

Documentation for the  
Global Human Built-up And Settlement Extent (HBASE)  
Dataset From Landsat

**October 2017**

Panshi Wang<sup>1</sup>, Chengquan Huang<sup>1</sup>, Eric C. Brown de Colstoun<sup>2</sup>, James C. Tilton<sup>3</sup>, Bin Tan<sup>4,5</sup>

<sup>1</sup> Department of Geographical Sciences, University of Maryland, College Park, MD USA

<sup>2</sup> Biospheric Sciences Laboratory, NASA Goddard Space Flight Center, Mailstop 618, NASA/GSFC, Greenbelt, MD 20771 USA

<sup>3</sup> Computational and Information Sciences and Technology Office, NASA/GSFC, Greenbelt, MD USA

<sup>4</sup> Science Systems and Applications, Inc., Lanham, MD USA

<sup>5</sup> Terrestrial Information Systems Laboratory, NASA Goddard Space Flight Center, Greenbelt, MD USA

**Abstract**

Urbanization is an important driver of change across our home planet. With over half of the world's population living in urban areas today, the mapping and monitoring of urbanization is critical to understanding these changes and their potential impacts. The availability of high resolution, free satellite imagery at multiple epochs from the Global Land Survey (GLS) Landsat archive provides great opportunities to map global man-made surfaces and extent in unprecedented detail. The Global Human Built-up And Settlement Extent (HBASE) Dataset From Landsat is derived from the GLS Landsat dataset for the target year 2010. The HBASE dataset consists of two components: 1) binary HBASE/non-HBASE classification; and 2) probability of HBASE. These layers are co-registered to the same spatial extent at a common 30m spatial resolution. The spatial extent covers the entire globe except Antarctica and some small islands. This dataset is one of the first global, 30m datasets of the extent of built-up area and settlements to be derived from the GLS data for 2010 and is a companion dataset to the Global Man-made Impervious Surface (GMIS) Dataset From Landsat.

**Data set citation:**

Wang, P., C. Huang, E. C. Brown de Colstoun, J. C. Tilton, and B. Tan. 2017. Global Human Built-up And Settlement Extent (HBASE) Dataset From Landsat. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). <https://doi.org/10.7927/H4DN434S>. Accessed DAY MONTH YEAR.

**Suggested citation for this document:**

Wang, P., C. Huang, E. C. Brown de Colstoun, J. C. Tilton, and B. Tan. 2017.  
Documentation for the Global Human Built-up And Settlement Extent (HBASE) Dataset  
From Landsat. Palisades, NY: NASA Socioeconomic Data and Applications Center  
(SEDAC). <https://doi.org/10.7927/H48W3BCM>. Accessed DAY MONTH YEAR.

We appreciate feedback regarding this dataset, such as suggestions, discovery of errors,  
difficulties in using the data, and format preferences. Please contact:

NASA Socioeconomic Data and Applications Center (SEDAC)  
Center for International Earth Science Information Network (CIESIN)  
Columbia University  
Phone: 1 (845) 365-8920  
Email: [ciesin.info@ciesin.columbia.edu](mailto:ciesin.info@ciesin.columbia.edu)  
Author Email: [eric.c.browndecolsto@nasa.gov](mailto:eric.c.browndecolsto@nasa.gov)  
Author Email: [pswang@umd.edu](mailto:pswang@umd.edu)

**Contents**

I.	Introduction.....	3
II.	Data and Methodology.....	4
III.	Data Set Description(s).....	6
IV.	How to Use the Data.....	7
V.	Potential Use Cases.....	7
VI.	Limitations.....	8
VII.	Acknowledgments.....	8
VIII.	Disclaimer.....	8
IX.	Use Constraints.....	9
X.	Recommended Citation(s).....	9
XI.	Source Code.....	9
XII.	References.....	9
XIII.	Documentation Copyright and License.....	11
	Appendix 1. Revision History.....	11
	Appendix 2. Contributing Authors & Documentation Revision History.....	11

## **I. Introduction**

Urban land cover only accounts for a small percentage of the Earth's land surface (estimates range from 0.2% ~ 3% [Liu et al. 2014; Schneider et al. 2009]), whereas more than half of the world's population now dwells in urban areas. Urban population is continuing to grow rapidly and is expected to reach two thirds of the world's population by 2050 (United Nations, 2015). This ongoing global urbanization process has broad impacts on the Earth's environmental systems including climatic systems (Arnfield 2003), hydrologic systems (Arnold and Gibbons, 1996) and ecosystems (Foley et al., 2005). Such impacts can go far beyond the physical footprint of urban areas (Lambin et al., 2001). Understanding the drivers, impacts and feedbacks of urban growth requires detailed and up-to-date information on the spatial extent of urban areas (Wentz et al., 2014).

Conventional data sources used for urban extent mapping include vector data and census population data based on administrative boundaries (Potere and Schneider, 2007). As valuable as these data sources are, they are too coarse for many applications and can become out of date quickly due to long updating cycles. Rapid development of earth observing technologies in recent decades has made it possible to mitigate these problems. Global datasets acquired by coarse to moderate resolution satellites, such as Defense Meteorological Satellite Program Operational Linescan System (DMSP-OLS) (Imhoff et al., 1997), Moderate Resolution Imaging Spectroradiometer (MODIS) (Schneider et al., 2009), Satellite Pour l'Observation de la Terre Vegetation (SPOT VEGETATION) (Bartholomé and Belward, 2005) and Medium Resolution Imaging Spectrometer (MERIS) (Arino et al., 2007), have been used to map urban extent at the global scale (Potere et al., (2009). However, many urban features are much smaller than the large pixel sizes of these datasets and hence cannot be mapped reliably using these datasets (Weng, 2012). Ideally, meter or sub-meter resolution images should be used in order to map the fine scale urban features accurately (Weng, 2012), but such high resolution images have yet to be acquired for many land areas of the globe. Landsat-class resolution satellites provide a viable option that balances the needs for global coverage and spatial details for urban monitoring, especially with the availability of global Landsat datasets known as the Global Landsat Survey (GLS) (Gutman et al., 2013).

We developed a Global Human Built-up And Settlements Extent (HBASE) mask using the GLS dataset for the 2010 epoch. The purpose of producing this dataset was twofold: 1) to address the need of a 30m urban area and settlements extent map, and 2) to address the problem of over-prediction of impervious cover in areas covered by dry soil, sands, rocks, etc. In using the HBASE mask, we have removed retrievals of impervious surface in such areas. These two datasets are available as individual layers here but are intimately connected.

## II. Data and Methodology

The methodology described below is only for the HBASE dataset. As mentioned above, HBASE was produced not only as a standalone product, but also as an effective way to remove or mask out errors of commission in the companion GMIS dataset. We refer the user to the appropriate documentation for the GMIS dataset for details on the methodology to create GMIS. Fig. 1 is the overall workflow of the methodology for the HBASE product.

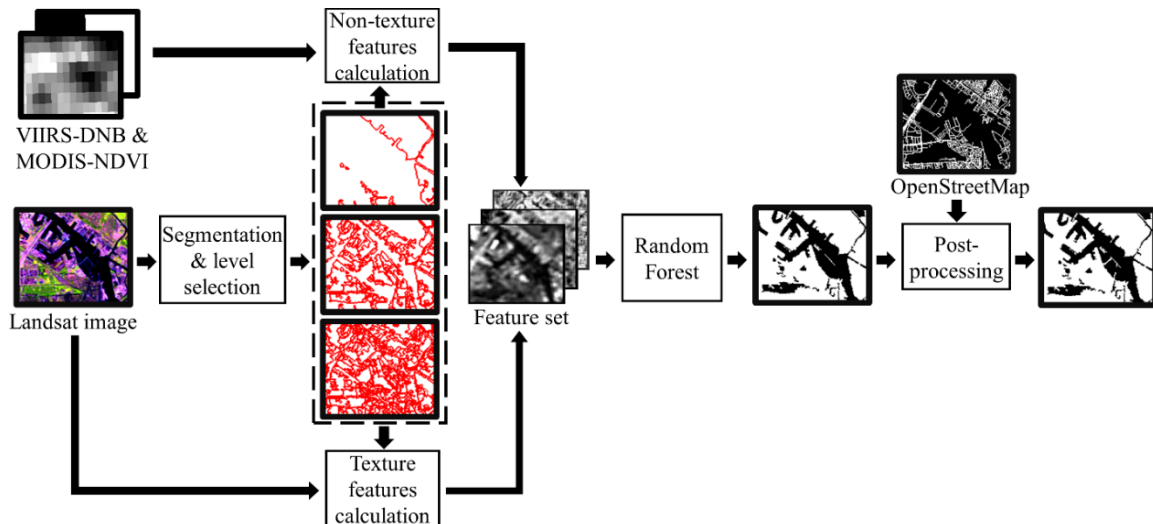


Figure 1. Overall flowchart for the GMIS product methodology.

### Step 1: GLS 2010 Surface Reflectance Dataset

The GLS dataset was developed through a joint NASA-USGS collaboration to provide a convenient basis for developing global land-cover and change products. It consisted of near complete coverage of Landsat data for all land areas of the globe for epochs centered on 1975, 1990, 2000, 2005, and 2010 (Gutman et al., 2013). The Thematic Mapper (TM) and Enhanced Thematic Mapper (ETM+) images have been atmospherically corrected and converted to surface reflectance (SR) using the Landsat Ecosystem Disturbance Adaptive Processing System (LEDAPS) (Masek et al., 2006) by the Global Land Cover Facility (GLCF). Global assessments revealed that LEDAPS SR values were highly consistent with MODIS SR (Feng et al., 2013).

### Step 2: Image Segmentation

The GLS SR images were segmented using the Recursive Hierarchical Image Segmentation (RHSeg) software package (Tilton et al., 2012), a recursive approximation of the Hierarchical Image Segmentation (HSeg) package. HSeg combines the power of the best merge region growing to delineate the boundaries between spatially

adjacent regions and spectral clustering to group spatially disjoint regions together. But its computational cost is very high for Landsat level (~8,000 by 8,000 pixels) or larger images. Using a divide-and-conquer approach, RHSeg was designed to improve the speed of the HSeg algorithm on cluster or cloud computing systems, which made it possible to process all Landsat images required by this study. The output of the RHSeg algorithm includes image objects at multiple scales, where finer scale objects are nested within coarser scale objects. In this study, we used object size thresholds to select three representative levels from the RHSeg segmentation hierarchy.

### **Step 3: Object-based Landsat Texture and Multi-source Features**

At each level, we calculated Gray Level Covariance Matrix (GLCM, (Haralick et al., 1973)) features including angular second moment, contrast, correlation, and variance using objects instead of windows as spatial units. In addition to traditional single-band texture features, we also extended the concept of GLCM by using the co-occurrence between two Landsat bands to calculate cross-band (color) texture features. The single-band textures are calculated for the Landsat spectral bands 5, 4 and 3. Since our implementation of GLCM is symmetrical, the cross-band textures are only calculated for the band combinations (5, 4), (4, 3) and (5, 3).

Three groups of variables were derived in addition to the textures. The first was a binary variable indicating whether an image was a Landsat 7 image, in order to separate images with and without Scan Line Corrector (SLC-off) gaps. We used metrics derived from MODIS Normalized Difference Vegetation Index (NDVI) datasets, including annual maximum, annual median, and the NDVI value for the month when the corresponding Landsat image was acquired. Finally, we used statistics derived from the Visible Infrared Imaging Radiometer Suite Day/Night Band (VIIRS DNB) datasets.

### **Step 4: Random Forest Classification**

The required training data were collected through visual interpretation of the Landsat images and high resolution images available from Google Earth. In addition to HBASE and non-HBASE samples, cloud/shadow samples were also selected as many GLS images had some level of cloud contamination. Training samples were selected based on the following rules: (1) an object was assigned to the HBASE class if it contained at least one HBASE pixel and did not have cloud/shadow pixels; (2) an object was assigned to the non-HBASE class if all its pixels were non-HBASE pixels; and (3) any object containing cloud/shadow pixels was assigned to the cloud/shadow class. This rule set was designed to produce an inclusive HBASE product that could capture low-density residential areas, meaning the final product would be optimized for minimizing omission of HBASE areas.

We used a random forest (RF) classifier to classify features described in Step 3 into HBASE/non-HBASE classes. The spatial units of classification are objects at the lowest level among the three selected segmentation levels. We adopted an iterative approach to derive the training data needed for this study. The initial set of training samples were

selected from across the globe to represent different urban types in different ecosystems. This training dataset was then used to train an RF model, which was used to produce global HBASE products. The derived product was then examined against the input Landsat images and high resolution images available from Google Earth. Additional training samples were collected over areas where large classification errors were found. This was iterated several times until no large errors were found and the remaining errors were more likely due to lack of separability than lack of representative training samples. In the end, a total of 1,658,805 training samples were collected as the final training data set.

### Step 5: Post-processing Steps

Due to limitations of Landsat spatial resolution, most roads were not successfully captured by the classification result. To mitigate this omission error, we used the OpenStreetMap (OSM) to provide information on major roads. OSM is a publically available dataset intended to provide up-to-date information on the global road network by incorporating open government data and users contributed data (OpenStreetMap contributors, 2016). Since some of the OSM users contributed datasets on local roads and their quality could be difficult to assess, only major roads (motorway, primary, and trunk roads according to OSM nomenclature), secondary roads, and airport runways from OSM were included. These features were rasterized and added to the final HBASE classification generated by the RF classifier.

## III. Data Set Description(s)

### Data set description:

The HBASE product is delivered in the format of GeoTIFF raster files with two layers (bands): non-HBASE/HBASE classification (200/201/202) and probability of the HBASE class (0-100).

The non-HBASE/HBASE classification (0/1) is given by the RF classifier. Note that pixels classified as non-HBASE have been set to 200 in accordance with the GMIS dataset. The values are coded as follows:

Value	Label
200	Non-HBASE
201	HBASE
202	Road
255	No data, clouds, shadows

The probability of the HBASE class (0-100) is estimated by the RF algorithm. Note that the HBASE probability for pixels classified as non-HBASE should be lower than 50%. The values are coded as follows:

Value	Label
0-100	probability of HBASE
255	No data, clouds, shadows

**Data set format:**

The data are available in GeoTIFF format as downloadable zip files. Users can access:

- Explore View - [sedac.ciesin.columbia.edu/mapping/gmis-hbase/explore-view/](http://sedac.ciesin.columbia.edu/mapping/gmis-hbase/explore-view/)
- Download View - [sedac.ciesin.columbia.edu/mapping/gmis-hbase/download-view/](http://sedac.ciesin.columbia.edu/mapping/gmis-hbase/download-view/)

**Data set downloads:**

Users can create a downloadable zip file via:

- Explore View by selecting the “Download View” button and will have the option to select the type of region to download (e.g. country, shapefile, rectangle).
- Download View with the option to select the type of region to download (e.g. country, shapefile, rectangle).

The data are provided in individual data files according to Universal Transverse Mercator (UTM) zones at native 30m, 250m, and 1km spatial resolutions.

## IV. How to Use the Data

Users are able to download data by country and region. Users should be aware that there are missing data or rather no data regions because of cloud cover, and Scan Line Corrector SLC gaps.

## V. Potential Use Cases

The dataset is expected to have a rather broad spectrum of users, from those wishing to examine/study the fine details of urban land cover over the globe at full 30m resolution to global modelers trying to understand the climate/environmental impacts of man-made surfaces at continental to global scales. For example, the data are applicable to local modeling studies of urban impacts on the energy, water, and carbon cycles as well as analyses at the individual country level.



## **VI. Limitations**

This dataset is based on the GLS dataset and may contain artifacts of that dataset such as cloud cover and issues related to the Landsat 7 Scan Line Corrector (SLC) failure. These features can cause some linear features to appear in the imagery and may cause boundary issues in the calculation of object-based features. Because of limitations of feature separability, it is also possible that some areas may have omission errors (e.g., small towns and villages) and commission errors (e.g., agricultural fields and bare soil). Coastlines and water bodies have been masked in from the best available source which may also contain small errors and/or omit small features.

## **VII. Acknowledgments**

This study was funded by NASA's Land Cover and Land Use Change (LCLUC) Program (grants 09-LCLUC09-2-0136-1 & NNX11AH67G). Additional support for Panshi Wang and Chengquan Huang were provided by NOAA and USGS. The NGA high resolution satellite images were processed through the GMIS project and were originally obtained by NASA under the NGA's NextView license agreement. The authors thank James Zhan, Mike Taylor and Sike Li for their efforts in deriving the impervious surface percentage data that was used for product assessment. The authors would also like to thank the Global Land Cover Facility at University of Maryland for providing the GLS-2010 surface reflectance dataset.

Funding for the dissemination of this dataset was provided under the NASA contract NNG13HQ04C for the continued operation of the Socioeconomic Data and Applications Center (SEDAC), which is operated by the Center for International Earth Science Information Network (CIESIN) of Columbia University.

## **VIII. Disclaimer**

CIESIN follows procedures designed to ensure that data disseminated by CIESIN are of reasonable quality. If, despite these procedures, users encounter apparent errors or misstatements in the data, they should contact SEDAC User Services at [ciesin.info@ciesin.columbia.edu](mailto:ciesin.info@ciesin.columbia.edu). Neither CIESIN nor NASA verifies or guarantees the accuracy, reliability, or completeness of any data provided. CIESIN provides this data without warranty of any kind whatsoever, either expressed or implied. CIESIN shall not be liable for incidental, consequential, or special damages arising out of the use of any data provided by CIESIN.



## IX. Use Constraints

Users are free to use, copy, distribute, transmit, and adapt the work for commercial and non-commercial purposes, without restriction, as long as clear attribution of the source is provided.

## X. Recommended Citation(s)

### Data set(s):

Wang, P., C. Huang, E. C. Brown de Colstoun, J. C. Tilton, and B. Tan. 2017. Global Human Built-up And Settlement Extent (HBASE) Dataset From Landsat. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC).  
<https://doi.org/10.7927/H4DN434S>. Accessed DAY MONTH YEAR.

## XI. Source Code


No source code is provided.

## XII. References

- Arino, O., D. Gross, F. Ranera, M. Leroy, P. Bicheron, C. Brockman, P. Defourny, C. Vancutsem, F. Achard, L. Durieux, L. Bourg, J. Latham, A. Di Gregorio, R. Witt, M. Herold, J. Sambale, S. Plummer, and J. L. Weber. 2007. Globcover: Esa Service for Global Land Cover from Meris. *Igarss: 2007 IEEE International Geoscience and Remote Sensing Symposium, Vols 1-12*, 2412-2415.
- Arnfield, A. J. 2003. Two Decades of Urban Climate Research: A Review of Turbulence, Exchanges of Energy and Water, and the Urban Heat Island. *International Journal of Climatology*, 23, 1-26.
- Arnold, C. L., and C. J. Gibbons. 1996. Impervious Surface Coverage: The Emergence of a Key Environmental Indicator. *Journal of the American Planning Association*, 62, 243-258.
- Bartholomé, E., and A. S. Belward. 2005. Glc2000: A New Approach to Global Land Cover Mapping from Earth Observation Data. *International Journal of Remote Sensing*, 26, 1959-1977.
- Feng, M., J. O. Sexton, C. Huang, J. G. Masek, E. F. Vermote, F. Gao, R. Narasimhan, S. Channan, R. E. Wolfe, and J. R. Townshend. 2013. A Global Land Survey surface reflectance product: assessment using coincident MODIS observations. *Remote Sensing of Environment*, 134, 276-293.
- Foley, J. A., R. Defries, G. P. Asner, C. Barford, G. Bonan, S. R. Carpenter, F. S. Chapin, M. T. Coe, G. C. Daily, H. K. Gibbs, J. H. Helkowski, T. Holloway, E. A. Howard,

- C. J. Kucharik, C. Monfreda, J. A. Patz, I. C. Prentice, N. Ramankutty, and P. K. Snyder. 2005. Global Consequences of Land Use. *Science*, 309, 570-574.
- Gutman, G., C. Huang, G. Chander, P. Noojipady, and J. G. Masek. 2013. Assessment of the NASA-USGS Global Land Survey (GLS) Datasets. *Remote Sensing of Environment*, 134, 249-265.
- Haralick, R. M., K. Shanmugam, and I. H. Dinstein. 1973. Textural Features for Image Classification. *IEEE Transactions on Systems, Man, and Cybernetics*, 3, 610-621.
- Imhoff, M. L., W. T. Lawrence, D. C. Stutzer, and C. D. Elvidge. 1997. A Technique for Using Composite DMSP/OLS "City Lights" Satellite Data to Map Urban Area. *Remote Sensing of Environment*, 61, 361-370.
- Lambin, E. F., B. L. Turner, H. J. Geist, S. B. Agbola, A. Angelsen, J. W. Bruce, O. T. Coomes, R. Dirzo, G. Fischer, C. Folke, P. S. George, K. Homewood, J. Imbernon, R. Leemans, X. Li, E. F. Moran, M. Mortimore, P. S. Ramakrishnan, J. F. Richards, H. Skånes, W. Steffen, G. D. Stone, U. Svedin, T. A. Veldkamp, C. Vogel, and J. Xu. 2001. The Causes of Land-Use and Land-Cover Change: Moving Beyond the Myths. *Global Environmental Change*, 11, 261-269.
- Liu, Z., C. He, Y. Zhou, and J. Wu. 2014. How Much of the World's Land Has Been Urbanized, Really? A Hierarchical Framework for Avoiding Confusion. *Landscape Ecology*, 29, 763-771.
- Masek, J.G., E.F. Vermote, N.E. Saleous, R.E. Wolfe, F.G. Hall, K.F. Huemmrich, F. Gao, J. Kutler, and T.K. Lim. 2006. A Landsat Surface Reflectance Dataset for North America, 1990–2000. *IEEE Geoscience and Remote Sensing Letters*, 3, 68-72.
- OpenStreetMap contributors. 2016. OpenStreetMap Planet dump [Data file from 04-22-2016 of database dump]. Available online at: <http://planet.openstreetmap.org>.
- Potere, D., and A. Schneider. 2007. A Critical Look at Representations of Urban Areas in Global Maps. *GeoJournal*, 69, 55-80.
- Potere, D., A. Schneider, S. Angel, and D. L. Civco. 2009. Mapping urban areas on a global scale: which of the eight maps now available is more accurate? *International Journal of Remote Sensing*, 30, 6531-6558.
- Schneider, A., M. A. Friedl, and D. Potere. 2009. A New Map of Global Urban Extent from MODIS Satellite Data. *Environmental Research Letters*, 4, 044003.
- United Nations. 2015. Urban and Rural Population Growth and World Urbanization Prospects. In, *World Urbanization Prospects: The 2014 Revision (St/EsA/Ser.A/366)*: United Nations, Department of Economic and Social Affairs, Population Division.
- Weng, Q. 2012. Remote Sensing of Impervious Surfaces in the Urban Areas: Requirements, Methods, and Trends. *Remote Sensing of Environment*, 117, 34-49.
- Wentz, E., S. Anderson, M. Fragkias, M. Netzband, V. Mesev, S. Myint, D. Quattrochi, A. Rahman, and K. Seto. 2014. Supporting Global Environmental Change Research: A Review of Trends and Knowledge Gaps in Urban Remote Sensing. *Remote Sensing*, 6, 3879-3905.

### **XIII. Documentation Copyright and License**

Copyright © 2017. The Trustees of Columbia University in the City of New York. This document is licensed under a Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>). 

#### **Appendix 1. Revision History**

No revisions have been made to this dataset.

#### **Appendix 2. Contributing Authors & Documentation Revision History**

Revision Date	Contributors	Revisions
October 25, 2017	Sri Vinay	This document is the 1 <sup>st</sup> instance of documentation.