

GLAD Tools V2.0

User's Manual



GLAD ARD data and tools are available at <https://glad.umd.edu/ard/home>

Suggested citation for the GLAD ARD (data and tools):

Potapov, P., Hansen, M.C., Kommareddy, I., Kommareddy, A., Turubanova, S., Pickens, A., Adusei, B., Tyukavina A., and Ying, Q., 2020. Landsat analysis ready data for global land cover and land cover change mapping. *Remote Sensing* 12, 426; doi:10.3390/rs12030426

<https://www.mdpi.com/2072-4292/12/3/426>

The latest document update: September 2023. This document is the work in progress and may contain errors.

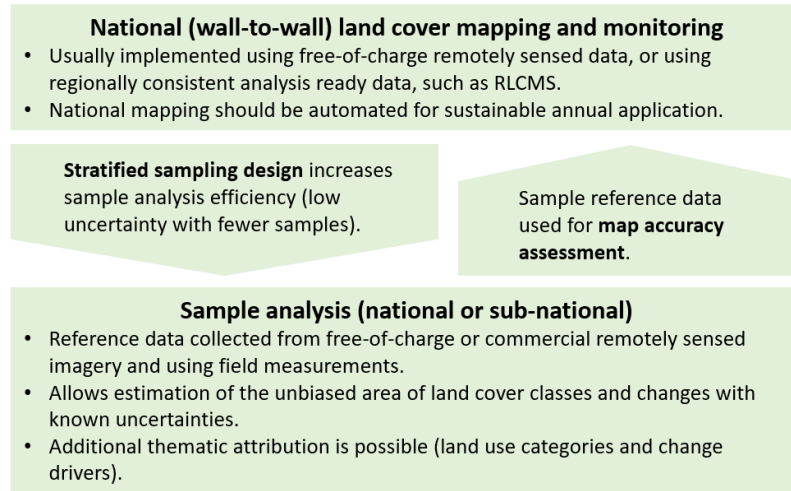
Preface

Timely land cover monitoring is a required precondition to the successful implementation of national policies and international agreements toward the goal of balancing economic development and environmental sustainability. The objectives for the national land cover monitoring include systematic collection, analysis, and dissemination of data. The information provided by the national monitoring supports decision-making at administrative, national, and international levels.

The following principles and good practice guidelines were recommended for national land cover monitoring system design:

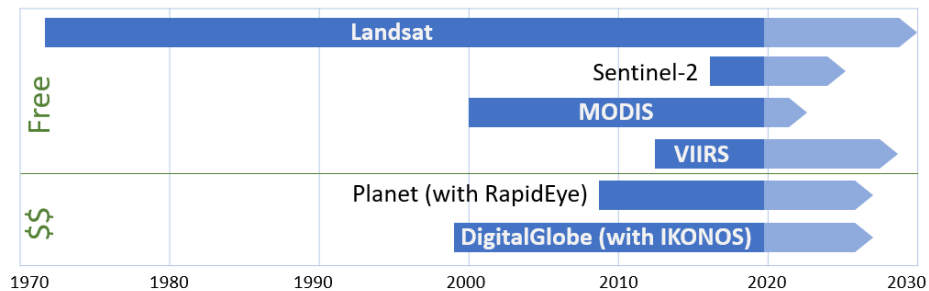
- Robust and consistent methods and data sources.
 - Information on land cover extent and change should be collected at regular intervals over time allowing quantification of changes. Land cover monitoring methods should be suitable for regular updates and historical analysis.
 - The holistic data collection within the theme of interest is recommended. For example, forest monitoring should consider all tree cover, including forest and trees outside forests.
 - The methodology should be flexible to produce information for a variety of users and to be adapted to national standards.
 - Results should be comparable beyond national, administrative, and forest use/ownership boundaries.
 - Quality control mechanisms should be established and implemented throughout the system.
 - Land cover mapping and monitoring results should be validated using sample reference data.
- Non-prohibitive data and data processing costs for national application.
 - Free-of-charge remotely sensed data is required for periodic national-scale updates.
 - Fast, semi-automatic data processing algorithms are preferable.
 - Methods and data sources should be suitable for national capacities and operational implementation.
- Transparency of methods and results.
 - Data analysis and validation methods should be transparent and replicable.
 - Mapping and monitoring results should be publicly available when possible.
- National ownership and responsibility. National implementation of the land cover monitoring system is the key to sustainability and reliability. The preconditions for national application include:
 - Adequate capacity building.
 - Long-term data and resource availability.
 - Preservation of expertise.
 - Gradual development/improvement of the system.

The good practices suggested that the national land cover monitoring system should consist of two components: (1) national scale mapping and (2) sample-based area estimation and uncertainty reporting. These components are interconnected. National mapping provides spatial information and facilitates sample analysis. Sample analysis provides area and uncertainty estimates as well as map accuracy information following established good practices.



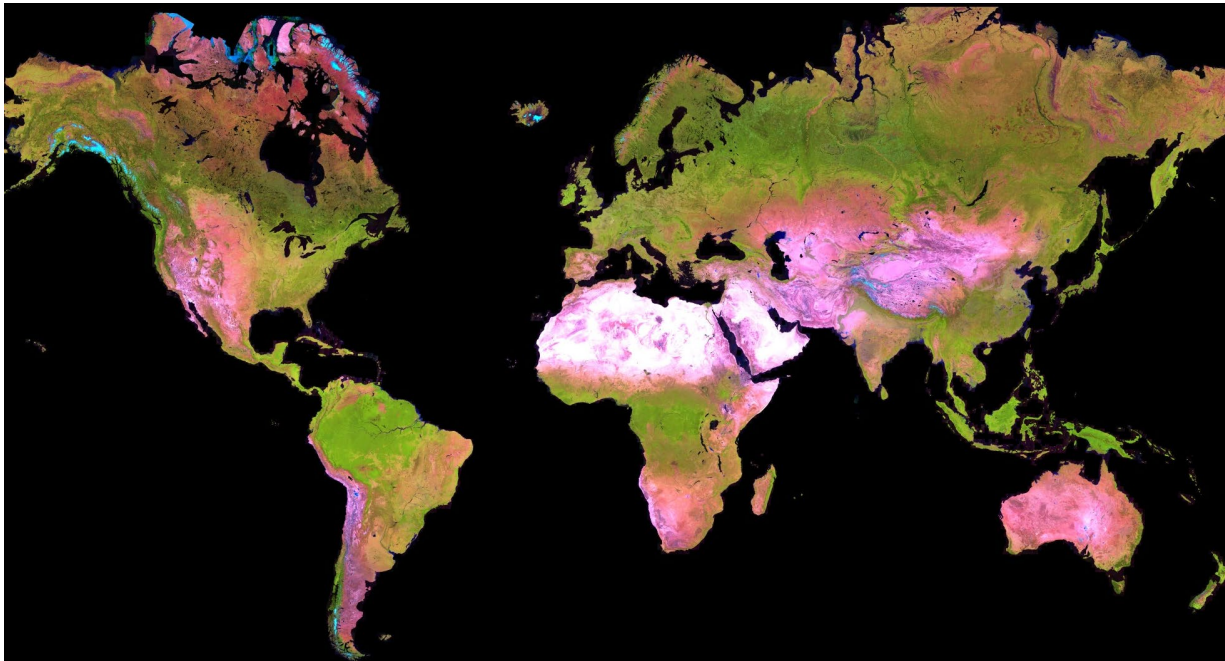
Components of the national land cover monitoring system.

The joint National Aeronautics and Space Administration (NASA) and the United States Geological Survey (USGS) Landsat program provides the longest continuous global archive of satellite earth observation data. Since the early 1980s, Landsat data have been collected at the same spatial resolution (30m per pixel) and with similar spectral bands, enabling a multi-decadal analysis of land cover and land use. The “time machine” capabilities of the Landsat archive are useful for multi-decadal and operational near-real-time land cover and land use change assessment at national to global extents. The free and open data policy and consistent Landsat imagery format support a variety of data applications. With Landsat 8 and 9 in operation and Landsat 10 in development, the mission has a high probability of continuation within the next decade.



Mission timeframe of major EO satellite systems.

The Global Land Analysis and Discovery (GLAD) team at the University of Maryland has developed and implemented an automated Landsat data processing system that generates globally consistent analysis-ready data (GLAD ARD) as inputs for land cover and land use mapping and change analysis. The data processing algorithms have been tested at a global extent for the forest, surface water, croplands, settlements, and vegetation extent mapping and annual monitoring. The GLAD ARD represents a 16-day time series of globally consistent, tiled Landsat normalized surface reflectance from 1997 to the present, updated every 16 days, and suitable for operational land cover change applications. The data are provided free of charge and are available through a dedicated Application Programming Interface (API) at <https://glad.umd.edu/ard/home>. In addition to the ARD dataset, the GLAD team has developed and provided users with a set of tools for time-series data processing, analysis, and machine-learning characterization. Together, the global GLAD ARD and GLAD Tools provide an end-to-end solution for national and regional users for no-cost Landsat-based natural resource assessment and monitoring. The GLAD Tools User's Manual supports the application of the latest version of the GLAD Tools for national and regional land cover characterization, change assessment, and area reporting using mapping and sample analysis techniques.



Cloud-free Landsat GLAD ARD composite for the year 2022.

1. GLAD Tools V2.0 Installation (Windows 10/11)

1.1. System Requirements

- Windows 10 or 11 (64-bit).
- 16 GB RAM (8GB RAM for limited capacity).
- Enough disk volume for data storage and processing. The disk volume requirement depends on the area of analysis and time interval. The following average data volumes may be used to estimate the required disk space:
 - ARD 16-day data for one tile, one year – 5 GB
 - Phenological metrics for one tile, one year – 6.5 GB
 - Change detection metrics for one tile, one year – 12 GB.
- Administrative privileges are required for software installation.

1.2. PERL

PERL is a programming language used for application management in GLAD Tools. Perl advantages include simple coding language, Linux and Windows applications, and no known issues related to compiler versions. Recommended PERL interpreter: <http://strawberryperl.com/>. We recommend downloading and installing the latest release of the 64-bit version. Restart the computer after installing PERL.

Alternative ActiveState PERL interpreter: <https://www.activestate.com/products/perl/>. Follow the website instructions for PERL download and installation. There are no known differences between Strawberry PERL and ActiveState PERL regarding GLAD Tools functionality.

1.3. QGIS and OSGeo4W

QGIS 3 is the recommended software for raster data visualization and collecting training data for supervised classification. OSGeo4W, a part of the QGIS 3 installation, is required for GLAD Tools. The following instructions are for using the latest version of QGIS 3. Due to the frequent update of QGIS and its plugins, the following instructions (May 2023) may be outdated.

1. Download the latest QGIS installer from <https://qgis.org/en/site/forusers/download.html>
2. Install QGIS 3 (default options).
3. Restart your computer.
4. Open QGIS and install the following plugins from the standard plugin depository:
 - Send2GE
 - QuickMapServices
5. Restart QGIS to implement changes.

1.4. Text Editor

The text editor should be suitable for Unix/Windows file editing and (optionally) PERL language style highlighting. We recommend Notepad++, an open-source editor. The latest version of the editor can be downloaded here: <https://notepad-plus-plus.org/downloads/>.

1.5. R

R is a programming language and free software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing. The GLAD Tools require R for sample analysis (sample allocation and reflectance profiles). We recommend installing the latest version of R.

1. Download and install the latest version of R from <https://cloud.r-project.org/> using default installation parameters.
2. Restart your computer after installation is complete.
3. After installing R, the user will need to install two packages required to generate reflectance profiles: **ggplot2** and **dplyr**. Open the R interface and execute the following commands:

```
install.packages("ggplot2", repos="http://cran.rstudio.com/")
install.packages("dplyr", repos="http://cran.rstudio.com/")
```

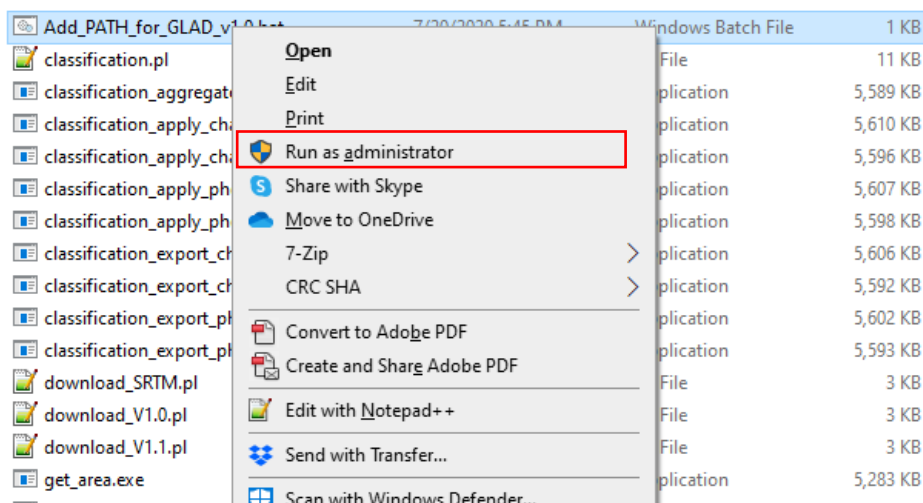
1.6. Google Earth Desktop

Google Earth Desktop provides access to high resolution image time-series that serves as an important reference data for LCLU classification and sample reference data collection. Use the standard software installation provided at <https://www.google.com/earth/versions/#earth-pro>.

1.7. GLAD Tools Installation

The following method is recommended for the installation of GLAD Tools on Windows 10. Users should have administrative privileges to install components.

1. Download the latest complete package of GLAD Tools: https://glad.umd.edu/Potapov/ARD/GLAD_Tools.zip
2. Create folder C:\GLAD_Tools and unpack the content of the zip file into this folder keeping the original subfolder structure and files and folder names. Any other folder structure (renaming the root folder, different drive letters, subfolders, etc.) may prevent the system from working correctly.
3. Add PATH variables to GLAD Tools components. Follow these instructions to complete this step:
 - Open the C:\GLAD_Tools
 - Right-click on the file `Add_PATH_for_GLAD_v1.1.bat` and select the option "Run as Administrator".



- Reboot your computer.

4. Add path to OSGeo4W and R installation. You will need to repeat this step every time the QGIS or R software is updated on your computer. To add, open the file "[C:\GLAD_Tools\dependencies.txt](#)" in a text editor. Find the paths to files OSGeo4w.bat and Rscript.exe and fill in the corresponding parameters in the text file:

```
ogr=C:/Program Files/QGIS 3.16/OSGeo4w.bat  
R=C:/R-4.1.2/bin/Rscript.exe
```

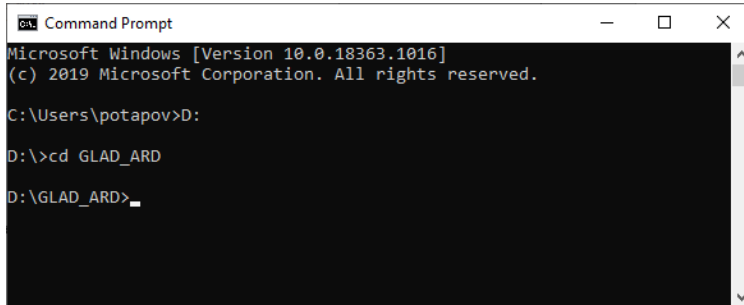
1.8. Software Operations and Troubleshooting

1.8.1. Data visualization in QGIS

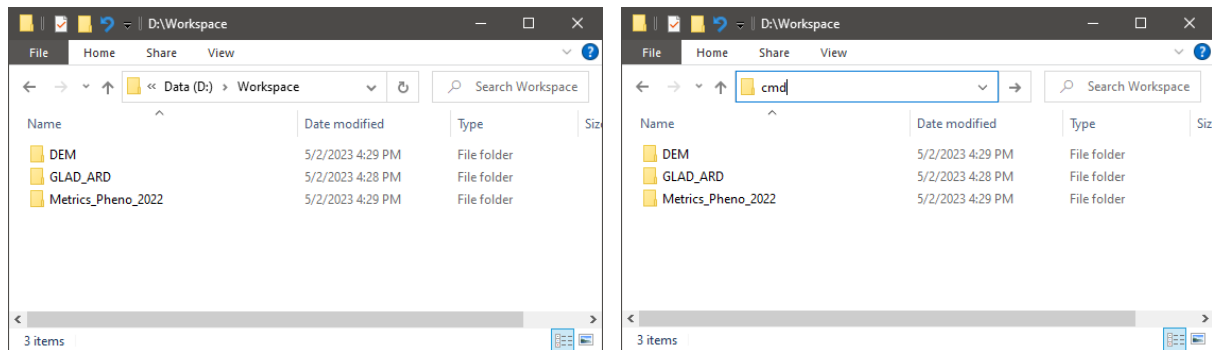
We recommend using QGIS to visualize metric composites, collect training data (shapefile editing) and check the model outputs. If you are not familiar with QGIS, we recommend you use online tutorials such as <https://www.qgis.org/en/site/forusers/trainingmaterial/index.html>

1.8.2. CMD

The CMD (command line interpreter) is required to operate GLAD Tools. The CMD can be opened using Windows Start Menu. To do this, type in the search window "cmd". In the new CMD terminal, you will need to change the drive path and change the directory within the drive to navigate to the work folder.



The fastest way to call the CMD window from a work directory is to type "cmd" in the address bar of the windows explorer. To do this, first, navigate to the work directory, then type "cmd" in the File Explorer address bar and hit "Enter". This will open the CMD window in this work directory.

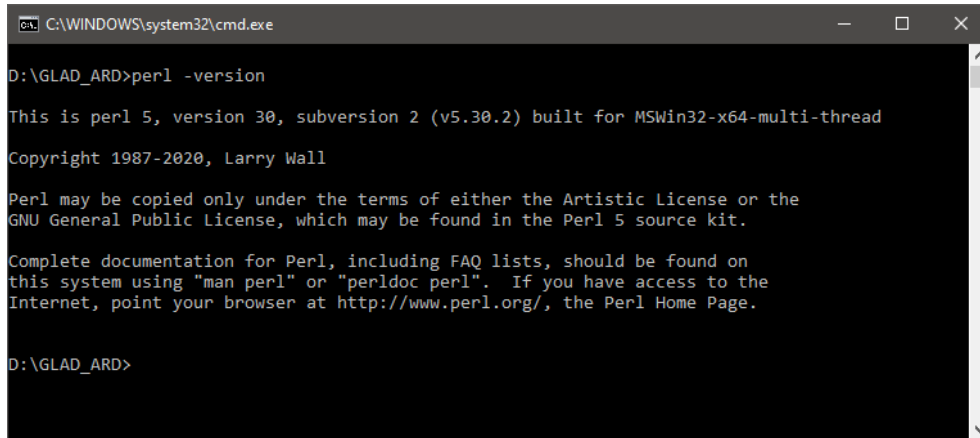


Alternatively, you may add the "Open command window here" command to Windows right-click context menu. To do this, run the file "[C:\GLAD_Tools\Add_CMD_to_Context_Menu.reg](#)" as administrator (you should have administrative privileges). Reboot your computer after running the tool. Now, you may open CMD from the work folder using the Ctrl-Right Click context menu.

1.8.3. Checking PERL Installation

If the GLAD Tools utilities are not working, check if the PERL language interpreter is installed correctly. To check the PERL installation, in the CMD window, type:

```
perl --version
```



```
C:\WINDOWS\system32\cmd.exe
D:\GLAD_ARD>perl -version
This is perl 5, version 30, subversion 2 (v5.30.2) built for MSWin32-x64-multi-thread
Copyright 1987-2020, Larry Wall

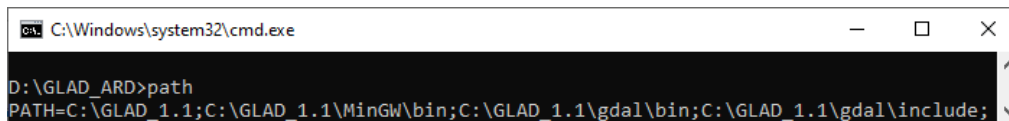
Perl may be copied only under the terms of either the Artistic License or the
GNU General Public License, which may be found in the Perl 5 source kit.

Complete documentation for Perl, including FAQ lists, should be found on
this system using "man perl" or "perldoc perl".  If you have access to the
Internet, point your browser at http://www.perl.org/, the Perl Home Page.

D:\GLAD_ARD>
```

1.8.4. Checking Environmental Variables

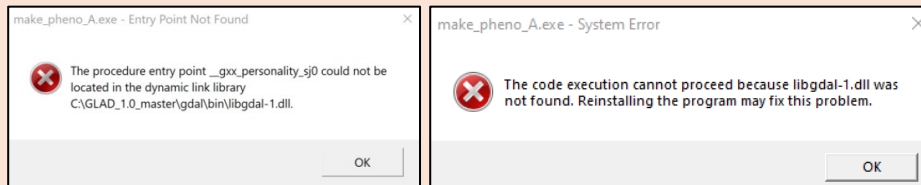
If the GLAD Tools utilities return errors (see examples below) or do not produce any data, check if the environmental variables are correctly set. Use the “path” command in the CMD to check if PATH variables were set correctly:



```
C:\Windows\system32\cmd.exe
D:\GLAD_ARD>path
PATH=C:\GLAD_1.1;C:\GLAD_1.1\MinGW\bin;C:\GLAD_1.1\gdal\bin;C:\GLAD_1.1\gdal\include;
```



A common indicator of incorrect PATH variable value is the following error messages which appeared when a classification or metric generation software is started:



This error indicates that the software was not correctly installed. Repeat installation, check environmental variables, and reboot your computer before using the software.

2. GLAD Landsat ARD

The GLAD ARD represents a 16-day time series of globally consistent, tiled Landsat normalized surface reflectance from 1997 to the present, updated every 16 days, and suitable for operational land cover change applications. The GLAD ARD is a source product for creating cloud-free annual composites and multitemporal metrics for land cover classification, change detection, and image interpretation. The GLAD ARD data is distributed free of charge and without restrictions on subsequent redistribution or use if the proper citation is provided as specified by the Creative Commons Attribution License (CC BY); see Licensing section. The methodology for the GLAD ARD production is provided in Appendix 1 and in Potapov et al., 2020 (the publication is available in the C:\GLAD_Tools\Documentation folder).

2.1. GLAD ARD Raster Data Format

The Landsat ARD data are stored as multi-layer raster tiles. The spatial resolution of the data is 0.00025 degrees per pixel, which corresponds to 27.83 m per pixel on the Equator. The size of one raster tile is 4004x4004 pixels, corresponding to an extent of 1.0005 by 1.0005 degrees. The tile system features a 2-pixel overlap to simplify the parallelization of the focal average computation. The 2-pixel buffer allows implementation of contextual analyses using 3 × 3 and 5 × 5 kernels without the need to read data from multiple tiles at a time.

The ARD product is stored in geographic coordinates using the World Geodetic System (WGS84). The coordinate system is defined by EPSG Geodetic Parameter Dataset:

EPSG:4326 (<https://spatialreference.org/ref/epsg/wgs-84/>)

Alternatively, it can be defined using the PROJ standard (<http://proj.org>):

`+proj=longlat +ellps=WGS84 +datum=WGS84 +no_defs`

The data for each 16-day interval for a tile are stored as 8-band, 16-bit unsigned, LZW-compressed GeoTIFF files. The list of image bands is provided below.

16-day ARD tile image layers

Image band	Image data	Units, data format
1	Blue band (0.48 μm)	Normalized surface reflectance scaled to the range from 1 to 40,000, UInt16
2	Green band (0.56 μm)	
3	Red band (0.66 μm)	
4	NIR band (0.86 μm)	
5	SWIR1 band (1.61 μm)	
6	SWIR2 band (2.20 μm)	
7	Normalized brightness temperature	K × 100, UInt16
8	Observation quality flag (QF)*	QF code, UInt16

* QF codes are provided in Appendix 1

Despite the global radiometric consistency of the 16-day GLAD ARD product, the direct application of this dataset as input to a land cover characterization model is hampered by the irregular frequency of clear-sky observation. The availability of clear-sky observations is a function of the Landsat orbital constellation, data acquisition strategy, and cloud cover. A 16-day composite contains the best quality observation data, but it is not equal to cloud-free data. If only cloud/shadow contaminated data exists for a particular 16-day interval, these data will be retained in the composite. If no data exist, the composite will contain zero values. Annual 16-day time-series for the same area may have dramatically different numbers of clear-sky observations between seasons and years. While 16-day time series data contain sufficient information to identify land cover types and land cover dynamics, the inconsistency of observation frequency may not allow calibration of a regional mapping model using solely ARD as source data.

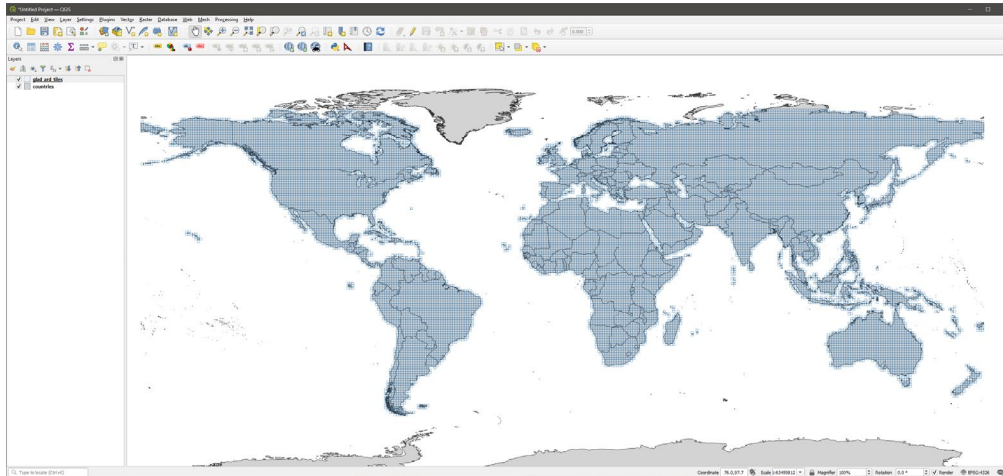
2.2. Global GLAD ARD Tile System

The global Landsat ARD product is provided as a set of 1x1 geographic degrees tiles. The tile format facilitates data handling and the parallelization of data processing, Tile names are derived from the tile center, and refer to their integer value of the tile center degrees. Tile naming example:

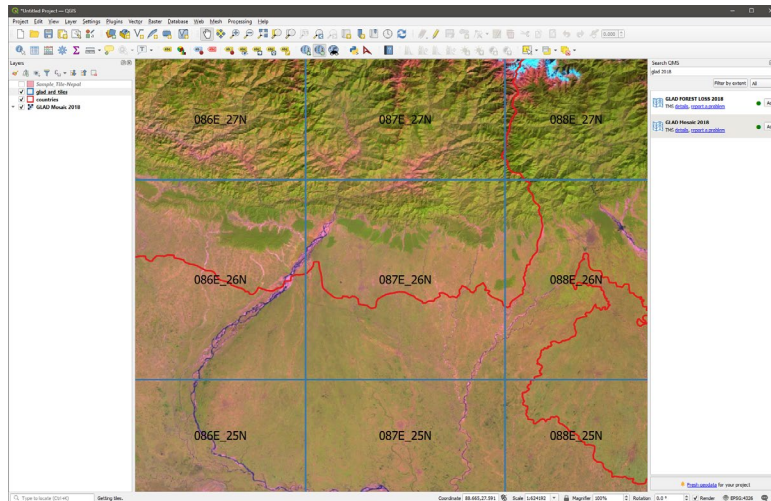
The name of a tile with center 17.5E and 52.5N is 017E_52N.

To select ARD data tiles for your area of analysis, use the tile boundary shapefile located in C:\GLAD_Tools\Data\Global_tiles\glad_ard_tiles.shp

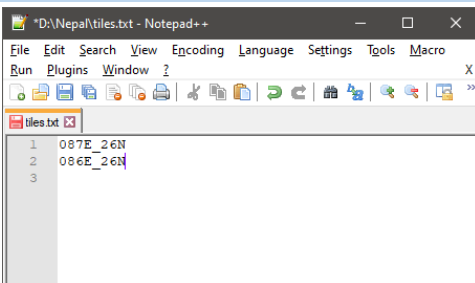
Open the file in QGIS and select the tiles that overlap your area of analysis.




The example on the right shows an overlay of the [glad_ard_tiles.shp](#) with GLAD Mosaic 2018 (annual clear-sky Landsat composite) that is available through the QGIS QMS plugin. The “Tile” field of [glad_ard_tiles.shp](#) provides the tile names. The list of tile names can be copied from QGIS as text or exported as a table.



Example



Selected tiles should be listed in the text file for data download (see the example on the left). We recommend keeping the text file with the tile list in the project work folder.

 Check that the text file does not have empty lines or spaces before/after the tile names.

2.3. GLAD ARD 16-day Interval IDs

Data collected for a single 16-day interval is stored in a single GeoTIFF file. There are 23 intervals per year. The ranges of dates for each interval are provided in the table below:

Start and end days of the year (DOY) for the GLAD ARD 16-day composite intervals.

Interval ID	DOY start	DOY end	Interval ID	DOY start	DOY end
1	1	16	13	193	208
2	17	32	14	209	224
3	33	48	15	225	240
4	49	64	16	241	256
5	65	80	17	257	272
6	81	96	18	273	288
7	97	112	19	289	304
8	113	128	20	305	320
9	129	144	21	321	336
10	145	160	22	337	352
11	161	176	23	353	365 (366)
12	177	192			

Each interval has a unique numeric ID, starting from the first interval of the year 1980. Use the 16-day interval ID table (C:\GLAD_Tools\Documentation\16d_intervals.xlsx) to select intervals for your analysis. The equation below shows how to obtain the interval identification number for a selected year and season.

$$ID = (\text{Year}-1980) \times 23 + \text{Interval}$$

ID – interval identification number (file name), Year – selected year (1980 and later),
Interval – selected annual 16-day interval (1–23).

We recommend using more than one year of data for phenological metrics to implement the gap-filling of missing data. For phenological metrics, at least one year of data is required, and up to four preceding years may be used for gap-filling. Change detection metrics require data for the target and three preceding years.

The following example demonstrates how to select the optimal time interval for land cover mapping using phenological metrics:

For the 2019 forest mapping, we select 16-day intervals for the year 2019 (898-920) and intervals for the four preceding years, 2015-2018 (806-897). The overall interval is 806-920.

16-day interval IDs for the years 1997-2027

Year	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1997	392	393	394	395	396	397	398	399	400	401	402	403	404	405	406	407	408	409	410	411	412	413	414
1998	415	416	417	418	419	420	421	422	423	424	425	426	427	428	429	430	431	432	433	434	435	436	437
1999	438	439	440	441	442	443	444	445	446	447	448	449	450	451	452	453	454	455	456	457	458	459	460
2000	461	462	463	464	465	466	467	468	469	470	471	472	473	474	475	476	477	478	479	480	481	482	483
2001	484	485	486	487	488	489	490	491	492	493	494	495	496	497	498	499	500	501	502	503	504	505	506
2002	507	508	509	510	511	512	513	514	515	516	517	518	519	520	521	522	523	524	525	526	527	528	529
2003	530	531	532	533	534	535	536	537	538	539	540	541	542	543	544	545	546	547	548	549	550	551	552
2004	553	554	555	556	557	558	559	560	561	562	563	564	565	566	567	568	569	570	571	572	573	574	575
2005	576	577	578	579	580	581	582	583	584	585	586	587	588	589	590	591	592	593	594	595	596	597	598
2006	599	600	601	602	603	604	605	606	607	608	609	610	611	612	613	614	615	616	617	618	619	620	621
2007	622	623	624	625	626	627	628	629	630	631	632	633	634	635	636	637	638	639	640	641	642	643	644
2008	645	646	647	648	649	650	651	652	653	654	655	656	657	658	659	660	661	662	663	664	665	666	667
2009	668	669	670	671	672	673	674	675	676	677	678	679	680	681	682	683	684	685	686	687	688	689	690
2010	691	692	693	694	695	696	697	698	699	700	701	702	703	704	705	706	707	708	709	710	711	712	713
2011	714	715	716	717	718	719	720	721	722	723	724	725	726	727	728	729	730	731	732	733	734	735	736
2012	737	738	739	740	741	742	743	744	745	746	747	748	749	750	751	752	753	754	755	756	757	758	759
2013	760	761	762	763	764	765	766	767	768	769	770	771	772	773	774	775	776	777	778	779	780	781	782
2014	783	784	785	786	787	788	789	790	791	792	793	794	795	796	797	798	799	800	801	802	803	804	805
2015	806	807	808	809	810	811	812	813	814	815	816	817	818	819	820	821	822	823	824	825	826	827	828
2016	829	830	831	832	833	834	835	836	837	838	839	840	841	842	843	844	845	846	847	848	849	850	851
2017	852	853	854	855	856	857	858	859	860	861	862	863	864	865	866	867	868	869	870	871	872	873	874
2018	875	876	877	878	879	880	881	882	883	884	885	886	887	888	889	890	891	892	893	894	895	896	897
2019	898	899	900	901	902	903	904	905	906	907	908	909	910	911	912	913	914	915	916	917	918	919	920
2020	921	922	923	924	925	926	927	928	929	930	931	932	933	934	935	936	937	938	939	940	941	942	943
2021	944	945	946	947	948	949	950	951	952	953	954	955	956	957	958	959	960	961	962	963	964	965	966
2022	967	968	969	970	971	972	973	974	975	976	977	978	979	980	981	982	983	984	985	986	987	988	989
2023	990	991	992	993	994	995	996	997	998	999	1000	1001	1002	1003	1004	1005	1006	1007	1008	1009	1010	1011	1012
2024	1013	1014	1015	1016	1017	1018	1019	1020	1021	1022	1023	1024	1025	1026	1027	1028	1029	1030	1031	1032	1033	1034	1035
2025	1036	1037	1038	1039	1040	1041	1042	1043	1044	1045	1046	1047	1048	1049	1050	1051	1052	1053	1054	1055	1056	1057	1058
2026	1059	1060	1061	1062	1063	1064	1065	1066	1067	1068	1069	1070	1071	1072	1073	1074	1075	1076	1077	1078	1079	1080	1081
2027	1082	1083	1084	1085	1086	1087	1088	1089	1090	1091	1092	1093	1094	1095	1096	1097	1098	1099	1100	1101	1102	1103	1104

2.4. Data Downloading using GLAD Tools

The global Landsat ARD is available for download using dedicated ARI. The API is a tool for a single 16-day image composite access. The curl command for data access through API is provided at the end of this section. To simplify data access, we provide a set of PERL codes that automate download for a selected area and time interval. The following instructions present the main steps of data access: defining dataset extent, selecting time intervals, and downloading ARD data and SRTM topography metrics using GLAD Tools.

The folder structure is important for the GLAD ARD because 16-day interval data files have the same file names and data for each tile must be stored in a separate folder. The root folder for the ARD data should be created before the download starts. If you are downloading data for the first time or for a new project, create a new folder where data will be stored. Make sure you have enough disk space. The data volume for one year for a single 1x1 degree tile is 5 GB on average. In the case when you are adding data to the existing project, use the name of the folder with previously downloaded 16-day composites. The code will skip data downloaded earlier and will only acquire new data. It is a good practice to keep all downloaded ARD data in the same destination root folder to avoid data duplication.

The tile list (text format, see section 2.1) and the list of 16-day intervals (see section 2.2) should be defined before the download begins. To start data downloading, open the CMD terminal and use the following command:

```
perl C:/GLAD_Tools/download_ARD.pl <tile list> <start int> <end int> <folder>
<tile list> - text file with the list of ARD tiles
<start int> <end int> - a range of 16-day intervals (start, end).
<folder> - output folder to store downloaded ARD data.
```

Example

```
>perl C:/GLAD_Tools/download_ARD.pl tiles.txt 806 920 D:/Data
```

2.5. Downloading SRTM Data

The SRTM data (extracted from NASA product SRTMGL1v003) includes elevation, slope, and aspect at Landsat pixel resolution. The data is used as inputs to most classification and change detection models. To download data, a user required the API credentials and the tile list in text format (the same tile list that was used for the ARD data download). It is recommended to keep the SRTM data in a separate root folder from the Landsat ARD data. To start data downloading, open the CMD terminal and use the following command:

```
perl C:/GLAD_Tools/download_SRTM.pl <tile list> <folder>
<tile list> - text file with the list of ARD tiles
<folder> - output folder to store downloaded SRTM data
```

Example

```
>perl C:/GLAD_Tools/download_SRTM.pl tiles.txt D:/SRTM
```


2.6. Data Access Through API

The following instructions are for advanced users who want to create their own data download codes and/or troubleshoot data access issues. The example of data download command provided for a single tile/single interval using curl (command line tool for data transfer that is available on Windows 10 since 2019).

```
curl -u <username>:<password> -X GET https://glad.umd.edu/dataset/glad_ard2/<lat>/<tile>/<interval>.tif -o <outfolder>/<interval>.tif
```

Required parameters:

<username> and <password> - default API credentials (Username: glad; Password: ardpas)

<tile> - ARD tile name

<lat> - tile latitude, the second half of the ARD tile name (e.g., for the 105E_13N, <lat> is 13N)

<interval> - unique 16-day interval ID.

<outfolder> - output folder. The folder must exist. Make sure that each tile is stored in a separate folder, otherwise, the data will be overwritten.

Example

```
>curl -u username:password -X GET https://glad.umd.edu/dataset/glad_ard2/26N/086E_26N/920.tif -o D:/Data/086E_26N/920.tif
```

2.7. Guidelines for Data Organization

The ARD data is sensitive to the way it is organized by the user after downloading. Incorrect data organization may cause errors in software applications. Here are several tips on how to organize the ARD data and manage ARD-based projects.

1. The downloaded ARD data should be stored in a separate folder and organized by sub-folders named by corresponding tiles (e.g., `D:\GLAD_Workspace\ARD\086E_26N`). The 16-day intervals for different tiles have the same names and cannot be stored in the same folder. The name of the folder should be exactly the same as the tile name, otherwise, the software will not be able to locate the files.

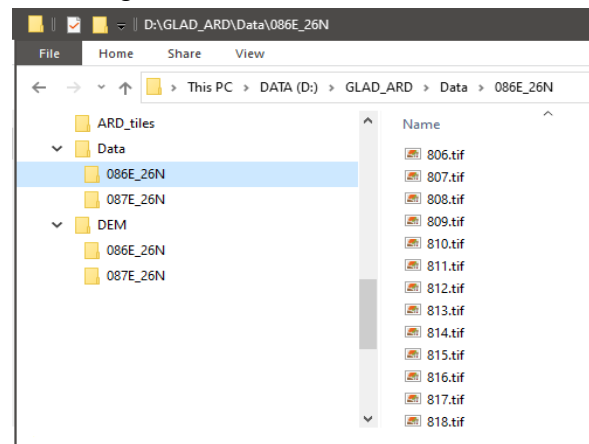
The SRTM data should be stored in a separate folder, with data separated by tile subfolders (e.g., `D:\GLAD_Workspace\DEM\086E_26N`).



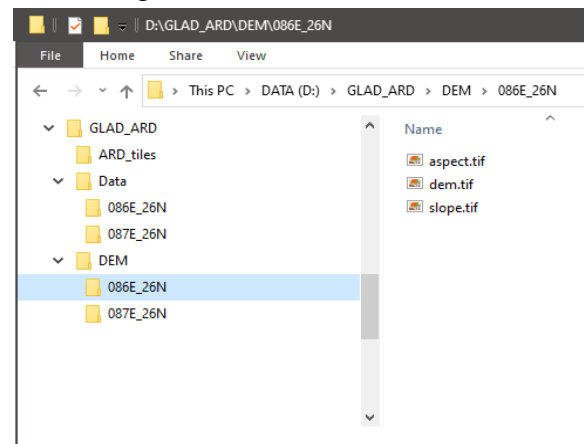
The path to ARD data storage or processing folders should not have spaces in folder names.

Example

ARD storage:



SRTM storage:



2. The Phenological and Change Detection metrics should be stored in separate subfolders for each year and each tile (e.g., [C:\GLAD_Workspace\Metrics_Pheno_2012\105E_09N](#)).
3. We recommend using a separate folder for each ARD-based project. The project should have all the required parameter files (tile list, parameter files for GLAD Tools, and temporal files created by the software). This is especially important for classification projects. Each classification project should be performed in a separate folder to avoid overwriting the results of another project. The classification folder should include training files, parameter files, and the list of tiles. Each sample set should also be created in a separate folder. Image mosaics may be stored in the same folder given that the user specified a unique output name for each mosaic. Multitemporal classifications are stored separately for each thematic class, but the years are stored together.
3. The GDAL Tools should be called either (a) from a folder with parameter files (for data download and creating metrics) or (b) from the folder specifically created for the output products (image mosaics, classification, analysis, and sampling). If a tool is called from the wrong folder, the output files may re-write existing data and corrupt your project. If a tool is called from [C:\GLAD_Tools](#) folder, that may corrupt the software.
4. The user should follow several rules when creating folder and file names and writing file path for the parameter files. Because the Tools were initially created for Linux OS, not all folder names permitted by the Windows system will be suitable. The folder and file names should only contain letters, numbers, dashes, and underscore symbols. The spaces in the folder or file names will cause the software to crash. For the parameter files or in-line command parameters, we recommend always using common slash "/" instead of backslash "\". While most of the tools are designed to handle backslash in the path names, it still may cause software errors.

3. Multitemporal Metrics

The multi-temporal metrics method is a time-series data transformation that improves spatial and temporal consistency, simplifies phenological analysis, and facilitates land cover mapping and change detection at large geographic extents. The metrics approach helps to overcome the inconsistency of clear-sky data availability that is typical for sensors with low observation frequency, such as Landsat. ARD-based multitemporal metrics represent a set of data distribution and vegetation phenology statistics extracted from a 16-day observation time series. The GLAD Tools provides software to calculate two types of metrics, annual phenological metrics and annual change detection metrics, designed for the two most common objectives: annual land cover mapping and detection of land cover changes between two consecutive years. A detailed explanation of the metric algorithm is provided in Appendix 2. The full list of metrics is provided in the MS Excel table format in [C:\GLAD_Tools\Documentation](#)

3.1. Phenological Metrics

The annual phenological metrics serve as source data for land cover, land use, and vegetation structure mapping models. The annual phenological metrics represent a set of reflectance data distribution and vegetation phenology statistics extracted from the observation time series. Using data distribution statistics, rather than 16-day observation time series, as input features for tree canopy structure modeling reduce the effect of inconsistent cloud-free image availability between regions and years and support empirical model calibration and application for multiple years. To create an annual phenological metric set we use only clear-sky observations from the 16-day GLAD ARD composites. If a gap in clear-sky observations is longer than two months, the software uses data from a preceding year to fill the gap to ensure the consistency of the land surface phenology information. The user can specify the number of preceding years that will be used for gap-filling, from 0 (no gap-filling) to 4 years.

The phenological metrics consist of four sets of data distribution statistics. The first set represents statistics extracted from the annual distribution of normalized surface reflectance and vegetation indices values and includes minimum and maximum values, quartiles, interquartile averages, and amplitudes. The second set

represents seasonality metrics, calculated as data distribution statistics derived from the observation time series ranked by the vegetation index and brightness temperature values (e.g., the value of the red band corresponding to the annual maximum surface temperature). The third set consists of the vegetation phenology statistics (start, peak, and end of the growing season) based on the time series of the normalized difference vegetation index (NDVI). The growing season boundaries for phenology statistics were defined as an interval between the beginning of NDVI consistent increase and the end of NDVI consistent decrease. The fourth set includes technical metrics that reflect data quality (observation number, water proportion, and quality flag). The metrics are stored as single-band 16-bit unsigned GeoTIFF files using the same tile system as the ARD (see Section 2.3). The metrics sets for each tile are stored in separate folders.

The metric naming convention is based on the bands, indices, and statistic abbreviation (shown in the table below in blue). Each metric name starts with the year of the metric set.

For the annual reflectance/index value distribution metrics, we use the following naming convention:

YYYY_Band_Statistic.tif

YYYY – Corresponding year.

Band – Spectral band or index.

Statistic – Statistic extracted from the observation time series.

The annual seasonality metrics use the following naming convention:

YYYY_Band_Statistic_Method.tif

YYYY – Corresponding year.

Band – Spectral band or index.

Statistic – Statistic extracted from the observation time series.

Method – Distribution method.

Example

2018_blue_max_RN.tif - The metric represents the value of the normalized surface reflectance of the Landsat blue band for the 16-day interval that has the highest red/NIR normalized ratio (also known as NDVI) value during the year 2018.

Not all the metrics are recorded to disk. Specifically, the amplitude metrics are calculated in memory during the classification procedure. To include spatial context in image classification, the focal mean for each of the metrics using a 3x3 kernel is calculated during the classification routine. The complete set of phenological metrics is explained in the following supplementary table: "C:\GLAD_Tools\Documentation\Metrics_pheno.xlsx".

Phenological metric naming convention

Annual spectral reflectance/index value distribution metrics

<i>Spectral Bands</i>	<i>Distribution statistics</i>
Blue (482 nm) [blue]	Minimum [min]
Green (561 nm) [green]	Maximum [max]
Red (654 nm) [red]	Median [median]
Near Infrared (864 nm) [nir]	Average between min and Q1 [avmin25]
Shortwave Infrared 1 (1609 nm) [swri1]	Average between Q3 and max [av75max]
Shortwave Infrared 2 (2201 nm) [swri2]	Average between Q1 and Q3 [av2575]
	Average of all values [avminmax]
	Standard Deviation [sd]
	Total Absolute Difference [absdif]

<i>Derived Indices</i>	<i>Amplitudes*</i>
$(\text{NIR}-\text{Green})/(\text{NIR}+\text{Green})$ [GN]	max - min
$(\text{NIR}-\text{Red})/(\text{NIR}+\text{Red})$ [RN]	av75max - avmin25
$(\text{NIR}-\text{SWIR1})/(\text{NIR}+\text{SWIR1})$ [S1N]	max - median
$(\text{NIR}-\text{SWIR2})/(\text{NIR}+\text{SWIR2})$ [S2N]	
$(\text{SWIR1}-\text{SWIR2})/(\text{SWIR1}+\text{SWIR2})$ [S1S2]	
Spectral variability index** [SVVI]	
Tasseled Cap Greenness*** [TCG]	

Annual seasonality metrics

<i>Spectral Bands</i>	<i>Distribution methods</i>	<i>Distribution statistics</i>
Blue [blue]	<i>(observation dates distributed by the following corresponding inputs)</i>	Minimum [min]
Green [green]		Maximum [max]
Red [red]		Average between min and Q1 [avmin25]
Near Infrared [nir]		Average between Q3 and max [av75max]
Shortwave Infrared 1 [swri1]	$(\text{NIR}-\text{Red})/(\text{NIR}+\text{Red})$ [RN]	
Shortwave Infrared 2 [swri2]	$(\text{NIR}-\text{SWIR2})/(\text{NIR}+\text{SWIR2})$ [S2N]	
	Brightness temperature [LST]	
		<i>Amplitudes*</i>
		max - min
		av75max - avmin25

Phenology metrics (based on the annual NDVI time series)

Start of season value [RNph_sos]
End of season value [RNph_eos]
Start of season slope [RNph_sos_slope]
End of season slope [RNph_eos_slope]
Start of season amplitude [RNph_sos_amp]
End of season amplitude [RNph_eos_amp]
Growing season average [RNph_ave]
Growing season total [RNph_sum]

Technical metrics

Number of observations used [TEC_count]
Percent water observations [TEC_prcwater]
Data quality flag [TEC_pf]

*Amplitude metrics are not recorded as files.

** See <https://www.sciencedirect.com/science/article/pii/S0034425716302693>

*** Tasseled Cap coefficients are from <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0147121#pone.0147121.s002>

3.2. Change Detection Metrics

The annual change detection metrics are designed to highlight inter-annual changes in spectral reflectance while reducing false detections due to reflectance fluctuations and inconsistent clear-sky observations availability. Each annual change detection metric set is created from the 16-day clear-sky observations of the corresponding and preceding years; the preceding year spectral data are aggregated over three years, taking the mean reflectance value for each 16-day interval. To create the interannual change detection metrics, the software calculates distribution statistics for spectral bands and index values separately for the corresponding and preceding years and the difference for each statistic value between these years. To highlight changes in seasonal reflectance, we compute differences in spectral reflectance and indices values between the corresponding and preceding years for each 16-day interval and extract selected data distribution statistics from the time series of difference values. Finally, we calculate the slope of the linear regression between the spectral value and observation date for the entire interval that includes preceding and corresponding years. The change detection metrics also include technical metrics that reflect data quality.

The metric naming convention is based on the bands, indices, and statistic abbreviation (shown in the table below in blue). Each metric name starts with the year of the metric set.

For the interannual reflectance/index value distribution metrics, we use the following naming convention:

YYYY_Band_Year_Statistic.tif

YYYY – Corresponding year.

Band – Spectral band or index.

Year – corresponding (c) or preceding (p) year identifier.

Statistic – Statistic extracted from the observation time series.

For the interannual seasonality metrics, we use the following naming convention:

YYYY_Band_Year_Statistic_Method.tif

YYYY – Corresponding year.

Band – Spectral band or index.

Year – corresponding (c) or preceding (p) year identifier.

Statistic – Statistic extracted from the observation time series.

Method – Distribution method.

Spectral data and indices

Spectral Bands

Blue [blue]
Green [green]
Red [red]
Near Infrared [nir]
Shortwave Infrared 1 [swri1]
Shortwave Infrared 2 [swri2]

Derived Indices

$(\text{NIR-Red})/(\text{NIR+Red})$ [NDVI]
$(\text{NIR-SWIR1})/(\text{NIR+SWIR1})$ [NDWI]
$(\text{SWIR1-SWIR2})/(\text{SWIR1+SWIR2})$ [SWSW]

Technical metrics

Number of observations used [count]
Percent water observations [prcwater]
Percent land observations [prcland]
Data usability code [code]
Data quality flag [pf]

Change detection metric naming convention
Statistics

Interannual metrics for the corresponding [c] and preceding [p] years

Distribution statistics

Minimum [min]
Maximum [max]
Second lowest value [smin]
Second highest value [smax]
Median [median]
Average of all values [ave]
Average of all values except min and max [avett]
Last annual observation [last]

Annual seasonality metrics

Distribution methods

$(\text{NIR-Red})/(\text{NIR+Red})$ [NDVI]
Brightness temperature [LST]

Distribution statistic

Minimum [min]
Maximum [max]
Second lowest value [smin]
Second highest value [smax]
Median [median]

Amplitudes*

Difference between all metrics extracted from corresponding and preceding years.

Seasonal reflectance change (calculated from the per-16-day difference time series)

Minimum [min]
Maximum [max]
Second lowest value [smin]
Second highest value [smax]
Average of all values [ave]
Average of all values except min and max [avett]
Value after minimum [amin]
Value after maximum [amax]

Interval time series

Slope of linear regression of band/index value vs. observation date [reg]
Standard deviation of the band/index value [sd]

*Amplitude metrics are not recorded as files.

For the seasonal reflectance change metrics, we use the following naming convention:

YYYY_Band_diff_Statistic.tif

YYYY – Corresponding year.

Band – Spectral band or index.

diff – identified seasonal change metric.

Statistic – Statistic extracted from the observation time series.

For the interval time-series metrics, we use the following naming convention:

YYYY_Band_diff_Statistic.tif

YYYY – Corresponding year.

Band – Spectral band or index.

Statistic – Statistic extracted from the observation time series.

3.3. Annual Clear Sky Composite

GLAD Tools provides an instrument to create cloud-free composites for data visualization. The annual composite represents a subset of phenological metrics that are designed for data visualization only, and not suitable for image classification. The composite includes spectral band reflectance average between Q1 and Q3 (av2575) metrics and technical metrics from the phenological metric set (section 3.1).

3.4. Using GLAD Tools to Calculate Metrics

Metric processing is fully automated and requires minimal user input. The following workflow is the same to generate any metric set:

- Download all required 16-day composites.
- Make a list of ARD tiles to process (single column, tile names only).
- Make a parameter file following a template
- Apply the metric generation code.

Before running the metric generation code, check that the disk space is sufficient for metric storage. The following estimates suggest the average size of a metric set for one year and one ARD tile:

- Clear sky composite: 210 MB
- Phenological metrics: 6.5 GB
- Change detection metrics: 12 GB

The parameter file defines all variables for metrics generation process. The file should have the following structure:

mettype=pheno	Metric type (possible values pheno , change , composite)
tilelist=tiles.txt	Name of the ARD tile list file
year=2022	Target year
input=D:/ARD_Data	Input ARD 16-day data folder
output=D:/Metrics_Pheno_2022	Output folder (will be created if new)
threads=1	Number of parallel processes *
gapfill=4	Number of years to use for gap-filling (values 0 ... 4). The default value is 4. Required only for metric types pheno and composite**

* The number of parallel processes should be set to 1 unless a computer has a multi-core processor (e.g., Intel Xeon) and available RAM is suitable for several processes simultaneously.

**Gap-filling is a process of filling missing 16-day interval data with clear-sky data from a preceding year. By default, four preceding years are used to fill the gap. The value 0-4 defined how many preceding years will be used. Value 0 means that gap-filling is not performed and only data from the current year is used.

The folder [C:\GLAD_Tools\Examples\metrics](#) provides a set of example files to generate all types of multitemporal metrics.

To start metric processing, open the CMD terminal and use the following command:

```
perl C:/GLAD_Tools/build_metrics.pl <parameter file>
<parameter file> - the name of the parameter file.
```

Example

Tile list file
tiles.txt

```
File Edit Format View Help
087E_26N
086E_26N
Windows (CRLF) UTF-8
```

Parameter file
metrics_pheno.txt

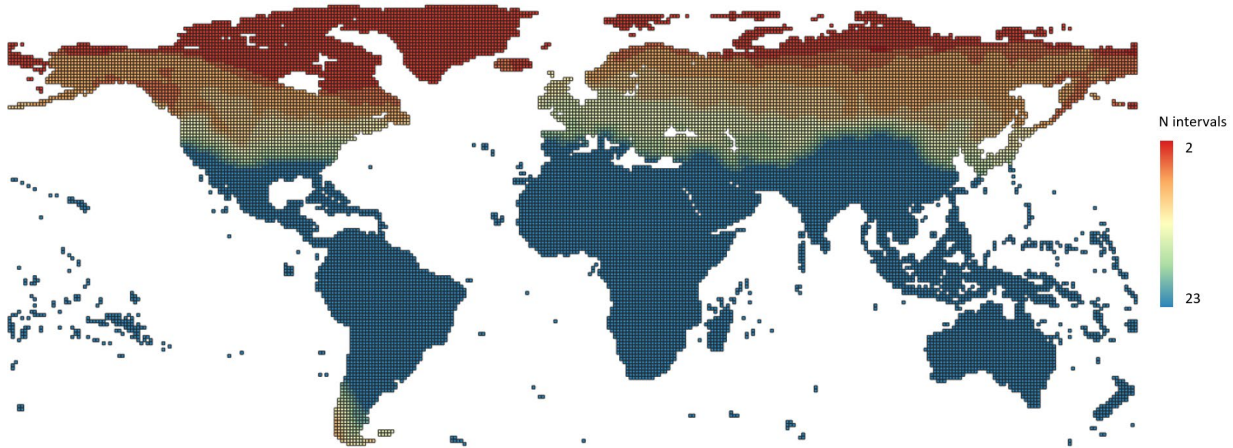
```
File Edit Format View Help
mettype=pheno
tilelist=tiles.txt
year=2022
input=D:/ARD
output=D:/Metrics_Pheno_2022
threads=4
gapfill=4
Ln 100% Unix (LF) UTF-8
```

CMD command

```
>perl C:/GLAD_Tools/build_metrics.pl metrics_pheno.txt
```

3.5. Seasonal Limits for Annual ARD Selection

The purpose of the multitemporal metrics is to map land cover and land use during the growing season, hence images affected by seasonal snow cover should be excluded from processing. To exclude 16-day intervals affected by seasonal snow cover we analyzed snow presence using the MODIS/Terra Snow Cover Monthly L3 Global product (<https://nsidc.org/data/MOD10CM/versions/6>) and Landsat imagery. We excluded all 16-day intervals that feature seasonal snow cover. The lookup table supplied with the tools ("C:\GLAD_Tools\Data\Global_tiles\tiles_seasonal_db.txt") contains information about the growing season interval for each tile.



The number of growing season intervals in the lookup table for each tile.

4. Multitemporal Metrics Visualization

The multitemporal metrics are stored as 1x1 degree tiles. To visualize data for a large region, tiles must be mosaicked together. OSGeo4W and GDAL Tools provide several solutions to mosaic the data as a Virtual Raster Table (VRT).

4.1. Metrics Mosaicking Using GLAD Tools

Before running the mosaicking tool, check the following preconditions:

- The metrics must be created before image mosaicking. Image mosaicking directly from the ARD 16-day data is not supported.
- The `C:\GLAD_Tools\dependencies.txt` file has a correct link to the `OSGeo4w.bat` software (a part of the QGIS installation) that is required for the GLAD Tools (see section 1.6).

The following instructions can be used to combine any set of phenological or change detection metrics into a seamless multi-band mosaic. The metrics should be created before the mosaicking. To specify the mosaic extent, GLAD Tools use the list of tiles (same format as for the ARD download and metric generation). The parameter file is required for the tool. The file should have the following structure:

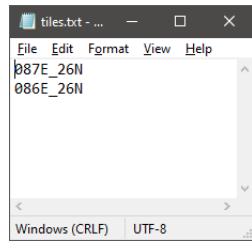
<code>source=D:/Metrics_pheno_2022</code>	Source folder
<code>list=tiles.txt</code>	The name of the tile list
<code>year=2022</code>	Year
<code>outname=median654</code>	Output name
<code>bands=swir1_av2575, red_av2575, nir_av2575</code>	List of metrics to aggregate (comma separated)

To create a VRT mosaic, open the CMD terminal and use the following command:

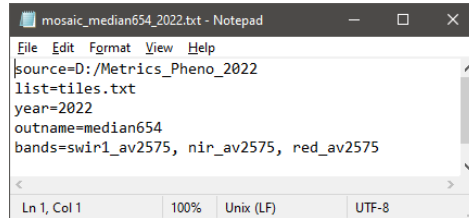
```
perl C:/GLAD_Tools/ mosaic_vrt.pl <parameter file>
<parameter file> - the name of the parameter file.
```

Example

Tile list file
tiles.txt

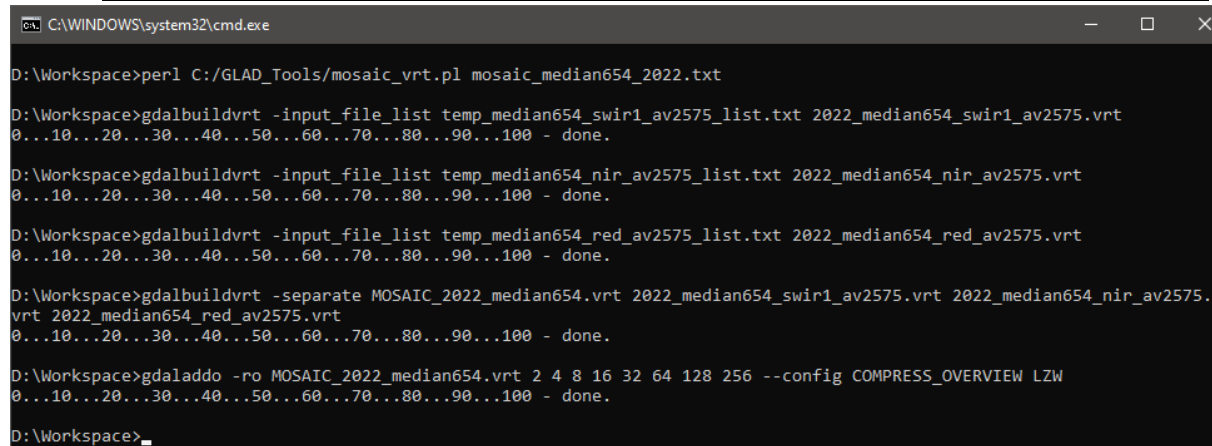


Parameter file
metrics_pheno_A.txt



CMD command

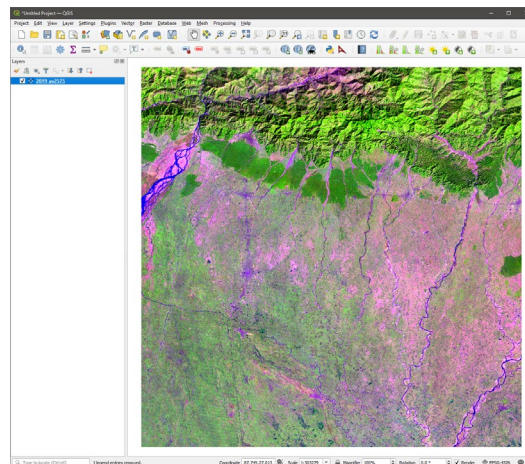
```
>perl C:/GLAD_Tools/mosaic_vrt.pl param_mosaic_av2775.txt
```



The output mosaic is stored as a set of VRT files for each band and the MOSAIC VRT file that includes all bands

- 2022_median654_nir_av2575.vrt – part of the mosaic dataset*
- 2022_median654_red_av2575.vrt– part of the mosaic dataset*
- 2022_median654_swir1_av2575.vrt– part of the mosaic dataset*
- MOSAIC_2022_median654.vrt.ovr– part of the mosaic dataset*
- MOSAIC_2022_median654.vrt – mosaic file**




Open the **MOSAIC_2022_median654.vrt** in QGIS

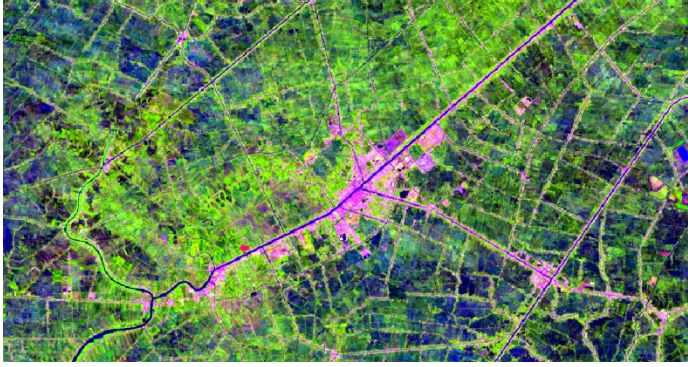



4.2. Phenological Metrics Visualization

Phenological metrics support the interpretation and mapping of different types of land use and land cover. These different land cover types required different metric visualization strategies. For example, visualization of high NDVI seasonality metrics supports mapping stable (minimal annual) surface water extent, while the low NDVI composite is the best to map the maximum surface water extent within the year. The users are encouraged to test different metric combinations to select the one that is most suitable for their applications.

A set of parameter files in C:\GLAD_Tools\Examples\phenological_mosaic provides examples for different visualization. Here are some examples of Landsat band combinations based on the interquartile average metric (av2575):

RGB combination	Mosaic parameter file	Visualization
R: red G: green B: blue	mosaic_median432_2022.txt	
R: NIR G: red B: green	mosaic_median543_2022.txt	
R: NIR G: SWIR1 B: SWIR2	mosaic_median567_2022.txt	

RGB combination	Mosaic parameter file	Visualization
R: SWIR1 G: NIR B: red	mosaic_median654_2022.txt	
R: SWIR2 G: SWIR1 B: red	mosaic_median764_2022.txt	

The seasonality metrics may be most important to map land cover and land use types that feature seasonal changes, e.g., crop types (both images in NIR-SWIR1-SWIR2 band combination)

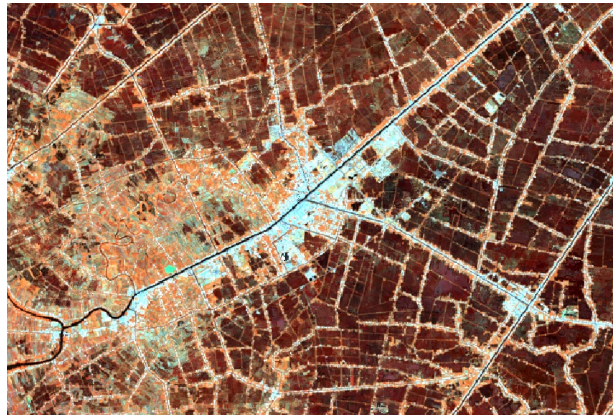
Metric: av75max_RN

Parameter file: mosaic_highNDVI567_2022.txt

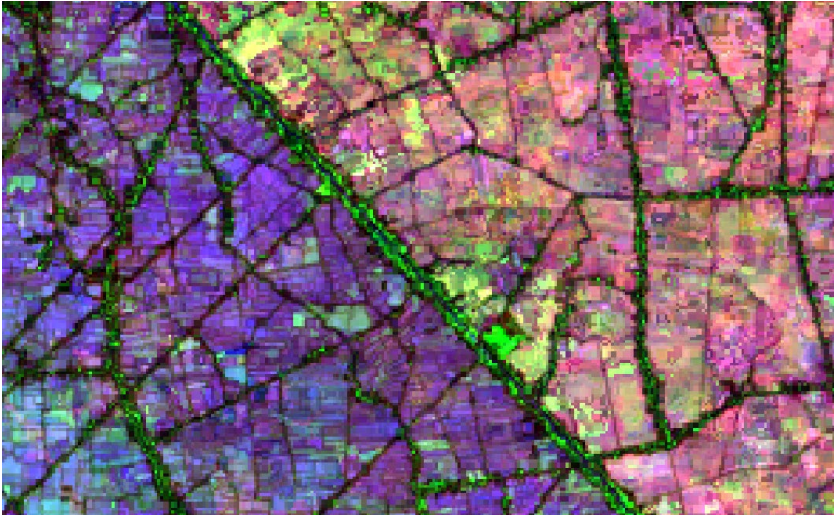


Metric: avmin25_RN

Parameter file: mosaic_lowNDVI567_2022.txt



Finally, the band composites that use amplitudes and indices may support the interpretation of certain land cover types, such as an example below where amplitude metrics highlight the difference between paddy rice and paddy rice/aquaculture rotation systems:



Parameter file:
mosaic_amplitude_2022.txt

R: nir_absdif
G: swir2_absdif
B: RN_absdif

4.3. Change Detection Metrics Visualization

Change detection metrics may be visualized using two approaches. First, the reflectance metrics of current and preceding years can be visualized separately and compared in the GIS system. Second, visualization of the metrics that highlight per-16-day composite (seasonal) changes in spectral reflectance simplifies change interpretation. A set of example parameter files in C:\GLAD_Tools\Examples\change_detection_mosaics provides examples for different visualization.

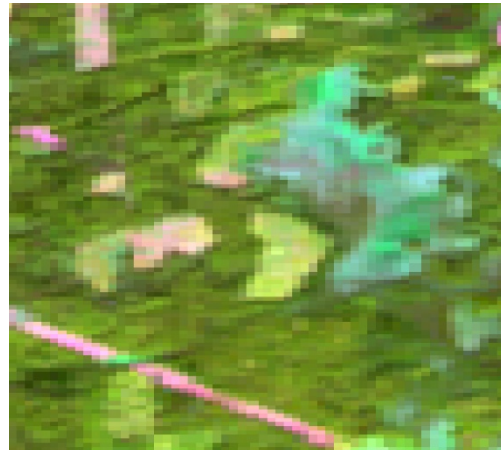
The following example shows two separate composites: one that displays the average annual reflectance of the current year and the other for the preceding year from the change metric dataset (band combination SIR1-NIR-red).

Parameter file: mosaic_c_avett654_2022.txt



Current year composite

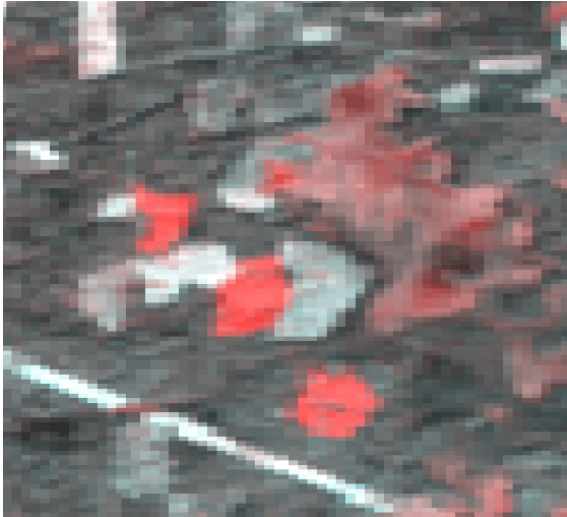
Parameter file: mosaic_p_avett654_2022.txt



Preceding year composite

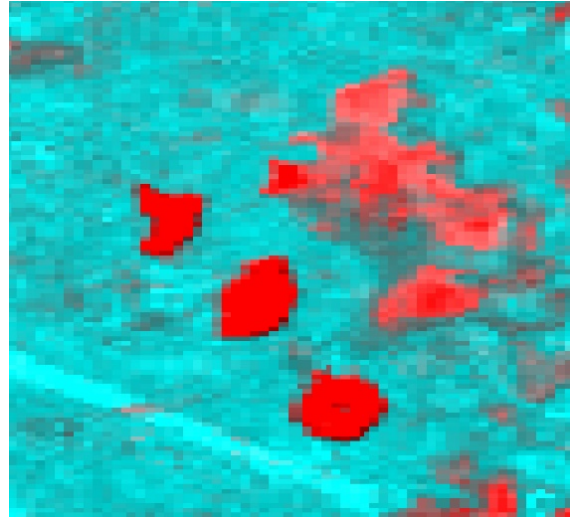
One of the most important features of the change detection metrics is the ability to highlight per-16-day composite (seasonal) changes in spectral reflectance. The following example illustrates two versions of such composites: one using a comparison of SWIR1 band reflectance between the years, and another using the seasonal reflectance change metrics.

Parameter file: mosaic_change1_2022.txt



An RGD composite shows the SWIR1 band difference between the current year (highest reflectance) and the preceding year (average reflectance). Note that not all changes in SWIR spectral reflectance represent land cover change.

Parameter file: mosaic_change2_2022.txt



A composite that displays the highest seasonal change of the SWIR1 band and the average seasonal change of the NIR/SWIR1 band ratio. Note that not all changes in spectral reflectance represent land cover change.

Another important change detection metrics visualization method is using the latest available observation stored in “p_last” and “c_last” metrics. While the data from these metrics may appear noisy (these metrics include the latest observation regardless of seasonality), these composites provide important information about the changes that occurred at the end of the corresponding year. The examples for “last” observation metrics aggregation are provided in the C:\GLAD_Tools\Examples\change_detection_mosaics.

4.4. Data Quality Metrics and Water Mask

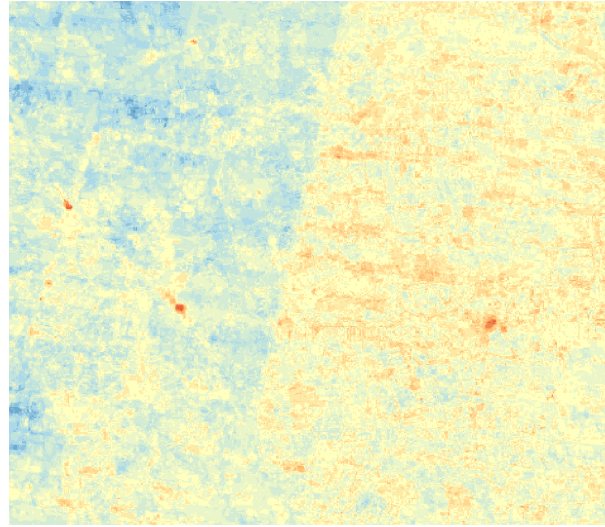
Data quality metrics provide information on data quality and clear-sky observation frequency. The water permanence metric (defined as the proportion of water detection of all available clear-sky observations) is particularly important as it allows a user to derive an annual water mask. The following examples illustrate data quality metric aggregation from the phenological metric set (C:\GLAD_Tools\Examples\phenological_mosaics).

Parameter file: mosaic_waterprc_2022.txt



Gray-scale image of water permanence. The pixel value (percent of water detection $\times 10$) indicates water duration within a year. The value 1000 indicates permanent water.

Parameter file: mosaic_imagecount_2022.txt



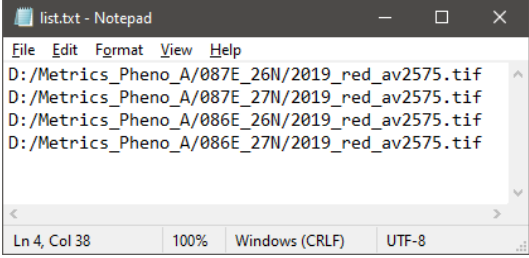
Observation number (0-23) as pseudo color raster. Note the effect of SLC-off artifacts and image path overlap. The observation number is usually higher over areas with intermittent water (where both land and water detection are used), and low for pixels with uncertain data quality (wetlands, permanently bright objects, etc.)

4.5. Metric Mosaicking using OSGeo4W

The following example illustrates the process of manually stitching the tiles into a mosaic (for advanced users that have experience working with OSGeo4W).

Example

Create a list of files to stitch and print it to a new text file. Use the full path to each file. The following example shows the list of files for red band av2575 (annual interquartile average).

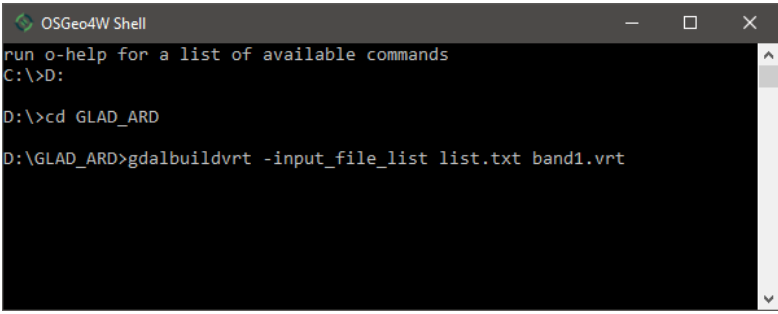


```
list.txt - Notepad
File Edit Format View Help
D:/Metrics_Pheno_A/087E_26N/2019_red_av2575.tif
D:/Metrics_Pheno_A/087E_27N/2019_red_av2575.tif
D:/Metrics_Pheno_A/086E_26N/2019_red_av2575.tif
D:/Metrics_Pheno_A/086E_27N/2019_red_av2575.tif
Ln 4, Col 38 100% Windows (CRLF) UTF-8
```

Open the OSGeo4w interface and navigate to the folder with the text file.

Execute the VRT generation command:

```
>gdalbuildvrt -input_file_list list.txt band1.vrt
```



```
OSGeo4W Shell
run o-help for a list of available commands
C:\>D:
D:\>cd GLAD_ARD
D:\GLAD_ARD>gdalbuildvrt -input_file_list list.txt band1.vrt
```

You may create a set of separate VRT files for each band, and then merge them together as multi-band image using the following command:

```
>gdalbuildvrt -separate mult-band.vrt band1.vrt band2.vrt band3.vrt
```

The VRT file can be directly loaded to QGIS. For use with other applications (i.e., ArcGIS), you will need to convert the VRT file into a raster image:

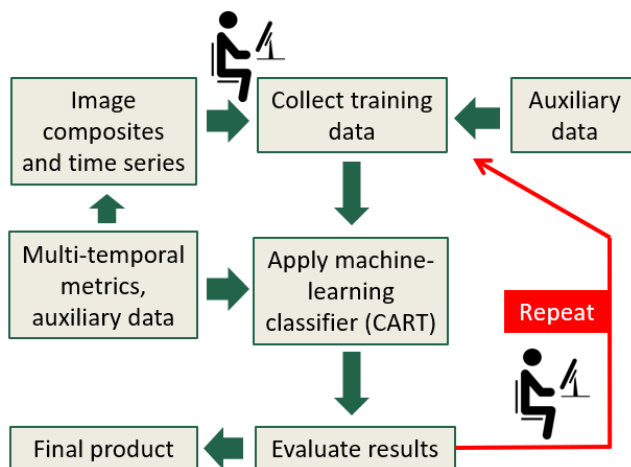
```
>gdal_translate -of "GTiff" -co "BIGTIFF=YES" -co "COMPRESS=LZW" mult-band.vrt mult-band.tif
```


5. Land Cover Classification

The GLAD Tools include a decision tree module (also called Classification and Regression Trees, or CART) for land cover/land use classification. The decision tree algorithm in GLAD Tools is based on the R's `rpart` tool, <https://cran.r-project.org/web/packages/rpart>. This algorithm estimates the probability of land cover classes using decision rules in the multispectral/multitemporal domain. The decision rules are generated automatically from a set of training pixels. The Appendix 3 section explains the decision tree algorithm and the tree model file structure.

The decision tree is a powerful tool for supervised image classification. An ensemble of tree models is built automatically using a population of training objects (image pixels manually mapped by an expert). Each pixel in the training population has a land cover class presence/absence label (defined by the image interpretation expert, called the “dependent variable”) and spectral reflectance and phenology data (multitemporal metrics, called “independent variables”). The tree model separates the multidimensional space of independent variables into compartments (hypervolumes called “nodes”) so that most training pixels within each node belong to the same land cover class. When the classification tree model is implemented for an entire image, it predicts the land cover class for every pixel.

The prediction, however, may be false due to inconsistent or insufficient training data. The model can be improved through the iterative process of adding training data, known as “active learning”. The active learning method consists of iterations of model construction, application, evaluation of results, and adding new training data until the desired map quality is achieved.



Active learning focuses on the interaction between the analyst and the classifier. After applying the classification, the results are checked by the analyst. The analyst assigns correct land cover labels to pixels classified incorrectly and adds these training pixels to the training set to improve the model. In this way, the model is optimized on well-chosen difficult examples, maximizing its generalization capabilities.

In the following section, we describe the application of the supervised decision tree classification tool that employs the dependent variable (training data) in the form of vector polygon layers and independent variables in the form of phenological metrics.



The GLAD classification tool is designed to map a single land cover/use class at a time. The training data for each classification includes only presence (target) and absence (background). The classification result represents the likelihood (probability) of the target class (in percent). For multi-class classification, refer to the “Hierarchical classification” section.

5.1. Collecting Training Data

Training data represent two polygon shapefiles, one with areas of land cover class presence (“target”), and the other with areas of land cover class absence (“background”). Both shapefiles should in the same coordinate system as phenological metrics (*+proj=longlat +ellps=WGS84 +datum=WGS84 +no_defs*). The classification tool uses only the object shape data, all attributes ignored. The shapefiles may contain overlapping polygons. The correct topology is not required as long as data can be correctly rasterized. The polygons in the “target” and “background” shapefiles may overlap. In case of overlap, the area under the “target” class polygons will be erased from the “background” layer (the “target” class has a higher priority during rasterization).

The polygon layers may be created in any GIS software. The following manual demonstrates the use of QGIS for shapefile editing. The following checklist summarizes the requirements for training data collection:

- QGIS with QMS, Freehand Editing, and Send2GE plugins (section 1.3).
- Image mosaic of selected metrics used for data visualization (section 4.2).
- Two empty shapefiles in geographic coordinates WGS84 (sample provided in C:\GLAD_Tools\Examples\classification_single_date).

To collect training data, follow the routine described below:

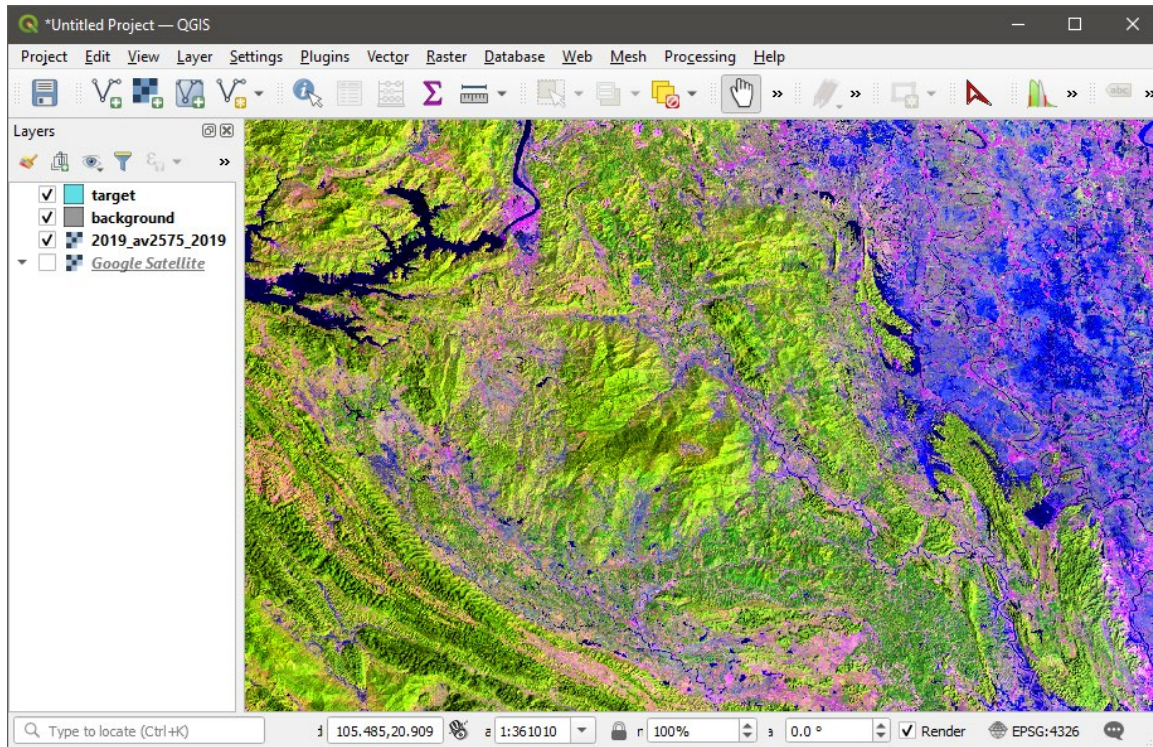
- Create the classification workspace. The workspace (folder) should include:
 - A list of tiles (single column, tile names only)
 - Two shapefiles for training data. Empty shapefiles are provided here: C:\GLAD_Tools\Examples\classification_single_date
 - classification parameter file (see below)
- Open QGIS (new project) and load the required plugins.
- Add raster layers (mosaics of selected metrics). Optionally: load the Google Satellite layer using the QMS plugin. To simplify tracing, use the Freehand plugin.
- Load [target.shp](#) and [background.shp](#) files. Put the target layer onto the top of the background layer in the Layer Panel.
- Start editing (Toggle Editing button) for both shapefiles.
- Use the “Add Polygon” or “Freehand Drawing” tools to add training samples. Avoid creating large training polygons. Distribute samples over the entire area of the image.
- Save layers and project (periodically)



Always use a new empty folder for a new classification project (new region or new thematic class).

Example

1. QGIS project with training shapefiles, image mosaic, and the Google Maps layer (through QMS).



2. Example of land cover class drawing using image mosaic

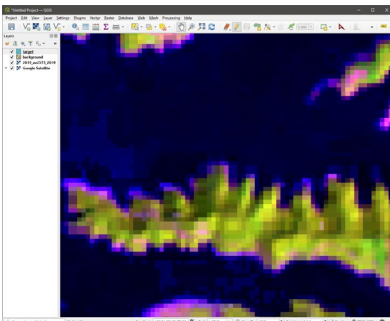
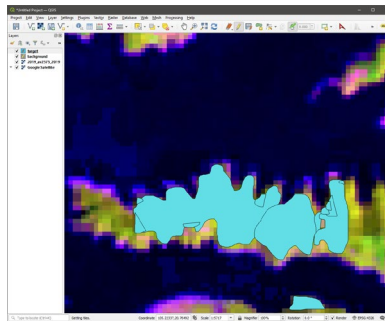
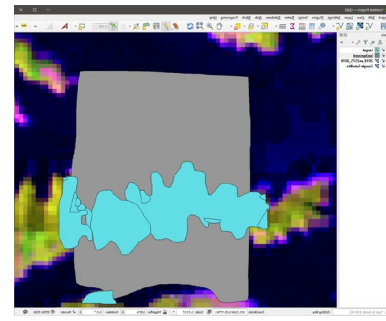


Image composite



Drawing the target class presence (forest).

The class can be mapped by any number of overlapping polygons. All pixels of the target class within the training area must be marked.



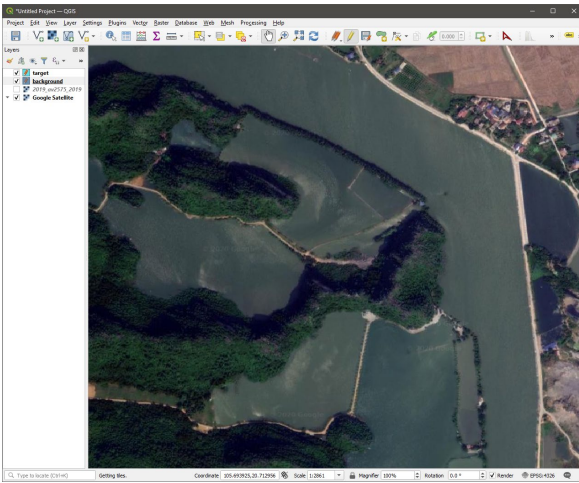
Drawing the class absence, or background (non-forest).

The polygons of the background layer can overlap with the polygons of the target class layer. The background layer polygons should not include any forest that is outside the target class map.

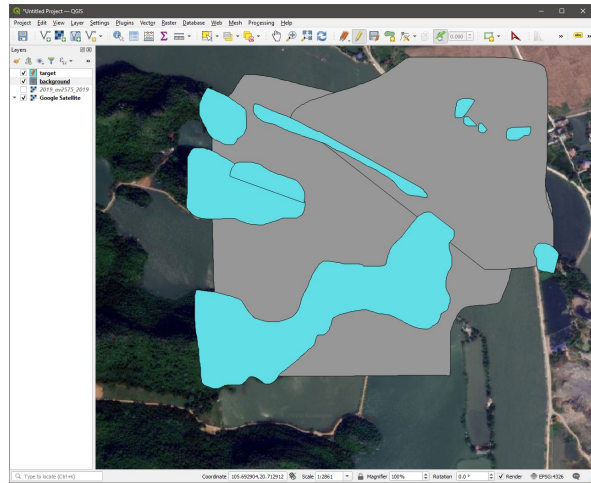


We do not recommend creating large polygons of training within homogeneous landscapes (i.e., over dense forests, water, etc.) The large number of spectrally similar training pixels will not improve classification output but will slow down the classification procedure and may cause the software to crash due to memory limitations.

3. Example of land cover class drawing using Google Maps layer.



Google Maps data

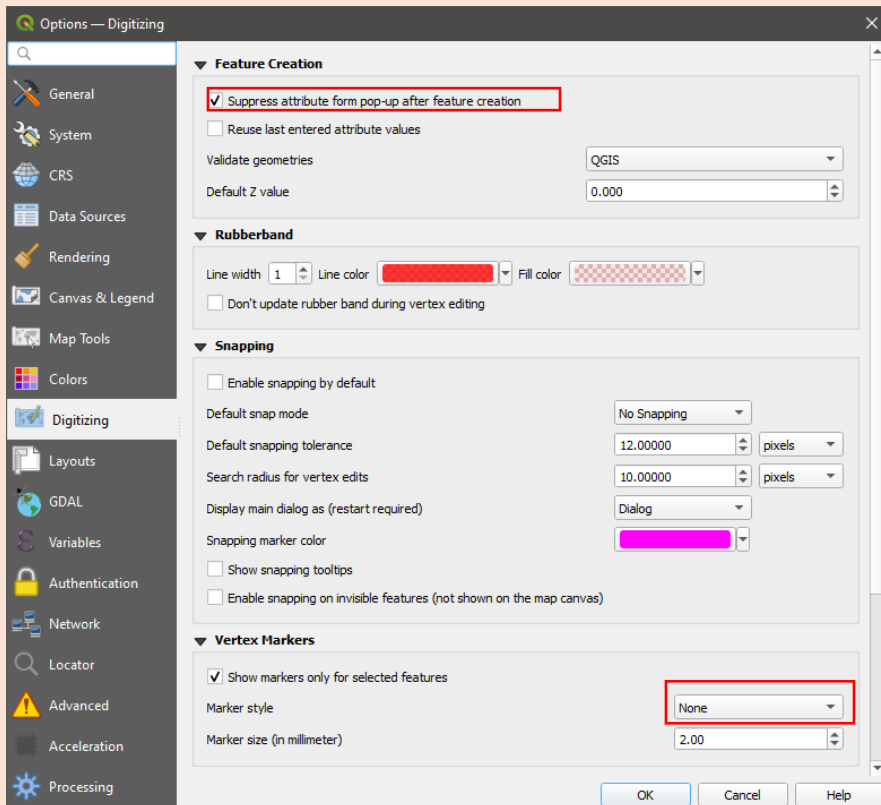


The same principle of target and background class drawing may be used to map classes at high resolution. During the rasterization, the nearest neighbor resampling will be used to select forest pixels.



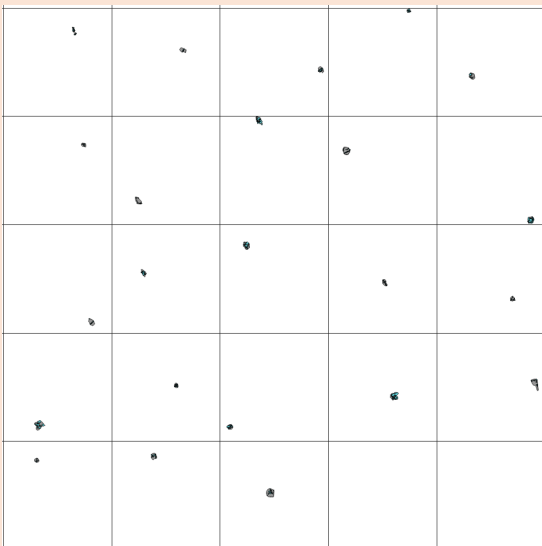
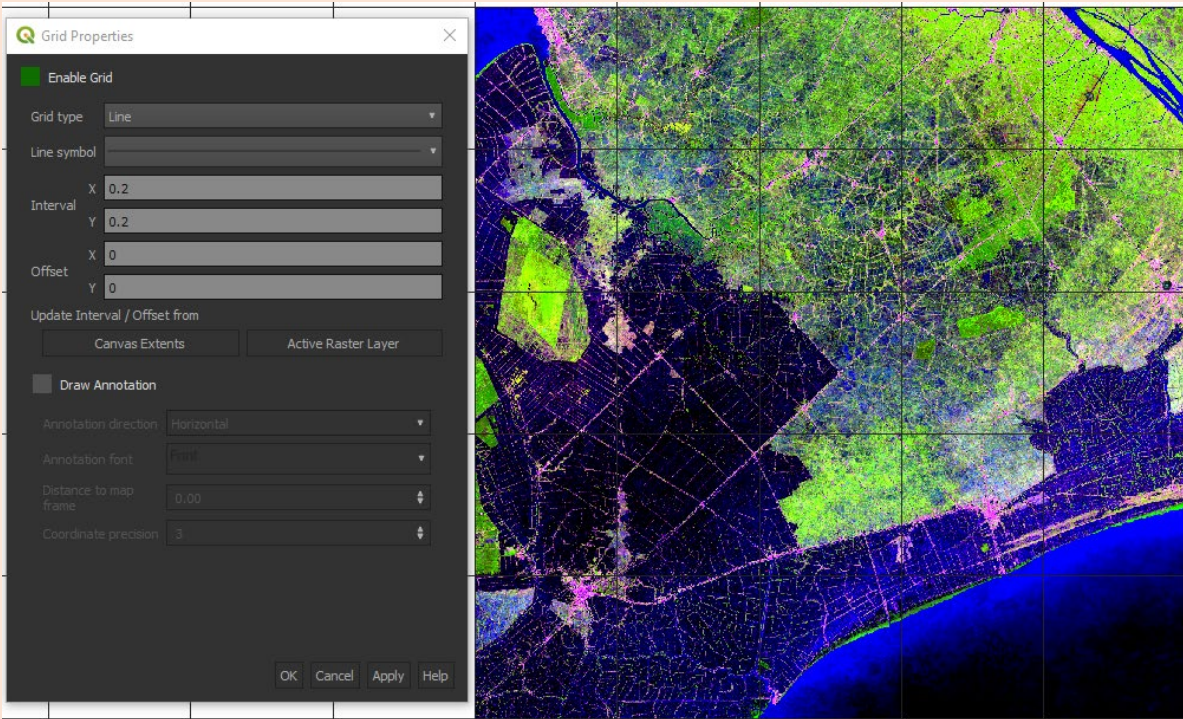
To simplify training selection and drawing, change the following setting of QGIS:

- Check the “Suppress attribute from pop-up after feature creation” box.
- Set Vertex marker style to “None.”





It is important to distribute training areas throughout the entire mapping region. We recommend using the QGIS grid tool to guide training allocation. The grid can be displayed over the metric mosaics using the QGIS menu option [View\Decorations\Grid](#). If your mapping area is smaller than 200x200 km, a grid spaced by 0.2 degrees is optimal to guide training data. In the image below, the grid overlaps a single GLAD ARD tile. The training samples should be collected within each grid cell. The placement of a sample inside a grid cell depends on the land cover class configuration.



The build-up land cover class training shapefiles. The training distribution is guided by the 0.2x0.2 degrees grid within the area of analysis.

5.2. Applying Classification

To apply classification, follow the routine described below:

- Save all edits and close the QGIS project.
- Edit the classification parameter file.

See the classification parameter file example here:

C:\GLAD_Tools\Examples\classification_single_date\classification_pheno_2022.txt

Classification parameter file structure

mettype=pheno	Metric type
metrics=D:/Metrics_pheno_2022	Multi-temporal metrics source folder
dem=D:/DEM	Topography metrics source folder
year=2022	Year of multi-temporal metrics
target_shp=target.shp	Target class shapefile name
bkgr_shp=background.shp	Background class shapefile name
tilelist=tiles.txt	Name of the tile list file
outname=settlements_2022	Output file name
mask=none	Mask file name (none – no mask)
maxtrees=25	Number of trees (odd number in the range 1-25)
sampling=10	Sampling rate (percent training data extracted for each tree)
mindev=0.0001	Tree pruning rule
threads=4	Number of parallel processes
treethreads=25	Number of parallel processes for a tree model
reuse_model=none	Use existing trees (none – create a new model)

You may modify parameters depending on the computer capacity, training size, etc. Specifically:

- Increasing the **maxtrees** parameter will slow classification but improve model generalization.
- Increasing the **mindev** will reduce tree complexity, while reducing it will increase tree complexity.
- Reduce **sampling** if sample areas are too large. Increase it if the “maxtrees” parameter is reduced.
- Reduce **threads** and **treethreads** parameters for a low-capacity computer (minimal value 1)

- Open cmd, navigate to the folder with the tile list, and run the program:

```
>perl C:/GLAD_Tools/classification.pl <parameter file>
<parameter file> - the name of the classification parameter file
```

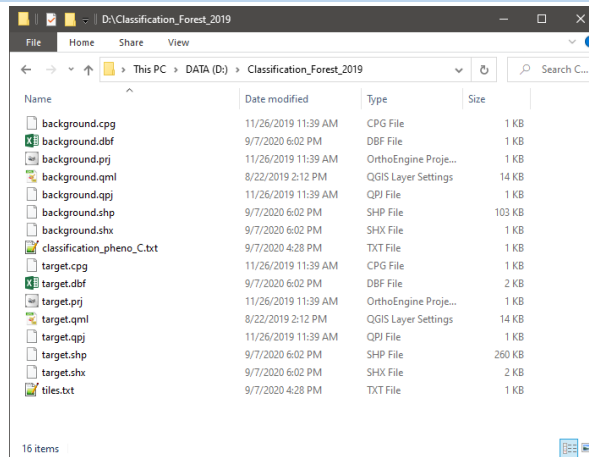
- Wait for the process to complete.
- Open QGIS and load the classification result (TIF file). To visualize the target class, use transparency threshold 0-49. To show only the background class, apply transparency to the interval 50-100.



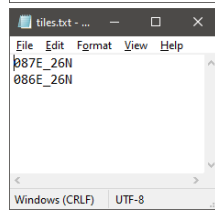
It is important to **save and close** the QGIS project before applying classification.

Example

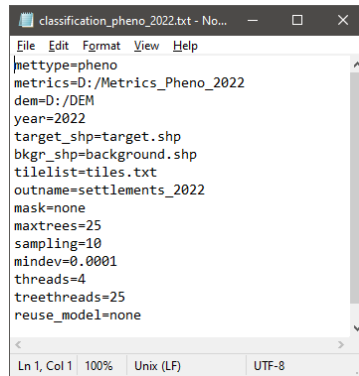
The work folder setup for image classification. The folder contains a copy of empty shapefiles (target and background), a list of tiles, and the classification parameter file



Tile list file
tiles.txt

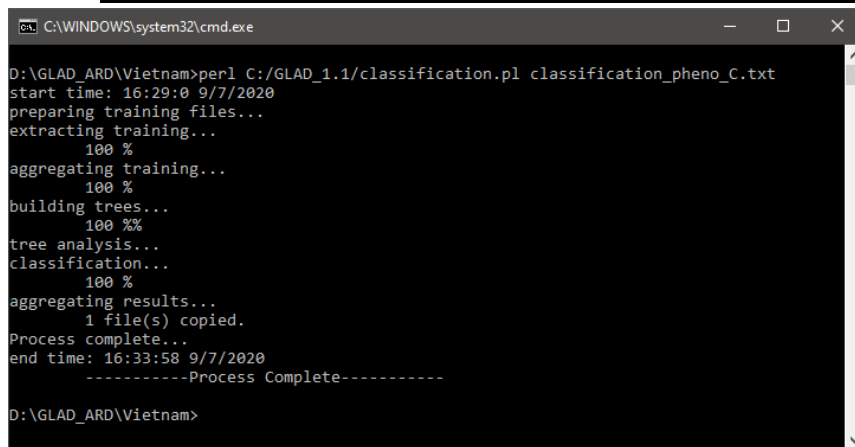


Parameter file
classification_pheno_C.txt



CMD command

```
>perl C:/GLAD_Tools/classification.pl classification_pheno_2022.txt
```

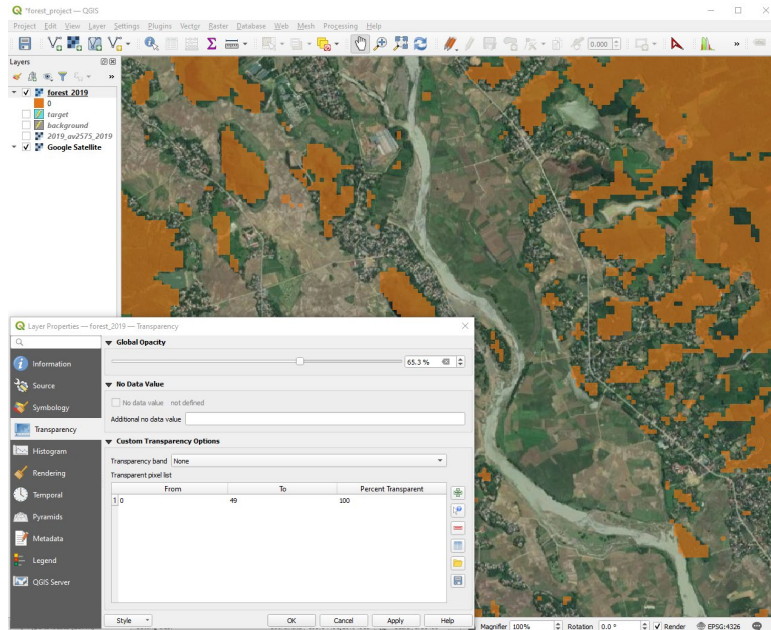


The classification output is stored as a raster file (LZW-compressed GeoTIFF). All tiles are mosaicked together. The pixel value is in the range of 0 – 100 and represents the target class likelihood. The commonly used threshold to identify the target class is 50% (values 0-49 represent the background class, and 50-100 - the target class). However, the threshold may be adjusted if needed. The likelihood should not be treated as “probability” or “similarity”, as it depends on the training population. The class likelihood does not represent the percentage of a target class within a pixel.

The output file name is specified by the parameter file (forest_2019.tif). The file is created (or updated) after running classification. The file can be loaded to QGIS or any other GIS software.

To convert the result into a categorical map, we use the threshold of 50%. In the QGIS, the layer can be turned into the target class map (forest map) by applying transparency in the interval 0-49.

Alternatively, applying transparency for the range of values 50-100, we may create a map of the background class (non-forest)



5.3. Iterating Classification

Due to high complexity of land cover, the accuracy of classification that is based on a small subjectively selected training population is (usually) low. To improve the classification accuracy, we implement an **active learning** method. After obtaining the initial classification output, we evaluate it and add new training sites in areas where commission or omission errors are evident. To perform active learning iterative training, follow this routine:

- Open the QGIS project and load classification results.
- Start editing for training shapefiles.
- Visually check the map (using both target and background class masks) and add training to shapefiles.
- Save shapefiles and the project and close QGIS.
- Perform classification. Classification results will be updated.



When using both image Landsat composite and Google Maps data to collect training, check for areas that may change recently. For example, a forest patch that is seen on Google Maps may be recently cleared, and Landsat data will show you that. Only Landsat image composite correctly represents land cover for a selected year. Always use both image sources to make the right decision before drawing training maps.

5.4. Hierarchical Classification

The classification tool allows mapping a single land cover type. To map multiple land cover types, the user must design a hierarchical classification system, where each class is mapped independently within the area outside already mapped classes. For example, a complete land cover classification to IPCC 6 classes may include the following steps:

- Mapping “settlements” class within the entire area of analysis.
- Mapping “water and wetlands” class within the mask outside settlements.
- Mapping “forests” class within the remaining area.
- Mapping “croplands” class within the remaining area.
- Mapping “grasslands” within the remaining area.
- The remaining area is assigned to the “other lands” class.

Another example is a thematic disaggregation of a general class, i.e., mapping forest types within the mask of the “forest” class.

The hierarchical classification method required a classification mask which represents an area that is not assigned to any land cover class. Another application of the classification mask is the reduction of the area of analysis, e.g., classification for land cover types within administrative boundaries.

The classification mask is a raster file that includes the entire area of analysis. The mask file should have only two classes (pixel values): values 1 (where classification will be applied) and 0 (where classification will output zero likelihood value). If the mask is used to restrict classification to a certain area, a vector file that contains an area polygon should be rasterized to the classification extent to serve as a classification mask (see section 9.6). For hierarchical classification, the mask should be created for every land cover class using the Image Modeler tool as described below. The Image Modeler instructions are provided in detail in section 9.2.

5.4.1. Making Classification Mask (Single Class)

To create a mask to map a new land cover class outside a class that was mapped earlier, use the following steps:

- Create a new classification workspace. Copy here empty training shapefiles and list of tiles.
- Create a text file with the model: mask_model.txt.
- The model has the following content (put the correct path and file name for the earlier classification result that will be used as a mask):

```
INPUT
%1=<path to earlier classification results>\<classification output name>.tif
END
MODEL
if (%1<50) {%0=1;}
END
```

- Open CMD and run the command:

```
>perl C:/GLAD_Tools/raster_model.pl mask_model.txt mask.tif
```
- The resulting file mask.tif may be used to collect training data and run classification.

5.4.2. Making Classification Mask (Multiple Classes)

In case several classes should be used together to make a single mask for hierarchical classification, use the same steps as in section 5.4.1, but with the following mask_model.txt file structure:

```
INPUT
%1=<path to earlier classification results1>\<classification output name1>.tif
%2=<path to earlier classification results2>\<classification output name2>.tif
%3=<path to earlier classification results3>\<classification output name3>.tif
END
MODEL
if (%1<50 and %2<50 and %3<50) {%0=1;}
END
```

The number of input classes is not limited.

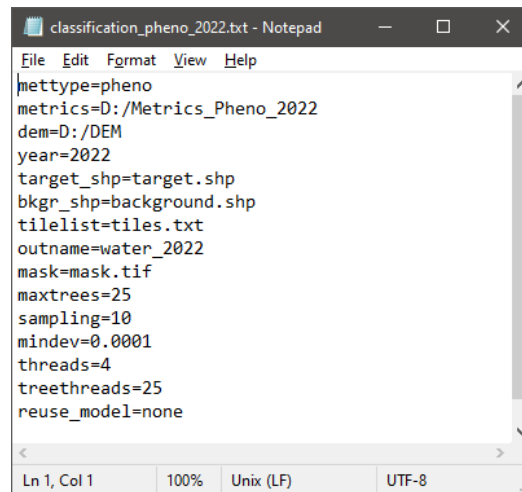
5.4.3. Applying Classification with Mask

To use the mask for classification, change the classification parameter file: replace “mask=none” with “mask=mask.tif”. See an example of the parameter file below. The training data collection and classification will be restricted to the areas with the value “1” in the mask.tif file.

Example

Changes to the parameter file from section 5.2:

```
mask=mask.tif
```



```
classification_pheno_2022.txt - Notepad
File Edit Format View Help
mettype=pheno
metrics=D:/Metrics_Pheno_2022
dem=D:/DEM
year=2022
target_shp=target.shp
bkgr_shp=background.shp
tilelist=tiles.txt
outname=water_2022
mask=mask.tif
maxtrees=25
sampling=10
mindev=0.0001
threads=4
treethreads=25
reuse_model=none
Ln 1, Col 1 100% Unix (LF) UTF-8
```

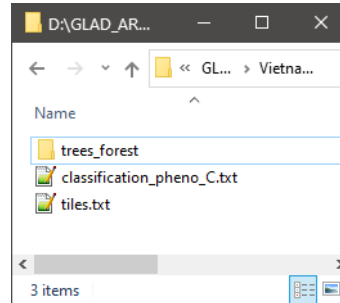
5.5. Applying an Existing Classification Model

One of the primary benefits of decision tree land cover classification models is the ability to re-use an existing model for a different region or different time interval (year). To reuse the model, first, create a new classification project folder. Copy the classification parameter file and the “trees” folder from the original classification folder. The tree folder can be renamed if needed. The list of tiles may be reused from the original classification or changed. The important changes to the parameter file are shown in the example below.

Example

1. The folder content for classification re-use:

- original tree model (may be renamed)
- list of tiles (original or new)
- parameter file (see below)



2. Important changes to the parameter file:

Check the year to apply the model:

```
year=2022
```

Check that the list of tiles is correct:

```
tilelist=tiles.txt
```

Put here the name of the tree model:

```
reuse_model=trees_forest
```

```
File Edit Format View Help
mettype=pheno
metrics=D:/Metrics_Pheno_2022
dem=D:/DEM
year=2022
target_shp=target.shp
bkgr_shp=background.shp
tilelist=tiles.txt
outname=forest_2022
mask=none|
maxtrees=25
sampling=10
mindev=0.0001
threads=4
treethreads=25
reuse_model=trees_forest
Ln 9, Col 10 100% Unix (LF) UTF-8
```

When the **reuse_model** parameter contains the name of the existing model, the classification code will skip model calibration and will implement the existing trees. The errors may indicate that the model’s name is incorrect, the metrics are missing, or the trees were copied incorrectly.

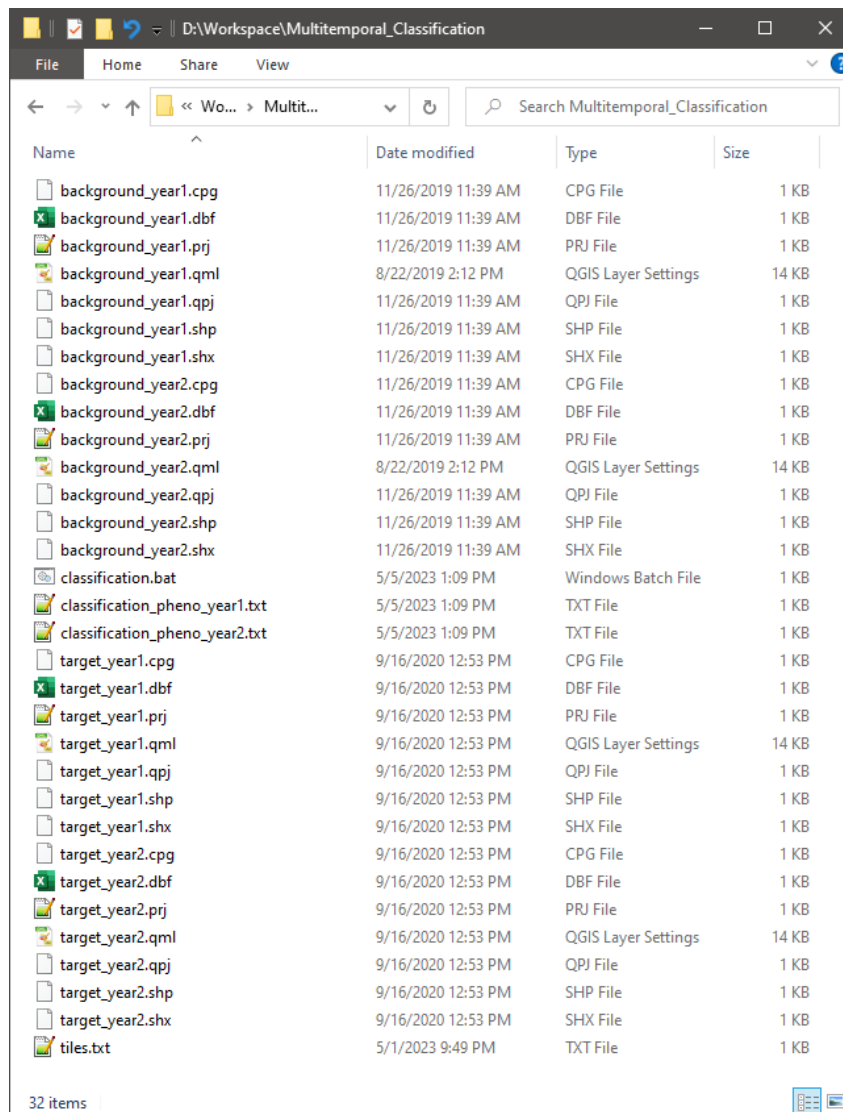
6. Multitemporal classification

The multitemporal classification tool is designed for calibrating temporally stable land cover and land use mapping models. The multitemporal classification model uses training data from different years to create a single classification model. This way, the classification model is trained to ignore differences between Landsat sensors and inconsistencies in annual data frequency. The number of intervals for model calibration and application is not limited; two to four intervals are recommended. The following example shows the model calibration using two annual intervals.

6.1. Multitemporal Classification Workspace

The metrics and image mosaics should be completed for both years (i.e., years 2008 and 2018 for the following example). The classification folder should contain:

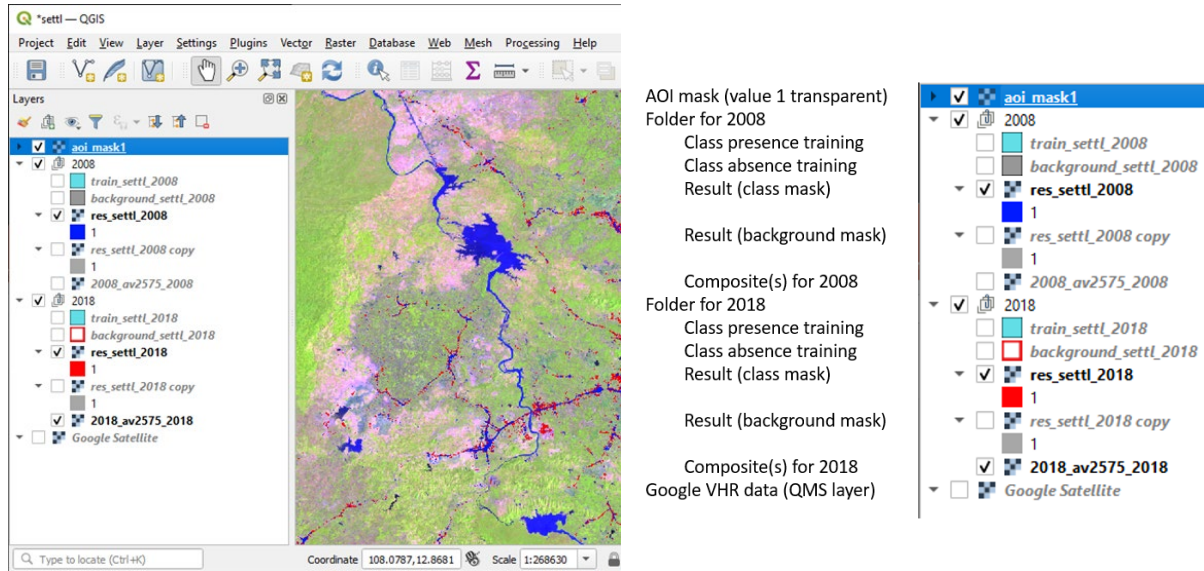
- A separate set of training shapefiles (target and background classes) for each year.
- Classification parameter files for each year.
- The list of tiles.
- Batch command file to run classification procedure.



6.2. QGIS Project Organization

We suggest using groups to organize the layers in the project's TOC, i.e., use a separate group for each year. The target training layer should be placed on top of the background layer. For more information about the training data collection see Section 5.1.

To collect training data, visualize data for one year at a time to add training sites. Visualizing classification results from both years allows for detecting change and highlights classification errors.



AOI mask (value 1 transparent)
 Folder for 2008
 Class presence training
 Class absence training
 Result (class mask)

 Result (background mask)

 Composite(s) for 2008
 Folder for 2018
 Class presence training
 Class absence training
 Result (class mask)

 Result (background mask)

 Composite(s) for 2018
 Google VHR data (QMS layer)

6.3. Applying Multitemporal Classification

Unlike single-year classification, multitemporal classification uses separate tools for sample extraction, model calibration, and model application. For each year, sample extraction and model application are implemented separately and are guided through parameter files. The parameter file structure is the same as described in section 5.2. For each year, the parameter files have different source metric sets, different training files, and different output file names (highlighted in red and blue in the example below).

```

classification_pheno_year1.txt
mettype=pheno
metrics=D:/Metrics_Pheno_2012
dem=D:/DEM
year=2012
target_shp=target_year1.shp
bkgr_shp=background_year1.shp
tilelist=tiles.txt
outname=settlements_2012
mask=none
maxtrees=25
sampling=10
mindev=0.0001
threads=4
treethreads=25
reuse_model=none
    
```

```

classification_pheno_year2.txt
mettype=pheno
metrics=D:/Metrics_Pheno_2022
dem=D:/DEM
year=2022
target_shp=target_year2.shp
bkgr_shp=background_year2.shp
tilelist=tiles.txt
outname=settlements_2022
mask=none
maxtrees=25
sampling=10
mindev=0.0001
threads=4
treethreads=25
reuse_model=none
    
```

The batch file simplifies the classification application. Here is an example of the batch file (for the two years):

```
classification.bat
perl C:/GLAD_Tools/multitemporal_class_export.pl classification_forest_year1.txt
perl C:/GLAD_Tools/multitemporal_class_export.pl classification_forest_year2.txt
perl C:/GLAD_Tools/multitemporal_class_model.pl 25 25 0.0001
perl C:/GLAD_Tools/multitemporal_class_apply.pl classification_forest_year1.txt
perl C:/GLAD_Tools/multitemporal_class_apply.pl classification_forest_year2.txt
```

The following list of commands is used for the classification procedure:

multitemporal_class_export.pl – export training data for the tree model. Each year is exported separately, all years are aggregated together in a single training file.

multitemporal_class_model.pl – build trees using aggregated training files. These trees use training from all input years. Parameters: <number of trees> <number of threads> <mindev>.

multitemporal_class_apply.pl – applies the classification model to each year. The model is the same for all years.

Example workflow to create a training data set and perform classification:

- Open the QGIS project and turn “on” the year 2018 folder.
- Draw training using Landsat composite and Google VHR data.
- Save both training files. Close project.
- Open CMD in the project folder (i.e., “D:\Settlements”) and call classification.bat.

```
>classification.bat
```

- Classification may take a long time. Ignore “process complete” messages; wait for the CMD prompt.
- After classification is complete, open QGIS and check the results.
- Add training for 2018 if needed. Save and stop editing before changing the year.
- Change the year 2018 folder “off” and the year 2008 “on”. Check the 2008 result. Add training to 2008 training files.
- Save and close QGIS before iterating the classification.
- If needed, iterate classification and check the results.

7. Mapping Continuous Variables

GLAD Tools supports continuous variable modeling, e.g., tree canopy cover, vegetation height, biomass, etc., using bagged regression tree ensembles. The regression tree model is similar to the decision tree used for classification and is implemented using the same code. The main difference between classification and regression tree application is training data: we use two discrete classes for the classification tree model and a continuous variable with the range from 0 to 100 for the regression tree model.

7.1. Training Data

The tool accepts only 8-bit unsigned (8u) raster data as training for the regression tree model. The variable should be in the range 0-100. The value 255 is used as a background to indicate the absence of training data. The training data area (number of pixels with values between 0 and 100) should not exceed the capacity of the tool. We estimated the maximum number of pixels in the training data between 5 and 10 million for a computer with at least 256GB RAM. For smaller RAM, we suggest using a smaller maximum number of training pixels.

The training file can be created by re-projecting and resampling the existing raster data or by rasterizing the vector data. The training file should be in the same projection (EPSG:4326) and preferably the same pixel grid as the ARD data. The data will be resampled (but not reprojected) by the sample extraction tool. GLAD Tools raster tools (see Section 9) may help to create the raster training data. The training file may be in any raster format compatible with GDAL (GeoTIFF, ERDAS IMG) or in a VRT format.

7.2. Classification Workflow

The following data should be prepared before running the regression tree model:

1. Phenological metrics (section 3.1).
2. Training file (8u EPSG:4326 raster file with values 0-100, and 255 indicating no training data).
3. List of tiles in text format.
4. Parameter file.

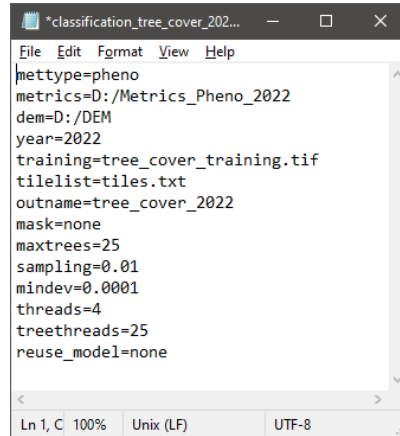
The classification parameter file is similar to the one used for the land cover classification. Instead of training shapefiles, the training file name should be listed as a “training” parameter.

Regression model parameter file structure

mettype=pheno	Metric type
metrics=D:/Metrics_pheno_2022	Multi-temporal metrics source folder
dem=D:/DEM	Topography metrics source folder
year=2022	Year (for multi-temporal metrics)
training=tree_cover_training.tif	Training raster file (or VRT)
tilelist=tiles.txt	Name of the tile list file
outname=tree_cover	Output file name
mask=none	Mask file name (none – no mask)
maxtrees=25	Number of trees (odd number in the range 1-25)
sampling=0.001	Sampling rate (percent training data extracted for each tree)
mindev=0.0001	Tree pruning rule
threads=1	Number of parallel processes
treethreads=25	Number of parallel processes for a tree model
reuse_model	Use existing trees (none – create a new model)

You may modify the parameter file depending on the computer capacity, training size, etc. Specifically:

- Increasing the **maxtrees** parameter will slow classification but improve model generalization.
- Increasing the **mindev** will reduce tree complexity, while reducing will increase tree complexity.
- Reduce the **threads** and the **treethreads** parameters for a low-capacity computer (minimal value 1)
- The “**sampling**” parameter should be set in order to extract sufficient, but not too high value of sample pixels from the training file. We recommend the maximum 50,000-100,000 total sample pixels for a tree model. Use the total number of training pixels (which can be calculated using image area tool) to estimate the optimal sampling rate.



```

*classification_tree_cover_202...
File Edit Format View Help
mettype=pheno
metrics=D:/Metrics_Pheno_2022
dem=D:/DEM
year=2022
training=tree_cover_training.tif
tilelist=tiles.txt
outname=tree_cover_2022
mask=none
maxtrees=25
sampling=0.01
mindev=0.0001
threads=4
treethreads=25
reuse_model=none
Ln 1, C 100% Unix (LF) UTF-8

```

The following workflow illustrates the regression model application:

- Open cmd, navigate to the folder with the tile list, and run the program:


```
>perl C:/GLAD_Tools/regression.pl <parameter file>
```

<parameter file> - the name of the regression model parameter file
- Wait for the process to complete.
- Open QGIS and load the classification result (TIF file). The values in the output raster will represent the continuous variable, 0-100.

8. Change Detection

Change detection metrics and classification algorithms are designed specifically to map the abrupt loss of tree and shrub canopy cover, such as logging, windfalls, wildfires, and others. For all other land cover changes, such as cropland extent change or urban expansion, users should implement multitemporal land cover mapping (see Section 6). The change detection metrics simplify the interpretation and mapping of the abrupt forest loss events that are indicated by significant differences in land surface reflectance between the same seasons of the corresponding (“current”) year and the preceding year(s). Similarly to Land Cover Classification, we apply a decision tree model that assigns the per-pixel likelihood (probability) to represent land cover change based on the application of statistical decision rules in the multispectral/multi-temporal domain. The decision tree model is calibrated using a training population of pixels with assigned change (target) and no-change (background) classes.

8.1. Collecting Training Data

Similarly to land cover classification (section 5.1), training data represent two polygon shapefiles, one with areas marking change class pixels (“target”), and the other marking other pixels (“background”). Both shapefiles should be in the same coordinate system as metrics (+proj=longlat +ellps=WGS84 +datum=WGS84 +no_defs). The classification tool uses only the object shape data, all attributes ignored. The shapefiles may contain overlapping polygons. The correct topology is not required as long as data can be correctly rasterized. The polygons in the “target” and “background” shapefiles may overlap. In case of overlap, the area under the “target” class polygons will be erased from the “background” layer.

The training for change detection should include areas that are marked as “change” and included in the “target” layer and “no change” in the background layer. The “change” is defined by the data analyst and may represent a specific spectrally detectable land cover transition (forest to non-forest). The annual change detection metrics are not designed to map slow processes like tree regrowth.

Thematically similar land cover changes may be represented by different trajectories of surface reflectance change. For example, a forest disturbance that causes the removal of woody vegetation may be indicated by contrasting changes in shortwave infrared reflectance (SWIR): SWIR will increase after logging and decrease after a forest fire or flooding. While the metrics are designed to allow detection of different indications of land cover changes, the change detection model requires a set of training data that represents different spectral responses. To examine land cover dynamics and to create a comprehensive training dataset, data analysts are encouraged to visualize a combination of different metrics.

Example of the training site for forest loss detection

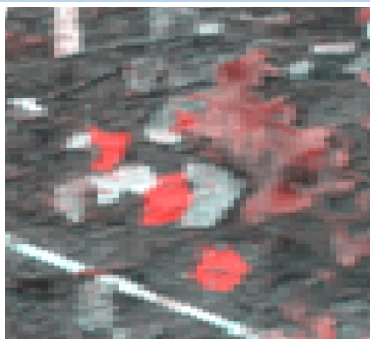
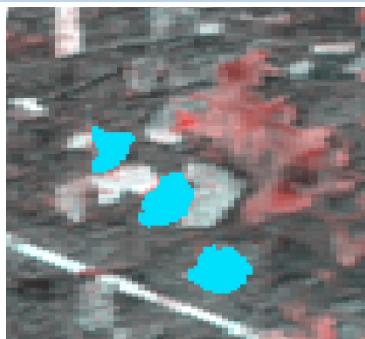
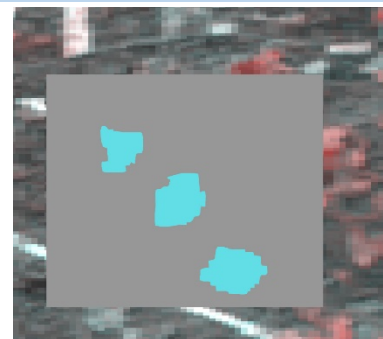


Image composite
(see Section 4.3)



Target training



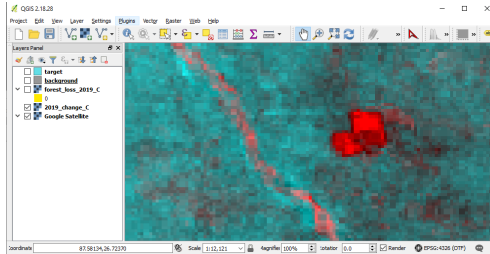
Background training (overlaid
with target training)

8.2. Applying Classification

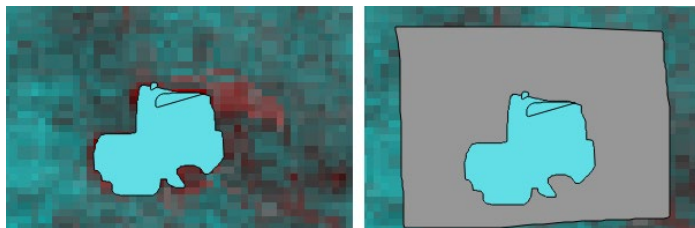
The model application for change detection is the same as for the land cover classification. See section 5.2 for parameter file structure and commands. Usually, the classification process must be iterated several times to obtain a sufficiently accurate map (see section 5.3).

Example of forest change classification

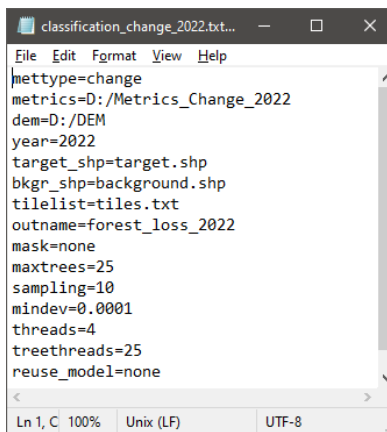
1. Image composites (see section 4.3) used as source data for change detection and drawing training sites in QGIS



2. Drawing change and no change training sites.



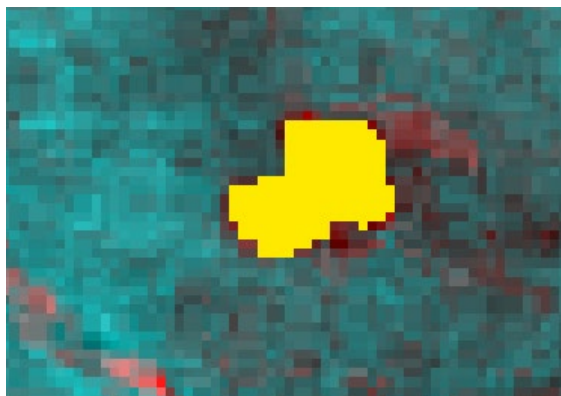
3. Classification parameter file.



4. CMD command

```
>perl C:/GLAD_Tools/classification.pl classification_change_2022.txt
```

5. Open QGIS and load the classification result (TIF file). To visualize the target class, use transparency threshold 0-49. To show only the background class, apply transparency to the interval 50-100.



8.3 Multi-year Change Detection Workflow

The ARD-based approach of change detection provided by GLAD Tools is designed for mapping land cover changes that occurred between the target year (end of the interval, e.g., 2020) and preceding years (three preceding years are considered, e.g., 2017-2019). The model that has been calibrated for a single year may be applied to other years using the approach provided in section 5.5. to create a time series of change detection results. The results can be integrated using the Image Modeler (section 9.2) to retain the date of the earliest detected change as shown in the example below.

Example

A change detection model was applied for three years, 2018-2020, and results are stored in separate files (change_2018.tif, etc.). The following model will output a single tif file with codes indicating the date of detected change (1 – 2018, 2 – 2019, 3 – 2020).

```
INPUT
%1=change_2018.tif
%2=change_2019.tif
%3=change_2020.tif
END
MODEL
If (%1>=50) {%0=1}
else if (%2>=50) {%0=2;}
else if (%3>=50) {%0=3;}
END
```

The users should be aware that while using three preceding years of data to create a change detection metrics set improves the classification quality, the metric set is sensitive to changes that happened not only between the corresponding and preceding years (e.g., 2019 and 2020) but also between the corresponding year and the years i-2 and i-3 (e.g., 2017 and 2018). There are several approaches to obtaining a map of change only between the latest preceding and the target years.

- (1) If a model is implemented for a series of years, and the results are aggregated to retain the date of the first detected change (as in the example above), the results from the third target year and later will exclude changes that happened in earlier years.
- (2) The “p_last” metric from the change metric set represents the latest observation from the preceding year, and the “c_last” represents the latest observation from the target year. The change detection composite that uses “p_last” and “c_last” (see section 4.3) can help select training to suppress the detection of the changes that were already happened by the end of the preceding year.
- (3) The change detection metrics can be created using the data for only two years (target and preceding). To do that, the user will need to store the ARD for only these two years. The change metric code will only use available composites to create metrics. This approach requires a large number of cloud-free images and it will only be suitable for regions with long cloud-free growing season and/or after the year 2014.

9. Raster Tools

9.1. Image Recode

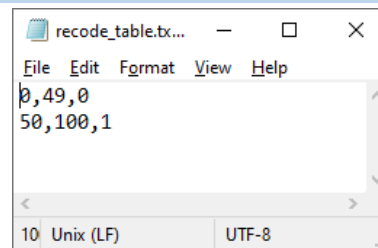
The image recode tool allows the user to create a new raster by changing the values of the original raster image to specified output values. The recode tool may be used to change the values of the output classification results (0-100, target class likelihood) to class codes (i.e., 0 and 1), or to change the values of a land cover map (i.e., to aggregate sub-classes to classes). The input raster must be in geographic coordinates on the WGS84 datum. Only 8-bit unsigned data in GeoTIFF format is supported.

The following example shows how to recode a classification output into a class mask for hierarchical classification (see section 5.4).

Example

Open a folder that contains the classification result file (e.g., "forest.tif").

Create a text file that contains a recode table (recode_table.txt). Each line of the recode table file has three elements: the minimal value in the input range, the maximal value in the input range, and the output value. The example file contains the table to recode the likelihood file into a mask with values 0 (background class) and 1 (target class).



Open CMD. Navigate to the new classification workspace. Perform the following command:

```
C:/GLAD_Tools/recode_8bit.exe <input>.tif <output>.tif recode_table.txt
```

where <input> and <output> names should be replaced with actual file names.

The output file (e.g., "mask.tif") will contain the values 1 (indicating class presence) and 0 (class absence).

9.2. Image Modeler

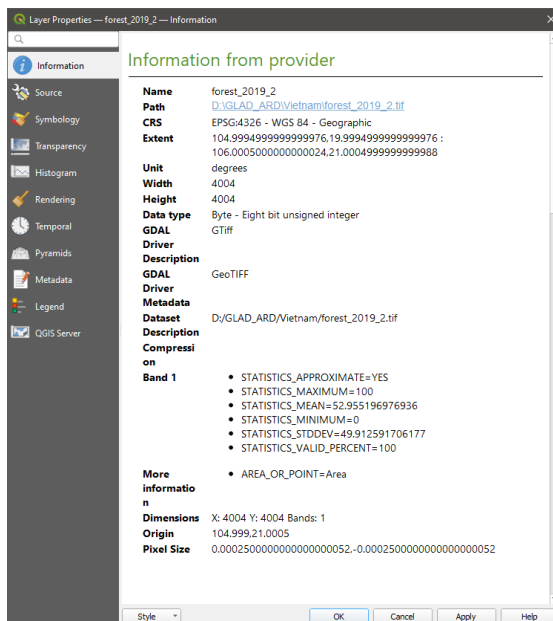
The Image Modeler tool allows the application of simple mathematical functions and conditional statements for a set of raster layers. The same operations (with more functionality) can be found in most raster GIS systems. This tool provides a limited set of tools to work specifically with the ARD products and classification results.

All input images must be in geographic coordinates on the WGS84 datum. All input images must have the same extent (in pixels), same pixel size, and same UL coordinates. The 8-bit unsigned, 16-bit unsigned, and 16-bit signed data in GeoTIFF format are supported for input images. The output image is an 8-bit unsigned raster (uncompressed GeoTIFF).

Source files must be stored in the same folder as the model text file, or a complete path to the file must be provided (e.g., "D:/folder/file.tif"). Before running the tool, confirm that all input rasters have the same coordinate system and extent. For that, you may use the QGIS image information tool, see an example below.

Example: checking image extent and georeferenced using QGIS.3xx

1. Open QGIS and load the raster file.
2. Right-click on the image in TOC and select "properties". Switch to the "Information" page



3. Check the image dimensions, origin, and pixel size. You can copy the information as text and save it for reference in Notepad++.
4. Repeat these steps for all images to make sure that they have the same projection and extent.

After all source images are checked, the user will need to make a model file that includes information on input data and processing algorithms. The parameter file includes two sections: input data definition and model definition. The sections are clearly marked (INPUT/END and MODEL/END). These marks are important for the code to interpret the parameters correctly. Each input file is associated with a variable denoted as %n, where n is the number of the input file. The output file is represented by variable %0. The Image Modeler supports only one output layer.

The format of the model section is similar to the C++ syntax for arithmetic and conditional statements. The following examples illustrate the most common operations:

```
Arithmetic operations:    %0= %1+%2;
                          %0= (%1-%2)/(%1+%2)*100;

Conditional statements:  if (%1==1) {%0=10;}
                          if (%1==1 or %2==1) {%0=1;}
                          else if (%3==2) {%0=2;}
                          else {%0=3;}
```

The following example shows all components of a simple parameter file for Image Modeler:

<pre>INPUT %1=forclass.tif %2=districts.tif END MODEL %0=%1+%2; END</pre>	<pre>The header of the input section (do not remove) First input file associated with variable %1 Second input file End of the input section (do not remove) The header of the model section (do not remove) Arithmetic model End of the model section (do not remove)</pre>
---	--

To run the Image Modeler, open the CMD in the folder that contains the parameter file and execute the following command:

```
perl C:/GLAD_Tools/raster_model.pl <parameter file> <output file>
<parameter file> - the name of the parameter file.
<output file> - the name of the output file (with extension .tif)
```

Example: creating land cover maps from ARD technical data and classification results

1. Making the water layer

The water permanence (percent water detection of annual clear-sky data) is stored in the metric TEC_prcwater for each tile. To mosaic tiles, we used the parameter file mosaic_water_pheno.txt.

```
source=D:/Metrics_Pheno_2019
list=Delta_tiles.txt
year=2019
outname=water
bands=TEC_prcwater
```

The command for the process:

```
perl C:/GLAD_Tools/mosaic_tiles.pl mosaic_water_pheno.txt
```

The output file (16-bit unsigned) 2019_water.tif shows the percentage of water detection (scaled by 10).

02. Make the analysis area mask

The analysis mask is created using a shapefile of the administrative area (or other polygonal data that delineates the analyzed region). We follow instructions from section 9.6. to rasterize the data mask. The output file (datamask.tif) has values 1 (within AOI) and 0

03. Source classification results

Source classification results (outputs of the land cover classification model) contain the target class likelihood of each class: tree cover, crop, and urban.

04. Creating the land cover map

The model lc_model.txt contains a list of the input files and rules to create the LC map classes.

```
INPUT
%1=D:/datamask.tif
%2=D:/urban.tif
%3=D:/2019_water.tif
%4=D:/crop.tif
%5=D:/trees.tif
END

MODEL
if (%1==1){
    if (%2>=50){%0=1;}
    else if (%3>=900){%0=2;}
    else if (%5>=50){%0=3;}
    else if (%4>=50){%0=4;}
    else if (%3>=100){%0=5;}
    else {%0=6;}
}
END
```

To run the model, we used the following command:

```
perl C:/GLAD_Tools/raster_model.pl lc_model.txt lc_2019.tif
```

The map can be checked using the QGIS project file in this folder

9.3. Focal Average

The focal average tool calculates the raster average in a moving window. Two types of moving windows can be used: circular (defined by the radius from the center point of a pixel) and square. Only 8-bit raster data is supported.

To run the tool (circular window), use the following CMD command:

```
C:\GLAD_Tools\ focal_average_circle.exe <input>.tif <output>.tif <radius in pixels>
<input> and <output> are names of in/out tif files (output file will be created), and radius
specifies moving window size.
```

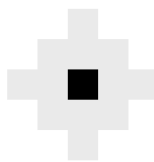
A similar command is used for square window:

```
C:\GLAD_Tools\ focal_average_square.exe <input>.tif <output>.tif <radius in pixels>
<input> and <output> are names of in/out tif files (output file will be created), and radius
specifies moving window size.
```

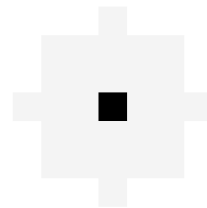
Here are examples of different window shape/size applications to a raster with values 100 and 0 using different parameters:



Circular window, radius = 1



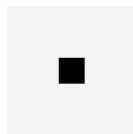
Circular window, radius = 2



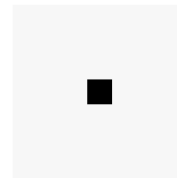
Circular window, radius = 3



Square window, radius = 1



Square window, radius = 2



Square window, radius = 3

Recommendations:

(a) Convert your classification result into a raster file with values 100 (target class) and 0. This way, the focal average result will represent the percent of the target class within a moving window.

(b) Use the square focal average tool for radius =1 and the circle focal average tool for a larger radius of the window to avoid visual artifacts.

(c) The tool can be used to create buffers following these steps:

1. Convert your raster to values 100 and 0 using Image Modeler
2. Perform focal average with a radius equal to the desired buffer distance (in pixels)
3. Using Image Modeler, assign a class label to all pixels with the value above 0. This will effectively expand your class area to the desired buffer distance.

9.4. Raster Area

The raster area tools are using an equation for the area of a spherical trapezoid on the WGS84 ellipsoid to correctly calculate the area of each pixel of the raster data in ARD geographic format. The tool is required to estimate the area of classes from the output classification maps. The input file should be in geographic coordinates on WGS84. Only 8-bit unsigned GeoTIFF files are supported. The tool outputs the area in square meters and the number of pixels for each class present on the map.

To run the tool, use the following CMD command:

```
C:\GLAD_Tools\get_area.exe <input>.tif <output>.txt
```

<input> is the name of the map file; <output> is the name of the output area report text file.

The output file contains information on the area and pixel count of each data pixel value. The area is in square meters. The “i” column denotes pixel values.

i	area,m2	count,pixels
0	5646992607.4	13721122
1	20695919151.7	50278878

9.5. Zonal Statistics

The tool calculates the area of target pixel classes within zones defined by another raster layer. Both input files should have the same extent, pixel size, and UL coordinate and both should be 8-bit unsigned GeoTIFF raster files. The input files should be in geographic coordinates on WGS84. See section 9.2 on how to check the image size and projection.

To run the tool, use the following CMD command:

```
C:\GLAD_Tools\zonal_stat.exe <input>.tif <zones>.tif <output>.txt
```

<input> is the name of the map file; <zones> is the name of the zones map file; <output> is the name of the output area report text file.

The output file contains information on the area and pixel count of each data pixel value. The area is in square meters. The “class” columns denote pixel values. The header row denotes columns corresponding to zones.

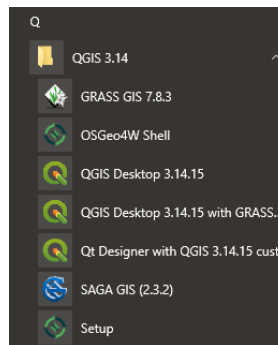
class	0	1	2	3	4
0	0.0	363759095.3	363481653.7	44602049.6	545098244.4
1	0.0	1338230723.6	2308410369.5	56567126.8	2228712247.0

9.6. Rasterizing Vector Data

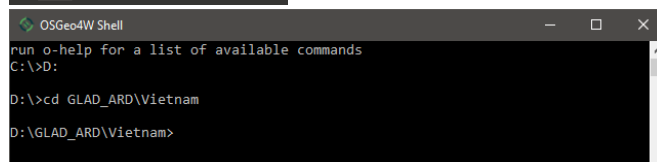
A raster version of a vector dataset (e.g., administrative boundary) may be required for some of the data processing operations such as classification masking, as input data for stratification, or to calculate zonal statistics. GLAD Tools does not include a separate tool for data rasterization because this operation can be easily performed using the GDAL tools provided by OSGeo4W. The following steps illustrate the data rasterization workflow.

1. Prepare the vector dataset. The dataset should have the following parameters:
 - a. ESRI shapefile format
 - b. Projection `+proj=longlat +ellps=WGS84 +datum=WGS84 +no_defs`. Projection information should be included in the shapefile prj file.
 - c. If an integer attribute is used to define raster values, the attribute value should be within the range 1-255.
2. Define the output raster extent from the existing raster file. In most cases, rasterized vector data is combined with the existing raster maps, and its extent and pixel size should match exactly the other data. To obtain information about the existing raster file, follow these steps:

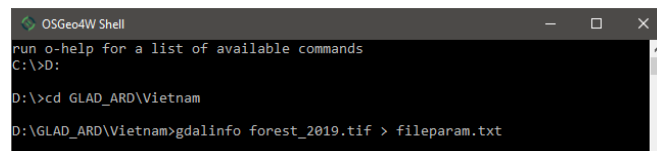
Open the OSGeo4W Shell interface using the Windows Start menu.



You can change the drive (D:) and path (cd GLAD_ARD_Vietnam) to navigate to the folder with the raster dataset.



Use the command `gdalinfo <input raster>.tif >fileparam.txt` to get the raster file parameter (replace <input raster> with the actual file name).



The “fileparam.txt” file contains information about the image extent and georeferencing. In the example below, the important lines are highlighted.

```
Driver: GTiff/GeoTIFF
Files: forest_2019.tif
Size is 4004, 4004
Coordinate System is:
GEOGCRS["WGS 84",
  DATUM["World Geodetic System 1984",
    ELLIPSOID["WGS 84",6378137,298.257223563,
      LENGTHUNIT["metre",1]],
    PRIMEM["Greenwich",0,
      ANGLEUNIT["degree",0.0174532925199433]],
    CS[ellipsoidal,2],
    AXIS["geodetic latitude (Lat)",north,
      ORDER[1],
      ANGLEUNIT["degree",0.0174532925199433]],
    AXIS["geodetic longitude (Lon)",east,
      ORDER[2],
      ANGLEUNIT["degree",0.0174532925199433]],
    ID["EPSG",4326]]
Data axis to CRS axis mapping: 2,1
Origin = (104.99949999999998,21.000499999999999)
Pixel Size = (0.000250000000000,-0.000250000000000)
Metadata:
  AREA_OR_POINT=Area
Image Structure Metadata:
  INTERLEAVE=BAND
Corner Coordinates:
Upper Left ( 104.9995000, 21.0005000) (104d59'58.20"E, 21d 0' 1.80"N)
Lower Left ( 104.9995000, 19.9995000) (104d59'58.20"E, 19d59'58.20"N)
Upper Right ( 106.0005000, 21.0005000) (106d 0' 1.80"E, 21d 0' 1.80"N)
Lower Right ( 106.0005000, 19.9995000) (106d 0' 1.80"E, 19d59'58.20"N)
Center ( 105.5000000, 20.5000000) (105d30' 0.00"E, 20d30' 0.00"N)
Band 1 Block=4004x2 Type=Byte, ColorInterp=Gray
```

Use the UL/LR boundaries and pixel size to define the output parameters for shapefile rasterization. The command format is (single line):

```
gdal_rasterize -te ulx lry lrx uly -tr 0.00025 0.00025 -ot Byte -of GTiff -co COMPRESS=LZW
-co BIGTIFF=IF_SAFER -a gird vector_zones.shp raster_zones.tif
-te ulx lry lrx uly          Raster extent (change according to fileparam.txt information)
-tr 0.00025 0.00025 Pixel size
-ot Byte -of GTiff -co COMPRESS=LZW -co BIGTIFF=IF_SAFER  output parameters of the raster file
-a <attribute name>        Attribute used as a pixel value.
```

If the raster should be used as a mask with a single class value (1), use `-burn 1` instead of `-a <attr>`

Example commands for rasterizing vector files

Using shapefile attribute “grid” to assign pixel value

```
gdal_rasterize -te 104.9995 19.9995 106.0005 21.0005 -tr 0.00025 0.00025 -ot Byte -of GTiff -co
COMPRESS=LZW -co BIGTIFF=IF_SAFER -a gird vector_zones.shp raster_zones.tif
```

Making the mask with value 1 for all polygons in the shapefile

```
gdal_rasterize -te 104.9995 19.9995 106.0005 21.0005 -tr 0.00025 0.00025 -ot Byte -of GTiff -co
COMPRESS=LZW -co BIGTIFF=IF_SAFER -burn 1 vector_zones.shp raster_mask.tif
```

10. Sample Analysis

10.1. Principles of Sample Analysis

National and global maps that are created using remotely sensed data are not suitable for direct calculation of land cover and change areas for official reporting. All maps derived from remotely sensed data are models. They contain modeling errors due to data limitations, classification/change detection algorithm limitations, and analyst errors. Map errors usually introduce bias that may be reflected by areas derived using pixel counting. Mixed pixels are one of the main causes of area estimation bias, especially affecting maps created at medium spatial resolution (10-30m). Because mixed pixel attribution can be subjective, small differences in training data may produce different results for the mixed pixels, changing the total class area. Satellite imagery does not always allow the mapping of similar land use types consistently. The uncertainties of certain mapped land use categories (e.g., tree plantations vs. orchards; crops vs. pastures) are usually too high for area reporting via map pixel counting.

Change detection possesses specific challenges for direct area estimation from the map. Because the change area is usually relatively small, its map-based estimation is uncertain. Certain techniques of change detection, like post-classification comparison, have high uncertainty and are usually not suitable for change area assessment.

The area uncertainty and map accuracy which are required for most national and international reporting cannot be estimated from the map alone. Map accuracy is also needed to compare the quality of different maps over the same area. While modern systems of satellite data characterization, including GLAD ARD Tools, simplify the production of new maps, map accuracy is essential to select the best data for national application.

Sample analysis that employs probability sampling allows estimation of the unbiased area of land cover classes and changes between these classes, estimation of the uncertainty of reported areas, estimation of map accuracy, and value-added thematic analysis based on sample reference data (e.g., differentiate land use types).

10.1.1. Sampling design

The sampling design specifies the method for selecting the reference sample. The entire area of interest (AOI) should be included in the sampling frame. Samples should be allocated using probability sampling, meaning that every population unit (e.g., pixel) should have a known and non-zero probability of being included in the sample. Some of the common probability sampling designs are simple random, systematic, stratified random, and stratified systematic sampling.

Stratified sampling is the recommended approach for most applications (Oloffson et al., 2014). Existing land cover and change maps support the creation of sampling strata that dramatically increase sampling efficiency (sample-based area estimation with low uncertainty while using fewer samples). The purpose of stratification is to:

- (1) minimize the within-stratum variance of target land cover and change classes;
- (2) target rare classes that would have been undersampled via simple random or systematic sampling;
- (3) target potential map errors (e.g., by creating buffers around rare classes, Oloffson et al., 2020).

The GLAD ARD Tools uses the Landsat pixel as a sampling unit. Sample size selection and allocation among strata depend on the expected precision of the resulting estimates and the existing knowledge about the distribution of target classes among strata. See examples of sample size computation below.

10.1.2. Response design

Response design specifies the protocol for obtaining “ground truth” information for each observation in the reference sample. The use of satellite data time series and commercial high-resolution data improves the precision of sample interpretation. The key requirements are the availability of reference satellite data for all sampled units, and the consistency of interpretation legend. Interpreted classes should be clearly defined,

mutually exclusive, and exhaustive, with physiognomic/structural attributes for land cover and observable traits of land uses that can be consistently estimated or identified from reference data. Every reference sample unit must be assigned to a unique category in the context of a set of definitions.

To ensure the quality of reference data collection, it is best to have all sample units independently interpreted by multiple experts. In cases of disagreements between interpreters, we are advocating for a consensus approach, where interpreters collectively revisit sample units with initial disagreements and work towards the final consensus interpretation. Even after such an iterative process, some sample units could be labeled as “low confidence”, e.g., due to the lack of cloud-free satellite data or difficulties associated with distinguishing certain land cover classes (e.g., natural grasslands vs. pastures). The high number of “low confidence” reference sample units might be an indicator that available satellite data, sample interpretation protocol, or class definitions are not sufficient/appropriate for the specific interpretation task.

10.1.3. Estimation and analysis protocol

The protocol for estimation of area, uncertainty, and quantification of the accuracy of the map is based on published and widely accepted principles (see References). The appropriate methods should be selected for each sampling design. Unbiased estimators of areas and map accuracy are available for all common probability sampling designs.

10.2. Sampling Design

A simple random or stratified random sampling design requires a source raster layer that defines sampling region (area of interest), which in case of stratified random sampling should be subdivided into sampling strata. The raster should be in the same format as ARD data, with pixel size and pixel boundary matching the ARD dataset. Only 8-bit GeoTIFF files are suitable for the following process. The raster defining sampling region (area of interest) can be obtained using the following methods:

- a. Rasterizing polygon vector layer. Follow instructions in section 9.6 for data rasterization. It is important to ensure that the pixel grid of the new raster is exactly the same as in ARD data.
- b. Using classification output. The classification results (likelihood) should be recoded to strata values using image recode tool (section 9.1) or image modeler (section 9.2).
- c. Strata can be defined using and intersection of different raster layers. An example of a process to create a land cover map from ARD data and classification outputs is provided in section 9.2.

Each pixel of the strata layer should have the value corresponding to the stratum ID (1– H). Pixels with values 0 are usually considered out of the sampling region. To calculate the strata area and pixels counts, use image area calculation tool (section 9.4). Use Excel (or other table editor) to create a sampling design based on strata area. For the guidance of selecting total sample size and allocating it among strata, see section 10.2.1. Make the output table following example below.

Example of sampling frame design in Excel.

1. Copy content of the area report text file to the table.
2. Create an output table using columns with stratum ID and pixel count (N_h). Zero values ignored.
3. Add column with sample size for each stratum (n_h , see section 10.2.1).

	A	B	C	D	E	F	G
1	area report (area_report.txt)						
2	i	area,m2	count,pixels				
3	0	73687031551	97039677			sampling design table	
4	1	1766608440	2331491		1	2331491	10
5	2	2573461390	3392517		2	3392517	10
6	3	5087855511	6710108		3	6710108	10
7	4	16419914442	21670502		4	21670502	10
8	5	5128351056	6755005		5	6755005	10
9	6	4699660554	6196716		6	6196716	10

To guide sample allocation procedure, user should make a parameter file following this format:

strata=lc_2019.tif	Strata – name of the sampling frame file.
R=C:/Program Files/R/R-4.0.2/bin/Rscript.exe	R – R installation (check your installation).

<pre>first=1 SAMPLING 1 233149110 2 339251710 END</pre>	<p>First – number of the first sample. Header of the sampling frame (do not remove) Each line (tab-separated): Strata_ID N_pixels n_samples End of the sampling frame (do not remove)</p>
---	---

Example of sampling parameter file.

Copy and paste the strata table to text file.

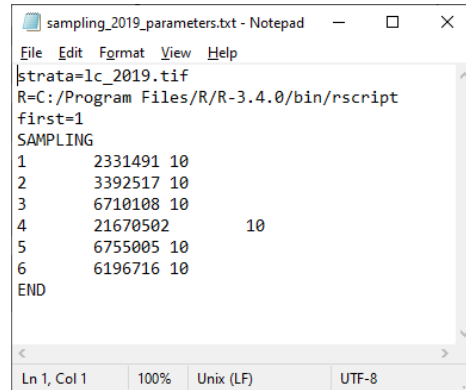
The table should start with the tag "SAMPLING" and end with the tag "END".

Add the following parameters:

strata – name of the file with sampling region (for simple random sampling) or sampling strata (for stratified random sampling).

R – R installation (check your installation).

first – ID of the first sample.

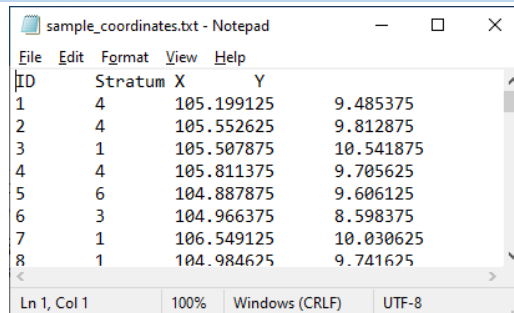


We recommend keeping the .tif file with sampling strata and parameter text file in the same folder. Open CMD in this folder and perform the following command:

```
perl C:/GLAD_Tools/samples_generate.pl <parameter file>.txt
```

The output file (sample_coordinates.txt) will contain the list of selected samples. The columns are: Sample ID; Stratum; Pixel center X; Pixel center Y. The samples are shuffled.

Example of sample table (sample_coordinates.txt)

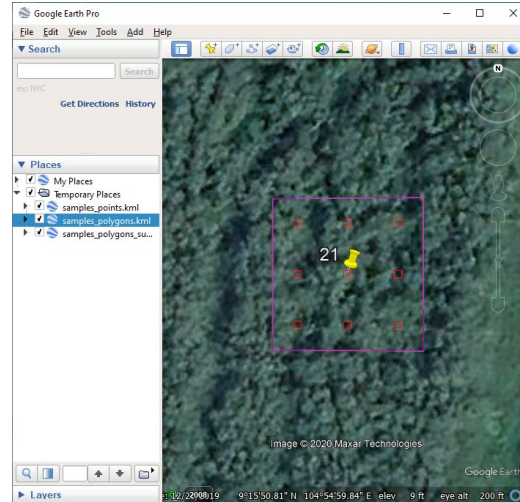


The sample table contains only the center point coordinates. To create a vector dataset with sample outlines, use sample_kml.pl tool. The tool can be executed using the following command:

```
perl C:/GLAD_Tools/samples_kml.pl sample_coordinates.txt
```

The output includes KML files of center points (samples_points.kml), sample outlines (samples_polygons.kml), and sample outlines with nine sub-plots for sub-pixel interpretation (samples_polygons_subplots.kml).

Example of sample represented by center point, outline, and outline with subplots KML layers



10.2.1. Sample Size Selection

Parameter “n_samples” in the parameter table in section 10.2 needs to be selected prior to allocating the sample. As a rule of thumb, a minimum of 30 sampling units should be selected in each sampling stratum. We suggest the following generic algorithm for computing required sample size and allocating it among strata, based on Cochran (1977):

Step 0 (optional): Allocate small initial number of sample units in each stratum (e.g., 30, 50 or 100 in each stratum), interpret them and derive target class proportions for each stratum (in case of binary sample labels, yes/no) or per-stratum sample variances;

Step 1: Estimate the total required sample size for the required precision under the optimal sample allocation assumption using stratum-specific class proportions or sample variances, either estimated from the initial sample (Step 0) or from the prior sampling and mapping experience (e.g. in Example 2 we could guess without initial sampling that “loss” stratum would have 85-90% of correctly mapped forest loss (0.85 class proportion from the initial sample), “gain” and “buffer” strata would have 5-10% of forest loss, omitted in a forest loss map used for stratification, and stable forest and no forest classes would have about 1% of forest loss).

Step 2: Compute optimal sample allocation among sampling strata or post-strata.

Step 3: Compare existing sample allocation with optimal sample allocation and increase sampling density in the strata or post-strata that have lower sample size compared to optimal.

Examples 1 and 2 below differ in terms of sample reference data: Example 1 uses proportional sample labels (e.g. 0, 50%, 100%), whereas in Example 2 sample reference classification is binary (yes/no). In both examples strata-specific sample variances s_h^2 (Example 1) or class proportions p_h (Example 2) are derived from the initial light sampling.

Example 1 (Proportional sample labels)

Step 1: Required sample size: Equation 5.47 from Cochran (page 105), for optimal allocation with fixed n:

$$n = \frac{(\sum W_h s_h)^2}{V + \frac{1}{N} \sum W_h s_h^2}$$

Where $W_h = N_h/N$

N_h – size of stratum h

N – total population size

s_h^2 – estimated sample variance of target class in stratum h

s_h – estimated sample standard deviation of target class in stratum h

V - target variance of the estimate of the mean proportion of target class

Current estimate of mean proportion of target class (from the study area) is 0.026836 with SE of 0.005262 (19.6%). Target SE = 10% or 0.002683556. Target variance (V) is hence 0.0000072015.

<i>Sampling strata</i>	<i>Initial n_h</i>	<i>N_h</i>	<i>W_h</i>	<i>s_h^2</i>	<i>s_h</i>	
Loss	1	43	3138841	0.021913031	0.228682171	0.478207246
Buffer	2	77	15079805	0.105275874	0.081168831	0.284901441
No loss*	3	119	125022198	0.872811096	0.00210084	0.045834925
Total			143240844			

Using these inputs in the equation above required $n = 899.33 \approx 899$

Step 2: Optimal sample allocation among strata (minimized variance for fixed n): Equation 5.26 from Cochran (page 98) and replacing true population standard deviation for each stratum with the one estimated from the sample:

$$n_h = n \frac{N_h s_h}{\sum N_h s_h}$$

Required sample size for each stratum:

$n_1 = 117$

$n_2 = 335$

$n_3 = 447$

Step 3: Initial sample sizes from the table above in each of the strata are smaller than required sample size with optimal allocation, so we need to add 74 samples to stratum 1, 258 samples to stratum 2, and 328 samples to stratum 3.



This might be a case where we would want to try out a slightly larger s_h for the big “No loss” stratum because if s_h is underestimated in the initial sample, the overall SE could be higher than what we’d want. In other words, a stratum that will be as influential as the “No loss” one in this case because it is 10 times bigger than any other stratum, it might be good to use a higher s_h just to be sure this stratum gets a large enough sample. There is no objective rule for this. For example, if we change s_h for “No loss” to 0.06 from 0.046, n_3 goes to 507, so in reality we might want to add a bit more to “No Loss”.

Example 2 (Binary sample labels)

Step 1: Required sample size: Equation 5.66 from Cochran (page 110), for the optimal allocation with fixed n:
For infinite populations

$$n_0 = \frac{(\sum W_h \sqrt{p_h(1-p_h)})^2}{V}$$

With finite population correction

$$n = \frac{n_0}{1 + \frac{1}{NV} \sum W_h p_h (1 - p_h)}$$

Where $W_h = N_h/N$

N_h – size of stratum h

N – total population size

p_h – proportion of target class in stratum h , estimated from the sample

V - target variance of the estimate of class proportion

Current estimate of mean proportion of target class (from the total study areas 0.029457 with SE of 0.006071 (20.6%). Target SE = 10% or 0.002945673. Target variance (V) is hence 0.0000086770

	<i>Stratum</i>	<i>N_h</i>	<i>Initial n_h</i>	<i>Samples, identified as target class</i>	<i>p_h</i>	<i>W_h</i>
1	stable non-forest	15583185	100	1	0.01	0.4648024
2*	core stable forest*	5997025	100	0	0	0.1788743
3	loss	552977	100	85	0.85	0.0164937
4	gain	843112	100	10	0.1	0.0251476
5	1-pix buffer around loss/gain	2213447	100	5	0.05	0.0660209
6	periphery stable forest (10-pix buffer inside)	8336728	100	2	0.02	0.2486661
Total		33526474				1

n₀ = 1366.3 ≈ 1366

n = 1366.2 ≈ 1366

There is no difference between n_0 and n , since our sampling population is very large ($N = 33,526,474$ pixels).

Step 2: Optimal sample allocation among strata (minimized variance for fixed n): Equation 5.60 from Cochran (page 108) and replacing true population class proportion for each stratum with the one estimated from the sample:

$$n_h = n \frac{N_h \sqrt{p_h(1-p_h)}}{\sum N_h \sqrt{p_h(1-p_h)}}$$

Required sample size for each stratum:

$n_1 = 580$ $n_2 = 0$ $n_3 = 74$ $n_4 = 95$ $n_5 = 181$ $n_6 = 437$

Step 3: We initially oversampled in strata 2, 3 and 4, but need to add more samples to strata 1, 5 and 6 (480, 81 and 337 samples respectively).

We might want to replace a 0 p_h for stratum 2 with something like 0.005 because if we use 0 the optimal allocation to this stratum will be sample size of 0.

10.4. Sample Data Visualization

10.4.1. Sample data extraction using ARD data

Sample reference data include (a) temporal profiles of NDVI $[(NIR-red)/(NIR+red)]$, NDWI $[(NIR-SWIR1)/(NIR+SWIR1)]$, and SWIR1 normalized surface reflectance; (b) annual and bi-monthly average normalized surface reflectance composites (which include the sample pixel surroundings providing landscape context for visual interpretation); and (c) high resolution Google Earth images. For the Google Earth data analysis, we provide sample outline in the KML format. All reference data are provided in HTML format that simplifies data access and image interpretation.

To create a set of reference data HTML pages, the following components are needed:

- ARD data for the entire analysis interval and entire area of analysis. E.g., if sample interpretation will be performed between the years 2000 and 2019, the entire ARD data from year 2000 to 2019 is required.
- The list of ARD tiles that cover the entire area of analysis (text format)
- The list of sample coordinates (sample_coordinates.txt) from the sampling tool (see section 10.1).

Sample data extraction controlled by a parameter file that has the following structure:

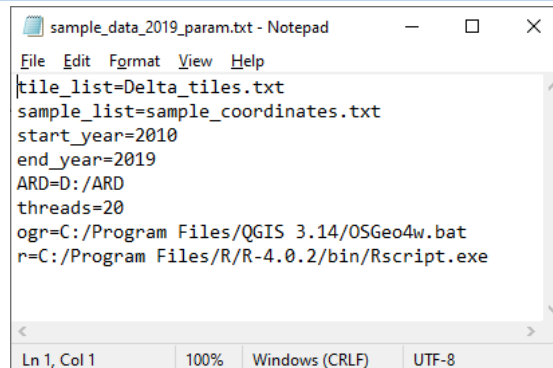
Sample reference data extraction parameter file structure

tile_list=tiles.txt	Name of the tile list file
sample_list=sample_coordinates.txt	Name of the sample coordinates file
start_year=2010	First year of reference data
end_year=2019	Last year of reference data
ARD=D:/ARD	Path to the ARD data folder
threads=20	Number of parallel processes
ogr=C:/Program Files/QGIS 3.14/OSGeo4w.bat	A link to OSGeo4w.bat file (check your local QGIS installation)
r=C:/Program Files/R/R-4.0.2/bin/Rscript.exe	A link to Rscript.exe file (check your local R installation)

* The reference sample extraction will not work if the first and the last year of the reference data are the same.

* Unlike metric generation and classification processes, the reference sample extraction does not require large RAM. We recommend to use all available cores for data processing.

Example of parameter file



```

sample_data_2019_param.txt - Notepad
File Edit Format View Help
tile_list=Delta_tiles.txt
sample_list=sample_coordinates.txt
start_year=2010
end_year=2019
ARD=D:/ARD
threads=20
ogr=C:/Program Files/QGIS 3.14/OSGeo4w.bat
r=C:/Program Files/R/R-4.0.2/bin/Rscript.exe
Ln 1, Col 1 100% Windows (CRLF) UTF-8

```

We recommend using the same folder as for sample allocation (section 9.1.) or to create a new folder that should contain:

- Tile list text file
- Parameter file as described above
- List of sample coordinates (output of sampling tool)

Open CMD in this folder and run the following command:

```
perl C:/GLAD_Tools/samples_data.pl sample_data_2019_param.txt
```

The command may take long time to complete. If errors are found, remove all temporary folders and files before re-running the command.

The reference data are stored in a new folder Sample_Data. For each sample, there are annual image composites, bi-monthly composites, temporal profile, and KML file. A set of html pages are created for each sample to display annual and monthly data. The image index is located in the root folder (image.html). The index file allows navigation to any sample. Each sample page contains temporal profiles of NDVI, NDWI, and SWIR1 reflectance. The indices and reflectance scaling were selected to simplify visual analysis: NDVI and NDWI values scaled to range 0-2; and SWIR2 reflectance scaled to range 1-3. The monthly composites open in pop-up window when clicking on image composite area. The KML may be opened automatically in Google Earth (check browser settings). We recommend using Chrome browser to work with sample data.

The link to Google Earth simplifies VHR data analysis

index.html allows navigation between samples.

Sample page provides navigation links, reflectance temporal profile and annual average reflectance image composites (SWIR1-NIR-red band combination).

Bi-monthly data provides average reflectance for every two month.

10.4.2. Sample data extraction using ARD and supplementary data.

In addition to the ARD data, each sample may be accompanied with the supplementary raster data extracted from the additional high-resolution images. External raster data should conform to the following format requirements:

1. The data should be stored locally.
2. The data should cover all samples (otherwise, sample image footprints will be blank).
3. The raster data should have a correctly set projection that is suitable for transformation to EPSG:4326 using the installed version of OSGeo4W.
4. Only 8-bit and 16-bit unsigned formats are supported. The first three image bands will be used for sample visualization as an RGB composite.
5. The data may be in any raster format supported by GDAL. VRT format is also supported.

The supplementary data extraction is guided by the data list. All data sources must be listed in the text file (suggested name "extra_datasets.txt"). The data list format is:

<output name>,<full path>,<format>,<scaling>

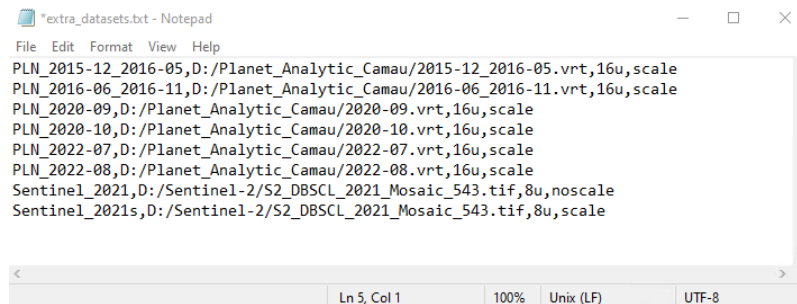
Output name: the name of the dataset in the visualization page.

Full path: path to the dataset file (raster or VRT)

Format: 8u or 16u

Scaling: "scale" or "noscale". The parameter is used only for 8u data processing. All 16u data will be scaled. Scaling refers to local (within image chip) histogram stretching to min/max range. We suggest using scaling when image data sources are not intended for visualizing (i.e., false-color Sentinel-2 image composites). Images designed for visualization (like Planet natural color image tiles) should not be scaled.

Example of external data file



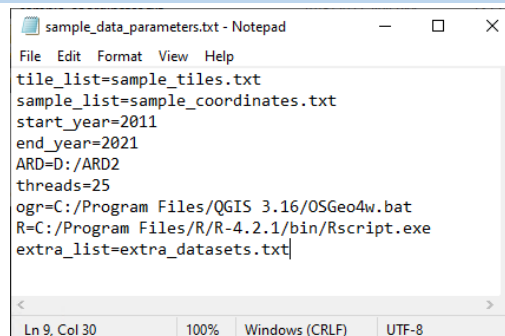
```

*extra_datasets.txt - Notepad
File Edit Format View Help
PLN_2015-12_2016-05,D:/Planet_Analytic_Camau/2015-12_2016-05.vrt,16u,scale
PLN_2016-06_2016-11,D:/Planet_Analytic_Camau/2016-06_2016-11.vrt,16u,scale
PLN_2020-09,D:/Planet_Analytic_Camau/2020-09.vrt,16u,scale
PLN_2020-10,D:/Planet_Analytic_Camau/2020-10.vrt,16u,scale
PLN_2022-07,D:/Planet_Analytic_Camau/2022-07.vrt,16u,scale
PLN_2022-08,D:/Planet_Analytic_Camau/2022-08.vrt,16u,scale
Sentinel_2021,D:/Sentinel-2/S2_DBSCL_2021_Mosaic_543.tif,8u,noscale
Sentinel_2021s,D:/Sentinel-2/S2_DBSCL_2021_Mosaic_543.tif,8u,scale
Ln 5, Col 1    100%    Unix (LF)    UTF-8

```

The sample extraction parameter file is similar to the one presented in section 10.4.1. The only difference is the addition of the path to the external dataset: **extra_list=extra_datasets.txt**

Example of data extraction parameter file



```

sample_data_parameters.txt - Notepad
File Edit Format View Help
tile_list=sample_tiles.txt
sample_list=sample_coordinates.txt
start_year=2011
end_year=2021
ARD=D:/ARD2
threads=25
ogr=C:/Program Files/QGIS 3.16/OSGeo4w.bat
R=C:/Program Files/R/R-4.2.1/bin/Rscript.exe
extra_list=extra_datasets.txt
Ln 9, Col 30    100%    Windows (CRLF)    UTF-8

```

Open CMD in this folder and run the following command:

```
perl C:/GLAD_Tools/samples_data_extra.pl sample_data_parameters.txt
```

The HTML pages for the samples will have an additional link [“Supplementary Data”](#) in the header for imagery access.

10.5. Sample Interpretation

Sample interpretation is the most important part of the sample analysis. Interpretation should be performed by analysts that have knowledge of regional land cover and the relationship between land cover types and their spectral properties. If several interpreters are working independently, a quality control and disagreement resolution protocols should be implemented. The classification legend should be selected taking in account the ability of analysts to discriminate classes. The protocol for dealing with mixed pixels should be implemented to avoid interpretation errors.

We recommend using spreadsheet editor, e.g. MS Excel or Google Sheets, to record sample interpretation results. The sample interpretation table should have sample IDs, but not information about strata as it may cause interpretation bias. The land cover or change classes may be recorded using different formats. Here are some examples of sample reference data recoding.

Examples of sample data recording.

Simple land cover classes. The entire pixel is assigned to the majority land cover class. Class abbreviations used to simplify data entry.

	A	B	C	D
1	ID	X	Y	category
2	1	130.6144	-23.4136	SNC
3	2	139.5139	-27.4239	SNC
4	3	143.6766	-37.8344	SNC
5	4	145.7281	-34.0169	CG
6	5	177.3454	-38.0679	SNC
7	6	146.7619	-31.9441	cl
8	7	130.8436	-28.8001	SNC
9	8	141.7591	-17.9009	SNC
10	9	115.5271	-28.3911	SNC
11	10	131.0376	-18.6859	SNC
12	11	148.3834	-26.6294	SNC
13	12	147.6321	-22.4626	CG

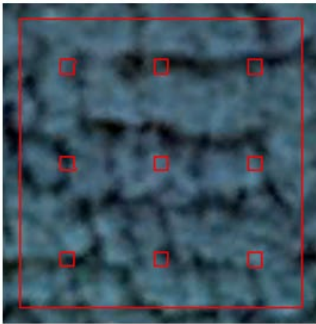
Land cover classes recoded for the first and last year of analysis. Net change can be interpreted as change between these classes. Gross disturbance dynamics recoded in between the first and last year.

	A	C	D	E	F	G	H	I	J	K	L	M
1	ID	X	Y	1988	loss1	loss2	loss3	loss4	loss5	loss6	max	2019
2	1	101.572	20.1446	Forest	2017						2017	Forest
3	2	106.559	15.9754	Forest							0	Forest
4	3	102.386	18.2826	Forest	2008						2008	Forest
5	4	104.129	19.6986	Forest	1995	2008	2015				2015	Forest
6	5	101.591	18.5159	Forest	2014						2014	Forest
7	6	103.439	20.1146	Forest							0	Forest
8	7	105.544	17.2099	Forest							0	Forest
9	8	100.562	20.8909	Forest	1998	2008					2008	Forest
10	9	105.984	17.1991	Forest							0	Forest
11	10	106.135	17.2386	Forest	2002	2016					2016	Forest
12	11	105.769	17.0231	Forest	1989	1994					1994	Pasture
13	12	104.016	20.3636	Forest	1993	2008	2019				2019	Agriculture

Sample data recoded as proportion of land cover classes within each pixel. The fraction if 0.25, 0.5, and 0.75 are recoded.

ID	Cropland	Rural Mosaic	Wetland	Aquaculture	Mangroves	Build-up
1	1					
2						1
3	0.5	0.5				
4			1			
5	1					
6		0.5		0.5		
7				0.5	0.5	
8	0.5					0.5
9			1			
10		0.5				0.5
11		0.5		0.5		

Example of sample interpretation (tree canopy cover) using sub-plots



9 sub-plots intersect canopy



3 sub-plots intersect canopy



0 sub-plots intersect canopy. Even there is trees within this sample none of the sub-plots intersect it.

The area of class recorded as the count of sub-plots that intersect with the class. We used 9 subplots. The percent can be calculated from the count (n) using equation: % tree cover = $n/9 \times 100$

	A	B	C
1	ID	Tree cover	Tree cover, %
2	1	0	0
3	2	0	0
4	3	0	0
5	4	0	0
6	5	2	22
7	6	0	0
8	7	5	56
9	8	9	100
10	9	8	89
11	10	9	100
12	11	0	0
13	12	3	33
14	13	1	11
15	14	0	0
16	15	0	0

10.6. Map Accuracy Estimation

10.6.1. Equations

The following equations are for stratified random sampling design (the stratum number denoted as h , total number of strata H). The sampling unit is defined as a Landsat pixel. Please note that equations below are not suitable for global applications when sampling GLAD ARD Lat/Long pixels, because the actual size of the Lat/Long pixel rapidly decreases close to the poles; one pixel on equator represents the same area as multiple pixels in the polar regions. Thus, when sampling pixels with equal probability, polar regions will be overrepresented in the global sample compared to equatorial regions. The equations below assume approximately the same sample size across the study area, and equal sampling probability for each pixel. Hence, they are suitable for large regional, national, and sub-national applications in tropical and sub-tropical regions, national and sub-national applications in temperate regions, and sub-national applications in boreal and polar regions. Both map and reference data attribute the entire pixel for one land cover class only. These equations are for estimating accuracy of one (target) class only. The estimation is repeated for each land cover class.

The inclusion probability for pixel u in stratum h is $\pi_u = n_h/N_h$ with a stratum population size (N_h) and stratum sample size (n_h). The number of pixels in the population (i.e., region of interest) is N . Stehman (2014) derived the stratified estimates of accuracy using indicator functions (Cochran, 1977) and x_u , where these observations obtained for pixel u have just two possible values, 0 or 1. For estimating user's accuracy of the target land cover class, define $y_u = 1$ if sample pixel u is correctly classified target land cover class otherwise $y_u = 0$, and define $x_u = 1$ if sample pixel u is classified as target land cover class, otherwise $x_u = 0$. For the producer's accuracy, define $y_u = 1$ if sample pixel u is correctly classified as target land cover class, otherwise $y_u = 0$, and define $x_u = 1$ if sample pixel u has reference target class, otherwise $x_u = 0$. The estimator of user's accuracy and producer's accuracy is then expressed as a ratio estimator:

$$\hat{R} = \frac{\sum_{h=1}^H N_h \bar{y}_h}{\sum_{h=1}^{25} N_h \bar{x}_h}$$

where h is the stratum index, $\bar{y}_h = \sum_{u \in h} y_u / n_h$ is the sample mean of the y_u values in stratum h , and \bar{x}_h is the sample mean of the x_u values of stratum h . The variance estimator for \hat{R} is:

$$\hat{V}(\hat{R}) = \frac{1}{\hat{X}^2} \sum_{h=1}^H N_h^2 (1 - n_h/N_h) (s_{y_h}^2 + \hat{R}^2 s_{x_h}^2 - 2\hat{R} s_{xy_h}) / n_h$$

where $\hat{X} = \sum_{h=1}^H N_h \bar{x}_h$, $s_{y_h}^2$ is the sample variance of the y_u values in stratum h , $s_{x_h}^2$ is the sample variance of the x_u values within stratum h , and s_{xy_h} is the sample covariance between x_u and y_u of stratum h (see equations below):

$$s_{y_h}^2 = \sum_{u \in h} (y_u - \bar{y}_h)^2 / (n_h - 1)$$

$$s_{x_h}^2 = \sum_{u \in h} (x_u - \bar{x}_h)^2 / (n_h - 1)$$

$$s_{xy_h} = \sum_{u \in h} (y_u - \bar{y}_h)(x_u - \bar{x}_h) / (n_h - 1)$$

To estimate overall accuracy, define $y_u = 1$ if pixel u is classified correctly and $y_u = 0$ if pixel u is classified incorrectly. The estimator for overall accuracy is then expressed as:

$$\hat{O} = \sum_{h=1}^H N_h \bar{y}_h / N$$

The variance estimator for \hat{O} is:

$$\hat{V}(\hat{O}) = \frac{1}{N^2} \sum_{h=1}^H N_h^2 (1 - n_h/N_h) s_{y_h}^2 / n_h$$

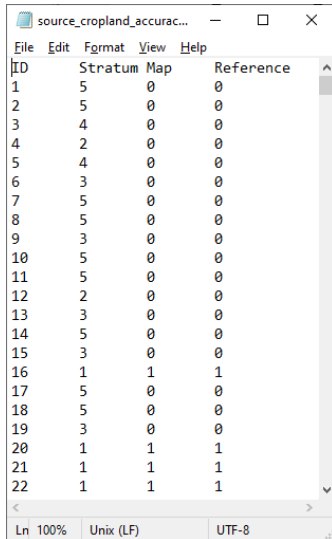
where $s_{y_h}^2$ is computed as shown above.

10.6.2. Accuracy of a Single LCLU Class

The GLAD Tools provide a code to estimate classification accuracy of a single land cover class. If the map has several classes, the accuracy of each class should be estimated separately. The input data must attribute the entire pixel to a land cover class. In case of reference data with proportional labels e.g., percent tree cover), user will need to convert this data into yes/no classes (e.g. using tree cover threshold to define "forest" class). The

accuracy of land cover change map is estimated using the same approach. The land cover change class in this case is treated as the target class.

The first step of map accuracy estimation is preparation of the sample table. The table should be prepared in Excel (or other spreadsheet editor) and exported (copy/paste) to a new text file. The output table have the following structure:



Columns are tab-separated. Table include a header line:

ID Stratum Map Reference

Each row has four columns (tab-separated):

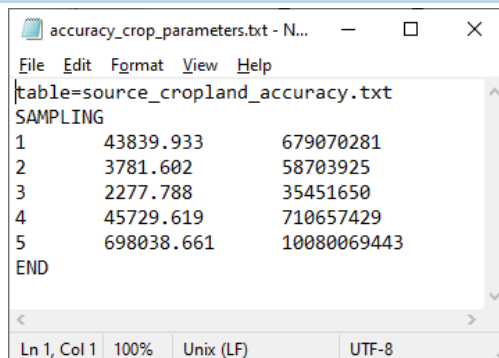
- Sample ID
- Sample stratum ID
- Map value (1/0 indicate Y/N of a pixel mapped as target land cover class)
- Reference value (1/0 indicate Y/N of a pixel interpreted as target land cover class)

The second step is to create a parameter file. The parameter file has the following structure:

Map accuracy parameter file structure

table=source_crop2000_area.txt	Name of the text file with sample table
SAMPLING	Header of the sampling frame (do not remove)
1 area pixel count	Each line (tab-separated):
2 area pixel count	Stratum_ID Area, A_h (any units) N_{pixels} , N_h
...	
END	End of the sampling frame (do not remove)

Example of the parameter file for map accuracy estimation



Both sample table and parameter file should be in the same folder. To run the estimation, open CMD in this folder and use the following command:

```
perl C:/GLAD_Tools/samples_estimate_accuracy.pl <parameter file name>
```


The map accuracy estimation report file will be named after the source sample table by adding suffix "Accuracy_report_". The report format:

Confusion matrix provides number of samples in each category of agreement between map and reference data

- **Map/Ref** – both map and reference assigned to the target class
- **Map/0** – commission errors
- **0/Ref** – omission errors
- **0/0** – both map and reference data show class absence

Strata	Map/Ref	Map/0	0/Ref	0/0
1	85	15	0	0
2	0	0	15	85
3	0	0	14	86
4	0	0	1	99
5	0	0	0	100

Accuracy estimates (in percent). SE stands for standard error (in percent). Classes are: class1 – target, class0 – other.

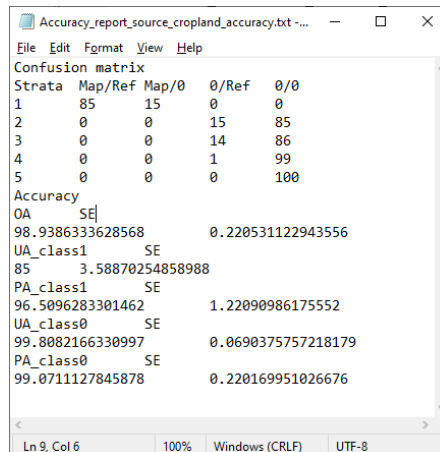
OA – Overall accuracy

UA – User’s accuracy (reflects map commission errors)

PA – Producer’s accuracy (reflects map omission errors)

OA	SE
98.93863336	0.220531123
UA_class1	SE
85	3.588702549
PA_class1	SE
96.50962833	1.220909862
UA_class0	SE
99.80821663	0.069037576
PA_class0	SE
99.07111278	0.220169951

Example of area report for a single land cover class



10.6.3. Overall Accuracy for Multiple Classes

For maps containing multiple LCLU classes, the accuracy tool must be implemented for each class separately to calculate the overall, user's and producer's accuracy statistics. The overall accuracy may also be estimated for all classes, considering the map and reference LCLU class agreement for each sample. The overall accuracy estimation is possible if:

1. The map and reference data have the same LCLU class legend.
2. The LCLU classes are presented as integer values in the table.

The source table should be exported as 4-column text file with map and reference class values (see example below).

ID	Stratum	map	ref
1	3	4	4
2	4	1	1
3	6	2	1
4	4	1	2
5	2	4	4
6	2	4	4
7	3	4	2
8	5	2	2
9	4	1	1
10	5	2	2

The parameter file should be created similar to the single class accuracy assessment (section 10.6.2). In the folder that contains both source table and parameter file, run the following command:

```
perl C:/GLAD_Tools/samples_estimate_oa.pl parameter_file.txt
```

The report file will contain the following information:

Strata	Map==Ref	Map!=Ref
1	17	33
2	45	5
3	31	19
4	38	12
5	45	5
6	24	26

Accuracy
 OA SE
 75.6719409054407 2.62711151607678

10.7. Area and Uncertainty Estimation

The sample-based area estimation was performed following Stehman et al. (2014). The area estimation requires information on total strata size, number of samples, and average proportion of the class area within samples of each strata. Equations below and computations scripts are suitable for both binary sample labels (with 0 corresponding to “no” and 100 to “yes” target class for each sample unit), and proportional sample labels (with 0-100% representing percentage of sample unit identified as target class).

Please note that equations below are not suitable for global applications when sampling GLAD ARD Lat/Long pixels, because the actual size of the Lat/Long pixel rapidly decreases close to the poles; one pixel on equator represents the same area as multiple pixels in the polar regions. Thus, when sampling pixels with equal probability, polar regions will be overrepresented in the global sample compared to equatorial regions. Equations below assume approximately the same sample size across study area, and equal sampling probability for each pixel. Hence, they are suitable for large regional, national and sub-national applications in tropical and

sub-tropical regions, national and sub-national applications in temperate regions, and sub-national applications in boreal and polar regions.

Area of land cover class (area units)

$$\hat{A} = A_{tot} * \sum_{h=1}^H \frac{N_h}{N} \bar{p}_h$$

A_{tot} – total area of interest (e.g., country area);
 H – number of sampling strata;
 N – total number of pixels in the AOI;
 n_h – sample size (number of sampled pixels) in stratum h ;
 N_h – total number of pixels in stratum h ;
 $\bar{p}_h = \frac{\sum_{u \in h} p_u}{n_h}$, mean proportion of the land cover class in stratum h , where p_u is the proportion of land cover class within each sampled pixel (p_u ranges from 0 to 1).

The standard error of the sample-based estimate is based on the variance of land cover proportion for samples in each stratum. The 95% confidence interval was obtained by multiplying standard error by 1.96.

Standard error of land cover class area (area units)

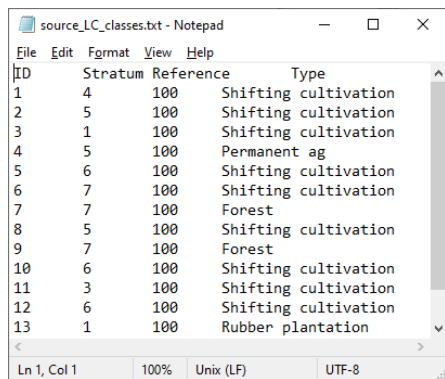
$$SE(\hat{A}) = A_{tot} * \sqrt{\frac{\sum_{h=1}^H N_h^2 (1 - \frac{n_h}{N_h}) \frac{S_{ph}^2}{n_h}}{N^2}}$$

where $S_{ph}^2 = \frac{\sum_{u \in h} (p_u - \bar{p}_h)^2}{n_h - 1}$, sample variance for stratum h .

The area is estimated separately for each land cover class. The input data may include all land cover classes if each sample assigned to a single land cover class. However, if a fraction of land cover class was recorded during interpretation, the area estimation procedure has to be performed for each class independently.

Land cover change is estimated using the same approach. The land cover change proportion of a pixel is used as a land category. Each land cover change transition can be analyzed separately.

The first step of land cover class area estimation is preparation of the sample table. The table should be prepared in Excel (or other spreadsheet editor) and exported (copy/paste) to a new text file. The output table have the following structure:



Columns are tab-separated. Table include a header line:

ID Stratum Reference Type

Each row has four columns (tab-separated):

- Sample ID
- Stratum ID
- Percent of land cover class (0-100 for proportional labels or 0/100 for yes/no)
- Name of land cover class (area of multiple classes can be estimated at the same time, if 100% of each pixel is assigned only to one class)

Two types of input files may be used regarding the land cover class interpretation:

1. When each pixel is attributed to a single land cover class (or land cover change class), the table should include value 100 for the "Percent of land cover class" column and a name or an individual class in "Name of land cover class" column as in the following example:

ID	Stratum	Reference	Type
1	4	100	Shifting cultivation
2	5	100	Shifting cultivation
3	1	100	Shifting cultivation
4	5	100	Permanent ag
5	6	100	Shifting cultivation

2. When proportion of land cover class was recoded, each land cover class processed separately. For each class, the table include percent of target land cover class in "Percent of land cover class" column and the class name in "Name of land cover class" column are the same for each sample:

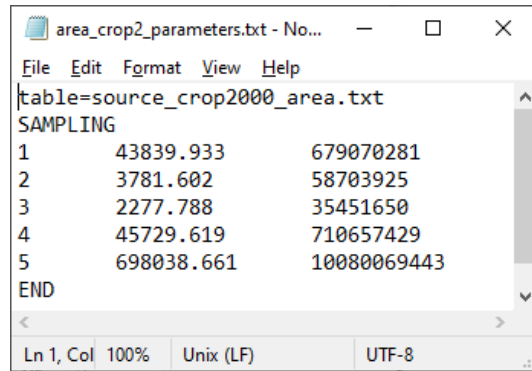
ID	Stratum	Reference	Type
1	1	100	CROP
2	2	0	CROP
3	1	50	CROP
4	2	0	CROP
5	1	100	CROP
6	2	0	CROP
7	2	0	CROP

The second step is to create a parameter file. The parameter file has the following structure:

Sample area parameter file structure

table=source_crop2000_area.txt	Name of the text file with sample table
SAMPLING	Header of the sampling frame (do not remove)
1 area pixel count	Each line (tab-separated):
2 area pixel count	Stratum_ID Area, A_h (any units) N_pixels, N_h
...	
END	End of the sampling frame (do not remove)

Example of the parameter file for area estimation



Both sample table and parameter file should be in the same folder. To run the estimation, open CMD in this folder and use the following command:

```
perl C:/GLAD_Tools/samples_estimate_area.pl <parameter file name>
```

The area estimation report file will be named after the source sample table by adding suffix "Area_report_". The report format:

Strata table

code	area	count	n_samples
1	43839.933	679070281	100
2	3781.602	58703925	100
3	2277.788	35451650	100
4	45729.619	710657429	100
5	698038.661	10080069443	100

Strata table is the copy of the table provided by the parameter file. The number of samples is calculated from the sample table. If this number is not correct, it is due to a possible error in sample table.

The area estimation table shows the following statistics (for each target estimation class)

- Proportion of the class within AOI
- Class area (in area units used in the parameter file)
- Standard error of area estimation (area units)
- 95% confidence interval of area estimation (area units)
- Percent 95% confidence interval of estimated class area (percent)

The "Total" row shows the same information if only one class was used, or shows the total AOI area if multiple classes were used for the analysis.

Type	proportion	area	SE	Conf95%	%Est
Total	0.052915408	41997.24501	1752.481273	3434.863295	8.178782427
CROP	0.052915408	41997.24501	1752.481273	3434.863295	8.178782427

Example of area report for a single land cover class

Area_report_source_crop2000_area.txt - Notepad

```
File Edit Format View Help
strata table
code area count N_samples
1 43839.933 679070281 100
2 3781.602 58703925 100
3 2277.788 35451650 100
4 45729.619 710657429 100
5 698038.661 10080069443 100

Type proportion area SE Conf95% %Est
Total 0.0529154079736394 41997.2450082055 1752.48127272526 3434.86329454152 8.17878242696732
CROP 0.0529154079736394 41997.2450082055 1752.48127272526 3434.86329454152 8.17878242696732
```

Ln 1, Col 1 100% Windows (CRLF) UTF-8

Example of an area report for multiple land cover classes

Area_source_types_area.txt - Notepad

```
File Edit Format View Help
strata table
code area count N_samples percent_interpreted number_50_and_above
1 180.671 2470745 50 50
2 98.850 1347837 50 50
3 1267.387 17315344 50 50 50
4 1775.874 24356351 100 100 100
5 2360.528 32445125 100 100 100
6 2377.028 32721687 100 100 100
7 14980.167 204914651 50 50 50

Type proportion area SE Conf95% %Est
Total 1 23040.505 0 0
Agroforestry 0.00620388245157821 142.940584645 57.6615886559602 113.016713765682 79.0655180586848
SC - Agroforestry 0.00419933632206737 96.7548295252748 48.2748887290679 94.6187819089731 97.792309048776
Agroforestry - Rubber 0.000156588482859714 3.60787772227164 3.60784121611687 7.07136878358906 195.998016782472
Forest 0.383289596051915 8831.18585428213 1056.36989645457 2070.48499705095 23.4451525674437
Forest - Agroforestry 0.000771816608166498 17.7830444195432 17.7830079135357 34.8546955105299 195.999597640465
Forest (disturbed) 0.0218956855895905 504.487653305388 305.892146512762 599.548607165014 118.843068455053
Forest - Perm Ag or Pasture 0.00812800379400259 187.273312055736 63.8005186401322 125.049016534659 66.7735381843636
Perm Ag or Pasture 0.0389606466662699 897.672974317426 507.584508518129 994.865636695532 110.827179291212
Forest - Rubber 0.00327723233392192 75.5090879758896 21.4649781562365 42.0713571862236 55.716945223411
```

Ln 1, Col 1 100% Windows (CRLF) UTF-8

Appendix 1. The GLAD ARD Methodology

A1.1. Source Landsat Imagery

We employ the archive of Landsat TM, ETM+, and OLI/TIRS data collected from the year 1997 to present available from the USGS EROS Data Center (<https://earthexplorer.usgs.gov/>). The Landsat Collection 2 Tier 1 data meets the highest geometric and radiometric standards, hence only those data are employed for ARD processing. We downloaded Tier 1 Landsat imagery for the 8352 World Reference System-2 (WRS) scenes which are located within ice-free land area. Small islands (where no Tier 1 data exist) and the high Arctic and Antarctic regions are excluded from ARD processing. Images affected by seasonal snow cover are excluded from processing. The seasonal snow cover was analyzed using the MODIS/Terra Snow Cover Monthly L3 Global product (<https://nsidc.org/data/MOD10CM/versions/6>) and Landsat imagery.

A1.2. Conversion to Radiometric Quantity

Due to the differences in spectral band configuration between Landsat sensors, only spectral bands with matching wavelengths between TM, ETM+, and OLI/TIRS sensors are processed. For the thermal infrared data, we use the high-gain mode thermal band (band 62) of the ETM+ sensor and 10.6–11.19 μ m thermal band (band 10) of the TIRS sensor. Landsat Collection 2 data contain radiation measurements for reflective visible/infrared bands in the form of scaled reflectance (OLI) or radiance (TM/ETM+) recorded as integer digital numbers (DNs). We convert the data into top-of-atmosphere (TOA) reflectance, scaled consistently across all Landsat sensors. Spectral reflectance (value range from zero to one) is scaled from 1 to 40,000 and recorded as a 16-bit unsigned integer value.

Landsat spectral bands used for ARD processing and corresponding MODIS spectral bands

Band name	Wavelength, nm			
	Landsat 5 TM	Landsat 7 ETM+	Landsat 8 OLI/TIRS	MODIS
Blue	450–520	441–514	452–512	459–479
Green	520–600	519–601	533–590	545–565
Red	630–690	631–692	636–673	620–670
Near-Infrared (NIR)	760–900	772–898	851–879	841–876
Shortwave Infrared 1 (SWIR1)	1,550–1,750	1,547–1,749	1,566–1,651	1,628–1,652
Shortwave Infrared 2 (SWIR2)	2,080–2,350	2,064–2,345	2,107–2,294	2,105–2,155
Thermal	10,410–12,500	10,310–12,360	10,600–11,190	10,780–11,280

TOA reflectance conversion method for TM and ETM+ sensors:

$$\rho = (\pi \times d^2 \times (G \times DN + B)) / (ESUN \times \sin(\text{sunelev} \times \pi / 180)) \times 40,000$$

ρ – scaled TOA reflectance; π – pi constant; d – Earth-Sun distance; G – gain factor; DN – original digital number; B – bias factor; $ESUN$ – mean exoatmospheric solar irradiance; sunelev – solar elevation angle. Parameters d , G , B , and $ESUN$ are taken from Chander et al. (2009). Parameter sunelev comes from the image metadata.

TOA reflectance conversion for the OLI sensor:

$$\rho = (0.0002 \times DN + 0.1) / (\sin(\text{sunelev} \times \pi / 180)) \times 40,000$$

ρ – scaled TOA reflectance; π – pi constant; DN – original digital number; sunelev – solar elevation angle from the image metadata.

The thermal band is converted into brightness temperature and recorded in Kelvin × 100 to preserve measurement precision:

$$T_B = K2 / \log(K1 / (G \times DN + B) + 1) \times 100$$

T_B – scaled brightness temperature; $K1$ and $K2$ – calibration coefficients; G – gain factor; DN – original digital number; B – bias factor. Parameters G , B , $K1$, and $K2$ are taken from Chander et al. (2009) for TM/ETM+ sensors and from the image metadata for the TIRS sensor.

A1.3. Observation Quality Assessment

The per-pixel observation quality assessment is used to highlight observations with a high probability of atmospheric contamination by clouds, haze, or cloud shadows. In addition, observation quality assessment performs generic snow/ice and water mapping. Observation quality assessment is based on the aggregation of the Landsat quality assessment band and GLAD quality assessment model output.

The Landsat Collection 2 data include a Quality Assessment (QA) band based on the globally consistent CFMask cloud and cloud shadow detection algorithm. The QA band contains the cirrus cloud (Landsat 8 only), clouds, cloud shadow, snow/ice, and radiometric saturation flags. The GLAD observation quality assessment model developed by our team represents a set of regionally adapted decision tree ensembles to map the likelihood of a pixel to represent cloud, cloud shadow, heavy haze, and, for clear-sky observations, water, or snow/ice. The model outputs represent likelihoods of assigning a pixel to the cloud, shadow, haze, snow/ice, and water classes. The masks were subsequently aggregated into an integral observation Quality Flag (QF) that highlights cloud/shadow contaminated observations, separates topographic shadows from likely cloud shadows, and specifies the proximity to clouds and cloud shadows. To derive QF, we implement buffering around cloud and shadow pixels, calculate the distance to clouds (along cloud shadow projection), and calculate areas affected by topographic shadows using the DEM and sun position. The QF is stored as band 8 of the GLAD ARD raster format. QF values 1, 2 and 15 indicate clear-sky observations. QF values 11–14 and 16–17 are considered clear-sky data with an indication of cloud/shadow proximity. QF values 5 and 6 indicate seasonal data quality issues (topographic shadows and snow cover). These observations may be removed from data processing if the number of clear-sky observations is sufficient. QF values 3, 4, and 7–10 are considered contaminated by clouds and shadows and are usually excluded from data processing.

Per-pixel observation quality flag (QF)

<i>QF</i>	<i>Observation quality</i>	<i>QF assignment criteria</i>
0	No Data	
1	Land	Clear-sky land observation.
2	Water	Clear-sky water observation.
3	Cloud	Cloud detected.
4	Cloud shadow	Shadow detected. The pixels are located within the projection of a detected cloud. Cloud projection defined using solar elevation and azimuth and limited to 9 km distance from the cloud.
5	Topographic shadow	Shadow detected. The pixel located outside cloud projections and within estimated topographic shadow (estimated using DEM and solar elevation and azimuth).
6	Snow/Ice	Snow or ice detected.
7	Haze	Dense semi-transparent clouds/fog detected.
8	Cloud proximity	Aggregation (OR) of two rules: (i) 1-pixel buffer around detected clouds. (ii) Above-zero cloud likelihood (estimated by GLAD cloud detection model) within 3-pixel buffer around detected clouds.
9	Shadow proximity	Shadow likelihood (estimated by GLAD shadow detection model) above 10% for pixels either (i) located within the projection of a detected cloud; OR (ii) within 3 pixels of a detected cloud or cloud shadow.
10	Other shadows	Shadow detected. The pixel located outside the projection of a detected cloud and outside of estimated topographic shadow.
11	Additional cloud proximity over land	Clear-sky land pixels located closer than 7 pixels of detected clouds
12	Additional cloud proximity over water	Clear-sky water pixels located closer than 7 pixels of detected clouds
14	Additional shadow proximity over land	Clear-sky land pixels located closer than 7 pixels of detected cloud shadows
15	Same as code 1. Land	
16	Same as code 11. Additional cloud proximity over land	
17	Same as code 14. Additional shadow proximity over land	

A1.4. Reflectance Normalization

Reflectance normalization is a required step that allows extrapolation of the image characterization models in time and space by ensuring spectral similarity of the same land-cover types. Normalization addresses several factors that affect surface reflectance measurement from space, including scattering and attenuation of radiation passing through the atmosphere, and surface anisotropy. We implemented a relative normalization procedure that is not computationally intensive and does not require synchronously collected or historical data on atmospheric properties and land-cover specific anisotropy correction factors. The normalized surface reflectance is not equal to surface reflectance derived using atmospheric transfer models and a solution for the Bidirectional Reflectance Distribution Function (BRDF). The GLAD ARD data was designed for land cover and land cover change mapping and should not be used as a source dataset for the analysis of surface reflectance

properties. The Landsat image normalization consists of four steps: production of the normalization target dataset; selection of pseudo-invariant objects; model parametrization; and model application.

A1.4.1. Normalization Target

We derived the target surface reflectance data from twelve years (2000–2011) of MODIS/Terra 16-day surface reflectance time-series. The normalization target represents the growing season average spectral reflectance calculated as the average spectral reflectance for all MODIS clear-sky observations with NDVI above the 75th percentile. The normalization data were re-scaled to match the Landsat TOA reflectance data (to the range from 1 to 40,000) and resampled to the Landsat spatial resolution.

Pseudo-Invariant Objects

The mask of pseudo-invariant objects is derived for each Landsat image automatically and used to calibrate the per-scene surface reflectance normalization model. The mask includes clear-sky land observations (pixels) that represent the same land cover type and phenology stage in the Landsat image and MODIS normalization target composite. Water and snow/ice observations are excluded from the mask due to different properties of surface anisotropy.

A1.4.2. Model Parametrization

To parametrize the reflectance normalization model, we calculate the bias between Landsat TOA reflectance and MODIS surface reflectance for each spectral band within the mask of pseudo-invariant objects. We collect per-band median bias for each 10 km interval of distance from the Landsat ground track. The set of median values is used to parametrize a per-band linear regression model using the least squares fitting method. For each image and each spectral band, we derive gain (G) and bias (B) coefficients to predict the reflectance bias as a function of the distance from the ground track.

A1.4.3. Model Application

After the gain and bias coefficients are derived for each spectral band, we apply the resulting models to the entire Landsat image. To apply the model, we use the raster layer of distances from the ground track (in meters) that is calculated for each WRS from the Landsat orbital parameters. The normalized surface reflectance is calculated per-pixel.

A1.5. Temporal Integration and Tiling

The 16-day composites are stored in geographic coordinates and organized in the form of 1×1-degree tiles (see Section 3). To create a 16-day composite, we first select all Landsat images within the date range that overlap a selected 1×1 degrees tile. All selected images are projected to geographic coordinates using the nearest neighbor resampling method to preserve reflectance values. If more than one image overlaps the composite area, we analyze the QF layers of these images. For each pixel with overlapping images, we select the best observations with the lowest probability of cloud and shadow contamination.

Appendix 2. Multitemporal Metrics Methodology

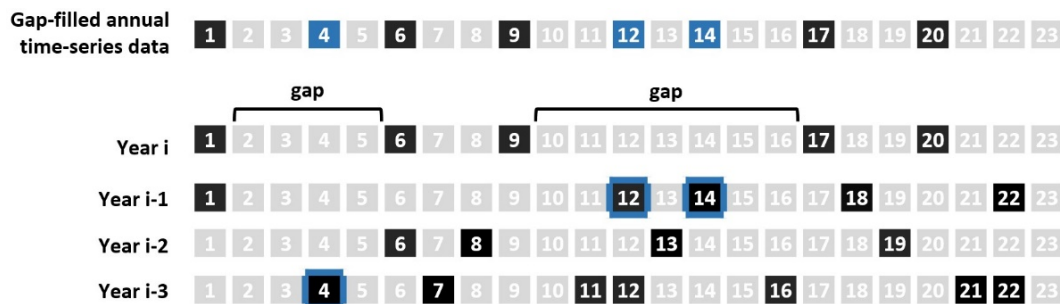
A2.1. Phenological Metrics

The phenological metric set is designed to allow annual land cover and vegetation structure mapping models extrapolation in space and time. This metric set is generated primarily using the observations collected during the target calendar year (January 1 – December 31). The data from the previous years may be used to fill gaps in the observation time-series. The gap-filling improves the year-to-year metric set consistency and is recommended for regions with frequent cloud cover. The process of phenological metrics construction includes two stages: (1) selecting clear-sky observations and filling gaps in the observation time series; and (2) extracting reflectance distribution statistics from the selected observation time-series.

A2.1.1. Data Selection and Gap-filling

The first stage of metric processing is compiling a time-series of annual observations with lowest atmospheric contamination. The per-pixel criterion for 16-day data selection is defined automatically based on the distribution of quality flags within the available data (the interval length depends on the gap-filling option and includes target and preceding years). If clear-sky land or water observations are present in the time-series data, only those are used for subsequent analysis. If no such observations are found, the code successively changes the quality threshold for data inclusion, first allowing observations with proximity to clouds and shadows, then allowing all available observations.

To create an annual gap-filled observation time-series for metric extraction, the code analyzes the duration of the gaps between existing 16-day observations of the current year (Year *i*). If a gap exceeded two months (four 16-day intervals), it will search for the clear-sky observations in the previous years within the gap date range, starting with Year *i*-1 and until the Year *i*-4. When clear-sky observations are found, they are added to the gap-filled time-series data, and the gap analysis is performed again until all gaps longer than two months are filled, or no available data are found within the four-year interval. The gap-filling for the pheno_C metric set uses adaptive rules that limit inclusion of two to four years old data to fill short gaps. It also uses linear regression to fill the remaining missing values in the time-series.



Schematic representation of the gap-filling algorithm implemented for pheno_A metrics. Year *i* stands for the corresponding year, and Years *i*-1 – *i*-3 for preceding years. Black squares are clear-sky observations and gray squares are 16-day intervals with no data. The blue squares in the gap-filled time-series are clear-sky observations filled from the Years *i*-1 – *i*-3 (highlighted by blue outlines) within the data gaps exceeding 2 months (four 16-day intervals).

After compilation of the annual gap-filled observation time-series, the code calculates selected normalized band ratios, or indices $(\text{Band A} - \text{Band B}) / (\text{Band A} + \text{Band B})$ for each selected observation.

$$NR^{AB} = (\rho^A - \rho^B) / (\rho^A + \rho^B) \times 10,000 + 10,000$$

NR^{AB} – Normalized ratio between bands A and B; ρ^A , ρ^B – normalized surface reflectance of bands A and B

A spectral variability vegetation index (SVVI, Coulter et al., 2016) is calculated using the standard deviation of spectral reflectance values.

$$SVVI = \sigma(\rho^{\text{Blue}}, \rho^{\text{Green}}, \rho^{\text{Red}}, \rho^{\text{NIR}}, \rho^{\text{SWIR1}}, \rho^{\text{SWIR2}}) - \sigma(\rho^{\text{NIR}}, \rho^{\text{SWIR1}}, \rho^{\text{SWIR2}}) + 10,000$$

SVVI – Spectral variability vegetation index; σ – standard deviation.

ρ^{Blue} , etc. – normalized surface reflectance.

A2.1.2. Spectral Reflectance Distribution Statistics

Multi-temporal metrics are generated from the time-series of normalized reflectance and indices using two independent ranking approaches. First, all observations are ranked by each spectral band reflectance or index value individually. From obtained individual ranks, we select the highest/lowest, second to the highest/lowest values and values corresponding to the first, second, and third quartiles. In addition to individual observations, we calculate averages for all observations between selected ranks and amplitudes between selected metrics. Second, we distribute observation dates by corresponding ranks of vegetation indices and brightness temperature. For these distributions, we extract observation dates corresponding to highest/lowest, second to highest/lowest and first, second, and third quartiles of the ranked variable. Phenology metrics that reflect salient points of phenology cycle (start, peak, end of the season; growing season average and total) were based on the normalized difference vegetation index (NDVI) time-series. For the metric set, we recorded normalized surface reflectance of these observations and calculated averages and amplitudes for observations between selected ranks. The amplitudes are not written to the files but calculated on the fly by classification software. To incorporate spatial features, in addition to each spectral metric we calculated the focal average of the metric value within the 3×3-pixel kernel (these values are also calculated on the fly during image classification).

A2.1.3. Data Quality Metrics

Each phenological metric set includes several data quality metrics that are used for data quality analysis. Data quality metrics are not included in the classification process as model inputs; however, the land cover classification is not applied on pixels with no data (count==0). The metrics are stored in the 16-bit unsigned format. **TEC_count** metric provides the number of selected observations. The best quality observations are selected using the rules presented in section 5.2.1. Gap-filling increases the number of available observations. The value range 0 (no data) – 23 (data for each 16-day interval exists). **TEC_prcwater** shows percent of water detections of all selected observations. The value is recorded as Percent × 10, value range 0 to 1000. **TEC_pf** is a Processing Flag (PF) that reflects the data quality. The PF shows the type of the data QF (quality flags) that were included in the 16-day time series for metric production. See QF table (section Appendix 1) for explanation. During the metric generation, the algorithm selects the best quality data, and lowers the inclusion threshold in case no clear-sky data is available. The PF provides the information on selected data quality. We do not recommend using data with PF 8 (it mostly includes permanently cloudy pixels). See table below for PF value interpretation.

Data quality layer (processing flag, PF)

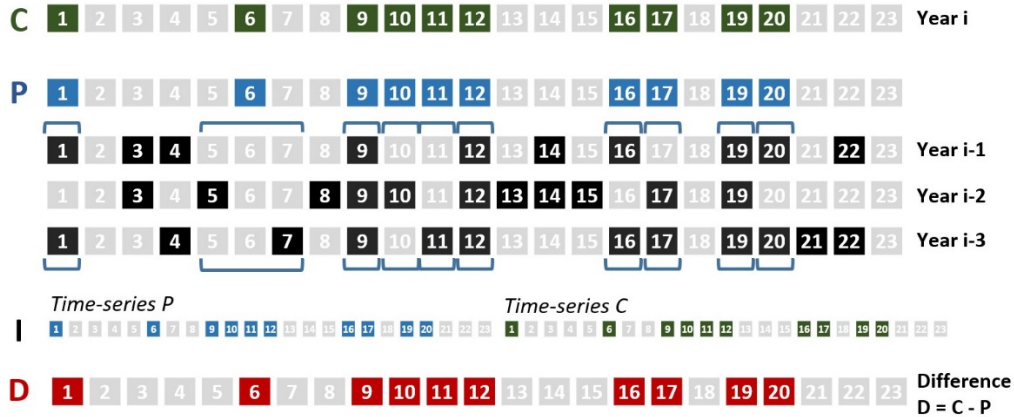
PF value	Data quality issues
0	No data
1	Selected data include only clear-sky land observations
2	Selected data include only clear-sky water observations
3	Selected data include clear-sky land and water observations
4	Selected data include clear-sky land and water, and observations affected by topographic shadows and moderate cloud shadow probability. This code is typically found in wetland and on shadow slopes.
5	Selected data have small number of clear-sky observations and include observations contaminated by topographic shadows, moderate probability cloud shadows, and snow/ice. Usually indicate mixed pixels in permanently cloudy regions and can be found within specific land cover (highlands, wetlands, urban areas) where QF model fails.
6	No clear-sky observations. Selected observations include data contaminated by clouds (proximity), shadows, and haze.
7	Selected data include only snow/ice observations
8	No clear-sky observations. Data consist entirely of cloud/cloud shadow contaminated observations.

A2.2. Change Detection Metrics Methodology

The annual change detection metrics are designed to facilitate land cover change mapping between the corresponding and previous years while reducing false change detections due to reflectance fluctuations and inconsistent clear-sky observations availability. Change detection metrics describe the surface reflectance within the corresponding and preceding years, spectral reflectance differences between these years, and surface reflectance trend within the time-series. The process of change detection metrics construction includes two stages: (1) selecting clear-sky observations and constructing data time-series, and (2) extracting reflectance and reflectance change distribution statistics from the time-series.

A2.2.1. Data Selection and Gap-filling

To build a set of change detection metrics, we utilize four years of data (one corresponding and three preceding) and select observations with the best available quality. The metric set can be generated with less than four years of data, but at least two consecutive years of data are required. Only observations with the lowest atmospheric contamination are used for metrics extraction. The per-pixel criterion for 16-day data selection is defined automatically based on the distribution of observation quality flags within the four years of data, similar to the phenological metrics algorithm. All other observations are discarded from further processing.



Schematic representation of the time-series data compilation for the change detection metrics. Green and black squares represent 16-day intervals with clear-sky observations, gray squares – 16-day intervals with no clear-sky observations. C stands for the corresponding year time-series (Year *i*); P for preceding year time-series (average of Years *i-1*, *i-2*, and *i-3*, selected observations highlighted in blue). Time-series I is compiled from time-series P and C. D stands for difference between 16-day observations of C and P time-series (intervals with difference values highlighted in red).

To facilitate extraction of the change detection data, we construct four different data time-series (time-series C, P, I, and D, see Figure 12). Time-series C comprised from the clear-sky observations of the corresponding year (Year *i*). To create a historical baseline for change detection (time-series P), we collect an average reflectance from the three preceding years (Year *i-1* – Year *i-3*) only for those 16-day intervals that have clear-sky observations in the time-series C. If no observations are found for a certain 16-day interval in historic data, we use clear-sky data from the closest observation before/after the missing 16-day composite interval. For each observation of time-series C and P, in addition to normalized reflectance, we calculate normalized ratios from selected bands. Time-series P and C are further aggregated into a single, 46-interval, time-series to calculate trend analysis metrics (time-series I). Finally, the per-16-day interval difference for all spectral band and index values between time-series P and C comprise the time-series D.

A2.2.2. Spectral Reflectance Distribution Statistics

To extract statistics, we use three different approaches:

- For the time-series C and P, we extract two independent sets of metrics that reflect annual phenology. Observations in each time-series are ranked by (a) spectral band or index value, and (b) corresponding NDVI and brightness temperature values. Similar to phenological metrics, we record selected ranks and average between ranks for each spectral variable.
- The time-series I is used to analyze unidirectional trend of spectral reflectance within a two-years interval. We use the least squares method to fit linear regression model that predicts spectral reflectance or index value from the observation date (date range is from 1 to 46) for clear-sky observations. We record the slope of linear regression as a metric value. In addition, we calculate and record standard deviation of spectral variable within the time-series I.
- The time-series D consists of per-16-day interval spectral reflectance or index differences. We rank different values and extract a set of statistics (selected ranks and averages) from these rankings.

A2.2.3. Change Detection Data Quality Metrics

Each change detection metric set includes several data quality metrics that are used for data quality analysis. The metrics are stored in the 16-bit unsigned format. Data quality metrics are not included in the classification process as model inputs.

The “code” metric is the most important data quality flag for change detection. It showed if the data were available for both current and preceding years. Code 0 indicates that the current year has no data. Code 1 indicates that the preceding three years used as the reference have no data. The change detection only applied on pixels with data in both current and preceding years (code==2).

Metric	
count	The number of selected observations for the current year. The best quality observations are selected using the rules presented in section 5.4.1. The value range 0 (no data) – 23 (data for each 16-day interval exists)
code	Data availability for change detection: 0 – no current year data available 1 – no reference (preceding years) data available 2 – both current and preceding year data available
prcwater	Percent of water detections of selected observations of the current year. The value is recorded as Percent × 10, value range 0 to 1000.
prcland	Percent of land detections of selected observations of the current year. The value is recorded as Percent × 10, value range 0 to 1000.
pf	The data quality layer (called processing flag, or PF) shows the type of the data QF (quality flags) that were included in the 16-day time series for metric production. See QF table (section 4) for explanation. The PF values table provided in section 5.3.3.

Appendix 3. Decision Tree Model

A3.1. Principles of the Decision Tree Model

A decision tree is a hierarchical classifier that predicts class membership by recursively partitioning a data set into more homogeneous subsets. This splitting procedure is followed until a perfect tree (one in which every pixel is discriminated from pixels of other classes, if possible) is created with all pure terminal nodes or until preset conditions are met for terminating the tree's growth. Our approach employs a deviance measure to split data into nodes that are more homogeneous with respect to class membership than the parent node. The reduction in deviance (D), is defined as:

$$D = D_s - D_t - D_u$$

where s is the parent node, and t and u are the splits from s .

Right and left splits along the digital counts for all data are examined. When D is maximized, the best split has been found, and the data are divided at that digital count and the process repeated on the two new nodes of the tree. The deviance for nodes is calculated from the following:

$$D_i = -2 \sum n_{ik} \log p_{ik}$$

where n is the number of pixels in class k in node i and p is the probability distribution of class k in node i .

If a decision tree model is implemented without constraints, it will attempt to create pure nodes (the nodes that consist of training pixels of the same class). However, due to noise and errors in training data, such a complex tree may produce incorrect results. To avoid over-fitting of a tree model, we implement pruning based on deviance decrease parameter (*mindev*). By increasing the deviance decrease parameter we may reduce the complexity of a tree which will produce a more generalized result.



For the first classification, we recommend using small *mindev* parameter value of 0.0001. The version of the tree code allows node numbers up to 2^{32} . If a child node number is above this limit, the tree growth is automatically terminated and the error message "maximum depth reached" will appear. The tree will be created, and it can be used for classification. However, the message indicates that the *mindev* parameter should be increased.

A3.2. Bootstrap Aggregation (Bagging)

To improve the model performance in case of errors in training data and noise in independent variables (metrics), we are implementing bagging (Bootstrap Aggregation) technique. The essence of bagging is to generate an ensemble of decision trees created using independent subsets of data from a training sample chosen randomly with replacement. The output class likelihood is calculated as the median output of all trees in the ensemble. A user may adjust the number of bagged trees (*maxtrees*). The number should not be too high (using more than 25 trees has a negligible effect on model performance). The number of trees should be odd (to simplify median calculation).

A3.3. Decision Tree Model Structure

The decision tree models are stored in the “trees” folder of the classification workspace.

The decision tree file structure

----- Input file: D:/xx_GLAD_ARD_WEB/04_classification/sample1.txt Rows: 7963 Columns: 255 Tree type: Classification tree mincut: 1 minsize: 2 mindev: 0.000100 Output type of y: label and probabilities Number of classes: 2 Number of all nodes: 369 Number of terminal nodes: 185 Totally 122 Variables used in model building: X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12 X13 X15 ... Overall misclassification rate on training data: 0.000000	Tree header Rows – number of sampled training pixels Columns – Number of metrics Classes IDs: 1 – background, 2 - target List of all metrics used to build the model. For metric names, refer to metrics_list_pheno_<type>.txt in C:/GLAD_1.0 Misclassification rate. 0 indicates a perfect tree (all training pixels were separated).
node), split, n, deviance, yval, (1 2) * denotes terminal node 1) root 7963 11036.273815 1 (0.509356 0.490644) 2) X133 < 1637.5 4206 3621.415894 2 (0.154541 0.845459) 4) X114 < 5264 345 24.589972 1 (0.994203 0.005797) 8) X173 < 1926.5 337 -0.000000 1 (1.000000 0.000000) * 9) X173 >= 1926.5 8 8.997362 1 (0.750000 0.250000) 18) X8 < 306 2 -0.000000 2 (0.000000 1.000000) * 19) X8 >= 306 6 -0.000000 1 (1.000000 0.000000) * 5) X114 >= 5264 3861 2143.461403 2 (0.079513 0.920487) 10) X102 < 1945.5 2843 377.376031 2 (0.012311 0.987689) 20) X141 < 1207.5 2009 45.036198 2 (0.001493 0.998507)	Tree model. First the root node (all training pixels), then child nodes and terminal nodes (marked with *). Parameters for each node: Node number Metric used to produce the split Threshold value Number of pixels after the split Deviance after the split Assigned class Target classes probability

A3.4. Metric Importance

The GLAD Tools include a tool to analyze all decision trees in the ensemble and to estimate the importance of each input metric in the model. The output file **tree_report.txt** includes the analysis of metric importance. The metric importance is calculated as the total deviance reduction from all nodes that uses a particular metric for a split. The deviance decrease is summarized for all decision tree models in the ensemble. The “percent_decrease_of_root” shows the percent of total deviance decrease for each metric of the root deviance. The higher the value, the higher the importance of a metric to separate the classes.

tree_report.txt

Internal Metric ID	Metric name	Mean deviance decrease for all trees	Percent root deviance decrease	
metric_ID	metric_name	deviance(decrease)	percent_decrease_of_root	Header
root		3561.1433106	NA	Root deviance
X2	blue_min	29.90149532	0.839659983101383	Total deviance decrease for each metric
X3	blue_max	17.81644588	0.500301288829575	
X4	blue_avmin25	7.59105876	0.213163529179089	

Licenses and Redistribution

The GLAD Tools and Landsat ARD data are available with no charges and no restrictions on subsequent redistribution or use, as long as the proper citation is provided as specified by the Creative Commons Attribution License (CC BY). The toolbox includes libraries and codes that were shared by other open-source software projects:

- MinGW - C++ compiler, GNU C Library (Open-source software; Copyright © Free Software Foundation)
- gdal - GDAL Core (Open-source software; Copyright © Frank Warmerdam and others)
- tree.exe – CART model (Open-source software; Copyright © B. D. Ripley and J. Ju)
- Other utilities – GLAD ARD Tools (Freeware; Copyright © GLAD UMD)

Copyright © Global Land Analysis and Discovery Team, University of Maryland

Suggested citation: Potapov, P., Hansen, M.C., Kommareddy, I., Kommareddy, A., Turubanova, S., Pickens, A., Adusei, B., Tyukavina A., and Ying, Q., 2020. Landsat analysis ready data for global land cover and land cover change mapping. *Remote Sens.* 2020, 12, 426; doi:10.3390/rs12030426 <https://www.mdpi.com/2072-4292/12/3/426>

GLAD Tools consists of compiled packages for use as described in the Tools manual. The GLAD team does not make any warranty, express or implied, including the warranties of merchantability and fitness for a particular purpose; nor assumes any legal liability or responsibility for the applications of the GLAD Tools.

GLAD Tools depend on several open-source third-party packages (PERL, curl, wget, OSGeo4W, R). The GLAD team is not responsible for supporting these packages. Users should refer to the licenses of the individual packages for redistribution policies.

References

Cochran W.G. (1977) *Sampling Techniques*. New York: John Wiley & Sons.

Olofsson P., Foody G.M., Stehman S.V., Woodcock C.E. (2013) Making better use of accuracy data in land change studies: Estimating accuracy and area and quantifying uncertainty using stratified estimation. *Remote Sensing of Environment* 129, 122-131.

Olofsson P., Foody G.M., Herold M., Stehman S.V., Woodcock C.E., Wulder, M.A. (2014) Good practices for estimating area and assessing accuracy of land change. *Remote Sensing of Environment*, 148, 42-57.

Olofsson P., Arévalo P., Espejo A.B., Green C., Lindquist E., McRoberts R.E., Sanz, M.J. (2020) Mitigating the effects of omission errors on area and area change estimates. *Remote Sensing of Environment*, 236, p.111492.

Potapov P., Hansen M.C., Kommareddy I., Kommareddy A., Turubanova S., Pickens A., Adusei B., Tyukavina A., Ying Q. (2020) Landsat analysis ready data for global land cover and land cover change mapping. *Remote Sensing* 12, 426; doi:10.3390/rs12030426

Stehman S.V. (2013) Estimating area from an accuracy assessment error matrix. *Remote Sensing of Environment* 132, 202-211.

Stehman, S. V. (2014) Estimating area and map accuracy for stratified random sampling when the strata are different from the map classes. *International Journal of Remote Sensing* 35, 13, 4923-4939