# Copernicus Global Land Operations "Vegetation and Energy" "CGLOPS-1" <br> Framework Service Contract № 199494 (JRC) 

## ALGORITHM THEORETICAL BASIS DOCUMENT

## MODERATE DYNAMIC LAND COVER <br> 100 м <br> Version (Collection) 3

## Issue 13.4

Organization name of lead contractor for this deliverable: VITO
Book Captain: Dr. Marcel Buchhorn (VITO)
Contributing Authors: Luc Bertels (VITO)
Bruno Smets (VITO)
Bert De Roo (VITO)
Myroslava Lesiv (IIASA)
Nandin-Erdene Tsendbazar (WUR)
Dainius Masiliunas (WUR)
Lin Linlin (WUR)

Please cite this publication the following:
Buchhorn, M.; Bertels, L.; Smets, B.; De Roo, B.; Lesiv, M.; Tsendbazar, N.E., Masiliunas, D., Linlin, L. (2020): Copernicus Global Land Service: Land Cover 100m: Version 3 Globe 2015-2019: Algorithm Theoretical Basis Document; Zenodo, Geneve, Switzerland, September 2020; doi:10.5281/zenodo. 3938968.

| Dissemination Level |  | X |
| :--- | :--- | :---: |
| PU | Public |  |
| PP | Restricted to other programme participants (including the Commission Services) |  |
| RE | Restricted to a group specified by the consortium (including the Commission Services) |  |
| CO | Confidential, only for members of the consortium (including the Commission Services) |  |

## Document Release Sheet

| Book captain: | Dr. Marcel Buchhorn |  | Date 31.03.2021 |
| :---: | :---: | :---: | :---: |
| Approval: | Roselyne Lacaze | Sign | Date 03.09.2020 |
| Endorsement: | Michael Cherlet | Sign | Date |
| Distribution: | Public |  |  |

## Change Record

| Issue/Rev | Date | Page(s) | Description of Change | Release |
| :---: | :---: | :---: | :--- | :---: |
|  | 25.09 .2017 | All | Initial Version | 11.00 |
|  | 09.10 .2017 | All | Revision after internal review | 11.10 |
| 11.10 | 26.08 .2019 | All | Version 2 of algorithm | 12.00 |
|  | 14.11 .2019 | All | Revision after internal review | 12.01 |
| 12.01 | 17.07 .2020 | All | Version 3 of algorithm | 13.2 |
| 13.2 | 03.09 .2020 | All | Revision after internal review | 13.3 |
|  | 31.03 .2021 | All | Revision after external review | 13.4 |

## Table Of Contents

Executive Summary ..... 19
1 Background of the document ..... 20
1.1 Scope and Objectives ..... 20
1.2 Content of the document ..... 20
1.3 Related documents ..... 20
1.3.1 Applicable documents ..... 20
1.3.2 Input ..... 20
1.3.3 Output ..... 21
1.3.4 External documents ..... 21
2 Review of Users Requirements ..... 22
3 Methodology Description ..... 26
3.1 Overview ..... 26
3.2 Input Data ..... 31
3.2.1 PROBA-V L1C archive (2013-2020) ..... 31
3.2.2 Land/sea mask datasets ..... 32
3.2.3 Sentinel-2 tiling grid ..... 33
3.2.4 ECMWF CAMS NRT datasets ..... 34
3.2.5 IMERG Precipitation dataset. ..... 34
3.2.6 Glacier and Ice Sheets datasets ..... 34
3.2.7 Global Mountain Explorer dataset ..... 35
3.2.8 Global ecological region datasets ..... 35
3.2.9 Circum-Polar Arctic Vegetation Map dataset ..... 36
3.2.10 Urban datasets ..... 36
3.2.11 JRC's Global Surface Water (GSW) dataset ..... 37
3.2.12 Digital Elevation Model datasets ..... 38
3.3 PROBA-V UTM Analysis Ready Data (ARD) Generation ..... 39
3.3.1 Overview ..... 39
3.3.2 Yearly Bright Surface and Permanent Snow detection ..... 40
3.3.3 Level 2A processing and Level 2A post-processing ..... 44
3.3.4 Level 2B processing ..... 48
3.3.5 Level 3 processing and final status mask ..... 51
3.3.6 Comparison PROBA-V versus PROBA-V UTM ..... 53
3.4 Epoch-wise Data Cleaning \& Compositing ..... 54
3.4.1 Overview ..... 54
3.4.2 The HANTS and madHANTS algorithms ..... 55
3.4.3 Median Composites (MC) generation ..... 57
3.4.4 Long-term 5-day Median Composite (HMC5) generation ..... 59
3.5 Epoch-Wise Input Data Quality Indicator Generation. ..... 60
3.5.1 Overview ..... 60
3.5.2 The PROBA-V theoretical observation mask and time series gap mask ..... 62
3.5.3 Calculating the time series statistics ..... 63
3.6 Epoch-Wise Data Fusion ..... 68
3.6.1 Overview ..... 68
3.6.2 Data Fusion pre-processing ..... 69
3.6.3 Principle of the Kalman-Filtering algorithm ..... 71
3.6.4 Data Fusion implementation using VITO's enhanced Kalman-Filtering approach ..... 72
3.6.5 Data Fusion post-processing. ..... 77
3.7 Ancillary Dataset Products Generation ..... 78
3.7.1 Global Biome Clusters product ..... 79
3.7.2 Yearly Built-up cover fraction product generation ..... 81
3.7.3 Yearly WetProducts generation ..... 82
3.8 Epoch-Wise Metrics Generation ..... 83
3.8.1 Overview ..... 83
3.8.2 Pre-Processing ..... 85
3.8.3 Metrics extraction ..... 90
3.8.4 Post-Processing ..... 93
3.9 Training Data Generation for BaseYears ..... 94
3.10 Epoch-Wise Classification / Regression ..... 96
3.10.1 Overview. ..... 96
3.10.2 The Random Forest Approach ..... 97
3.10.3 Training data and classifier/regressor optimization for BaseYears. ..... 98
3.10.4 Epoch-Wise Scenario-based Classification using the BaseYearModels ..... 100
3.10.5 Epoch-Wise Scenario-based Regression using the BaseYearModels ..... 101
3.11 Epoch-Wise Pre-Products Generation ..... 102
3.11.1 Overview. ..... 102
3.11.2 Epoch-wise cover fraction pre-products generation ..... 104
3.12 Concept of Temporal Cleaning in the CGLS-LC100 product workflow ..... 107
3.12.1 Overview. ..... 107
3.12.2 Break Detection using BFAST. ..... 110
3.12.3 Collection of change training data to optimize BFAST parameters ..... 112
3.12.4 Implementation of BFAST in the CGLS-LC100 workflow ..... 113
3.12.5 HMM as binary change detector. ..... 120
3.13 Annual CGLS-LC100 Cover Fraction Products Generation ..... 120
3.13.1 Overview. ..... 120
3.13.2 Annual CGLS-LC100 cover fraction products generation ..... 121
3.14 Annual CGLS-LC100 Land Cover Map Generation ..... 124
3.14.1 Overview. ..... 124
3.14.2 Epoch-wise discrete land cover pre-map generation ..... 125
3.14.3 Annual CGLS-LC100 Land cover maps generation ..... 136
3.15 The Annual CGLS-LC100 Products ..... 137
4 Limitations ..... 142
5 Future development of the product. ..... 145
6 References ..... 146

## List of Figures

Figure 1: Workflow diagram for the annual CGLS Dynamic Land Cover layers at 100 m resolution (CGLS-
LC100 products) generation. ................................................................................................................ 26
$\qquad$
Figure 2: Workflow diagram for Part A of the CGLS-LC100 products - generation of the epoch-wise classification and cover fraction layers pre-products. Note: dotted lines indicate extra processing steps when new training data is available and model updates have to be carried out.

28
Figure 3: Workflow diagram for Part B of the CGLS-LC100 products - temporal post-processing to generate consistent annual land cover maps and cover fraction layers with the corresponding Quality layers. ... 30

Figure 4: PROBA-V L1C processing chain flowchart (adapted from [PROBA-V User Manual]) .................... 31
Figure 5: The USGS shoreline vector layer (Sayre et al., 2013). Zoom in to the Strait of Gibraltar showing the shoreline vector layer (red) overlay to a Google Earth image (left) and to the PROBA-V land/sea mask (right). Note that the PROBA-V land/sea mask has a buffer around the land masses in order to map land/sea transitions. 32

Figure 6: Global shoreline layer generated out of the GSHHG and USGS datasets (Wessel et al., 1996; Sayre et al., 2013)

Figure 7: Sentinel-2 tiling grid overlaid on the global shoreline layer. Green rectangles show $6 \times 8$ deg UTM Grid Zone Designators (GZD's) overlaid on single 100,000-meter grid squares within the GZD's which form the single tiles.33

Figure 8: Global Mountainous Area Mask (MAM) showing mountain areas in white. Note: for better visualization the global shoreline layer is overlaid in red. ............................................................................................ 35
Figure 9: Circum-Polar Arctic Bioclimate Subzones. Image derived from http://www.ArcticAtlas.org ............ 36
Figure 10: The DLR World Settlement Footprint layer (Marconcini et al., 2017b). Zoom in to Strait of Gibraltar showing the urban areas in black. Note: in red the shoreline vector layer is shown. 37

Figure 11: The JRC Global Surface Water product (Pekel et al., 2016). Zoom in to the Strait of Gibraltar showing the water seasonality mask for 2015.

38
Figure 12: The Copernicus COP-DEM-GLO-90 dataset. Zoom in to the Strait of Gibraltar. Note: in red the USGS shoreline vector layer is shown. 38

Figure 13: Workflow chart showing the generation of the PROBA-V UTM Analysis Ready Data (ARD) archive out of the PROBA-V L1C archive. 39

Figure 14: Yearly Bright Surface and Permanent Snow detection - Part A: Raw mask generation using the PROBA-V L2A archive. 41

Figure 15: Distribution of global test sites used in the PROBA-V UTM ARD generation. Test sites following the Sentinel-2 naming convention. 42

Figure 16: Yearly Bright Surface and Permanent Snow detection - Part B: Mask cleaning and final Sentinel-2 tile generation.43

Figure 17: Examples for detected bright surfaces and permanent snow for year 2015 during the PROBA-V UTM ARD pre-processing. Upper left: bright desert sand in tile 11SNA; upper right: bright roof tops in tile 31UES; lower left: permanent snow (yellow) and bright rock (red) in tile 32TNS; lower right: bright roof top in tile 31UES.
Figure 18: Sentinel-2 tiling grid overlaid on PROBA-V imaging area. Shown are the 18,595 tiles covering the land masses which form the PROBA-V UTM ARD archive ..... 45
Figure 19: PROBA-V UTM L2A post-processing - Part A: improved snow and cloud detection - raw detection by thresholding. ..... 46
Figure 20: PROBA-V UTM L2A post-processing - Part B: improved snow and cloud detection - cleaning by masks and resolving confusions ..... 47
Figure 21: PROBA-V UTM L2A post-processing - Part C: updating of status mask ..... 47
Figure 22: PROBA-V UTM L2B processor - improved SMAC routine and quality indicator generation for atmospheric correction. ..... 48
Figure 23: Comparison of used ancillary data used in the improved and original PROBA-V L2B processor. Left:ECMWF CAMS NRT data with 6-hour temporal resolution; right: long-term monthly data used in originalL2B processor. Note: visualizations are shown for October 15, 201450
Figure 24: PROBA-V UTM L3 processing - generation of 1-day synthesis. From left to right: partly covered tilein first and second overpass, 1-day synthesis, and Google Earth image for comparison.52
Figure 25: Comparison between PROBA-V and PROBA-V UTM cloud/snow detection ..... 53
Figure 26: Comparison between PROBA-V (left) and PROAB-V UTM (right) geometric correction (L2A).Already in mid-latitudes a reduced distortion is viewable. Area shown: province Niedergoersdorf, Teltow-Flaeming, Germany.54
Figure 27: General overview of the data cleaning and compositing section in the CGLS-LC100 productworkflow. Note: Numbers in the upper left corner of data container indicate the number of layers - herenumber of observations - for one epoch used to create the yearly land cover map55

Figure 28: Example for the madHANTS temporal outlier detection algorithm for the epoch2015 (reference year 2015 +/- 1 year). Note: blue stars mark all original pixel values in the time series (invalid observation values are set to -1 ), the red line shows the harmonized time series by applying the amplitudes and phases of the identified frequencies during the HANTS transformation, red squares show the detected outliers which are more than $\sim 3.5$ standard deviations away from the harmonized pixel value57

Figure 29: Example showing the temporal outlier detection using the madHANTS algorithm on the blue and SWIR reflectance values of a pixel at location $9.459^{\circ}$ lon, $6.562^{\circ}$ lat. Note: red curve shows the harmonized time series; blue shows the valid pixel reflectance values; green shows flagged outliers cumulative detected in the blue or SWIR reflectance band58

Figure 30: Example for status mask and temporal data cleaning for part of a PROBA-V UTM TOC-S1 100 m tile (observation taken 2016-03-06). Image is shown as false colour composite (RGB = SWIR, NIR, blue) for a sample area in tile in Nigeria: raw image (left), status masked cleaned image (status mask flagged areas are shown in red - left), madHANTS cleaned image (additional pixels which are flagged as outliers or clouds are shown in blue - right).59

Figure 31: Exemplary temporal profiles for PROBA-V UTM ARD+ time series with a high Data Density Indicator (DDI) (upper graph) and low DDI (lower graph) showing less available valid data. Note: x-axis shows the 5-day time step starting from the epoch. Values different than zero on the y-axis shows that data is available on the corresponding time step61
Figure 32: Data Density Indicator processing workflow. ..... 61


#### Abstract

Figure 33: Calculating the theoretical maximum number of observations (Theo. Obs. mask) and the longest concurrent gap in data time series data (T-GAP) is the first step in calculating the DDI characteristics. Calculating the Theo. Obs. Mask requires three parameters: a threshold on the SZA as used by SMAC, the satellite overpass time and the camera viewing angles


Figure 34: The raw time series statistics calculated on the PROBA-V UTM ARD+ data. After the raw time
series statistics are calculated the PROBA-V UTM ARD+ data is checked for sensor specific observation
limitations and updated in this case by imprinting the Theo. Obs. Mask. ..... 64
Figure 35: The cleaned time series statistics - Part A ..... 65
Figure 36: The cleaned time series statistics - Part B ..... 66
Figure 37: Final scoring of time series statistics to compute the Data Density Indicator (DDI) and output of the PROBA-V UTM ARD+ statistics. ..... 67
Figure 38: Visual RGB representation of UTM tile 32NPM with some of the calculated DDI statistics for the 100$m$ resolution for epoch2015, i.e. Data Density Indicator, Observation ratio, Observation frequency, Gapweight and Longest concurrent gap67
Figure 39: Average cloud cover over a 13 year period from July 2002 to April 2015 captured by the MODISAqua RS sensor. Note: Colours range from dark blue (no clouds) to light blue (some clouds) to white(frequent clouds). Image credits: NASA Earth Observatory, 201968
Figure 40: Number of cloud free data observations for PROBA-V UTM 100 m (left) and PROBA-V UTM 300 m (right) in the reference year 2015 for Africa ..... 68
Figure 41: General overview of the data fusion section in the CGLS-LC100 product workflow. Note: Numbers in the upper left corner of data container indicate the number of layers - here number of 5-day observations - for one epoch used to create the yearly land cover map ..... 69
Figure 42: Example of data fusion pre-processing results for PROBA-V UTM TOC-MC5 300 m time series atpixel location $9.459^{\circ}$ Ion, $6.562^{\circ}$ lat. top) continuous gap length in the 300 m time series (dashed greenline indicates threshold for small gaps, dashed red line indicates threshold for big gaps), middle) timeseries before pre-processing (blue line shows original TOC-MC5 time series, green line shows originalTOC-MC10 time series, red line shown the long term harmonized time series for the full PROBA-Varchive), bottom) TOC-MC5 300m time series after pre-processing (blue line shows the original TOC-MC5 time series, orange line shows the final pre-processed time series which will be used for the datafusion).70
Figure 43. Basic concept of the Kalman filtering approach. Adapted from Aimonen (2011). ..... 71
Figure 44: Principle of VITO's enhanced Kalman-filtering approach for EO data fusion based on the pktools algorithm. Note: Image adapted from Kempeneers et al. (2016). ..... 72
Figure 45: Example of 100 m time series reconstruction of one pixel using VITO's improved Kalman filter.Upper image shows original 100 m timeseries in orange (orange line shows existing values and orangeshading shows missing time steps), the corresponding 300 m time series in blue and the final 100 m gapfilling in green plus uncertainty in light green areas; second image shows the whole reconstructed 100 mtime series for forward (blue), backward (orange) and smooth mode (green); third image shows the 100 mforward reconstruction with uncertainty; and fourth image shows the 100 m backward reconstruction withuncertainty.74


#### Abstract

Figure 46: Visualization of data fusion with the improved Kalman filter for one $3 \times 3100 \mathrm{~m}$ pixel block in PROBAV UTM ARD+ datasets. Note: red rectangles show the zoom into selected area of 100 m and corresponding 300 m tile, and the selected 300 m pixel for selected time series step (Note: 100 m has no valid data); blue rectangles show forward (top) and backward (center) unmixing of 300 m time step into 100 m resolution for zoom and pixel, respectively; magenta rectangles show the 100 m reconstructed time series steps out of 300 m pixel for zoom and $3 \times 3$ pixel area, respectively (Note: shown is forward (top), backward (center) and smooth (bottom) mode of Kalman filter)

76


Figure 47: top: full PROBA-V UTM TOC-MC5 100 m tile for red reflectance channel (areas with missing data are shown in white); bottom-left: PROBA-V UTM TOC-MC5 100 m data fused image; bottom-right: zoom in to full PROBA-V UTM 100 m resolution of left image over red box shown in top image (the area in the red box had no observations at all for that time stamp and shows, after data fusion, consistent image data). Example of PROBA-V UTM tile in Nigeria on 2016-03-06.

78

Figure 48: Global stratification into 73 biome clusters to improve regional classification/regression in the CGLS
LC100 workflow ..... 80

Figure 49: General overview of the yearly WetProducts generation. Besides the PROBA-V UTM ARD++ derived VI data products, i.e. HUE, VALUE and NDVI, two ancillary layers are used as input for the WetProducts generation. Note: Numbers in the upper left corner of data container indicate the number of layers - here number of 5-day observations - for one reference year used to create the yearly product.

Figure 50: General overview of the metrics generation section for each epoch in the CGLS-LC100 product workflow. Note: datasets marked in red are re-used in following sub-steps. Numbers in the upper left corner of data containers indicate the number of layers (time steps in the ARD++ where 2165 -day time steps represent the three year epoch or product layers in the resulting metrics).

84
Figure 51: The HSV colour space and the formulas for transforming the RGB colour space into the HSV colour space. Note: V=Value, S= Saturation, H=Hue, R=red, G=green, B=blue............................................ 88

Figure 52: Spectral Area Indices explained. The sum of the FWHM reflectance (reflectance over the full band width) for all reflectance bands can be used as single scalar to compare spectral signatures

89
Figure 53: Example for a harmonized time series profile. The result of the HANTS model is shown in red dots, where the blue dots show the real observations

91
Figure 54: Examples of the 183 derived metrics for a tile in Africa for the epoch2015. a) texture metric (the lighter the colour the more homogeneous is the pixel compared to its surrounding pixels), b) standard deviation of the Structure Intensive Pigment Index for the vegetation season (the greener the colour the higher is the SIPI change within the epoch), c) sum of the Enhanced Vegetation Index for the epoch (the redder the colour the more vegetation) 94
Figure 55: Training dataset for the BaseYear2015 used in the CGLS-LC100 workflow showing the $\sim 168 \mathrm{~K}$
training locations together with their sampled land cover class at 100 m ................................. 96 training locations together with their sampled land cover class at 100 m .

96
Figure 56: The Random Forest classifier principle. Image by Niraj (2016)................................................. 98
Figure 57: Rule matrix calculated for the total of $N$ training points which belong to four different LC classes. For each training point the RMSE value is calculated against all the other training points, i.e. c11 with c11, c12, c13 till c4n, subsequently c12 with c11, c12, c13 till c4n and so on
Figure 58: General overview of the epoch-wise pre-products generation section in the CGLS-LC100 productworkflow.103
Figure 59: Rule block 1 to 4 for the epoch-wise cover fraction pre-products generation - generation of snow, built-up, permanent and seasonal water cover fraction and cleaning by CL1 and CAVM. ..... 105
Figure 60: Rule block 5 and 6 for the epoch-wise cover fraction pre-products generation - cover fraction normalization and cleaning by the land/sea mask ..... 107
Figure 61: Processing mode of land cover change maps. Note: the rectangles represent the land cover class of a certain pixel in the map. Near-real time mapping doesn't allow the direct identification of land cover changes for certain classes which is indicated by red pattern. ..... 108
Figure 62: Example of MODIS time series of NIRv over an area north of San Pablo Bay, California (longitude:-122.454, latitude: 38.12883 ) which is now an outlet of the bay. MODIS time series spans over a decadeof imagery with a single sensor and allows identifying stable and unstable conditions, in this caseagriculture before 2016 and water after109
Figure 63: BFAST family fundamentals: a time series is subdivided into several segments and a linear (in thisexample harmonic 3rd order) model is fit onto each segment to detect a break. The magnitude of thebreak (change in model parameters between segments) is also reported.110
Figure 64: Overview of Change reference data locations to calibrate BFAST change detection. ..... 113
Figure 65: Workflow for the calibration of BFASTON and BFAST Monitor ..... 114
Figure 66: The construction site of the Chengdu Tianfu International Airport (Sichuan province, China;longitude: 104.463, latitude: 30.326). Left: Sentinel-2 imagery before (summer 2016) and after (summer2017) the start of the construction, right: an aerial photograph in later stages of construction.116
Figure 67: Time series of the area of the Tianfu airport. The break is correctly identified by BFASTON, and the reference time of land cover change is less than one year away from the detected break ..... 116
Figure 68: The time series of the Tianfu airport with break detection using BFAST Monitor. Each panel showsone year for which break predictions are made, in sequence (2016-2017, 2017-2018, and 2018 -2019). The break is correctly identified, but another spurious break is identified one year later, as themodel does not yet have enough data to adjust to the new land cover state. Note: blue line indicates timestep of detected break, red line time step of real break, dotted line indicates start year of monitoring period.117
Figure 69: True color Sentinel-2 imagery of forest in the Philippines (longitude: 122.177, latitude: 7.412) beingreplaced by a tree plantation.118
Figure 70: Detection of breaks for the Phillipines deforestation case using BFASTON. Land cover change in2016 is detected correctly and more precisely than the reference data indicates, but the change in 2018is missed as it does not include a break.118
Figure 71: Detection of breaks for the Phillipines deforestation case using BFAST Monitor. Each panel showsone year for which break predictions are made, in sequence (2016-2017, 2017-2018, and 2018 -2019). The result is similar to BFASTON, but the detected break is indicated later than in BFASTON. Note:blue line indicates time step of detected break, red line time step of real break, dotted line indicates startyear of monitoring period119
Figure 72: Smoothing of cover fractions over the years 2015-2019. The original cover fractions (square) are smoothened by setting to the value obtained by applying a linear regression
Figure 73: General overview of the epoch-wise discrete Land Cover pre-map generation section in the CGLS- LC100 product workflow ..... 125
Figure 74: Flowchart of the classification post-processing to produce additional input layers for the land cover pre-product generation. ..... 126
Figure 75: Initialization and Expert Rule block 1 for the discrete pre-map and map probability layer generation - imprinting hard masks ..... 129
Figure 76: Expert Rule block 2 to 4 for the discrete pre-map and map probability layer generation - imprintingthe derived cover fraction layers for classes, resolving classification scenarios, and filling of the remainingland pixels130
Figure 77: Expert Rule block 5 part A for the discrete pre-map and map probability layer generation - resolvingconfusion between closed forest, open forest, shrub and grassland on continental level. *See Figure 48and Table 6 for biome cluster location and naming133
Figure 78: Expert Rule block 5 part B for the discrete pre-map and map probability layer generation - resolving confusion between closed forest, open forest, shrub and grassland on continental level ..... 134
Figure 79: Expert Rule block 6 and 7 for the discrete pre-map and map probability layer generation - wetland class improvements and final separation of the open/closed forest class by forest type ..... 135
Figure 80: The CGLS-LC100 discrete map product for epoch 2015 with 23 discrete classes (detailed legend in Figure 82). ..... 138
Figure 81: The cover fraction layers for the 9 base land cover classes and the seasonal inland water cover fraction of the CGLS-LC100 product for epoch 2015 (shown as a collage on global scale) ..... 138
Figure 82: The three level legend of the CGLS-LC100 discrete map product. The raster code together with the RGB colour code and discrete class name is shown ..... 139

## List of Tables

# Table 1: Usefulness of information on LC and LC change processes for different international actions and programmes. <br> 23 

Table 2: List of land cover classes requested by users ..... 23
Table 3: Spectral characteristics of the PROBA-V bands showing the four bands, i.e. BLUE, RED, Near Infra- Red (NIR) and Short Wave Infra-Red (SWIR) ..... 31
Table 4: Status mask bit mapping of the PROBA-V L2A data ..... 44
Table 5: PROBA-V UTM L3 status mask. New flags marked in bold ..... 52
Table 6: Global biome cluster by continents ..... 81
Table 7: A schematic view on rule set 1 which is applied on stable and reliable cover fractions. The cover fractions will only be adapted in case of the triggering of rule 03,04 or 05. ..... 122
Table 8: A schematic view on rule set 2 which is applied on unstable cover fractions. Note that in most cases when a break has been detected no adaptation to the cover fraction has been done. ..... 123
Table 9: The 23 discrete classes of the CGLS-LC100 discrete product. Note: final classes are shown in bold together with numerical code in brackets. ..... 127
Table 10: Overview of existing CGLS-LC100 products produces by algorithm version ..... 140

Issue: I3.4

## List of Acronyms

| Acronym | Meaning |
| :--- | :--- |
| AD | Applicable document |
| aNIR | Angle at NIR |
| AOD | Aerosol Optical Depth |
| AOT | Aerosol Optical Thickness |
| API | Application Programming Interface |
| ARD | Analysis Ready Data |
| ATBD | Algorithm Theoretical Basis Document |
| AUC | Area Under the Curve |
| AVHRR | Advanced Very-High-Resolution Radiometer |
| BFAST | Breaks For Additive Season and Trend |
| CAMS | Copernicus Atmosphere Monitoring Service |
| CAVM | Circum-Polar Arctic Vegetation Map |
| CCI | Climate Change Initiative |
| CF | Climate \& Forecast conventions |
| CGLS | Copernicus Global Land service |
| CGLS-LC100 | Dynamic Land Cover layers at 100 m resolution |
| CLC | Corine Land Cover |
| CORRA | Combined Ku Radar-Radiometer Algorithm |
| DDI | Data Density Indicator |
| DEM | Digital Elevation Model |
| DLR | Deutschen Zentrums für Luft- und Raumfahrt |
| DOI | Digital Object Identifier |
| DSM | Digital Surface Model |
| ECMWF | European Centre for Medium-Range Weather Forecasts |
| EEA | European Environmental Agency |
| EO | Earth Observation |
| EPSG | European Petroleum Survey Group Geodesy |
| ESA | European Space Agency |
| EUMETSAT | European Organisation for the Exploitation of Meteorological Satellites |
| EVI | Enhanced Vegetation Index |
| FAO | Food and agriculture organization of the united nation |
| FAPAR | Fraction of Absorbed Photosynthetically Active Radiation |
| FFT | Fast Fourier Transformation |
| FWHM | Full Width at Half Maximum |
| GEMS | Geostationary Environmental Monitoring System |
| GIMP | Greenland Ice Mapping Project |
| GIO | GMES/Copernicus initial operations land |
| GIS | Geographic Information System |
| GLIMS | Global Land Ice Measurements from Space |
| GOME | CGLOPS Monitoring Experiment |
| Document-No. | Issue: I3.4 |

Issue: I3.4

| GPCP | Global Precipitation Climatology Project |
| :---: | :---: |
| GPROF | Goddard Profiling Algorithm |
| GPU | Graphical Processor Unit |
| GSD | Ground Sampling Distance |
| GSHHG | Global Self-consistent, Hierarchical, High-resolution Geography Database |
| GSW | Global Surface Water |
| GTCO3 | GEMS total column ozone |
| GUF | Global Urban Footprint |
| GZD | Grid Zone Designators |
| HANTS | Harmonic Analysis of Time Series |
| HMC5 | Harmonized 5-daily Median Composite |
| HMM | Hidden Markow Model |
| HSV | Hue Saturation Value colour system |
| HUE | Chromacity |
| IIASA | International Institute for Applied Systems Analysis |
| IASI | Infrared Atmospheric Sounding Interferometer |
| IMERG | Integrated Multi-satellitE Retrievals for GPM |
| JRC | Joint Research Center |
| LAI | Leaf Area Index |
| KML | Keyhole Markup Language |
| L1C | Level-1 C product (unprojected data, Top-of-Atmosphere reflectance) |
| LC | Land Cover |
| LC100 | Land Cover map at 100m resolution |
| LCCS | Land Cover Classification System |
| MAD | Median Absolute Deviation |
| madHANTS | Median Absolute Deviations of HANTS |
| MAM | Mountainous Area Mask |
| MC10 | 10-daily median composite |
| MC5 | 5-daily median composite |
| ME | Moving Estimate |
| MMU | Minimum Mapping Unit |
| MODIS | Moderate Resolution Imaging Spectroradiometer |
| NALCMS | North American Land Change Monitoring System |
| NASA | National Aeronautics and Space Adminstration |
| NAUC | Normalized Area Under the Curve |
| NLCD | National Land Cover Database |
| NDMI | Normalized Difference Moisture Index |
| NDVI | Normalized Difference Vegetation Index |
| NIR | Near Infra-Red |
| NIRv | Near-Infrared reflectance of vegetation |
| NRT | Near-Real-Time |
| OECD | Organisation for Economic Co-operation and Development |


| Document-No. | CGLOPS1_ATBD_LC100m-V3 | © C-GLOPS1 Lot1 consortium |
| :--- | :--- | ---: |
| Issue: 13.4 | Date: 31.03 .2021 | Page: 16 of 152 |


| OLS-MOSUM | Ordinary Least Squares - moving window sum |
| :--- | :--- |
| OSM | OpenStreetMap |
| PGISM | Permanent Glacier and Ice Sheet Mask |
| PMAp | Polar Multi-Sensor Aerosol Properties Product |
| PMW | Passive microwave |
| PPSIM | Potential Permanent Snow and Ice mask |
| PROBA-V | Vegetation instrument on board of PROBA satellite |
| PSD | Product Specification Document |
| PUM | Product User Manual |
| REED+ | Reducing Emissions from Deforestation and Forest Degradation and the role |
|  | of conservation, sustainable management of forests and enhancement of |
|  | forest carbon stocks in developing countries |
| RF | Random Forest classifier |
| RGB | Red-Green-Blue colour space |
| RMSE | Root Mean Square Error |
| RS | Remote Sensing |
| RSS | Residual sum of Squares |
| S1 | 1-daily synthesis |
| SAA | Sun Azimuth Angle |
| SDG | Sustainable Development Goals |
| SEEA/FAO | System of Environmental-Economic Accounting / Food and Agriculture |
|  | Organization |
| SIPI | Structure Intensive Pigment Index |
| SG | Satellite-Gauge |
| SM | Status Map |
| SMAC | Simplified Model for Atmospheric Correction |
| SQL | Structured Query Language |
| SSD | Service Specifications Document |
| SVP | Service Validation Plan |
| SWIR | Short Wave Infra-Red |
| SZA | Sun Zenith Angle |
| tcwv | Total column water vapor |
| t-GAP | timeseries gap |
| Theo. Obs. mask | Theoretical maximum number of Observations masks |
| TOA | Top Of Atmosphere |
| TOC | Top Of Canopy |
| TPI | Topographic Position Index |
| TPR | Temporal Regression Post-Processing |
| UN | United Nations |
| UNCCD | United Nations Convention to Combat Desertification |
| UNFCCC | Unirements Document |
| URD |  |
|  |  |
|  |  |


| Document-No. | CGLOPS1_ATBD_LC100m-V3 | © C-GLOPS1 Lot1 consortium |
| :--- | :--- | ---: |
| Issue: 13.4 | Date: 31.03 .2021 | Page: $\mathbf{1 7}$ of $\mathbf{1 5 2}$ |


| US | United States |
| :--- | :--- |
| USGS | United States Geological Survey |
| UTM | Universal Transverse Mercator |
| VAA | Viewing Azimuth Angle |
| VI | Vegetation Index |
| VITO | Flemish Institute for Technological Research |
| VR | Validation Report |
| VRT | Virtual Raster Translator |
| VZA | Viewing Zenith Angle |
| WGS84 | World Geodetic System 1984 |
| WLS | Weighted least square |
| WUR | Wageningen University \& Research |
| WSF | Word Settlement Footprint |
| XML | Extensible Markup Language |

## Executive Summary

The Copernicus Global Land Service (CGLS) is earmarked as a component of the Land service to operate "a multi-purpose service component" that provides a series of bio-geophysical products on the status and evolution of land surface at global scale. Production and delivery of the parameters take place in a timely manner and are complemented by the constitution of long-term time series.

From $1^{\text {st }}$ January 2013, the Copernicus Global Land Service is providing a series of bio-geophysical products describing the status and evolution of land surface at global scale, i.e Leaf Area Index (LAI), the Fraction of Absorbed Photosynthetically Active Radiation (FAPAR), the Land Surface Temperature, the soil moisture, and other vegetation indices on an hourly to 10 days interval from Earth Observation satellite data.

The CGLS was complemented through the provision of an annual dynamic global Land Cover product at 100 m spatial resolution (CGLS-LC100). Land cover plays a major role in the climate and biogeochemistry of the Earth system. The CGLS Land Cover product provides a primary land cover scheme at three classification levels with class definitions according to the Land Cover Classification System (LCCS) scheme. Next to these discrete classes, the product also includes continuous field layers or "fraction maps" for all basic land cover classes that provide proportional estimates for vegetation/ground cover for the land cover types. This continuous classification scheme may depict areas of heterogeneous land cover better than the standard classification scheme and, as such, can be tailored for application use (e.g. forest monitoring, rangeland management, crop monitoring, biodiversity and conservation, monitoring environment and security, climate modelling, etc.)

The first Land Cover map, produced with algorithm V1.0, was provided in July 2017, derived from the Vegetation instrument on board of PROBA satellite (PROBA-V) 100 m time-series, a database of high quality land cover training sites and several ancillary datasets. This collection 1 was solely covering the African continent and the mapping year 2015 with an overall accuracy of $75 \%$. For the global up-scaling of the product for the reference year 2015, the algorithm V2.0 was improved not only to map the entire globe, but also to improve the map quality, reaching now $80 \%$ of accuracy at class level 1 on each continent, and its usability. Therefore, the production of the map has switched to the Universal Transverse Mercator (UTM) coordinate system aligned to the Sentinel-2 tiling grid in order to improve mapping quality in the high latitudes. Moreover, the Sentinel-2 tiling grid facilitates continuation in the production as soon the primary data input has to switch from PROBA-V to Sentinel-2 after June 2020. Collection 2 was delivered in May 2019. Algorithm version 3 adds the change detection methodology to the CGLS workflow. We put particular emphasis on the temporal consistency of the annual land cover maps. Specific algorithms were applied to increase the stability of the annual classifications with the effect that differences between the annual maps resulting from inconsistent classifications are reduced while areas of probable land cover change are considered as different classes in the various annual maps. Therefore, collection 3 of this product provides the annual global land cover maps and cover fraction layers for the reference years 2015 to 2019.

This Algorithm Theoretical Basis Document (ATBD) describes the methods used for classifying and temporal cleaning the $3^{\text {rd }}$ collection of annual land cover layers at 100 m resolution at global scale.

| Document-No. | CGLOPS1_ATBD_LC100m-V3 | © C-GLOPS1 Lot1 consortium |
| :--- | :--- | ---: |
| Issue: 13.4 | Date: 31.03 .2021 | Page: $\mathbf{1 9}$ of $\mathbf{1 5 2}$ |

## 1 BACKGROUND OF THE DOCUMENT

### 1.1 Scope and Objectives

The scope of this document is to describe the theoretical basis and justification that underpins the implementation of the dynamic land cover product at 100 m resolution Version 3 (collection 3) provided in the Copernicus Global Land Service. It details the methodology to be applied on the PROBA-V data, together with an evaluation of the algorithm performance and a description of the limitations.

A thorough validation of the products is completed by a full validation exercise [CGLOPS1_VR_LC100m-V3.0] according to the Service Validation Plan [CGLOPS1_SVP].

### 1.2 Content of the document

This document is structured as follows:

- Chapter 2 recalls the users requirements, and the expected performance
- Chapter 3 describes the retrieval methodology
- Chapter 4 identifies the limitations
- Chapter 5 outlines some risks of failure and their mitigation


### 1.3 RelAted documents

### 1.3.1 Applicable documents

AD1: Annex I - Technical Specifications JRC/IPR/2015/H.5/0026/OC to Contract Notice 2015/S 151277962 of $7^{\text {th }}$ August 2015

AD2: Appendix 1 - Copernicus Global land Component Product and Service Detailed Technical requirements to Technical Annex to Contract Notice 2015/S 151-277962 of $7^{\text {th }}$ August 2015

AD3: GIO Copernicus Global Land - Technical User Group - Service Specification and Product Requirements Proposal - SPB-GIO-3017-TUG-SS-004 - Issue I1.0-26 May 2015.

### 1.3.2 Input

## Document ID

CGLOPS1_SSD

CGLOPS1_SVP

## Descriptor

Service Specifications Document of the Copernicus Global Land Service.

Service Validation Plan of the Copernicus Global Land Service.

Issue: I3.4

CGLOPS1_URD_LC100m

CGLOPS1_PSD_LC100m

CGLOPS1_TrainingDataReport_LC100 m

### 1.3.3 Output

## Document ID

CGLOPS1_VR_LC100m-V3.0

CGLOPS1_PUM_LC100m-V3.0

User Requirements Document dynamic land cover 100m product

Product Specifications Document of the dynamic land cover 100 m product.

Report describing the training dataset used for Dynamic Land Cover 100 m product

## Descriptor

Validation Report describing the results of the scientific quality assessment of the dynamic moderate land cover 100 m version 3 product

Product User Manual Document of the dynamic moderate land cover 100m version 3 product

### 1.3.4 External documents

| PROBA-V | http://proba-v.vgt.vito.be/ |
| :--- | :--- |
| PROBA-V User Manual | User Guide of the PROBA-V data, available on <br>  <br>  <br> eodata.be/PDF/image/PROBAV-Products User Manual.pdf |

## 2 Review of Users Requirements

According to the applicable document [AD2] and [AD3], the users requirements relevant for Dynamic Moderate Land Cover are:

- Definition: Dynamic global land cover products at 300 m and/or 100 m resolution using (United Nations) UN Land Cover Classification System (LCCS)
- Geometric properties:
- Pixel size of output data shall be defined on a per-product basis so as to facilitate the multi-parameter analysis and exploitation.
- The baseline datasets pixel size shall be provided, depending on the final product, at resolutions of 100 m and/or 300 m and $/$ or 1 km .
- The target baseline location accuracy shall be $1 / 3$ of the at-nadir instantaneous field of view.
- Geographical coverage:
- geographic projection: lat long
- geodetical datum: World Geodetic System 1984 (WGS84)
- pixel size: minimum $1 / 112^{\circ}$ - accuracy: min 10 digits
- coordinate position: pixel centre for netCDF and upper left corner for GeoTIFF
- minimum global window coordinates:
- Upper Left: $180^{\circ} \mathrm{W}-75^{\circ} \mathrm{N}$
- Bottom Right: $180^{\circ} \mathrm{E}, 56^{\circ} \mathrm{S}$
- Accuracy requirements: Overall thematic accuracy of dynamic land cover mapping products shall be $>80 \%$. The overall accuracy assessment (including confidence limits) will be based on a stratified random sampling design and the minimum number of sampling points per land cover class relevant to the product shall be calculated as described in Wagner and Stehman, 2015.

A few workshops were held in 2016 to consult different stakeholders to understand users' needs for global land cover maps. A feasibility study was performed to define the guidelines to create the first CGLS-LC100 map. More details can be found in [CGLOPS1_URD_LC100m]. Larger consultations in 2017 and 2018 allowed collecting the requirements of wide user communities which were translated in product specifications [CGLOPS1_PSD_LC100m].

Table 1 summarizes the usefulness of information on LC and LC change processes for different international actions and programmes. Table 2 includes the LC classes that are required by different JRC units and are marked by (" $X$ "), respectively. The last column provides the following information: either a class is included in the CGLS-LC100 collection 3 product's legend, or it can be derived by users from the fraction layers, or additional research is needed.

Table 1: Usefulness of information on LC and LC change processes for different international actions and programmes.

|  | LC types | Related <br> change <br> processes | land | UNFCCC | UNCCD | OECD | SEEAFAO |
| :---: | :--- | :--- | :---: | :---: | :---: | :---: | :---: | SDGs

Table 2: List of land cover classes requested by users.

|  | $\begin{aligned} & \text { N } \\ & \hline 0 \\ & \hline 0 \\ & 0 \\ & 0 \\ & 0 \\ & \hline \end{aligned}$ |  |  |  | 6u!̣ои!uou do. כ | 7 <br> $\frac{2}{4}$ <br> $\frac{3}{0}$ <br> $\frac{2}{0}$ <br> 0 |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 10 |  | A12A3A20B2 | Forest/ree cover | X |  | X | X | X | Yes |  |
|  | 11 | $\begin{aligned} & \text { A12A3A20B2 } \\ & \text { D2E1 } \end{aligned}$ | Evergreen Needleleaf forest | X |  |  | X | X | Yes |  |
|  | 12 | $\begin{aligned} & \text { A12A3A20B2 } \\ & \text { D1E1 } \end{aligned}$ | Evergreen Broadleaf forest | X |  |  | X | X | Yes |  |
|  | 13 | $\begin{aligned} & \text { A12A3A20B2 } \\ & \text { D2E2 } \end{aligned}$ | Deciduous <br> Needleleaf forest | X |  |  | X | X | Yes |  |
|  | 14 | $\begin{aligned} & \text { A12A3A20B2 } \\ & \text { D1E2 } \end{aligned}$ | Deciduous <br> Broadleaf forest | X |  |  | X | X | Yes |  |

Document-No.
Issue: I3.4

Date: 31.03.2021
© C-GLOPS1 Lot1 consortium
Page: 23 of 152

Issue: I3.4

|  | 15 | $\begin{aligned} & \text { A12A3A20B2 } \\ & \text { D1D2 } \end{aligned}$ | Mixed forest | X |  | X |  |  | Yes |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | 16 | $\begin{aligned} & \text { A12A3A10B2 } \\ & \text { XXXX } \\ & \text { (assuming that } \\ & \text { an intact forest } \\ & \text { is a very } \\ & \text { dense forest) } \end{aligned}$ | Intact forest | X |  | X |  | X | To map these classes, addition R\&D is required for methodology. Either we develop an expert rule based on other datasets such as: <br> http://www.intactforests.org/dat a.ifl.html or new training and validation datasets. These classes could potentially be included in the next product evolutions. <br> Yes <br> These classes could be potentially included in the next product evolutions. We will have to collect corresponding training and validation data. <br> Yes <br> To map these classes, addition R\&D is required for methodology. Also we will have to develop new training and validation datasets. These classes could potentially be included in the next product evolutions. <br> The product includes such a LC class as open forest, which is a mix of trees (more than 15\%), shrubs and grassland. This class only partly corresponds to savannas because a 100 m x100m pixel may include less trees but still be considered as savanna. <br> However, users are encouraged to use the fraction |  |  |
|  | 17 | - | Secondary forest | X |  | X |  | X |  |  |  |
|  | 18 | A11A1 | Managed forest | X |  | X |  | X |  |  |  |
|  |  | A11A1 | Plantation forest/tree crops | X | X | X |  | X |  |  |  |
|  |  | A11A1 | Oil palm plantation | X | X |  |  |  |  |  |  |
|  |  | - | Forest logging | X | X | X |  |  |  |  |  |
|  |  | A12A3 | Dominant tree species, e.g. spruce, pine, birch | X |  | X |  |  |  |  |  |
|  |  | A11A1(A2/A3) | Shifting cultivation system | X | X |  |  | X |  |  |  |
| 20 |  | $\begin{aligned} & \text { AA12A4A20B } \\ & 3 \text { (B9) } \end{aligned}$ | Shrub |  |  | X | X | X |  |  |  |
|  | 21 | $\begin{aligned} & \text { A12A4A20B(B } \\ & \text { 9)XXE1 } \end{aligned}$ | Evergreen shrubs |  |  | X |  |  |  |  |  |
|  | 22 | $\begin{aligned} & \text { A12A4A20B3( } \\ & \text { B9)XXE2 } \end{aligned}$ | Deciduous shrubs |  |  | X |  |  |  |  |  |
| 30 |  | $\begin{aligned} & \text { A12A2(A6)A2 } \\ & \text { 0B4 } \end{aligned}$ | Herbaceous vegetation |  |  | X | X | X |  |  |  |
|  |  | A12A6A10 // A11A1A11B4 XXXXXXF2F4 F7G4-F8 | Pasture/managed grassland |  |  |  |  | X |  |  |  |
|  |  | A122(A6)A10 | Natural grassland |  |  | X |  | X |  |  |  |
|  |  | A12A2 | Grass types for Western Africa |  |  | X |  |  |  |  |  |
|  |  | $\begin{aligned} & \text { A12A3A11B2 } \\ & \text { XXXXXXF2F4 } \\ & \text { F7G4-A12; } \\ & \text { A12A3A11B2- } \\ & \text { A13; } \\ & \text { A12A1A11 } \end{aligned}$ | Savannas |  |  | X |  |  |  |  |  |
| Document-No. CGLOPS1_ATBD_LC100m-V3 © C-GLOPS1 Lot1 consortium <br> Issue: 13.4 Date: 31.03 .2021 Page: $\mathbf{2 4}$ of $\mathbf{1 5 2}$ |  |  |  |  |  |  |  |  |  |  |  |

Issue: I3.4

|  |  |  |  |  |  |  |  |  | layers to produce their own savanna layer by applying specific thresholds for tree, shrub and grass cover. |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 40 |  | A11A3 | Cultivated and managed vegetation/agricult ure |  | X | X | X | X | Yes |
|  | 41 | $\begin{aligned} & \text { A11A3XXXXX } \\ & \text { XD3(D9) } \end{aligned}$ | Irrigated cropland |  | X |  |  | X | To map these classes, addition R\&D is required for methodology. Also we will have to develop new training and validation datasets. These classes could potentially be included in the next product evolutions. |
|  | 42 | $\begin{aligned} & \text { A11A3XXXXX } \\ & \text { XD1 } \end{aligned}$ | Rainfed cropland |  | X |  |  | X |  |
|  | 43 | A11A3 | Big and small farming/field size |  | X |  |  |  |  |
|  | 44 | A11A1-W8/A2 | Permanent crops |  | X |  |  | X |  |
|  | 45 | A11A3 | Row crops |  | X |  |  |  |  |
|  |  | A11A2 | Crop types: long/short cycle or winter/summer crops |  | X |  |  |  |  |
|  |  | A11A2 | Multiple crop cycles |  | X |  |  |  |  |
| 50 |  | B15A1 | Urban/built up |  |  | X | X | X | Yes |
| 60 |  | B16A1(A2) | Bare/sparse vegetation |  |  |  | X | X | Yes |
| 70 |  | B28A2(A3) | Snow and Ice |  |  |  | X | X | Yes |
| 80 |  | B28A1 | Open water |  |  |  | $X$ | X | Yes |
| 90 |  | $\begin{aligned} & \text { A24A1(A2/A3/ } \\ & \text { A4) } \end{aligned}$ | Wetland |  |  | X | X | X | Yes |
|  |  | A24A3 | Mangroves | X |  | X |  |  | To map this class, addition R\&D is required for methodology. Also we will have to develop new training and validation datasets. These classes could potentially be included in the next product evolutions. |

## 3 Methodology Description

### 3.1 Overview

The CGLS Dynamic Land Cover Layers at 100 m resolution (CGLS-LC100) product is generated by combining several proven individual methodologies through:

1. Data pre-processing including atmospheric \& geometric correction
2. Data cleaning by sensor specific status masks and (temporal) outlier detection techniques,
3. Applying data fusion techniques at multiple levels,
4. Supervised classification,
5. Including established third party datasets via expert rules, and
6. Temporal cleaning via break detection analysis of the landcover trends.

The workflow, shown in Figure 1, can be divided into two main parts: the generation of epoch-wise classification and cover fraction layers pre-products for the single reference years, and the temporal post-processing to generate consistent annual land cover maps and cover fraction layers. The temporal coverage of the annual products is one nominal year each (reference years) but includes three years of data as input for each map (epochs spanning +/- 1 year around the reference year).


Figure 1: Workflow diagram for the annual CGLS Dynamic Land Cover layers at 100 m resolution (CGLS-LC100 products) generation.

Figure 2 shows in detail the epoch-wise generation of the pre-products (classification results and cover fraction layers) for each reference year (upper part of Figure 1) which can be divided into the following sections:

1. PROBA-V UTM Analysis Ready Data (ARD) generation for full PROBA-V archive,
2. data cleaning \& compositing for full ARD archive,
3. yearly quality indicator generation for PROBA-V UTM ARD+ input data,
4. data fusion of PROBA-V UTM 100 m and 300 m ARD+ archive,
5. metrics generation/extraction per three year epoch around reference year,
6. yearly ancillary datasets products generation,
7. training data generation for BaseYear (if needed),
8. classification / regression using BaseYear models,
9. epoch-wise classification and cover fraction layers pre-products generation.

To reduce distortion in the High North, to make our land cover products better usable with other data sets and to allow continuity of the service, the PROBA-V archive used as current main input data source and normally distributed in the geographic coordinate system World Geodetic System 1984 (WGS84) - known as EPSG:4326 - was reprocessed with a new geometric correction and an improved atmospheric correction. The complete PROBA-V L1C archive (October 2013 - March 2020) (unprojected, Top-of-Atmosphere data) was in this way translated into the PROBA-V UTM Analysis Ready Data (ARD) archive which is projected to the Universal Transverse Mercator (UTM) coordinate system, fully aligned with the Sentinel-2 tiling grid in tiling naming as well as tile dimensions, and corrected to surface reflectance (Top-of-Canopy).

The PROBA-V UTM ARD main product, the PROBA-V multi-spectral daily-synthesis surface reflectance Top Of Canopy (TOC) data (L3 or also known as TOC-S1) with a Ground Sampling Distance (GSD) of 100 m is used as primary Earth Observation (EO) data, and PROBA-V UTM multispectral daily-synthesis surface reflectance data with a GSD of 300 m secondarily. Next to a Status Map (SM) cleaning using the internal quality flags of the PROBA-V UTM L3 data, a temporal cloud and outlier filter built on a Fourier transformation is applied to clean the data. Moreover, a median compositing is carried out in order to archive regular 5-day time steps in the 100 m and 300 m PROBA-V EO time series. This composing is needed since the PROBA-V satellite provides a daily global coverage for the 300 m L3 product corresponding to 5 -day for the 100 m L3 product. This cleaning and compositing step upgrades the PROBA-V UTM ARD archive into the PROBA-V UTM ARD+ archive. From this cleaned and outlier screened 5-day time series a Data Density Indicator (DDI) is calculated for each reference year which is used as an quality indicator for the used input data as well as weight factor in the supervised learning process of the model generation.

Next, to improve the data density in the 5-day 100 m time series, the 100 m and 300 m EO datasets are fused using a Kalman filtering approach. The Kalman-filled 100 m time series are then automatically checked for consistency. This Kalmann-filling upgrades the PROBA-V UTM ARD+ archive into the PROBA-V UTM ARD++ archive, which gives temporal cleaned, consistent and dense 5-day image stacks for all global land masses at 100 m resolution (Note: 300 m data is not existing in the ARD++). Following, several Vegetation Indices (VI's) are generated out of the PROBA-

V UTM ARD++ surface reflectance data for each time step in the archive. Since this data would be too much to be used in a supervised learning, the time dimension in the data stack has to be condensed. Again, for the generation of the land cover maps and cover fraction layers for each reference year, the data of three years (+/- 1 year around the reference year) is used. The three year period around the reference year is called "epoch" in this documentation. Therefore, several metrics explaining the time series of an epoch are extracted as input for the supervised classification and regression for each reference year. Thus for each epoch a harmonic model is fitted through the time steps of each of the derived vegetation indices. Next to the parameters of the harmonic model which are used as metrics for the overall level and seasonality of the time series of the corresponding epoch, descriptive statistics and textural metrics are generated. Overall, 183 metrics are extracted from the PROBA-V UTM ARD++ archive for each epoch (Note: to distinguish the epochs in the documentation they are labelled by their reference year e.g. epoch2015, epoch2016, ...).


Figure 2: Workflow diagram for Part A of the CGLS-LC100 products - generation of the epoch-wise classification and cover fraction layers pre-products. Note: dotted lines indicate extra processing steps when new training data is available and model updates have to be carried out.

To use external ancillary data in the classification/regression, the datasets have been checked for consistency and if needed re-warped to the UTM coordinate system, resampled to 100 m , retiled into the Sentinel-2 tiling grid, and post-processed to usable ancillary data products. Some of the used external ancillary data is valid for all reference years whereas some have to be generated for each reference year. The training data for the supervised learning is collected through manual
classification using Google Maps and Bing images at 10 m spatial resolution using the Geo-Wiki Engagement Platform (http://www.geo-wiki.org/). Therefore, the training data not only includes the land cover type, but also the cover fractions of the main land cover classes can be derived for 100 m resolution. The training data was initially collected for the reference year 2015 and will be updated every 5 years. Therefore the 5 -yearly intervals are called BaseYears (e.g. BaseYear2015 represents the training data / model for the reference year 2015 with a duration of 1 year). To optimize the CGLS-LC100 products to regional patterns, the classification/regression are handled eco-region specific (called biome clusters). In the classification and regression preparation, the metrics of each biome cluster of the training points of a BaseYear are analysed for inter-specific outliers in the pure endmembers, as well as screened for the best metrics combinations (best band selection) in order to reduce redundant information. The optimized training data together with the quality indicator of the input data (DDI dataset) for the corresponding epoch of the BaseYear are input in the training of the supervised classifier/regressor using Random Forest (RF) techniques. The generated biome cluster specific models are optimized via a 5 -folded cross-validation in order to estimate the optimal classifier/regressor parameters. These models are therefore only produced in 5 -year intervals for the BaseYears and used for the classification/regression of all reference years starting with the BaseYear up but not including the next BaseYear (e.g. the models of BaseYear2015 is used for the processing of the reference years 2015 - 2019). Finally, after applying the biome cluster specific classification/regression models and running a post-processing, epoch-wise pre-products are generated which include the epoch-wise classification results and epoch-wise cover fraction layers.

Since all epoch-wise pre-products are generated independently for each reference year, a second processing flow (lower part of Figure 1) is needed to generate consistent annual land cover maps and cover fraction layers over time (Figure 3). Therefore, the pre-products (classification and cover fraction layers) of every epoch are post-processed in several steps:

1. Generation of break-maps with BFAST algorithm;
2. Generation of secondary break-map with a Hidden Markov model;
3. Temporal post-processing of epoch-wise cover fraction layer pre-products;
4. Applying expert rules to generate annual land cover maps and applying temporal expert rules to make them consistent.

In the first pre-processing step, break maps are created based on a Breaks For Additive Season and Trend (BFAST) algorithm, additional training data to optimize the break detection and MODIS 300 m EO time series. The long-term MODIS RS archive was used for the break detection due to the fact that the PROBA-V archive spans only 6 years. These break maps show areas where change occurred between years. The BFAST family of break detection algorithms are unsupervised change detection methods that use a univariate VI time series with irregular timesteps as input. From this input, several components are computed: a monotonously increasing trend component, a nominal season component that indicates the season that each observation belongs to, and harmonic components that are sine and cosine functions for each harmonic order. The second pre-processing step consist of creating a break map based on a hidden Markov model (HMM). This model takes the cover fractions of all years and predict from these cover fractions the discrete map. The differences in the discrete maps between the years are then used as break map. BFAST was used as our main
predictor for breaks and HMM was used as ancillary data set to improve the confidence for all detected breaks. Next, in the temporal regression post-processing (TPR) all epoch-wise cover fraction pre-products are stacked together to obtain a cover fraction time series. These time series (one for each LC class) are then evaluated in terms of stability based on the deviation of the cover fractions from their respective means. Depending on the stability certain rules are applied in order to smooth the cover fraction depending on whether or not BFAST breaks has been detected. Applying these rules implies consistent cover fraction over all year unless BFAST detected some breaks.

In a second processing step, we build upon the success of previous global mapping efforts and/or other ancillary datasets which are included via expert rules in the annual land cover maps generation. Based on the classification pre-products of the single epochs and on the annual temporal cleaned cover fraction, expert rules are used to obtain annual land cover pre-maps. To improve our temporal stability, the change in the land cover maps is aligned with the prediction of BFAST and imprinted in the final annual land cover maps. Also a 3-level confidence mask is generated by implementing the HMM break maps.

The produced annual land cover maps use a hierarchical legend based on the United Nations (UN) and Land Cover Classification System (LCCS). Compatibility with existing global land cover products is hereby taken into account. Still a novelty of this product is the generation of continuous cover fields that allow proportional estimates of cover fractions of all main land cover classes for each year.


Figure 3: Workflow diagram for Part B of the CGLS-LC100 products - temporal post-processing to generate consistent annual land cover maps and cover fraction layers with the corresponding Quality layers.

### 3.2 INPUT DATA

Several input data sets are used in the CGLS-LC100 product workflow, i.e. PROBA-V L1C EO data, land/sea mask, ecological zones dataset, global urban and water masks, etc. Overall, these datasets can be grouped into 12 categories which serve as input for the different algorithm steps. This section list all used external datasets together with a short description. Information regarding needed postprocessing for the usage in the PROBA-V UTM ARD generation or the generation of the ancillary dataset products can be found in section 3.3 , and section 3.7 , respectively.

### 3.2.1 PROBA-V L1C archive (2013-2020)

Main EO input into the processing line is the PROBA-V level 1C (L1C) collection 1 archive (May 2013 to April 2020). PROBA-V multi-spectral L1C satellite data is radiometrically corrected to Top Of Atmosphere (TOA) reflectance and given unprojected per strip/camera (Francois et al., 2014).

A summary regarding the PROBA-V platform, the data format and L1C generation can be found in the PROBA-V product user manual [PROBA-V User Manual]. A more detailed description to the used level 1 processor can be found in Sterckx et al. (2014) and is summarized in Figure 4.


Figure 4: PROBA-V L1C processing chain flowchart (adapted from [PROBA-V User Manual])

Table 3: Spectral characteristics of the PROBA-V bands showing the four bands, i.e. BLUE, RED, Near Infra-Red (NIR) and Short Wave Infra-Red (SWIR).

| Spectral band | Wavelength | Central Wavelength | Full Width at Half <br> Maximum (FWHM) |
| :---: | :---: | :---: | :---: |
| BLUE | $0.447-0.493 \mu \mathrm{~m}$ | $0.460 \mu \mathrm{~m}$ | $0.042 \mu \mathrm{~m}$ |
| RED | $0.610-0.690 \mu \mathrm{~m}$ | $0.658 \mu \mathrm{~m}$ | $0.082 \mu \mathrm{~m}$ |
| NIR | $0.770-0.893 \mu \mathrm{~m}$ | $0.834 \mu \mathrm{~m}$ | $0.121 \mu \mathrm{~m}$ |
| SWIR | $1.570-1.650 \mu \mathrm{~m}$ | $1.610 \mu \mathrm{~m}$ | $0.089 \mu \mathrm{~m}$ |

The PROBA-V TOA reflectance is available for four spectral bands corresponding to the selected measurements (Table 3). The L1C data is stored in a szip compressed hdf5 file and delivered together with datasets for geometric viewing and illumination geometries, observation timestamps, and sensor-specific metadata.

### 3.2.2 Land/sea mask datasets

The land/sea mask is mainly used to distinguish between land and open sea water. Thus, used in the workflow for reducing the amounts of data to be processed.

In CGLS-LC100 algorithm version 1.0, we used the 30 m shoreline vector layer of the U.S. Geological Survey (USGS) which was produced from Landsat 7 EO data for the Africa Ecosystem Project (Sayre et al., 2013). We use the USGS shoreline layer instead of the internal PROBA-V land/sea mask since the PROBA-V land/sea mask consist of a buffer around the land masses in order to show the land/sea transition zones. Figure 5 shows an example of the USGS shoreline vector layer for the Strait of Gibraltar overlaid to a Google Earth image and the internal PROBA-V land/sea mask.


Figure 5: The USGS shoreline vector layer (Sayre et al., 2013). Zoom in to the Strait of Gibraltar showing the shoreline vector layer (red) overlay to a Google Earth image (left) and to the PROBA-V land/sea mask (right). Note that the PROBA-V land/sea mask has a buffer around the land masses in order to map land/sea transitions.

Anyhow, for the global up-scaling of the CGLS-LC100 product, the land/sea mask produced out of the Sayre et al. (2013) dataset is not sufficient since it covers only the African continent. Therefore, we used the Global Self-consistent, Hierarchical, High-resolution Geography (GSHHG) Database to generate a high resolution global land/sea mask (Wessel et al., 1996). Since the Sayre et al. (2013) dataset showed a better delineation of the shoreline than the GSHHG dataset, we fused the African continent shoreline of the Sayre et al. (2013) dataset into the GSHHG vector dataset (Figure 6).


Figure 6: Global shoreline layer generated out of the GSHHG and USGS datasets (Wessel et al., 1996; Sayre et al., 2013)

### 3.2.3 Sentinel-2 tiling grid

To improve the quality, usability and comparability of the CGLS-LC100 product, the production grid was switched from the geographic coordinate system WGS84 in algorithm V1.0 to UTM in version 2.0. Moreover, we decided to adapt to the tiling grid of the Sentinel-2 mission of the EU Copernicus programme. The Sentinel-2 tiling grid can be downloaded from the European Space Agency (ESA) webpage in the Sentinel-2 download section (ESA, 2018). The file is provided in Keyhole Markup Language (KML) format and can be easily transferred into a vector file. Figure 7 shows the derived vector file filtered to the global land masses using the shoreline layer described in section 3.2.2.


Figure 7: Sentinel-2 tiling grid overlaid on the global shoreline layer. Green rectangles show 6x8 deg UTM Grid Zone Designators (GZD's) overlaid on single 100,000-meter grid squares within the GZD's which form the single tiles.

### 3.2.4 ECMWF CAMS NRT datasets

To improve the PROBA-V atmospheric correction (level 2 processor) (see section 3.3), we implemented the usage of European Centre for Medium-Range Weather Forecasts (ECMWF) Copernicus Atmosphere Monitoring Service (CAMS) Near-Real-Time (NRT) data. The CAMS NRT data can be downloaded from the ECMWF data servers in batch via an Application Programming Interface (API) and is provided by the CAMS Information 2019.

Overall, we used 4 global data products downloaded in a $0.125 \times 0.125$ degree spatial resolution and with a 6-hour temporal resolution:

- GEMS total column ozone (gtco3) given in $\mathrm{kg} / \mathrm{m}^{2}$;
- Total column water vapor (tcwv) given in $\mathrm{kg} / \mathrm{m}^{2}$;
- Total aerosol optical depth at 550 nm (aod550);
- Surface temperature at 2 m given in Kelvin.


### 3.2.5 IMERG Precipitation dataset

The usage of precipitation datasets can help in the classification to distinguish land cover classes from each other or to recognize and adapt the classifier to seasonal effects i.e. dry years versus wet years. We use the Integrated Multi-satellitE Retrievals for GPM (IMERG) product which is the unified U.S. algorithm that provides the multi-satellite precipitation product for the U.S. GPM team (Huffman et al., 2019). The current version of the dataset is Version 06B.

The definition of the dataset is stated by Huffman et al. (2019): "The precipitation estimates from the various precipitation-relevant satellite passive microwave (PMW) sensors comprising the GPM constellation are computed using the 2017 version of the Goddard Profiling Algorithm (GPROF2017), then gridded, intercalibrated to the GPM Combined Ku Radar-Radiometer Algorithm (CORRA) product, and merged into half-hourly $0.1^{\circ} \times 0.1^{\circ}$ (roughly $10 \times 10 \mathrm{~km}$ ) fields. Note that CORRA is adjusted to the monthly Global Precipitation Climatology Project (GPCP) Satellite-Gauge (SG) product over high-latitude ocean and tropical land to correct known biases.".

### 3.2.6 Glacier and Ice Sheets datasets

To improve the permanent snow and ice detection during the L2B pre-processing of the PROBA-V data (see section 3.3), we generated a Permanent Glacier and Ice Sheet Mask (PGISM) out of the following datasets:

- The Global Land Ice Measurements from Space (GLIMS) Randolph Glacier Inventory 6.0, a global inventory of glacier outlines provided as vector dataset (Pfeffer et al., 2014);
- Greenland Ice Mapping Project (GIMP) land classification and surface elevation datasets, a regional high resolution mapping of the Greenland ice sheet (Howat et al., 2014).

Moreover we generated a Potential Permanent Snow and Ice mask (PPSIM) out of the following global and regional land cover classifications:

- ESA Climate Change Initiative (CCI) land cover map for year 2015 (CCI Land Cover, 2017);
- US National Land Cover Database 2011 (NLCD 2011) (Homer et al., 2015);
- North America classification based on MODIS data (NALCMS, 2010);
- Corine Land Cover (CLC) 2012 (European Environmental Agency, 2012).

For the PPSIM, the classified permanent snow and ice areas of the single input dataset were merged and buffered. Where the PGISM is used as an inclusion dataset, the PPSIM is used as an exclusion dataset - areas outside the PPSIM cannot be covered by permanent snow and ice.

### 3.2.7 Global Mountain Explorer dataset

To improve the snow detection during the L2B processing of the PROBA-V data (see section 3.3), we used the K1 dataset of the USGS Global Mountain Explorer (Kapos et al., 2000) to generate a global Mountainous Area Mask (MAM). The MAM together with surface temperature data is used to reset false detected snow in non-mountainous areas.


Figure 8: Global Mountainous Area Mask (MAM) showing mountain areas in white. Note: for better visualization the global shoreline layer is overlaid in red.

### 3.2.8 Global ecological region datasets

In order to group EO data for faster processing or adaptation of algorithms to specific regions, we combined several external ecological zone datasets into so-called biome clusters. The combination of the ecological zone datasets was done by expert knowledge in a Geographic Information System (GIS) environment and is described in more detail in section 3.7.1.

Main input into the biome cluster vector and raster file generation were the following datasets:

- Ecoregions RESOLVE2017 global dataset (Dinerstein et al., 2017);
- Global tree line layer (Alaska Geobotany Center, 2003);
- Geiger-Koeppen global ecozones after Olofsson update (Olofsson et al., 2012);
- Global land/sea mask (see section 3.2.2)
- Sentinel-2 tiling grid and PROBA-V imaging extent (see section 3.2.3, [PROBA-V User Manual]);
- Global Digital Elevation Model (DEM) (see section 3.2.12);
- Global ecological zones for Food and Agriculture Organization of the united nation (FAO) forest reporting - update 2010 (FAO, 2012).


### 3.2.9 Circum-Polar Arctic Vegetation Map dataset

To improve the forest classification in High Latitudes we used the Circum-Polar Arctic Vegetation Map (CAVM) (CAVM team, 2003). In detail, we used the Arctic Bioclimate Subzones Zone A to Zone E for generating a mask of forest limits in the High North (Figure 9).


Zone B
Zone C
Zone D
Zone E
Non Arctic
$\square$ Glaciers

Figure 9: Circum-Polar Arctic Bioclimate Subzones. Image derived from http://www.ArcticAtlas.org

### 3.2.10 Urban datasets

In order to generate an urban mask in the CGLS-LC100 workflow, DLR's Word Settlement Footprint (WSF) dataset for 2015 (Marconcini et al., 2017a, Marconcini et al., 2017b) is used. The WSF layer used mainly multi-temporal Sentinel-1 radar data in combination with multi-temporal Landsat-8 multispectral optical data to detect urban structures with a spatial resolution of 10 m . Figure 10 shows an example of the WSF 2015 layer for the Strait of Gibraltar.

| Document-No. | CGLOPS1_ATBD_LC100m-V3 | © C-GLOPS1 Lot1 consortium |
| :--- | :--- | ---: |
| Issue: 13.4 | Date: 31.03 .2021 | Page: $\mathbf{3 6}$ of $\mathbf{1 5 2}$ |



Figure 10: The DLR World Settlement Footprint layer (Marconcini et al., 2017b). Zoom in to Strait of Gibraltar showing the urban areas in black. Note: in red the shoreline vector layer is shown.

To provide updates for the WSF layer after 2015, we decided to integrate yearly OpenStreetMap data (OSM). The OSM data is retrieved from the database in yearly intervals (OpenStreetMap contributors, 2015), post-processed and then combined with the WSF dataset. More details regarding additional processing can be found in section 3.7.2.

### 3.2.11 JRC's Global Surface Water (GSW) dataset

In order to generate a water product in the CGLS-LC100 workflow, JRC's Global Surface Water (GSW) product is used (Pekel et al., 2016). The GSW uses the 30 m Landsat 5, 7 and 8 EO data archive to generate a mask showing the water bodies visible from space, including natural and artificial water. Therefore, the thermal properties and spectral properties of water and other features in the Landsat spectral bands were used to separate pixels acquired over open water from those acquired over other surfaces (Pekel et al., 2016). Within the CGLS-LC100 workflow, we use the water seasonality layers. The water seasonality layer provides information regarding the intra-annual distribution of surface water (number of months when the pixel is detected as water). Figure 11 shows an example of the water seasonality 2015 layer for the Strait of Gibraltar. More details regarding additional processing can be found in section 3.7.3.


Figure 11: The JRC Global Surface Water product (Pekel et al., 2016). Zoom in to the Strait of Gibraltar showing the water seasonality mask for 2015.

### 3.2.12 Digital Elevation Model datasets

To include topographic metrics (elevation, slope, aspect) in the CGLS-LC100 product workflow, the Copernicus COP-DEM-GLO-90 dataset is used as primary input. The Copernicus DEM is a Digital Surface Model (DSM) which represents the surface of the Earth including buildings, infrastructure and vegetation. This DEM is derived from an edited DSM named WorldDEMTM and provided as a CORE product of the Copernicus service (Copernicus, 2019) for the reference year 2015. We are using the global product with a $\sim 90 \mathrm{~m}$ spatial resolution. Figure 12 shows an example of the COP-DEM-GLO-90 dataset for the Strait of Gibraltar.


Figure 12: The Copernicus COP-DEM-GLO-90 dataset. Zoom in to the Strait of Gibraltar. Note: in red the USGS shoreline vector layer is shown.

| Document-No. | CGLOPS1_ATBD_LC100m-V3 | © C-GLOPS1 Lot1 consortium |
| :--- | :--- | ---: |
| Issue: 13.4 | Date: 31.03 .2021 | Page: $\mathbf{3 8}$ of $\mathbf{1 5 2}$ |

### 3.3 PROBA-V UTM Analysis Ready Data (ARD) Generation

### 3.3.1 Overview

To improve the quality, usability and comparability of the CGLS-LC100 product, the complete PROBA-V L1C archive (unprojected, Top-of-Atmosphere data) (section 3.2.1) was translated into the PROBA-V UTM Analysis Ready Data (ARD) archive. Figure 13 shows the processing workflow developed to generate the ARD in a fully automatic mode in VITO's data centre.


Figure 13: Workflow chart showing the generation of the PROBA-V UTM Analysis Ready Data (ARD) archive out of the PROBA-V L1C archive.

The term ARD was defined by Dwyer et al. (2018) as: "Data that have been processed to allow analysis with a minimum of additional user effort ... [and therefore,] ... reduce the burden of pre-Document-No. CGLOPS1_ATBD_LC100m-V3 © C-GLOPS1 Lot1 consortium Issue: I3.4
processing on users ...". The advantage of ARD archives are that now also users with less specific knowledge in geometric and atmospheric correction can directly use high quality data and focus fully on the production of added-value products (Dwyer et al., 2018). The PROBA-V UTM ARD follows the specifications set by Dwyer et al. (2018) and "...are provided as tiled, geo-registered, ... atmospherically corrected products defined in a common equal area projection, accompanied by spatially explicit quality assessment information, and appropriate metadata to enable further processing while retaining traceability of data provenance." (Dwyer et al., 2018).

The main steps (Figure 13) are that the production grid was switched from the geographic coordinate system WGS84 in algorithm V1.0 to UTM in version 2.0. The UTM system divides the Earth's surface into 60 zones. Most UTM zones have a horizontal width of $6^{\circ}$ of longitude with an own designated central meridian. This longitudal restriction allows a projection of the EO data with less over-/undersampling compared to the geographic coordinate system WGS84. Moreover, we decided to adapt also to the tiling grid of the Sentinel-2 mission of the EU Copernicus programme (ESA, 2018). This allows the continuation of the CGLS-LC100 mapping service since the operational phase of the PROBA-V mission has ended in April 2020. This aim was reached with an update in the warping and tiling of the existing PROBA-V Level 2A processor ([PROBA-V User Manual]; Sterckx et al., 2014). Since the existing PROBA-V Level 2A processor had issues in the snow/cloud detection over bright surfaces, the update implemented also the usage of a yearly bright surface mask (see section 3.3.2), an improved cloud detection and snow/ice detection.

Subsequently, an improved atmospheric correction based on the original PROBA-V Level 2B processor (Sterckx et al., 2014; Dierckx et al., 2014; [PROBA-V User Manual]) is applied to obtain Top-Of-Canopy (TOC) reflectance values. Next to the improved Simplified Method for the Atmospheric Correction (SMAC) algorithm based on Rahman \& Dedieu (1994) using ECMWF CAMS NRT data (see section 3.2.4), the bright surface and permanent snow mask, described in section 3.3.2, is used to avoid errors over bright surfaces such as salt pans, permanent glaciers etc. Moreover, a quality indicator for the atmospheric correction using SMAC was developed and implemented as status flags in the PROBA-V UTM L2B status mask (Figure 13).

Afterwards, all single PROBA-V UTM L2B EO images were post-processed to generate 1-day synthesis (S1) on 100 m and 300 m GSD. The complete PROBA-V archive was in this way translated into the PROBA-V UTM ARD archive which allows more flexibility and faster data processing, is now fully aligned with the Sentinel-2 tiling grid in tiling naming as well as tile dimensions, and corrected to surface reflectance (Top-of-Canopy).

### 3.3.2 Yearly Bright Surface and Permanent Snow detection

To improve the snow / cloud detection in the existing PROBA-V L2A processor ([PROBA-V User Manual]; Sterckx et al., 2014), a yearly bright surface mask is generated to reduces the false snow detection in bright areas (Figure 13). Therefore, we use the existing PROBA-V L2A collection 1 archive to run the yearly analysis. Main reasons to use the existing PROBA-V archive and only compile yearly masks are that bright surface areas do not change significantly over a year, that bright areas are mostly bigger areas and so difference between the UTM and geographic coordinate
Document-No. CGLOPS1_ATBD_LC100m-V3 © C-GLOPS1 Lot1 consortium

[^0]system WGS84 are marginal, and the processing time and costs. The PROBA-V UTM archive was build-up sequentially during cluster off-times, therefore to use the existing archive was more economic. As a by-product during the processing also a permanent snow mask is generated which is used in the later status mask update as well as in the final CGLS-LC100 map generation.

Figure 14 shows the first part of the workflow from the single PROBA-V L2A acquisition handling to generation of the raw bright surface and permanent snow masks. In detail, in each calendar year all existing single Collection 1 PROBA-V L2A image acquisitions in 100 m and 300 m are processed to obtain for each pixel the number of valid observations, number of snow count, value of the maximum Normalized Difference Vegetation Index (NDVI), and number of bright surface count. Therefore, the PROBA-V L2A status mask and the single reflectance bands are analysed. A valid observation exists when all four reflectance bands have a value higher than 0 and the status mask doesn't indicate a "undefined" flag. For the maximum NDVI value (more detailed description of the NDVI see section 3.8.2.1) a pre-check on the red reflectance band ensures that no extremes are taken into account since we do not use the PROBA-V L2A flags for snow/ice and cloud/cloud shadows. For the number of bright surface counts and the number of snow counts, we are using the existing PROBA-V thresholding approach to detect snow/ice and clouds of the PROBA-V L2A processor of collection 0 . Note: we use the cloud detection algorithms for the calculation of the number of bright surface counts following the assumption that a pixel cannot be detected as cloud in all valid observations over a calendar year. Finally, empirical thresholds on these four indicators are used to generate the raw bright surface and permanent snow masks. The empirical thresholds were extracted by visual checks using global distributed test areas (Figure 15).


Figure 14: Yearly Bright Surface and Permanent Snow detection - Part A: Raw mask generation using the PROBA-V L2A archive.

Document-No.
Issue: 13.4


Figure 15: Distribution of global test sites used in the PROBA-V UTM ARD generation. Test sites following the Sentinel-2 naming convention.

Figure 16 shows the second part of the workflow where the raw bright surface and permanent snow masks are post-processed and finally resampled/re-warped to PROBA-V UTM resolution and tiling grid. For the post-processing, four ancillary datasets are used: the PGISM and PPSIM (see section 3.2.6), the land/sea mask (see section 3.2.2) and the Sentinel-2 tiling grid (see section 3.2.3). The PGISM is used as inclusion mask on the raw permanent snow mask, where the PPSIM is used as exclusion mask and all detected snow pixels outside the mask are reset. Moreover, also all areas in sea following the land/sea mask are reset in the raw permanent snow mask. Further, a spatial filtering is done to reduce artefacts and get consistent closed permanent snow areas (snow areas without single pixel holes). Afterwards, an iterative approach cleans overlapping areas between the raw permanent snow and raw bright surface masks. Therefore snow/bright surface area image statistics on tile basis are used to clean detected permanent snow and bright surface pixels. Finally, the land/sea mask is applied again to reset sea areas in case the spatial filtering generated artefacts over the sea.

Last step in the mask generation is the re-warping and resampling of the masks generated in the geographic coordinate system WGS84 - known as EPSG:4326 - to UTM and 100 m GSD , 300 m respectively. Since the UTM system consists out of 120 single regional coordinate systems (ESPG: 32601-32660 and 32701-32760), this has to be done zone wise. Finally, the zones are tiled up into Sentinel-2 tiles following the naming convention and tile size as also later used in the PROBAV UTM ARD product (Figure 16). The resulting yearly 100 m and 300 m bright surface masks and permanent snow masks can be now used in the improved L2A processing to reduce confusion in the snow/ice and cloud detection. Figure 17 shows some examples of the detected bright surfaces and permanent snow surfaces for the reference year 2015.


Figure 16: Yearly Bright Surface and Permanent Snow detection - Part B: Mask cleaning and final Sentinel-2 tile generation.


Figure 17: Examples for detected bright surfaces and permanent snow for year 2015 during the PROBA-V UTM ARD pre-processing. Upper left: bright desert sand in tile 11SNA; upper right: bright roof tops in tile 31UES; lower left: permanent snow (yellow) and bright rock (red) in tile 32TNS; lower right: bright roof top in tile 31UES.

| Document-No. | CGLOPS1_ATBD_LC100m-V3 | © C-GLOPS1 Lot1 consortium |
| :--- | :--- | ---: |
| Issue: 13.4 | Date: 31.03 .2021 | Page: $\mathbf{4 3}$ of $\mathbf{1 5 2}$ |

### 3.3.3 Level 2A processing and Level 2A post-processing

The existing PROBA-V Level 2A processor ([PROBA-V User Manual]; Sterckx et al., 2014), was designed to geometric correct L1C segments into a regularly grid in the geographic coordinate system WGS84 - known as EPSG:4326. Moreover, the L2A processor runs next to the mapping also a pixel-based snow/ice detection, cloud, and cloud shadow detection on every single acquisition. This information is written as flags into a status mask which is delivered together with Aerosol Optical Thickness (AOT) reflectance bands and separate datasets containing information for each pixel to the Sun-Sensor-Object geometry as well as the overpass time. The PROBA-V L2A status mask bit mapping provides the following information for each pixel (Table 4): the observation status (bits 1 to 3 ) giving results of snow/ice, cloud, and cloud shadow detection; the land/sea mask bit (bit 4); bits 5-8 are quality flags for radiometric quality; bits $9-12$ are flags showing the pixel coverage by the reflectance bands, and bits 13-16 are currently unused.

A more detailed explanation regarding the PROBA-V L2A processor can be found in Sterckx et al. (2014), this documentation will focus on the changes to improve the L2A correction.

Table 4: Status mask bit mapping of the PROBA-V L2A data

| Bit | Name | Description |
| :---: | :---: | :---: |
| 1-3 | Observation | 000: clear <br> 001: shadow <br> 010: undefined <br> 011: cloud <br> 100: snow/ice |
| 4 | Land/sea mask | 0 : sea <br> 1: land |
| 5 | SWIR quality flag | 0 : invalid data <br> 1: valid data |
| 6 | NIR quality flag | 0 : invalid data <br> 1: valid data |
| 7 | RED quality flag | 0 : invalid data <br> 1: valid data |
| 8 | BLUE quality flag | 0 : invalid data <br> 1: valid data |
| 9 | SWIR coverage | $\begin{aligned} & 0: \text { no } \\ & \text { 1: yes } \end{aligned}$ |
| 10 | NIR coverage | $\begin{aligned} & 0: \text { no } \\ & \text { 1: yes } \end{aligned}$ |
| 11 | RED coverage | $\begin{aligned} & 0: \text { no } \\ & \text { 1: yes } \end{aligned}$ |
| 12 | BLUE coverage | $\begin{aligned} & 0: \text { no } \\ & \text { 1: yes } \end{aligned}$ |

The major update on the L2A processor for the PROBA-V UTM processing was to adjust the geometric correction to the UTM system and tiling the L1C segments into Sentinel-2 conform tiles in
the 120 single regional UTM coordinate systems (ESPG: 32601-32660 and 32701-32760) spanning the globe (Figure 13). Therefore a Sentinel-2 tiling grid (see section 3.2 .3 ) adjusted to the PROBA-V imaging area and global buffered land/sea mask (see section 3.2.2) was generated. This "PROBA-V UTM tiling grid" consists of 18,595 tiles with a dimension of $110 \times 100 \mathrm{~km}$ and covering all global land masses (Figure 18). Note: the tiling grid is still called "Sentinel-2 tiling grid.


Figure 18: Sentinel-2 tiling grid overlaid on PROBA-V imaging area. Shown are the 18,595 tiles covering the land masses which form the PROBA-V UTM ARD archive.

The remaining PROBA-V L2A processor stayed untouched and provided for the PROBA-V UTM images L2B processing next to the status mask files also the needed Sun-Sensor-Object geometries as well as the overpass time files. Anyhow, due to noticed issues in the L2A cloud and snow/ice detection - misclassification over bright surfaces, bad cloud and snow/ice separation over mountainous areas - we decided to implement a L2A post-processing step into the workflow (Figure 13).

The PROBA-V UTM L2A post-processing includes an improved snow/ice detection as well as cloud detection module (Figure 19 and Figure 20), which results in an update of the PROBA-V UTM status mask and therefore replacement of the old one (Figure 21). The snow/ice and cloud detection modules are empirical thresholds based detection algorithms which were already used in the L2A processor of PROBA-V Collection 0 (Sterckx et al., 2014). By combining the existing threshold approaches with additional ancillary data sets, the snow/ice and cloud separation was improved (Figure 19). The ancillary datasets are ECMWF CAMS NRT surface temperature data in a 6 hours temporal resolution (see section 3.2.4), a mountainous area mask (see section 3.2.7), the land/sea

| Document-No. | CGLOPS1_ATBD_LC100m-V3 | © C-GLOPS1 Lot1 consortium |
| :--- | :--- | ---: |
| Issue: 13.4 | Date: 31.03 .2021 | Page: $\mathbf{4 5}$ of $\mathbf{1 5 2}$ |

mask (see section 3.2.2), and the yearly bright surface and permanent snow masks (see section 3.3.2).

First, for every pixel in the PROBA-V UTM L2A single images, the status mask is analysed for valid pixels (pixels with good radiometry for all 4 reflectance bands, and that all 4 reflectance bands have a valid coverage). Hereafter, the thresholds for snow and clouds are applied and the decision taken if the pixel belongs to clear, cloud or snow (threshold are shown in Figure 19). To reset false detected snow, a combination of surface temperature and mountainous areas mask is used. This follows the assumption that in non-mountainous areas pixels with a daily minimum temperature above $2^{\circ} \mathrm{C}$ or a daily minimum temperature above $-2^{\circ} \mathrm{C}$ in combination with a daily maximum temperature above 4 ${ }^{\circ} \mathrm{C}$ cannot contain snow.


Figure 19: PROBA-V UTM L2A post-processing - Part A: improved snow and cloud detection - raw detection by thresholding.

Second, the yearly permanent snow mask is used to update the snow detection by imprinting the mask. Similarly, the yearly bright surface mask is used to reset false detected clouds in bright areas. Afterwards, the land/sea mask is used to reset detected snow or clouds over sea (these pixels are ignored overall and set to undefined in the status mask). All remaining land pixels are handled by analyzing the image statistics, where the ratio of detected cloud pixels versus defined land pixel triggers different decision trees (see decision boxes in Figure 20).


Figure 20: PROBA-V UTM L2A post-processing - Part B: improved snow and cloud detection cleaning by masks and resolving confusions.


Figure 21: PROBA-V UTM L2A post-processing - Part C: updating of status mask.

After the improved ice/snow and cloud detection for the single PROBA-V UTM L2A tile was carried out, the cloud shadow detection module of the original PROBA-V L2A processor is used to detect cloud shadows (Figure 21). The final step is the updating of the PROBA-V UTM L2A status mask with the newly snow/ice, cloud and cloud shadow information detected on the tile acquisition. Moreover, an additional bit flag compared to the original PROBA-V L2A status mask (Table 4) is introduced in the "observation" section which identify bright surface pixels by imprinting the yearly

```
Document-No.
Issue: I3.4
```

CGLOPS1_ATBD_LC100m-V3
Date: 31.03.2021
© C-GLOPS1 Lot1 consortium
Page: 47 of 152
bright surface mask for this tile (Table 5). Then the new land/sea mask is imprinted. The inner tile boundary ( 3 pixels at 100 m and 1 pixel at 300 m ) is reset to "undefined" in order to avoid cloud and cloud shadow detection at the tile boundary. Note: due to the fact that the Sentinel-2 tiling grid has a tile overlap of $20 \%$, the information at the tile boundary is saved in a neighbouring tile and combined during the L3 post-processing (see section 3.3.5). In the last step, the PPSIM is used to detect saturated pixels in non-bright surface areas and non-permanent snow areas by checking if the remaining "clear" land pixels in non-PPSIM areas do not have reflectance values above 1.0 (scaled value of 2000).

### 3.3.4 Level 2B processing

The PROBA-V UTM L2B processor, applying the atmospheric correction to convert TOA reflectance into TOC (also known as surface reflectance), is built on the proven atmospheric processor of PROBA-V using SMAC (Figure 13, Sterckx et al., 2014; Dierckx et al., 2014; [PROBA-V User Manual]). To improve the atmospheric correction for PROBA-V UTM, the SMAC 4.0 algorithm (Rahman and Dedieu, 1994) was enhanced by some extra status mask checks, a fall-back mechanism to avoid errors over bright surfaces such as salt planes, permanent glaciers etc., and the calculation of TOC quality indicators (Figure 22). The atmospheric correction is carried out pixelbased for every single PROBA-V UTM L2A 100 m and 300 m acquisition in the archive separately.

A detailed explanation regarding the SMAC atmospheric correction can be found in Rahman \& Dedieu (1994), this documentation will focus on the changes to improve the L2A correction.


Figure 22: PROBA-V UTM L2B processor - improved SMAC routine and quality indicator generation for atmospheric correction.

Main input into the PROBA-V UTM L2B processor (Figure 22) are the following datasets:

- The four PROBA-V UTM L2A reflectance datasets (for band overview see Table 3);
- The updated PROBA-V UTM L2A status mask;
- The PROBA-V UTM L2A image viewing geometry (saved in the geometric parameter files):
- Viewing Azimuth Angle (VAA) for each pixel
- Viewing Zenith Angle (VZA) for each pixel;
- The PROBA-V UTM L2A sun geometry (also saved in the geometric parameter files):
- Sun Azimuth Angle (SAA) for each pixel
- Sun Zenith Angle (SZA) for each pixel;
- Scene (tile) topographic information in form of a DEM (saved in geometric parameter files provided by PROBA-V and not from DEM in section 3.2.12);
- Scene (tile) information regarding the aod550, tcwv, and gtco3 concentrations;
- Scene (tile) aod550 fall-back algorithm;
- Sensor specific SMAC coefficient files for each of the four reflectance bands and three cameras of PROBA-V.

One important change was the switch from low resolution, monthly long-term CLIMATOLOGY statistics (2003 - 2012) for aod550 and gtco3 used in the original PROBA-V L2B processor towards ECMWF CAMS NRT datasets with a 6-hours temporal resolution. The ECMWF CAMS NRT datasets for tcwv and gtco3 are real measurements, where aod550 comes from a model-driven approach with two CAMS forecast model runs per day, at 00 UTC and 12 UTC. Satellite aod550 retrievals from MODIS Deep Blue and EUMETSAT PMAp (retrievals based on AVHRR, GOME, and IASI observations) are assimilated in these models (see section 3.2.4). Figure 23 shows a visual comparison of the higher spatial resolution of the ECMWF CAMS NRT data compared to the CLIMATOLOGY and MeteoServices datasets. Since the ECMWF CAMS NRT datasets come in a 6 -hour temporal resolution, the needed ancillary data for the SMAC 4.0 routine can be better matched to the time stamp of the image acquisition. This higher temporal resolution allows the adaptation of the SMAC atmospheric processing to regional and temporal phenomena compared to the usage of monthly long-term statistics.

The first step in the PROBA-V UTM L2B processing is the evaluation of the status mask. Only the reflectance's of clear, snow, or bright pixels are input into the SMAC 4.0 routine together with the geometric parameters of the Sun-Object-Sensor geometry for each reflectance band, the DEM, the ECMWF CAMS NRT data, and the sensor specific SMAC coefficients (Figure 22). In a following step, the resulting surface reflectance values are checked pixelwise for negative reflectance and/or if the pixel was not SMAC processed due to another error. These pixels are sent back to the SMAC 4.0 routine but processed with an aod550 fall-back procedure, since tests have shown that negative surface reflectance values were mainly produced due to too high values in the ECMWF CAMS NRT aod550 data. A more conservative aod550 estimation based on latitude position is therefore used (Berthelot and Dedieu,.1997). The surface reflectance values of the reprocessed pixel is then again
checked for negative reflectance values and flagged with the status "undefined" in case of positive detection.


Figure 23: Comparison of used ancillary data used in the improved and original PROBA-V L2B processor. Left: ECMWF CAMS NRT data with 6-hour temporal resolution; right: long-term monthly data used in original L2B processor. Note: visualizations are shown for October 15, 2014.

Another enrichment of PROBA-V UTM is the delivery of a quality indicator regarding the estimated TOC reflectance using the SMAC 4.0 routine. This quality indicator is an empirical decision tree developed on the research results of Proud et al. (2010a; 2010b). The quality indicator uses thresholds on the pixel-based available SZA, VZA, and tcwv and gtco3 concentrations to categorise the TOC quality into 4 groups: good, acceptable, unreliable, and bad. The thresholds were retrieved from regression graphs calculated on global differences between SMAC reflectance values and Meteosat Second Generation TOC reflectance values (Proud et al., 2010a; Proud et al., 2010b). The following pseudo-code shows the used decision tree to determinate the TOC quality class:

- Good: (SZA <= $45^{\circ}$ ) AND (VZA <= $45^{\circ}$ ) AND (SZA + VZA <= $70^{\circ}$ );
- Acceptable: $\left(\left(S Z A>45^{\circ}\right)\right.$ AND $\left.\left(S Z A<=60^{\circ}\right)\right)$ OR ((VZA > $\left.45^{\circ}\right)$ AND $\left(V Z A<=60^{\circ}\right)$ ) OR ((SZA + VZA > $70^{\circ}$ ) AND (SZA + VZA $\left.<=100^{\circ}\right)$ );
- Unreliable: $\left(\left(\mathrm{SZA}>60^{\circ}\right)\right.$ AND $\left.\left(\mathrm{SZA}<=75^{\circ}\right)\right)$ OR ((VZA > $\left.60^{\circ}\right)$ AND $\left(\mathrm{VZA}<=75^{\circ}\right)$ ) OR $\left(\left(S Z A+V Z A>100^{\circ}\right)\right.$ AND $\left.\left(S Z A+V Z A<=140^{\circ}\right)\right)$;
- Bad: $\left(S Z A>75^{\circ}\right) \mathrm{OR}\left(V Z A>75^{\circ}\right) \mathrm{OR}\left(S Z A+V Z A>140^{\circ}\right)$.

An extra rule applies on pixel already classified as "good". In case the pixel values for gtco3 is higher than $0.35 \mathrm{~atm}-\mathrm{cm}$ or the pixel value for tcwv is higher than $2.25 \mathrm{~g} / \mathrm{cm}^{2}$, the TOC quality indicator for this pixel is downgraded to "acceptable" (Proud et al., 2010b).

The TOC quality level as well as a status bit showing if the L2B processor was successfully run is pixelwise coded in the enhanced L2B status mask (Figure 22 and Table 5).

### 3.3.5 Level 3 processing and final status mask

The last step in the PROBA-V UTM ARD generation (Figure 13) is the compositing - called L3 processing - of the single observations of all sensor overpasses per day into 1-day synthesis at 100 m and 300 m , respectively. The 1-day synthesis - also called TOC-S1 - is done following these rules in order to archive the highest possible data quality:

- first: the synthesis is done per TOC quality level, meaning pixels of overpasses with higher TOC quality will be preferred and finally stacked in the order bad, unreliable, acceptable, and good overwriting lower TOC qualities;
- second: camera 2 (center camera) has priority within a TOC quality level, meaning that even when pixels from different overpasses have the same TOC quality level, the pixels taken with the PROBA-V center camera are preferred due to the fact that less steep viewing geometries can be expected;
- third: combination within each TOC quality level is done via calculating the median, meaning that in cases when pixels from different overpasses have the same TOC quality level and are either all from the center camera or only from the off-view cameras, the reflectance values are combined by calculating the median within the TOC quality level;
- fourth: pixel with no valid observations are set to "undefined" in the status mask and to nodata in the reflectance bands.

Figure 24 shows an example for the generation of a PROBA-V UTM TOC-S1 synthesis product. The first two images show the same tile at two different sensor overpasses at the same day at 300 m since each camera strip (overall three) is processed as segment separately, this constellation is quite common - where the third image shows the seamless 1-day synthesis.

In the L3 post-processing redundant data between a tile and its neighbouring tiles within an UTM zone is resolved by applying the same synthesis rules as for the 1 -day compositing. Due to the fact that in the original Sentinel-2 tiling grid a tile overlap of 10 km to each side is defined, this processing step saves overall $20 \%$ data storage. Moreover, this step fills the introduced "nodata" inner tile boundary during the L2A processing with valid data from the neighbouring tiles.

Finally, the PROBA-V UTM L3 (TOC-S1) status mask is generated. Next to the L2A and L2B status mask information, an extra status bit is added which shows if all PROBA-V UTM processing steps were successfully carried out and this pixel holds valid data to be included in the ARD (ARD status flag). The PROBA-V UTM ARD archive consists therefore of the PROBA-V UTM L3 reflectance bands and status mask at 100 m and 300 m for all single acquisitions between 2013 - 2020.


Figure 24: PROBA-V UTM L3 processing - generation of 1 -day synthesis. From left to right: partly covered tile in first and second overpass, 1-day synthesis, and Google Earth image for comparison.

Table 5: PROBA-V UTM L3 status mask. New flags marked in bold.

|  | Bit | Name | Description |
| :---: | :---: | :---: | :---: |
|  | 1-3 | Observation | 000: clear <br> 001: shadow <br> 010: undefined <br> 011: cloud <br> 100: snow/ice <br> 111: bright surface |
|  | 4 | Land/sea mask | $\begin{aligned} & 0: \text { sea } \\ & \text { 1: land } \end{aligned}$ |
|  | 5 | SWIR quality flag | 0 : invalid data <br> 1: valid data |
|  | 6 | NIR quality flag | 0 : invalid data <br> 1: valid data |
|  | 7 | RED quality flag | 0 : invalid data <br> 1: valid data |
|  | 8 | BLUE quality flag | 0 : invalid data <br> 1: valid data |
|  | 9 | SWIR coverage | $\begin{aligned} & \text { 0: no } \\ & \text { 1: yes } \end{aligned}$ |
|  | 10 | NIR coverage | $\begin{aligned} & \text { 0: no } \\ & \text { 1: yes } \end{aligned}$ |
|  | 11 | RED coverage | $\begin{aligned} & \text { 0: no } \\ & \text { 1: yes } \end{aligned}$ |
|  | 12 | BLUE coverage | $\begin{aligned} & \text { 0: no } \\ & \text { 1: yes } \end{aligned}$ |
|  | 13-14 | TOC quality | 00: bad <br> 01: unreliable <br> 10: acceptable <br> 11: good |
|  | 15 | SMAC flag | 0 : not carried out <br> 1: successful |
|  | 16 | ARD flag | 0 : not usable <br> 1: all steps successful |
| Document-No. Issue: I3.4 |  | $\begin{aligned} & \text { ATBD_LC100m-V3 } \\ & 2021 \end{aligned}$ | © C-GLOPS1 Lot1 Page: 52 |

Copernicus Global Land Operations - Lot 1 Date Issued: 31.03.2021

### 3.3.6 Comparison PROBA-V versus PROBA-V UTM

Figure 25 gives examples showing the improvements in the L2A snow/ice and cloud detection between PROBA-V Collection 1 and PROBA-V UTM. The better snow delineation is visible in the improved algorithm.


Figure 25: Comparison between PROBA-V and PROBA-V UTM cloud/snow detection.

Document-No.
Issue: I3.4

CGLOPS1_ATBD_LC100m-V3
Date: 31.03.2021
© C-GLOPS1 Lot1 consortium Page: 53 of 152

Figure 26 shows the improvement in the geometric quality due to the switch from EPSG:4326 to UTM projection. The fields are now comparable due to the minimal distortion related to WGS84 the UTM projection is designed to have a scale error from less than 0.1 percent.


Figure 26: Comparison between PROBA-V (left) and PROAB-V UTM (right) geometric correction (L2A). Already in mid-latitudes a reduced distortion is viewable. Area shown: province Niedergoersdorf, Teltow-Flaeming, Germany.

## 3.4 Еросн-wise Data Cleaning \& Compositing

### 3.4.1 Overview

After generation of the PROBA-V UTM ARD archive, the next step in the CGLS-LC100 product workflow (Figure 2) is the data cleaning and compositing (Figure 27). Input data are the multi-spectral PROBA-V UTM TOC-S1 datasets as described in section 3.3 for an epoch around the reference year. An epoch thereby consist of three consecutive years of satellite data around the reference year for the land cover map (reference year $+/-1$ year). The cleaning and composing is done sequentially for all epochs from 2015 to 2019.

Next to the status mask cleaning of the 100 m and 300 m PROBA-V UTM TOC-S1 observations for the reference year plus/minus 1 year (overall 3 years or up to 1,096 daily observations), a temporal outlier cleaning algorithm called Median Absolute Deviations of HANTS (madHANTS) is applied. In order to reduce the noise in the 1-day synthesis products as well as to generate regular time steps, a median compositing algorithm is used to generate 5 -day and 10-day median composites (MC5 and MC10, respectively). For the 300 m PROBA-V UTM ARD data also a Harmonized 5-day Median Composite (HMC5) for the whole PROBA-V data archive (2013-2020) is produced. Therefore, a harmonization algorithm called Harmonic Analysis of Time Series (HANTS) is used.

After the data cleaning and compositing step, the archive is upgraded into the PROBA-V UTM ARD+ status where the plus (+) indicates the additional temporal outlier/cloud screening (Figure 2). Note: the PROBA-V UTM ARD + archive is only generated out of the PROBA-V UTM TOC-MC5 datasets in 100 m and 300 m as well as the PROBA-V UTM TOC-MC10 datasets in 300 m . The PROBA-V

UTM TOC-HMC5 datasets in 300 m as shown in Figure 27 are only produced temporary as input for the next processing steps.


Figure 27: General overview of the data cleaning and compositing section in the CGLS-LC100 product workflow. Note: Numbers in the upper left corner of data container indicate the number of layers - here number of observations - for one epoch used to create the yearly land cover map.

### 3.4.2 The HANTS and madHANTS algorithms

The HANTS algorithm and derived madHANTS algorithm are frequently used in the CGLS-LC100 product workflow.

The HANTS algorithm is an evolution of the Fast Fourier Transform (FFT) algorithm to perform Fourier analysis (Verhoef, 1996; Roerink et al., 2000). It was developed to deal with time series data with irregularly spacing and to identify and remove outliers as well as reconstruct the removed data in one step (Verhoef, 1996). Since we only use the HANTS algorithm to identify the most significant frequencies and/or to generate a gap-less time series profile by applying the identified frequencies, we developed an algorithm to screen the time series observations for temporal outliers (mainly clouds and haze) based on the HANTS algorithm. Therefore, we combined the HANTS algorithm with an outlier test based on Median Absolute Deviations (MAD) (Walker, 1931) - which created the madHANTS algorithm. Next to a performance gain compared to the original HANTS algorithm used for outlier detection, the madHANTS algorithm is more adaptable to the data. The madHANTS algorithm works in two steps:

1. First, the HANTS algorithm is applied on the original time series data and for each time step (index) the harmonized reflectance value is calculated out of the phases and amplitudes of the identified frequencies. The usage of 3 frequencies above the zero-frequency (so overall 4 frequencies) showed the best performance-to-accuracy ratio in the CGLS-LC100 product
workflow where the time series profiles for each pixel are 3 years long (reference year plus/minus 1 year).
2. In the second step, the outlier detection is performed - the original pixel reflectance values along the time dimension are evaluated against the corresponding harmonized reflectance values via the MAD outlier test. Therefore, an evaluation score is calculated for each pixel value along the time series profile. The evaluation score is calculated using equation 1. In the original MAD test (Leys et al., 2013), this evaluation score was calculated by subtracting the median of all pixel reflectance values from the current pixel reflectance value $x_{i}$ in the numerator of the division. The adaptation in equation 1 using instead the median of the corresponding harmonized pixel reflectance values in the numerator allows a better adaptation of the outlier detection algorithm to seasonality.

$$
\begin{equation*}
s=\frac{\left|x_{i}-x_{i \text { HANTS }}\right|}{\operatorname{median}\left(\left|x_{1}-x_{1 \text { HANTS }}\right|, \ldots,\left|x_{n}-x_{\text {nHANTS }}\right|\right)} \tag{Eq. 1}
\end{equation*}
$$

, where $s$ is the evaluation score to determine if a pixel reflectance value along a time series is an outlier compared to the neighbouring time steps, $x_{i}$ is the current pixel reflectance value of the to test time series step, $x_{i H A N T s}$ is the harmonized pixel reflectance value of the current time step using the HANTS algorithm.

A threshold of 3.5 (comparable to $\sim 3.5$ standard deviations) is used on the evaluation score to determine if a pixel reflectance value has to be characterized as outlier compared to the temporal neighbouring pixel reflectance values. The MAD outlier test was chosen since the MAD, as a measure of statistical dispersion, is more resilient to outliers in a data set than the standard deviation. Where in the standard deviation the distances from the mean are squared (thus outliers can heavily influence it), in the MAD the deviations of a small number of outliers are irrelevant (Leys et al., 2013). Figure 28 shows an example of the applied madHANTS algorithm on a PROBA-V UTM TOC-S1 100 m time series of a pixel at location $9.459^{\circ}$ lon, $6.562^{\circ}$ lat.

Figure 28 also shows a shortcoming of the reconstructed time series using the HANTS algorithm (red line). Since the HANTS algorithm needs several observations in the beginning and end of the time series to find the optimal solution, the amplitudes and phases of the identified frequencies don't produce always reliable results for the reconstruction of the original time series in the beginning and end of the time series (up to 10 time series steps).


Figure 28: Example for the madHANTS temporal outlier detection algorithm for the epoch2015 (reference year $2015+/-1$ year). Note: blue stars mark all original pixel values in the time series (invalid observation values are set to -1), the red line shows the harmonized time series by applying the amplitudes and phases of the identified frequencies during the HANTS transformation, red squares show the detected outliers which are more than $\sim 3.5$ standard deviations away from the harmonized pixel value.

### 3.4.3 Median Composites (MC) generation

The general workflow for the data cleaning and compositing section (outlined in Figure 27) can be split in two steps. The first is the generation of status mask cleaned and temporal outlier cleaned 5day and 10-day median composites for the 100 m and 300 m time series data. The median compositing reduces the random component of the noise compared to the 1-day TOC products.

Overall three processing sub-steps are needed in order to generate the 5-day (MC5) and 10-day (MC10) composites from the PROBA-V UTM ARD archive for an epoch (Figure 27). First, all multispectral PROBA-V UTM TOC-S1 image acquisitions are identified in the ARD archive via a Structured Query Language (SQL) query for the reference year plus 1 year before and after the reference year (called epoch). This query can give back up to 1,096 observations (time steps) for each single pixel in the 300 m and 100 m tiles of the PROBA-V UTM ARD archive. The time series density for each pixel not only depends on the GSD (PROBA-V has a daily revisit time at the equator at 300 m and a 5 -day at 100 m ), but also on the latitude (the higher the latitude the more overlapping orbits exists, but also the fewer observations can be done in the winter period). Each PROBA-V UTM TOC-S1 dataset consists of four spectral bands and the corresponding status mask file (see section 3.3.5) which are handled simultaneously. The status mask cleaning sets pixel flagged as noise ("undefined"), cloud, cloud shadow, sea, or with an invalid TOC quality flag to a "no data" value (see Table 5). The datasets can be identified by the new naming: PROBA-V UTM TOC-S1 SM cleaned.

In the next processing sub-step, the madHANTS algorithm (see 3.4.2) is applied to clean the reflectance time series from remaining haze and undetected clouds during the PROBA-V UTM L2A processing (see section 3.3.3). Therefore, the madHANTS outlier test is conducted for the blue and SWIR reflectance bands of the PROBA-V UTM TOC-S1 SM cleaned 100 m time series ( 300 m respectively). The combination of detected outliers in the blue as well as SWIR reflectance bands showed the best overall detection of temporal outliers. The madHANTS outlier test is pixel-based which means that the time series profile of each pixel is evaluated independently. As soon a time step in the blue or SWIR reflectance band was detected as an outlier, this time step is flagged in all four reflectance bands of that pixel. Figure 29 shows an example of the madHANTS temporal outlier cleaning algorithm. In order to optimize the workflow, the madHANTS algorithm is applied simultaneously on all pixels within an image line of a PROBA-V UTM tile using multi-core computing as well as several PROBA-V UTM tiles are processed simultaneously via cloud computing. Output of the madHANTS processing step are PROBA-V TOC-S1 SM and outlier cleaned 100 m time series for the four reflectance bands ( 300 m respectively). Figure 30 shows an example of the data quality improvement by applying the SM cleaning and temporal outlier cleaning processing steps on a PROBA-V UTM TOC-S1 100 m tile.


Figure 29: Example showing the temporal outlier detection using the madHANTS algorithm on the blue and SWIR reflectance values of a pixel at location $9.459^{\circ}$ Ion, $6.562^{\circ}$ lat. Note: red curve shows the harmonized time series; blue shows the valid pixel reflectance values; green shows flagged outliers cumulative detected in the blue or SWIR reflectance band.


Figure 30: Example for status mask and temporal data cleaning for part of a PROBA-V UTM TOC-S1 100 m tile (observation taken 2016-03-06). Image is shown as false colour composite (RGB = SWIR, NIR, blue) for a sample area in tile in Nigeria: raw image (left), status masked cleaned image (status mask flagged areas are shown in red - left), madHANTS cleaned image (additional pixels which are flagged as outliers or clouds are shown in blue - right).

The last processing sub-step is the generation of the MC5 for the 100 m and 300 m as well as MC10 composites for the 300 m time series reflectance cubes. The MC5 (MC10 respectively) composites for the four spectral bands are obtained by calculating the median of the PROBA-V UTM TOC-S1 SM and outlier cleaned daily reflectance values over a 5 -days window (10-days window respectively). The median composites are calculated across the four spectral bands for each pixel and all pixels in a full tile simultaneously. Final output are the following three datasets:

- PROBA-V UTM TOC-MC5 100 m reflectance data cubes with 216 regular time steps;
- PROBA-V UTM TOC-MC5 300 m reflectance data cubes with 216 regular time steps;
- PROBA-V UTM TOC-MC10 300 m reflectance data cubes with 108 regular time steps.


### 3.4.4 Long-term 5-day Median Composite (HMC5) generation

The second step in the general workflow for the data cleaning and compositing section (outlined in Figure 27) is the generation of the long-term 5-day median composite (HMC5) for PROBA-V UTM

TOC-S1 300 m . Since the HMC5 is produced out of the whole PROBA-V UTM ARD archive (2013 - 2020), the HMC5 workflow had to be adapted to the data volume and differs slightly from the MC5 and MC10 generation workflow (section 3.4.3).

First, all PROBA-V UTM TOC-S1 300 m image acquisitions are identified in the ARD archive via a SQL query, which can be over 2,100 observations (time steps) for each pixel. Next follows the status mask cleaning analogue to the step in section 3.4.3.

Instead of proceeding with the temporal outlier screening of the PROBA-V UTM TOC-S1 SM cleaned reflectance cubes, the 300 m MC5 composites are generate first in the HMC5 workflow (Figure 27). This change was needed since the madHANTS algorithm becomes slower in the outlier detection the more observations (time steps) in a pixel time series exist, which made a time effective processing impossible. The 5-day compositing compress the data amount by factor 5. Again, the MC5 composites for the four spectral bands are obtained by calculating the median of the PROBAV UTM TOC-S1 SM cleaned 300 m daily reflectance values over a 5 -days window. Output are PROBA-V UTM TOC-MC5 SM cleaned 300 m reflectance cubes spanning the full temporal profile of the ARD archive. Next, the temporal outlier screening and removal using the madHANTS workflow as explained in section 3.4 .3 is carried out. Output of that are the PROBA-V UTM TOC-MC5 SM and outlier cleaned 300 m time series cubes for the four reflectance bands.

The final sub-step is the generation of the harmonized composites. Therefore, the HANTS algorithm is used (see section 3.4.2). For each pixel and reflectance band independently, the HANTS algorithm identifies the four most significant frequencies (3 frequencies above the zero-frequency) in the time series profile of the PROBA-V UTM TOC-MC5 SM and outlier cleaned 300 m data cubes. Then the phases and amplitudes of the identified frequencies are used to generate a harmonized time series for each pixel and reflectance band. Finally, the HMC5 is cut to the same time interval as the PROBA$V$ UTM TOC-MC5 300 m data set for the specific epoch. Final output is the following dataset:

- PROBA-V UTM TOC-HMC5 300 m reflectance data cubes with 216 regular time steps.


### 3.5 Epoch-Wise Input Data Quality Indicator Generation

### 3.5.1 Overview

To provide the product user with a quality indicator for the input data as well as to account in the supervised learning for training points from areas with a low temporal data density, we developed the Data Density Indicator (DDI) (Figure 32). In the latter case the DDI assigned to the training point locations is used as a weight factor in the training of the supervised classification/regression models for the BaseYears. The DDI is calculated for the reference year $+/-1$ year from the PROBA-V UTM ARD+ archive for 100 m and 300 m , so before the data fusion (Figure 2). Overall 19 single time series statistics are calculated and combined via a scoring approach into the DDI which ranges from 0 to 100, where 0 indicates no existing input data and 100 the best data availability in the PROBA-V UTM ARD+ 100 m and 300 m time series. Figure 31 shows two representative temporal profiles for a pixel time series with a high DDI and low DDI, respectively.


Figure 31: Exemplary temporal profiles for PROBA-V UTM ARD+ time series with a high Data Density Indicator (DDI) (upper graph) and low DDI (lower graph) showing less available valid data. Note: x-axis shows the 5 -day time step starting from the epoch. Values different than zero on the $y$-axis shows that data is available on the corresponding time step.

The overall processing workflow, shown in Figure 32, consists of three main steps: calculation of the theoretic observations and time series gap (t-GAP) mask; calculation of 19 time series statistics; and finally scoring the statistics and extracting the data.


Figure 32: Data Density Indicator processing workflow.

Document-No.
Issue: 13.4

CGLOPS1_ATBD_LC100m-V3
Date: 31.03.2021
© C-GLOPS1 Lot1 consortium Page: 61 of 152

### 3.5.2 The PROBA-V theoretical observation mask and time series gap mask

The first step in calculating the DDI is deriving the Theoretical maximum number of Observations masks (Theo. Obs. mask) for 100 m and 300 m (Figure 33), and the generation of the t-GAP mask from the valid/invalid observations in the PROBA-V UTM ARD+ time series (Figure 33). Only the information of the red reflectance band was used since all reflectance bands in the PROBA-V UTM ARD+ show the same number of valid observations in the time series.

The theoretical observation mask is a flag mask identifying time steps in the time series of a pixel where given by the PROBA-V sensor specification at certain times in the calendar year no data acquisitions are possible (e.g. winter months in the High Latitudes). Therefore, the following three parameter sets are required:

- the PROBA-V camera viewing angles and orbit activation times depending on the day of the year;
- the development of the satellite overpass time from mission start up to the end of the operational phase;
- thresholds on the SZA at which the PROBA-V UTM L2B processor cannot give any usable data anymore.


Figure 33: Calculating the theoretical maximum number of observations (Theo. Obs. mask) and the longest concurrent gap in data time series data (T-GAP) is the first step in calculating the DDI characteristics. Calculating the Theo. Obs. Mask requires three parameters: a threshold on the SZA as used by SMAC, the satellite overpass time and the camera viewing angles.

The t-GAP mask is a binary quality layer in 100 m resolution that indicates when the gap between two valid observations within the time series of a pixel is larger than a threshold. The t-GAP mask was originally developed as quality indicator to evaluate the influence of the data fusion (Kalman filling approach - see section 3.6) on the data quality by flagging pixels in the image which have possible a higher data uncertainty due to the Kalman-filling. The basic principle is to calculate the longest concurrent gap between two valid observations in the 100 m as well as 300 time series of a pixel, and then to evaluate the length of these gaps against thresholds. These thresholds (gap lengths in days) were chosen experimentally by artificially increasing gap lengths in the 100 m and corresponding 300 m time series up to the moment the data fusion could not successfully reconstruct the 100 m time series within given specifications. For the calculation of the t-GAP mask, the PROBAV UTM ARD+ 100 m and 300 m pixel time series are converted into boolean arrays transforming the reflectance values solely into valid and invalid observation time steps. The masks are then cleaned by the corresponding theoretical observation masks to account for time steps which can never exist in the PROBA-V configuration. This cleaning is needed since otherwise the gap lengths of pixels in the High North during the winter months would be artificially too long. Finally, the longest concurrent gaps are calculated and evaluated. If, in the time series of a 100 m pixel, the longest concurrent gap exceeds 100 days and, in the corresponding 300 m time series, the longest concurrent exceeds 30 days, then this pixel is flagged as t-GAP.

### 3.5.3 Calculating the time series statistics

The calculation of the pixel-based time series statistics for the PROBA-V UTM ARD+ 100 m and 300 m are done individually. For the total number of observations, number of valid observations and the resulting observation ratio the PROBA-V UTM ARD+ datasets are used but are converted into boolean arrays transforming the reflectance values solely into valid and invalid observation time steps. For the remaining time series statistics, we account for the sensor specific observation limits of PROBA-V (see section 3.5.2). Therefore, the Theo. Obs. mask is checked for every pixel if the pixel's time series data has observation limits and therefore has to be updated for correct statistic calculation.

The following is a short description on how the statistical values are derived. Note: the word "raw" means that, for this statistic, the PROBA-V UTM ARD+ data was used, and the word "cleaned" indicates that the Theo. Obs. Mask was applied before the calculation.

- Total $\mathbf{N}^{\circ}$ of observations (Figure 34, equation 2). Calculated on the "raw" time series. This number is the total number of 5 -day observations (valid or invalid) over the epoch (reference year $+/-1$ year) and currently fixed to 216 observations.
- $\mathbf{N}^{\circ}$ of valid observations (Figure 34, equation 3). Calculated on the "raw" time series. This number is the number of valid observations over the epoch.
- Observation ratio (Figure 34, equation 4). This number is the ratio of the $N^{\circ}$ of valid observations to the Total $N^{\circ}$ of observations.


Figure 34: The raw time series statistics calculated on the PROBA-V UTM ARD+ data. After the raw time series statistics are calculated the PROBA-V UTM ARD+ data is checked for sensor specific observation limitations and updated in this case by imprinting the Theo. Obs. Mask.

- $\mathbf{N}^{\circ}$ of theoretical observations (Figure 35, equation 5). Calculated on the "cleaned" time series. This number is the total number of possible observations from the theoretical observation mask - indicating how many theoretical observation would be even possible given the PROBA-V sensor specifications (maximum would equal the total $\mathrm{N}^{\circ}$ of observations).
- Theoretical observation vs maximum observation ratio (Figure 35, equation 6). This number is the ratio of the $N^{\circ}$ of theoretical observations to the Total $N^{\circ}$ of observations.
- Theoretical observation ratio (Figure 35, equation 7). This number is the ratio of the $N^{\circ}$ of valid observations to the $N^{\circ}$ of theoretical observations.
- Gap weight (Figure 35, equation 8). Calculated on the "cleaned" time series. The gap weight is a scoring on the pixelwise gap histogram. For that the lengths of the gaps between valid observations within the cleaned time series are identified and counted. Depending on the gap length and occurrence of the different gap length classes a predefined score is assigned (see table in Figure 35).


Figure 35: The cleaned time series statistics - Part A.

- Observation frequency score (Figure 36, equation 9). Calculated on the "cleaned" time series. This statistical parameter depends on a few logical reasoning's. We will get a maximum observation frequency score if we have more valid observations than it was theoretically calculated or if no invalid observations are existing (the full time series exists only of valid observations). In all other case the observation frequency score is calculated. Therefore the observation frequency which depends on the ratio of 'the number of valid observations changes from False to True' (multiplied by two because this can only be half of the number of theoretical observations) to the number of theoretical observations is calculated. Afterwards, a final observation frequency score is calculated depending on the relationship between valid observations and invalid observations.


Figure 36: The cleaned time series statistics - Part B.

After all time series statistics are calculated for the 100 m and 300 m PROBA-V UTM ARD+ datasets, the statistics are stacked and saved as raster dataset to be used as quality indicators (Figure 37). Overall 19 time series statistics are provided:

- $\mathrm{N}^{\circ}$ of valid observations for 100 m and 300 m ;
- Total $\mathrm{N}^{\circ}$ of observations for 100 m and 300 m ;
- Observation ratio for 100 m and 300 m ;
- Observation frequency score for 100 m and 300 m ;
- Gap weight for 100 m and 300 m ;
- $\mathrm{N}^{\circ}$ of theoretical observations for 100 m and 300 m ;
- Theoretical observation ratio for 100 m and 300 m ;
- Theoretical observation vs maximum observation ratio for 100 m and 300 m ;
- Longest concurrent gap for 100 m and 300 m (see section 3.5.2);
- t-GAP mask (see section 3.5.2).

The last step is the calculation of the DDI by scoring the results of the time series statistics for the 100 m and 300 m PROBA-V UTM ARD+ datasets. The nine needed statistics and corresponding scores can be seen in the left part of Figure 37. Besides a visual representation of the PROBA-V UTM ARD+ for tile 32NPM (over central Africa), Figure 38 shows also some of the statistical images calculated on the time series.

| Document-No. | CGLOPS1_ATBD_LC100m-V3 | © C-GLOPS1 Lot1 consortium |
| :--- | :--- | ---: |
| Issue: 13.4 | Date: 31.03 .2021 | Page: $\mathbf{6 6}$ of $\mathbf{1 5 2}$ |



Figure 37: Final scoring of time series statistics to compute the Data Density Indicator (DDI) and output of the PROBA-V UTM ARD+ statistics.


Figure 38: Visual RGB representation of UTM tile 32NPM with some of the calculated DDI statistics for the 100 m resolution for epoch2015, i.e. Data Density Indicator, Observation ratio, Observation frequency, Gap weight and Longest concurrent gap.

Document-No.
Issue: I3.4

CGLOPS1_ATBD_LC100m-V3
Date: 31.03.2021
© C-GLOPS1 Lot1 consortium
Page: 67 of 152

### 3.6 Eросн-Wise Data Fusion

### 3.6.1 Overview

High temporal cloud coverage in several regions of the globe (Figure 39), e.g. Sahel zone, Central America, are challenging for all land cover classification approaches depending on optical data.


Figure 39: Average cloud cover over a 13 year period from July 2002 to April 2015 captured by the MODIS Aqua RS sensor. Note: Colours range from dark blue (no clouds) to light blue (some clouds) to white (frequent clouds). Image credits: NASA Earth Observatory, 2019.

In order to overcome the low data density in these areas and therefore data gaps in the PROBA-V UTM ARD+ 100 m time series, we developed a data fusion approach, based on a Kalman filtering approach (Kalman, 1960), which uses PROBA-V UTM 300 m data with a daily revisit time to fill time series gaps. Figure 40 illustrates the PROBA-V UTM 100 m and 300 m data density for Africa.


Figure 40: Number of cloud free data observations for PROBA-V UTM 100 m (left) and PROBA-V UTM 300 m (right) in the reference year 2015 for Africa.

Figure 41 illustrates an overview of the data fusion section in the CGLS-LC100 product workflow carried out for each epoch independently (Figure 2). Overall three processing steps are needed in order to produce high resolution ( 100 m ) time series with the temporal observation density of the low resolution ( 300 m ) time series. Input data is the PROBA-V UTM ARD+ archive (see section 3.4.3) for all four PROBA-V reflectance bands. In a first step, small data gaps in the 300 m MC5 time series are filled with corresponding existing observations in the MC10 time series, and large data gaps (> 25 days) are additionally filled with the corresponding observations in the HMC5. In the next step, the Kalman-filtering approach is applied in order to generate dense, high resolution MC5 time series. The last step is the screening for introduced outliers as well as the consistency of the data is checked.

This upgrades the PROBA-V UTM ARD+ archive into the PROBA-V UTM ARD++ archive, which gives temporal cleaned, consistent and dense 5-day image stacks for all global land masses at 100 m resolution (Note: 300 m data is not existing in the ARD++).


Figure 41: General overview of the data fusion section in the CGLS-LC100 product workflow. Note: Numbers in the upper left corner of data container indicate the number of layers - here number of 5day observations - for one epoch used to create the yearly land cover map.

### 3.6.2 Data Fusion pre-processing

Input data in the data fusion pre-processing step are the generated PROBA-V UTM TOC-MC5, TOCMC10 and TOC-HMC5 time series cubes in 300 m for all four reflectance bands (see section 3.4). In detail:

- PROBA-V UTM TOC-MC5 300 m for the blue, red, NIR and SWIR band;
- PROBA-V UTM TOC-MC10 300 m for the blue, red, NIR and SWIR band;
- PROBA-V UTM TOC-HMC5 300 m for the blue, red, NIR and SWIR band.

In the data fusion pre-processing, small gaps (5 - 10 day gaps) in the 300 m 5-day time series product are filled with the pixel values of the 10-day time series product for the corresponding time steps. In a second step, larger gaps (> 25 day gaps) in the 300 m 5 -day time series product are filled via interpolation with the HMC5 300 m long term trend product for the corresponding time series steps (see Figure 42). This gap filling is needed in order to guide the Kalman filtering approach in cases were no PROBA-V 100 m and 300 m TOC-MC5 data is available for more than 1 months in a row for a pixel. Note: no gaps are pre-filled in the 100 m data - therefore the Kalman-Filtering approach is used. The pre-processing works tile based meaning that all pixels and corresponding pixel time series are processed simultaneously for the 300 m datasets. Moreover, all four PROBA-V reflectance bands are processed simultaneously.

Output of the pre-processing is a modified 300 m TOC-MC5 datasets, the PROBA-V UTM TOC-MC5-TOC 300 m gap-filled time series for the four reflectance bands (Figure 41).


Figure 42: Example of data fusion pre-processing results for PROBA-V UTM TOC-MC5 300 m time series at pixel location $9.459^{\circ}$ lon, $6.562^{\circ}$ lat. top) continuous gap length in the 300 m time series (dashed green line indicates threshold for small gaps, dashed red line indicates threshold for big gaps), middle) time series before pre-processing (blue line shows original TOC-MC5 time series, green line shows original TOC-MC10 time series, red line shown the long term harmonized time series for the full PROBA-V archive), bottom) TOC-MC5 300 m time series after pre-processing (blue line shows the original TOC-MC5 time series, orange line shows the final pre-processed time series which will be used for the data fusion).

### 3.6.3 Principle of the Kalman-Filtering algorithm

The Kalman-filtering approach, also known as linear quadratic estimation, is a recursive algorithm that combines a model of a series of measurements of different variables observed over time with existing measurements with partly or fully missing variables in order to produce estimates of the unknown variables at each time step in the time series, even if no measurements are available for certain time steps (Welch \& Bishop, 2006; Kempeneers et al., 2016). Goal is to produce a complete time series for all variables. A good introduction into the Kalman-filtering approach can be found in Welch \& Bishop (2006).

Figure 43 illustrates the basic concept of the Kalman-filtering approach. The algorithm works in two steps, a prediction step and an update step (Aimonen, 2011; Welch \& Bishop, 2006). In the prediction step, the Kalman filter produces estimates of the variables for a time step $k$, along with their uncertainties, using the prior knowledge state at time step k-1. Thus, the Kalman filter keeps track of the estimated states of the system and the uncertainty of the estimate for all time steps using a dynamic model. Once the outcome of the next measurement is observed, the update step is initialised. The estimate for time step $k$ is updated by incorporating the real measurement at time step k into the model. In detail, the estimates of all variables of time step k are updated using a weighted average, with more weights being given to estimates with higher certainty. Output is an estimate of all variables at time step $k$, even when the real measurement of the variables at time step k is incomplete (missing variables), noisy, or completely missing. The estimate of the variables for time step $k$ is then set as new current knowledge state. The algorithm is recursive. It can run in real time, using only the present input measurements and the previously calculated state and its uncertainty matrix; no additional past information is required.


Figure 43. Basic concept of the Kalman filtering approach. Adapted from Aimonen (2011).

### 3.6.4 Data Fusion implementation using VITO's enhanced Kalman-Filtering approach

The enhanced Kalman filter algorithm used in our workflow was introduced in its basic implementation by Sedano et al. (2014) and is available as open source software in the pktools (http://pktools.nongnu.org). Kempeneers et al. (2016) applied their Kalman filter implementation successfully on PROBA-V data. The pktools implementation of the Kalman-filtering approach for data fusion shows a slight difference compared to the original approach stated by Sedano et al. (2014): instead of interpreting the trend between two time stamps of the pixels of the higher resolution data, which are falling within one pixel of the coarse resolution data (9 PROBA-V UTM 100 m pixels are within one PROBA-V UTM 300 m pixel at the same geographic location), as independent variables of the coarse pixel trend (spectral unmixing), all nine fine resolution pixels get the same trend between two time stamps as the estimated trend between these two time stamps in the coarse resolution data (Kempeneers et al., 2016). Neglecting spatial context can be an issue, in particular for heterogeneous landscapes (Kempeneers et al., 2016), but has a considerable influence on the processing speed. In the first version of the CGLS-LC100 workflow this short-coming was accepted, but for the second version we decided to follow closer the original implementation idea by Sedano et al. (2014). Therefore, we implemented a spectral unmixing step for the coarse resolution dataset in order to compute nine independent trajectories for the fine resolution. Figure 44 illustrates this concept and shows that the reconstructed pixel values of the fine resolution image between time step $k-1, k$, and $k+1$ are independent from each other (heterogeneous surfaces are reconstructed more correctly).


Figure 44: Principle of VITO's enhanced Kalman-filtering approach for EO data fusion based on the pktools algorithm. Note: Image adapted from Kempeneers et al. (2016).

Document-No.
Issue: I3.4

CGLOPS1_ATBD_LC100m-V3
Date: 31.03.2021

Input data into the Kalman-Filtering processing step are the PROBA-V UTM TOC-MC5 100 m time series for the four reflectance bands (see section 3.4.3) and the PROBA-V UTM TOC-MC5 300 m gap-filled time series for the four reflectance bands (see section 3.6.2). The four reflectance bands are processed pixel-wise and independently from each other, but multi-core processing is used to process all data simultaneously. The Kalman-filter algorithm implementation runs in three modes which have to be processed sequentially. The following algorithm explanation is done for the time series of one 100 m pixel in the red reflectance band following Figure 45. The upper part in Figure 45 shows the 100 m TOC-MC5 time series in orange where the orange line shows existing valid time series steps and the orange shaded area shows missing time series steps. Moreover, the existing gap-filled TOC-MC5 time series for the corresponding 300 m pixel is shown in blue.

In the first run, the forward mode of the Kalman filter implementation, the pixel values for all time steps in the fine resolution data (PROBA-V UTM TOC-MC5 100m) are predicted for the current date meaning the prediction model is built up by walking forward through the time steps and executing also the spectral unmixing of the corresponding 300 m time series in the forward direction. The third image in Figure 45 shows the Kalman forward reconstruction together with the calculated uncertainty, where the valid data of the 100 m time series steps is shown as orange line, the spectral unmixed values of the 300 m time series as orange dashed line and the reconstructed 100 m time steps in blue with the uncertainty in green. The uncertainty increases, the longer the data gaps are in the 100 m time series as well as the further away we are from the starting time step.

In the second run, the backward mode, the algorithm is running back in time and predicts all pixel values for all time steps in the fine resolution data. Therefore, the spectral unmixing of the corresponding 300 m time series is done in the backward direction. The fourth (lower) image in Figure 45 shows the Kalman backward reconstruction together with the calculated uncertainty, where the colour schema in this image is the same as this of the forward run.

In the third run, the smooth mode, the pixel information and calculated uncertainties for each time step in the time series profiles of the forward and backward modes are combined by calculating weighted averages taking the uncertainties of the modes into account, which result in more reliable estimates with lower uncertainties. The second image in Figure 45 shows the output of the forward mode in blue, of the backward mode in orange and the final smooth result in green. The final result with its combined uncertainty is also shown in the upper image in Figure 45 where the final smooth time series are introduced in green (the uncertainty in light green, respectively) into the original TOCMC5 100 m time series shown in orange where time steps were missing.


100 m pixel time series and corresponding 300 m


Kalman fusion (forward, backward and smooth mode)


Kalman forward reconstruction with uncertainty


Kalman backward reconstruction with uncertainty


Figure 45: Example of 100 m time series reconstruction of one pixel using VITO's improved Kalman filter. Upper image shows original 100 m timeseries in orange (orange line shows existing values and orange shading shows missing time steps), the corresponding 300 m time series in blue and the final 100 m gap filling in green plus uncertainty in light green areas; second image shows the whole reconstructed 100 m time series for forward (blue), backward (orange) and smooth mode (green); third image shows the 100 m forward reconstruction with uncertainty; and fourth image shows the 100 m backward reconstruction with uncertainty.

Figure 46 is used to visualize this whole Kalman filling process for a bigger image area to show the power of the algorithm to successfully reconstruct also heterogeneous areas. The following explanations therefore explain Figure 46 and the colour coding. The left side shows the full PROBAV UTM TOC-MC5 100 m tile for a selected MC5 time step (here 136, mid December 2015) and the lower part the corresponding gap-filled 300 m time step. In the lower left, the upper graph shows the full time series for the selected pixel in 100 m in orange, for the corresponding 300 m pixel in blue, and the final reconstructed missing 100 m time series steps in green (same colour schema as in Figure 45 upper image). The lower graph in the lower left shows the nine 100 m forward unmixed time series for the corresponding 300 m pixel of the selected 100 m one. This helps to see if the selected 300 m pixel has a more homogeneous or heterogeneous surface (in this example minimum eight of the underlying sub-pixels are having the same trends where one sub-pixel shows trend differences). Overall, these four images (two tiles overviews on the left and the two graphs in the lower left) give a spatial and temporal overview for a selected 100 m pixel (corresponding 300 m pixel, respectively). Next, the marked red area in Figure 46 gives a zoom in for the 100 m and 300 m tile as well as the $3 \times 3$ pixel area of the selected pixel and the corresponding 300 m pixel (right part of image) at the selected time step showing that for the selected time step no 100 m time series data is available. The marked blue area visualize the forward and backward spectral unmixing results of the 300 m data into 100 m for the zoomed-in tile as well as for the $3 \times 3$ pixel area of the selected pixel and selected time step showing that the 300 m pixel is spatially not fully homogeneous. Finally, the magenta marked area visualize the results of the three Kalman filtering modes for the zoomedin tile as well as for the $3 \times 3$ pixel area of the selected pixel and time step showing that the missing 100 m time series data could be successfully reconstructed without introducing extrema.

Output of the data fusion step is a continuous, gap-free 5-day time series product in fine resolution - called PROBA-V UTM TOC-MC5 100 m Kalman-filled - for the four reflectance bands.















of pixel
of pixe
into 9 100 m pixels


100m Kalman reconstructed time step of 9 pixels out of one 300 m pixel
Figure 46: Visualization of data fusion with the improved Kalman filter for one $3 \times 3100 \mathrm{~m}$ pixel block in PROBA-V UTM ARD+ datasets. Note: red rectangles show the zoom into selected area of 100 m and corresponding 300 m tile, and the selected 300 m pixel for selected time series step (Note: 100 m has no valid data); blue rectangles show forward (top) and backward (center) unmixing of 300 m time step into 100 m resolution for zoom and pixel, respectively; magenta rectangles show the 100 m reconstructed time series steps out of 300 m pixel for zoom and $3 \times 3$ pixel area, respectively (Note: shown is forward (top), backward (center) and smooth (bottom) mode of Kalman filter)

### 3.6.5 Data Fusion post-processing

The last step in the data fusion workflow is the post-processing in which the Kalman filled PROBAV UTM TOC-MC5 100 m data is screened for consistency and introduced outliers (Figure 41). This screening is needed since long concurrent gaps in the 100 m time series profiles can lead to extreme values (artefacts) when also in the 300 m data not enough observations are available. The gap filling with harmonized HMC5 data in the 300 m time series profiles can guide the Kalman-Filter over these gaps, but still can lead to unexpected predictions.

The post-processing step can be split into three sub-steps which have to be processed sequentially for each reflectance band, but all four PROBA-V UTM reflectance bands can be processed simultaneously via multi-core processing. The sub-steps are:

- data clamping: the predicted reflectance values for the four PROBA-V UTM bands have to be clamped to the maximal reflectance range from 0 to 1 . That is needed since introduced extreme values can be higher or lower than the maximal reflectance values range.
- Re-injection: Since the Kalman-filled time series is completely generated from the estimated/predicted reflectance values for each time step, the original measured values (PROBA-V UTM TOC-MC5 100 m time series values) have to be re-injected into the time series profiles. This process is needed in order to only fill the missing time steps in the measured data with the predicted reflectance values of the Kalman-filtering approach.
- outlier detection: introduced artefacts in the time series does not have to be always extreme values (issues in spectral unmixing can create spatial artefacts), therefore the estimated reflectance values have to be checked against their neighbours and overall for consistency. We use the madHANTS algorithm for the temporal outlier screening (see section 3.4.2). All identified outliers are flagged as "no data" values in the time series profiles.

Output of the data fusion post-processing is a consistent PROBA-V UTM TOC-MC5 100 m time series (called PROBA-V UTM TOC-MC5 100 m data fused) for the TOC reflectance data in the blue, red, NIR and SWIR wavelength regions. An example for tile X18Y06 before and after applying the data fusion approach is shown in Figure 47.

Also note: Since the data fusion improves the data quality, we decided that this processing upgrades the PROBA-V UTM ARD+ archive into the PROBA-V UTM ARD++ archive, which gives temporal cleaned, consistent and dense 5-day image stacks for all global land masses at 100 m resolution (Note: 300 m data is not existing in the ARD++).


Figure 47: top: full PROBA-V UTM TOC-MC5 100 m tile for red reflectance channel (areas with missing data are shown in white); bottom-left: PROBA-V UTM TOC-MC5 100 m data fused image; bottom-right: zoom in to full PROBA-V UTM 100 m resolution of left image over red box shown in top image (the area in the red box had no observations at all for that time stamp and shows, after data fusion, consistent image data). Example of PROBA-V UTM tile in Nigeria on 2016-03-06.

### 3.7 Ancillary Dataset Products Generation

In order to include ancillary datasets into the classification/regression part of the CGLS-LC100 product workflow (Figure 2), the external datasets have to be pre-processed. This pre-processing includes one or all of the following steps:

- Projection of the input dataset to the UTM coordinate system;
- Resampling of the input dataset to 100 m ;
- Tiling of the dataset into the Sentinel-2 tiling grid and naming of tiles conform to Sentinel-2 (see section 3.2.3).

If the ancillary datasets have been only processed with these steps, then they are called "ancillary data masks" in the following sections. These ancillary masks are:

- The land/sea mask created from the input datasets shown in section 3.2.2. Note: since this dataset was already used as a buffered version in the PROBA-V UTM ARD generation (see section 3.3), the physical processing of this ancillary data mask was already done during the PROBA-V UTM generation and is here stated for completeness.;
- The Potential Permanent Snow and Ice mask (PPSIM) created from the input datasets shown in section 3.2.6. Note: since this dataset was already used in the PROBA-V UTM ARD generation (see section 3.3), the physical processing of this ancillary data mask was already done during the PROBA-V UTM generation and is here stated for completeness.;
- The Circum-Polar Arctic Vegetation Map (CAVM) (see section 3.2.9 for more details).;
- The DEM (see section 3.2.12 for more details).;
- The yearly precipitation and temperature products (see section 3.2.5 and 3.2.4 for more details).;
- The Ecoregion layer generated from the RESOLVE2017 dataset (see section 3.2.8 for more details).

Some ancillary dataset have been processed on top with additional algorithms and/or processing steps and therefore are called "ancillary dataset products" in the following sections. Overall three ancillary dataset products are used in the CGLS-LC100 product:

- Global biome clusters product;
- Yearly Built-up cover fraction product;
- Yearly WetProducts permanent and seasonal water cover fractions products which are generated from the Global Surface Water (GSW) datasets together with generation of a wetland indication mask as well as water and vegetation occurrence layers..

All ancillary datasets are not only input into the LC map generation process, but also in the training data generation and the classification/regression section of the workflow. The following three subsections explain the generation of the ancillary dataset products in more detail.

### 3.7.1 Global Biome Clusters product

In order to group EO data for faster processing or adaptation of algorithms to specific regions, we use a Global Biome Cluster layer (Figure 48). The term "biome cluster" hereby refers to a geographic area which has similar bio-geophysical parameters and, therefore, can be grouped for processing. In other words, the biome cluster layer can be seen as an ecological regionalisation which outlines areas of similar environmental conditions, ecological processes, and biotic communities (Coops et al., 2018). There are already several global regionalisation layers existing, e.g. Ecoregions 2017 global dataset (Dinerstein et al., 2017), Geiger-Koeppen global ecozones after Olofsson update (Olofsson et al., 2012), Global ecological zones for FAO forest reporting with update 2010 (FAO, 2012). But several tests in the CGLS-LC100 workflow have shown that the existing layers did not provide the required global and continental classification accuracy. These finding go along with Coops et al. (2018) who stated that "Most regionalisations are made based on subjective criteria,
and cannot be readily revised, leading to outstanding questions with respect to how to optimally develop and define them".


Figure 48: Global stratification into 73 biome clusters to improve regional classification/regression in the CGLS-LC100 workflow.

Therefore, we decided to develop a customized ecological regionalisation layer which performs best with the given PROBA-V RS data and the specifications of the CGLS-LC100 product. It groups spectral similar areas and helps to optimize the later classification/regression to regional patterns. Input into the layer creation were well-known existing datasets (see section 3.2.8) which were combined, re-grouped and advanced based on prior CGLS-LC100 classification results and local mapping knowledge of the workflow developer. To ensure that this layer is clearly separable from other existing regionalisations and not mistakenly interpreted as an eco-region layer, we decide to call it "biome clusters" layer. The following steps outline the global biome clusters layer generation:

- Spatial union of Ecoregions 2017 dataset (Dinerstein et al., 2017), Geiger-Koeppen dataset (Olofsson et al., 2012) and Global FAO eco-regions datasets (FAO, 2012);
- Regrouping and dissolving by using experience from first global CGLS-LC100 mapping results and subjective mapping experience of the developer;
- Refinement of the biome clusters in the High North latitudes via incorporation of a Global tree-line layer (Alaska Geobotany Center, 2003);
- Manual improvement of borders between biome clusters to reduce classification artefact by using a DEM (see section 3.2.12) and mapping experience from previous projects and continental test runs;
- Usage of a global land/sea mask (see section 3.2.2), the Sentinel-2 tiling grid (see section 3.2.3) and PROBA-V imaging extent ([PROBA-V User Manual]) to extend the borders of the biome clusters into the sea to make sure that also small islands on the coastline are correctly processed.

When developing a regionalisation, the definition of the clusters and the boundaries that delineate them in time and space is the key challenge. Since a full automated processing was not possible, the quality of these layers depends on several subjective decisions. Overall, we separated 73 global biome clusters from each other which can be seen in Figure 48. The continental assignment of the biome clusters can be seen in Table 6. The final vector layer is directly used in several processing steps (e.g. sections 3.9, 3.10.3), but also a raster mask following the Sentinel-2 tiling and naming was generated for the final CGLS-LC100 discrete map generation (see section 3.14).

Table 6: Global biome cluster by continents.

| continent | \#biome clusters | Biome cluster names |
| :---: | :---: | :---: |
| North America | 16 | ZA01 - ZA16 |
| South America | 6 | ZA17 - ZA22 |
| Europe | 10 | ZB01, ZB06 - ZB11, ZB18, ZB20, ZB21 |
| Africa | 15 | ZB22 - ZB36 |
| Asia | 19 | ZB02 - ZB05, ZB12 - ZB17, ZB19, ZB37 - ZB44 |
| Oceania | 7 | ZB45 - ZB51 |

### 3.7.2 Yearly Built-up cover fraction product generation

The detection of urban structures is one of the most challenging tasks in LC classification processes. Instead of handling this task in our workflow by an own process, we decided to incorporate existing knowledge. The Built-up cover fraction product used in the CGLS-LC100 workflow is a combination of the DLR's World Settlement Footprint (WSF) dataset for 2015 (Marconcini et al., 2017, Marconcini et al., in review) (see section 3.2.10) showing an overall accuracy between 83.27 - $89.33 \%$ depending on the "built-up' definition (Marconcini et al., in review) and the OpenStreetMap (OSM) layers (see section 3.2.10). Since the WSF layer is only provided for the reference year 2015, the OSM layers are used to provide the yearly updates in the Built-Up cover fraction layer. Note: since the OSM layers are produced by a citizen science project, the crowd-mapping application OpenStreetMap also has disadvantages. Mainly the quantity and quality of the data hold in OSM is depending on the quality checks of other citizen. Volunteers' motivations and their spatial literacy skills have a considerable impact on the quality of these layers. Several issues with the built-up class and cover fractions in the CGLS-LC100 products can be explained by this fact.

Since the original WSF 2015 raster layer is provided in 10 m spatial resolution, we first resampled and projected the dataset to 10 m pixels in the UTM system. Then the OSM vector data layers for the corresponding reference year are imprinted (rasterized). For the rasterization of the OSM datasets, the OSM database is filtered for buildings, roads, train tracks, and traffic related structures (airports, ports, etc.) and line layers are buffered. Finally, a moving window approach is applied on the 10 m raster dataset in order to transform the binary mask into built-up cover fractions at 100 m spatial resolution (count of number of 10 m pixel in corresponding $100 \times 100 \mathrm{~m}$ box). The final step is then the tiling and naming of the dataset following the Sentinel- 2 tiling convention. Output are yearly Built-Up cover fraction datasets which are used in the yearly map generation.

### 3.7.3 Yearly WetProducts generation

The detection of surface water bodies is another challenging task in LC classification processes. Again, we decided to incorporate existing knowledge in the way of using external data and algorithms to form an own PROBA-V UTM 100 m water product (called WetProducts) (Figure 49). Main input are the yearly GSW seasonal and permanent water cover fraction layers which are derived from JRC's GSW water seasonality 2015-2019 layers (Pekel et al., 2016). The GSW water seasonality layer gives information on the number of months water was detected for the concerned year (see section 3.2.11). In the pre-processing step (Figure 49) the GSW layer is first projected to UTM and resampled to 10 m spatial resolution and subsequently this raster layer is used to calculated the two GSW cover fraction layers on 100 m resolution via a moving window approach, i.e. GSW seasonal water cover fraction layer and GSW permanent water cover fraction layer. The pre-processing ends with the tiling and naming of the layers following the Sentinel-2 tiling grid.

The original JRC's GSW water seasonality layer shows an overall accuracy between 95.79 to $97.01 \%$ where permanent water detection shows the better accuracy with up to $99 \%$ and seasonal water detection a slightly lower accuracy with 75 - 98 \% (Pekel et al., 2016). Since this accuracy could not reached with an own algorithm based on multi-spectral PROBA-V EO data alone, we only implemented an add-on step to detect wetlands and calculate the yearly water and vegetation occurrence layers. This additional processing step is based on the water detection algorithm by Bertels et al. (2016).

Bertels's water body detection algorithm for the single reference years uses the PROBA-V UTM ARD++ derived VI data layers HUE, VALUE and NDVI for a per pixel water and vegetation occurrence detection (see section 3.8.2.1). The three output layers, i.e. water occurrence layer, vegetation occurrence layer and the wetland indication mask, are calculated by applying specific thresholds (empirically defined, cross-checked against reference points and Pekel et al., 2016) on HUE, VALUE and NDVI. Finally the Land/Sea mask (see section 3.2.2) as well as GSW cover fraction layers are used to clean-up the three generated layers.

In detail, the actual water body detection is done only on pixels which have HUE and VALUE greater than zero and the NDVI may not have a 'No data' value. Subsequently, the following checks are performed:
(i) NDVI must be less than 0.32 , or the NDVI is greater than or equal to 0.32 and its VALUE is less than or equal to 0.11 ; and
(ii) its HUE must be greater than 120 or its VALUE must be greater than 0 and less than or equal to 0.14 .

This allows the calculation of a per pixel water occurrence which is the percentage that water was detected for the reference year to all valid observations in this year. The generation of the vegetation occurrence is done similar by calculating the percentage that vegetation was detected (NDVI for an observation above 0.32 ) for the reference year to all valid observations in this year. Note: At least 10 valid observations for a pixel has to exist in the reference years to get a valid pixel result.

The detection of wetland is done on the detected seasonal water pixels. If the cover fraction is at least $25 \%$ and the total NDVI (sum of all NDVI values of the valid detections in the reference year) is over a pre-defined threshold (17.5) the pixel is re-defined as wetland.

To separate the final output of the independent processing line, we bundled the output layers under the term "WetProducts" which consists of:

- WetProducts permanent water cover fractions layer;
- WetProducts seasonal water cover fractions layer;
- WetProducts wetland indication mask,
- WetProducts water occurrence layer, and vegetation occurrence layer.


Figure 49: General overview of the yearly WetProducts generation. Besides the PROBA-V UTM ARD++ derived VI data products, i.e. HUE, VALUE and NDVI, two ancillary layers are used as input for the WetProducts generation. Note: Numbers in the upper left corner of data container indicate the number of layers - here number of 5-day observations - for one reference year used to create the yearly product.

### 3.8 Epoch-Wise Metrics Generation

### 3.8.1 Overview

Figure 50 gives an overview of the metrics generation section in the CGLS-LC100 product workflow (Figure 2). The term "metrics" hereby refers to the case of LC classifications of time series to
quantitative indicators or proxies which can be used to describe the time series or time series behaviour. These metrics are the main input into the classification / regression algorithm in order to produce LC maps or cover fraction layers, respectively.

Overall three processing steps are needed for each epoch and all epochs (epoch2015 - epoch2019) are processed sequentially. The first step is the calculation of 6 additional vegetation indices out of the four reflectance bands of the PROBA-V UTM ARD++ archive for each time series step of an epoch (216 time steps). Moreover, a Hue, Saturation, Value colour system (HSV) transformation and spectral area indices calculation is performed in the first processing step. During the second processing step, the metrics extraction, the descriptive statistics for the time series profiles of the 10 layers (six calculated vegetation indices (VI), two layers of the colour transformation, and two spectral area indices) are extracted for each pixel. Furthermore, the harmonic metrics as attributes for the overall level and seasonality of the time series are extracted for the 10 layers. Also a textural metric showing the uniformity of a pixel compared to its neighbouring pixels ( $3 \times 3$ box) is generated. Additional metrics are extracted from ancillary datasets. In the last processing step, the postprocessing, the overall 183 generated metrics for each PROBA-V UTM 100 m pixels are combined into one container.


Figure 50: General overview of the metrics generation section for each epoch in the CGLS-LC100 product workflow. Note: datasets marked in red are re-used in following sub-steps. Numbers in the upper left corner of data containers indicate the number of layers (time steps in the ARD++ where 216 5-day time steps represent the three year epoch or product layers in the resulting metrics).

### 3.8.2 Pre-Processing

The pre-processing step of the metrics generation section can be subdivided into three sub-steps (Figure 50): in the first, additional vegetation indices for each time step in the PROBA-V TOC-MC5 100 m data fused time series (PROBA-V UTM ARD++ archive) are generated for each epoch; in the second, the HSV colour transformation is carried out; and in the third the spectral area indices are calculated.

### 3.8.2.1 PROBA-V UTM Vegetation Indices generation

We decided to use only already established and successfully proven VI's in the CGLS-LC100 product workflow. Since the multi-spectral PROBA-V satellite provides only four spectral bands (see section 3.2.1), the list of usable VI's is limited. Overall, we identified six VI's in the literature:

1. Normalized Difference Vegetation Index (NDVI) using the red and the NIR reflectance bands (Tucker, 1979). The NDVI is one of the oldest, most well-known, and most frequently used VI. The combination of its normalized difference formulation and use of the highest absorption and reflectance regions of chlorophyll make it robust over a wide range of conditions. It can, however, saturate in dense vegetation conditions when the Leaf-Area-Index (LAI) becomes high. The value of this index ranges from -1 to 1 . The common range for green vegetation is 0.2 to 0.8 . Equation 10 shows the formula for the NDVI:

$$
\begin{equation*}
N D V I=\frac{N I R-\text { red }}{N I R+\text { red }} \tag{Eq. 10}
\end{equation*}
$$

, where NIR and red refers to the reflectance value of the corresponding PROBA-V reflectance bands.
2. Enhanced Vegetation Index (EVI) using the blue, red, and NIR reflectance bands (Huete, et al., 2002). The EVI was developed to improve the NDVI by optimizing the vegetation signal in high LAI regions by using the blue reflectance to correct for soil background signals and reduce atmospheric influences, including aerosol scattering. This VI is therefore most useful in high LAI regions, where the NDVI may saturate. The value of this index ranges from -inf to inf. The common range for green vegetation is 0.2 to 0.8 , and value over +2 or under -1 are already extremes. Equation 11 shows the formula to calculate the EVI for PROBA-V:

$$
\begin{equation*}
E V I=2.5 * \frac{N I R-\text { red }}{(N I R+6 * \text { red }-7.5 * \text { blue }+1)} \tag{Eq. 11}
\end{equation*}
$$

, where NIR, red and blue refers to the reflectance value of the corresponding PROBA-V reflectance bands.
3. Structure Intensive Pigment Index (SIPI) using the blue, red, and NIR reflectance bands (Blackburn, 1998). The SIPI is designed to maximize the sensitivity of the index to the ratio of bulk carotenoids (for example, alpha-carotene and beta-carotene) to chlorophyll while decreasing sensitivity to variation in canopy structure (for example, leaf area index).

Increases in SIPI are thought to indicate increased canopy stress (carotenoid pigment). Applications include vegetation health monitoring, plant physiological stress detection, and crop production and yield analysis. The value of this index ranges from -infinity to infinity. The common range for green vegetation is 0.8 to 1.8 , and value over +2 or under 0 are already extremes. Equation 12 shows the calculation formula for PROBA-V:

$$
\begin{equation*}
\text { SIPI }=\frac{\text { NIR }- \text { blue }}{\text { NIR - red }} \tag{Eq. 12}
\end{equation*}
$$

, where NIR, red and blue refers to the reflectance value of the corresponding PROBA-V reflectance bands.
4. Normalized Difference Moisture Index (NDMI) using the NIR and the SWIR reflectance bands (Gao, 1996). The NDMI is a commonly used index used to detect vegetation water content and therefore also an important indicator for water stress. Where SWIR reflects changes in both the vegetation water content and the spongy mesophyll structure in vegetation canopies, the NIR reflectance is affected by leaf internal structure and leaf dry matter content but not by water content. Their combination removes variations induced by leaf internal structure and leaf dry matter content, improving the accuracy in retrieving the vegetation water content. The value of this index ranges from -1 to 1 . The common range for green vegetation is 0.0 to 1.0. Equation 13 shows the calculation formula:

$$
\begin{equation*}
N D M I=\frac{N I R-S W I R}{N I R+S W I R} \tag{Eq. 13}
\end{equation*}
$$

, where NIR and SWIR refers to the reflectance value of the corresponding PROBA-V reflectance bands.
5. Near-Infrared reflectance of vegetation (NIRv) using the red and the NIR reflectance bands (Badgley et al., 2017). The NIRv is one of the newest Vl's and is the product of total scene NIR reflectance and the normalized difference vegetation index (NDVI). Badgley et al. (2017) state that "...from a physical perspective, NIRv represents the proportion of pixel reflectance attributable to the vegetation in the pixel.". The formula adapted for PROBA-V data is shown in Equation 14:

$$
\begin{equation*}
N I R_{V}=\left(\frac{N I R-\text { red }}{N I R+\text { red }}-0.08\right) * \text { red } \tag{Eq. 14}
\end{equation*}
$$

, where NIR and red refers to the reflectance value of the corresponding PROBA-V reflectance bands.
6. Angle at NIR (aNIR) using the red, NIR and SWIR reflectance bands (Khanna et al., 2007). This index is suited to detect dry plant matter in the presence of soil and green vegetation. It ranges in value from 0 to $2 \pi$ radians (Equation 15).

$$
\begin{equation*}
a N I R=\arccos \left(\frac{a^{2}+b^{2}+c^{2}}{2 a b}\right) \tag{Eq. 15}
\end{equation*}
$$

Where $a, b$ and $c$ are Euclidean distances in a cartesian plane where $x$ represents the wavelength in micrometers and $y$ the reflectance value following Equations 16 -18:

$$
\begin{align*}
& a=\sqrt{(\text { red }-N I R)^{2}+\left(W L_{\text {red }}-W L_{\text {NIR }}\right)^{2}}  \tag{Eq. 16}\\
& b=\sqrt{(N I R-S W I R)^{2}+\left(W L_{\text {NIR }}-W L_{S W I R}\right)^{2}}  \tag{Eq. 17}\\
& c=\sqrt{(\text { red }-S W I R)^{2}+\left(W L_{\text {red }}-W L_{S W I R}\right)^{2}} \tag{Eq. 18}
\end{align*}
$$

, with $W L_{B x}$ representing the position of the central wavelength in nm of a reflectance band (see Table 3) and red, NIR and SWIR refers to the reflectance value of the corresponding PROBA-V reflectance bands.

Input data into the calculation of the VI's are the PROBA-V TOC-MC5 100 m data fused time series (PROBA-V UTM ARD++ archive) for all four reflectance bands. The calculation of the VI's is carried out independently for each MC5 time step in the time series profiles and follows the Equations 10 to 18. Thus, all six VI's and time steps can be processed simultaneously. Output are VI time series profiles which are named:

- PROBA-V UTM MC5-NDVI 100 m ,
- PROBA-V UTM MC5-EVI 100 m,
- PROBA-V UTM MC5-SIPI 100 m,
- PROBA-V UTM MC5-NDMI 100 m ,
- PROBA-V UTM MC5-NIRv 100 m , and
- PROBA-V UTM MC5-aNIR 100 m


### 3.8.2.2 PROBA-V HSV colour transformation

The HSV colour transformation is an approach that transforms the RGB (Red, Green, Blue) colour space into the HSV (Hue, Saturation, and Value) colour space which decouples chromaticity and luminance. The HSV colour space is commonly used in image processing. It is a nonlinear transformation of the RGB colour space using Equations 19, 20 and 21 presented in Figure 51.

Current remote sensing applications which use the HSV colour transformation are e.g. surface water detection algorithms. Therefore, by analyzing the distribution of the pixels in the Hue - Value space, the pixels can be classified as "water" and as "no water" based on thresholds (Pekel et al., 2014; Bertels et al., 2016). Also the water product generation algorithm (called WetProducts) within the CGLS-LC100 product workflow uses the HSV colour transformation (see section 3.7.3). Also, we decided to use the time series of Hue and Value as indices in the metrics extraction process.

Input in the HSV colour transformation are the red, NIR and SWIR reflectance bands of the PROBAV TOC-MC5 100 m data fused time series (PROBA-V UTM ARD++ archive) with R mapped to SWIR band, G to NIR band and B to the RED band (Figure 51). The Hue (H) is defined as the dominant wavelength of the perceived colour. It is the visual perceptual property corresponding to the categories called yellow, blue, green, etc. Hue is considered as an angle between a reference line and the colour point, going from $0^{\circ}$ to $360^{\circ}$. The Saturation (S) is defined as the degree of purity of the colour and may be intuitively considered as the amount of white mixed in a colour. This component represents the radial distance from the cone center going from 0 to 1 . The nearer the point is to the center, the lighter is the colour. The Value (V), which is a brightness approximation, represents the height of the axis of the HSV cone, going from 0 to 1 . This axis describes the gray levels.

Again, the calculation of Hue and Value is carried out independently for each MC5 time step in the time series profiles. Output are Hue and Value time series profiles which are named:

- PROBA-V UTM MC5-HUE 100 m , and
- PROBA-V UTM MC5-VALUE 100 m .


$$
\begin{equation*}
V=\max (R, G, B) \tag{Eq. 19}
\end{equation*}
$$

$S=\frac{V-\min (R, G, B)}{V}$

$$
H=\left\{\begin{array}{l}
\left(60^{\circ} * \frac{G-B}{V-\min (R, G, B)}+360^{\circ}\right) \bmod 360^{\circ} \text { if } V=R  \tag{Eq. 20}\\
60^{\circ} * \frac{B-R}{V-\min (R, G, B)}+120^{\circ} \quad \text { if } V=G \\
60^{\circ} * \frac{R-G}{V-\min (R, G, B)}+240^{\circ} \quad \text { if } V=B
\end{array}\right.
$$

Figure 51: The HSV colour space and the formulas for transforming the RGB colour space into the HSV colour space. Note: V=Value, S= Saturation, H=Hue, R=red, G=green, B=blue.

### 3.8.2.3 PROBA-V Spectral Area Indices

Spectral area indices, in this case Area-Under-The-Curve and Normalized-Area-Under-The-Curve, are integrated measurement of a measurable effect or phenomenon. It is used as a cumulative measurement as a means to compare features or better spectral reflectance signatures by one single scalar (Figure 52). The advantage against a two or multi-band VI is that these indices take all reflectance bands into account and therefore allow a full comparison between reflectance signatures.


Figure 52: Spectral Area Indices explained. The sum of the FWHM reflectance (reflectance over the full band width) for all reflectance bands can be used as single scalar to compare spectral signatures.

Overall we implemented two spectral area indices in the CGLS-LC100 workflow:

1. Area Under the Curve (AUC) using the reflectance values of all four PROBA-V reflectance bands and the Full Width at Half Maximum (FWHM) for each reflectance band as measurement of the band width. Since the AUC represent the sum of the reflectance over all bands, it shows the overall reflectance of a surface (Figure 52). Equation 22 shows the used formula:

$$
\begin{equation*}
A U C=\sum_{x_{B}}^{n} x_{B} * F W H M_{x_{B}} \tag{Eq. 22}
\end{equation*}
$$

, where $x_{B}$ is one of the PROBA-V bands blue, red, NIR and SWIR and $F W H M_{x_{B}}$ is the bandwidth of that band (Table 3).
2. Normalized Area Under the Curve (NAUC) uses the same methodology as the AUC, but normalize all reflectance values to the minimum reflectance before the area calculation. The normalization allows the removal of offsets between spectral signatures and therefore compares the real features instead of just the overall reflectance of a spectral signature. Therefore, spectra of same materials can be easier grouped and identified. Equation 23 shows the calculation in the CGLS-LC100 workflow:

$$
\begin{equation*}
N A U C=\sum_{x_{B}}^{n}\left(x_{B}-M\right) * F W H M_{x_{B}} \tag{Eq. 23}
\end{equation*}
$$

, where $x_{B}$ is one of the PROBA-V bands blue, red, NIR and SWIR and $F W H M_{x_{B}}$ is the bandwidth of that band (Table 3), and $M=\operatorname{argmin}($ blue,red, NIR,SWIR) is the minimum value across all the bands for each pixel.

Input data into the calculation of the spectral area indices are the PROBA-V TOC-MC5 100 m data fused time series (PROBA-V UTM ARD++ archive) for all four reflectance bands. The calculation of the indices is carried out independently for each MC5 time step in the time series profiles and follows the Equations 22 to 23 . Thus, all two spectral area indices and time steps can be processed simultaneously. Output are spectral area indices time series profiles which are named:

- PROBA-V UTM MC5-AUC 100 m , and
- PROBA-V UTM MC5-NAUC 100 m.


### 3.8.3 Metrics extraction

The metrics extraction, the second processing step in the metrics generation section (Figure 50), generates the main input data for the LC classification and regression process. The metrics condense the information content of the time series profiles for each pixel. This information reduction (in detail, all valid observations of the epoch spanning three years of data are condensed in a few variables) allows so the analysis of complex inter-annual behaviours of vegetated and non-vegetated land areas via easy to interpretable and comparable variables. Input for the metrics extraction step are the six PROBA-V UTM MC5 Vl's time series profiles at 100 m for each epoch generated in the pre-processing step (see section 3.8.2.1), the PROBA-V UTM MC5-HUE and MC5-VALUE time series profiles at 100 m of the HSV colour transformation pre-processing step (see section 3.8.2.2), and the two PROBA-V UTM MC5 spectral indices time series profiles at 100 m (see section 3.8.2.3). So overall 10 input layers, for which the harmonic, descriptive and textural metrics are separately extracted.

Several methods can be applied to derive metrics information from RS time series. In the simplest approaches, time is just treated as an identifier and multiple observations are directly used as input features for the classification, but that would increase dramatically the data volume for the classifier. An easy solution to condense the information content is that statistical metrics are derived for specified intervals of the time series and then used as input for the classifier. An additional solution is the fitting of models to the time series profiles and using either the model parameters for classification directly, or deriving statistical metrics from these models with their harmonized time series profiles. The advantage of using harmonic model parameter next to descriptive statistics in a supervised classification approach has been shown by Eberenz et al. (2016). Within the CGLSLC100 product workflow we decided to combine the methods. We derive descriptive statistics from the time series profiles, but also fit a harmonic model through the time series profiles and use the

| Document-No. | CGLOPS1_ATBD_LC100m-V3 | © C-GLOPS1 Lot1 consortium |
| :--- | :--- | ---: |
| Issue: 13.4 | Date: 31.03 .2021 | Page: 90 of $\mathbf{1 5 2}$ |

model parameters as attributes. The model generation and descriptive statistic calculation has to be processed sequentially, therefore we used multi-core processing to process all pixel of an image and all 10 input layers simultaneously for each epoch.

### 3.8.3.1 Harmonic metrics

The first sub-step is the generation of the harmonic metrics. Thus, a harmonic model is fitted through all values of the time series of an epoch for each pixel for all 10 input layers separately. The harmonic model is again based on the HANTS algorithm using a Fourier transformation (see section 3.4.2 for a detailed HANTS description). We use 3 frequencies above the zero-frequency in the HANTSmodelling process in order to capture annual and inter-annual variations of the time series. The seven model parameters of the harmonic model - meaning the phases and amplitude of the identified most significant frequencies in the time series profiles - are used as metrics for the overall level and seasonality of the time series.

Output of this sub-step are overall 70 harmonic metrics representing the temporal behaviour of the three year epoch ( 7 model parameters representing the phases and amplitudes of the Fourier transformation for each of the 10 input layers) which are directly used in the classification/regression approach. Moreover, for each of the input bands/layers, a harmonized time series profile in 5 -day time steps is generated using the identified phases and amplitudes which are then used as additional input in the descriptive metrics generation sub-step. Figure 53 shows an example of a harmonized time series profile of the NDVI.


Figure 53: Example for a harmonized time series profile. The result of the HANTS model is shown in red dots, where the blue dots show the real observations.

### 3.8.3.2 Descriptive metrics

Descriptive statistics are quantitative parameters which provide simple summaries about the time series or parts of the time series. We decided to use the following 9 parameters in our analysis:

1. mean: for identifying the central value of a time series;
2. standard deviation: to quantify the amount of variation within the pixel values of a time series profile;
3. sum: to quantify the overall quantity of the time series in a given time length. Note: for this metric the harmonized time series profile is used in order to be comparable between pixels;
4. median: to identify the "middle" value which separates the higher half of the pixel values in the time series from the lower one;
5. $10^{\text {th }}$ percentile: to identify the pixel value below which $10 \%$ of the pixel values of the time series can be found;
6. $90^{\text {th }}$ percentile: identify the pixel value below which $90 \%$ of the pixel values of the time series can be found;
7. $10^{\text {th }}-90^{\text {th }}$ percentile range: to identify the interval in which $80 \%$ of the pixel values of a time series are located;
8. Month of the Year in which the $10^{\text {th }}$ percentile of the time series shortened to the reference year appears (MOY_q10): to identify the month of the year when the minima values appear. Note: for this metric we use the harmonized time series profile cut to the reference year (72 time steps) in order to have always enough data steps; and
9. Month of the Year in which the $90^{\text {th }}$ percentile of the time series shortened to the reference year appears (MOY_q90): to identify the month of the year when the maxima values appear. Note: for this metric we use the harmonized time series profile cut to the reference year (72 time steps) in order to have always enough data steps.

The 9 descriptive statistics are independently calculated for each of time series profiles of the 10 input layers. In order to guarantee comparability, the sum metric has to be calculated on the harmonised time series profiles of the input bands/layers (see section 3.8.3.1 - reconstructed 5 -day time series profiles using the phases and amplitudes of the identified frequencies in the original time series). This is needed since not all pixels have otherwise the same amount of valid time step values and this can lead to false interpretation.

Output of this sub-step are overall 90 descriptive metrics which are directly used in the classification/regression approach.

### 3.8.3.3 Textural metrics

In order to describe the uniformity of a pixel compared to its neighbours, we created a textural metric. Input in the calculation is the median (descriptive statistic, see 3.8.3.2) parameter of the time series profiles for all 10 layers, where the textural parameter is calculated for each independently. The textural metrics can be seen as an additional descriptive metric since it is generated by calculating the standard deviation of a $3 \times 3$ moving window for each pixel (in detail: we calculated the standard

| Document-No. | CGLOPS1_ATBD_LC100m-V3 | © C-GLOPS1 Lot1 consortium |
| :--- | :--- | ---: |
| Issue: 13.4 | Date: 31.03 .2021 | Page: $\mathbf{9 2}$ of $\mathbf{1 5 2}$ |

deviation of the median values of the centre raster cell and its eight neighbours). Thus, low values show that a pixel is in a homogeneous area compared with the neighbouring pixels, where high values show a more heterogeneous land cover.

Output of this sub-step are overall 10 textural metrics (one for each of the 10 input layers) which are directly used in the classification/regression approach.

### 3.8.4 Post-Processing

The last processing step in the metrics generation section (Figure 50), the post-processing, handles all calculated metrics. In the last step, the topographic metrics of the DEM (see section 3.2.12) and temperature metrics (see section 3.2.4) are extracted and combine with the precipitation metric (see section 3.2.5), water and vegetation occurrence metrics (see section 3.7.3.), and geographical metric as well as all harmonic, descriptive and textural metrics.

The separation of classes only by metrics from spectral data is often imperfect, and the integration of ancillary data into the classification/regression can significantly improve the results (Santos, 2011). Especially the integration of topographic data, including elevation, slope and aspect, into land cover classification are assumed to boost the separability via the fact that several land cover classes have topographical preferences (Santos, 2011). Santos (2011) even showed that improvement of $10 \%$ in overall accuracy for the classification with the integration of topographical data over the one that depended on spectral data only are possible.

Therefore, we decided to integrate 4 topographic metrics (i.e. elevation, slope, aspect, and purity) into the CGLS-LC100 classification/regression. Input data is the external DEM (see section 3.2.12). In a pre-processing step the global DEM was reprojected into the UTM coordinate system, resampled to 100 m via a bilinear filter and tiled into Sentinel-2 equivalent tiles (see section 3.7 for more details). Next, the slope, aspect and "purity" for each 100 m pixel was calculated and stored as separate datasets. Where slope is the degree of inclination of a feature relative to the horizontal plane, shows the aspect the compass direction that a slope faces. The direction a slope faces (so the aspect) can affect the physical and biotic features of the slope, known as a slope effect (Bennie et al., 2006). The purity metric should not be confused with the common Topographic Position Index (TPI). The purity metric is calculated by identifying the highest absolute difference between the centre pixel and its eight neighbours in a $3 \times 3$ box around this centre pixel. The lower the purity value the more homogeneous is the area around a pixel, and the higher the purity value the more heterogeneous is the area around the pixel.

Moreover, we added 5 metrics describing the ground temperature regime for a pixel over the year. Input is the yearly temperature data (see section 3.2.4) for which the median, $10^{\text {th }}$ percentile, $90^{\text {th }}$ percentile, Month of the Year in which the $10^{\text {th }}$ percentile temperature occurs, and Month of the Year in which the $90^{\text {th }}$ percentile temperature occurs metrics are calculated. Next, the sum of the yearly precipitation (see section 3.2.5) is used as metrics to distinguish yearly anomalies. Also a geographical metric was added to help the classifier to group areas of similar vegetation (ecoregions) even within a biome cluster (see section 3.7.1 for more information to biome clusters) - therefore
the ecoregion ID of the RESOLVE2017 dataset was used (see section 3.2.8). Finally, the water and vegetation occurrences generated during the WetProducts generation (see section 3.7.3) are added as metrics.

The huge amount of metrics data is easier to handle when all metrics are bundled into one file meaning that the metrics can be accessed as bands in an image. Therefore, we decide to generate Virtual Raster datasets (VRT) out of the single metrics files. The VRT file is an Extensible Markup Language (XML) encoded file which stores the location of the metric files on the server for each pixel and can also hold additional metadata for each file.

Overall, 183 metrics -7 harmonic metrics plus 9 descriptive metrics plus 1 textural metric for the 10 time series profiles ( 6 vegetation indices, 2 colour transformed bands, and 2 spectral area indices) plus 13 additional metrics - are combined in the VRT file. This VRT file is then input in the classification/regression section of the automated processing chain. Figure 54 shows a visualization of some selected metrics.


Figure 54: Examples of the 183 derived metrics for a tile in Africa for the epoch2015. a) texture metric (the lighter the colour the more homogeneous is the pixel compared to its surrounding pixels), b) standard deviation of the Structure Intensive Pigment Index for the vegetation season (the greener the colour the higher is the SIPI change within the epoch), c) sum of the Enhanced Vegetation Index for the epoch (the redder the colour the more vegetation).

### 3.9 Training Data Generation for BaseYears

Training data has been collected through the Geo-Wiki engagement platform. A specific branch of Geo-Wiki (http://geo-wiki.org/) was developed for collecting reference data at the required resolution and grid (PROBA-V UTM 100 m pixels). It shows the pixels to be interpreted on top of Google Earth and Bing imagery, where each pixel is further subdivided into 100 sub-pixels of approximately 10 m x 10 m each. Using visual interpretation of the underlying very high resolution imagery, experts (a group of people trained by International Institute for Applied Systems Analysis (IIASA) staff) interpret
each sub-pixel based on the land cover type visible, which includes trees, shrubs, water objects, arable land, burnt areas, etc. This information is then translated into different legends using the UN LCCS (United Nations Land Cover Classification System) as a basis. A detailed description of the training data collection and sampling layout can be found in Lesiv et al. (2017) and [CGLOPS1_TrainingDataReport_LC100m]. In short: The 'core' distribution of sample sites is systematic, with the same distance between sample sites, which is approximately 35 km . However, land cover data are not collected at every sample site as the frequency depends on the heterogeneity of land cover types by region and availability of valid PROBA-V UTM 100 m imagery. In addition to the "core" training data set, we collected more training sample sites for rare classes (e.g. wetland) and for areas of low accuracy (e.g. Africa).

In total, the experts have classified 178,405 unique locations, including 28,000 new locations in the Northern Hemisphere through integrating Digital Globe imagery to overcome gaps in the free Microsoft Bing and Google imagery. The quality of the data has been checked by revisiting locations that were either inter- or intra- land cover class outliers from a remote sensing perspective. This screening was done by creating a rule matrix of calculating the (Root Mean Square Error) RMSE between the metrics of all ground reference data. This analysis is comparable with the data screening in the training data optimization step (see section 3.10.3), but instead of directly removing suspicious training points, the training points were check by visual interpretation. Moreover, additionally to the rule matrix check, also disagreeing training points with Australian land cover map and Corine Land Cover for Europe were identified. Training points that were wrongly classified or those where it was impossible to identify the land cover class by visual interpretation, were removed. Final training dataset consists of approximately 168,000 sample sites (Figure 55).

Important to note is that the training data for collection 3 is collected and quality checked for the reference year 2015. A yearly update is not feasible, therefore we decided that training data updates will be carried out in 5 -yearly intervals. Meaning, every 5 years (i.e. 2015, 2020, 2025,..) all training locations will be revisited and the labels/description will be updated if needed. Moreover, new training locations will be added based on user feedback or the global validation assessment. Since the training data is fixed to a reference date, we call the years in which training data was collected BaseYears (i.e. BaseYear2015 for the training data used in generation of collection 3 of the CGLSLC100 product). The influences on the classification/regression approach are explained in detail in section 3.10.

Since the classification/regression is conducted in geographical subsets (see section 3.10.1) to optimize the classification/regression process to regional pattern, in the next step the training data was split up in those subsets. Therefore, we used the generated 73 global biome clusters (see section 3.7.1). Even the biome clusters were optimized to show no border artefacts, we used a 1 degree buffer around each biome cluster for training point selection to reduce possible border effects during the classification/regression processing step. Moreover, the DDI mask for the corresponding reference year (see section 3.5) was applied to ensure that all training points have sufficient time series data (e.g. the DDI of epoch 2015 is used to check the training data of BaseYear2015). In detail, training points in locations with a DDI value of 0 , meaning that no PROBA-V UTM time series
data for the epoch in which the BaseYear falls was available for this location, is removed from further processing.

In the next step, the PROBA-V UTM time series metrics (see section 3.8.4) for the epoch corresponding to the training data BaseYear are extracted for each 100 m raster pixel corresponding to the geographic location of each training point. Output is a complete training data set for each biome cluster containing the land cover ground reference information (discrete class and cover fraction percentages for each discrete class) and the spectral information of the EO time series data condensed in the form of the metrics for the BaseYear.


Figure 55: Training dataset for the BaseYear2015 used in the CGLS-LC100 workflow showing the $\sim 168 \mathrm{~K}$ training locations together with their sampled land cover class at 100 m .

### 3.10 Еросн-WISE CLASSIFICATION / Regression

### 3.10.1 Overview

In order to adapt the classification/regressor algorithms to regional patterns, the classification/regression of the epoch data is carried out in data subsets ( 73 biome cluster subsets) using the created global biome clusters layer (see section 3.7.1). Important to note is that since training data is only available in certain BaseYears (see section 3.9), the classification/regression is carried out by using pre-trained models which are updated only in following BaseYears. In detail, training data is only collected/updated in 5 -yearly intervals (starting with 2015). Therefore, only classification/regression models can be generated in these BaseYears and are applied on all epochwise classifications/regressions up to the next BaseYear (i.e. the classification/regression models for the BaseYear2015 are used in collection 3 for the epochs 2015 - epoch 2019).

| Document-No. | CGLOPS1_ATBD_LC100m-V3 | © C-GLOPS1 Lot1 consortium |
| :--- | :--- | ---: |
| Issue: 13.4 | Date: 31.03 .2021 | Page: $\mathbf{9 6}$ of $\mathbf{1 5 2}$ |

Prior to the classification/regression of the EO data (Figure 2), additional pre-processing and optimization steps of the training data are needed for BaseYears in order to generate the classification/regression models (BaseYearModels). The BaseYear training datasets provided for each biome cluster are screened for inter-class outliers for each biome cluster independently. Next, the best features (metrics) for each biome clusters and classification/regression scenario are identified and ranked via an all relevant feature selection process in order to optimize the classification/regression algorithms for each biome cluster (see 3.10.3). Moreover, during the optimization phase also the classification algorithm parameter for the "pure" scenario are optimized for each biome cluster using a combined random and grid search approach. Note: this optimization of the classifier and regressor to the biome cluster areas can produce small regional artefacts on the border between two biome clusters. Detailed checks for these border artefacts were conducted on global scale for collection 2 of the CGLS-LC100 product and removed via optimization of the biome cluster boundaries in collection 3 of the product.

Finally, the land cover classification and regression to estimate the cover fractions for each pixel is conducted scenario-based for each epoch separately using the BaseYearModels. A wide range of algorithms are available as classifiers in LC mapping approaches. Due to its relatively simple parameterization, computation efficiency, and high accuracy, we decided to use the random forest (RF) classification and regression algorithm (Ho, 1995) within the CGLS-LC100 product workflow. Moreover, the RF algorithm was successfully applied to derive LC from seasonal models and metrics of PROBA-V time series (Eberenz et al., 2016).

### 3.10.2 The Random Forest Approach

Random forests or random decision forests are supervised machine learning methods for classification, regression and other tasks. It is called "supervised" learning because the algorithm knows the correct classification answers for the input training data. The algorithm iteratively makes predictions on the training data classes and is corrected by the input data - learning stops when the algorithm achieves an acceptable level of performance.

The random forest algorithm operates by constructing a multitude of decision trees during the training phase and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees during the classification/regression phase. In easy words, the RF algorithm grows many classification trees based on bootstrapped samples of the input training data. To classify a new pixel from the input dataset (metrics), the input dataset is put down each of the trees in the forest. Each tree gives a classification or regression result plus a "vote" for that class. The RF algorithm chooses the classification/regression result having the most votes (over all the trees in the forest) as the final result for that pixel (Figure 56). Advantages of the RF method over other classifiers includes the ability to accommodate many predictor variables (metrics), as well as the fact that it is a non-parametric classifier which does not assume any underlying distribution in the training samples (Eberenz et al., 2016).


Figure 56: The Random Forest classifier principle. Image by Niraj (2016).

### 3.10.3 Training data and classifier/regressor optimization for BaseYears

Since we use biome clusters to subset the EO data, the RF classifier/regressor algorithms should be optimized for each biome clusters training dataset prior the prediction phase. This optimization ensures that the problem - unclassified pixels within each ecozone - is optimally solved by the RF algorithm giving the different training data classes distribution in each biome cluster. Note: as already mentioned in section 3.10.1, this optimization of the classifier and regressor to the biome cluster areas can produce small regional artefacts on the border between two biome clusters.

The training data used in the CGLS-LC100 workflow was selected by experts (see section 3.9). For the classification of "pure" endmembers (see section 3.10.4) we had to make sure that inter-class confusion is minimized for training points showing a $100 \%$ cover fraction in a class of the ground reference data (training points). Therefore, a ground reference data analysis is performed individually for each biome cluster-based training data set of the BaseYears to exclude confusing data from each dataset. This task is accomplished in several sub-steps:

1. The input metrics of all training points within one biome cluster are scaled to make all features, individual metrics, comparable. For this approach, the combination of the metrics for one training point is interpreted as bands in a spectra (as in hyper-spectral RS);
2. A rule matrix is created by calculating the RMSE between the spectra (function of the scaled metric values over the amount of metrics) of all training points within the biome cluster. This results in a square $\mathbf{N}$ by $\mathbf{N}$ matrix, where $\mathbf{N}$ is the total number of training point spectra in the biome cluster (Figure 57);


Figure 57: Rule matrix calculated for the total of N training points which belong to four different LC classes. For each training point the RMSE value is calculated against all the other training points, i.e. c 11 with $\mathrm{c} 11, \mathrm{c} 12, \mathrm{c} 13$ till c 4 n , subsequently c 12 with $\mathrm{c} 11, \mathrm{c} 12, \mathrm{c} 13$ till c 4 n and so on.
3. The rule matrix is used to exclude inter-class outliers. E.g. when the median RMSE value of "training point $\mathbf{c} 1_{1}$ calculated with the remaining spectra of its own class (Class 1 )" is greater than the median RMSE value of "training point $\mathbf{c 1} 1_{1}$ calculated with all the spectra of the other classes" then training point $\mathbf{c} 1_{1}$ could be excluded from the dataset. This check is repeated for each ground reference spectrum in each class. Output is a list of suspicious inter-class outliers and a calculated impact score.
4. All inter-class outliers (training points within the biome cluster) with an impact score over 50\% are removed from further processing.
Note: This outlier removal step is only done for training points showing a $100 \%$ cover fraction and not for all training points, otherwise too many training points on class boundaries would be removed. But the identified outliers are removed for all classification/regression scenarios.

In order to overcome the Hughes phenomenon (Hughes, 1968) due to the fact that the 183 metrics include a high information redundancy, an all-relevant feature selection approach is used to select those best metrics for each classification/regression scenario within the biome cluster-based training datasets which solve the classification/regression problem for this scenario best. Note: this optimization step is carried out for all classification/regression scenarios and biome clusters but only for the BaseYears. In a first sub-step the metrics are sorted by their separability. In detail, the best metrics (features) to separate the LC classes for each classification/regression scenario are identified for each biome cluster-based training dataset by using the inter-class outlier cleaned datasets. The best metrics are defined as those which have the highest separability compared to the other metrics. For each metric, the separability is calculated by comparing the metric values of one LC class to the metric values of another LC class. All LC class combinations are handled. The separability is calculated following Equation 24:

$$
\mathrm{S}_{\mathrm{AB}}=\frac{\mathrm{nA}+\mathrm{nB}}{\operatorname{tot} \mathrm{~A}+\operatorname{tot} \mathrm{B}}
$$

Eq. 24

Here $\mathbf{S}_{\mathrm{AB}}$ is the separability between LC class $A$ and LC class $B$ with $\mathbf{n A}$ and $\mathbf{n B}$ the number of training points metrics from resp. LC class $A$ and LC class $B$ that fall outside the common values of LC class $A$ and $B$. Finally, totA and totB are the total number of training point metrics for resp. LC class $A$ and $B$. The final training point separability is the mean of the calculated separabilities for all LC class combinations. And, subsequently, the best metrics are those which have the highest final training point separability. In the second sub-step, the sorted metric bands are reduced since the identified and sorted best metrics can still be redundant (carry the same information content). Therefore we use an all-relevant feature selection approach. The advantage of an all-relevant feature selection approach compared to the usual minimal-optimal selection approach is that "it tries to find all features carrying information usable for prediction, rather than finding a possibly compact subset of features on which some classifier has a minimal error" (Kursa, 2017). We use the Boruta package by Kursa et al. (2010) in our automated workflow. The Boruta algorithm is as a wrapper around a RF classification algorithm. It iteratively removes the metrics which are proved by a statistical test to be less relevant than random probes. Output of this processing step is a list for each classification/regression scenario within each biome cluster-based training dataset of the minimum 50 best metrics which are non-redundant. These best metrics datasets are then used in the next optimization step as well as in the final classifier/regressor training. Note: The best metrics lists per biome cluster and classification/regression scenario for input into the classifier/regressor training as well as the final metrics importance lists after the training are available on request.

The last optimization step is the hyper-parameter search for each biome cluster-based training dataset of the "pure" classification scenario. In detail, the optimal parameters to identify the best machine learning RF model created by the training data for each biome cluster are called hyperparameter. Since these hyper-parameters cannot be learned directly from the training data in the standard model training process, we used a combined grid and random search with a five folded cross-validation to identify the optimal model parameter for each biome cluster-based training data set of the "pure" scenario (Note: hyper-parameter optimization is only used in the "pure" scenario and all other scenarios run with the standard parameters).

### 3.10.4 Epoch-Wise Scenario-based Classification using the BaseYearModels

For the supervised classification, the RF classifier implemented in the scikit-learn package was used (Pedregosa et al., 2011). This implementation has the advantage that the ensemble of decision trees classifiers which build up the random forest classifier are combined by averaging their probabilistic prediction (sklearn, 2017), and not let each classifier vote for a single class (Breiman, 2001). The RF classification is conducted for each of the 73 biome cluster independently using the biome cluster layer (see section 3.7.1) to split up the input data (generated metrics for the PROBA-V UTM time series - see section 3.8.4), and using the biome cluster-specific generated BaseYearModels. The BaseYearModels are pre-generated using the biome cluster-specific training datasets, selected best
metrics and hyper-parameters. Moreover, the DDI (see section 3.5) for each training point showing the availability and quality of the PROBA-V UTM input data for the BaseYear is used as weight factor during the RF classification model training. The DDI allows to further group the training data by their quality, meaning that training points with a high DDI have a higher influence in the RF model training than training points with a low DDI.

Next to the classification results showing the discrete class for each pixel, also the predicted class probability for each pixel is generated. Overall three Random Forest classification scenarios for each biome cluster and epoch with different settings are carried out (Note: meaning that for each scenario an own BaseYearModel was generated and then applied epoch-wise):

1. "pure class" scenario (CL1): in this scenario, only training points with a cover percentage of 100 \% in the bare, herbaceous vegetation, shrub, forest, herbaceous wetland, snow and ice, moss and lichen or agriculture class are used. This scenario can be interpreted as endmember selection (extreme sample reduction in terms of purity) and classification. A pixel's metrics/spectral profile is matched to the metric/spectral signature of a specified land cover type (endmember). By incorporating the predicted class probability, the pixels with "pure" land cover classes can be identified (e.g., a pixel classified as forest with $90 \%$ predicted class probability would mean that the classifier is to $90 \%$ certain that the pixel is forest with a $100 \%$ of forest cover).
2. "discrete class" scenario (CL2): in this scenario, all training points which were classified as forest, shrub, grassland, agriculture or bare are used.
3. "forest type" scenario (CL4): in this scenario, only training points with a forest cover percentage over $15 \%$ and a valid forest type attribute (e.g. evergreen needleleaf, evergreen broadleaf, deciduous needleleaf, deciduous broadleaf) are used. The resulting map is therefore a forest type map and later used to subdivide the forest class.

The scenario-based classification was implemented in order to overcome weakness in the all-in-one classification scheme. Instead of using all training classes at once, the scenario based classification produces thematic maps which can be fused by analysing the class probabilities and LC extents of the different scenarios.

### 3.10.5 Epoch-Wise Scenario-based Regression using the BaseYearModels

A novelty of the CGLS-LC100 product is the generation of vegetation continuous fields - also called cover fractions - that provide proportional estimates for vegetation cover for all 9 discrete base classes. The input data are the cover fractions collected for all training points which are used in a Random Forest regression.

The RF regression is conducted for each of the main land cover types (forest, shrub, herbaceous vegetation, agriculture, moss \& lichen, bare) and for each of the 73 biome cluster independently using the biome cluster layer (see section 3.7.1) to split up the input data (generated metrics for the PROBA-V UTM time series - see section 3.8.4). Again, the regression models are pre-generated by using the biome cluster-specific generated training datasets for the BaseYears, the selected best
metrics and hyper-parameters, and the DDI (see section 3.5) for each training point is used as weight factor during the RF regression model training. These regression BaseYearModels are then applied on the epoch-wise data.

An important change compared to algorithm version 2 is the improved adaptation of the RF regressor implemented in the scikit-learn package. Pre-studies had shown that the implemented scikit-learn RF regressor operates suboptimal in the $0 \%$ and $100 \%$ cover fraction ranges. Therefore, we implemented Masiliunas et al. (2020, in review) approach for a 3-step regression in the CGLS-LC100 workflow. The three steps thereby consists of two RF pre-classifications for the $0 \%$ and $100 \%$ cover fraction range of a scenario and is followed by a RF regression to determine the $1-99 \%$ cover fractions. Moreover, for further calculations we generate an internally used RF quality indicator for each scenario by calculating the standard deviation of all single regression results in the trees of a forest for each pixel. Therefore, a low standard deviation indicates that the RF regressor has a higher probability that the estimates cover fraction is correct, where a high standard deviation indicates a low probability. Since the regression models are trained biome cluster-wise, these internal quality indicators can be only used to compare the quality within its biome cluster.

Overall six regression scenarios for each biome cluster are carried out (Note: meaning that for each scenario an own BaseYearModel was generated and then applied epoch-wise):

1. Forest (R1): the forest cover percentages of all training points are used in the regression model;
2. Shrub (R2): the shrub cover percentages of all training points are used in the regression model;
3. Herbaceous vegetation (R3): the herbaceous vegetation and herbaceous wetland cover percentages of all training points are used in the regression model;
4. Bare (R4): the bare cover percentages of all training points are used in the regression model;
5. Agriculture (R5): the agriculture cover percentages of all training points are used in the regression model;
6. Moss/lichen (R6): the moss and lichen cover percentages of all training points are used in the regression model.

### 3.11 Еросн-Wise Pre-Products Generation

### 3.11.1 Overview

After the scenario-based RF regression of the six main vegetation classes (forest, shrub, herbaceous vegetation, agriculture, moss \& lichen, and bare) for each epoch (see section 3.10.5), two additional processing steps are needed in order to generate the 10 epoch-wise cover fraction pre-products (9 discrete base classes and the seasonal water cover fraction layer). The two processing steps are outlined in Figure 58. The epoch-wise pre-products are accomplished by the three classification and classification probabilities datasets (see section 3.10.4). Note: These datasets are called epoch-wise pre-products since no temporal cleaning and consistency checks were applied (see section 3.12. for more details).

Issue: I3.4

Main input data layers come from the epoch-wise scenario-based classification results, the scenariobased regression results together with their corresponding internally quality indicators (standard deviations), some products from earlier processing steps, and the ancillary data masks as well as ancillary dataset products (see section 3.7).


Figure 58: General overview of the epoch-wise pre-products generation section in the CGLS-LC100 product workflow.

In the first processing step - the pre-processing-, the DDI statistics are analysed to generate the PROBA-V UTM 100 m NoData mask for each epoch. Therefore, the " $\mathrm{N}^{\circ}$ of valid observations for the PROBA-V UTM 100 m time series" layer of the DDI statistics (see section 3.5.3) is transferred into a binary mask showing areas with no valid observations over the full time series (raster value zero) of an epoch. The PROBA-V UTM 100 m NoData mask is then added to the other input data sets and gets part of the post-regression processing step (Figure 58). The next sub-section explains the postregression in more detail (see section 3.11.2).

Output of the epoch-wise pre-product generation section are the scenario classification results (classification and classification probability), the epoch-wise cover fractions consisting each of ten thematic cover fraction layers, the internally used cover fraction quality layers for six of the thematic layers, and the PROBA-V UTM 100m NoData mask.

### 3.11.2 Epoch-wise cover fraction pre-products generation

In the regression post-processing, the epoch-wise scenario-based regression results of the six thematic classes are post-processed and checked for consistency. Moreover, the cover fraction layers for the four remaining thematic classes (built-up, snow\&ice, permanent water bodies, and seasonal water bodies) are generated out of the ancillary masks and dataset products (see section 3.7) as well as classification results from selected scenario-based classifications (see section 3.10.4). The detailed processing is shown in Figure 59 and Figure 60.

The five input dataset groups for each epoch are in detail (see also Figure 59):

- Three products of the epoch-wise scenario-based classifications (see section 3.10.4):
- PROBA-V UTM 100 m CL1 ("pure" class scenario) classified map;
- PROBA-V UTM 100 m CL1 ("pure" class scenario) class probabilities; and
- PROBA-V UTM 100 m CL2 ("discrete class" scenario) classified map;
- The six PROBA-V UTM 100 m regression result (cover fraction) layers for the classes forest, shrub, herbaceous vegetation, agriculture, moss \& lichen, and bare (see section 3.10.5);
- The six regression quality indicator (cover fraction standard deviation) layers of the PROBAV UTM 100 m regression results showing the standard deviation of the regression decision tree voting (see section 3.10.5);
- The yearly permanent snow mask generated during the PROBA-V UTM ARD generation (see section 3.3.2);
- The PROBA-V UTM 100 m NoData mask generated during the pre-processing step;
- The CAVM, PPSIM, and land/see mask as part of the ancillary masks (see section 3.7); and
- The WetProducts permanent water cover fraction layer and the WetProducts seasonal water cover fraction layer (see section 3.7.3) as well as the built-up cover fraction layer (see section 3.7.2) as part of the ancillary dataset products.


Figure 59: Rule block 1 to 4 for the epoch-wise cover fraction pre-products generation - generation of snow, built-up, permanent and seasonal water cover fraction and cleaning by CL1 and CAVM.

The cover fraction pre-products and the internally quality indicators are calculated in six sub-steps for each epoch independently. First, the permanent snow\&ice cover fraction layer is generated by combining the yearly permanent snow \& ice mask obtained during the PROBA-V UTM generation, the 'snow \& ice' classified areas from the CL2 map and the 'snow \& ice' classified areas from the CL1 map for which the classifier probability was higher than $95 \%$. Then the cumulative areas are restricted by the PPSIM in order to reduce snow \& ice misclassifications in the High North (mainly ice covered lakes during the winter months were misclassified as permanent snow \& ice in the scenario-based classifications). The permanent snow \& ice cover fraction layer is then cleaned by the NoData mask and the land/sea mask (rule I in also Figure 59).

Second, the epoch cover fraction layers for built-up, permanent water and seasonal water are generated. Therefore, the WetProducts permanent water cover fraction layer, WetProducts seasonal water cover fraction layer, and the built-up cover fraction layer are cleaned against the final permanent snow\&ice cover fraction layer. This cleaning is needed to avoid cover fractions above $100 \%$ in permanent snow \& ice areas, since the snow\&ice class has the highest priority. Afterwards the WetProducts permanent water cover fraction layer of the epoch is cleaned against the built-up cover fraction layer for the corresponding epoch. And finally the WetProducts seasonal water cover fraction layer is cleaned against the WetProducts permanent water cover fraction layer, since the
permanent water is also interpreted as $100 \%$ water occurrence next to the spatial extent. Note: the WetProducts seasonal water cover fraction layer is not cleaned against the built-up cover fraction layer, since the term seasonal water describes the temporal domain and not in the spatial.

Next, in sub-step 3 the scenario-based regression results for the epoch are updated by the class information of the classified pixels from CL1 which have a classification probability of $100 \%$. The $100 \%$ classification probability ensures that the selected pixels with their class are "pure" endmembers, or in other words: have a $100 \%$ cover fraction for the classified discrete class. This sub-step was introduced after additional test showed that still the $100 \%$ cover fraction range is underestimated in the RF regressions (Masiliunas et al., 2020). Subsequently, in sub-step 4 the CAVM is used to clean-up vegetation mis-classifications at high latitudes. Here, we make sure that the biomes above the High Latitude tree line do not show trees (including a transition buffer to account for tree line shifts due to global warming).

Sub-step 5 is the most important step in the whole post-regression section (Figure 60). Here we apply a regression normalization to ensure that the total cover fraction for a pixel cannot exceed the $100 \%$. For this, the normalization uses the regression standard deviations of the six scenario-based retrieved cover fractions to calculate conversion factors for each pixel following Equation 26. These conversion factors are then used to normalize the cover fraction percentages of forest, shrub, herbaceous vegetation, agriculture, moss \& lichen, and bare up to the maximum cover fraction for this pixel which takes into account reduction due to snow\&ice, built-up or permanent water portions in this pixel (see Equation 25).

```
\(100 \%\) - builtup cover fraction - snow cover fraction - permanent water cover fraction
                                    \(=\)
    tree cf. * tree conversion + shrub cf. * shrub conversion factor +
grass cf. * grass conversion factor + bare cf. * bare conversion factor +
    crop \(c f . *\) crop conversion factor + moss \(c f . *\) moss conversion factor
```

        conversion factor \({ }^{\oplus}=\frac{\text { regression std.dev. } \odot}{\text { } \text { regression std.dev. } \odot^{\ominus}}\)
    Eq. 26
    $\circledast$ calculated per cover type

Finally sub-step 6 deals with outputting the results, in detail the tree, shrub, herbaceous vegetation, crops, moss\&lichen and bare cover fractions and their corresponding standard deviation as well as the permanent snow\&ice, built-up, permanent and seasonal water cover fractions are cleaned by the PROBA-V UTM 100m NoData mask and the land/sea mask.


Figure 60: Rule block 5 and 6 for the epoch-wise cover fraction pre-products generation - cover fraction normalization and cleaning by the land/sea mask.

### 3.12 Concept of Temporal Cleaning in the CGLS-LC100 product workflow

### 3.12.1 Overview

The annual land cover maps are generated using an 'epoch', where an epoch consists of 3 years of input data (one year before and one year after). The classification/regressor model is generated on the BaseYear (Base reference year), in this case 2015 for collection 3 - therefore this land cover map is called the Base map (processing mode of the generated land cover map).

The machine learning model (BaseModel) is re-used in the subsequent generation of the change maps. As shown in Figure 61, the actual land cover can change over time and it depends on when the CGLS-LC100 processing workflow is run (point in time of the year) if a change compared to the previous land cover map can be already detected. Therefore, the annual land cover map updates have two modes. This mode depends on the availability of remote sensing data after the possible change that may have occurred.

For a timely delivery of yearly maps, Near-Real-Time mode is generated. In this mode, the updated maps are delivered at the end of the $1^{\text {st }}$ quarter of the following year. Therefore for the land cover map for a specific year (e.g., 2019) not enough input data is available to fill the full epoch (one year before and one year after). Hence changes occurred during the latter part of that year can be detected but the new class cannot yet be identified (depicted in red crosses in Figure 61). This mode is called Near-Real-Time.

The second mode is called Consolidated mode. In this mode, the NRT mode is regenerated but this time with data of full year after the end of the specific year. With such additional data, it is possible to identify the land cover classes after change which may have occurred at a late stage of the year.


Figure 61: Processing mode of land cover change maps. Note: the rectangles represent the land cover class of a certain pixel in the map. Near-real time mapping doesn't allow the direct identification of land cover changes for certain classes which is indicated by red pattern.

The result of this 'mode' processing implies that during the beginning of a new year (Yr), two maps are provided (the latter map consolidates the detected changes of the NRT map and hence substitutes the former map):

1. Near-Real-Time (NRT) map of previous year (Yr-1)
2. Consolidated (CONSO) map of the year before (Yr-2)

To minimise the number of changes in land cover marked by the classifier, which is often due to small variations in the input data that sway the vote in a different direction and do not indicate actual land cover change, a land cover change method was used to constrain such spurious change. Various methods for detecting breaks in a time series are available, such as the BFAST family, LandTrendr, CCDC, npphen, in addition to simple methods such as post-classification image differencing. The BFAST family was selected for this purpose, as it is well-established and applied widely to different areas of the world, as well as on various classes of land cover, ranging from semiarid regions of Australia (Watts and Laffan, 2014) to various ecozones in Canada (Fang et al., 2018) and to tropical agriculture and rainforests in the Amazon region (Arantes et al., 2017). In addition, this algorithm was deemed more suitable due to its ability to identify all breaks within a time series, as well as to indicate the timing of change within the year. It also has comparatively high computing

| Document-No. | CGLOPS1_ATBD_LC100m-V3 | © C-GLOPS1 Lot1 consortium |
| :--- | :--- | ---: |
| Issue: 13.4 | Date: 31.03 .2021 | Page: 108 of 152 |

efficiency, with the possibility of running some of the algorithms with GPU acceleration (von Mehren et al., 2018).

For detecting potential changes in land cover, a univariate time series of a vegetation index is needed. The algorithm works by segmenting the time series of the vegetation index into stable periods, and as such it can detect changes that have a big effect on the time series of the vegetation index, such as flooding (any land cover class getting replaced by water), but it is more difficult to detect changes between vegetation (e.g. change between shrubs and herbaceous vegetation) and it cannot detect breaks within gradual change (e.g. forest regrowth, i.e. succession from herbaceous to shrub to trees, is seen as a single process of continuous increase in vegetation with no breaks in between).

MODIS was chosen as the data source for the time series, as it provides a consistent long-term archive of surface reflectance with high temporal frequency (Figure 62). The length of the time series and its consistency is very important so that the algorithm can learn what variability is considered normal at each location separately. The high temporal frequency is also vital in areas with high cloud cover, as it increases the chance of having cloud-free imagery and thus being able to tell whether and when a break happened. Lastly, break detection is a very computationally intensive procedure, and thus the lower spatial resolution allows for global land cover change detection at a reasonable time. Unfortunately, it also means that change is detected at a coarser resolution ( 250 m ) than the map ( 100 m ) itself, therefore small areas of change or change in heterogeneous areas may be missed.

## MODIS time series



Figure 62: Example of MODIS time series of NIRv over an area north of San Pablo Bay, California (longitude: -122.454, latitude: 38.12883) which is now an outlet of the bay. MODIS time series spans over a decade of imagery with a single sensor and allows identifying stable and unstable conditions, in this case agriculture before 2016 and water after.

Finally, to improve the coarse change detection Minimum Mapping Unit (MMU) from BFAST, we integrated a secondary change detection algorithm based on the Hidden Markov Model which is applied on the full time series of land cover maps in order to improve the temporal consistency of each pixel and reduce the number of illogical and spurious changes (Gong et al., 2017; Abercrombie \& Friedl, 2016).

### 3.12.2 Break Detection using BFAST

### 3.12.2.1 Introduction to BFAST (general concept)

The BFAST family of break detection algorithms are unsupervised change detection methods that use a univariate vegetation index time series with irregular timesteps as input (Figure 63). From this input, several components are computed: a monotonously increasing trend component, a nominal season component that indicates the season that each observation belongs to, and harmonic components that are sine and cosine functions for each harmonic order. The order parameter controls the number of these harmonic orders to be computed, whereas the sbins parameter controls the number of bins to divide the seasons in a year into. Optionally, an autoregressive (lag) term can be added, also for seasonality. The formula parameter allows choosing which components are then used to fit the response variable, i.e. the input vegetation index time series. A seasonal bin parameter controls how many seasons are estimated each year for the season component.

Before the break detection algorithm is run, there is an option to run a structural change test, which tests whether there is at least one break in the time series altogether, following structural change theory (Zeileis, 2006). This additional test is an optimisation that allows avoiding costly runs of the full algorithm in areas where no change is expected altogether. However, it may also result in missing some change.


Figure 63: BFAST family fundamentals: a time series is subdivided into several segments and a linear (in this example harmonic 3rd order) model is fit onto each segment to detect a break. The magnitude of the break (change in model parameters between segments) is also reported.

### 3.12.2.2 BFASTON

For the CONSO maps, the BFASTON algorithm was used. This algorithm is based on a piecewise linear regression and can detect all breaks in a time series, except for a period at the start and end of the time series. The duration of this period, and the minimum distance in time between two subsequently detected breaks, is controlled by the $h$ parameter, which denotes the minimal segment size in the piecewise linear regression. Since the CONSO map is a yearly-updated map, the $h$
parameter was set to one year at minimum (in some cases it could be higher due to cloud cover causing data gaps).

The algorithm works by estimating the optimal allocation of breaks within the time series to minimise the residual sum of squares (RSS) of the piecewise linear regression, and this algorithm is performed iteratively from zero up to the maximum possible number of breaks depending on the $h$ parameter, using the algorithm of Bai and Perron (2002). Finally, the optimal number of breaks is estimated using an information criterion such as BIC or LWZ, implemented in the R package "strucchange" (Zeileis et al., 2003).

### 3.12.2.3 BFAST Monitor

As BFASTON cannot detect breaks at the end of the time series, for the NRT map the BFAST Monitor algorithm (Verbesselt et al., 2012) was used instead. This algorithm is made for detecting whether there is at least one break at the end of the time series. This detection is achieved by segmenting the input time series into two parts: the history period, which is used to determine the normal variability of the vegetation index in the past; and the monitoring period, which is the timeframe that is tested for a break by checking if it deviates from what is expected from the history period. In other words, a model is fitted onto the history period and is then extrapolated into the monitoring period, and a test is made to determine whether the deviation of the actual observed data from the model is large enough to consider that a break has happened.

For the NRT map, the start of the monitoring period is defined as the start of the previous calendar year. To determine the length of the history period, a number of options are available: a reverseordered CUSUM test (with adjustable p-value threshold), segmentation by the algorithm of Bai and Perron mentioned above, or just using all available observations. For the detection of the break in the monitoring period, also a number of monitoring processes are available (Zeileis, 2005), and their $p$-value thresholds can also be customised. From these monitoring processes, Ordinary Least Squares - moving window sum (OLS-MOSUM) and Moving Estimates (ME) tests also make use of an h parameter, in this case referring to the size of the moving data window.

### 3.12.2.4 Selection of VI

As the input to the change detection algorithms is a time series of a VI from the MODIS sensor, three possible MODIS VI were tested: EVI, NDMI and NIRv (for detailed VI definition see section 3.8.2.1). In addition, the VI compositing over 16 days was compared to compositing over 8 days, as well as storage using 2 bytes per pixel vs 1 byte per pixel. The test results showed little difference between the vegetation indices; NDMI resulted in more breaks detected (sensitivity 0.44 , specificity 0.87 ), NIRv resulted in fewer (sensitivity 0.41, specificity 0.91), thus NIRv was selected, as it does not require the use of the lower resolution SWIR band. 16-day compositing was selected over 8-day, as there was no difference in the results for NIRv (for other VIs, 16-day compositing resulted in smoother time series that slightly reduced the number of predicted breaks), and it improved processing time
and storage space requirements. Lastly, 1 byte per pixel was selected, since in addition to lower storage and memory requirements, the reduction in possible values had a positive effect of smoothing the time series of low-vegetation areas such as deserts, reducing the number of false positives there. All in all, 16 -day NIRv with 1 byte per pixel was selected.

### 3.12.3 Collection of change training data to optimize BFAST parameters

The Geo-WIKI tool was updated to generate this unique change reference database. Next to the $10 \times 10(100 \mathrm{~m})$ box as used to collect the 2015 reference training dataset, a 300 m box was added to capture change information that represents to the MODIS dataset as used in the break detection. To supplement the existing Geo-Wiki data sources, we have purchased additional VHR images (Digital Globe at a 50 cm resolution) that fill the data gaps (spatial and time gaps) in Google maps and Microsoft Bing images. These VHR images were visualized on Geo-Wiki to facilitate land cover change detection. Moreover, we have added a tool visualising time series of Sentinel images, which are more frequent than Digital Globe images and therefore helpful to detect approximate time of land cover change.

The IIASA experts have classified 12,000 sample sites to the 2015-2019 change reference, as shown in Figure 64. The locations were sampled in 100 m resolution as well as 300 m resolution to evaluate the influence of MODIS coarser resolution on the detectability of change in the 100 m product. These locations were sampled in the following areas:
I. stratified over the entire Africa continent ( $\sim 1,000$ sample sites of change, $\sim 3,000$ sample sites of no-change);
II. stratified over six priority regions outside the Africa continent ( $\sim 2,000$ sample sites of change, $\sim 6,000$ sample sites of no-change). These priority areas were chosen to capture different change processes:

- Guyana/Brazil (deforestation)
- Indonesia (agro-forestation)
- Ukraine (crop/land abandonment)
- Spain (crop expansion/wetland degradation, water reservoirs)
- Russia SE-Jamal area (low vegetation fires)
- California and greater-Shangai (urbanization)

To assure the quality of the training data set on change, we followed the same approach as for the training data set for the base year 2015. These quality measurements included weekly quality check meetings; regular discussions of various land cover cases; and revisiting of the majority of the classifications.


Figure 64: Overview of Change reference data locations to calibrate BFAST change detection.

### 3.12.4 Implementation of BFAST in the CGLS-LC100 workflow

To be able to get coherent and reproducible results from BFAST a new workflow was setup. This workflow is described in the following part. The workflow starts from the generation of the NIRv VI and ends with the resulting BFAST break maps (BFASTON and BFASTmon). The calibration of BFASTON and BFASTmon is described. The results on the quality checks and ensurance are also included to give evidence on the quality of the workflow and the results break maps. The break detection is run with the final parametrization to obtain the final break maps.

### 3.12.4.1 MODIS VI time-series generation

To have a sufficient length of the time series to detect breaks, a MODIS 250 m derived NIRv VI from 2009 to 2019 (+ first 3 months of 2020) is used (MODIS product MOD09Q1 v006 was used). First the NIRv VI was calculated (see equation 14 in section 3.8.2.1). To ensure minimal impact of resampling, it was done before converting to the UTM grid. Cloudy/sea/missing/snow and invalid areas were masked out using of the quality layer of MOD09Q006. The NIRv VI was averaged 8 -day composites to 16-day composites.
These composites were resampled into the 300 m UTM (bilinear model) grid. All these composites were stacked together into a data cube per tile corresponding to 23 time steps per year per raster cell.

### 3.12.4.2 Calibration of BFASTON

As the BFAST family has a number of parameters, they were calibrated by comparing the break detection results against land cover change training data (Figure 65). First, the algorithm was run

| Document-No. | CGLOPS1_ATBD_LC100m-V3 | © C-GLOPS1 Lot1 consortium |
| :--- | :--- | ---: |
| Issue: 13.4 | Date: 31.03.2021 | Page: $\mathbf{1 1 3}$ of $\mathbf{1 5 2}$ |

over all of the land cover change training points, and then the predicted breaks in the time series were compared to breaks identified by experts in the training data. This comparison was done by checking each entry (year) in the training data to see if there was or was no break detected by the BFAST methods, $\pm 1$ year of the time identified by the experts. Thus a true positive is when training data indicated a break, and BFAST methods predicted a break within 1 year; a true negative is when the data shows no break and no break is predicted by BFAST methods within 1 year. Incorrect predictions of the BFAST methods were false negatives and false positives, respectively. The training data contained information about the change in three years: 2016, 2017 and 2018, and thus for each unique location, there were three observations.

The reference data is highly imbalanced, as land cover change is a rare phenomenon. Out of the total of 101,643 observations in the training data, only 3,055 ( $3 \%$ ) were identified as change. Out of those, one third $(1,011)$ was in Asia, followed by Africa (875) and North America (716), with the rest of the world having little identified change. As such, the objective was to maximise the F1 score statistic, which is a geometric mean of sensitivity (also known as recall or producer's accuracy of change) and precision (also known as user's accuracy of change). This results in a conservative model where the lower of the two statistics, in this case precision, has a higher effect on the F1 score, and therefore avoiding false positives in given a higher priority over detecting more real changes, as the overestimation of change was a bigger challenge to overcome in this case.


Figure 65: Workflow for the calibration of BFASTON and BFAST Monitor.

Due to the size of the dataset and the time needed to run change detection for each unique location, a guided iterative approach was used for selecting optimal parameters. Each parameter was changed step by step to several values within a reasonable range and its effect was evaluated each iteration. If there are no further gains to be had by changing the parameter further, the next parameter is tested, and so on.

The parameters that were chosen in the end for BFASTON included the structural change test at a p-level threshold of 0.01 to decrease processing time and increase specificity; the components used were trend and seasonal dummies ( 7 bins per year), which allows a flexible fit which is still not overfitting the data, and LWZ was the statistic used for estimating the number of breaks in the time series to reduce the number of false detections. This parametrisation results in 0.31 sensitivity, 0.96 specificity and 0.25 F 1 score, compared to the default model's sensitivity of 0.45 , specificity of 0.80 and F1 score of 0.11.

### 3.12.4.3 Calibration of BFASTmon

For BFAST Monitor, a similar optimisation procedure was carried out as for BFASTON (Figure 65). One difference is that due to BFAST Monitor being made to only detect a single break at the end of the time series, the input time series was altered for each observation to cut off the end of the time series 1.25 years after the start of the observation year, so that the monitoring period is always 1.25 year: the year of interest for the observation, plus one quarter that is expected to be available by the time the NRT map is being produced. That way, a near-real-time situation is simulated using historical data.

In addition, BFAST Monitor includes more parameters than BFASTON, and also it runs faster, thus more parameters can be tested. In the end, the chosen parametrisation for BFAST Monitor was also conservative, but distinctly different from BFASTON. The highest F1 score of 0.21 was obtained when using only the trend component (thus only detecting breaks when the trend is significantly different than usual), using all observations for the history period (thus assuming that there was no change throughout the time series up to that point), and setting the monitoring $p$-value threshold to 0.001 to only detect breaks that the model is confident in. This resulted in a sensitivity of 0.44 and specificity of 0.92 , thus even despite all the options making the predictions more conservative, it is not as conservative as BFASTON. Nevertheless, the parametrisation is much more conservative compared to the default settings, which result in a sensitivity of 0.68 , specificity of 0.67 and F 1 score of 0.11 .

### 3.12.4.4 Examples on Quality checks and ensurance

As an example, we take the construction of the Chengdu Tianfu International Airport, a massive project in the Sichuan province of China (Figure 66). Previously the area was rural with agriculture as the dominant land cover class, but in 2016 the construction started and the agricultural land was excavated, revealing bare soil. In the future, the bare soil will be paved and the land cover will transition to the built-up class.

The land cover change reference data indicates the land cover transition happening in late 2017, likely due to the lack of high-resolution imagery at the time of data collection, as in reality the transition happened in early 2017 (official reports dated the project and construction start to beginning of 2017). Running the BFASTON algorithm with tuned parameters on the MODIS NIRv
time series results in the correct identification of the break, and at a more correct time of the year compared to the reference data (Figure 67).

Running BFAST Monitor on the data results in three steps, as NRT data is simulated by cutting off the end of the time series and detecting breaks every year (Figure 68). BFAST Monitor correctly identified no break in 2016 and a break in 2017, although later in the year than it happened in reality. It also incorrectly identified a break in 2018, as it is expected that the land cover would still be more similar to the agricultural state seen earlier in the history, rather than the new bare soil state. This shows a limitation of the NRT data, where the short period of time at the end of the time series makes it more difficult to establish whether there is a change from the normal state, as well as a limitation of BFAST Monitor in that it tends to estimate breaks later in the time series than they actually occur (up to 6 months later).


Figure 66: The construction site of the Chengdu Tianfu International Airport (Sichuan province, China; longitude: 104.463, latitude: 30.326). Left: Sentinel-2 imagery before (summer 2016) and after (summer 2017) the start of the construction, right: an aerial photograph in later stages of construction.

BFASTON


Figure 67: Time series of the area of the Tianfu airport. The break is correctly identified by BFASTON, and the reference time of land cover change is less than one year away from the detected break.


BFAST Monitor


BFAST Monitor
Break detected at: 2018.78


Figure 68: The time series of the Tianfu airport with break detection using BFAST Monitor. Each panel shows one year for which break predictions are made, in sequence (2016-2017, 2017-2018, and 2018 - 2019). The break is correctly identified, but another spurious break is identified one year later, as the model does not yet have enough data to adjust to the new land cover state. Note: blue line indicates time step of detected break, red line time step of real break, dotted line indicates start year of monitoring period.

A different example of land cover change is the replacement of natural forest with tree plantations in the Philippines (Figure 69). A natural forest is cut down or burnt, and cash crops such as oil palms are planted in its stead. The forest clearing event is clearly visible in the MODIS time series (Figure 70), followed by the growth of the new tree crops. The land cover change reference data identifies two changes: from trees to grass between 2015 and 2016, and from grass to shrubs between 2017 and 2018, as the newly planted trees are not yet mature by that point. Another change, from shrubs to trees, may have happened between 2019 and 2020, but reference data was only collected until 2018, and thus this change is not marked in the reference data.

BFASTON detects the clearing event correctly, but does not detect the land cover transition to shrubs, as it is a natural regrowth process that does not involve any break in the time series (Figure 70). This shows a limitation of the approach, as land cover change does not always involve breaks in the time series of a vegetation index. BFAST Monitor likewise detects only the clearing event, and identifies no break during the regrowth phase (Figure 71).


Figure 69: True color Sentinel-2 imagery of forest in the Philippines (longitude: 122.177, latitude: 7.412) being replaced by a tree plantation.


Figure 70: Detection of breaks for the Phillipines deforestation case using BFASTON. Land cover change in 2016 is detected correctly and more precisely than the reference data indicates, but the change in 2018 is missed as it does not include a break.



BFAST Monitor for 2018
No break detected


Figure 71: Detection of breaks for the Phillipines deforestation case using BFAST Monitor. Each panel shows one year for which break predictions are made, in sequence (2016-2017, 2017-2018, and 2018 - 2019). The result is similar to BFASTON, but the detected break is indicated later than in BFASTON. Note: blue line indicates time step of detected break, red line time step of real break, dotted line indicates start year of monitoring period.

### 3.12.5 HMM as binary change detector

It is well known that by the independent classification of annual LC maps, the change detection is negatively affected by classification errors of the individual maps (Liu et al., 2006). The Hidden Markov Model is applied on the full time series of land cover maps in order to improve the temporal consistency of each pixel and reduce the number of illogical and spurious changes (Gong et al., 2017; Abercrombie \& Friedl, 2016). The Hidden Markov Model makes use of the Viterbi algorithm. The model is applied towards the future and the 2015 BASE land cover map classification is considered the start of the sequence. The Hidden Markov Model works with a transition probability matrix. To incorporate all possible change into the model the probability changes between different classes is set to $1 \%$. An exception is made for urban areas which are considered to be monotonically increasing over time (Gong et al., 2017; Hermosilla et al., 2018). Last, a covariation matrix between different land cover classes is calculated. This matrix calculation is done on the biome level, to obtain biome specific results.

Based on the sequence of annual temporal cleaned land cover fractions, the transition probability matrix and the covariation matrix, the most likely land cover class for each time period is inferred. This resulting time series is used for break detection. The results of this method is only used to increase or decrease the level of the confidence in the BFAST generated break mask and has not been used for any other purposes. Since the last year of series (2019) is less predictable due to the NRT status, it is not used in the confidence break mask for the NRT year.

### 3.13 Annual CGLS-LC100 Cover Fraction Products Generation

### 3.13.1 Overview

In a first step to obtain better and more coherent results over time, the epoch-wise cover fraction pre-products are temporally post-processed. Therefore the different years were stacked together and cover fractions were analysed together. The cover fraction that are under investigation are tree, shrub, bare ,moss-lichen, grass and crop cover fraction (6 in total - note: since the built-up and permanent plus seasonal water cover fraction layers are derived from external sources they are not further temporally cleaned but used in the cleaning process of the other). When BFAST did not detect any breaks, cover fractions are expected to be stable over the years. By applying some rules, this is enforced on the cover fractions. These rules must ensure a smooth transition of the cover fractions over the years when no break has been detected, and should limit the increase or decrease of these cover fractions. By applying the rules also the noise of the cover fraction can be diminished. In Figure 72 , an example is given of a pixel with and without smoothing (where the dots represent the original data and the linear line the possible results of the smoothing). When BFAST detected a break or breaks then the cover fractions are mainly unchanged with regard to some basic cleaning rules.


Figure 72: Smoothing of cover fractions over the years 2015-2019. The original cover fractions (square) are smoothened by setting to the value obtained by applying a linear regression.

### 3.13.2 Annual CGLS-LC100 cover fraction products generation

The following steps are carried out to obtain the annual temporal corrected cover fraction products.

- Applying a linear regression;
- Label the temporal stability of the pixel;
- Apply temporal post-regression rules;
- Normalization of the cover fractions per epoch.


### 3.13.2.1 Applying linear regression

To obtain a baseline from which a comparison can be made about the stability of the pixels, a linear regression has been performed per cover (6 in total) over the 5 years (2015-2019). This linear regression was not an ordinary least square (OLS) but a weighted least square (WLS) linear regression which has the advantage that weights can be assigned to points. In this case the correctness of our BaseYear map is assumed to be very high and the correctness of the NRT year is assumed to be very low. In this way the following weights were assigned to the individual years:

- Base years : 20
- Conso years : 2
- NRT years: 1

Mathematically, the weights can be seen as a multiple occurrence of the same point.

### 3.13.2.2 Pixel Stability

The stability of a pixel (per cover) can be obtained by comparing the deviation of the cover fraction to the median of the cover fraction. To do this 2 thresholds were defined:

- stable threshold : 3
- reliable threshold : 15

A pixel is stable if the cover fraction is within a stable threshold distance of the median of the cover fraction. A pixel is reliable if the cover fraction is not stable and within the reliable threshold distance of the median of the cover fraction. A pixel is unreliable if it is not stable nor reliable.

### 3.13.2.3 Temporal post-regression rules

This part is the main part of the temporal cleaning workflow. It will adapt the cover fraction according to the pixel stability, presence of a break and the slope of the linear regression.

In total 13 rules have been applied. These rules can be split into two sets:

- Pixels where only stable or reliable cover fractions are present (5 rules);
- Pixels where at least one of the cover fractions is unstable (8 rules).

The influence of rule set one on the cover fractions is limited: Only smoothing will be applied for the cover fractions that are reliable. Set one has the following rules:

If the sum of all covers (tree, shrub, grass, crop, bare, moss-lichen) is smaller than $10 \%$ then no changes to the cover fraction are done and all other rules are skipped. The next rules checks if all covers are labelled as reliable; if so then the cover fractions, which have a larger error than the average error, is adapted (smoothed). The cover fractions are smoothed by replacing them by the weighted linear regression results. The tree cover fraction has a special rule in this set. If the tree cover fraction is reliable and everywhere zero then the tree cover fraction for all years is pushed to zero, except for the last year or the penultimate year when a break has been detected in this year. In Table 7 a summary can be found for set 1 of the temporal post-regression rules.

Table 7: A schematic view on rule set 1 which is applied on stable and reliable cover fractions. The cover fractions will only be adapted in case of the triggering of rule 03, 04 or 05.

| Rule | Description | Cover updated? |
| :--- | :--- | :--- |
| $\mathbf{0 0}$ | No Rule has been triggered | No, skip all other rules |
| $\mathbf{0 1}$ | sum(covers) < 10\% | No, skip all other rules |
| $\mathbf{0 2}$ | all covers are stable | Yes, smooth values where standard error is larger than the <br> average standard error of all covers. |
| $\mathbf{0 3}$ | all covers are reliable | Yes, set tree outlier to zero except when outlier occurs in NRT <br> year or when outlier occurs in the year before the NRT year and <br> a BFAST break occurs this year |
| $\mathbf{0 4}$ | tree outlier (all zero except one) |  |
| $\mathbf{0 5}$ | mix of stable/reliable covers | Yes, smooth reliable covers where standard error is larger than <br> the average standard error of the reliable covers. |

The second set of rules is applied on all pixels where at least one of the cover fractions is unstable. These rules are more severe and will adapt the cover fractions more than rule set 1 . They will in most cases smooth the cover fraction and limit the slope of the increase/decrease to a reliable slope this reliable slope has been defined as the maximum slope that can be present in case of no errors before a pixel is declared unstable. In this case with a reliable threshold of 15 , it means that the slope of the cover fraction can't be more than $15 /($ amount of years) or 3 . The following rules are applied when one or more of the cover fractions has been labelled unstable. First, for all unstable covers where the cover fraction is every year zero except for 1 year. If the year with the non-zero cover fraction is not the last year or the penultimate year if the penultimate year has a break, then the value of the year with the non-zero cover fraction will put to zero. The same rules apply for the inverse outliers (all non-zero except for one year), but then the cover fraction will be smoothed. If these rules were triggered no other rule of set 2 will be applied. The next rules in set 2 are some cover specific rules. The first rule applies to decreasing forests without a break. The tree cover fraction is smoothed and the slope of the cover fraction of the years will be limited to the reliable slope. The next part of the rules handle confusions when no break has been detected by BFAST (shrub-grass and shrubOpenForest confusion). If only shrub and grass are unstable and no break is present, only smooth cover series with highest standard error and limit their slope to a reliable slope. For OpenForest/Shrub confusion, If both shrub cover fractions and forest fraction are unstable then the forest cover fraction is smoothed and limited to a reliable slope. If the tree cover fraction is unstable and for every even year below the predict cover fraction by the linear regression and every uneven year above the predicted cover fraction (or vice versa) then the tree cover fraction is smoothed. The next step is to check all unstable covers. When there is no break present and the slope of the linear regression of the unstable points is larger than the reliable slope then the cover fraction is smoothed and limited to the reliable slope. In Table 8, a summary can be found for set 2 of the temporal postregression rules.

Table 8: A schematic view on rule set 2 which is applied on unstable cover fractions. Note that in most cases when a break has been detected no adaptation to the cover fraction has been done.

| Rule | Description | Cover updated when no break? | Cover updated when break? |
| :---: | :---: | :---: | :---: |
| 41 | Outlier (all years are zero except one) | Yes, set tree outlier to zero except when outlier occurs in NRT year | Yes, set outlier to zero except when outlier occurs in NRT year or in the year before the NRT year |
| 42 | inverse outlier | Yes, smooth outlier if the pixel becomes stable when the outlier is excluded | Yes, smooth outlier if the pixel becomes stable when the outlier is excluded. Except when outlier occurs in NRT year or in the year before the NRT year |
| 45 | tree decrease <br> (tree unstable + neg. slope) | Yes, smooth tree cover and limit tree to reliable slope | No |
| Document-No. Issue: 13.4 | I3.4. CGLOPS1_ATBD_LC <br>  Date: 31.03 .2021 | CGLOPS1_ATBD_LC100m-V3 | © C-GLOPS1 Lot1 consortium |


| 46 | shrub-grass confusion <br> (shrub and grass are unstable) | Yes, smooth the (shrub or <br> grass) cover fractions with <br> highest standard error and <br> limit the increase to a reliable <br> slope <br> Shrub-OpenForest confusion <br> (tree and shrub unstable) | Yes, smooth the shrub cover <br> fraction and limit the increase <br> to a reliable slope |
| :--- | :--- | :--- | :--- |
| $\mathbf{4 7}$ | No |  |  |
| $\mathbf{4 8}$ | jumpy trees <br> (tree cover fraction unstable and the <br> tree cover fraction alternates <br> Above/Below median) | Yes, smooth the tree cover <br> fraction and limit the increase <br> to reliable slope | Yes fraction |
| $\mathbf{6 9}$ | true unstable <br> (all covers are unstable) | Yes, smooth all cover <br> fractions and limit the <br> increase to a reliable slope | No |
| $\mathbf{7 2}$ | at least one unstable cover, but non <br> of the previous rules has applied | Yes, smooth the cover <br> fraciont and limit the increase <br> of the unstable cover to a <br> reliable slope | No |

However, after applying these rules there is no guarantee that the sum of all cover fractions is again equal to $100 \%$ for each pixel and annual layer. For these cases the cover fractions are normalized as previously explained in section 3.11.2. This will result in the final annual cover fraction layers.

### 3.14 Annual CGLS-LC100 Land Cover Map Generation

### 3.14.1 Overview

Finally, we build upon the success of previous global mapping efforts and/or other ancillary datasets to generate the annual CGLS-LC100 land cover maps. Main input are the epoch-wise classification pre-products with their different scenarios (see section 3.11), the temporal post-processed annual cover fraction layers for the corresponding epoch (see section 3.13.2), and the epoch-wise generated ancillary data masks and products (see section 3.7).

To generate consistent land cover maps two main processing steps have to be carried out: first the epoch-wise generation of land cover pre-maps, and second the temporal post-processing of all epochs. Based on the classification pre-products of the single epochs and on the temporal cleaned cover fraction, expert rules are used to obtain annual land cover pre-maps. To improve our temporal stability, the change in the land cover maps is aligned with the prediction of BFAST which will lead to annual consistent land cover maps. Also a 4-level confidence mask is generated by implementing the HMM break maps in order to quantify and qualify the change areas over time.

| Document-No. | CGLOPS1_ATBD_LC100m-V3 | © C-GLOPS1 Lot1 consortium |
| :--- | :--- | ---: |
| Issue: 13.4 | Date: 31.03 .2021 | Page: $\mathbf{1 2 4}$ of $\mathbf{1 5 2}$ |

### 3.14.2 Epoch-wise discrete land cover pre-map generation

In order to generate the epoch-wise discrete land cover pre-maps, two processing steps are needed (Figure 73). Main input data layers come from the epoch-wise scenario-based classification results, the temporal post-processed cover fraction layers for the corresponding epoch, some products from earlier processing steps, and the ancillary data masks as well as ancillary dataset products for the corresponding epoch (see section 3.7).


Figure 73: General overview of the epoch-wise discrete Land Cover pre-map generation section in the CGLS-LC100 product workflow.

### 3.14.2.1 Pre-processing step

In the first processing step - the pre-processing-, additional input layers for the second processing step are generated. In detail, a discrete map and maximum natural vegetation map are generated for each epoch out of the corresponding annual cover fraction layers as well as the forest type map for the epoch is generated. Moreover, the results of the CL2 scenario-based classification are updated into the CL2p5 classification map by cleaning areas in the High North latitudes with the CAVM (equivalent as the step in the cover fraction layer generation section - see section 3.11.2). Figure 74 shows the pre-processing for the new layers in more detailed. Note: again all processing is done epoch-wise.


Figure 74: Flowchart of the classification post-processing to produce additional input layers for the land cover pre-product generation.

First, in order to incorporate the vegetation cover fraction layers in the LC map generation process, a discrete map is generated by applying the training data rules on the annual temporal cleaned cover fraction layers (see section 3.13). In detail, during the training data collection (see section 3.9), a set of rules has been established to assign a training point with its by visual interpretation estimated cover fraction percentages to a discrete class (e.g. training point with cover percentages of $65 \%$
forest and 35 \% shrubs is classified as an "open forest" training point). The final discrete LC map can be seen as a reverse classification process, since the LC class was not directly assigned by a classification algorithm. The applied rules are shown in detail in Figure 74. This discrete LC map is called Cover-2-discrete (CL5) during the expert rules.

Next, the maximum natural vegetation mask is generated. Therefore, the annual temporal cleaned cover fraction layers excluding the agriculture (crops) thematic layer are processed and, for each pixel, the LC class with the maximum cover fraction is assigned. This discrete LC map is called max. natural vegetation map (CL7) during the expert rules. Again, the applied rules are shown in detail in Figure 74. The last pre-processing step is the generation of the forest type map for the epoch, whereby the CL4 scenario-based classification map is only cleaned up by the tree cover fraction layer of the corresponding epoch. Therefore, areas with a tree cover fraction percentage of $0 \%$ in the tree cover fraction layer are set to the nodata value in the CL4 forest type map (Figure 74).

The next sub-section explains expert rules - the second processing step in the discrete pre-map generation - in more detail. There we combine existing knowledge represented by ancillary datasets with the classification and regression results for each epoch (see section 3.14.2.2).

### 3.14.2.2 Epoch-wise discrete land cover pre-map generation via expert rules

The final processing step in the generation of the epoch-wise discrete land cover pre-maps is the combination of all classification scenarios and ancillary datasets into a final LC map vial expert rules (for each epoch independently). The expert rules are implemented as a decision tree which is applied pixel-based on the input data. Thereby, the predicted class probabilities of the classification scenarios are used as thresholds in the decision tree in order to generate the class discrete map product, but are not the only deciding factor as explained in the next paragraphs. Figure 75 to Figure 79 illustrate the applied expert rules in detail.

Overall 23 discrete classes have to be compiled out of the input data. Table 9 give an overview of the classes, class groups and the final assigned raster values in the discrete land cover map of the pre-map each epoch. Note: this table is also valid for the final annual LC maps.

Table 9: The $\mathbf{2 3}$ discrete classes of the CGLS-LC100 discrete product. Note: final classes are shown in bold together with numerical code in brackets.

| Level 1 | Level 2 | Level 3 |
| :---: | :---: | :---: |
| forest (10) | closed forest (11) | closed forest, needle-leaved, evergreen (111) |
|  |  | closed forest, needle-leaved, deciduous (113) |
|  |  | closed forest, broadleaved, evergreen (112) |
|  |  | closed forest, broadleaved, deciduous (114) |
|  |  | closed forest, mixed type (115) |
|  |  | closed forest, unknown type (116) |
|  | open forest (12) | open forest, needle-leaved, evergreen (111) |
| Document-No. Issue: I3.4 | CGLOPS1_ATBD_LC100m-V3 <br> Date: 31.03.2021 | © C-GLOPS1 Lot1 consortium Page: 127 of 152 |


|  |  | open forest, needle-leaved, deciduous (113) |
| :--- | :--- | :--- |
|  |  | open forest, broadleaved, evergreen (112) |
| open forest, broadleaved, deciduous (114) |  |  |
| open forest, mixed type (115) |  |  |
| open forest, unknown type (116) |  |  |

Overall 16 input datasets which can be grouped into five input dataset groups are used in the expert rules for each epoch. In detail (see also Figure 75):

- Two products of the epoch-wise scenario-based classifications (see section 3.10.4):
- PROBA-V UTM 100 m CL1 ("pure" class scenario) classified map; and
- PROBA-V UTM 100 m CL1 ("pure" class scenario) class probabilities;
- Five layers of the temporal cleaned annual cover fraction layers (permanent snow\&ice, builtup, tree, shrub, and permanent water) (see section 3.13.2);
- Five layers produced during the pre-processing:
- PROBA-V UTM 100 m CL2p5 (updated "discrete" class scenario) classified map;
- PROBA-V UTM 100 m CL2p5 (updated "discrete" class scenario) class probabilities;
- PROBA-V UTM 100 m CL5 (Cover-2-discrete) classified map;
- PROBA-V UTM 100 m CL7 (max. natural vegetation map) classified map; and
- CGLS-LC100 forest type layer;
- PROBA-V UTM 100 m NoData mask;
- The land/see mask as part of the ancillary masks (see section 3.7); and
- The WetProducts wetland indicator mask (see section 3.7.3) as well as the biome cluster layer (see section 3.7.1) as part of the ancillary dataset products.

Issue: I3.4

Before the first expert rule (Figure 75) is applied on the input data, two output datasets are initialized with a "No Data" value for each epoch. For simplification the following explanations are done for one epoch. Expert rule 1 imprints the hardcoded masks, i.e. the 'no data mask' and the 'land/sea mask', into the 'Final Classification' dataset. The 'Final Probability' dataset is updated for the corresponding pixels, resp. 'No Data' and '100 \%' (Figure 75).

Expert rule 2 deals with imprinting the 'derived' LC classes. Meaning that LC classes are not obtained from direct classification and instead derived from thresholding the cover fraction pre-products, i.e. the permanent snow\&ice cover fraction, permanent water cover fraction and the built-up cover fraction for the corresponding epoch (Figure 76). For this, three checks are performed:

- Pixels for which a permanent snow\&ice cover fraction greater than zero was found obtain the final class 'permanent snow \& ice'.
- Pixel having a permanent water cover fraction equal to or over $60 \%$ obtain the class 'permanent water'.
- Pixel having a built-up cover fraction equal to or over $60 \%$ obtain the class 'built-up.

For each matching case the corresponding pixel's final probability is set to 'No Data'. Main reason these pixels have no classification probability is that these pixels come partly from external data and therefore a probability based on the classification training cannot be given.


Figure 75: Initialization and Expert Rule block 1 for the discrete pre-map and map probability layer generation - imprinting hard masks.

In the next step, expert rule 3 checks for high probabilities in the classification maps CL1 and CL2p5 and if checked positive the corresponding classes are imprinted in the final map (Figure 76). The
rules are only applied for those pixels which are not classified as 'permanent snow \& ice' in the CL1 or CL2p5 map and have still a 'No Data' value in the final map. Rule 3 is applied in three steps:

- If corresponding pixels in CL1 and CL2p5 have the same class and if those pixels have a probability over 95 \% in CL1 and 60 \% or higher in CL2p5 then this Ic class is taken over in the final map. The final probability for those pixels is set to the highest value found for the corresponding probabilities in CL1 and CL2p5.
- If the CL2p5 probability is equal or over $60 \%$ and the final pixel is not yet classified, it will obtain the class from CL2p5 as well as the corresponding CL2p5 probability.
- If the CL1 probability is over $90 \%$ and the final pixel is not yet classified, it will obtain the class from CL1 as well as the corresponding CL1 probability.


Figure 76: Expert Rule block 2 to 4 for the discrete pre-map and map probability layer generation imprinting the derived cover fraction layers for classes, resolving classification scenarios, and filling of the remaining land pixels.

Subsequently, rule 4 resolves land pixels which are not yet classified in the final map (Figure 76). Most important in this filling step is that we have to guarantee that no 'permanent snow\&ice' pixels from CL2p5 or CL5 are copied to the final product since the 'permanent snow\&ice' class was imprinted in Rule 2. These pixels are evaluated in three steps:

- If the CL2p5 class is different from 'permanent snow \& ice' the unclassified pixels in the final map and probability layer get the corresponding class and probability from CL2p5.
- If still pixels are left unclassified in the final map they get the class and probability from CL5 unless they were classified as 'permanent snow \& ice'.
- Finally, if still pixels are left unclassified in the final map they get the class and probability from CL7 which is a mask showing the max. natural vegetation (evaluated by the highest value in cover fraction of all vegetation classes including bare). Note: this rule was implemented as pure fallback in cases of artefacts in the classification of CL1, CL2p5 and CL5. The logged statistics of the assigned expert rules show that this rule was never triggered for any pixels in the global LC maps of 2015.

At this stage, all pixels in the final map have a class and probability assigned.

To minimize misclassifications between closed forest, open forest, shrub and herbaceous vegetation in the discrete pre-map, another set of expert rules is applied by rule 5 (Figure 77, Figure 78). These rules are continent as well as biome cluster specific. Figure 48 and Table 6 show the used assignment in order to transfer biome cluster into continents. The following paragraphs explain the applied rule sets by continent:

- Africa
- Rule 5b: check confusion between shrub and open forest. If a pixel in the final map is classified as 'shrub' with a probability of $40 \%$ or less and the class of the corresponding pixel in the CL5 map indicates 'open forest' then the class in the final map is changed to 'open forest'. If a pixel in the final map is classified as 'shrub' and this pixel has a tree cover fraction of $25 \%$ or higher then the class in the final map is changed to 'open forest'.
- Rule 5c: check confusion between grassland and shrub. If a pixel in the final map is classified as 'grassland' with a probability of $40 \%$ or less and the class of the corresponding pixel in the CL5 map indicates 'shrub' then the class in the final map is changed to 'shrub'. If a pixel in the final map is classified as 'grassland' and this pixel has a shrubland cover fraction of $30 \%$ or higher and the biome type is ZB32 or ZB34 (two areas in the Southern part of Africa) then the class in the final map is changed to 'shrub'.
- Oceania
- Rule 5a: check for confusion between closed and open forest. If a pixel in the final map is classified as 'closed forest' with a probability of $40 \%$ or less and the class of the corresponding pixel in the CL5 map indicates 'open forest' then the class in the final map is changed to 'open forest'.
- Rule 5b: check confusion between shrub and open forest. If a pixel in the final map is classified as 'shrub' with a probability of $40 \%$ or less and the class of the corresponding pixel in the CL5 map indicates 'open forest' then the class in the final map is changed to 'open forest'.
- Rule 5c: check confusion between grassland and shrub. If a pixel in the final map is classified as 'grassland' with a probability of $40 \%$ or less and the class of the
corresponding pixel in the CL5 map indicates 'shrub' then the class in the final map is changed to 'shrub'. If a pixel in the final map is classified as 'grassland' and this pixel has a shrub cover fraction of $30 \%$ or higher then the class is changed to 'shrub'.
- South America, Europe and Eurasia
- Rule 5b: check confusion between shrub and open forest. If a pixel in the final map is classified as 'shrub' and this pixel has a tree cover fraction of $30 \%$ or higher then the class in the final map is changed to 'open forest.
- North America
- Rule 5a: check for confusion between closed and open forest. If a pixel in the final map is classified as 'closed forest' with a probability of $40 \%$ or less and the class of the corresponding pixel in the CL5 map indicates 'open forest' then the class in the final map is changed to 'open forest'.
- Rule 5b: check confusion between shrub and open forest. If a pixel in the final map is classified as 'shrub' with a probability of $40 \%$ or less and the class of the corresponding pixel in the CL5 map indicates 'open forest' then the class in the final map is changed to 'open forest'. If a pixel in the final map is classified as 'shrub' and this pixel has a tree cover fraction of $30 \%$ or higher then the class in the final map is changed to 'open forest'.
- Rule 5c: check confusion between grassland and shrub. If a pixel in the final map is classified as 'grassland' with a probability of $40 \%$ or less and the class of the corresponding pixel in the CL5 map indicates 'shrub' then the class in the final map is changed to 'shrub'. If a pixel in the final map is classified as 'grassland' and this pixel has a shrubland cover fraction of $25 \%$ or higher then the class is changed to 'shrub'.
- Asia
- Rule 5a: check for confusion between closed and open forest. If a pixel in the final map is classified as 'closed forest' with a probability of $40 \%$ or less and the class of the corresponding pixel in the CL5 map indicates 'open forest' then the class in the final map is changed to 'open forest'.
- Rule 5b: check confusion between shrub and open forest. If a pixel in the final map is classified as 'shrub' with a probability of $40 \%$ or less and the class of the corresponding pixel in the CL5 map indicates 'open forest' then the class in the final map is changed to 'open forest'.
- Rule 5c: check confusion between grassland and shrub. If a pixel in the final map is classified as 'grassland' with a probability of $40 \%$ or less and the class of the corresponding pixel in the CL5 map indicates 'shrub' then the class in the final map is changed to 'shrub'.


Figure 77：Expert Rule block 5 part A for the discrete pre－map and map probability layer generation－resolving confusion between closed forest，open forest，shrub and grassland on continental level．＊See Figure 48 and Table 6 for biome cluster location and naming．


Figure 78: Expert Rule block 5 part B for the discrete pre-map and map probability layer generation - resolving confusion between closed forest, open forest, shrub and grassland on continental level.

Issue: I3.4

In the final stage of the expert rules, two additional checks are applied. Rule 6 checks for herbaceous wetland. If the final class for a pixel is grassland and the corresponding pixel in the WetProducts wetland mask indicates "wetland" then the final class for this pixel is changed to 'herbaceous wetland' (Figure 79).


Figure 79: Expert Rule block 6 and 7 for the discrete pre-map and map probability layer generation wetland class improvements and final separation of the open/closed forest class by forest type.

Rule 7 imprints the forest types. If a pixel in the final map is classified as 'closed-' or 'open forest' and the corresponding pixel in the forest type map has a meaningful value then the pixel's class is updated accordingly. This rule means that depending on the value in the forest type map:

- 'closed forest' is differentiated into:
- evergreen needleleaf closed forest
- evergreen broadleaf closed forest
- deciduous needleleaf closed forest
- deciduous broadleaf closed forest
- 'open forest' is differentiated into:
- evergreen needleleaf open forest
- evergreen broadleaf open forest
- deciduous needleleaf open forest

Document-No.
Issue: I3.4

CGLOPS1_ATBD_LC100m-V3
Date: 31.03.2021
© C-GLOPS1 Lot1 consortium Page: 135 of 152

- deciduous broadleaf open forest

If the corresponding pixel in the forest type map has no meaningful value then the class 'closed forest' is changed to 'undefined closed forest' and the class 'open forest' is changed to 'undefined open forest' (Figure 79).

### 3.14.3 Annual CGLS-LC100 Land cover maps generation

To further improve the temporal consistency of the discrete LC maps, the epoch-wise land cover pre-maps are stacked temporally and cleaned with some temporal expert rules. The final annual land cover maps are created with a rule set that only affect CONSO and NRT years (Note: epochwise LC pre-maps generated in processing mode conso and nrt). The BaseYear land cover pre-map is not changed since this map is considered as our source of truth.

The following rules apply for CONSO and NRT pre-maps:

1. A break and therefore a possible land cover change process is considered in year X if a BFAST break (see section 3.12.2) was detected in this year $X$ or in the last six months of year $X$-1.
2. If there is no satellite data in this year then the data of the year before is taken except if a break has been detected then the value is put to the no data value.
3. If built-up, permanent water, snow or wetland changes has been detected, it is kept.
4. If a break is detected in the first half of the NRT year (BFASTmon), this break is used and the land cover altered to the NRT discrete class.
5. If a break is detected in the second half of the NRT year (BFASTmon), this break is discarded and the land cover of the year before is used.
6. For all other data points if there is no break then the land cover of the year before is used.

To provide the user with a quality layer regarding the change detection, a change confidence mask for the break detection has been created based on the amount of evidence there is for the break detection. It is a 3 level confidence mask for all CONSO and NRT LC maps based on the following value definitions:

0 : no change;
1: (only for NRT), BFASTmon detected break in second half of NRT year - potential change;
2: Imprint of urban, permanent water, snow or wetland OR change detected by BFAST but HMM model didn't confirm this break in higher resolution OR change detected by BFASTmon in the first half of NRT year

3: (only non-NRT), BFAST detected a change and HMM confirmed this change in higher resolution.

### 3.15 The Annual CGLS-LC100 Products

The final step in the annual CGLS-LC100 products generation is the injection of metadata attributes compliant with version 1.6 of the Climate \& Forecast conventions (CF V1.6) as well as injection of the colour bars translating the discrete class code into the legend. A full explanation of the injected metadata attributes can be found in the CGLS-LC100 Product User Manual ([CGLOPS1_PUM_LC100m-V3.0]).

Overall fourteen layers are provided for each mapped year (BASE, CONSO and NRT mode) as well as an additional layer for all CONSO and NRT processed years and delivered as the CGLS-LC100 collection 3 product:

1. CGLS-LC100 discrete map;
2. CGLS-LC100 discrete classification probability;
3. CGLS-LC100 forest-type layer;
4. CGLS-LC100 bare/sparse vegetation cover fraction layer;
5. CGLS-LC100 cropland cover fraction layer;
6. CGLS-LC100 herbaceous grassland cover fraction layer;
7. CGLS-LC100 moss \& lichen cover fraction layer;
8. CGLS-LC100 shrubland cover fraction layer;
9. CGLS-LC100 tree cover fraction layer;
10. CGLS-LC100 permanent snow \& ice cover fraction layer;
11. CGLS-LC100 built-up cover fraction layer;
12. CGLS-LC100 permanent water cover fraction layer;
13. CGLS-LC100 seasonal water cover fraction layer;
14. CGLS-LC100 DataDensityIndicator for PROBA-V UTM 100m input data; and
15. CGLS-LC100 Change confidence mask (only for CONSO and NRT processed epochs).

Figure 80 shows an overview of the discrete map product on global scale for the BaseYear 2015, where Figure 82 demonstrates the legend for the CGLS-LC100 discrete map with 23 classes together with the RGB colour codes and assigned byte values for each LC class. Moreover, Figure 82 shows the additional grouping possibilities of the CGLS-LC100 discrete map into 3 levels. The grouping can be mainly done on "forest" level and is used in the validation ([CGLOPS1_VR_LC100m-V3.0]) and distribution of the product.

Figure 81 shows the 10 provided CGLS-LC100 cover fraction layers for the BaseYear 2015 in a collage.


Figure 80: The CGLS-LC100 discrete map product for epoch 2015 with 23 discrete classes (detailed legend in Figure 82)


Figure 81: The cover fraction layers for the 9 base land cover classes and the seasonal inland water cover fraction of the CGLS-LC100 product for epoch 2015 (shown as a collage on global scale).

Copernicus Global Land Operations - Lot 1
Date Issued: 31.03.2021
opernicus
Europe's eyes on Earth
Issue: I3.4

| LEVEL 2 |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| value map | grouping | group <br> code | R | G | B | class name |
| 0 |  | 0 | 40 | 40 | 40 | unknown (no input data available) |
| 111 | $\boldsymbol{x}$ | 11 | 0 | 120 | 0 | closed forest |
| 113 |  |  |  |  |  |  |
| 112 |  |  |  |  |  |  |
| 115 |  |  |  |  |  |  |
| 116 |  |  |  |  |  |  |
| 121 | $\boldsymbol{x}$ | 12 | 100 | 140 | 0 | open forest |
| 123 |  |  |  |  |  |  |
| 122 |  |  |  |  |  |  |
| 124 |  |  |  |  |  |  |
| 125 |  |  |  |  |  |  |
| 126 |  |  |  |  |  |  |
| 20 |  | 20 | 255 | 187 | 34 | shrubland |
| 30 |  | 30 | 255 | 255 | 76 | herbaceous vegetation |
| 90 |  | 90 | 0 | 150 | 160 | herbaceous wetland |
| 100 |  | 100 | 250 | 230 | 160 | moss \& lichen |
| 60 |  | 60 | 180 | 180 | 180 | bare / sparse vegetation |
| 40 |  | 40 | 240 | 150 | 255 | cropland |
| 50 |  | 50 | 250 | 0 | 0 | built-up |
| 70 |  | 70 | 240 | 240 | 240 | snow \& ice |
| 80 |  | 80 | 0 | 50 | 200 | permanent inland water bodies |
| 200 |  | 200 | 0 | 0 | 128 | sea |


| LEVEL 3 |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| value map | grouping | group code | R | G | B | class name |
| 0 |  | 0 | 40 | 40 | 40 | unknown (no input data available) |
| 111 |  | 111 | 88 | 72 | 31 | closed forest, needle-leaved, evergreen |
| 113 |  | 113 | 112 | 102 | 62 | closed forest, needle-leaved, deciduous |
| 112 |  | 112 | 0 | 153 | 0 | closed forest, broadleaved, evergreen |
| 114 |  | 114 | 0 | 204 | 0 | closed forest, broadleaved, deciduous |
| 115 |  | 115 | 78 | 117 | 31 | closed forest, mixed type |
| 116 |  | 116 | 0 | 120 | 0 | closed forest, unknown type |
| 121 |  | 121 | 102 | 96 | 0 | open forest, needle-leaved, evergreen |
| 123 |  | 123 | 141 | 116 | 0 | open forest, needle-leaved, deciduous |
| 122 |  | 122 | 141 | 180 | 0 | open forest, broadleaved, evergreen |
| 124 |  | 124 | 160 | 220 | 0 | open forest, broadleaved, deciduous |
| 125 |  | 125 | 146 | 153 | 0 | open forest, mixed type |
| 126 |  | 126 | 100 | 140 | 0 | open forest, unknown type |
| 20 |  | 20 | 255 | 187 | 34 | shrubland |
| 30 |  | 30 | 255 | 255 | 76 | herbaceous vegetation |
| 90 |  | 90 | 0 | 150 | 160 | herbaceous wetland |
| 100 |  | 100 | 250 | 230 | 160 | moss \& lichen |
| 60 |  | 60 | 180 | 180 | 180 | bare / sparse vegetation |
| 40 |  | 40 | 240 | 150 | 255 | cropland |
| 50 |  | 50 | 250 | 0 | 0 | built-up |
| 70 |  | 70 | 240 | 240 | 240 | snow \& ice |
| 80 |  | 80 | 0 | 50 | 200 | permanent inland water bodies |
| 200 |  | 200 | 0 | 0 | 128 | sea |

Figure 82: The three level legend of the CGLS-LC100 discrete map product. The raster code together with the RGB colour code and discrete class name is shown.

Finally, to identify all collections and dataset versions of the annual CGLS-LC100 product, we decided to assign unique Digital Object Identifier (DOI) to each annual produced dataset. All products are uploaded on the Zenodo platform for long-term archiving and assigned a concept DOI as well as a version DOI. The concept DOI will resolve all the time into the newest version of the dataset within a collection where the version DOI can be used if a specific dataset wants to be accessed or cited. The DOI information for all collections and dataset versions of the CGLS-LC100 product will be always updated in the Product User Manual ([CGLOPS1_PUM_LC100m-V3.0]). Table 10 shows the provided information exemplary for Collection 1 to 3 .

Table 10: Overview of existing CGLS-LC100 products produces by algorithm version

| Dataset name | Collection 1 | Collection 2 | Collection 3 |
| :---: | :---: | :---: | :---: |
| Dataset version | V1.0.2 | V2.0.2 | V3.0.1 |
| Temporal range | 2015 | 2015 | 2015-2019 |
| Processor algorithm version for Dataset | V1.0 | V2.2 | V4.1 |
| Concept DOI (always resolves in newest version) | - | 10.5281/zenodo. 3243508 | 2015: 10.5281/zenodo. 3243508 <br> 2016: 10.5281/zenodo. 3518025 <br> 2017: 10.5281/zenodo. 3518035 <br> 2018: 10.5281/zenodo. 3518037 <br> 2019: 10.5281/zenodo. 3939049 |
| Version DOI | - | 10.5281/zenodo. 3243509 | 2015: 10.5281/zenodo. 3939038 2016: 10.5281/zenodo. 3518026 2017: 10.5281 /zenodo. 3518036 2018: 10.5281 /zenodo. 3518038 2019: 10.5281/zenodo. 3939050 |
| Citation | - | Marcel Buchhorn, Bruno Smets, Luc Bertels, Myroslava Lesiv, Nandin-Erdene Tsendbazar, Martin Herold, \& Steffen Fritz. (2019). Copernicus Global Land Service: Land Cover 100m: Collection 2: epoch 2015 (Version V2.0.2) [Data set]. Zenodo. <br> 10.5281/zenodo. 3243509 | Marcel Buchhorn, Bruno Smets, Luc Bertels, Myroslava Lesiv, NandinErdene Tsendbazar, Dainius Masiliunas, Lin Linlin, Martin Herold, \& Steffen Fritz. (2020). Copernicus Global Land Service: Land Cover 100m: Collection 3: epoch <YEAR>: Globe (Version V3.0.1) [Data set]. Zenodo. DOI: <VERSION DOI FOR YEAR> |

Copernicus Global Land Operations - Lot 1
Date Issued: 31.03.2021

Issue: I3.4
Europe's eyes on Earth

| Direct access | - | https://land.copernicus.eu/global <br> Icviewer | https://land.copernicus.eu/global//cvie <br> wer |
| :--- | :--- | :--- | :--- |
| Data layers <br> per epoch | 8 | 20 | 14 for BASE years; <br> 15 for CONSO and NRT years |
| File Size | 6.6 GB | 63.3 GB | 187.4 GB |
| Note | only Africa | global | global |

## 4 Limitations

Although minimization of omission and commission errors is achieved by the usage of ancillary dataset, they are sometimes inevitable. An overview of reasons for omission and commission errors include:

- Due to the usage of optical RS data, classifications in areas with high yearly cloud cover can have lower classification accuracies.
- Artefacts at boundaries of biome clusters can appear due to the used biome cluster vector layer as well as the biome cluster-specific generated hyper-parameter for the Random Forest classifier and regressor.
- Remaining shadowed pixels in the time series not filtered out during the data cleaning process can lead to misclassifications.
- Fires and more specifically burned areas were not yet taken into account and therefore could lead to misclassifications.
- Highly fragmented landscapes, in particular mixed areas with very small cropland fields (less $<0.5 \mathrm{ha}$ ), are very difficult to map because of the resolution of 100 m (e.g. Nigeria, Ghana). This could lead to an overestimation of croplands.
- Areas with low cropland fragmentation (very sparse cropland fields of a very small size) are difficult to map because of the resolution of 100 m . This could lead to an underestimation of croplands.
- Very small villages or bigger houses are difficult to map, especially when not detected by the WSF layer at 10 m resolution or the OpenStreetMap layers, which could lead to an underestimate of built-up areas.
- Through the usage of the OpenStreetMap layer all limitation of the layer apply, e.g. wrongly mapped buildings in location or size. In certain areas urban commission errors were detected due to scaling issues in the OSM layer. This OSM issue can lead also to errors in the change detection between years.
- Some limitations are due to the legend or class definition, e.g.:
- In the southern part of Africa, there are huge areas with kind of tundra type of vegetation (low vegetation created out of mosses and grass), NDVI values are very low in these areas and can confuse the classifier to misclassify between grassland or bare land.
- In Africa, there are a lot of riparian forests, which are evergreen. A lot of pixels were noticed with mixed deciduous trees and riparian evergreen forest which can confuse the classifier to misclassify the forest type.
- The permanent snow and ice class was limited by the PPSIM layer in order to reduce permanent snow\&ice misclassifications especially over temporary frozen lakes (mostly in the High Latitudes where high cloud cover is frequent and therefore images of the thawed lakes are rare). This masking can led to underestimation of newly formed ice\&snow areas outside the PPSIM.
- The CGLS-LC100 permanent snow and ice cover fraction layer shows only cover fractions of $0 \%$ or $100 \%$ at the moment. This is due to the fact that not sufficient amounts of training points with a reported permanent snow\&ice cover fraction were available. Therefore, we focused on areas with $100 \%$ covered permanent snow\&ice. This can lead to an underestimation of permanent snow and ice in partly covered 100 m pixels.
- The usage of PROBA-V multi-spectral data limits the possibilities to successfully distinguish between herbaceous vegetation and shrubland, as well as between open forest and shrubland, especially when the class borders within a pixel are not clear (meaning the cover fraction of the neighbouring discrete classes are close and therefore the voting in the classifier is difficult). We implemented an extra rule part in the Expert Rules section to deal with these classification confusions, but still misclassifications between these classes are possible. This confusion is expected to be reduced when switching to the Sentinel-2 multispectral data input - see chapter 5 for an outlook of the future development of the product.
- Since the PROBA-V time series is limited at the moment (2014 to 2020), misclassifications can happen especially in temporary ultra-dry and temporary ultra-wet areas. This misclassification is due to the fact that some regional effects with a low frequency (e.g. El Niño and La Niña) could not be correctly captured with the training data reference base of 2015
- Internal change process from one class to another and back (mainly shrub - grass) can appear due to the usage of pre-trained classification / regression models. Since the models are trained on 2015 reference data and if this reference year was an unexpected wetter or dryer year than usual in a region, this can cause confusion in shrub - grass detection in certain regions (e.g. Central Australia). Therefore this leads to false change detection.
- Due to the pixel-based classification and regression, spatial artefacts can occur. We encourage users to apply cluster/sieving techniques in non-fragmented areas to further improve the regional classification results.
- Due to the low training data density in the High Latitudes, the classification results above 70 degree North can be less accurate.
- Small-scale changes may be missed due to MODIS data being used for change detection, as its MMU is 300 m (MODIS 250 m regridded to Proba-V 300 m grid).
- The change detection algorithms were calibrated using training data that was created using very high resolution image interpretation. This may not always scale well to what is visible at the 300 m level, as changes indicated from the former are not visible (and cannot be detected) in the time series of the latter. This impacts the statistics from BFAST calibration.
- A change detection approach based on a vegetation index time series cannot detect transitional stages in degradation or regrowth, e.g. the succession of grasses to shrubs to trees would always get detected as two breaks: before the start of the regrowth and after the start of the regrowth. These breaks may be offset from the corresponding land cover change times, depending on the speed of regrowth.
- As a global parametrisation for break detection is used, its performance may vary over different continents or ecozones.
- Detecting changes close to the end of the time series may be less accurate in cases where a rapid change is ongoing. For example, the endpoint of deforestation (time when the deforestation switches into a regrowth) may not be detected as a break because the model can assume that negative vegetation index values are possible, and it takes a longer time series to make sure that the new situation is stable.
- Changes in the time period January - February of a year could be recorded in the change of the year before due to BFAST break detection (time shifting problem).


## 5 Future development of the product

With ending of the PROBA-V mission on June $30^{\text {th }}, 2020$, no input data from PROBA-V will be available anymore. Continuation will be guaranteed using the Sentinel-2 dataset and using Sentinel1 for gap filling.

Sentinel 1 \& 2 data has become globally available since 2018 and offers a new data source which can be used as input for the CGLOPS classification routine. Limited changes to the routine are necessary to work with Sentinel input.

The main difference between Sentinel and PROBA-V input data is that Sentinel $1 \& 2$ bands record data at $10+20 \mathrm{~m}$. In order to achieve 100 m resolution land cover maps from this data there are two options. First, the aggregation of these reflectance to 100 m is an option. As such, both the classification and regression modules can be executed using the original training data and highly consistent results can be obtained regarding cover fractions compared to PROBA-V. Second, the classification can be executed on the 10 m reflectance bands directly. As such, the cover fractions at 100 m will be obtained by calculating the amount of 10 mx 10 m pixels within the 100 m aggregated pixel. In this way, an optimal use of the Sentinel information is obtained.

Both aggregation of these reflectance to 100 m and the direct classification at 10 m resolution are options to continue the generation of Copernicus 100 m global land cover layers from 2020 onwards. Furthermore, the additional spectral bands available on the Sentinel satellites compared to PROBAV allow us to calculate additional vegetation indices, which adds more information to the classification / regression routines.

Anyhow, the increased spectral resolution of Sentinel-1 and Sentinel-2 compared to the PROBA-V satellite will allow additional separation possibilities.

Preliminary tests of these routine using Sentinel input data for the year 2019 have been executed, showing overall accuracies at a global scale in line with the results of the obtained using PROBA-V. Additionally, preliminary land cover classification maps are obtained using Sentinel as input for several countries in both Africa and Europe.

A workflow to ensure consistency between the results of Sentinel and PROBA-V is currently in design.

## 6 References

Abercrombie, S.P.; Friedl, M.A. (2016): Improving the Consistency of Multitemporal Land Cover Maps Using a Hidden Markov Model, in IEEE Transactions on Geoscience and Remote Sensing, vol. 54, no. 2, pp. 703-713, Feb. 2016, DOI: 10.1109/TGRS.2015.2463689.

Aimonen, P. (2011). Basic concept of Kalman filtering.svg. Image published under the Creative Commons CCO 1.0 Universal Public Domain Dedication. Webpage: https://upload.wikimedia.org/wikipedia/commons/a/a5/Basic concept of Kalman filtering.svg.

Alaska Geobotany Center (2003): Circumpolar Arctic Coastline and Treeline Map, http://arcticatlas.geobotany.org/catalog/entries/5421-circumpolar-arctic-coastline-and-treelinemap.

Arantes, T.; Chaves, M.; Bastos, R.; Carvalho, L.; Oliveira, M. (2017). Effectiveness of BFAST algorithm to characterize time series of dense forest, agriculture and pasture in the Amazon region. Theoretical and Applied Enginnering - TAAE (ISSN: 2594-3863). 1. 10-19.

Badgley, G., Field, C. B., \& Berry, J. A. (2017). Canopy near-infrared reflectance and terrestrial photosynthesis. Science Advances, 3(3), e1602244. DOI: 10.1126/sciadv. 1602244.

Bai, J.; Perron, P. (2003). Computation and analysis of multiple structural change models, Journal of Applied Econometrics, 18, 1, DOI: 10.1002/jae. 659

Bennie, Jonathan; Mark O. Hill; Robert Baxter; Brian Huntley (2006). "Influence of slope and aspect on long-term vegetation change in British chalk grasslands". Journal of Ecology. 94 (2): 355-368. doi:10.1111/j.1365-2745.2006.01104.x.

Bertels, L., Smets, B., Wolfs, D. (2016). Dynamic Water Surface Detection Algorithm Applied on PROBA-V Multispectral Data. Remote Sens., 8, 1010. DOI: $10.3390 / \mathrm{rs} 8121010$.

Berthelot, B. and Dedieu, G. (1997). Correction of atmospheric effects for VEGETATION data. In: G.a.P. Guyot, T. (Editor), Physical Measurements and Signatures in Remote Sensing, Courchevel, France, pp. 19-25.

Blackburn, G.A. (1998). Spectral indices for estimating photosynthetic pigment concentrations: a test using senescent tree leaves. International Journal of Remote Sensing, 19 (4), 657-675. DOI: 10.1080/014311698215919.

Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32.
CAVM Team. 2003. Circumpolar Arctic Vegetation Map. Scale 1:7,500,000. Conservation of Arctic Flora and Fauna (CAFF) Map No. 1. U.S. Fish and Wildlife Service, Anchorage, Alaska.

CCI Land Cover (2017). CCI Land Cover - Release of a 1992-2015 time series of annual global land cover maps at 300 m , (2017). webpage: https://www.esa-landcovercci.org/index.php?q=webfm send/88.

Coops, N. C., Kearney, S. P., Bolton, D. K., \& Radeloff, V. C. (2018). Remotely-sensed productivity clusters capture global biodiversity patterns. Scientific reports, 8(1), 16261. doi:10.1038/s41598-018-34162-8

Copernicus Atmosphere Monitoring Service Information (2019). Usage of CAMS NRT data in 6 hour resolution for ozone, water vapor, AOT and surface temperature. Retrieved via API available from
webpage:
https://confluence.ecmwf.int/display/WEBAPI/Accessing+ECMWF+data+servers+in+batch.
Copernicus (2019). Copernicus COP-DEM-GLO-90 dataset, produced using Copernicus WorldDEM $^{\text {TM }}-90$ © DLR e.V. 2010-2014 and © Airbus Defence and Space GmbH 2014-2018 provided under COPERNICUS by the European Union and ESA; all rights reserved.

Dierckx, W., Sterckx, S., Benhadj, I., Livens, S., Duhoux, G., Van Achteren, T., Francois, M., Mellab, K., Saint, G. (2014). PROBA-V mission for global vegetation monitoring: standard products and image quality. International Journal of Remote Sensing, 35:7, 2589-2614, DOI: 10.1080/01431161.2014.883097.

Dinerstein, E., Olson, D., Joshi, A., Vynne, C., Burgess, N. D., Wikramanayake, E., Hahn, N., Palminteri, S., Hedao, P., Noss, R., Hansen, M., Locke, H., Ellis, E. C., Jones, B., Barber, C. V., Hayes, R., Kormos, C., Martin, V., Crist, E., Sechrest, W., Price, L., Baillie, J. E. M., Weeden, D., Suckling, K., Davis, C., Sizer, N., Moore, R., Thau, D., Birch, T., Potapov, P., Turubanova, S., Tyukavina, A., de Souza, N., Pintea, L., Brito, J. C., Llewellyn, O. A., Miller, A. G., Patzelt, A., Ghazanfar, S. A., Timberlake, J., Klöser, H., Shennan-Farpón, Y., Kindt, R., Lillesø, J. B., van Breugel, P., Graudal, L., Voge, M., Al-Shammari, K. F., Saleem, M. (2017). An Ecoregion-Based Approach to Protecting Half the Terrestrial Realm, BioScience, Volume 67, Issue 6, Pages 534-545, DOI: org/10.1093/biosci/bix014.

Dwyer, J.L., Roy, D.P., Sauer, B., Jenkerson, C.B., Zhang, H.K., Lymburner, L. (2018). Analysis Ready Data: Enabling Analysis of the Landsat Archive. Remote Sens, 10, 1363.

Eberenz, J., Verbesselt, J., Herold, M., Tsendbazar, N.-E., Sabatino, G., Rivolta, G. (2016). Evaluating the Potential of PROBA-V Satellite Image Time Series for Improving LC Classification in Semi-Arid African Landscapes. Remote Sens., 8, 987. DOI: 10.3390/rs8120987.

EEA (2012). Corine land cover classification 2012. European Environment Agency (EEA) under the framework of the Copernicus programme. Webpage: https://land.copernicus.eu/pan-european/corine-land-cover/clc-2012.

ESA (2018). Sentinel-2 tiling grid dataset. Webpage: https://sentinel.esa.int/documents/247904/1955685/S2A OPER GIP TILPAR MPC 2015120 9 T095117 V20150622T000000 21000101 T000000 B00.kml. Retrieved: January 15, 2018.

Fang, X.; Zhu, Q.; Ren, R.; Chen, H.; Wang, K.; Peng, C. (2018), Large-scale detection of vegetation dynamics and their potential drivers using MODIS images and BFAST: A case study in Quebec, Canada, Remote Sensing of Environment, Volume 206, 2018, Pages 391-402, DOI: https://doi.org/10.1016/j.rse.2017.11.017.

FAO (2012). Global ecological zones for FAO forest reporting: 2010 Update. Forest Resources Assessment Working Paper 179, FAO, Rome, Italy.

Francois, M., Santandrea, S., Mellab, K., Vrancken, D., Versluys, J. (2014). The PROBA-V mission: the space segment, International Journal of Remote Sensing, 35:7, 2548-2564, DOI: 10.1080/01431161.2014.883098.

Gao, B. C. (1996). NDWI-a normalized difference water index for remote sensing of vegetation liquid water from space. Remote Sensing of Environment, 58: 257-266.

Gong, W.; Fang, S.; Yang, G.; Ge, M. (2017). Using a Hidden Markov Model for Improving the Spatial-Temporal Consistency of Time Series Land Cover Classification. ISPRS Int. J. Geo-Inf. 2017, 6, 292, DOI: 10.3390/ijgi6100292.
Hermosilla, T.; Wulder, M.A.; White, J.C.; Coops, N.C.; Hobart, G.W. (2018): Disturbance-Informed Annual Land Cover Classification Maps of Canada's Forested Ecosystems for a 29-Year Landsat Time Series, Canadian Journal of Remote Sensing, 44:1, 67-87, DOI: 10.1080/07038992.2018.1437719.

Ho, Tin Kam (1995). Random Decision Forests. Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14-16 August 1995. pp. 278-282. Archived from the original (PDF) on 17 April 2016. Retrieved 5 June 2016.

Homer, C.G., Dewitz, J.A., Yang, L., Jin, S., Danielson, P., Xian, G., Coulston, J., Herold, N.D., Wickham, J.D., and Megown, K., 2015, Completion of the 2011 National Land Cover Database for the conterminous United States-Representing a decade of land cover change information. Photogrammetric Engineering and Remote Sensing, v. 81, no. 5, p. 345-354, DOI: 10.14358/PERS.81.5.345.

Howat, I.M., A. Negrete, B.E. Smith, 2014, The Greenland Ice Mapping Project (GIMP) land classification and surface elevation datasets, The Cryosphere, 8, 1509-1518, DOI: 10.5194/tc-8-1509-2014.

Huete, A., et al. (2002). Overview of the Radiometric and Biophysical Performance of the MODIS Vegetation Indices. Remote Sensing of Environment, 83,195-213. DOI: 10.1016/S0034-4257(02)00096-2.

Huffman, G.J., E.F. Stocker, D.T. Bolvin, E.J. Nelkin, Jackson Tan (2019), GPM IMERG Final Precipitation L3 1 month 0.1 degree x 0.1 degree V06, Greenbelt, MD, Goddard Earth Sciences Data and Information Services Center (GES DISC), Accessed: 2020/06/01, 10.5067/GPM/IMERG/3B-MONTH/06

Hughes, G. F. (1968). On the Mean Accuracy of Statistical Pattern Recognizers. IEEE Transactions on Information Theory, IT-14:55-63. DOI: 10.1109/TIT.1968.1054102.

Kalman, R. E. (1960). A New Approach to Linear Filtering and Prediction Problems. Journal of Basic Engineering. 82: 35. DOI: 10.1115/1.3662552.

Kapos, V., Rhind, J., Edwards, M., Prince, M. and Ravilious, C. (2000). Developing a map of the world's mountain forests. In: M. Price and N. Butt (eds), Forests in Sustainable Mountain Development, IUFRO Research Series 5, CABI Publishing, Wallingford, UK; pp. 4-9.

Kempeneers, P., Sedano, F., Piccard, I., Eerens, H. (2016). Data Assimilation of PROBA-V 100 and 300 m . In IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 9, no. 7, pp. 3314-3325. DOI: 10.1109/JSTARS.2016.2527922.

Khanna, S., Palacios-Orueta, A., Whiting, M. L., Ustin, S. L., Riaño, D., Litago, J. (2007). Development of angle indexes for soil moisture estimation, dry matter detection and land-cover discrimination. Remote Sensing of Environment, 109.2, 154-165, DOI: 10.1016/j.rse.2006.12.018.

Kursa, M.B. (2017): Boruta. Webpage: https://m2.icm.edu.pl/boruta/. Retrieved: July 17, 2017.
Kursa, M.B., Rudnicki, W.R., (2010). Feature Selection with the Boruta Package. Journal of Statistical Software, [S.I.], v. 36, Issue 11, p. 1-13. DOI: $10.18637 / j$ jss.v036.i11.

Lesiv, M., Fritz, S., and See, L. (2017). A New Open Reference Global Dataset for Land Cover Mapping at a 100 m Resolution (Frascati, Italy).

Leys, C., et al. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. Journal of Experimental Social Psychology, Volume 49, Issue 4, pp. 764-766. DOI: 10.1016/i.jesp.2013.03.013.

Liu, D.; Kelly, M.; Gong, P. (2006). A spatial-temporal approach to monitoring forest disease spread using multi-temporal high spatial resolution imagery, Remote Sensing of Environment, Volume 101, Issue 2, Pages 167-180, DOI: https://doi.org/10.1016/j.rse.2005.12.012.

Marconcini, M., Uereyen, S., Esch, Y., Metz, A., Zeidler, J. (2017). Mapping urban areas globally by jointly exploiting optical and radar imagery - the GUF+2015 layer. Presentation at the Worldcover 2017 Conference, ESA-ESRIN, Frascati, Italy, 14-16 March 2017.

Marconcini, M., Metz-Marconcini, A., Üreyen, S., Palacios-Lopez, D., Hanke, W., Bachofer, F., Zeidler, J., Esch, T., Gorelick, N., Kakarla, A., Strano, E. (in review). Outlining where Humans Live - The World Settlement Footprint 2015. Electrical Engineering and Systems Science. https://arxiv.org/abs/1910.12707v1

Masiliunas, D., Tsendbazar, N.-E., Herold, M., Lesiv, M., Verbesselt, J., Buchhorn, M. (2020, in review). Fractional land cover classification method assessment using PROBA-V satellite data. Remote Sensing of Environment.

NASA Earth Observatory (2019): webblog „Cloudy Earth". Retrieved: October 4 ${ }^{\text {th }}, 2019$ from webpage: https://earthobservatory.nasa.gov/images/85843/cloudy-earth.

Niraj (2016): Data Science, ML \& AI Tutorials - Machine Learning. Webpage: https://sites.google.com/site/nirajatweb/

North American Land Change Monitoring System (NALCMS). North American Land cover at 30m. Collaborative initiative between Natural Resources Canada/Canada Centre for Mapping and

Earth Observation (NRCan/CCMEO); the United States Geological Survey (USGS); and three Mexican organizations: Insituto Nacional de Estadística y Geografía (INEGI); Comisión Nacional para el Conocimiento y Uso de la Biodiversidad (CONABIO); and Comisión Nacional Forestal (CONAFOR). The project is facilitated by the Commission for Environmental Cooperation (CEC). Webpage: http://www.cec.org/tools-and-resources/map-files/land-cover-2010-landsat-30m.
Olofsson, P., Stehman, S. V., Woodcock, C. E., Sulla-Menashe, D., Sibley, A. M., Newell, J. D., Friedl, M. A., and Herold, M., (2012). A global land-cover validation data set, part I: fundamental design principles, International Journal of Remote Sensing, 33:18, 5768-5788, DOI: 10.1080/01431161.2012.674230.

OpenStreetMap contributors. (2015) Planet dump [Data file from 2020-01-01]. Retrieved from https://planet.openstreetmap.org

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, p. 2825-2830.

Pekel, J.F., Cottam, A., Gorelick, N., Belward, A.S. (2016). High-resolution mapping of global surface water and its long-term changes. Nature 540, 418-422. DOI: 10.1038/nature20584.

Pekel, J.-F., Vancutsem, C., Bastin, L., Clerici, M., Vanbogaert, E., Bartholomé, E., \& Defourny, P. (2014). A near real-time water surface detection method based on HSV transformation of MODIS multi-spectral time series data. Remote Sensing of Environment, 140, 704-716. DOI: 10.1016/j.rse.2013.10.008.

Pfeffer, W. T., Arendt, A. A., Bliss, A., Bolch, T., Cogley, J. G., Gardner, A. S., Hagen, J.-O., Hock, R., Kaser, G., Kienholz, C., Miles, E. S., Moholdt, G., Mölg, N., Paul, F., Radic, V., Rastner, ' P., Raup, B. H., Rich, J., and Sharp, M. J. (2014). The Randolph Glacier Inventory: a globally complete inventory of glaciers. Journal of Glaciology 60:537-552, DOI: 10.3189/2014JoG13J176.

Proud, S. R., Fensholt, R., Rasmussen, M. O., and Sandholt, I. (2010b). A comparison of the effectiveness of 6 S and SMAC in correcting for atmospheric interference of Meteosat Second Generation images, J. Geophys. Res., 115, D17209, DOI: 10.1029/2009JD013693.

Proud, S. R., Rasmussen, M. O., Fensholt, R., Sandholt, I., Shisanya, C., Mutero, W., Mbow, C., Anyamba, A. (2010a). Improving the SMAC atmospheric correction code by analysis of Meteosat Second Generation NDVI and surface reflectance data. Remote Sensing of Environment, Volume 114, Issue 8, Pages 1687-1698, DOI: 10.1016/j.rse.2010.02.020.

Rahman, H.; Dedieu, G. (1994). SMAC: a simplified method for the atmospheric correction of satellite measurements in the solar spectrum. International Journal of Remote Sensing, 15(1): 123-143. DOI: $10.1080 / 01431169408954055$.

Roerink, G. J.; Menenti, M.; Verhoef, W. (2000). Reconstructing cloudfree NDVI composites using Fourier analysis of time series. International Journal of Remote Sensing, 21 (9), pp. 1911-1917. DOI: 10.1080/014311600209814.

Santos, Teresa. (2011). A METHODOLOGY TO PRODUCE GEOGRAPHICAL INFORMATION FOR LAND PLANNING USING VERY-HIGH RESOLUTION IMAGES. available at: https://run.unl.pt/handle/10362/19737

Sayre, R., Comer,P., Hak, J., Josse, C., Bow, J., Warner, H., Larwanou, M., Kelbessa, E., Bekele, T., Kehl, H., Amena, R., Andriamasimanana, R., Ba, T., Benson, L., Boucher, T., Brown, M., Cress, J., Dassering, O., Friesen, B., Gachathi, F., Houcine, S., Keita, M., Khamala, E., Marangu, D., Mokua, F., Morou, B., Mucina, L., Mugisha, S., Mwavu, E., Rutherford, M., Sanou, P., Syampungani, S., Tomor, B., Vall, A., Vande Weghe, J., Wangui, E, Waruingi, L. (2013). A New Map of Standardized Terrestrial Ecosystems of Africa. Washington, DC: Association of American Geographers, 24 pages.

Scikit-learn (sklearn) (2017). User guide for ensemble module: ensemble methods. Webpage: http://scikit-learn.org/stable/modules/ensemble.html\#forest. Retrieved: July 18, 2017.

Sedano, F., Kempeneers, P., Hurtt, G. (2014). A Kalman Filter-Based Method to Generate Continuous Time Series of Medium-Resolution NDVI Images. Remote Sens., 6, 12381-12408. DOI: 10.3390/rs61212381.

Sterckx, S., Benhadj, I., Duhoux, G., Livens, S., Dierckx, W., Goor, E., Adriaensen, S., Heyns, W., Van Hoof, K., Strackx, G., Nackaerts, K., Reusen, I., Van Achteren, T., Dries, J., Van Roey, T., Mellab, K., Duca, R., Zender, J. (2014). The PROBA-V mission: image processing and calibration. International Journal of Remote Sensing, 35:7, 2565-2588. DOI: 10.1080/01431161.2014.883094.

Tucker, C.J. (1979). Red and Photographic Infrared Linear Combinations for Monitoring Vegetation. Remote Sensing of Environment, 8(2),127-150. DOI: 10.1016/0034-4257(79)90013-0.

Verbesselt, J.; Zeileis, A.; Herold, M. (2012). Near real-time disturbance detection using satellite image time series, Remote Sensing of Environment, Volume 123, Pages 98-108, DOI: https://doi.org/10.1016/j.rse.2012.02.022.

Verhoef, W. (1996). Application of Harmonic Analysis of NDVI Time Series (HANTS). In Fourier Analysis of Temporal NDVI in the Southern African and American Continents, edited by S. Azzali and M. Menenti, DLO Winand Staring Centre, Wageningen, The Netherlands, Report 108, pp. 19-24.
von Mehren, M.; Gieseke, F.; Verbesselt, J.; Rosca, S.; Horion, S.; Zeileis, A. (2018). Massivelyparallel break detection for satellite data. In Proceedings of the 30th International Conference on Scientific and Statistical Database Management (SSDBM '18). Association for Computing Machinery, New York, NY, USA, Article 5, 1-10. DOI:https://doi.org/10.1145/3221269.3223032

Wagner, J. E.; Stehman, S.V. (2015). Optimizing sample size allocation to strata for estimating area and map accuracy, Remote Sensing of Environment, 168, 126-133. DOI: 10.1016/j.rse.2015.06.027.

Walker, H. (1931). Studies in the History of the Statistical Method. Baltimore, MD: Williams \& Wilkins Co. pp. 24-25

Watts, L.M. and Laffan, S.W. (2014). Effectiveness of the BFAST algorithm for detecting vegetation response patterns in a semi-arid region, Remote Sensing of Environment, Volume 154, 2014, Pages 234-245, DOI: https://doi.org/10.1016/j.rse.2014.08.023.

Welch, G., Bishop, G. (2006). An Introduction to the Kalman Filter. Report, University of North Carolina at Chapel Hill, Department of Computer Science, TR 95-041.

Wessel, P., Smith, W. H. F. (1996), A global, self-consistent, hierarchical, high-resolution shoreline database, J. Geophys. Res., 101(B4), 8741-8743, DOI: 10.1029/96JB00104.

Zeileis, A. (2005). A Unified Approach to Structural Change Tests Based on ML Scores, F Statistics, and OLS Residuals, Econometric Reviews, 24:4, 445-466, DOI: 10.1080/07474930500406053

Zeileis, A. (2006). Implementing a class of structural change tests: An econometric computing approach, Computational Statistics \& Data Analysis, Volume 50, Issue 11, Pages 2987-3008, DOI: https://doi.org/10.1016/j.csda.2005.07.001.

Zeileis, A.; Kleiber, C.; Krämer, W.; Hornik,K. (2003). Testing and dating of structural changes in practice, Computational Statistics \& Data Analysis, Volume 44, Issues 1-2, Pages 109-123, DOI: https://doi.org/10.1016/S0167-9473(03)00030-6.


[^0]:    Issue: I3.4

