Methods for Cloud- and Gap-Filling Landsat Data Using Regression Trees

Eric Brown de Colstoun[1,3] and Jim Storey[2,3]

[1] Science Systems and Applications, Inc., Lanham, MD email: ericbdc@ltpmail.gsfc.nasa.gov

[2] Science Applications International Corporation, USGS National Center for EROS, Sioux Falls, SD

[3] Biospheric Sciences Branch, NASA/Goddard Space Flight Center, Greenbelt, MD

Abstract- *We have examined the performance of the CUBIST regression tree in filling cloud covered and/or areas affected by the Scan-Line Corrector (SLC) failure on Landsat Imagery. Tests were performed on two terrain- and atmospherically-corrected yet cloudy Landsat 5 scenes and a subset of three L7 images each acquired a month apart over the Upper Delaware River Basin. Regression tree results are within 0.5% reflectance in the visible wavelengths, between 1.5 and 2.7% for Landsat Band 4, and between 0.6 and 1.4% reflectance in Bands 5 and 7. Visual examination of the regression-filled images shows that, IF clouds and shadows in the input scenes can be accurately detected before processing of the data, regression treeS are an effective tool to mitigate not only the image gaps due to the SLC failure, but also clouds and cloud shadows.*

Background

- On May 31, 2003, an instrument malfunction occurred onboard Landsat 7.
- Without a functioning Scan-Line Corrector (SLC), approximately 22% of each scene is lost.
- This has significant implications for land cover/use monitoring activities with Landsat 7.
- Gap-Filled data are available from USGS National Center for EROS based on multi-scene histogram matching algorithms.

Regression Trees

- During Phase I of SLC mitigation activities, we found that regression trees offered advantages over other gap-filling methods:

  -Could use multiple scenes as opposed to one.

  -Regressions could focus on specific land cover "clusters".

  -Could potentially describe some land cover changes in scenes.

  -Methods outperformed even current histogram matching algorithm.

Divide and Conquer approach to handle non-linearity and high order interactions+: Cubist* example

```
 Rule 1/1: [10230 cases, mean 85.3, range 60 to 122, est err 4.1]

   if
         band01 <= 99
         band04 > 52
         band04 <= 77
   then
         dep = 10.2 + 0.66 band01 + 0.42 band02 - 0.12 band03 - 0.05 band05

  Rule 1/2: [610 cases, mean 85.7, range 65 to 134, est err 5.6]

   if
         band01 <= 103
         band04 <= 52
   then
         dep = 4.4 + 0.65 band01 + 0.48 band02 + 0.18 band06 - 0.15 band03
               - 0.21 band04
```

* http://www.rulequest.com/cubist-info.html

Regression Tree Approach

1) Develop classical regression tree
   - -all nodes mutually exclusive
   - -subdivide data into subsets which minimize the simple linear model weighted standard deviation of residuals

2) Develop Generalized rules from the regression tree in step 1
   - -makes trees easier to interpret
   - -less "rules" than tree "leaves" (a.k.a. terminal nodes)
   - -generalized by deleting excessive conditions
   - -rule sets can overlap, predictions are averaged in overlap areas for smoother output
   - -usually as accurate as a pruned regression tree

3) Generalized rules are used to make predictions across the image

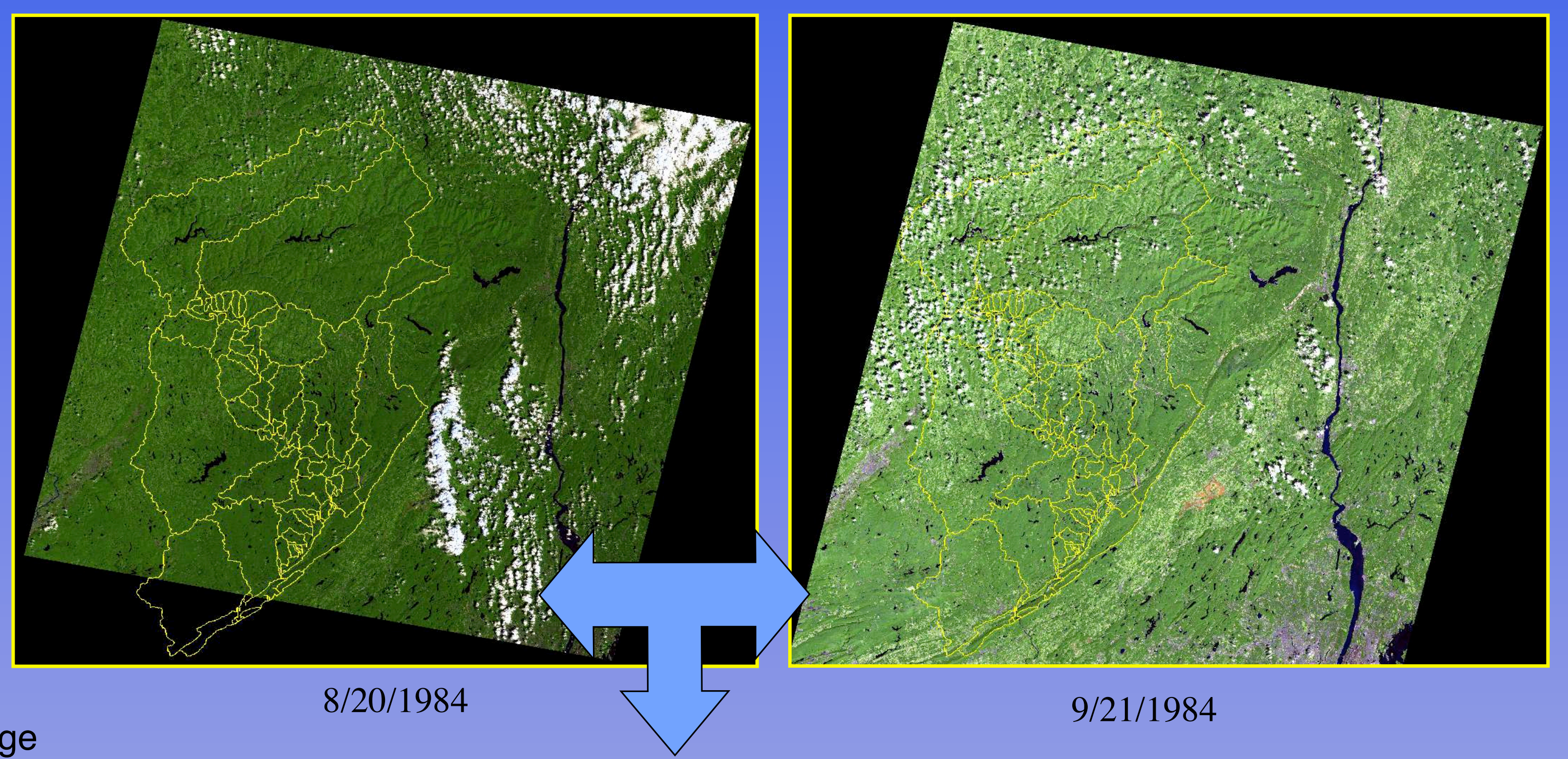
Data

- Two Landsat 5 scenes (p014r031) acquired on 8/20/84 and 9/21/84.
- Three SLC-off scenes (p014r031) acquired:
  - 7/2/04
  - 8/3/04
  - 9/20/04
  - One SLC-On scene for areas with no data acquired 10/01/02.
- Area of coverage is the Upper Delaware River Basin
  - Entire Upper Basin for 1984.
  - 1000 by 1000 pixel subset for 2004 tests.

Methods

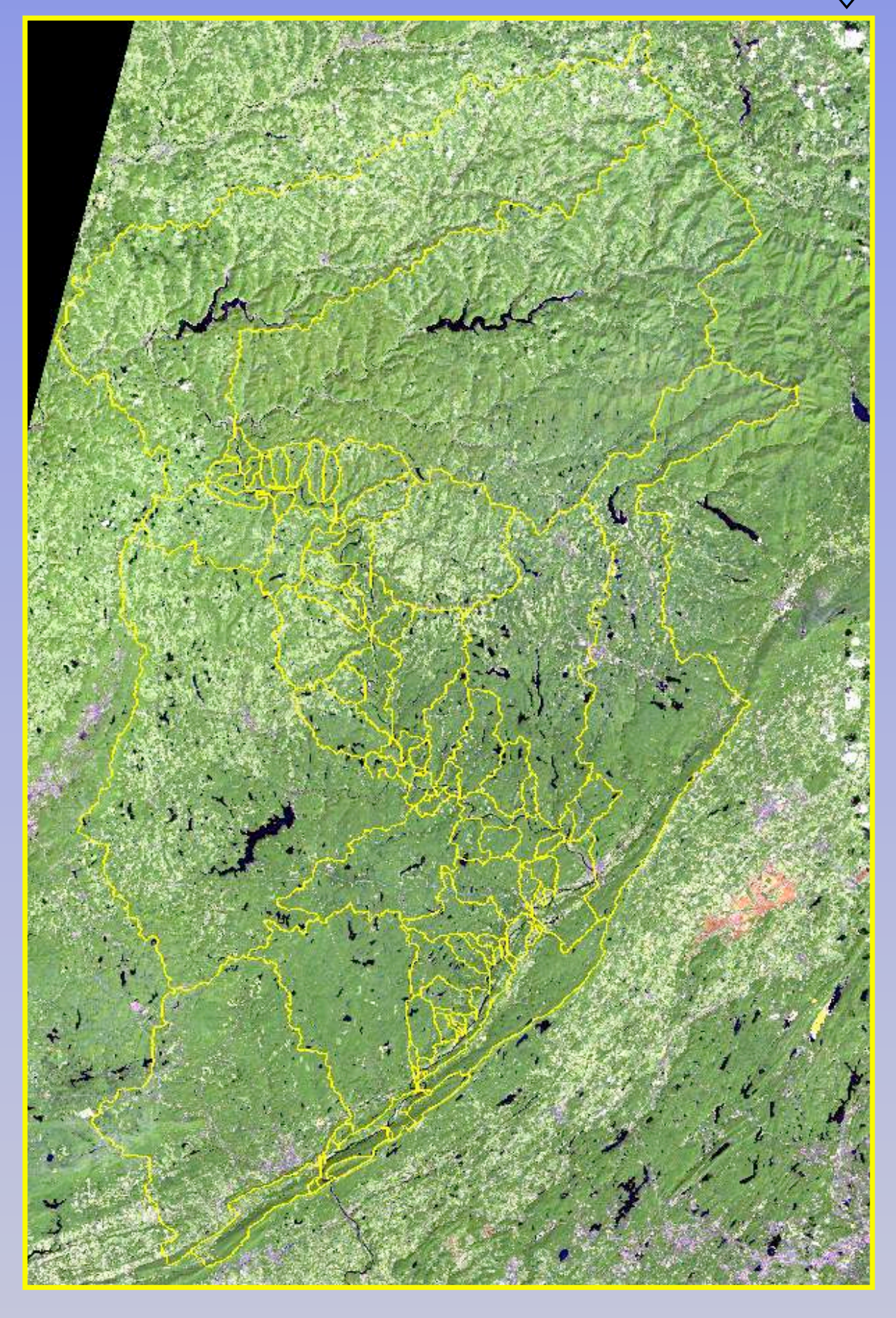
- Landsat 5 data terrain-corrected through NLAPS, L7 through IAS.
- Data atmospherically-corrected through Landsat Ecosystem Disturbance Assessment Processing System (LEDAPS) at GSFC.
- New radiometric coefficients for L5 used.
- Cloud and shadow masks developed starting from 5% differences between images. Fully-automated approach not implemented.
- Cloud and shadow masks were dilated to capture cloud and shadow edges.
- Training data for regression trees were taken from all non-cloud, non-shadow, non-gap areas common to the target and slave images.
- Regression blocks selected with a moving local window (500x500 pixels with 250 pixel overlap for Upper Basin and 100x100 pixel regions with 50 pixel overlap for subsets).
- Results evaluated on 80% training data and 20% test data.

BEFORE



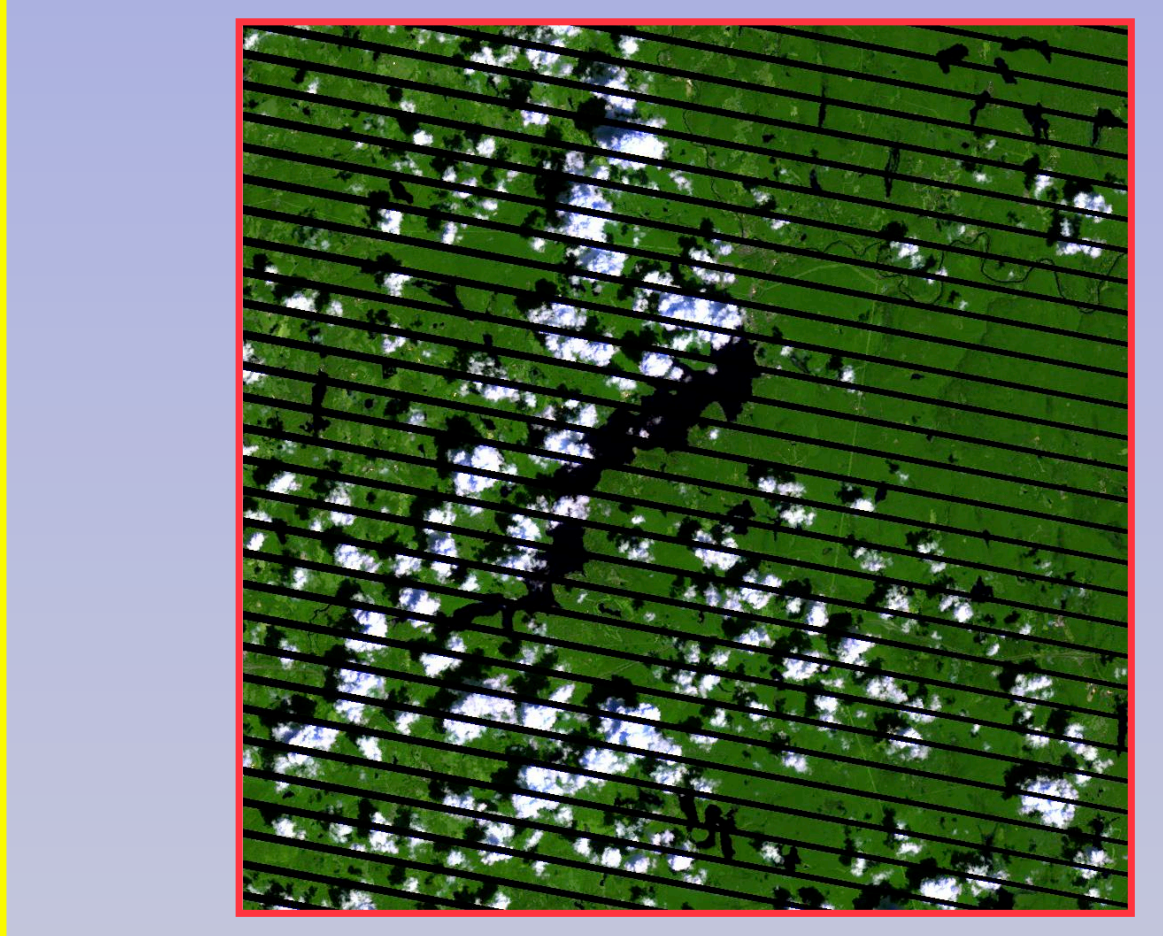
8/20/1984

9/21/1984

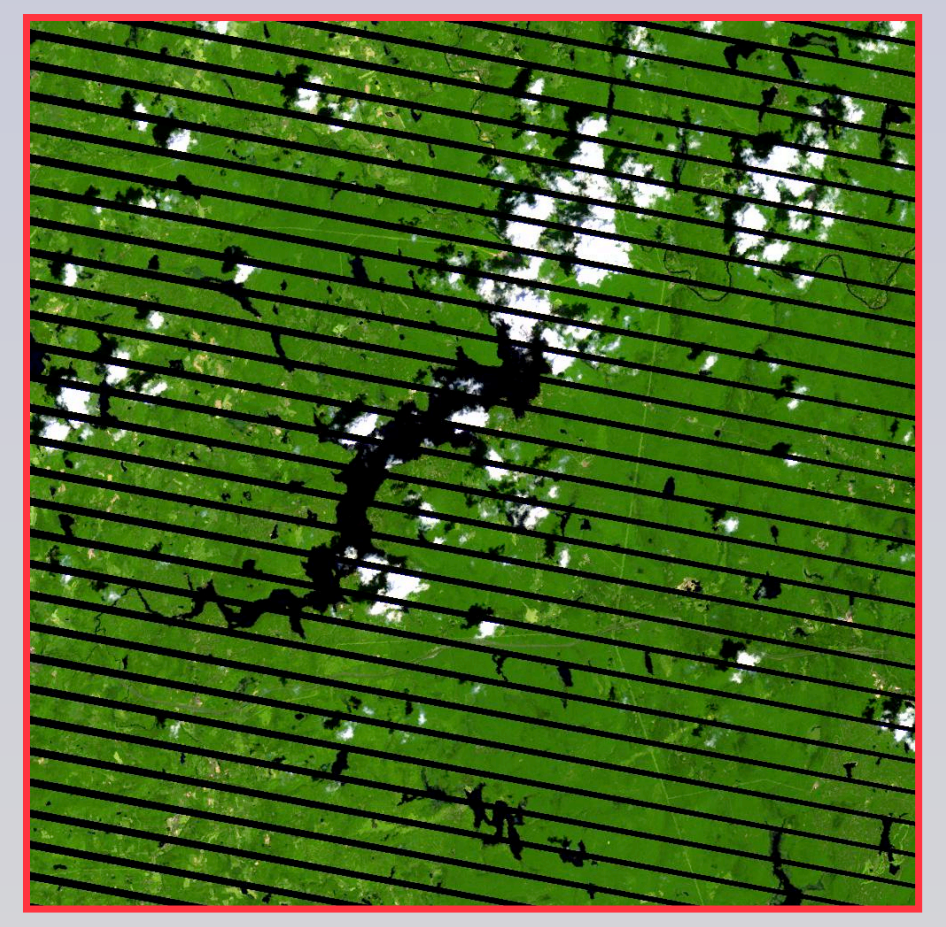


AFTER

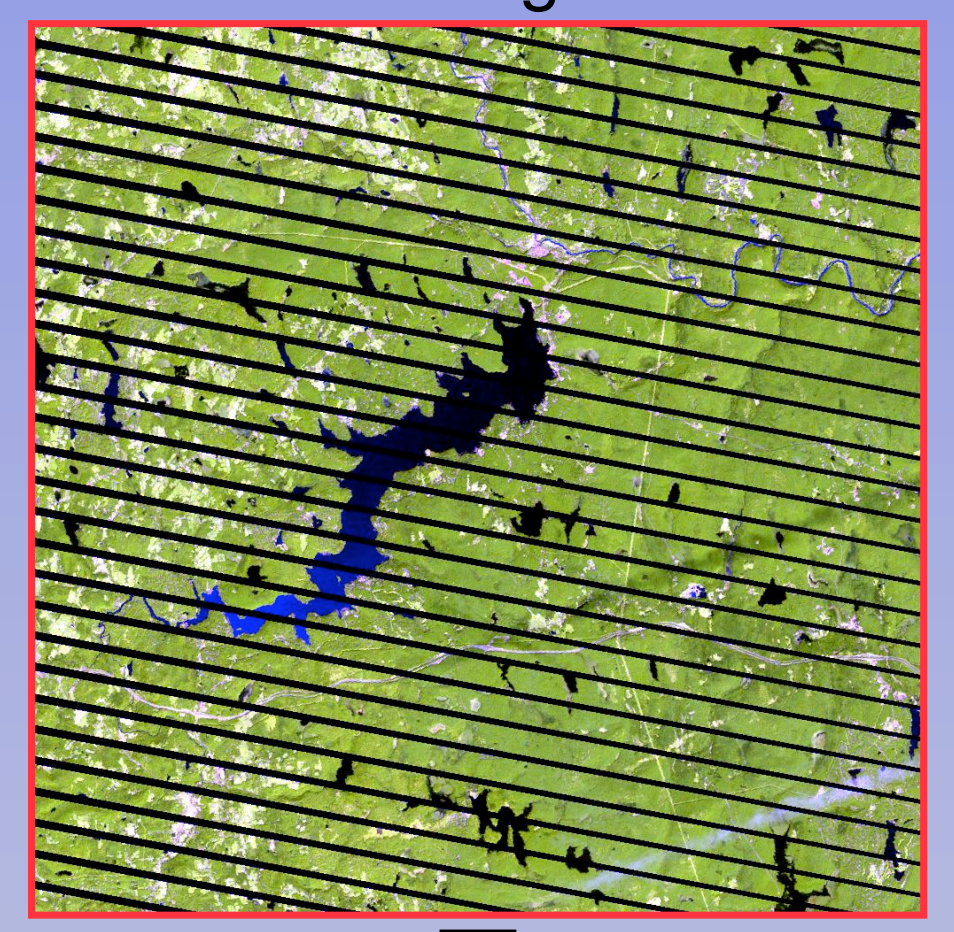
08/03/04



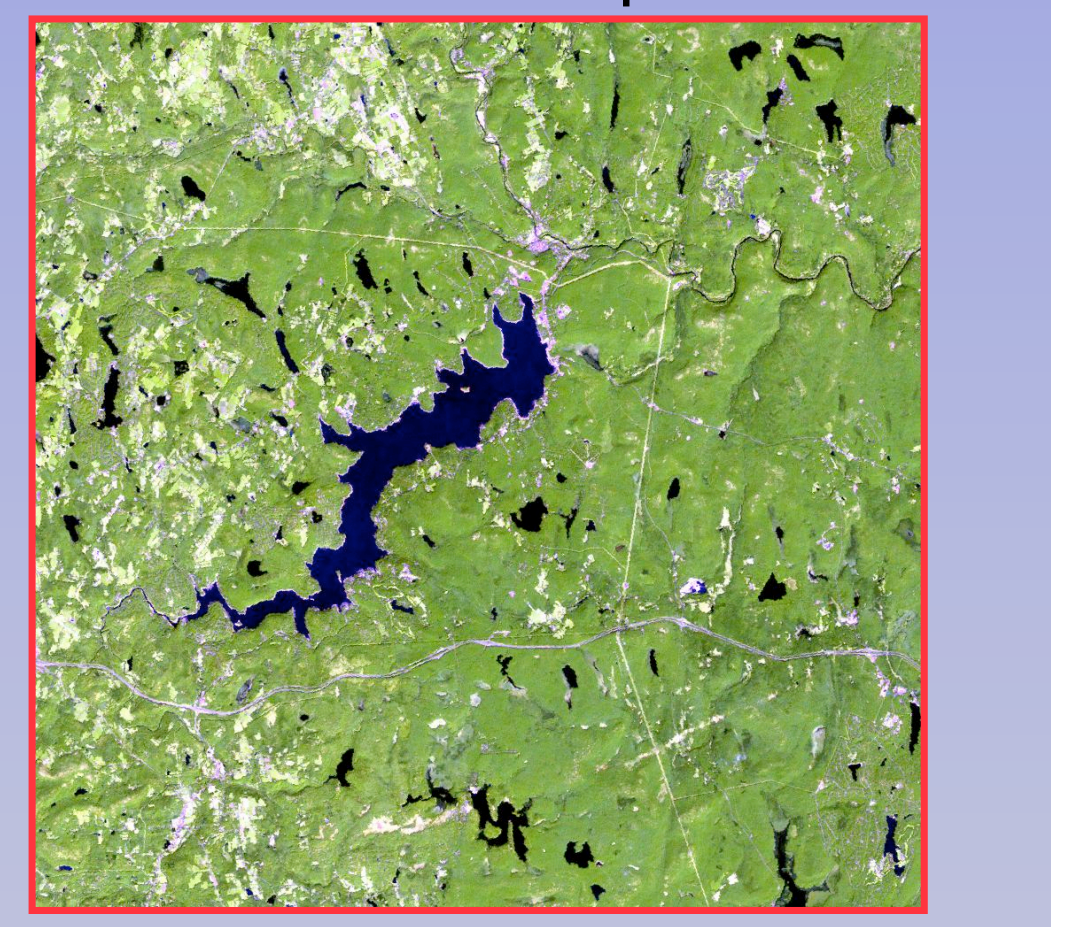
07/02/04

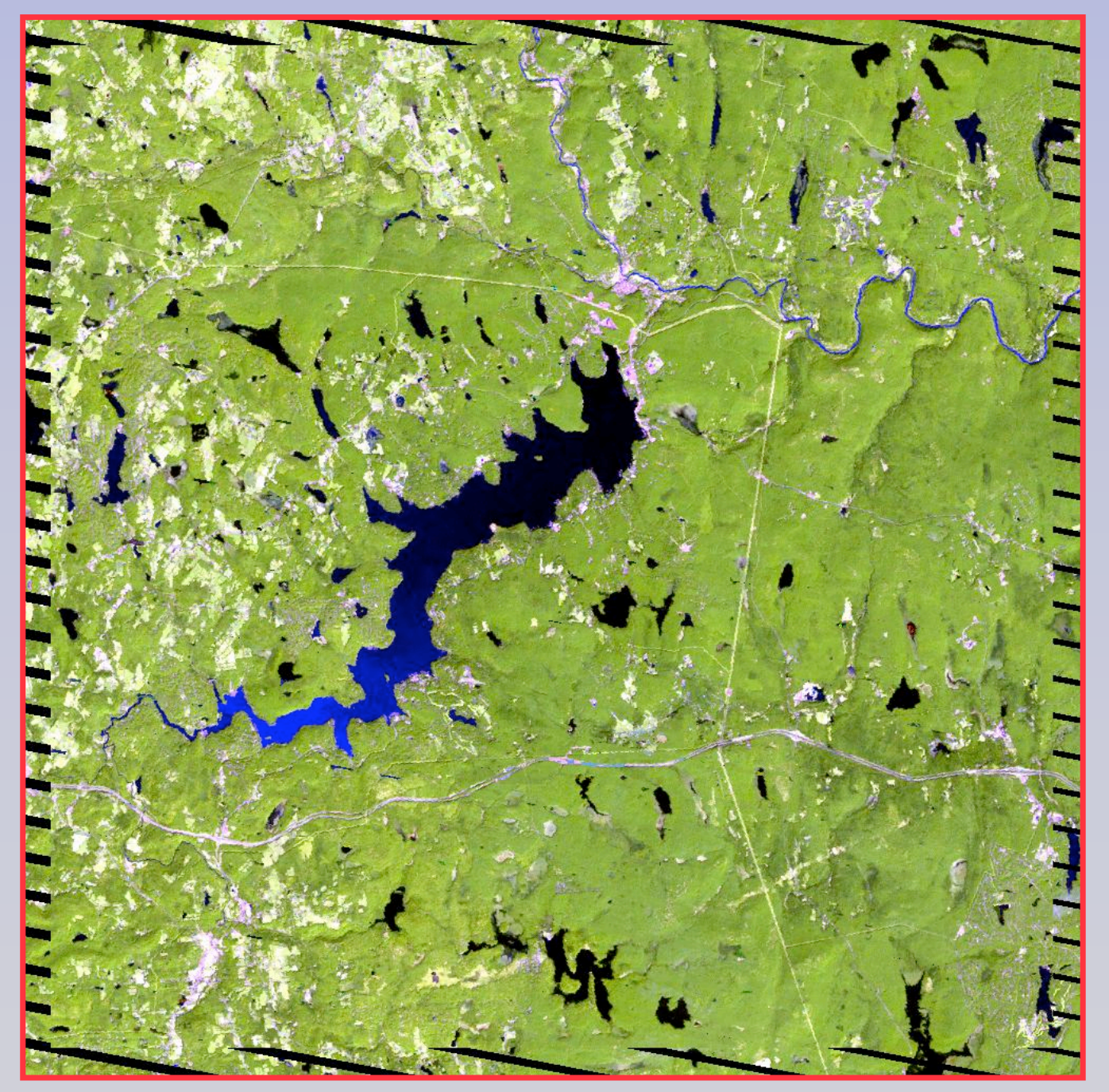


9/20/04 Target Scene



10/01/02 Backup Scene





Cloud- and Gap-Filled Result

Results

	Upper Delaware Basin	Subsets	
	1984	07/02/04	08/03/04
Band1	0.427	0.296	0.303
Band2	0.570	0.343	0.337
Band3	0.690	0.370	0.374
Band4	2.662	2.604	1.738
Band5	1.347	1.444	0.979
Band7	1.016	0.815	0.616

The results above are errors in Reflectance units and were obtained from combined training and test data.

Next Steps

- Incorporate two surrounding scenes to better describe land cover change.
- Test and Integrate with automated cloud and shadow masks.
- Test on test SLC-on imagery using cloud, shadow and gap masks
- Evaluate impact on land cover change detection products.