LSDS-2320
Version 1.0

Department of the Interior
U.S. Geological Survey

Land Change Monitoring, Assessment, and Projection (LCMAP) Collection 1.2 Continuous Change Detection and Classification (CCDC) Algorithm Description Document (ADD)

Release 1.0

Version 1.0

November 2021

USGS
science for a changing world

Land Change Monitoring, Assessment, and Projection (LCMAP) Collection 1.2 Continuous Change Detection and Classification (CCDC) Algorithm Description Document (ADD)

Release 1.0

November 2021

Approved By:

JESSLYN BROWN Digitally signed by JESSLYN BROWN Date: 2021.11.16 11:05:39 -06'00'

_____

Jesslyn Brown Date
LCMAP CCB Chair
U.S. Geological Survey

EROS
Sioux Falls, South Dakota

Executive Summary

This Algorithm Description Document (ADD) defines the Continuous Change Detection and Classification (CCDC) algorithm used for the generation of Land Change Monitoring, Assessment, and Projection (LCMAP) Collection 1.2 Science Products. The CCDC algorithm is implemented as part of the LCMAP initiative at the USGS Earth Resources Observation and Science (EROS) Center. This document provides a high-level overview of the LCMAP implementation of CCDC and technical details about how the subsequent LCMAP Collection 1.2 Science Products are derived.

LCMAP is a USGS science initiative that harnesses the remotely sensed Landsat data record to provide state-of-the-art land surface change information needed by scientists, resource managers, and decision-makers. LCMAP uses a modernized, integrated approach to map, monitor, synthesize, and understand the complexities of land use, cover, and condition change.

Basic foundational elements of the LCMAP project include:

- Landsat Collection 1 U.S. Analysis Ready Data (ARD)
- Land surface change and land cover data
- Independent reference data for validation and area estimation
- Scenario-driven projections of future land use and land cover extents and patterns
- Assessments focused on land change processes, characteristics, and consequences

This document is under LCMAP Configuration Control Board (CCB) control. Please submit changes to this document, as well as supportive material justifying the proposed changes, via Change Request (CR) to the Process and Change Management Tool.

Document History

Document Number	Document Version	Publication Date	Change Number
LSDS-2320	Version 1.0	November 2021	CR 20831

Contents

Executive Summary ..... **iii**
Document History ..... **iv**
Contents ..... **v**
List of Figures ..... **vi**
List of Tables ..... **vii**
Section 1 Introduction **1**
1.1 Background 1
1.2 Purpose 2
1.3 LCMAP Project Documentation Suite 3
1.4 Document Organization 3
Section 2 Continuous Change Detection **4**
2.1 Description 4
2.2 Dependencies 4
2.3 CCD Inputs 4
2.4 CCD Output 5
2.5 CCD Key Algorithm Concepts 5
2.5.1 Data Filtering 5
2.5.2 Harmonic Modeling 6
2.5.3 Regression 7
2.5.4 Data Variability 7
2.5.5 Change Detection Thresholds 8
2.5.6 Forward Processing 9
2.6 CCD Step-by-Step Walk-through 10
2.6.1 Data Setup and Procedure Selection/Run 11
2.6.2 Individual Procedure Operations 11
2.7 Known Issues 17
Section 3 Classification **19**
3.1 Description 19
3.2 Dependencies 19
3.3 Classification Inputs 20
3.4 Classification Outputs 20
3.5 Classification Key Concepts 20
3.5.1 Algorithm 20
3.5.2 Training Data and Sample Size 21
3.5.3 Prediction 22
3.5.4 Forward Processing 23
3.6 Classification Step-by-Step Walk-through 24
3.6.1 Model Training 24
3.6.2 Prediction 26
3.7 Known Issues 27
Section 4 Product Definitions **28**
4.1 Description 28

4.2 Definitions 28
4.2.1 Time of Spectral Change (SCTIME) 29
4.2.2 Change Magnitude (SCMAG) 30
4.2.3 Spectral Stability Period (SCSTAB) 31
4.2.4 Time Since Last Change (SCLAST) 32
4.2.5 Spectral Model Quality (SCMQA) 33
4.2.6 Land Cover (LCPRI and LCSEC) and Land Cover Confidence (LCPCONF and LCSCONF) 34
4.2.7 Annual Land Cover Change (LCACHG) 40
Appendix A Acronyms 42
Appendix B Default CCD Parameters 44
References 46

List of Figures

Figure 1-1. Time series for pixel in SWIR-1 band from U.S. Landsat ARD Surface Reflectance 1
Figure 1-2. Stable harmonic periods that CCD found in time series 2
Figure 1-3. CCD harmonic segments without original observations from which they were derived 2
Figure 1-4. CCD harmonic segments combined with land cover information, where green signifies periods of Tree Cover and tan represents Grass/Shrub 2
Figure 1-5. Annual July 1st dates overlaid with classified CCD harmonic segments 2
Figure 2-1. Overview of CCD Processing 10
Figure 3-1. Moving from left to right, tile h03v10 is cross-walked from original NLCD 2001 (2011 Edition) to LCMAP level 1 classes, then eroded by one pixel 23
Figure 3-2. Cross-walked and eroded NLCD data (left) is filtered to include only pixels with CCD segments that encompass target training date (2001-07-01) 23
Figure 3-3. Information from 8 surrounding tiles are used to help inform classification modeling 26
Figure 3-4. Predictions for First Segment in Example Time Series 27
Figure 4-1. Legend for product definition examples, where green indicates time classified as Tree Cover, tan indicates time classified as Grass/Shrub, gray dashed lines are annual July 1st dates, and magenta dashed lines are spectral breaks 29
Figure 4-2. The DOY a spectral change caused a “break” in a CCDC time series model. When a spectral change occurs, causing a divergence from model predictions, a new time series model begins. 29
Figure 4-3. Example Python Function for Time of Spectral Change 30
Figure 4-4. Change Magnitude when Spectral Break Occurs 30
Figure 4-5. Example Python Function for Change Magnitude 31
Figure 4-6. Spectral Stability Period during a time series model starting on July 1st start date for product year 31

Figure 4-7. Spectral Stability Period between CCDC time series models, beginning on the July 1st start date for product year to the most recent harmonic segment characterized 32
Figure 4-8. Example Python Function for Spectral Stability 32
Figure 4-9. Time Since Last Change measured from July 1st start date to DOY of last time series model break 33
Figure 4-10. Example Python Function for Time Since Last Change 33
Figure 4-11. Spectral Model Quality providing information regarding type of time series model available for each annual July 1st 34
Figure 4-12. Example Python Function for Determining Spectral Model Quality 34
Figure 4-13. Primary Land Cover Values for Example Time Series 35
Figure 4-14. Secondary Land Cover Values for Example Time Series 35
Figure 4-15. Primary Land Cover Confidence Values for Example Time Series 35
Figure 4-16. Secondary Land Cover Confidence Values for Example Time Series 35
Figure 4-17. Example of how land cover is determined from initial classifier for first time series model 36
Figure 4-18. Example Predictions for Second Time Series Model in Example Pixel 38
Figure 4-19. Land cover classes are temporally interpolated between time series models, and extrapolated at ends of time series 39
Figure 4-20. Annual Land Cover Change Values for Example Time Series 41

List of Tables

Table 1-1. LCMAP Collection 1.2 Documentation Suite 3
Table 3-1. LCMAP Level 1 Land Cover Classes 19
Table 3-2. NLCD 2001 (2011 Edition) to LCMAP Land Cover Cross-walk 23
Table 4-1. Description of Pixel Values for Spectral Model Quality 34
Table 4-2. Description of Pixel Values for Primary and Secondary Land Cover Confidence 40
Table 4-3. Description of Pixel Values for Land Cover Change 41

Table B-1. Default CCD Parameters 45

Section 1 Introduction

1.1 Background

The Continuous Change Detection and Classification (CCDC, Zhu & Woodcock, 2014a) algorithm was developed at the Center for Remote Sensing (Department of Geography and Environment) at Boston University. The Land Change Monitoring, Assessment, and Projection (LCMAP) project selected CCDC to demonstrate the capabilities for time-series data analysis across the historical Landsat archive.

The CCDC algorithm uses a robust methodology to identify when and how the land surface changes through time. It employs every observation in a time series of Landsat Collection 1 U.S. Analysis Ready Data (ARD) to determine whether change has occurred at any given point in the observation record. The algorithm further classifies the pixel to indicate what land cover type(s) were observed before and after a detected change on the land surface.

The original implementation of CCDC was written in the MATLAB (The MathWorks, Inc.) programming language. The Continuous Change Detection (CCD) portion has since been translated into an open source library as Python code. The full implementation joins the CCD Python library with the classification methodology and combines them with data delivery/processing services made available through the U.S. Geological Survey (USGS) LCMAP project.

The algorithm uses a pixel time series (Figure 1-1) to characterize the stable periods with harmonic regression fits to each input band (Figure 1-2), forming temporal model segments. These harmonic characterizations can then be used in place of the original observations (Figure 1-3) for approximating the landscape at any given time during the time series. The temporal segments are subsequently used to classify the land cover (Figure 1-4) to help better define the overall pixel history (i.e., how the pixel changes through time). Finally, characteristic properties of the model segments are extracted as LCMAP Science Products, using an annual July 1st date to capture a snapshot of the dynamic state of the model in the middle of a calendar year, as shown in Figure 1-5.

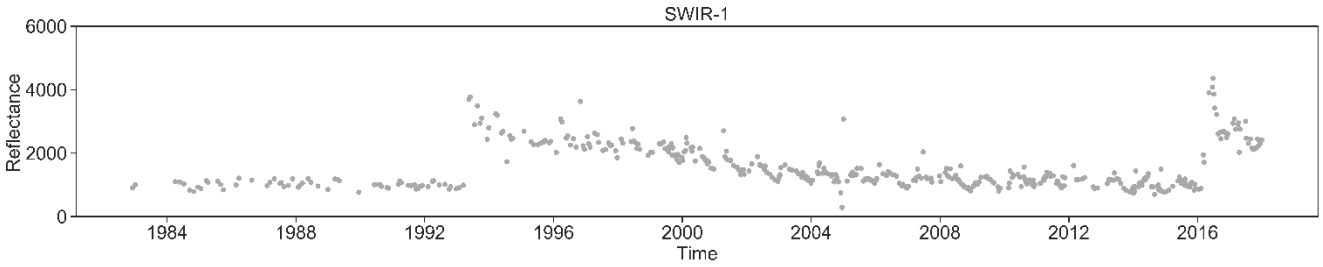


Figure 1-1. Time series for pixel in SWIR-1 band from U.S. Landsat ARD Surface Reflectance

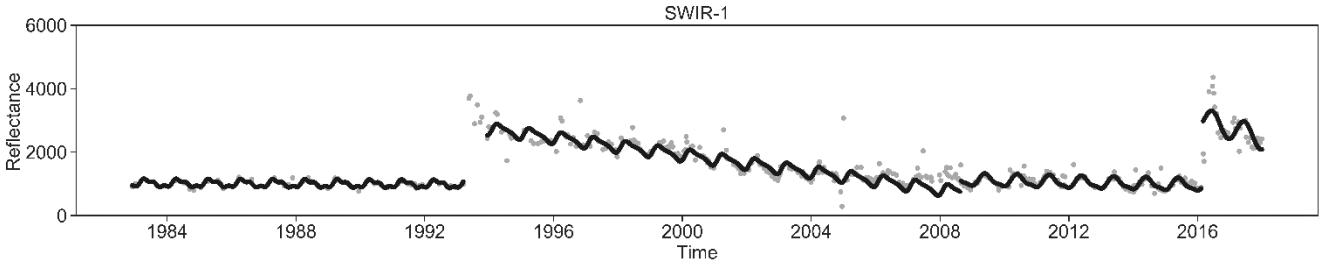


Figure 1-2. Stable harmonic periods that CCD found in time series

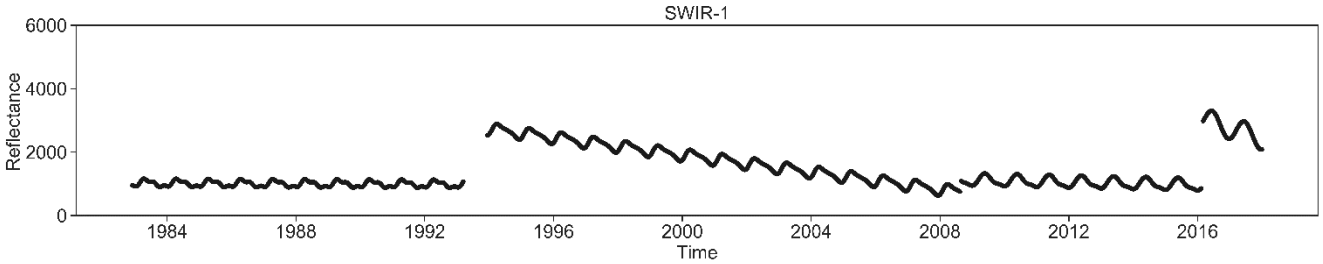


Figure 1-3. CCD harmonic segments without original observations from which they were derived

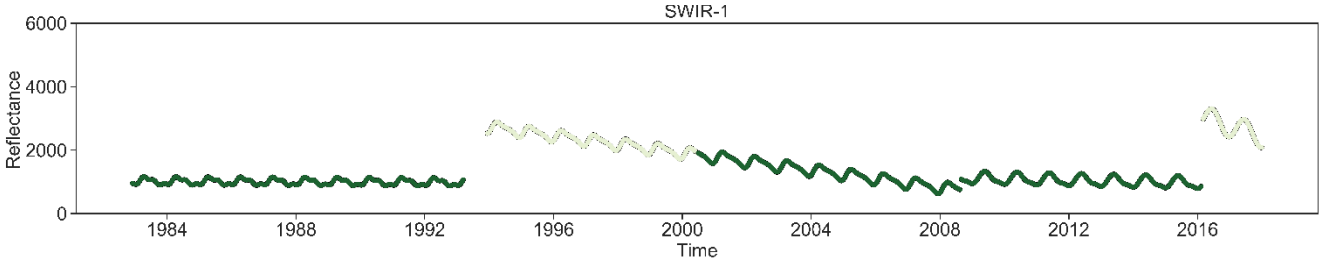


Figure 1-4. CCD harmonic segments combined with land cover information, where green signifies periods of Tree Cover and tan represents Grass/Shrub

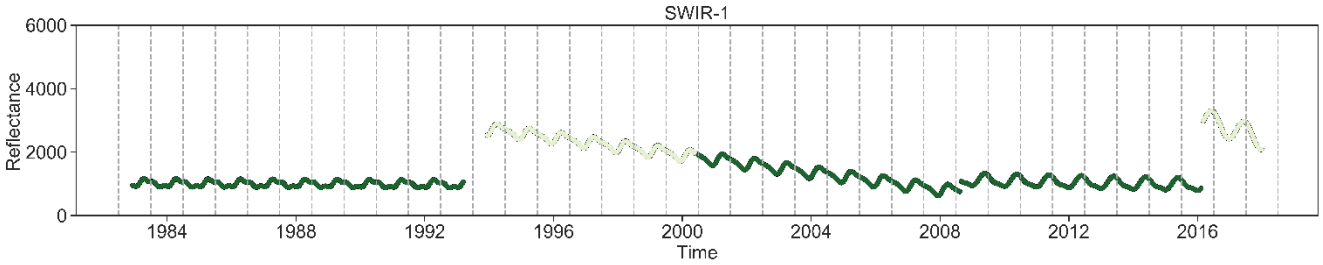


Figure 1-5. Annual July 1st dates overlaid with classified CCD harmonic segments

1.2 Purpose

The purpose of this document is to provide a high-level overview and technical details for the CCDC algorithm used for Collection 1.2.

1.3 LCMAP Project Documentation Suite

In accompaniment to the LCMAP Collection 1.2 Science Products, the LCMAP project provides a documentation suite, described in Table 1-1. All LCMAP documentation can be found on the LCMAP Website.

Document Type	Document Contents
LSDS-2320 LCMAP Collection 1.2 CCDC Algorithm Description Document (ADD)	Describes the Continuous Change Detection and Classification (CCDC) algorithm that is used to produce the associated science products part of the LCMAP project. The ADD gives in-depth descriptions of how various components of the CCDC operate and how the products and product values are derived.
LSDS-2321 LCMAP Collection 1.2 Data Format Control Book (DFCB)	Provides detailed information on data formats for the LCMAP Science Products. This includes information on product and file specifications, product packaging, and metadata file examples.
LSDS-2322 LCMAP Collection 1.2 Science Product Guide	Provides an overview of the current LCMAP approach, descriptions of the science products and their characteristics, and other relevant information to facilitate the use of LCMAP Science Products in the land change and land cover science community.

Table 1-1. LCMAP Collection 1.2 Documentation Suite

1.4 Document Organization

This document provides the technical details of the LCMAP implementation of the CCDC algorithm and how the subsequent product values are derived. In addition to this introduction, the document contains the following sections:

- Section 2: The CCD component of the algorithm, describing the harmonic modeling, the determination of spectral breaks, and the time series segment data returned by the USGS Python-based CCD (PyCCD) software
- Section 3: Classification, including the datasets that comprise the training data, how tile-based models are created, and how land cover predictions are generated
- Section 4: Product Definitions, describing how each of the land surface change products (i.e., those derived from the CCD results directly) and the land cover products (derived from classification) are produced
- Appendix A: Acronyms
- Appendix B: Default PyCCD parameters, with an explanation of each parameter (LCMAP Science Products were created using these default settings)
- References: Supporting documents

Section 2 Continuous Change Detection

2.1 Description

This section describes the CCD component of the CCDC algorithm and gives a brief synopsis of the USGS usage of the Python-based CCD (PyCCD) library. PyCCD is a port of the original MATLAB implementation of CCD. This open-source library has been made available to the public at the following location:

https://code.usgs.gov/lcmap/pyccd

CCD utilizes all available surface reflectance, brightness temperature, and associated quality data (together, these data are the “observations”) to create harmonic regression fits (models) for each input band to characterize the spectral response of every pixel. The harmonic regression fits are then used to categorize each pixel time series into temporal segments of stable periods and to estimate the dates at which the spectral time-series data diverge from past responses or patterns. Spectral time-series data divergence from past responses or patterns in a temporal segment indicates a model “break” (or “spectral break”). This is generally the result of an abrupt change (e.g., wildfire, logging, and mining), but can also result from a gradual shift (e.g., forest growth, insect infestation, disease). When a break occurs, a new temporal segment or model becomes established for the subsequent data points.

2.2 Dependencies

Landsat Collection 1 U.S. Landsat ARD provide the required level of data consistency to enable CCD to identify change based on spectral response.

These conditions are:

- Uniform mapping grid
- Geometric and radiometric consistency
- Per-pixel/observation quality information

2.3 CCD Inputs

A history is gathered from Landsat Collection 1 U.S. ARD for each pixel location consisting of the following information:

- Landsat Level 2 Surface Reflectance (SR) for the Thematic Mapper (TM), Enhanced Thematic Mapper Plus (ETM+), and Operational Land Imager (OLI)
- Landsat Level 2 Brightness Temperature (BT) for TM, ETM+, and the Thermal Infrared Sensor (TIRS)
- Landsat Level 2 Pixel Quality Assessment (PIXELQA) for TM, ETM+, and OLI
- Observation dates as ordinal values, where January 1 of the year 1 has ordinal value of 1 (proleptic Gregorian calendar)

2.4 CCD Output

The PyCCD library is a per-pixel algorithm, and the fundamental outputs are the spectral characterizations (segments) of the input data.

The output for each pixel is the following:

- Algorithm version
- Processing mask identifying which observations were used for harmonic calculations
- The percentage of observations that have been characterized as snow, water, and cloud by the pixel Quality Assessment (QA)
- List of harmonic regression fits (segments) for the stable temporal regions for the input data

Segments consist of the following attributes:

- Date that the segment begins
- Date that the segment ends
- Date of the first observation that statistically does not fit with the segment's harmonic regression (break date)
- The number of observations used between the start and end dates
- A value indicating whether the segment ended in a spectral break
- Quality information (“model QA”) for the harmonic regression fits
- Per band spectral information:
    - Seven harmonic regression fit coefficients
    - Intercept value for the regression fit
    - Root Mean Square Error (RMSE)
    - Magnitude of the spectral difference at the break date

As part of the operational LCMAP system, additional information is stored:

- Coordinate information (x,y) in the U.S. Landsat ARD grid system
- Dates of the ARD observations used in PyCCD

2.5 CCD Key Algorithm Concepts

Throughout this section and the next, software parameters are referred to in capital letters. Their descriptions and default values can be found in Appendix B.

2.5.1 Data Filtering

The removal of invalid and cloud-contaminated data points is important for deriving model coefficients that accurately represent the phenology of the surface, and for the correct identification of model break points. The CCD algorithm makes use of the U.S. Landsat ARD PIXELQA values to mask observations identified as cloud, cloud shadow, fill, or (in some cases) snow. Additional cirrus and terrain occlusion bits are provided for Landsat 8 OLI-TIRS ARD that are not available in the Landsat 4–7 TM/ETM+ quality assessment band. To maintain consistency across the historical archive, the algorithm does not use these Landsat 8-only QA flags to filter out observations. For more

information on PIXELQA, see the U.S. Landsat Collection 1 (C1) Analysis Ready Data (ARD) Data Format Control Book (DFCB).

Observations containing invalid data values are also removed. For the surface reflectance bands (which are not unscaled from their INT16 values), the valid data range is between 0 and 10000. Brightness temperature values are converted to 100 × °C (stored as 10 × temperature (kelvin) in the ARD), and observations are filtered for values within the range -9320 and 7070 (-93.2–70.7°C). This scaling is done to put the brightness temperature values into a roughly similar numerical range as the surface reflectance bands.

The above steps take place before any model is fit to the data points. Despite such filtering, pixel values with some degree of atmospheric obscuration will unavoidably remain in the dataset. During model fitting within the standard procedure outlined below, additional outlier removal is performed by using the multitemporal observation record as a further aid in identifying values that deviate from the overall phenology curve. Model initialization uses an adaptation of the multiTemporal Mask (Tmask) algorithm (Zhu & Woodcock, 2014b) to identify and exclude outliers within the initialization window (where the "window" is the span of observations covered by the procedure). That algorithm uses a specific harmonic model to perform an initial fit to the phenology, defined below:

$$\hat{p} = c_0 + c_1 \cos \omega t + c_2 \sin \omega t + c_3 \cos \frac{\omega t}{N} + c_4 \sin \frac{\omega t}{N} + c_5 t \qquad (1)$$

where,

ω the base annual frequency ($2\pi/\text{AVG\_DAYS\_YR}$),

N the smallest integer value greater than or equal to the fractional length of time covered by the observations in question relative to the base period, $\left\lceil \frac{t_0 - t_n}{AVG\_DAYS\_YR} \right\rceil$, where t_0, t_n are the first and last ordinal dates of the observations in the initialization window.

When moving through the time series, the model uses a threshold defined by the $OUTLIER\_THRESHOLD$ parameter to exclude values that deviate significantly from the established time series models. The default value of this threshold is derived from an inverse χ^2 distribution.

2.5.2 Harmonic Modeling

The time series is fit by harmonic models whose sinusoidal components are frequency multiples of the base annual frequency. A constant and linear term characterize the surface reflectance or brightness temperature offset value and overall slope, respectively. The full harmonic model is defined as follows:

$$\hat{p}(i,t) = c_{0,i} + c_{1,i} t + \sum_{n=1}^{3} (a_{n,i} \cos \omega n t + b_{n,i} \sin \omega n t) \qquad (2)$$

where,

ω the base annual frequency ($2\pi/\text{AVG\_DAYS\_YR}$)

t	the ordinal of the date, where January 1st of the year 1 has ordinal value of 1 (proleptic Gregorian calendar)
i	the i^{th} Landsat band
$a_{n,i}$, $b_{n,i}$	the estimated n^{th} order harmonic coefficients for the i^{th} Landsat band
$c_{0,i}$, $c_{1,i}$	the estimated intercept and slope coefficients for the i^{th} Landsat band
$\hat{p}(i,t)$	the predicted value for the i^{th} Landsat band at ordinal date t

The number of coefficients, defined to include c_0 and c_1 as well as the a_n and b_n harmonic coefficients used to model each spectral band, are parameterized by the software via the $COEFFICIENT\_MIN$, $COEFFICIENT\_MID$, and $COEFFICIENT\_MAX$ variables. For LCMAP, the values of these parameters are four, six, and eight, respectively, such that a four-coefficient (or "simple") model contains the constant term, linear term, and one sine and cosine term.

Model initialization, as well as certain special-case regression fits such as at the beginning/end of the time series, use the simple four-coefficient model. Outside of these cases, the choice of coefficient depends on the number of observations used for the regression. The "Look Forward" step (see procedure below) requires that the number of coefficients used, multiplied by NUM_OBS_FACTOR (3), cannot exceed the number of observations under consideration (or else, must be $COEFFICIENT\_MIN$). For a full model (eight coefficients), there must be at least twenty-four observations covered by the regression. The number of coefficients used is recorded by the model QA and is reported by the Spectral Model Quality (SCMQA) product as described in Section 4.2.5. Note that this QA does not report whether the remaining coefficients were zeroed during regression. The fit parameters returned by PyCCD always include eight values (seven coefficients and an intercept; all are referred to as coefficients in this document), with unused coefficients reported as zeroes.

2.5.3 Regression

Least Absolute Shrinkage and Selection Operator (LASSO) regression is used to calculate best-fit coefficients for the time series model. In contrast to Ordinary Least Squares (OLS), LASSO penalizes the sum of the absolute values of coefficients, in some cases, forcing a subset of the coefficients to zero. Together with the explicit limits enforced on the number of coefficients by the NUM_OBS_FACTOR parameter, this has the effect of reducing instances of overfitting, including in cases when observations are too sparse or unevenly distributed through time to constrain the model to actual phenological features. The Tmask algorithm uses a different regression algorithm, Robust Iteratively Reweighted Least Squares (RIRLS), which is designed to be less sensitive to outliers. Tmask and RIRLS are used only during model initialization.

2.5.4 Data Variability

In order to detect change, the algorithm must distinguish between a substantive deviation from model prediction, and deviations that result from variability inherent in the data (due to incomplete atmospheric removal and/or other sources of natural variation). The algorithm relies on two metrics to estimate the variability of data for each spectral band. The first is a comparison RMSE, defined as the RMSE of the twenty-four

observations covered by the model which are closest in day of year to the last observation in the "peek window" (see below), or over all observations covered by the model if there are fewer than twenty-four. This value is recalculated at each step in the time series.

A second value (referred to as var below) is computed once, at the beginning of the standard procedure, using all non-masked observations in the time series. This value, which is intended to describe the overall variability of the data values, is defined as the median of the absolute value of the differences between each observation and the i^{th} successive observation, where i is the smallest value such that the majority of these observation pairs are separated by greater than 30 days, if possible (otherwise, $i = 1$).

2.5.5 Change Detection Thresholds

As the algorithm steps through the time series, it checks for a model break (spectral change) with respect to its last determined best-fit harmonic model. Change detection sensitivity depends on the value of $CHANGE\_THRESHOLD$. This threshold is drawn from an inverse χ^2 distribution, with degrees of freedom equal to the number of elements in $DETECTION\_BANDS$. For LCMAP, the detection bands are the green, red, Near Infrared (NIR), Shortwave Infrared Band 1 (SWIR-1), and Shortwave Infrared Band 2 (SWIR-2) bands (the blue and thermal bands are not used in change detection).

Observations not yet incorporated into the model are evaluated as a group of no fewer than the $PEEK\_SIZE$ parameter value; this is the "peek window", which "slides" along the time series by one observation at a time. Each iteration, a value is calculated for each individual observation within the peek window, as follows:

$$mag_n = \sum_{i \in D} \left(\frac{resid_{n,\,i}}{max(var_i,\ RMSE_i)} \right)^2 \qquad (3)$$

where,

$resid_{n,\,i}$ — the residual relative to the LASSO models for each band i, for each observation n within the $PEEK\_SIZE$ window

$var_i,\ RMSE_i$ — the measures of dispersion as described above, for each band i

This summation is carried out across all bands i in the set of $DETECTION\_BANDS$ (D). This produces a scalar magnitude, representing the deviation from model prediction across these bands, for each observation. The detection of a model break requires this value to be above the $CHANGE\_THRESHOLD$ value for all observations in the window. Note that this magnitude is different from the value that is reported as a per-band magnitude when a change is detected in the time series.

The algorithm may, at the beginning of the standard procedure, adjust both the $PEEK\_SIZE$ and $CHANGE\_THRESHOLD$ from their default values, in response to the data density. Specifically, if valid observations occur, on average, more frequently in the time series (due to, e.g., a pixel being located within the side-lap regions of image swaths), the algorithm will increase the $PEEK\_SIZE$ (examine more observations within

a peek window) and decrease the change threshold required to trigger a model break (to adjust for the reduced likelihood that the increased number of observations will all clear the threshold).

2.5.6 Forward Processing

Because Landsat observations are continually acquired, and one of the goals of the LCMAP project is to provide timely updates of changes in the land surface, extending the model using newly acquired observations is an important capability for the algorithm. An important consideration in forward processing is maintaining consistency with previous PyCCD processing. As of version 2020.10.10, the PyCCD software has been updated with a new parameter (STAT_ORD), and the optional capability to accept a previous set of results.

The STAT_ORD parameter establishes the end ordinal date (inclusive) to be evaluated for many calculations that normally take the entire time series into consideration. This is a required parameter, and must be set with consideration to the time series information being provided. This parameter affects the time span used for the following calculations (see Section 2.6 for more details):

- Fit procedure to be used by the model (based on the proportion of cloud, snow, and other QA bits)
- PEEK_SIZE and CHANGE_THRESHOLD adjustments performed at the beginning of the Standard procedure (based on average data density over the base period)
- The value of var, which characterizes the overall scatter of the data points
- Additional masking used for the Insufficient Clear procedure (described further below)

When a previous set of results are supplied, PyCCD will try to start building models from the last identified spectral break. This generally provides a boost to processing time in relation to the total number of breaks found previously, but also helps ensure floating point consistency. Utilizing a set of previous results affects the algorithm's behavior in the following ways:

- The procedure to be used (Standard, Insufficient Clear, or Permanent Snow, described below) is set to that used in the previous set of results.
- During the Standard procedure, previous results from before the most recent break day are incorporated into the returned results array and the procedure skips ahead to begin at the last break day reported.

Conceptually, forward processing extends unfinalized segments (those that have not yet encountered a spectral break) and replaces "end fits", basic models at the end of a time series where there was an insufficient number of observations to properly initialize a model. No changes occur to model segments in previous results among those that have already seen a change. For unfinalized segments, multiple years' worth of spectral product values (see Section 4) will sometimes be modified relative to those present in

previous results. In addition to extending the segments and potentially finding new break points, model re-fitting will alter the values of the harmonic coefficients; see Section 3.5.4 for a discussion of the implications this will have for classification and land cover product values.

2.6 CCD Step-by-Step Walk-through

The following is an in-depth step-by-step walk-through of how the PyCCD software operates. This is organized first by data setup and procedure selection, then broken down by how the individual procedures operate on the data. Information on known issues is also provided. Refer to Figure 2-1 for a diagram of the PyCCD algorithm.

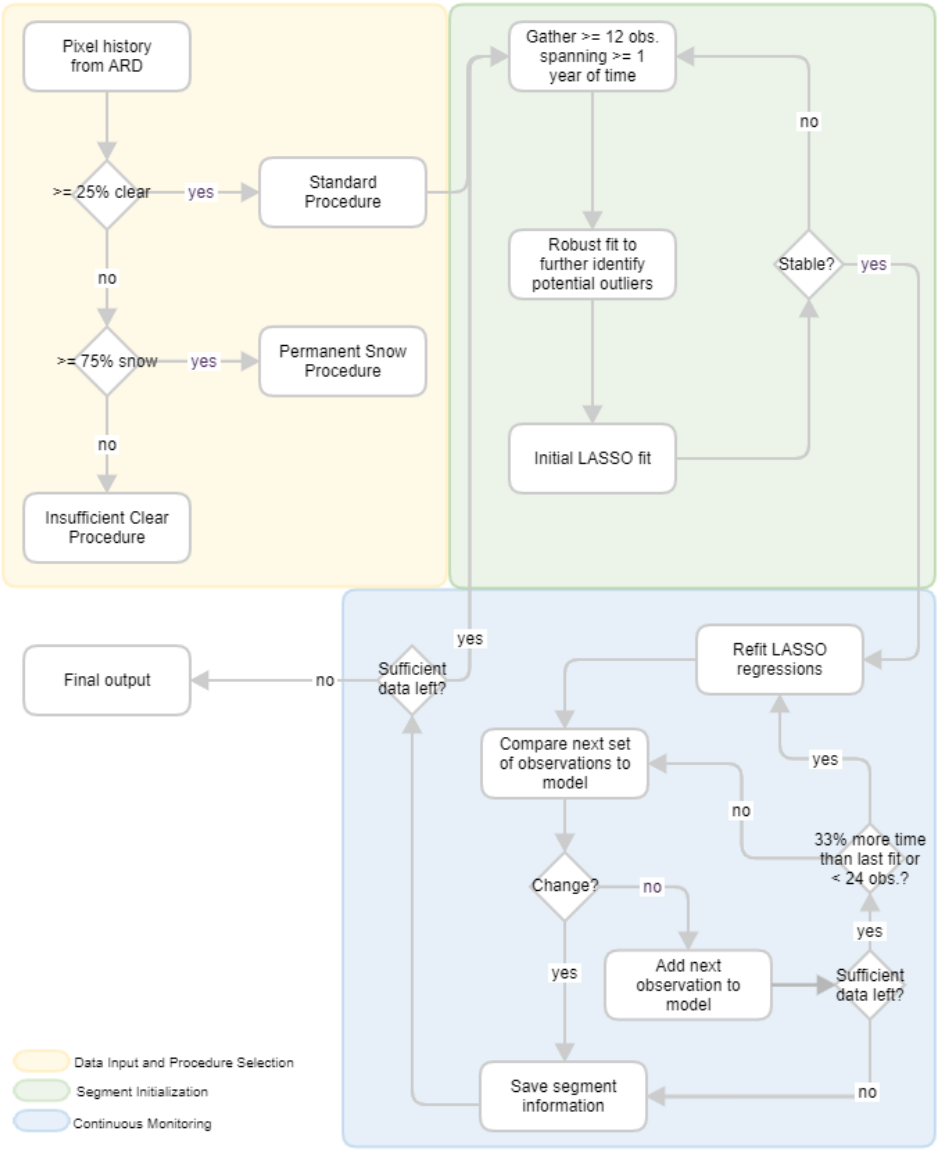


Figure 2-1. Overview of CCD Processing

2.6.1 Data Setup and Procedure Selection/Run

1. Update the default processing parameters (refer to Appendix B) with user-specified overrides.
2. Compare all input arrays (dates, spectra, QA) to ensure they are of the same length.
3. Sort all input by date.
4. Check $QA\_BITPACKED$:
    a. If True, use the offsets identified in the other QA processing parameters to identify if an observation is fill, cloud, shadow, snow, water, or clear.
        - If both $QA\_CIRRUS1$ and $QA\_CIRRUS2$ bits are set, QA_CLEAR will be returned.
        - Similarly, if the $QA\_OCCLUSION$ bit is set, QA_CLEAR will be returned.
    b. If $QA\_BITPACKED$ is False, then the values in the QA array will be taken as is and must be consistent with the other QA processing parameters.
5. Calculate the probability that an observation from the time series is cloud, water, or snow based on the QA array:

$$Percent\ cloud = \frac{n_{cloud}}{n_{non\text{-}fill}}$$

$$Percent\ water = \frac{n_{water}}{n_{clear} + n_{water} + 0.01}$$

$$Percent\ snow = \frac{n_{snow}}{n_{clear} + n_{water} + n_{snow} + 0.01}$$

6. Select a procedure based on the QA array, filtered to include only observations with ordinal dates ≤ STAT_ORD:
    a. If ≥ 25% of the observations are identified as clear or water, use the Standard Procedure.
    b. If ≥ 75% of the observations are identified as snow, use the Permanent Snow Procedure.
    c. Failing the above tests, use the Insufficient Clear Procedure.
7. Run the selected procedure.
8. Attach information about the version of the algorithm and the cloud/water/snow probabilities to the results before returning the final output to the user.

2.6.2 Individual Procedure Operations

2.6.2.1 Standard Procedure

The standard procedure is the default CCD method, used for all pixels that pass the PIXELQA check described above (which is expected to be the majority of all pixels). The standard procedure detects spectral change by comparing surface reflectance predicted by the model to measured surface reflectance. Models may use different numbers of harmonic terms to fit the data, depending on the number of observations covered by the model period. A variable number of segments may be returned for the set of input observations, depending on how many times the model breaks and re-initializes a new segment. The sequential steps of the standard processing procedure are as follows:

1. Convert the thermal band brightness temperature from kelvin to degrees Celsius, further scaled by 10 from normal U.S. Landsat ARD output. This puts the values into a similar range as the surface reflectance bands:
   $$BT_{scaled\,℃} = BT_{ARD} \times 10 - 27315$$
2. Filter for usable observations based on the following criteria:
   a. The observation is identified from QA as water or clear.
   b. All surface reflectance bands are within the valid range:
      $$0 < SR_{ARD} < 10000$$
   c. The brightness temperature value is within the valid range:
      $$-9320 < BT_{scaled\,℃} < 7070$$
   d. There are no duplicate (same date) observations already accepted through the above criteria.
3. If previous results have been provided, initialize the array of results with finalized segments from the previous run.
   a. Removed any unfinished segments and end fits.
   b. Adjust the model's starting position to begin where the last finalized previous model encountered a spectral break.
4. If fewer than $MEOW\_SIZE$ observations remain after filtering, do not attempt to fit a model to the data, and return zero segments.
5. Adjust the number observations used to detect change ($PEEK\_SIZE$, six by default) and the statistical change threshold ($CHANGE\_THRESHOLD$):
   a. $PEEK\_SIZE$ adjustment:
      $$PEEK\_SIZE_{adj} = round\left(PEEK\_SIZE \times \frac{16}{median(T_{diff})}\right),$$
      where $T_{diff} = \{t_{n+1} - t_n\}$ is the set of successive differences between observation dates for all $n \in \{1, \ldots, N-1\}$, where N is the total number of valid observations with ordinal dates ≤ STAT_ORD. If $PEEK\_SIZE_{adj} < PEEK\_SIZE$, then use the original $PEEK\_SIZE$.
   b. If $PEEK\_SIZE$ was adjusted, then $CHANGE\_THRESHOLD$ needs to be recalculated based on the new size:
      $$CHANGE\_THRESHOLD = Inv\chi^2\left(1 - (1 - 0.99)^{\frac{PEEK\_SIZE}{PEEK\_SIZE_{adj}}}\right),$$
      where $Inv\chi^2$ is an inverse χ^2 distribution with five degrees of freedom (i.e., the number of $DETECTION\_BANDS$).
6. Compute a value representing the overall variability of the data values in the time series, as follows:
   a. Calculate var_i for each of the x spectral bands, as the median of the absolute value of the differences between successive observations in each band:
      $$var_i = median\left(Obs_{diff,\,i}\right),$$
      where $Obs_{diff,\,i} = \{|obs_{i,n+1} - obs_{i,n}|\}$ is the set of the absolute values of successive differences between observations, for all $n \in \{1, \ldots, N-1\}$, where N is the total number of valid observations.
   b. Adjust for temporal variability by making the majority of observation dates under comparison greater than 30 days apart:

i. Find the smallest index offset j, where $j > 0$, such that the majority (mode) of the shifted temporal differences (T_{diff}) is greater than 30:

$$T_{diff} = \{t_{n+j} - t_n\},$$

where $\{t_{n+j} - t_n\}$ is the set of differences between each observation date and the j^{th} successive date for all $n \in \{1, \ldots, N - j\}$, where N is the total number of valid observations. Increment j until the condition $mode(T_{diff}) > 30$ is met. If no such j is found, use var_i as calculated above and skip the following two steps.

ii. Identify the observations (obs^*) where the shifted temporal difference is greater than 30.

iii. Calculate an adjusted spectral variability value for each band, using observations > 30 days apart:

$$var_i = median\left(Obs^*_{diff,\,i}\right),$$

where $Obs^*_{diff,\,i} = \{|obs^*_{i,n+1} - obs^*_{i,n}|\}$ is the of the absolute values of successive differences between filtered observations, for all $n \in \{1, \ldots, N^* - 1\}$, where N^* is the total number of valid filtered observations.

7. While there are at least twelve ($MEOW\_SIZE$) observations left to initialize a new model, do the following:
   a. **Initialize**. Try to initialize a new harmonic segment for continuous monitoring:
      i. Create a window of at least $MEOW\_SIZE$ observations that span at least one year of time.
      ii. Use the green and SWIR-1 bands ($TMASK\_BANDS$) to further identify outliers through the Tmask filtering:

      Outliers are identified if a residual ($resid_{i,n}$) exceeds the distance set by the var_i scaled by T_CONST:

      $$\left|resid_{i,n}\right| > var_i \times T\_CONST,$$

      for all $i \in TMASK\_BANDS$ and for all $n \in \{1, \ldots, N\}$, where N is the total number of observations in the window.

      iii. If, after excluding these outliers, there are not at least twelve observations ($MEOW\_SIZE$) or the observations do not span at least one year, include the next observation in the time series and return to step ii above.
      iv. Mask the outliers identified by Tmask from future processing.
      v. Create initial LASSO harmonic fits using a simple, four-coefficient model ($COEFFICIENT\_MIN$).
      vi. Conduct a stability check to ensure the initial regression fits are not over a potential landscape disturbance:

      $$\sum_{i \in D} \frac{|c_{1,i}t_1 - c_{1,i}t_N| + |resid_{1,i}| + |resid_{N,i}|}{max(var_i, RMSE_i)} < CHANGE\_THRESHOLD\ ,$$

      where c_1 is the spectral slope coefficient, t_1, t_N are the first and last dates for observations $\{1, \ldots, N\}$ covered by the regression fit, $resid_{1,i}, resid_{N,i}$ are the first and last residuals

for the same, $RMSE_i$ is the per-band root-mean-square error over the window, for each spectral band $i \in D$, where D are the $DETECTION\_BANDS$.

vii. If the stability check fails, skip the first observation and start initialization again.

b. **Look Back**. If initialization has skipped observations, then look back at those observations with the initialized model to determine if they statistically belong:
   i. Gather the most recent $PEEK\_SIZE$ observations that were skipped (or all, if fewer).
   ii. Calculate the per-band residuals ($resid_{n,i}$) relative to the LASSO models for each observation n within the $PEEK\_SIZE$ window.
   iii. Calculate a magnitude (mag_n) as $mag_n = \sum_{i \in D} \left(\frac{resid_{n,i}}{max(var_i, RMSE_i)} \right)^2$ (Equation 3) for each spectral band $i \in D$, where D are the $DETECTION\_BANDS$.
   iv. If $mag_n > CHANGE\_THRESHOLD$ for all n in the $PEEK\_SIZE$ window, do not add any further observations to the segment, and exit Look Back.
   v. If the most recent observation in the Look Back window meets outlier criteria ($mag_n > OUTLIER\_THRESHOLD$), exclude it from further processing. Otherwise, add it to the segment.
   vi. Continue checking the skipped observations while any remain.

c. If there are no finalized segments for the time series (i.e., the model just initialized for the first time), and initialization has skipped at least $PEEK\_SIZE$ observations (after Look Back), then perform a harmonic regression fit to those observations using a simple, four-coefficient model ($COEFFICIENT\_MIN$), and set the model QA value to 14 ($CURVE\_QA: START$), indicating a "start fit". Report the model coefficients and RMSE from this regression. Use the window of observations covered by the model to report the start, end, and observation count. Set the change Boolean value to 0 and record the break day as the date of the first observation not covered by the start fit. Report the magnitude of change as zero for each band.

d. **Look Forward**. While there are still at least $PEEK\_SIZE$ observations remaining after the end of the current segment, look forward in the time series trying to add observations to the segment, until a spectral break is detected.
   i. If there are < 24 observations or 33% more time has passed than that covered by the last regression fit, refit the per-band LASSO regressions:
      1. For the harmonic models, use the number of coefficients n_coeff determined by the number of observations N covered by the model, as follows:

$$n\_coeff = \begin{cases} 4, & 12 \le N < 18 \\ 6, & 18 \le N < 24 \\ 8, & N \ge 24 \end{cases}$$

i.e., such that $n\_coeff \times NUM\_OBS\_FACTOR$ (3) is at least equal to N.

ii. Calculate the residuals ($resid_{n,i}$) for the next $PEEK\_SIZE$ observations in the time series.

iii. Calculate an $RMSE_i$ for the per-band regression models based on the 24 observations currently covered by the model (or all observations, if there are fewer than 24) that are closest by day-of-year to the first observation in the $PEEK\_SIZE$ window, as follows:

$$RMSE_i = \sqrt{\frac{\sum_{n \in N}(resid_{n,\,i})^2}{N - n_{coeff}}},$$

where N is the number of observations covered by the model (or 24, if more than that number), and n_coeff is the number of coefficients used. (This equation also describes the RMSE calculation that is reported in the output, with the exception that N represents the full number of observations, not just ≤ 24.)

iv. Using $resid_{n,i}$, the $RMSE_i$, and var_i, calculate a mag_n, for all observations in the $PEEK\_SIZE$ window, as per Equation 3:

$$mag_n = \sum_{i \in D}\left(\frac{resid_{n,\,i}}{max(var_i,\ RMSE_i)}\right)^2$$

v. If $mag_n < CHANGE\_THRESHOLD$ for any n in the $PEEK\_SIZE$ window, add the most recent observation to the segment:
   1. Shift the $PEEK\_SIZE$ window one observation forward in the time series.
   2. Start the next iteration of Look Forward.

vi. If $mag_n > CHANGE\_THRESHOLD$ for all n in the $PEEK\_SIZE$ window, this is considered a spectral break; finalize the segment:
   1. Report the magnitude of change for each band:

      $$magnitude_i = median(\{resid_{n,i}\})$$

      for all $n \in \{1, \ldots, N\}$, where N is the number of observations covered by the $PEEK\_SIZE$ window, for each spectral band i.

      Report the model coefficients and RMSE determined via the last regression, for each spectral band; also, report the number of coefficients used (the model QA). Use the window of observations covered by the model to report the start, end, and observation count. Set the change Boolean value to 1 and record the break day as the date of the first observation in the peek window.

2. Exit Look Forward.

vii. If $mag_n > OUTLIER\_THRESHOLD$ for the most recent observation in the peek window, exclude it from further processing:
   1. Adjust the $PEEK\_SIZE$ window to account for the excluded observation.
   2. Start the next iteration of Look Forward.

e. Iterate over the remaining observations in the time series.

8. If there are still at least $PEEK\_SIZE$ valid observations remaining at the end of the time series, then perform a harmonic regression fit to those observations using a simple, four-coefficient model ($COEFFICIENT\_MIN$), and set the model QA value to 24 ($CURVE\_QA$: END), indicating an "end fit". Report the model coefficients and RMSE from this regression. Use the window of observations covered by the model to report the start, end, and observation count. Set the change Boolean value to 0 and record the break day as the last observation date. Report the magnitude of change as zero for each band.

2.6.2.2 Permanent Snow Procedure

Selection of the Permanent Snow procedure indicates that too few clear or water observations exist to robustly detect change, and a large fraction are snow. The algorithm will return at most one segment, fit through the entire time series, provided the filtered observations number at least twelve ($MEOW\_SIZE$). The model will, under the default settings, fit only four coefficients (i.e., characterizing the reflectance and brightness temperature bands using only a simple harmonic with no higher frequency terms). Unlike other procedures, snow observations are not filtered out. Instead, they are used to characterize the seasonal behavior of the land surface.

1. Filter for usable observations based on the following criteria:
   a. Either the value is identified from QA as snow, or all of the following conditions are met:
      i. The observation is identified from QA as water or clear.
      ii. All surface reflectance bands are within the valid range: $0 < SR_{ARD} < 10000$
      iii. The brightness temperature value is within the valid range: $-9320 < BT_{scaled\,℃} < 7070$
   b. There are no duplicate (same date) observations already accepted through the above criteria.
2. If fewer than twelve ($MEOW\_SIZE$) observations remain after filtering, do not attempt to fit a model to the data, and return zero segments.
3. Perform a LASSO regression fit for the entire time series using four coefficients ($COEFFICIENT\_MIN$), and set the model QA to 54 ($CURVE\_QA$: $PERSISTENT\_SNOW$). Report the model coefficients and RMSE from this regression. Use the window of observations covered by the model to report the start, end, and observation count. Set the change Boolean value to 0 and record the break date as the last observation date. Report the magnitude of change as zero for each band.

2.6.2.3 Insufficient Clear Procedure

Selection of the Insufficient Clear procedure indicates that too few clear or water observations exist to robustly detect change. The algorithm will return at most one segment, fit through the entire time series, provided the filtered observations number at least twelve ($MEOW\_SIZE$). The model will, under the default settings, fit only four coefficients (i.e., characterizing the reflectance and brightness temperature bands using only a simple harmonic with no higher frequency terms) to avoid overfitting. In addition to excluding snow observations, this procedure differs from the Permanent Snow procedure by the inclusion of an additional filtering step on the $GREEN\_IDX$ bands, designed to further reduce cloud/outlier data points that may negatively affect a model fit through a sparse time series.

1. Filter for usable observations based on the following criteria:
    a. The observation is identified from QA as water or clear.
    b. All surface reflectance bands are within the valid range:
        $0 < SR_{ARD} < 10000$
    c. The brightness temperature value is within the valid range:
        $-9320 < BT_{scaled\,℃} < 7070$
    d. There are no duplicate (same date) observations already accepted through the above criteria.
    e. The $GREEN\_IDX$ band value is less than the sum of the median $GREEN\_IDX$ value of all the observations with ordinal dates ≤ STAT_ORD, and a constant (400) defined by $MEDIAN\_GREEN\_FILTER$.
2. If fewer than twelve ($MEOW\_SIZE$) observations remain after filtering, do not attempt to fit a model to the data, and return zero segments.
3. Perform a LASSO regression fit for the entire time series using four coefficients ($COEFFICIENT\_MIN$), and set the curve QA to 44 ($CURVE\_QA$: $INSUF\_CLEAR$). Report the model coefficients and RMSE from this regression. Use the window of observations covered by the model to report the start, end, and observation count. Set the change Boolean value to 0 and record the break day as the last observation date. Report the magnitude of change as zero for each band.

2.7 Known Issues

The issues listed below are unintentional modifications to the above procedures that arose from programming errors. These were discovered during or subsequent to production processing efforts. They are documented here both for user awareness, as well as for reference with respect to any potential changes in future product releases. The ramifications of each of these are believed to be minor.

- The Permanent Snow and Insufficient Clear procedures do not transform the brightness temperature band values. This leads to different model fits for a small minority of pixels that can affect downstream usages, such as classification procedures.
- When determining whether there are enough observations while moving forward through the time series with a segment that has not broken, one of the stop

conditions is missing an equality. This can lead to misreporting of the resulting end date for the final time series model.

- The number of coefficients is sometimes misreported when moving through the Look Forward procedure. This may result in an incorrect model QA value indicating a higher number of coefficients than were actually used in the fit. A small portion of Spectral Model Quality pixels will be affected within the resulting products.

Section 3 Classification

3.1 Description

This section describes the classification component of the CCDC algorithm, and the methodology used to determine land cover classification of a pixel at any point in the Landsat 4–8 record using the LCMAP time series model approach.

The classification element of CCDC produces a land cover classification for every pixel based on data from the time series models (e.g., model coefficients). Land cover classifications are generated on an annual basis, using July 1st as a representative date. A list of land cover classes and descriptions is provided in Table 3-1.

Code	Land Cover Class	Description
1	Developed	Areas of intensive use with much of the land covered with structures (e.g., high-density residential, commercial, industrial, mining, or transportation), or less intensive uses where the land cover matrix includes vegetation, bare ground, and structures (e.g., low-density residential, recreational facilities, cemeteries, transportation/utility corridors, etc.), including any land functionality related to the developed or built-up activity.
2	Cropland	Land in either a vegetated or unvegetated state used in production of food, fiber, and fuels. This includes cultivated and uncultivated croplands, hay lands, orchards, vineyards, and confined livestock operations. Forest plantations are considered as forests or woodlands (Tree Cover class) regardless of the use of the wood products.
3	Grass/Shrub	Land predominantly covered with shrubs and perennial or annual natural and domesticated grasses (e.g., pasture), forbs, or other forms of herbaceous vegetation. The grass and shrub cover must comprise at least 10% of the area and tree cover is less than 10% of the area.
4	Tree Cover	Tree-covered land where the tree cover density is greater than 10%. Cleared or harvested trees (i.e., clearcuts) will be mapped according to current cover (e.g., Barren, Grass/Shrub).
5	Water	Areas covered with water, such as streams, canals, lakes, reservoirs, bays, or oceans.
6	Wetland	Lands where water saturation is the determining factor in soil characteristics, vegetation types, and animal communities. Wetlands are composed of mosaics of water, bare soil, and herbaceous or wooded vegetated cover.
7	Ice/Snow	Land where accumulated snow and ice does not completely melt during the summer period (i.e., perennial ice/snow).
8	Barren	Land comprised of natural occurrences of soils, sand, or rocks where less than 10% of the area is vegetated.

Table 3-1. LCMAP Level 1 Land Cover Classes

3.2 Dependencies

Classification is dependent on the following criteria:

- CCD model results are available for the training and classification extent
- CCD results and other input data are geometrically aligned

- The data are relevant to the training time period, circa 2001

3.3 Classification Inputs

Input for CCDC classification is the output of CCD combined with ancillary data. CCDC classification utilizes the following datasets:

- LCMAP CCD model results:
  - CCD temporal segment coefficients
  - CCD temporal segment Root Mean Square Error (RMSE)
- National Elevation Dataset (NED, Gesch et al., 2002) Digital Elevation Model (DEM) and derivatives:
  - DEM-derived slope
  - DEM-derived aspect
  - DEM-derived topographic position index
- National Land Cover Database (NLCD, 2011 ed.) Wetland Potential Index (WPI), a measure of agreement among the following datasets:
  - The U.S. Department of Agriculture (USDA) Natural Resources Conservation Service Soil Survey Geographic database (SSURGO) for hydric soils
  - The National Wetlands Inventory (NWI)
  - NLCD (2006 ed., Fry et al., 2011) 2006 land cover
- National Land Cover Database 2001 land cover (2011 ed., Homer et al., 2015)

3.4 Classification Outputs

The outputs of the classification methodology are a trained classification model for an ARD tile location, and annual July 1st predictions of the land cover class.

Trained classification models consist of:

- Trained gradient boosted decision tree

Annual July 1st predictions consist of:

- Locational information
- Prediction date
- Start date for the applicable change detection segment
- End date for the applicable change detection segment
- A list consisting of the model confidence for each of the 8 land cover classes

3.5 Classification Key Concepts

3.5.1 Algorithm

XGBoost (“eXtreme Gradient Boosting”; Chen & Guestrin, 2016) is a scalable implementation of gradient tree boosting, which is a supervised learning method that can be used to develop a classification model when provided with an appropriate training dataset. Training data comprise a set of features (here, per-pixel data at

Landsat 30 m spatial scale), with a classification label that allows the model to learn the features that are predictive of the land cover. The resultant trained model can then be applied to a larger dataset to generate predictions and probability scores that are the basis for LCMAP primary and secondary land cover, and primary and secondary land cover confidence values.

3.5.2 Training Data and Sample Size

The LCMAP Collection 1 land cover products are produced by an XGBoost classification model that is trained using the NLCD (2011 ed.) land cover for 2001, CCD model results, and several ancillary data layers. NLCD uses a finer-grained Anderson Level ll-based legend, in contrast to LCMAP's broader Level I-derived classes (Anderson et al., 1976). With this finer granularity, the NLCD data must first be cross-walked to LCMAP classes, as shown in Figure 3-1 and Table 3-2. The extent of each land cover in the cross-walked NLCD layer is eroded by one pixel. This step aims to reduce "noise" in the classifier by removing pixels that may be heavily mixed, or whose land cover label may be less reliable. It also removes the narrow linear low intensity developed pixels corresponding to road networks mapped by NLCD, which were found to have registration issues with ARD.

Ancillary data is comprised of two main source datasets: the USGS NED (Gesch et al., 2002), 1 arc-second DEM, and a WPI layer created as USGS in-house ancillary data for NLCD 2011 land cover production (Zhu et al., 2016). The DEM was re-projected to the ARD Albers Equal Area (AEA) conic map projection to produce elevation data gridded consistently with the NLCD and LCMAP data. The WPI layer is a ranking (0–8) of wetland likelihood from a comparison of the NWI, the USDA SSURGO database for hydric soils, and the NLCD 2006 wetlands land cover classes.

Three calculated DEM derivative values are also supplied as ancillary data: topographic slope, aspect, and position index. Slope was calculated using the maximum change in elevation between the pixel in question and the eight surrounding pixels, expressed as a continuous variable (0–90°). The aspect layer identifies the downslope direction of the maximum rate of change in topographic elevation. These values indicate the compass direction that the sloped surface is facing at that pixel location. Aspect is expressed as a categorical value, with values that range from 1–16 and represent clockwise degrees from north, in intervals of 22.5° (i.e., a value of 1 indicates a downslope direction of 0–22.5° from north). Flat areas have no downslope direction and are given a value of –1. Position index is an indicator of the topographic context of a pixel (hilltop, valley bottom, exposed ridge, flat plain, upper or lower slope, etc.). It is defined as:

$$\frac{z_i - z_{min}}{z_{max} - z_{min}} \times 100 \qquad (4)$$

where,

z_{max}, z_{min} the highest and lowest elevation within a 7x7 window

z_i the elevation of a pixel at the center of a 7x7 window

For each pixel, CCD segment data for the segment that includes the July 1st, 2001, date is used. Pixels that do not include a segment with a stable CCD model for that date are excluded from the training pool (Figure 3-2). The model data used in training are the surface reflectance and brightness temperature model coefficients (except the intercepts) for each band, the model RMSE value for each band, and an average reflectance value for each band for the July 1st, 2001, year that is calculated as:

$$avg\_refl_i(t) = c_{0,i} + c_{1,i}t \quad (5)$$

where,

t the ordinal date for July 1st, 2001

$c_{0,i}$, $c_{1,i}$ the estimated intercept and slope coefficients for the i^{th} Landsat band

$avg\_refl_i(t)$ the average reflectance for the i^{th} Landsat band on ordinal day t

For training as well as prediction, higher-order terms for “simple” and “advanced” models (four and six coefficient models, respectively), are provided to the XGBoost routines as zeroes.

Training takes place at the tile level, using a random sample drawn from the tile to be classified as well as the eight surrounding tiles (or fewer, if near the edge of Conterminous United States (CONUS) ARD). Cross-walked and eroded NLCD data are used as classification labels, while the CCD results and ancillary data are provided as independent variables. Based on testing of different sample sizes for the training dataset (Zhou et al., 2020), a target sample size of 20 million pixels was chosen, requiring approximately proportional representation of classes with the added constraint that no class be represented by fewer than 600,000 or more than 8 million samples. If there are fewer than 600,000 samples available for a class, then all the available samples are used without any oversampling.

3.5.3 Prediction

Once the model for a given tile is trained, prediction is relatively straightforward. The input data provided to the classifier are the DEM elevation and derivatives, the WPI, and the CCD segment data (coefficients and RMSE for every segment). Predictions are generated for each July 1st date that has an associated CCD segment, replacing the model intercept value with an average annual reflectance as given by Equation 5 before being passed to the classifier. (Note that many of these predictions will land within the time period of the same CCD segment; in these cases, the input will be very similar, but the average reflectance value may differ, depending on the value of the spectral slope.) The resulting prediction information is supplied to the production generation step (see Section 4) for the creation of land cover tiffs. The process is repeated for each tile in the LCMAP CONUS ARD extent.

NLCD Value	LCMAP Value
11	5
12	7

NLCD Value	LCMAP Value
21, 22, 23, 24	1
31	8
41, 42, 43	4
51, 52, 71, 72, 73, 74	3
81, 82	2
90, 95	6

Table 3-2. NLCD 2001 (2011 Edition) to LCMAP Land Cover Cross-walk

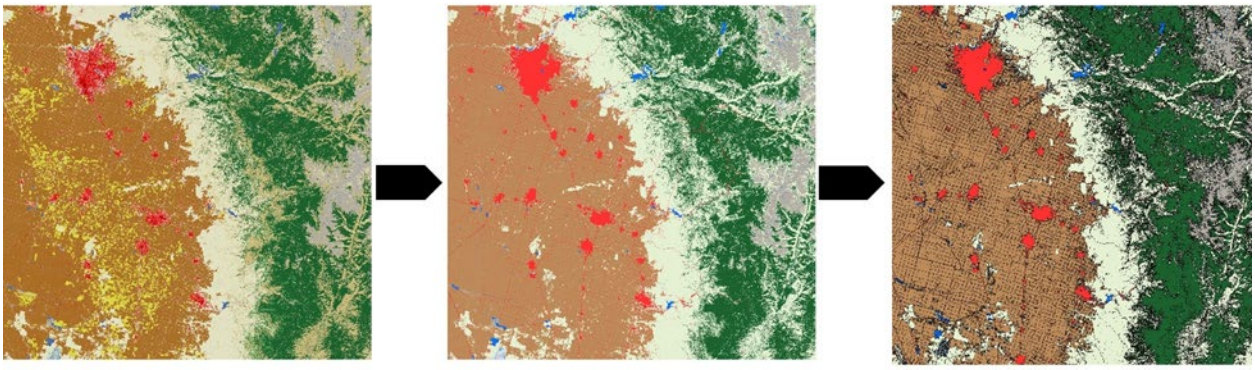


Figure 3-1. Moving from left to right, tile h03v10 is cross-walked from original NLCD 2001 (2011 Edition) to LCMAP level 1 classes, then eroded by one pixel

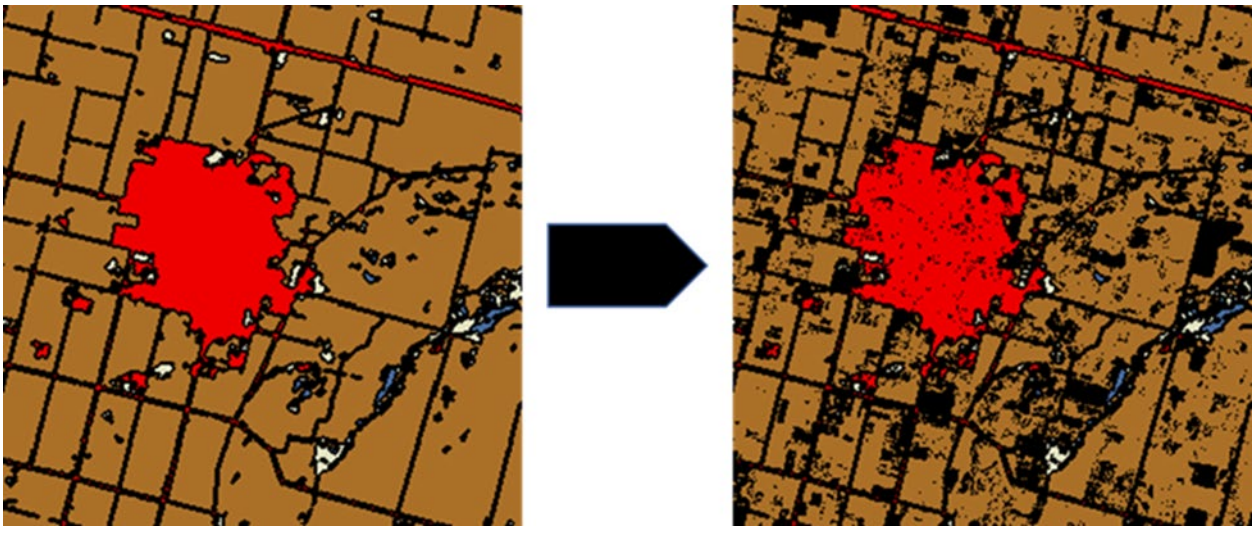


Figure 3-2. Cross-walked and eroded NLCD data (left) is filtered to include only pixels with CCD segments that encompass target training date (2001-07-01)

3.5.4 Forward Processing

The acquisition and processing of additional observations allows land cover classification to be extended forward in time. The procedure for extending the time

series harmonic models from a previous set of results is covered in Section 2.5.6. To maintain consistency, the classification models previously trained for LCMAP Collection 1.0 are used for land cover predictions for Collection 1.2.

Note that, because previously unfinalized (unbroken) harmonic models may be re-fit to include new observations, different fit coefficients may be provided to the classifier and land cover results may change for previously released dates. Land cover product values are determined in the same manner as in LCMAP Collection 1.1, described in Section 4. This means that product values have the potential to change, possibly as far back as the first year of the dataset, although changes in primary land cover are not common for the long-running segments that extend far back into the study period. On the other hand, short segments at the end of the time series in previous results (i.e., newly initialized, unbroken segments, and “end fits”) are more likely to change land cover values, due to the potential for more substantial changes in the coefficients that are provided to the classifier.

3.6 Classification Step-by-Step Walk-through

The following is an in-depth step-by-step walk-through of how the harmonic change segments of a pixel history characterized by PyCCD are used to classify the land cover. At a high level, there are two major procedures: model training and prediction.

3.6.1 Model Training

Trained classification models from LCMAP Collection 1.0 are used to maintain consistency. Classification models were developed for each tile, using data from that tile and the eight (or fewer) surrounding tiles, to allow for regional variations to be captured more accurately.

Classification training steps:

1. Gather the required data from the target tile and its surrounding neighbors:
    - Identify neighboring tile locations in addition to the target tile for gathering training data (Figure 3-3).
    - For each tile, gather the following data:
        - Cross-walked and eroded NLCD 2001
        - Ancillary layers
        - LCMAP CCD results
    - Filter the data by the following attributes:
        - Cross-walked and eroded NLCD 2001 is not the no-data value (0)
        - Pixels that have a CCD segment which encompasses the target training date (2001-07-01), as shown in Figure 3-2.
    - Extract relevant values from the CCD segments that encompass the target training date:
        - Per-band model coefficients
        - Per-band model RMSEs

2. Replace the per-band intercept values with the average annual reflectance value for the target training date, using Equation 5.
3. Randomly sample the filtered data points to generate the dataset to be trained on:
    - Target a sample size of 20 million samples.
    - Sample proportionately to the class size represented in the total training dataset for the nine-tile region.
    - Require no fewer than 600,000 and no more than 8 million samples.
4. Set the following hyperparameters for XGBoost (see https://xgboost.readthedocs.io/en/latest/parameter.html):
    - Maximum tree depth: 8
    - Tree method: fast histogram optimized approximate greedy algorithm
    - Evaluation metric: multiclass logloss
    - Maximum number of rounds: 500
5. Pass the following data values corresponding to the sampled pixels to classifier (XGBoost):
    - Cross-walked and eroded 2001 NLCD (2011 ed.)
    - Ancillary layers
    - Change detection per-band model coefficients (sans intercept), average reflectance values, and RMSEs
6. Save the resulting model.

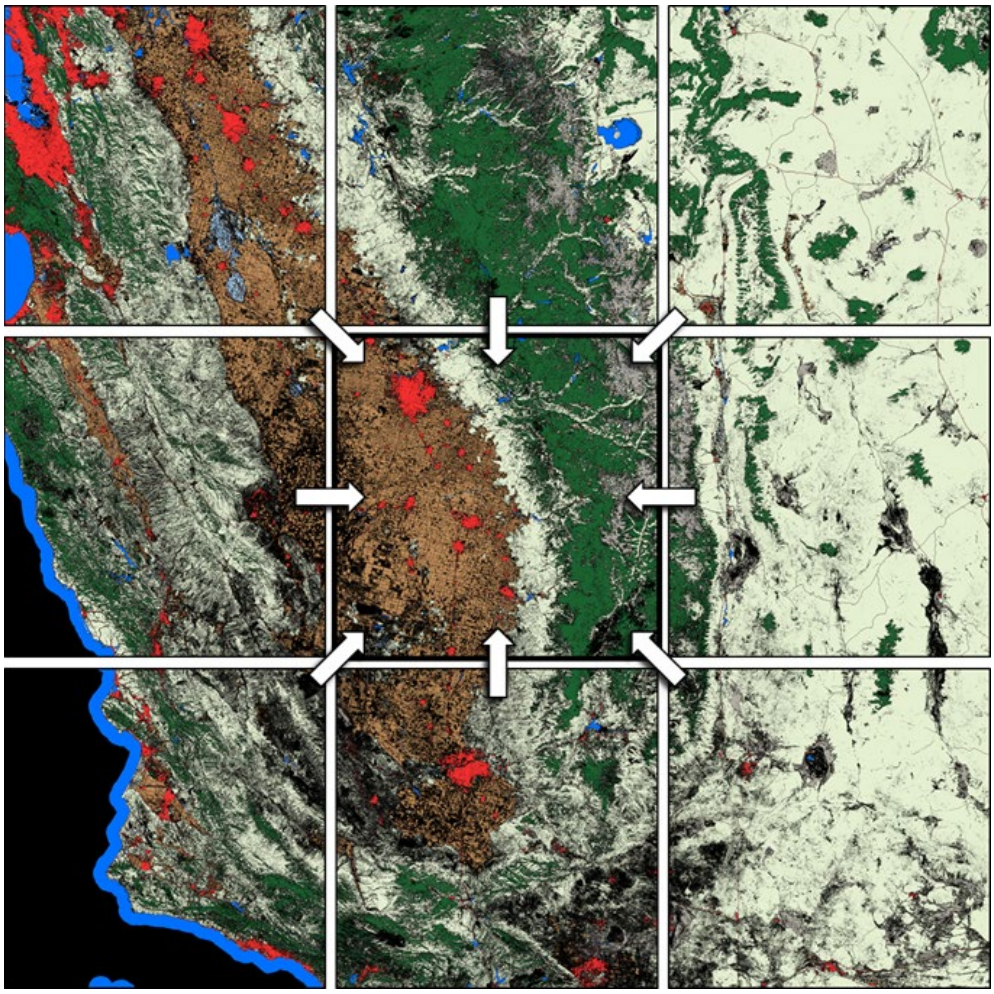


Figure 3-3. Information from 8 surrounding tiles are used to help inform classification modeling

3.6.2 Prediction

Once a trained classifier model is ready to be used, it can then be applied to all segments in the tile. This process is similar to the training step, except the 2001 NLCD layer is no longer needed, and now all July 1st dates that intersect a CCD segment can be classified (Figure 3-4). Unlike training, which was done per-tile using the neighboring tiles information, prediction is a single pixel, or time-series operation.

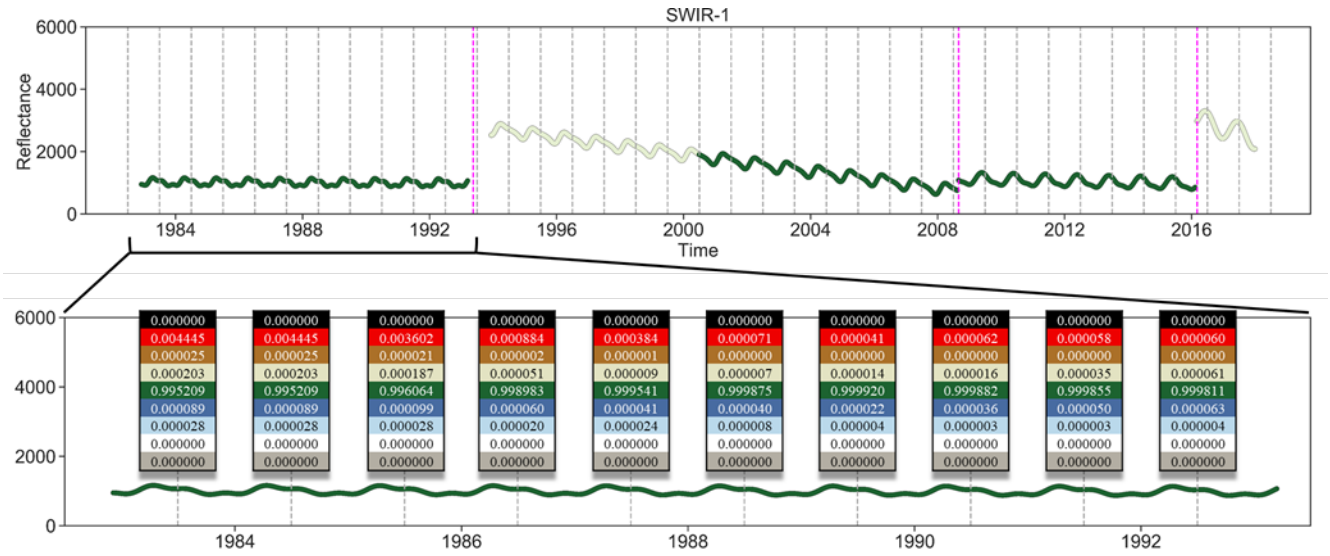


Figure 3-4. Predictions for First Segment in Example Time Series

Prediction steps:

1. For each pixel, gather the following data:
    - Ancillary layer values
    - LCMAP CCD results
    - Trained classifier associated with the tile containing the pixel
2. For each change detection model:
    a. Extract the following information:
        - Change detection per-band model coefficients
        - Change detection per-band model RMSEs
    b. For each annual July 1st data that the model crosses:
        - Replace the model intercept with a predicted average or overall reflectance value for each band, as per Equation 5.
        - Pass all information to the classifier:
            - Ancillary layer information
            - Change detection per-band model coefficients (sans intercept), average reflectance values, and RMSEs
3. Save the resulting prediction values.

3.7 Known Issues

- Training data were originally intended to be filtered to those pixels for which the PyCCD segment spans the two years around the July 1st, 2001, date, to ensure temporal consistency with NLCD land cover. Due to a programming error, the current implementation only requires a segment to intercept that target date. Analyses by the team did not find a net detrimental effect on the resulting products.

Section 4 Product Definitions

4.1 Description

This section gives detailed descriptions of the LCMAP Collection 1.2 Science Products and how the CCDC algorithm generates the products.

The CCDC approach provides a continuous capability to identify the state of land cover and land surface change at any point in the Landsat 4–8 temporal record. The LCMAP Science Products consists of five annual land cover and five annual land surface change products that are produced at an annual time step. Land cover products represent the annual status of each pixel on July 1st, a representative date of each year. Land surface change products provide additional information about spectral and temporal change (e.g., magnitude or date) that occurred during an annual period.

Land cover products:

- Primary Land Cover (LCPRI)
- Primary Land Cover Confidence (LCPCONF)
- Secondary Land Cover (LCSEC)
- Secondary Land Cover Confidence (LCSCONF)
- Annual Land Cover Change (LCACHG)

Land surface change products:

- Time of Spectral Change (SCTIME)
- Change Magnitude (SCMAG)
- Spectral Stability Period (SCSTAB)
- Time Since Last Change (SCLAST)
- Spectral Model Quality (SCMQA)

LCMAP Collection 1.2 Science Products are available for the CONUS from 1985–2020. Re-releases contain data products for all years covered by the dataset, and may contain different product values relative to a previous release for the same map year (especially for later years), for reasons described in Sections 2.5.6 and 3.5.4.

4.2 Definitions

Figures of abstracted surface reflectance time series are used here to clarify the methods used to produce LCMAP Science Products. For simplicity, harmonic regressions are shown without the original pixel observations. Figure 4-1 provides a legend for the pixel time series.

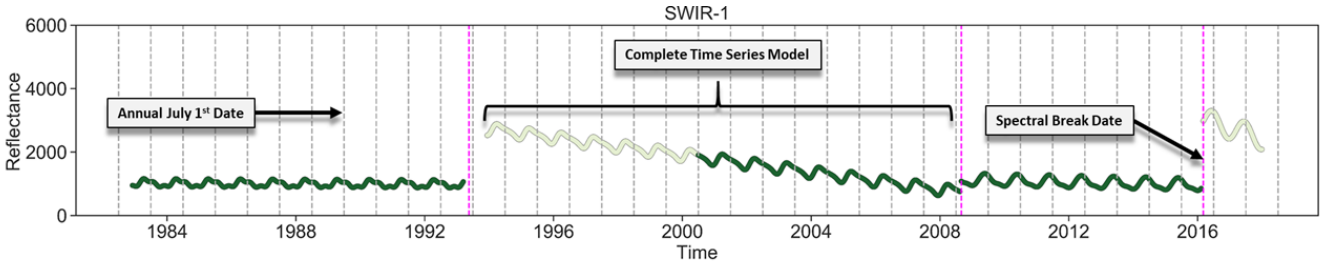


Figure 4-1. Legend for product definition examples, where green indicates time classified as Tree Cover, tan indicates time classified as Grass/Shrub, gray dashed lines are annual July 1st dates, and magenta dashed lines are spectral breaks

4.2.1 Time of Spectral Change (SCTIME)

Time of Spectral Change represents the timing of a spectral change within the current product year as the Day of Year (DOY) the change occurred (Figure 4-2). A spectral change is defined as a "break" in a CCDC time series model where spectral observations have diverged from the model predictions. These breaks may be indicative of a change in thematic land cover (e.g., fire, urbanization) or may represent more subtle conditional surface changes (e.g., forest growth, insect infestation). A value of zero indicates there was no recorded model break in the current year. SCTIME and SCMAG are both characteristics of the same model breaks and will always be coincident in space and time. An example Python function for generating Time of Spectral Change values is provided in Figure 4-3.

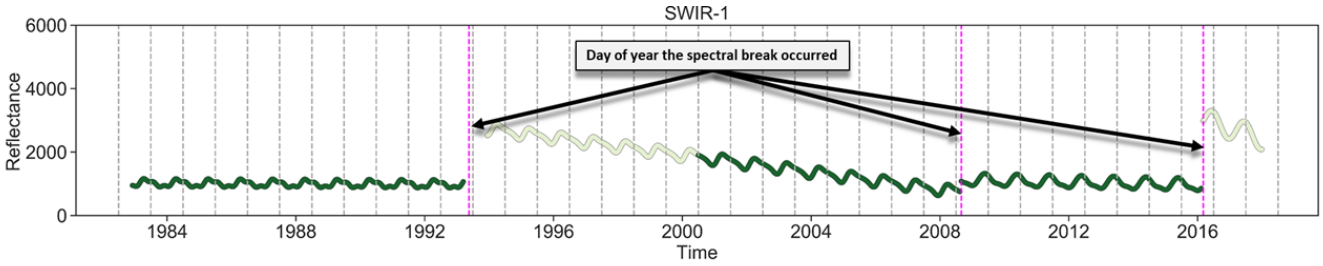


Figure 4-2. The DOY a spectral change caused a "break" in a CCDC time series model. When a spectral change occurs, causing a divergence from model predictions, a new time series model begins.

```python
import datetime as dt
from typing import List

def spectral_breaks(pyccd_result: dict) -> List[dt.datetime]:
    """Example pseudo function for generating the dates
    that a pixel had spectral breaks."""
    break_models = filter(lambda x: x['change_probability'] == 1,
                          pyccd_result['change_models'])

    dates = []
    for model in break_models:
        dates.append(dt.datetime.fromordinal(model['break_day']))

    return dates
```

Figure 4-3. Example Python Function for Time of Spectral Change

4.2.2 Change Magnitude (SCMAG)

Change Magnitude provides information on the spectral strength or intensity of a time series model break when spectral observations have diverged from the model predictions (Figure 4-4). Change Magnitude is calculated as the square root of the sum of the squared per-band median residuals (excluding the blue and BT bands) between the observed per-band Landsat surface reflectance (scaled) and CCDC predictions at the time of a detected CCDC model break. Change Magnitude is unitless and generally ranges between 1–10,000. A value of zero indicates there is no recorded model break in the current year. An example Python function for generating Change Magnitude values is provided in Figure 4-5.

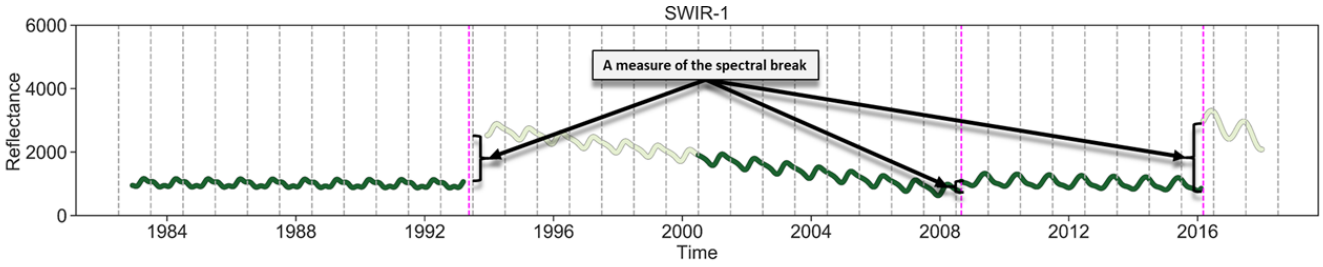


Figure 4-4. Change Magnitude when Spectral Break Occurs

```python
import datetime as dt
from typing import List
import numpy as np

def change_magnitudes(pyccd_result: dict) -> List[float]:
    """Example pseudo function for generating the associated
    change magnitude values for a pixel's spectral breaks"""
    break_models = filter(lambda x: x['change_probability'] == 1,
                          pyccd_result['change_models'])

    magnitudes = []
    for model in break_models:
        mag = np.linalg.norm([model['green']['magnitude'],
                              model['red']['magnitude'],
                              model['nir']['magnitude'],
                              model['swir1']['magnitude'],
                              model['swir2']['magnitude']])
        magnitudes.append(mag)

    return magnitudes
```

Figure 4-5. Example Python Function for Change Magnitude

4.2.3 Spectral Stability Period (SCSTAB)

Spectral Stability Period is a measure of the amount of time in days a pixel has been in its current spectral state. Spectral state can refer to a pixel's state during a time series model (Figure 4-6), and the temporal ranges between time series models (Figure 4-7). An example Python function for generating Spectral Stability values is provided in Figure 4-8.

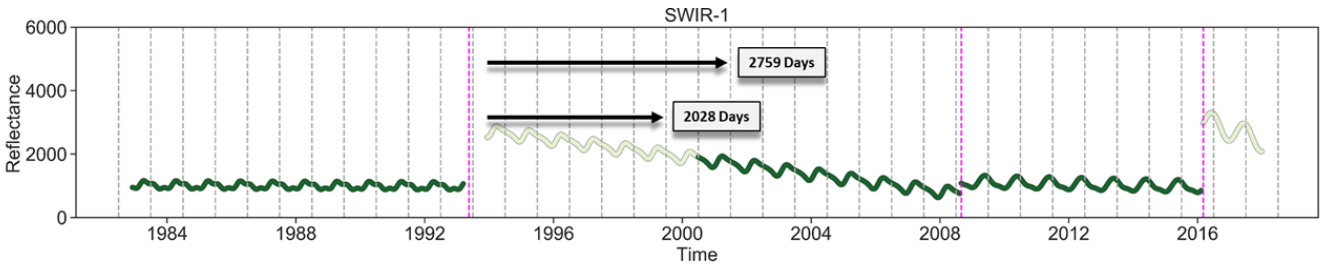


Figure 4-6. Spectral Stability Period during a time series model starting on July 1st start date for product year

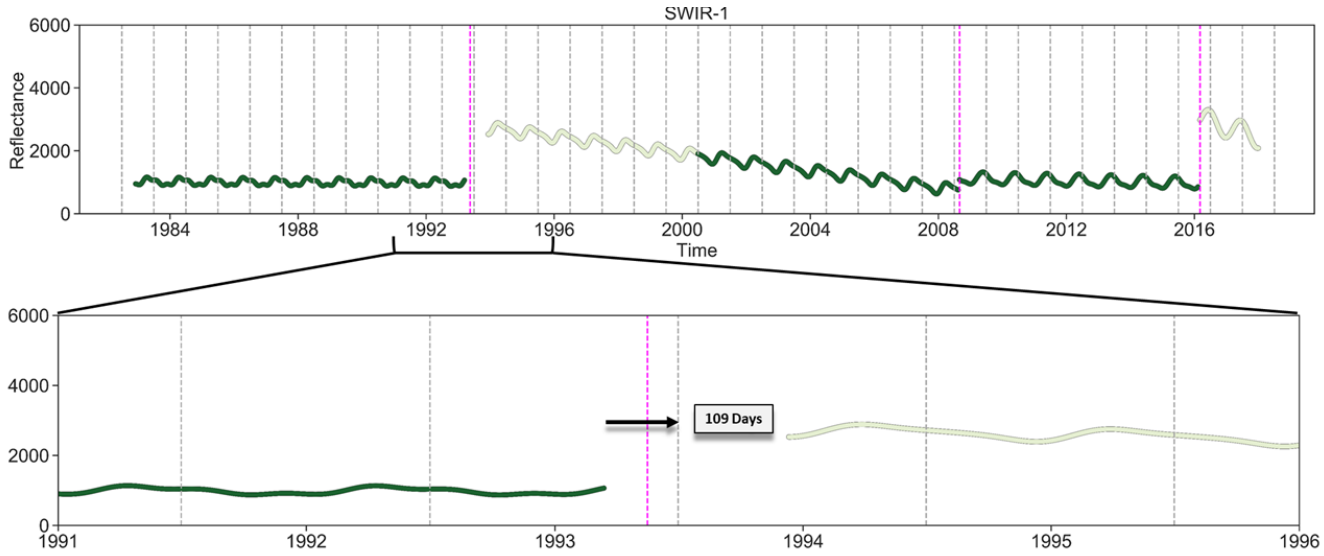


Figure 4-7. Spectral Stability Period between CCDC time series models, beginning on the July 1st start date for product year to the most recent harmonic segment characterized

```
def stability(pyccd_result: dict, ordinal: int,
              begin_ordinal: int) -> int:
    """Example pseudo function for how long, in days, has the
    pixel been in its current state, where begin_ordinal is
    the first day to begin counting"""
    diff = [ordinal - begin_ordinal]

    for model in pyccd_result['change_models']:
        if ordinal > model['end_day']:
            diff.append(ordinal - model['end_day'])
        else:
            diff.append(ordinal - model['start_day'])

    return min(filter(lambda x: x >= 0, diff), default=0)
```

Figure 4-8. Example Python Function for Spectral Stability

4.2.4 Time Since Last Change (SCLAST)

Time Since Last Change represents the time, in days, from July 1st of the current product year back to the most recent time series model break where spectral observations diverged from CCDC model predictions (Figure 4-9). This can also be expressed as the time, in days, since the last recorded result in both Time of Spectral Change and Change Magnitude. An example Python function for generating Time Since Last Change values is provided in Figure 4-10.

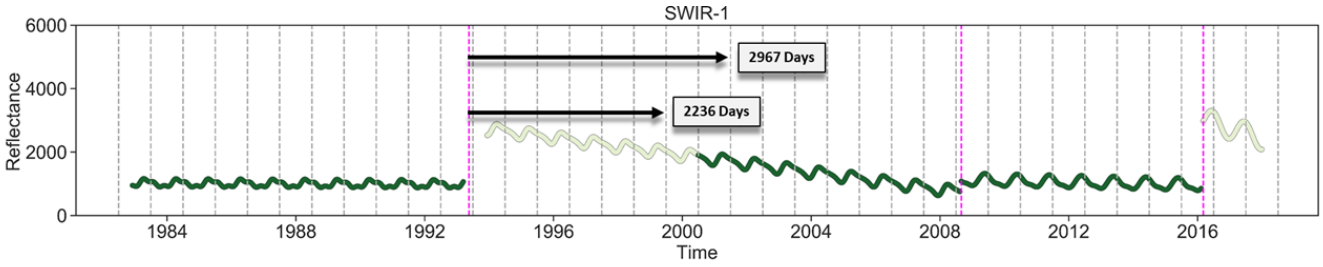


Figure 4-9. Time Since Last Change measured from July 1st start date to DOY of last time series model break

```
def last_change(pyccd_result: dict, ordinal: int) -> int:
    """How long ago, in days, was the last spectral break
    from the given date"""
    break_models = filter(lambda x: x['change_probability'] == 1,
                          pyccd_result['change_models'])

    diff = []
    for model in break_models:
        diff.append(ordinal - model['break_day'])

    return min(filter(lambda x: x >= 0, diff), default=0)
```

Figure 4-10. Example Python Function for Time Since Last Change

4.2.5 Spectral Model Quality (SCMQA)

Spectral Model Quality provides additional information regarding the type of time series model available in the current product year (Figure 4-11). SCMQA reflects the type of time series model present on July 1st of the current year and can be useful for interpreting results in other LCMAP Science Products. SCMQA pixel values and descriptions are provided in Table 4-1. An example Python function for generating Spectral Model Quality values is provided in Figure 4-12.

Pixel Value	Model Type	Description
0	No Model	No model established for July 1st of current year.
4	Simple Model	A partial, 4-coefficient harmonic model.
6	Advanced Model	A partial, 6-coefficient harmonic model.
8	Full Model	A full, 8-coefficient harmonic model.
14	Start Fit	A simple model at the beginning of a time series where sparse and/or highly variable spectral measurements prevent establishment of a harmonic model.
24	End Fit	A simple model at the end of a time series where there are insufficient observations and/or time to establish a new harmonic model following a model break.

Pixel Value	Model Type	Description
44	Insufficient Clear	A simple model for the entire time series in cases where fewer than 25% of input observations are labeled as “Clear” or “Water” by the Landsat ARD per-pixel quality band (PIXELQA).
54	Persistent Snow	A simple model for the entire time series in cases where 75% or more of input observations are labeled as “Snow” by the Landsat ARD per-pixel quality band (PIXELQA).

Table 4-1. Description of Pixel Values for Spectral Model Quality

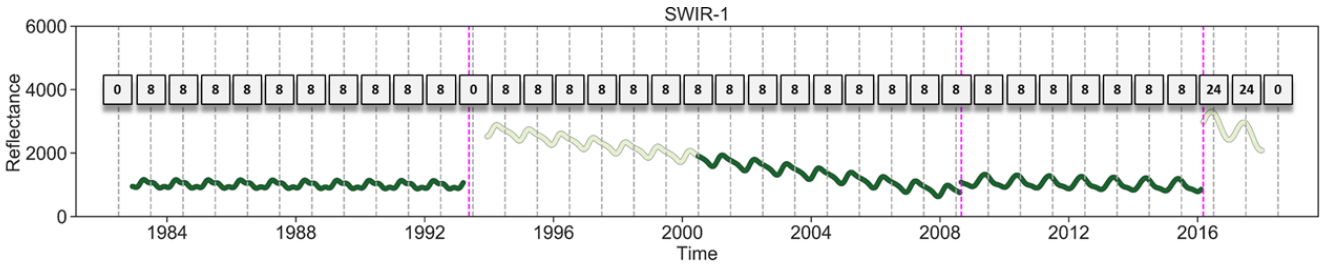


Figure 4-11. Spectral Model Quality providing information regarding type of time series model available for each annual July 1st

```
def model_quality(pyccd_result: dict, ordinal: int) -> int:
    """Example pseudo function for determining the Spectral
    Model Quality value for a given date"""
    for model in pyccd_result['change_models']:
        if model['start_day'] <= ordinal <= model['end_day']:
            return model['curve_qa']

    return 0
```

Figure 4-12. Example Python Function for Determining Spectral Model Quality

4.2.6 Land Cover (LCPRI and LCSEC) and Land Cover Confidence (LCPCONF and LCSCONF)

Primary and Secondary Land Cover classifications and the associated Primary and Secondary confidence values are determined through a combination of three approaches: initial classifier, secondary analysis, and rule-based assignment.

The result of applying any of these methods, which are described in detail below, is a land cover value from 1–8 for the LCMAP primary and secondary land cover class (see Table 3-1). If the assigned land cover value is derived directly from the initial classifier, confidence values will be in the range 1–100. If from secondary analysis of trends, values will be either 151 or 152. Finally, rule-based assignments will be values of 201–214 (see Table 4-2 at the end of this section). Only values from 1–100 represent a continuous confidence score (low confidence to high confidence); values greater than 100 are categorical indicators.

Example plots of a time series showing Primary Land Cover, Secondary Land Cover, Primary Land Cover Confidence, and Secondary Land Cover Confidence are provided in Figure 4-13 to Figure 4-16. As can be seen from the confidence codes, this example time series contains three segments (first, third, and fourth) classified across multiple years using the initial classifier, one segment (second) classified as a Grass/Shrub to Tree Cover transition by secondary analysis, and several individual years (1982, 1993, and 2018) whose July 1st date does not intersect with a time series model. These latter values are classified by rule-based assignment.

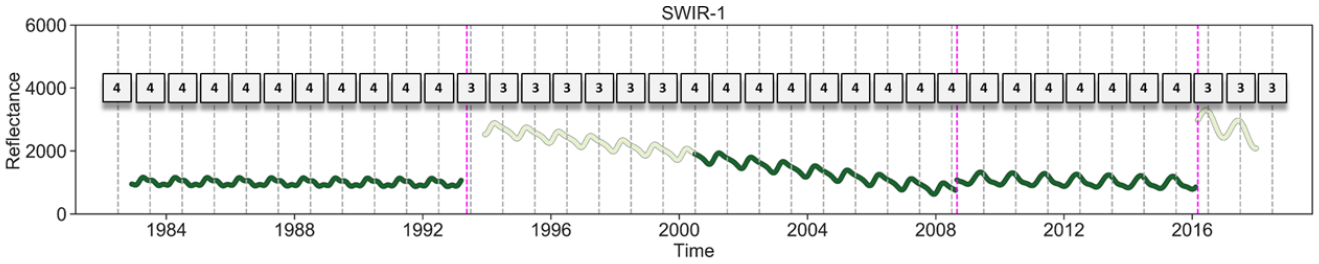


Figure 4-13. Primary Land Cover Values for Example Time Series

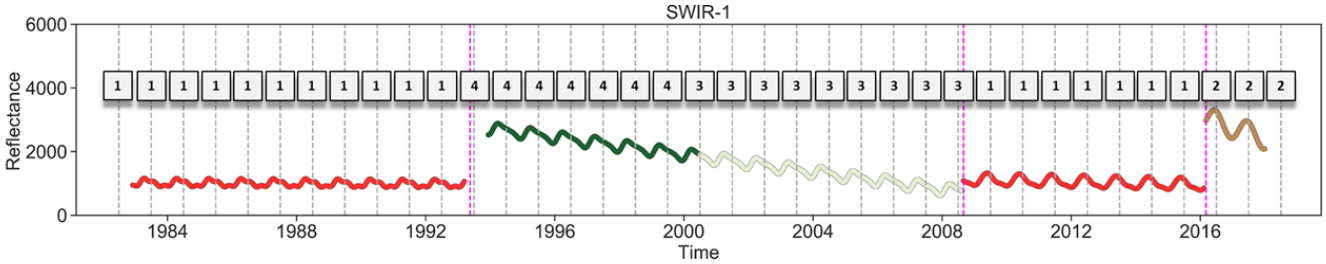


Figure 4-14. Secondary Land Cover Values for Example Time Series

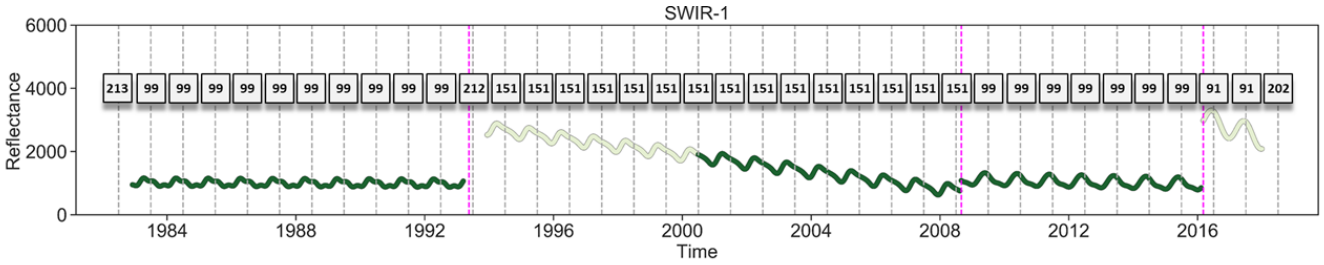


Figure 4-15. Primary Land Cover Confidence Values for Example Time Series

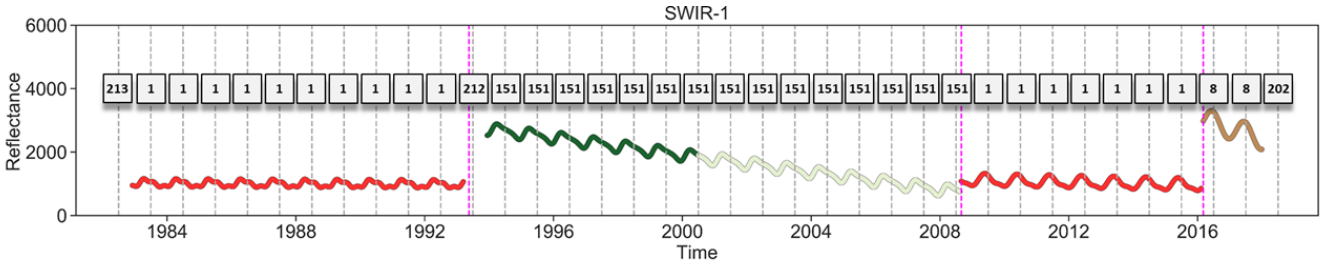


Figure 4-16. Secondary Land Cover Confidence Values for Example Time Series

4.2.6.1 Initial Classifier

Land cover is determined by calculating a per-class mean across all the predictions for a time series model (Figure 4-17). The land cover classes with the highest and second highest confidence values become the primary and secondary land cover, while the associated confidence values are used to determine primary and secondary confidence. The 0.0 to 1.0 floating point mean confidence value is scaled by 100 and converted to an integer (floor with a minimum value of 1) for the LCPCONF and LCSCONF product values. For the first time series model shown in Figure 4-17, mean annual prediction confidence values favor Tree Cover (primary land cover class code 4, primary confidence 99), while secondary cover and confidence are assigned 1 (for the Developed class) and a confidence value that is (coincidentally) also 1 (low confidence).

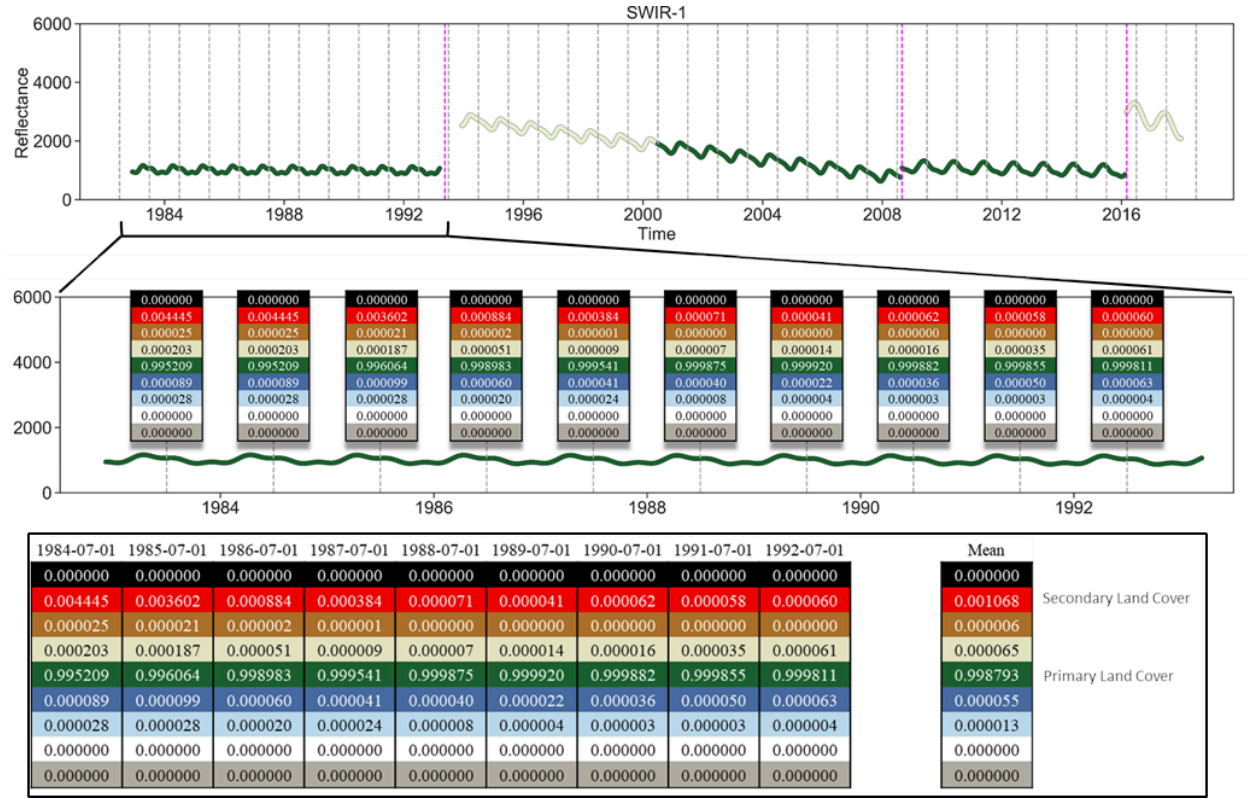


Figure 4-17. Example of how land cover is determined from initial classifier for first time series model

4.2.6.2 Secondary Analysis

Grass/Shrub to Tree Cover and Tree Cover to Grass/Shrub transitions are handled by looking at the first and last set of predictions for the time series model, along with the average reflectance values at the beginning and end of the model for a specific combination of spectral bands. Specifically, we define a normalized band ratio at ordinal date t using the NIR and SWIR-1 bands, as follows:

$$BR(t) = \left(\frac{avg\_refl_{nir}(t) - avg\_refl_{swir1}(t)}{avg\_refl_{nir}(t) + avg\_refl_{swir1}(t)}\right) \quad (6)$$

where,

$avg\_refl_{nir}(t)$	the average reflectance value (Equation 5) for the NIR band at ordinal date t
$avg\_refl_{swir1}(t)$	the average reflectance value (Equation 5) for the SWIR-1 band at ordinal date t
$BR(t)$	the predicted normalized ratio at ordinal date t

The differenced equation between start and end dates is defined as:

$$\Delta BR = BR_{end\_date} - BR_{start\_date}$$

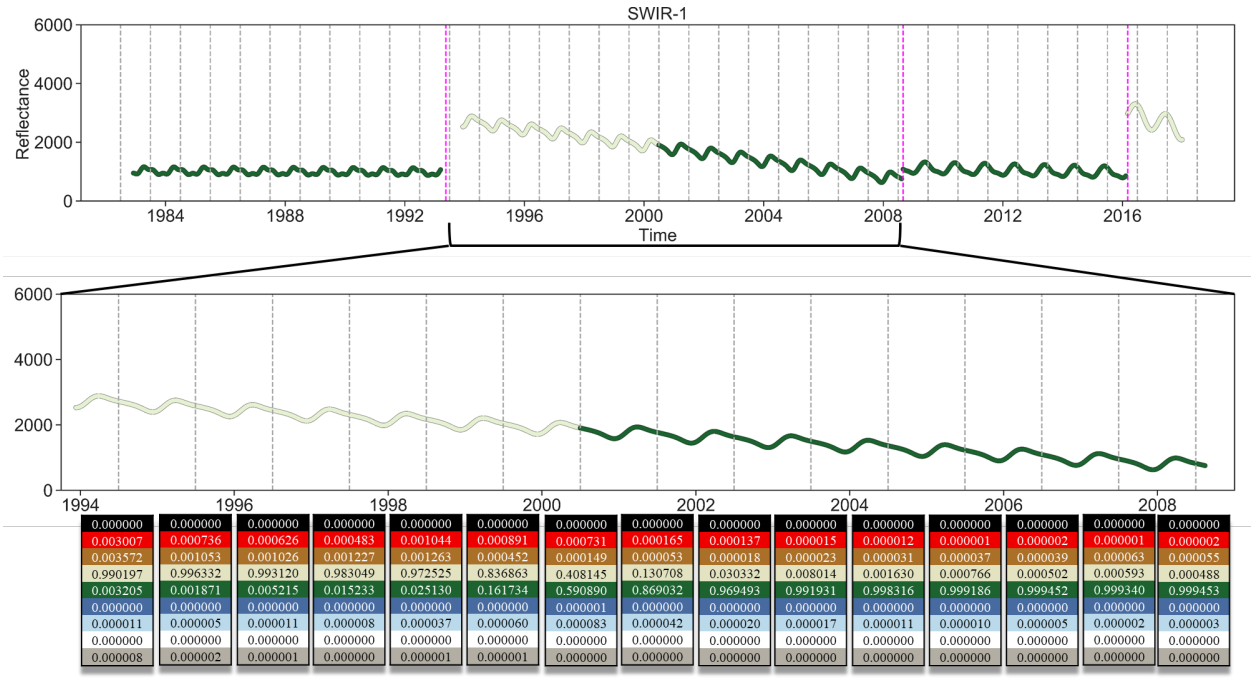
where,

BR_{end_date}	Equation 6 based on the end date for the time series model
BR_{start_date}	Equation 6 based on the start date for the time series model
ΔBR	the differenced normalized band ratio

A segment is considered a Grass/Shrub to Tree Cover transition segment if the first annual prediction from the classifier indicates Grass/Shrub, the last indicates Tree Cover, and $\Delta BR > 0.05$. In these cases, all dates for the time series model before the first annual prediction indicating Tree Cover are assigned to Grass/Shrub, and that date and all subsequent dates covered by the model are assigned to Tree Cover. For these segments, land cover confidence (both primary and secondary) is assigned the value 151.

A more detailed view of the example transition segment from the plots above is provided in Figure 4-18. Starting with the July 1st, 2001, date, land cover is assigned to Tree Cover, while prior years' land cover is assigned to Grass/Shrub. Note that the land cover change occurs without a break in the time series model, a unique feature of transition segments.

For Tree Cover to Grass/Shrub transitions, the reverse set of conditions apply: the first annual prediction must indicate Tree Cover, the last Grass/Shrub, and we require $\Delta BR < 0.05$. For these cases, the primary and secondary land cover confidence is set to 152.



0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.003007	0.000736	0.000626	0.000483	0.001044	0.000891	0.000731	0.000165	0.000137	0.000015	0.000012	0.000001	0.000002	0.000001	0.000002
0.003572	0.001053	0.001026	0.001227	0.001263	0.000452	0.000149	0.000053	0.000018	0.000023	0.000031	0.000037	0.000039	0.000063	0.000055
0.990197	0.996332	0.993120	0.983049	0.972525	0.836863	0.408145	0.130708	0.030332	0.008014	0.001630	0.000766	0.000502	0.000593	0.000488
0.003205	0.001871	0.005215	0.015233	0.025130	0.161734	0.590890	0.869032	0.969493	0.991931	0.998316	0.999186	0.999452	0.999340	0.999453
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000001	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000011	0.000005	0.000011	0.000008	0.000037	0.000060	0.000083	0.000042	0.000020	0.000017	0.000011	0.000010	0.000005	0.000002	0.000003
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000008	0.000002	0.000001	0.000000	0.000001	0.000001	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

Figure 4-18. Example Predictions for Second Time Series Model in Example Pixel

4.2.6.3 Rule-Based Assignment

For rule-based assignment, land cover is determined by interpolation or extrapolation of other information available in the time series. This is used when trying to determine land cover that is not immediately attributable to a time series model. This happens either between time series models, at the ends of the time series, or if there are not enough observations to build a time series model.

The following describes how different scenarios are handled:

- If there is no stable time series model available for the entire time series (i.e., in cases with few observations):
  - Use a cross-walked NLCD 2001 (2011 Edition) for both primary and secondary land cover for all dates.
  - Assign value 201 to land cover confidence (primary and secondary).

- If the date falls before all time series models start dates (e.g., July 1st, 1982, from Figure 4-18):
  - Assign the primary and secondary land cover class from the subsequent time series model.
  - Assign value 213 to land cover confidence (primary and secondary).

- If the date falls after all time series models end dates (e.g., July 1st, 2018, from Figure 4-18), and the final time series model did not end with a spectral break:
  - Assign the last identified primary and secondary land cover class.
  - Assign value 202 to land cover confidence (primary and secondary).

- If the date falls after all time series models end dates, and the final time series model did end with a spectral break:
    - Assign the last identified primary and secondary land cover class.
    - Assign value 214 to land cover confidence (primary and secondary).

- If the date falls in a gap between two time series models (for example, July 1st, 1993, from Figure 4-18), then for primary and secondary land cover *separately*:
    - If the surrounding time series models indicate the same land cover:
        - Assign the common land cover value.
        - Assign value 211 to land cover confidence.
    - If the surrounding time series models indicate differing land cover:
        - Use the break date between the two time series models to interpolate the land cover values (Figure 4-19):
            - Before the break date, assign the previous land cover value.
            - At or after the break date, assign the subsequent land cover value.
        - Assign value 212 to land cover confidence.

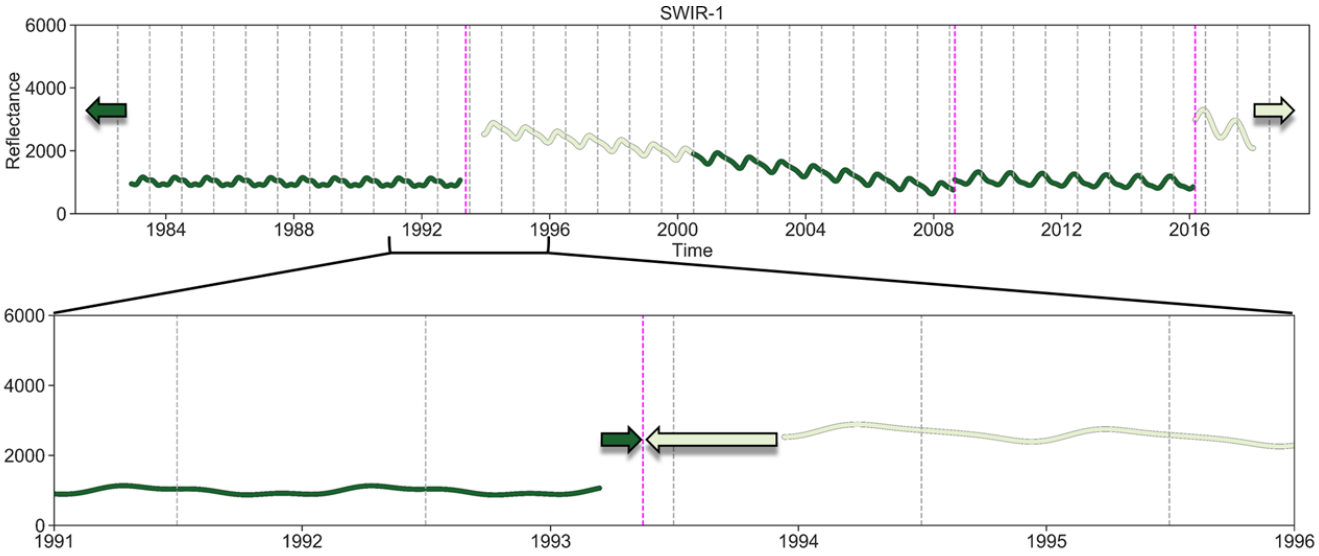


Figure 4-19. Land cover classes are temporally interpolated between time series models, and extrapolated at ends of time series

Pixel Value	LC Label Source	Description
1-100	Initial classifier	Measure of confidence that the Primary Land Cover label matches the training data.
151	Secondary analysis	Time series model identified as transition from a Grass/Shrub class to a Tree Cover class. Primary Land Cover class assignment based on secondary analysis.

Pixel Value	LC Label Source	Description
152	Secondary analysis	Time series model identified as transition from a Tree Cover class to a Grass/Shrub class. Primary Land Cover class assignment based on secondary analysis.
201	Rule-based	No stable time series models were produced for this location. Primary Land Cover assigned the land cover class present in NLCD-2001 (cross-walked to LCMAP Level-1 classification schema, *see Table 3-2*).
202	Rule-based	Insufficient data available to extend most recent time series model past July 1st of current year. Primary Land Cover assigned the last identified cover class from earlier year.
211	Rule-based	July 1st falls in a gap between two stable time series models of the same land cover class. Primary Land Cover assigned the land cover class of those before/after models.
212	Rule-based	July 1st falls in a gap between two stable time series models of differing land cover class. If July 1st is before the "break date" of the earlier model, Primary Land Cover is assigned the land cover class of that earlier model. Otherwise, Primary Land Cover is assigned the land cover class of the subsequent, later model.
213	Rule-based	Insufficient data available to establish a stable time series model at the beginning of the time series prior to July 1st of the current year. Primary Land Cover assigned the land cover class of 1st subsequent model.
214	Rule-based	Insufficient data available to establish a new stable time series model following a break near the end of the time series prior to July 1st of the current year. Primary Land Cover assigned the last identified cover class from earlier year.

Table 4-2. Description of Pixel Values for Primary and Secondary Land Cover Confidence

4.2.7 Annual Land Cover Change (LCACHG)

Annual Land Cover Change is a synthesis product derived from the Primary Land Cover (LCPRI) of the current product year and the LCPRI of the previous year (Figure 4-20). Values 1–8 correspond to the integer values representing land cover classes presented in Table 3-1 and are assigned to pixels if no change was identified between the two years. If change between years was identified, the resulting two-digit pixel value is a concatenation of the previous & current land cover class code (based on the values in Table 3-1). For example, a pixel value of 18 is a concatenation of a one (1), representing the Developed class, and an eight (8), representing the Barren class. Therefore, the resulting value of 18 represents a change from Developed to Barren. Table 4-3 provides a partial list of these concatenated change codes as examples along with the land cover changes they represent.

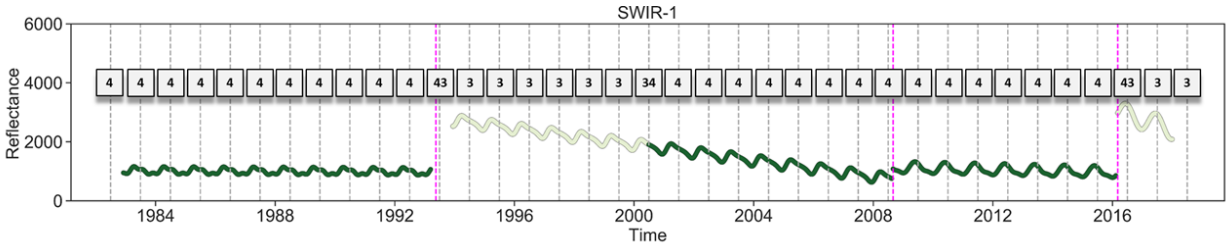


Figure 4-20. Annual Land Cover Change Values for Example Time Series

Pixel Value	Previous Land Cover Class	Current Land Cover Class	Land Cover Change
1-8	*(as Table 3-1)*	*(as Table 3-1)*	No change
18	Developed	Barren	Developed to Barren
21	Cropland	Developed	Cropland to Developed
32	Grass/Shrub	Cropland	Grass/Shrub to Cropland
42	Tree Cover	Cropland	Tree Cover to Cropland
43	Tree Cover	Grass/Shrub	Tree Cover to Grass/Shrub
65	Wetlands	Water	Wetlands to Water
73	Snow/Ice	Grass/Shrub	Snow/Ice to Grass/Shrub

Note: Pixel Value examples in table are not all-inclusive.

Table 4-3. Description of Pixel Values for Land Cover Change

Appendix A Acronyms

ADD	Algorithm Description Document
AEA	Albers Equal Area
ARD	Analysis Ready Data
BT	Brightness Temperature
CCB	Configuration Control Board
CCD	Continuous Change Detection
CCDC	Continuous Change Detection and Classification
CONUS	Conterminous United States
CR	Change Request
DEM	Digital Elevation Model
DFCB	Data Format Control Book
DOI	Department of the Interior
DOY	Day of Year
EROS	Earth Resources Observation and Science Center
ETM+	Enhanced Thematic Mapper Plus
INT	Signed Integer
LASSO	Least Absolute Shrinkage and Selection Operator
LCACHG	Annual Land Cover Change
LCMAP	Land Change Monitoring, Assessment, and Projection
LCPCONF	Primary Land Cover Confidence
LCPRI	Primary Land Cover
LCSCONF	Secondary Land Cover Confidence
LCSEC	Secondary Land Cover
LSDS	Land Satellites Data System
MEOW	Minimum Expected Observation Window
NED	National Elevation Dataset
NIR	Near Infrared
NLCD	National Land Cover Database
NWI	National Wetlands Inventory
OLI	Operational Land Imager
OLS	Ordinary Least Square
PIXELQA	Level 2 Pixel Quality Assessment Band
PyCCD	Python-based CCD
QA	Quality Assessment
RIRLS	Robust Iteratively Reweighted Least Squares
RMSE	Root Mean Square Error
SCLAST	Time Since Last Change
SCMAG	Change Magnitude
SCMQA	Spectral Model Quality

SCSTAB	Spectral Stability Period
SCTIME	Time of Spectral Change
SR	Surface Reflectance
SSURGO	Natural Resources Conservation Service Soil Survey Geographic Database
SWIR-1	Shortwave Infrared Band 1
SWIR-2	Shortwave Infrared Band 2
TIRS	Thermal Infrared Sensor
TM	Thematic Mapper
Tmask	multiTemporal Mask
UINT	Unsigned Integer
USDA	U.S. Department of Agriculture
USGS	U.S. Geological Survey
WPI	Wetland Potential Index
XGBoost	eXtreme Gradient Boosting

Appendix B Default CCD Parameters

Parameter	Default Value	Description
$MEOW\_SIZE$	12	Minimum Expected Observation Window; minimum number of observations required to build a harmonic model.
$PEEK\_SIZE$	6	Minimum number of observations to use to evaluate if a change has occurred in the time series; may be adjusted by the algorithm.
DAY_DELTA	365	Minimum time span, in days, required to build a harmonic model by the standard procedure.
AVG_DAYS_YR	365.2425	Length (in days) of an annual period to use for the lowest order harmonic.
$COEFFICIENT\_MIN$	4	Number of model coefficients to fit, chosen based on the number of observations
$COEFFICIENT\_MID$	6	
$COEFFICIENT\_MAX$	8	
NUM_OBS_FACTOR	3	Factor determining the number of coefficients to use for the model; coefficients above $COEFFICIENT\_MIN$ require at least $NUM\_OBS\_FACTOR \times COEFFICIENT\_(MID, MAX)$ observations
$BLUE\_IDX$	0	Indices of spectral and QA bands within observation input sequence
$GREEN\_IDX$	1	
RED_IDX	2	
NIR_IDX	3	
$SWIR1\_IDX$	4	
$SWIR2\_IDX$	5	
$THERMAL\_IDX$	6	
QA_IDX	7	
$DETECTION\_BANDS$	[1,2,3,4,5]	Spectral band indices used to detect change
$TMASK\_BANDS$	[1,4]	Spectral band indices used by the Tmask algorithm for outlier removal during model initialization
$QA\_BITPACKED$	True	Boolean value indicating if the QA values are bit-packed (stored as bit flags)
QA_FILL	0	ARD PIXELQA bit-pack offsets
QA_CLEAR	1	
QA_WATER	2	
QA_SHADOW	3	
QA_SNOW	4	
QA_CLOUD	5	

$QA\_CIRRUS1$	8	
$QA\_CIRRUS2$	9	
$QA\_OCCLUSION$	10	
$CURVE\_QA: PERSISTENT\_SNOW$	54	PyCCD model output QA codes for special fits
$CURVE\_QA: INSUF\_CLEAR$	44	
$CURVE\_QA: START$	14	
$CURVE\_QA: END$	24	
$CLEAR\_PCT\_THRESHOLD$	0.25	The ratio of clear or water QA values to other values, excluding fill, in a time series of observations, at or above which the standard model procedure shall be used
$SNOW\_PCT\_THRESHOLD$	0.75	The ratio of snow QA values to clear or water values, in a time series of observations, at or above which the permanent snow procedure shall be used in lieu of the insufficient clear procedure, when the clear threshold check is failed
$OUTLIER\_THRESHOLD$	35.888186879610423	Threshold above which a "change magnitude" (measure of deviation from model prediction across all detection bands) will be ruled an outlier and the observation excluded from processing
$CHANGE\_THRESHOLD$	15.086272469388987	Threshold "change magnitude" value that all observations within a window must exceed for a positive change detection; may be adjusted by the algorithm
T_CONST	4.89	Value by which to scale the calculated band variability in determining outlier threshold
$MEDIAN\_GREEN\_FILTER$	400	Value added to median value of $GREEN\_IDX$ band, used as a threshold to filter out additional observations with equal or higher values in this band within the insufficient clear procedure
$FITTER\_FN$	'ccd.models.lasso.fitted_model'	String name of the function used to find the best-fit between time series models and observations
$LASSO\_MAX\_ITER$	1000	Within the LASSO function, the maximum number of iterations to undergo in order to find convergence

Table B-1. Default CCD Parameters

References

Anderson, J.R., Hardy, E.E., Roach, J.T., and Witmer, R.E. (1976). A land use and land cover classification system for use with remote sensor data. Professional Paper 964, U.S. Geological Survey. Washington, DC: U.S. Government Printing Office. https://doi.org/10.3133/pp964

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794. https://doi.org/10.1145/2939672.2939785

Dwyer, J., Roy, D., Sauer, B., Jenkerson, C., Zhang, H., and Lymburner, L. (2018). Analysis Ready Data: Enabling analysis of the Landsat archive. Remote Sensing. https://doi.org/10.3390/rs10091363

Fry, J.A., Xian, G., Jin, S.M., Dewitz, J.A., Homer, C.G., Yang, L.M., Barnes, C.A., Herold, N.D. and Wickham, J.D. (2011). Completion of the 2006 National Land Cover Database for the conterminous United States. PE&RS, Photogrammetric Engineering & Remote Sensing, 77(9), pp.858-864.

Gesch, D., Oimoen, M., Greenlee, S., Nelson, C., Steuck, M. and Tyler, D. (2002). The National Elevation Dataset. Photogrammetric Engineering & Remote Sensing, 68(1), pp.5-32.

Homer, C., Dewitz, J., Yang, L., Jin, S., Danielson, P., Xian, G., Coulston, J., Herold, N., Wickham, J. and Megown, K. (2015). Completion of the 2011 National Land Cover Database for the conterminous United States–representing a decade of land cover change information. Photogrammetric Engineering & Remote Sensing, 81(5), pp.345-354.

LSDS-1873. U.S. Landsat Collection 1 (C1) Analysis Ready Data (ARD) Data Format Control Book (DFCB) https://www.usgs.gov/media/files/landsat-analysis-ready-data-ard-data-format-control-book-dfcb

Zhou, Q., Tollerud, H., Barber, C., Smith, K., and Zelenak, D. (2020). Training data selection for annual land cover classification for the LCMAP initiative, Remote Sensing 12(4), 699. https://doi.org/10.3390/rs12040699

Zhu, Z., and Woodcock, C.E. (2014a). Continuous change detection and classification of land cover using all available Landsat data: Remote Sensing of Environment 144: 152–171. https://doi.org/10.1016/j.rse.2014.01.011

Zhu, Z., and 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 Sensing of Environment 152: 217-234. https://doi.org/10.1016/j.rse.2014.06.012

Zhu, Z., and Woodcock, C.E. (2012). Object-based cloud and cloud shadow detection in Landsat Imagery, Remote Sensing of Environment 118: 83-94. https://doi.org/10.1016/j.rse.2011.10.028

Zhu, Z., Gallant, A.L., Woodcock, C.E., Pengra, B., Olofsson, P., Loveland, T.R., Jin, S., Dahal, D., Yang, L., and Auch R.F. (2016). Optimizing selection of training and auxiliary data for operational land cover classification for the LCMAP initiative, ISPRS Journal of Photogrammetry and Remote Sensing 122: 206-221. https://doi.org/10.1016/j.isprsjprs.2016.11.004

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 Sensing of Environment 162: 67-83. https://doi.org/10.1016/j.rse.2015.02.009