Final Report for NASA grant # NNX08AH05A

"Enhanced Land Cover and Land Cover Change Products"

PI: John Townshend

Co-I: Robert Sohlberg

Co-I: Matthew Hansen

Co-I: Chengquan Huang

Co-I: Ruth DeFries

University of Maryland

1. Table of Contents

2. Introduction

Characterization of the land surface from satellite data has been performed for over three decades. The Vegetation Continuous Fields (VCF) product is a global representation of the Earth's surface as gradations of three components of ground cover: percent tree cover, percent non-tree vegetation and percent bare (figure 1) (Carroll et al, 2010). Each pixel is shown as a sub-pixel mixture of cover with each of the three components expressed as a percentage of ground cover.

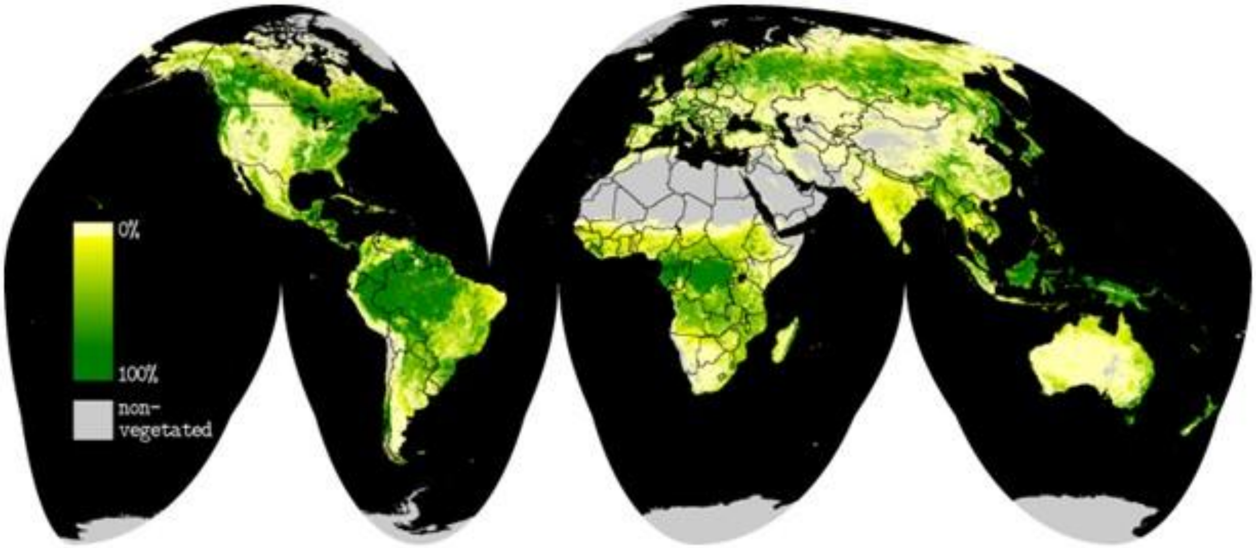

Figure 1Global Vegetation Continuous Fields percent tree cover. Darker greens indicate denser tree cover, pale colors indicate light tree cover, and gray indicates completely bare.

The three components are each stored in separate layers so they can be used independently to look at a particular type of ground cover or collectively to look at the entire surface.

When originally proposed the VCF product represented a revolutionary new approach to the characterization of vegetative land cover (DeFries and Townshend, 1994; DeFries et al, 1997). Conventional land cover classifications suffer from the imposition of arbitrary thresholds between classes, and the characterization of the land surface is highly dependent on the a priori class boundaries which are chosen (Hansen et al, 2002). Moreover when land cover products are used in models, parameterization has to be carried out, which is often crude and inaccurate. By depicting each pixel as a proportion of characteristics such as percentage tree cover, non-tree vegetation cover and bare ground a genuinely quantitative depiction of land cover became possible. The advantages of this approach have been recognized by the widespread adoption of the VCF product by many users in the modeling and monitoring communities. The VCF product has also been identified as an Earth System Data Record (ESDR) by the science community (Masek et al, 2006).

Historically the creation of the algorithms for the production of global land cover maps has been largely hand crafted as human intervention was required to help the algorithm make distinctions between certain land cover types, such as wetlands distinction from forest. The current version of the VCF algorithm endeavors to minimize the human element and allow the algorithm to make final decisions. The early MODIS VCF algorithms were developed using a semi-automated process where the regression trees were created using machine learning software. These trees were then evaluated by an operator, who could then add training at certain branches of the tree or simply eliminate parts of that tree. This human interaction was necessary because the training data, though very good at the time it was created, had some inconsistencies. In the new approach, the training data has been completely updated using Landsat Geocover data and has been revised and refined using the plethora of fine and ultra-fine resolution data available through the NASA science data purchase, Google Earth, among many others. The improved training data and the implementation of new and improved data mining software have resulted in much greater accuracy in the final product without human intervention.

The final algorithm for VCF operates in a completely automated fashion with the results published upon completion. The following pages describe in detail how this algorithm came to be and basically how it works.

3. Algorithm

The first step in the process was to develop an updated training data set. The spatial resolution for MODIS data used in the VCF Collection 5 product is 250m. The training data that were used for previous versions of the VCF product were developed in the late 1990’s and revised circa 2001. This data set represented a high quality data set at the time, but relied heavily on Landsat 5 Thematic Mapper data from the 1980’s. With the availability of the ortho-rectified Landsat Geocover and globally available fine resolution data it was advantageous to create a completely new training data set that better matched the acquisition dates of the MODIS data (2000 to present).

Training data are created by performing a discrete classification on the Landsat data into 4 classes of relative percent tree cover (0, 25, 50, 80+). These relative percentages are verified by overlaying the scenes with fine and ultra-fine resolution imagery from Ikonos, Quickbird, and other data as available. In addition, comparisons were made to Google Earth where tree crowns can be seen distinctly. Adjustments were made to the discrete values as necessary to match observed conditions from the fine resolution data. The 30m data are then averaged to 250m spatial resolution yielding a continuous representation of the surface for that scene from 0 – 100 percent.

The “0” class is dominant in the final training data set. To combat this, the training data with 0’s in desert regions and permanent snow is minimized. This allows more training pixels to be chosen from problem areas such as transition zones in semi-arid regions and cropped areas. In addition, it helps to normalize the training set so that 0 is not as dominant in the final set.

The production algorithm for VCF runs in three parts: sampling inputs under the training data; creating models; and applying the models to the output. These three steps are accomplished with open-source software (Weka data mining software, Java, GDAL) and custom software written in C. A global annual VCF data set can be produced in the

MODIS Adaptive Processing System (MODAPS), the PI led processing system for MODIS products, in about 5 days with the full ten year record able to be processed in under 6 weeks. By way of comparison, previous VCF products were generated at the PI's Science Computing Facility (SCF) at a rate of 1 global annual VCF product created in about a month. Model development in the SCF usually took an additional month or more.

The new process employs bagging where 30 independent regression trees are created and the final result is the average of the 30 independent results. The regression tree models are created using the "M5 regression tree with pruning". This process has been used in regression tree models with remote sensing data for over a decade and has been shown to produce more reliable results as compared to a single tree model (Chang et al, 2007; Hansen et al, 2003).

4. Proposed Enhancements

New proposed to the VCF products include the following:

1) Create and deliver operational code to MODAPS
2) 250 meter and spatially degraded versions of all current layers;
3) Enhancement to training data;

4.1. Operational code development

During negotiations with the program manager after the project was selected for funding it was decided that a necessary result of this project was to deliver operational code that could be run inside the MODIS Adaptive Processing System (MODAPS). Prior to this, the products were created at the Science Computing Facility (SCF) and only the products were delivered to NASA for distribution. The project accepted this task and as a consequence had to eliminate a task that was in the original proposal. The task that was eliminated was the disturbance mapping from multiple annual products. The differences between the code that was to be run at the SCF and that which could be delivered to MODAPS are substantial as the version that runs in MODAPS must be completed automated and not monitored by an operator. In addition, the code that runs at the SCF utilized commercial software unavailable in the MODAPS resulting in the need to find an open source replacement and re-write the VCF source code to utilize the new software. This task became the primary focus of the project.

4.2. New spatial resolutions for products.

The collection 5 version of VCF will be produced using the MOD44C 250m 16-day surface reflectance composites. All data layers will be created using this input which has previously been used to create the Vegetative Cover Conversion results. The four times improvement in detail revealed by this product is expected greatly to increase scientific value of the product because of the relatively fine scale of land cover variation and especially of cover change (Townshend and Justice 1990).

The modeling community has requested spatially degraded versions of the data layers to use as inputs for their models. We will deliver and make available via the GLCF, a spatially degraded version of all VCF data sets for use by the modeling community. Due to the proportional nature of the VCF cover estimates, it is a simple task to aggregate and average the fine resolution data to any desired spatial resolution. Our nominal delivery will include resolutions of 5 km, 20 km and 0.5º. Alternative requests will be accommodated.

4.3. Enhancements to training data.

Initially, the training data will have to be re-evaluated and resampled from the original fine resolution to 250 meter spatial resolution to match the new input data. To improve the VCF products, we propose to substantially enhance the quality of the training data needed to generate the VCF models through two approaches. One is to greatly increase the number of training samples and improve their spatial distribution across the globe. The other is to automatically screen the training samples and remove those that may have experienced changes since they were delineated. The training samples for generating the VCF models were delineated based on Landsat images (DeFries *et al* 2000; Hansen *et al* 2002a). We will select the areas where training data are needed based on deficiencies in the distribution of our existing training samples and global forest ecosystem maps (e.g. Olson *et al* 2001).

5. Milestones for each project year

Year 1

- Data acquisition for 8 years of MODIS 250 meter Collection 16-day composites of visible bands plus thermal
- Code development for automation of production of training data from Landsat data inputs
- Code development for automation of VCF product generation

Year 2

- Testing of code to create new training data
- Create new global training data set for use with VCF
- Generate initial 250m VCF output products
- Begin product evaluation

Year 3

- Final 250m VCF output products created
- Final evaluation of products
- Product delivery to the Land Processes DAAC
- Code and product documentation
- Code delivery to MODAPS

6. Results

Annual results for the VCF product using MODIS Terra data from 2000 to 2009 have been produced for percent tree cover. These results (figure 1) show expected patterns of tree cover extent. There remain some minor confusion with some cropped areas, high latitude mountain shadows, and some wetlands, but overall the output is substantially better than the previous 500m version in spatial detail and coherence.

In the image pairs in figure 2, the image on left is the old 500m product and the image on right is the new 250m product. Both are shown in a 250m grid to emphasize the improvement in spatial detail. Figures 2 a and b show improvements in the representation of the ridge and valley system in southern Pennsylvania in the US. Figures 2 c and d show clearings in southern Mato Grosso state in Brazil where the new VCF shows values approaching 0% tree cover in the clearings and the old VCF product showed values between 10% and 30% in many cases. Finally, figures 2 e and f show agricultural areas in southern Brazil. The old 500m product showed these areas as having between 10 and 25% tree cover, where the new 250m product indicates that the tree cover is near 0%, and the forested areas are highly fragmented.

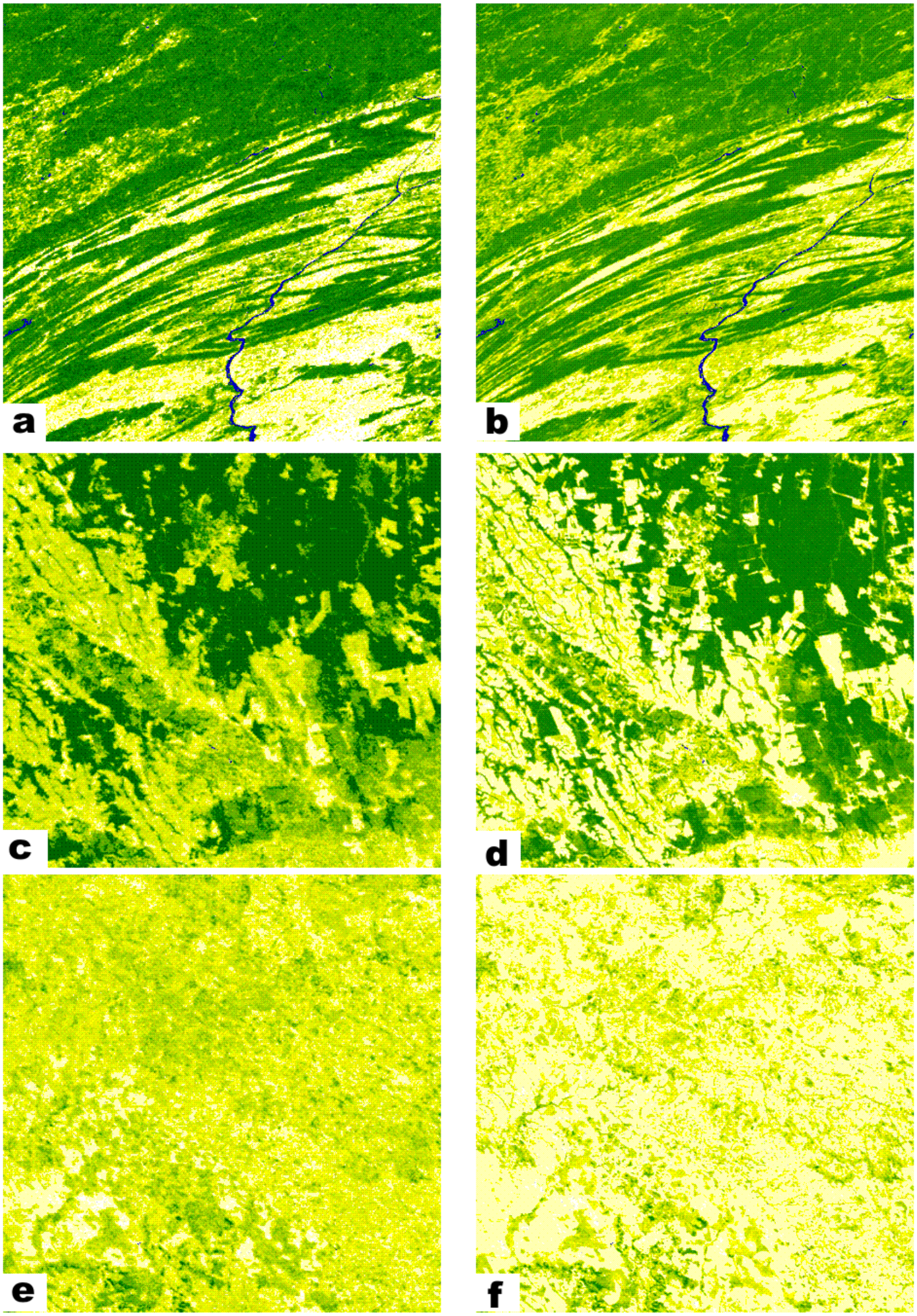

Figure 2 Image pairs showing the Collection 3 500m VCF on the left and the new Collection 5 250m VCF on the right. Darker green color indicates denser tree cover.

7. Validation

A limited amount of validation has been performed using field data from two sites in Maryland, and three sites in Brazil, South America (Table 1). Initial results show that the new C5 VCF product is substantially more accurate compared to ground based measurements of canopy cover with as much as a 50% improvement in RMSE between the two versions. Although these results are preliminary we are encouraged by the overall improvement in the VCF tree cover product with available ground based validation data.

Site	Field data	Old VCF	New VCF
Maryland			
SERC 1	29	16	34
SERC 2	48	61	51
SERC 3	33	40	50
SERC 4	59	61	46
SERC 5	69	40	57
GB 1	67	74	59
GB 2	69	66	68
GB 3	33	74	37
RMSE		19.27%	9.47%
Mean Absolute Error		14.37%	7.87%
Mato Grosso			
Explorada 1	64		49
Explorada 2	80		78
Moth	63		76
Disturbed	64		74
Logged 2	72		79
Logged	55		79
Ik-log	50		80
Tower	0		
RMSE			10.46%
Mean Absolute Error			9.40%

Table 1 Validation data from field sites in Maryland, United States.

8. Overall project achievements

The primary accomplishments of this project were the generation of an updated training data set using ortho-rectified Landsat data from the Geocover collection, and the generation of fully automated source code that can generate the regression tree models "on the fly" and produce final outputs in NASA MODIS production system (MODAPS).

The new training data set uses more recent landsat data which matches the MODIS acquisition dates more closely. Previous training data for the VCF product came from scenes from as far back as the early 1980’s and were not ortho-rectified leading to unknown contamination from collocation problems.

The production code that is now running in MODAPS to generate the VCF product is the first in the MODIS suite to generate an annual product from 16-day composites and the first to generate regression tree models on the fly and then apply them to all inputs in a single software stream. This software is extremely flexible and will require only minimal updates to incorporate planned updates to the VCF product including the generation of percent bare and percent non-tree vegetation.

The new QA layers which identify the impact of poor quality inputs, cloudy inputs, and the standard deviation between the 30 models offer substantial improvement in the end user’s ability to interpret the data. Prior versions of the VCF product did not include QA layers due to the complexity of the input metrics that are used to generate the product. With this version we have simplified the process and as a result have been able to generate meaningful QA that provide per pixel quality information which can be used to estimate the accuracy of the data. This is a powerful new tool for the end users giving them more explanatory power.

In addition to improved training data, we have improved the creation and selection of attributes used as model inputs. These attributes are calculated from surface reflectance and thermal data with the aim of reducing regional variance in seasonality. These improvements together with the finer resolution of the outputs have decreased errors in the products substantially. Comparisons of our new outputs with high resolution data (table 1) show as much as 50% improvement in accuracy.

We are exploring the hosting of the product in Open Geospatial Consortium (OGC) and Keyhole Markup Language (KML) formats with collaborators in the NASA Advanced Information Systems Technology (AIST) program to meet the needs of global first responders and other users. In addition, the new 250m Land/Water mask has been incorporated into the layers of the VCF product to help users in areas where water is prevalent. We are also polling the modeling community to determine the most appropriate projections and resolutions to create the Climate Modeling Grid outputs, these products will be forthcoming pending the results of that query.

9. Citations

Carroll, M., Townshend, J., Hansen, M., DiMiceli, C., Sohlberg, R., Wurster, K. 2011. Vegetative Cover Conversion and Vegetation Continuous Fields. In Ramachandran,, B., Justice, C.O., Abrams, M. (eds.) Land Remote Sensing and Global Environmental Change: NASA’s Earth Observing System and the Science of ASTER and MODIS *Springer-Verlag*.

DeFries, R., Hansen, M., Townshend, J.R.G., Janetos, A.C. and Loveland, T.R. (2000). Continuous Fields 1 Km Tree Cover. College Park, Maryland: The Global Land Cover Facility.

DeFries, R., Field, C. B., Fung, I., Justice, C. O., Matson, P. A., Matthews, M., Mooney, H. A., Potter, C. S., Prentice, K., Sellers, P. J., Townshend, J., Tucker, C. J., Ustin, S. L. and Vitousek, P. M. (1995). Mapping the land surface for global atmosphere-biosphere models: toward continuous distributions of vegetation's functional properties, *Journal of Geophysical Research,* 100:20,867-20,882.

Hansen, M., Stehman, S, Potapov, P., Loveland, T., Townshend, J., DeFries, R., Pittman, K., Arunarwati, B., Stolle, F., Steininger, M., Carroll, M. and DiMiceli, C. (2008). Humid Tropical Forest Clearing from 2000 to 2005 Quantified Using Multi-temporal and Multi-resolution Remotely Sensed Data, *Proceedings National Academy of Sciences*, 10, (27), pp. 9439–9444.

Hansen, M., Townshend, J., Stehman, S., Mayaux, P. and Morisette, J. (in preparation). Recommendations on the validation of vegetation continuous fields cover maps. Report from a joint CEOS-WGCV and GOFC-GOLD workshop on validation of vegetation continuous fields products, October 27-28, 2005, Brookings, South Dakota.

Hansen, M., Townshend, J., DeFries, R., and Carroll, M. (2005). Estimation of tree cover using MODIS data at global, continental and regional/local scales. *International Journal of Remote Sensing*, 26(19):4359-4380.

Hansen, M.C., DeFries, R. S., Townshend, J. R. G., Carroll, M., Dimiceli, C., and Sohlberg, R. A. 2003. Global Percent Tree Cover at a Spatial Resolution of 500 Meters: First results of the MODIS Vegetation Continuous Fields Algorithm. *Earth Interactions*, 7, 7 – 007.

Hansen, M.C., Sohlberg, R., Dimiceli, C., Carroll, M., DeFries, R.S. and Townshend, J.R.G. (2002). Towards an operational MODIS continuous field of percent tree cover algorithm: Examples using AVHRR and MODIS data. *Remote Sensing of Environment*, 83(1-2): 303-319.

Hansen, M. C., DeFries, R.S., Townshend, J.R.G., and Sohlberg, R. (2000). Global land cover classification at 1km spatial resolution using a classification tree approach, *International Journal of Remote Sensing,* 21, 1331-1364.

Olson, D.M., Dinerstein, E., Wikramanayake, E.D., Burgess, N.D., Powell, G.V.N., Underwood, E.C., D'Amico, J.A., Itoua, I., Strand, H.E., Morrison, J.C., Loucks, C.J., Allnutt, T.F., Ricketts, T.H., Kura, Y., Lamoreux, J.F., Wettengel, W.W., Hedao, P. and Kassem, K.R. (2001). Terrestrial Ecoregions of the World: A New Map of Life on Earth. *BioScience,* 51(11): 933-938.

Townshend, J.R.G. and Justice, C.O. 1990. The spatial variation of vegetation at very large scales. *International Journal of Remote Sensing*, 11, 149-157.