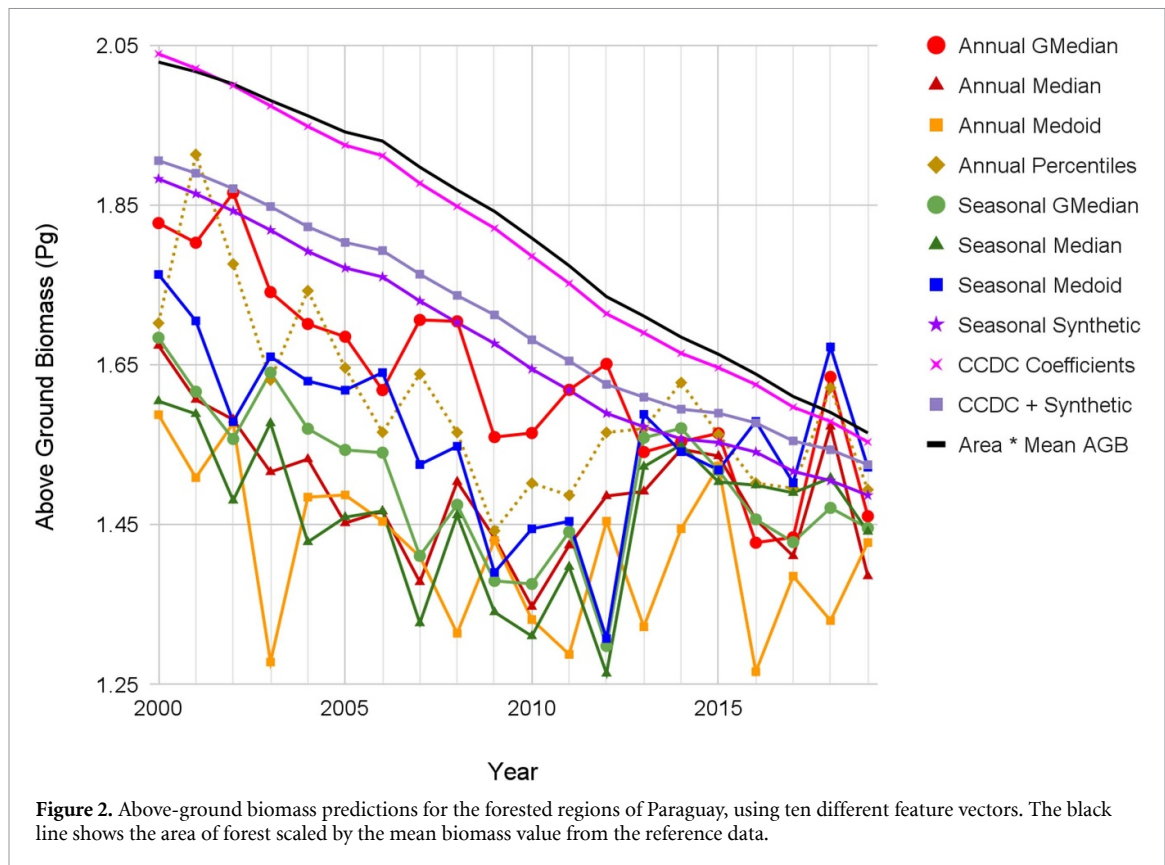
| Compositing method | NRMSE |
|--------------------|-------|
| Annual GMedian     | 0.510 |
| Annual median      | 0.475 |
| Annual medoid      | 0.472 |
| Annual percentiles | 0.448 |
| Seasonal GMedian   | 0.470 |
| Seasonal median    | 0.460 |
| Seasonal medoid    | 0.453 |
| Seasonal synthetic | 0.443 |
| CCDC coefficients  | 0.454 |
| CCDC + Synthetic   | 0.435 |

change tests compared to the seasonal compositing methods, with the medoid performing poorly in both cases. While not performing as well as the other seasonal methods in terms of overall accuracy, the seasonal synthetic method produced a much smoother curve over time, reflected in the significantly better average number of breaks, exact matches, and better performance in the change test.

### 3.2. Above-ground biomass over time in Paraguay

The CCDC-based feature vectors in this experiment produced the lowest normalized RMSE error

compared to the reference data (table 4). They also produced a smoothly varying annual biomass that closely approximates the nearly linear decline in forest area computed from the Hansen Global Tree Cover dataset (figure 2), while the models built from the annual and seasonal composites show year-to-year variations that are grossly unrealistic. All non-CCDC models exhibited noise over time in their prediction of AGBD, including an impossible one-year 10% increase in forest biomass in 2018 in six out of seven of the models.



#### 4. Discussion and conclusions

For countries reporting under the REDD+ framework, less accurate maps and higher uncertainties can result in less money for results-based payments. Our experiments show that using a CCDC-based feature vector produces better results with higher accuracy and lower uncertainty over the other methods but with substantial additional gains in stability and accuracy over time.

Operational change detection methods are typically more complex and accurate than the simple post-classification delta approach we used to test feature vector effects. In some cases, many change detection algorithms are used in parallel, with a secondary ensemble model used to weight the inputs in an optimal way for independent training data (Cohen *et al* 2020). In these ensemble models and the post-stratified estimation methods (Olofsson *et al* 2013, 2014), map accuracy may not be directly proportional to output accuracy or precision (respectively). However, accuracies reported for our simple tests suggest that time series processing using CCDC improves land use and biomass signal strength as well as the stability of resulting classifications over time.

While the differences in the computed uncertainties between the methods we tested are small, this is primarily a consequence of reusing the same reference data to compute uncertainty for each method. Of the 22 066 reference data points used in the analysis, only 80 (0.3%) represent forest loss. This small sample and

the use of a post-stratified estimator resulted in only a few reference points intersecting the forest change class in each map. While the CCDC + Synthetic feature vector performed best, all the methods produced similar and relatively high (~11%) uncertainty estimates for forest loss.

We expect the differences in uncertainty between maps to be more pronounced and weighted in favor of the more accurate algorithms if areas are instead estimated using a stratified random approach targeting the change areas using an optimal sample allocation approach. Quoting from Stehman and Foody (2019), ‘As a general rule, the more accurate the map, the greater the reduction in the standard errors of the area estimates compared to a strategy that does not use strata determined from the map.’ This method is the more common approach for countries seeking to report their carbon emissions from forest loss, but would require generating an independent set of interpretations per map, an analysis outside of the scope of this paper. Uncertainty in the adjusted areas could be further minimized by carefully constructing the strata to reduce the impact of errors of omission of change (Olofsson *et al* 2020).

As an additional observation, we expected the medoid and geometric median compositing methods to outperform the regular median in these experiments, as they both end up choosing a single observation instead of potentially mixing spectral bands between observations. But the medoid performed poorly in all cases, and the geometric median was

roughly tied with the median. Both the medoid and geometric median involve computing a single distance metric over all bands, and Housman *et al* (2022) suggests the medoid results might be improved by omitting the blue band from the distance calculations due to residual atmospheric contamination.

#### 4.1. Landsat as a time machine

While the Sentinel-2 platform may offer a higher spatial resolution and active sensors such as GEDI may offer a more direct observation of forest biomass, the deep time series of Landsat is appealing as a form of time travel; it allows practitioners to estimate physical properties for periods when these other instruments did not exist (e.g. Powell *et al* 2010, Potapov *et al* 2019). Our results show that in these ‘time-machine’ applications, a method like CCDC that exploits the temporal context is likely to outperform methods applied year-by-year, and the temporal coherence (accuracy in sequence in addition to single-date classification) directly affects the interpretability of the data in a REDD+ reporting context. Unrealistic AGBD trajectories may compromise the credibility of a monitoring system and may mask progress or regression in the area of forest degradation. Likewise, ‘bouncing’ of land use from one class to another because of ephemeral image artifacts, as measured here through the number of land use segments, can inflate the number of transitions, with direct implications in the precision with which processes like deforestation may be estimated.

#### 4.2. Alternative methods

CCDC is just one example of a preprocessing algorithm that uses the temporal context to help produce a smoother time series. Another prominent example is LandTrendr (Kennedy *et al* 2010), which iteratively fits linear segments to a time series to find breakpoints and smooth the inputs between breakpoints. LandTrendr has also been made available in Google Earth Engine (Kennedy *et al* 2018), and several studies compare CCDC and LandTrendr results (Cohen *et al* 2018, Healey *et al* 2018, Pasquarella *et al* 2022), so it was not included here. In addition to its use in preprocessing, LandTrendr has also been used as a post-processing method to help produce smooth results (Powell *et al* 2010).

Other notable examples of time-series smoothing through post-processing include the C2C algorithm (Hermosilla *et al* 2022), which uses a preprocessing scoring function to rank and select optimum pixel observations for inclusion in annual composites representing mid-summer conditions (White *et al* 2014), but then uses a post-classification process involving a Hidden Markov Model to eliminate unlikely class transitions. Similarly, the BULC algorithm (Cardille and Fortin 2016) uses Bayesian statistics to eliminate improbable transitions in time-series classifications. Future work is needed to understand the interaction

of these post-classification smoothing methods and the kind of pre-classification trend-fitting achieved with CCDC.

#### 4.3. Agroforestry versus deforestation

The ability to distinguish between natural forest removal and agroforestry activities is important for accurately assessing environmental impact, ensuring policy effectiveness, promoting transparency, addressing climate change implications, and holding governments accountable for their NDCs. While this study does not directly address the discrimination of these activities, other research has shown that classifications based on CCDC coefficients are able to differentiate between the them (Zhou *et al* 2023). That research also points towards improvements that could be made to future versions of the dataset to further facilitate the separation of plantation versus natural forest through the inclusion of additional band transformations.

#### 4.4. The global dataset

The computation to generate the CCDC coefficients dataset was substantial, on the order of 10 million CPU hours and touching more than 4.4 million Landsat images and 1.6 petabytes. Due to the dynamic nature of CCDC’s thresholding, it is currently only feasible to update the pre-generated dataset with data after 2019 by re-running the entire computation from scratch. We are investigating methodologies that might allow annual or sub-annual updates to CCDC results without incurring the costs of a full recomputation; however, due to the discontinuation of the Collection-1 processing by the USGS, it will be impossible to update this version of the dataset beyond the end of 2021.

Work has already begun developing a second version of this dataset based on Landsat Collection-2 (Friedl *et al* 2022), and that product is expected to be updated periodically. However, translating historical reference data, such as the data used in our land use experiment, between the two collections is problematic due to the geolocation refinements applied to Collection-2. That leaves a gap that this Collection-1-based dataset can help fill for historic FRELS, even once a Collection-2-based version exists. Expansion using the Harmonized Landsat Sentinel dataset (Claverie *et al* 2018) may also improve change mapping accuracy and robustness (Zhou *et al* 2019).

The benefits we demonstrate here in the context of forest carbon accounting—stabilization of reflectance signals in a way that creates acuity in classifying surface elements and identifying change—are likely transferable to other land cover, land use, and change detection tasks. CCDC has been effectively used to map habitat over time (Witt *et al* 2022), characterize urban growth (Aljaddani *et al* 2022), and support deep learning of agricultural patterns (Zhang *et al* 2022). Public release of the dataset we describe should

reduce redundant (and expensive) time series processing, supporting forest carbon monitoring in a globally consistent way and stimulating activity in new sectors.

## Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: [https://developers.google.com/earth-engine/datasets/catalog/GOOGLE\\_GLOBAL\\_CCDC\\_V1](https://developers.google.com/earth-engine/datasets/catalog/GOOGLE_GLOBAL_CCDC_V1).

## ORCID iDs

Noel Gorelick  <https://orcid.org/0000-0002-5548-2436>

Paulo Arévalo  <https://orcid.org/0000-0001-7566-7561>

Eric L Bullock  <https://orcid.org/0000-0003-3279-6771>

Sean P Healey  <https://orcid.org/0000-0003-3498-4266>

## References

- Aljaddani A H, Song X P and Zhu Z 2022 Characterizing the patterns and trends of urban growth in Saudi Arabia's 13 capital cities using a Landsat time series *Remote Sens.* **14** 2382
- Arévalo P, Olofsson P and Woodcock C E 2020 Continuous monitoring of land change activities and post-disturbance dynamics from Landsat time series: a test methodology for REDD+ reporting *Remote Sens. Environ.* **238** 111051
- Brooks E B, Wynne R H, Thomas V A, Blinn C E and Coulston J W 2013 On-the-fly massively multitemporal change detection using statistical quality control charts and Landsat data *IEEE Trans. Geosci. Remote Sens.* **52** 3316–32
- Bullock E L, Woodcock C E and Olofsson P 2020 Monitoring tropical forest degradation using spectral unmixing and Landsat time series analysis *Remote Sens. Environ.* **238** 110968
- Cardille J A and Fortin J A 2016 Bayesian updating of land-cover estimates in a data-rich environment *Remote Sens. Environ.* **186** 234–49
- Chile 2021 Methodological protocol for the elaboration of cartographies of uses and changes of uses of the land (available at: <https://simef.minagri.gob.cl/biblioteca/digital/handle/20.500.12978/19497>)
- Claverie M, Ju J, Masek J G, Dungan J L, Vermote E F, Roger J C, Skakun S V and Justice C 2018 The harmonized Landsat and Sentinel-2 surface reflectance data set *Remote Sens. Environ.* **219** 145–61
- Cohen W B, Healey S P, Yang Z, Zhu Z and Gorelick N 2020 Diversity of algorithm and spectral band inputs improves Landsat monitoring of forest disturbance *Remote Sens.* **12** 1673
- Cohen W B, Yang Z, Healey S P, Kennedy R E and Gorelick N 2018 A LandTrendr multispectral ensemble for forest disturbance detection *Remote Sens. Environ.* **205** 131–40
- Czaplewski R L and Patterson P L 2003 Classification accuracy for stratification with remotely sensed data *For. Sci.* **49** 402–8
- Deng C and Zhu Z 2020 Continuous subpixel monitoring of urban impervious surface using Landsat time series *Remote Sens. Environ.* **238** 110929
- Dominica 2022 Forest reference level 2018–2025 (available at: [https://redd.unfccc.int/files/2022\\_dom\\_frel\\_report\\_jan\\_5.pdf](https://redd.unfccc.int/files/2022_dom_frel_report_jan_5.pdf))
- Drusch M *et al* 2012 Sentinel-2: ESA's optical high-resolution mission for GMES operational services *Remote Sens. Environ.* **120** 25–36
- Dwyer J L, Roy D P, Sauer B, Jenkerson C B, Zhang H K and Lymburner L 2018 Analysis ready data: enabling analysis of the Landsat archive *Remote Sens.* **10** 1363
- Ecuador 2020 Nivel de Referencia de Emisiones Forestales por Deforestación del Ecuador Período 2001–2014 (available at: [https://redd.unfccc.int/files/06.01.2020\\_nivel\\_de\\_referencia\\_de\\_emisiones\\_forestales\\_de\\_ecuador\\_2001-2014.pdf](https://redd.unfccc.int/files/06.01.2020_nivel_de_referencia_de_emisiones_forestales_de_ecuador_2001-2014.pdf))
- Espejo A, Federici S, Green C, Amuchastegui N, d'Annunzio R, Balzter H and Wulder M 2020 *Integration of Remote-sensing and Ground-based Observations for Estimation of Emissions and Removals of Greenhouse Gases in Forests: Methods and Guidance from the Global Forest Observations Initiative Edition 3.0* (UN Food and Agriculture Organization) p 300
- Flood N 2013 Seasonal composite Landsat TM/ETM+ images using the medoid (a multi-dimensional median) *Remote Sens.* **5** 6481–500
- Friedl M A *et al* 2022 Medium spatial resolution mapping of global land cover and land cover change across multiple decades from Landsat *Front. Remote Sens.* **3** 894571
- GFOI 2020 Integration of remote-sensing and ground-based observations for estimation of emissions and removals of greenhouse gases in forests, version 3 (available at: [www.reddcompass.org/mgd/resources/GFOI-MGD-3.1-en.pdf](http://www.reddcompass.org/mgd/resources/GFOI-MGD-3.1-en.pdf))
- Gómez C, White J C and Wulder M A 2016 Optical remotely sensed time series data for land cover classification: a review *ISPRS J. Photogramm. Remote Sens.* **116** 55–72
- Gorelick N, Hancher M, Dixon M, Ilyushchenko S, Thau D and Moore R 2017 Google earth engine: planetary-scale geospatial analysis for everyone *Remote Sens. Environ.* **202** 18–27
- Haldane J B S 1948 Note on the median of a multivariate distribution *Biometrika* **35** 414–7
- Hansen M C *et al* 2013 High-resolution global maps of 21st-century forest cover change *Science* **342** 850–3
- Hansen M C *et al* 2014 Monitoring continuous United States (CONUS) land cover change with web-enabled Landsat data (WELD) *Remote Sens. Environ.* **140** 466–84
- Healey S P *et al* 2018 Mapping forest change using stacked generalization: an ensemble approach *Remote Sens. Environ.* **204** 717–28
- Hermosilla T, Wulder M A, White J C and Coops N C 2022 Land cover classification in an era of big and open data: optimizing localized implementation and training data selection to improve mapping outcomes *Remote Sens. Environ.* **268** 112780
- Houghton R A, House J I, Pongratz J, Van Der Werf G R, Defries R S, Hansen M C, Le Quéré C and Ramankutty N 2012 Carbon emissions from land use and land-cover change *Biogeosciences* **9** 5125–42
- Housman I W, Campbell L S, Goetz W E, Finco M V, Pugh N and Megown K 2022 US forest service landscape change monitoring system methods version 2020.5 *GTAC-10252-RPT1* (U.S. Department of Agriculture, Forest Service, Geospatial Technology and Applications Center) p 25
- Indonesia 2022 National forest reference level for deforestation, forest degradation and enhancement of forest carbon stock (available at: [https://redd.unfccc.int/files/2nd\\_frl\\_indonesia\\_final\\_submit.pdf](https://redd.unfccc.int/files/2nd_frl_indonesia_final_submit.pdf))
- Jin S, Yang L, Danielson P, Homer C, Fry J and Xian G 2013 A comprehensive change detection method for updating the National Land Cover Database to circa 2011 *Remote Sens. Environ.* **132** 159–75
- Kennedy R E, Yang Z and 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** 2897–910

- Kennedy R E, Yang Z, Gorelick N, Braaten J, Cavalcante L, Cohen W B and Healey S 2018 Implementation of the LandTrendr algorithm on google earth engine *Remote Sens.* **10** 691
- Nesha M K, Herold M, De Sy V, Duchelle A E, Martius C, Branthomme A, Garzuglia M, Jonsson O and Pekkarinen A 2021 An assessment of data sources, data quality and changes in national forest monitoring capacities in the global forest resources assessment 2005–2020 *Environ. Res. Lett.* **16** 054029
- Olofsson P, Arévalo P, Espejo A B, Green C, Lindquist E, McRoberts R E and Sanz M J 2020 Mitigating the effects of omission errors on area and area change estimates *Remote Sens. Environ.* **236** 111492
- Olofsson P, Foody G M, Herold M, Stehman S V, Woodcock C E and Wulder M A 2014 Good practices for estimating area and assessing accuracy of land change *Remote Sens. Environ.* **148** 42–57
- Olofsson P, Foody G M, Stehman S V and 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** 122–31
- Paraguay 2015 Nivel de Referencia de las Emisiones Forestales por Deforestación en la República del Paraguay para pago por resultados de REDD+ bajo la CMNUCC (available at: [https://redd.unfccc.int/files/2016\\_submission\\_frel\\_paraguay\\_es.pdf](https://redd.unfccc.int/files/2016_submission_frel_paraguay_es.pdf)) (Accessed 1 October 2022)
- Pasquarella V J, Arévalo P, Bratley K H, Bullock E L, Gorelick N, Yang Z and Kennedy R E 2022 Demystifying LandTrendr and CCDC temporal segmentation *Int. J. Appl. Earth Obs. Geoinf.* **110** 102806
- Pengra B W *et al* 2020 LCMAP reference data product 1984–2018 land cover land use and change process attributes US Geological Survey Data Release (<https://doi.org/10.5066/P9ZW0XJ7>)
- Peru 2016 Peru's submission of a forest reference emission level (FREL) for reducing emissions from deforestation in the Peruvian Amazon (available at: [https://redd.unfccc.int/files/frel\\_submission\\_peru\\_modified.pdf](https://redd.unfccc.int/files/frel_submission_peru_modified.pdf))
- Potapov P, Hansen M C, Kommareddy I, Kommareddy A, Turubanova S, Pickens A, Adusei B, Tyukavina A and Ying Q 2020 Landsat analysis ready data for global land cover and land cover change mapping *Remote Sens.* **12** 426
- Potapov P, Tyukavina A, Turubanova S, Talero Y, Hernandez-Serna A, Hansen M C and Nguyen Q H 2019 Annual continuous fields of woody vegetation structure in the Lower Mekong region from 2000–2017 Landsat time-series *Remote Sens. Environ.* **232** 111278
- Powell S L, Cohen W B, Healey S P, Kennedy R E, Moisen G G, Pierce K B and Ohmann J L 2010 Quantification of live aboveground forest biomass dynamics with Landsat time-series and field inventory data: a comparison of empirical modeling approaches *Remote Sens. Environ.* **114** 1053–68
- Roberts D, Mueller N and McIntyre A 2017 High-dimensional pixel composites from earth observation time series *IEEE Trans. Geosci. Remote Sens.* **55** 6254–64
- Sato T *et al* 2015 Development of allometric equations for tree biomass in forest ecosystems in Paraguay *Jpn. Agric. Res. Q.* **49** 281–91
- Small C G 1990 A survey of multidimensional medians *Int. Stat. Rev.* **58** 263–77
- Stehman S V and Foody G M 2019 Key issues in rigorous accuracy assessment of land cover products *Remote Sens. Environ.* **231** 111199
- Tang X, Bullock E L, Olofsson P, Estel S and Woodcock C E 2019 Near real-time monitoring of tropical forest disturbance: new algorithms and assessment framework *Remote Sens. Environ.* **224** 202–18
- Thailand 2022 Report on the technical assessment of the proposed forest reference emission level and forest reference level of Thailand submitted in 2021 (available at: <https://unfccc.int/documents/470451>)
- Van Doninck J and Tuomisto H 2017 Influence of compositing criterion and data availability on pixel-based Landsat TM/ETM+ image compositing over Amazonian forests *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **10** 857–67
- Wang J A, Sulla-Menashe D, Woodcock C E, Sonnentag O, Keeling R F and Friedl M A 2020 Extensive land cover change across Arctic–Boreal Northwestern North America from disturbance and climate forcing *Glob. Change Biol.* **26** 807–22
- White J C *et al* 2014 Pixel-based image compositing for large-area dense time series applications and science *Can. J. Remote Sens.* **40** 192–212
- Wilson B T, Knight J F and McRoberts R E 2018 Harmonic regression of Landsat time series for modeling attributes from national forest inventory data *ISPRS J. Photogramm. Remote Sens.* **137** 29–46
- Witt C *et al* 2022 Linking robust spatiotemporal datasets to assess and monitor habitat attributes of a threatened species *PLoS One* **17** e0265175
- Wulder M A *et al* 2022 Fifty years of Landsat science and impacts *Remote Sens. Environ.* **280** 113195
- Xian G Z, Smith K, Wellington D, Horton J, Zhou Q, Li C, Auch R, Brown J F, Zhu Z and Reker R R 2022 Implementation of the CCDC algorithm to produce the LCMAP collection 1.0 annual land surface change product *Earth Syst. Sci. Data* **14** 143–62
- Zhang H, He B and Xing J 2022 Mapping paddy rice in complex landscapes with Landsat time series data and superpixel-based deep learning method *Remote Sens.* **14** 3721
- Zhou Q, Rover J, Brown J, Worstell B, Howard D, Wu Z, Gallant A L, Rundquist B and Burke M 2019 Monitoring landscape dynamics in central us grasslands with harmonized Landsat-8 and Sentinel-2 time series data *Remote Sens.* **11** 328
- Zhou Q, Wang L, Tang F, Zhao S, Huang N and Zheng K 2023 Mapping spatial and temporal distribution information of plantations in Guangxi from 2000 to 2020 *Front. Ecol. Evol.* **11** 1201161
- Zhu Z and Woodcock C E 2014 Continuous change detection and classification of land cover using all available Landsat data *Remote Sens. Environ.* **144** 152–71
- Zhu Z, Woodcock C E, Holden C and Yang Z 2015 Generating synthetic Landsat images based on all available Landsat data: predicting Landsat surface reflectance at any given time *Remote Sens. Environ.* **162** 67–83