User Guide for the MODIS Vegetation Continuous Fields product Collection 5 version 1

PI: John Townshend
Co-I: Matthew Hansen
Co-I: Mark Carroll
Co-I: Charlene DiMiceli
Co-I: Robert Sohlberg
Co-I: Chengquan Huang
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. 2011; Hansen et al. 2000, 2002, 2003, 2005). Each pixel is shown as a sub-pixel mixture of cover with each of the three components expressed as a percentage of ground cover.
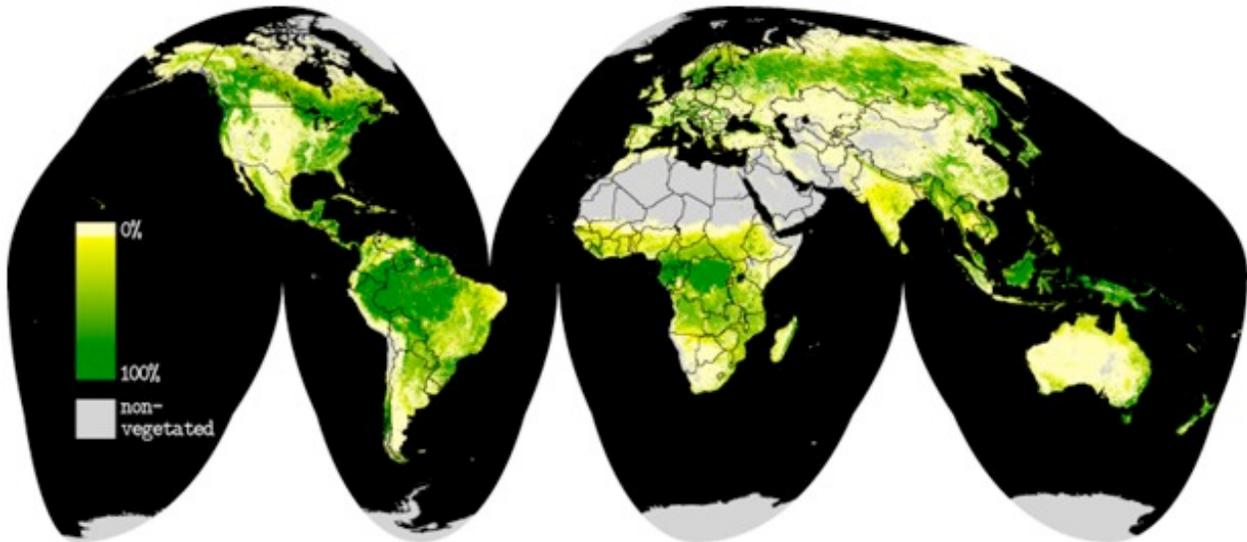

Figure 1: Global Vegetation Continuous Fields percent tree cover for 2001. Darker greens indicate denser tree cover, pale colors indicate light tree cover, and gray indicates completely bare.

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).

3.1.　　Training Data

Training data were 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 were 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 were then averaged to 250m spatial resolution yielding a continuous representation of the surface for that scene from $0 - 100$ percent.

Once created, the training data is a static data set. However, if specific problem areas emerge, the training data will be augmented with new information.

3.2.　　Data Inputs

The inputs for the MODIS VCF product are a 16-day surface reflectance composite which includes MODIS bands $1 - 7$ and brightness temperature from MODIS bands 20, 31,32; the training data (described above); and the MODIS Global 250m Land/Water Map. The surface reflectance composites that are used are an intermediate product generated inside the MODAPS production facility for the VCF product (for further information see Carroll et al. 2011). There are 23 16-day composites for each year of data. One year of 16-day composites are further composited to yield 8 composites per year in order to minimize clouds and as a data reduction step. These 8 final composites are used in the data production algorithm.

3.3. Data Algorithm

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) and custom software written in C programming language. In order to run in the MODIS Adaptive Processing System (MODAPS) it was necessary to adopt the open source software in lieu of proprietary software that was used in previous versions of the VCF product.

The algorithm begins with step 1 creating 30 independent samples from the training data set and organizing them for step 2, the model creation step. The model creation is performed by the Weka data mining software. The 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). Once all 30 models are created they are applied to the MODIS data to yield 30 independent results. The 30 independent results are averaged together to yield one result for any given pixel. Standard deviation from the 30 results is retained in a QA layer for the end user to understand the amount of agreement between the independent models. 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.

There are certain areas where persistent clouds or other factors preclude accurate prediction. In those cases, a reasonable value is inserted from a "basemap" created by the researchers. These values are flagged with a value of -100 in the Percent_Tree_Cover_SD SDS.

The surface reflectance composites that form the basis of the VCF product contain a wealth of per pixel QA information. This information is used to combine the 23 16- day composites into 8 composites and is retained during processing to be passed on to the end user. Poor quality input data obviously results in poor quality outputs so this information is saved in two quality assurance layers in the final product (see section 4 for full description).

3.4. File Naming Convention

The MODIS Vegetation Continuous Fields product is a "standard" MODIS product called MOD44B. Filenames contain the product ID, product date, tileID, collection number, and production date/time stamp; ex.

MOD44B.A2005065.h09v05.005.2011110122251.hdf

and are produced in hdf-eos file format with internal compression. The product date refers to the start date of the annual period so a product with ID "2005065" was produced with data from 2005065 – 2006064. The start date of all MODIS VCF products is yyyy065 (where yyyy refers to the 4 digit year). This originally derived from the first full 16-day composite period in the MODIS data record which begins with day of year 2000065. However, it has been continued because it relates better to seasons than the "Gregorian" calendar. If the product were generated starting January 1 and ending with December 31 it would result in splitting the southern hemisphere summer between 2 product years which is less desirable.

4. Data Layers

In the Collection 5 version 1 release of the MODIS VCF product (MOD44B) there are 4 science data sets (SDS'): 1) percent tree cover; 2) QA bad data; 3) QA cloudy data; 4) Standard Deviation of models. The first layer is the primary data layer with the ensuing 3 layers providing the user with indications of the overall quality of the data for any given pixel.

4.1.   *Percent_tree_cover*

The percent tree cover layer is a primary data layer which describes the percent of a pixel which is covered by tree canopy. This is defined as light penetration to the ground as compared to "crown" cover which describes the amount of the ground which is encompassed by the tree's crown regardless of whether light penetrates. The information in this layer can be used to identify forested areas for a variety of applications from resource management to the creation of plant functional types for climate modeling.

Valid values in the VCF percent tree cover layer are:

0 – 100   percent tree cover

200   water

253   fill / outside of projection

4.2.   *Quality*

The "Quality" sds is an 8-bit packed bit layer which describes, per pixel, those inputs that had poor quality defined by the MODIS surface reflectance quality assurance. In this case we define poor quality as those pixels which are cloudy, high aerosol, cloud shadow, or view zenith >45°. The bit field is described in table 1 below:

Bit	Input layers	State
0	DOY 065 - 097	0 clear; 1 bad
1	DOY 113 - 145	0 clear; 1 bad
2	DOY 161 - 193	0 clear; 1 bad
3	DOY 209 - 241	0 clear; 1 bad
4	DOY 257 - 289	0 clear; 1 bad
5	DOY 305 - 337	0 clear; 1 bad
6	DOY 353 - 017	0 clear; 1 bad
7	DOY 033 - 045	0 clear; 1 bad

Table 1: Description of bit field for "Quality" sds in VCF product.

Essentially each bit in the field represents 1 of the 8 input surface reflectance files to the model. If the value for that time period had only bad data (as defined above) the bit is turned "on" indicating that data for that time period was bad. The user should take this information into consideration when evaluating the usefulness of data for a given pixel. If the data are "bad" for 2 or more of the 8 time periods the user should be cautious with the value as it may be biased due to the poor inputs. This layer can be used in conjunction with the 4th sds which identifies those "bad" data pixels which were cloudy.

4.3.   Percent_tree_cover_SD

The "percent tree cover SD" layer provides the standard deviation (SD) of the 30 models that were used to generate the pixel value in the percent tree cover data layer. This information can be used to determine the level of agreement between the models in the production of the tree cover value. Values in this field represent percent cover and can be read as +/- the value shown. Obviously, a lower value indicates better agreement between the independent models. A fill value of -100 indicates that there was not sufficient data to make an accurate prediction and the pixel contains a reasonable value derived from other sources.

Valid Range: 0 – 10,000 (scaled by 100)

Fill Value: -100

4.4.   Cloud

The "cloud" layer is an 8-bit packed bit layer which clarifies the "Quality" layer to give the user an indication that the "bad" data refers to cloudy input data. The bit field is described in table 2 below:

Bit	Input layers	State
0	DOY 065 - 097	0 clear; 1 cloudy
1	DOY 113 - 145	0 clear; 1 cloudy
2	DOY 161 - 193	0 clear; 1 cloudy
3	DOY 209 - 241	0 clear; 1 cloudy
4	DOY 257 - 289	0 clear; 1 cloudy
5	DOY 305 - 337	0 clear; 1 cloudy
6	DOY 353 - 017	0 clear; 1 cloudy
7	DOY 033 - 045	0 clear; 1 cloudy

Table 2: Description of bit field for "Cloud" sds in VCF product.

As with the "Quality" layer, each bit in the field represents1 of the 8 surface reflectance inputs. The cloud is provided as a clarification to the quality because cloudy data is likely to depress the tree cover value in the model. This information is provided to help the user understand potential reasons why values seen in the percent tree cover data layer are high or low and provides the user a tool for determining whether they should trust the result.

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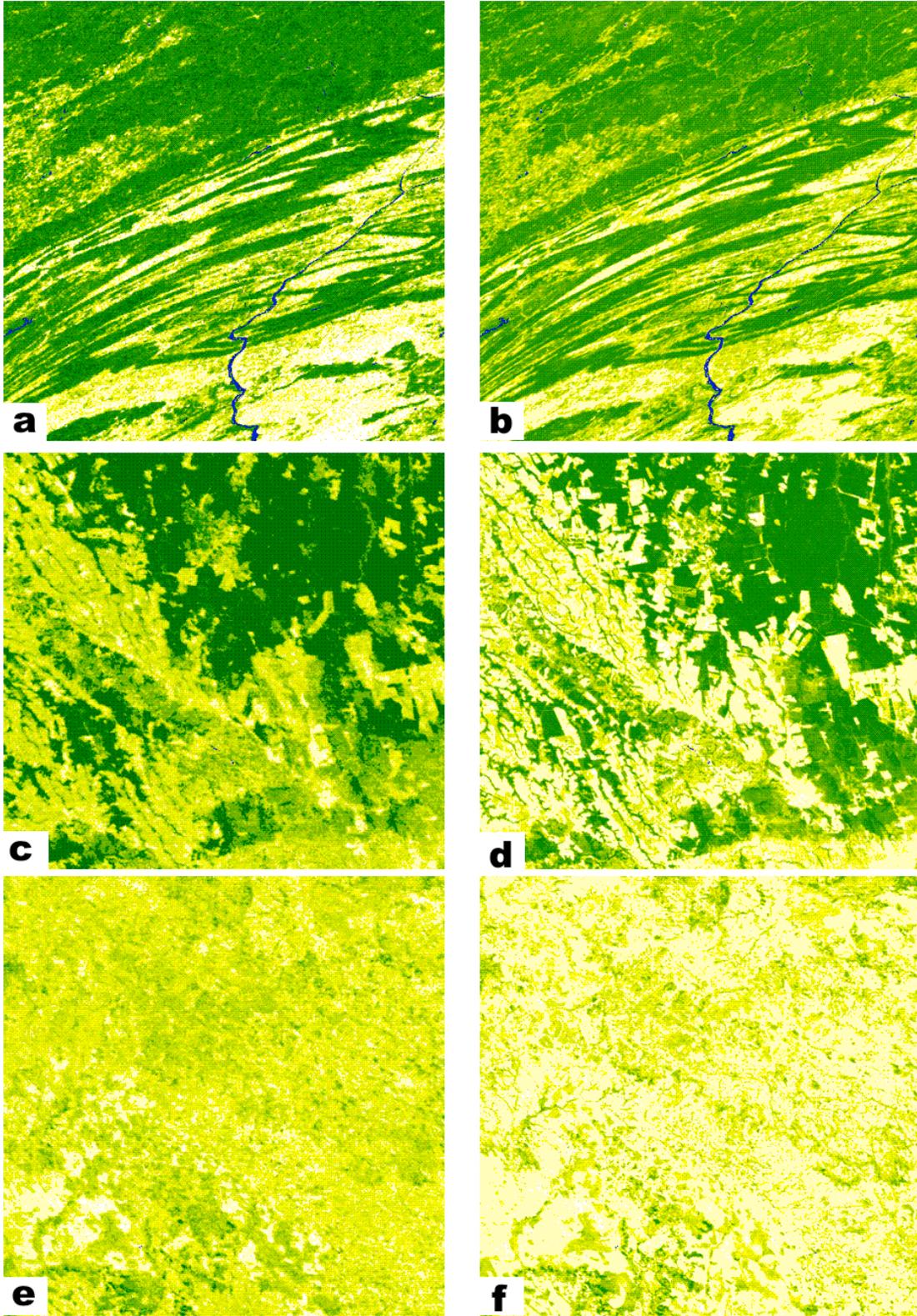
Figure 2: Image pairs showing the Collection 3 500m VCF on the left and the new Collection 5 250m VCF on the right. Darker green color indicates denser tree cover.

6. 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 3). 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 3: Validation data from field sites in Maryland, United States and Mato Grosso, South America.

7. Accessing and citing the data

The VCF data are provided free of charge. The data can be accessed from the Land Processes DAAC (https://lpdaac.usgs.gov/lpdaac/products/modis_products_table) under the product name MOD44B in MODIS tile format, and in alternate formats from the Global Land Cover Facility (http://landcover.org). Data should be cited as follows:

Full Citation Example: Townshend, J.R.G., M. Carroll, C. Dimiceli, R. Sohlberg M. Hansen, and R. DeFries. (2011), Vegetation Continuous Fields MOD44B, <mark>2001 Percent Tree Cover</mark>, Collection 5, University of Maryland, College Park, Maryland, 2001. (digital data accessed MM/DD/YYYY from <URL>)

The highlighted portion should change to reflect the data year and data layer used.

8. References

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*.

Chang J., Hansen M.C., Pittman K., Carroll M. & DiMiceli C., (2007). Corn and soybean mapping in the united states using MODIS time-series data sets, Agronomy Journal. 99(6):1654-1664.

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.

Masek, J.G., Friedl, M., Loveland, T., Brown de Colstoun, E., Townshend, J., Hansen, M. and Ranson, K.J. (2006). ESDR Community White Paper on Land Cover/Land Cover Change. NASA, Greenbelt, MD, pp. 9. (ftp://ftp.iluci.org/Land_ESDR/Landcover-change_Masek_whitepaper.pdf)

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.