

Copernicus Global Land Operations “Vegetation and Energy”

”CGLOPS-1”

Framework Service Contract N° 199494 (JRC)

ALGORITHM THEORETICAL BASIS DOCUMENT

MODERATE DYNAMIC LAND COVER

COLLECTION 100 M

VERSION 2

Issue I2.00

Organization name of lead contractor for this deliverable: VITO

Book Captain: Dr. Marcel Buchhorn (VITO)

Contributing Authors: Luc Bertels (VITO)


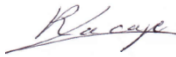
Bruno Smets (VITO)

Myroslava Lesiv (IIASA)

Nandin-Erdene Tsendbazar (WUR)

Dissemination Level		
PU	Public	X
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:	Marcel Buchhorn	Sign 	Date 26.08.2019
Approval:	Roselyne Lacaze	Sign 	Date 25.11.2019
Endorsement:	Michael Cherlet	Sign	Date
Distribution:	Public		

Change Record

Issue/Rev	Date	Page(s)	Description of Change	Release
	25.09.2017	All	Initial Version	I1.00
	09.10.2017	All	Revision after internal review	I1.10
I1.10	26.08.2019	All	Version 2 of algorithm	I2.00

TABLE OF CONTENTS

Executive Summary	17
1 Background of the document.....	18
1.1 Scope and Objectives.....	18
1.2 Content of the document.....	18
1.3 Related documents	18
1.3.1 Applicable documents	18
1.3.2 Input.....	18
1.3.3 Output.....	19
1.3.4 External documents	19
2 Review of Users Requirements.....	20
3 Methodology Description	25
3.1 Overview	25
3.2 Input Data.....	27
3.2.1 PROBA-V L1C archive	27
3.2.2 Land/sea mask datasets.....	28
3.2.3 Sentinel-2 tiling grid	30
3.2.4 ECMWF CAMS NRT datasets.....	30
3.2.5 Glacier and Ice Sheets datasets	31
3.2.6 Global Mountain Explorer dataset.....	31
3.2.7 Global ecological region datasets	32
3.2.8 Circum-Polar Arctic Vegetation Map dataset	32
3.2.9 DLR's World Settlement Footprint dataset.....	33
3.2.10 JRC's Global Surface Water (GSW) dataset	33
3.2.11 Digital Elevation Model datasets.....	34
3.3 PROBA-V UTM Analysis Ready Data (ARD) Generation	35
3.3.1 Overview	35
3.3.2 Yearly Bright Surface and Permanent Snow detection	36
3.3.3 Level 2A processing and Level 2A post-processing.....	40
3.3.4 Level 2B processing.....	44
3.3.5 Level 3 processing and final status mask	47
3.3.6 Comparison PROBA-V versus PROBA-V UTM.....	49
3.4 Data Cleaning & Compositing.....	51
3.4.1 Overview	51
3.4.2 The HANTS and madHANTS algorithms	52
3.4.3 Median Composites (MC) generation.....	54
3.4.4 Long-term 5-daily Median Composite (HMC5) generation	57

3.5	Input Data Quality Indicator Generation	57
3.5.1	Overview	57
3.5.2	The PROBA-V theoretical observation mask and time series gap mask	59
3.5.3	Calculating the time series statistics	61
3.6	Data Fusion	66
3.6.1	Overview	66
3.6.2	Data Fusion pre-processing	68
3.6.3	Principle of the Kalman-Filtering algorithm	69
3.6.4	Data Fusion implementation using VITO's enhanced Kalman-Filtering approach	70
3.6.5	Data Fusion post-processing	75
3.7	Ancillary Dataset Products Generation	76
3.7.1	Global Biome Clusters product	77
3.7.2	WSF Built-up cover fraction product generation	79
3.7.3	WetProducts permanent and seasonal cover fraction products generation	79
3.8	Metrics Generation	81
3.8.1	Overview	81
3.8.2	Pre-Processing	82
3.8.3	Metrics extraction	88
3.8.4	Post-Processing	91
3.9	Training Data Generation	92
3.10	Classification / Regression	94
3.10.1	Overview	94
3.10.2	The Random Forest Approach	95
3.10.3	Training data and classifier/regressor optimization	95
3.10.4	Scenario-based Classification	98
3.10.5	Scenario-based Regression	98
3.11	CGLS-LC100 Cover Fraction Products Generation	99
3.11.1	Overview	99
3.11.2	Cover fraction products generation	102
3.12	CGLS-LC100 Land Cover Map Generation	105
3.12.1	Overview	105
3.12.2	Discrete map generation via expert rules	108
3.13	CGLS-LC100 Products	116
4	Limitations	122
5	Risk of failure and Mitigation measures	124
6	References	125

List of Figures

Figure 1: Workflow diagram for the CGLS Dynamic Land Cover layers at 100 m resolution (CGLS-LC100 products) generation.....	26
Figure 2: PROBA-V L1C processing chain flowchart (adapted from [PROBA-V User Manual]).....	28
Figure 3: The USGS shoreline vector layer (Sayre et al., 2013). Zoom in to the Street 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.....	29
Figure 4: Global shoreline layer generated out of the GSHHG and USGS datasets (Wessel et al., 1996; Sayre et al., 2013).....	29
Figure 5: 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.	30
Figure 6: Global Mountainous Area Mask (MAM) showing mountain areas in white. Note: for better visualization the global shoreline layer is overlaid in red.	31
Figure 7: Circum-Polar Arctic Bioclimate Subzones. Image derived from http://www.ArcticAtlas.org	32
Figure 8: The DLR World Settlement Footprint layer (Marconcini et al., 2017b). Zoom in to Street of Gibraltar showing the urban areas in black. Note: in red the shoreline vector layer is shown.	33
Figure 9: The JRC Global Surface Water product (Pekel et al., 2016). Zoom in to the Street of Gibraltar showing the maximum water extent layer (left) and the water seasonality mask for 2015 (right). Note: in red the USGS shoreline vector layer is shown in the maximum water extent layer (left image).....	34
Figure 10: The NASA Shuttle Radar Topography Mission Global 1 arc second dataset. Zoom in to the Street of Gibraltar. Note: in red the USGS shoreline vector layer is shown.	34
Figure 11: Workflow chart showing the generation of the PROBA-V UTM Analysis Ready Data (ARD) archive out of the PROBA-V L1C archive.....	35
Figure 12: Yearly Bright Surface and Permanent Snow detection – Part A: Raw mask generation using the PROBA-V L2A archive.....	37
Figure 13: Distribution of global test sites used in the PROBA-V UTM ARD generation. Test sites following the Sentinel-2 naming convention.	38
Figure 14: Yearly Bright Surface and Permanent Snow detection – Part B: Mask cleaning and final Sentinel-2 tile generation.	39
Figure 15: 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.	39
Figure 16: 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.	41
Figure 17: PROBA-V UTM L2A post-processing – Part A: improved snow and cloud detection – raw detection by thresholding.	42
Figure 18: PROBA-V UTM L2A post-processing – Part B: improved snow and cloud detection – cleaning by masks and resolving confusions.....	43
Figure 19: PROBA-V UTM L2A post-processing – Part C: updating of status mask.....	43
Figure 20: PROBA-V UTM L2B processor – improved SMAC routine and quality indicator generation for atmospheric correction.....	44
Figure 21: 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.....	46
Figure 22: PROBA-V UTM L3 processing – generation of daily synthesis. From left to right: partly covered tile in first and second overpass, daily synthesis, and Google Earth image for comparison.	48
Figure 23: Comparison between PROBA-V and PROBA-V UTM cloud/snow detection.	50
Figure 24: Comparison between PROBA-V (left) and PROBA-V UTM (right) geometric correction (L2A). Already in mid-latitudes a reduced distortion is viewable. Area shown: province Niedergoersdorf, Teltow-Flaeming, Germany.....	51
Figure 25: 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.	52
Figure 26: Example for the madHANTS temporal outlier detection algorithm. 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 3.5 standard deviations away from the harmonized pixel value.....	54
Figure 27: Example showing the temporal outlier detection using the madHANTS algorithm on the blue and SWIR reflectance values of a pixel at location 9.459° lon, 6.562° 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.	55
Figure 28: 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).	56
Figure 29: Exemplary temporal profiles for PROBA-V UTM ARD+ time series with a high Data Density Indicator (DDI) (upper graph) and low DDI (lower graph).	58
Figure 30: Data Density Indicator processing workflow.	59
Figure 31: 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.	60
Figure 32: 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.	62
Figure 33: The cleaned time series statistics – Part A.	63
Figure 34: The cleaned time series statistics – Part B.	64
Figure 35: Final scoring of time series statistics to compute the Data Density Indicator (DDI) and output of the PROBA-V UTM ARD+ statistics.	65
Figure 36: Visual RGB representation of UTM tile 32NPM (2016) with some of the calculated DDI statistics for the 100 m resolution, i.e. Data Density Indicator, Observation ratio, Observation frequency, Gap weight and Longest concurrent gap.	65
Figure 37: 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.	66
Figure 38: 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.	67
Figure 39: 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-daily observations.	68
Figure 40: Example for data fusion pre-processing results for PROBA-V UTM TOC-MC5 300 m time series at pixel location 9.459° lon, 6.562° 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 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 data fusion).	69
Figure 41. Basic concept of the Kalman filtering approach. Adapted from Aimonen (2011).	70

Figure 42: 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).	71
Figure 43: Example for 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. ...	72
Figure 44: Visualization of data usion with the improved Kalman filter for one 3x3 100 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 3x3 pixel area, respectively (Note: shown is forward (top), backward (center) and smooth (bottom) mode of Kalman filter)	74
Figure 45: 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.	76
Figure 46: Global stratification into 73 biome clusters to improve regional classification/regression in the CGLS-LC100 workflow.	78
Figure 47: General overview of the water bodies detection algorithm. Besides the PROBA-V UTM derived VI data products, i.e. HUE, VALUE and NDVI, three ancillary layers are used as input for the WetProducts generation.....	81
Figure 48: General overview of the metrics generation section 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.	82
Figure 49: 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.	86
Figure 50: 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.	87
Figure 51: 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.	89

Figure 52: Examples of the 270 derived metrics for tile X18Y06. 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).	92
Figure 53: Training dataset used in the CGLS-LC100 workflow showing the ~141K training locations together with their sampled land cover class at 100 m.	94
Figure 54: The Random Forest classifier principle. Image by Niraj (2016).	95
Figure 55: 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.	96
Figure 56: General overview of the cover fraction layers generation section in the CGLS-LC100 product workflow.	101
Figure 57: Rule block 1 to 4 of the CGLS-LC100 cover fraction layer generation – generation of snow, built-up, permanent and seasonal water cover fraction and cleaning by CL1 and CAVM.	103
Figure 58: Rule block 5 and 6 of the CGLS-LC100 cover fraction layer generation – cover fraction normalization and cleaning by the land/sea mask.	105
Figure 59: General overview of the Land Cover map generation section in the CGLS-LC100 product workflow.	106
Figure 60: Flowchart of the classification post-processing to produce additional input layers for the CGLS-LC100 discrete map generation.	107
Figure 61: Initialization and Expert Rule block 1 of the CGLS-LC100 discrete map and map probability layer generation – imprinting hard masks.	110
Figure 62: Expert Rule block 2 to 4 of the CGLS-LC100 discrete map and map probability layer generation – imprinting the derived cover fraction layers for classes, resolving classification scenarios, and filling of the remaining land pixels.	111
Figure 63: Expert Rule block 5 part A of the CGLS-LC100 discrete map and map probability layer generation – resolving confusion between closed forest, open forest, shrub and grassland on continental level. *See Figure 46 and Table 6 for biome cluster location and naming.	114
Figure 64: Expert Rule block 5 part B of the CGLS-LC100 discrete map and map probability layer generation – resolving confusion between closed forest, open forest, shrub and grassland on continental level.	115
Figure 65: Expert Rule block 6 and 7 of the CGLS-LC100 discrete map and map probability layer generation – wetland class improvements and final separation of the open/close forest class by forest type.	116

Figure 66: The CGLS-LC100 discrete map product for epoch 2015 with 23 discrete classes (detailed legend in Figure 68) 118

Figure 67: 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..... 119

Figure 68: 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)..... 120

List of Tables

Table 1: Usefulness of information on LC and LC change processes for different international actions and programmes.....	21
Table 2: List of land cover classes requested by users.	21
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).	28
Table 4: Status mask bit mapping of the PROBA-V L2A data	40
Table 5: PROBA-V UTM L3 status mask. New flags marked in bold.	48
Table 6: Global biome cluster by continents.....	79
Table 7: The 23 discrete classes of the CGLS-LC100 discrete product. Note: final classes are shown in bold together with numerical code in brackets.....	108
Table 8: Overview of existing CGLS-LC100 products produces by algorithm version	120

List of Acronyms

Acronym	Meaning
AD	Applicable document
AIST	Advanced Industrial Science and Technology
aNIR	Angle at NIR
AOD	Aerosol Optical Depth
AOT	Aerosol Optical Thickness
API	Application Programming Interface
ARD	Analysis Ready Data
ASTER	Advanced Spaceborne Thermal Emission and Reflection Radiometer
ATBD	Algorithm Theoretical Basis Document
AUC	Area Under the Curve
AVHRR	Advanced Very-High-Resolution Radiometer
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
DDI	Data Density Indicator
DEM	Digital Elevation Model
DLR	Deutsches Zentrum für Luft- und Raumfahrt
DOI	Digital Object Identifier
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
GHS	Global Human Settlement
GIMP	Greenland Ice Mapping Project
GIS	Geographic Information System
GLIMS	Global Land Ice Measurements from Space
GOME	Global Ozone Monitoring Experiment
GSD	Ground Sampling Distance

GSHHG	Global Self-consistent, Hierarchical, High-resolution Geography Database
GSW	Global Surface Water
GUF	Global Urban Footprint
GZD	Grid Zone Designators
HANTS	Harmonic Analysis of Time Series
HMC5	Harmonized 5-daily Median Composite
HSV	Hue Saturation Value colour system
HUE	Chromaticity
IIASA	International Institute for Applied Systems Analysis
IASI	Infrared Atmospheric Sounding Interferometer
JRC	Joint Research Center
LAI	Leaf Area Index
KML	Keyhole Markup Language
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
METI	Ministry of Economy, Trade, and Industry of Japan
MODIS	Moderate Resolution Imaging Spectroradiometer
NALCMS	North American Land Change Monitoring System
NASA	National Aeronautics and Space Administration
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
PGISM	Permanent Glacier and Ice Sheet Mask
PMAp	Polar Multi-Sensor Aerosol Properties Product
PPSIM	Potential Permanent Snow and Ice mask
PROBA-V	Vegetation instrument on board of PROBA satellite
PUM	Product User Manual
PVC	Percentage Vegetation Cover
RF	Random Forest classifier
RGB	Red-Green-Blue colour space
RMSE	Root Mean Square Error
RS	Remote Sensing

S1	1-daily synthesis
SAA	Sun Azimuth Angle
SIPI	Structure Intensive Pigment Index
SM	Status Map
SMAC	Simplified Model for Atmospheric Correction
SQL	Structured Query Language
SRTM	Shuttle Radar Topography Mission
SRTMGL1	Shuttle Radar Topography Mission Global 1 arc second
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
UN	United Nations
URD	User Requirements Document
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
WB	Water Bodies
WGS84	World Geodetic System 1984
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 1st 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. Essential Climate Variables like Leaf Area Index (LAI), the Fraction of Absorbed Photosynthetically Active Radiation (FAPAR), the surface albedo, the Land Surface Temperature, the soil moisture, the burnt areas, the areas of water bodies, and additional vegetation indices, are generated every hour, every day or every 10 days from Earth Observation satellite data. Production and delivery of the parameters are performed on a reliable, automatic and timely manner and are complemented by the constitution of long term time series.

The CGLS delivers a yearly 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 levels, 12 classes at level 1 and up to 23 classes at level 3, with classes according to the Land Cover Classification System (LCCS). 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 in Africa, climate modelling, etc.)

The first Land Cover map, produced with algorithm V1.0, was provided for the 2015 reference year over the African continent 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.

For the global up-scaling of the product, the algorithm was improved (Version 2) 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 (estimated by mid 2020).

This Algorithm Theoretical Basis Document (ATBD) describes the methods used for classifying the 2nd version land cover layers on PROBA-V data at 100 m resolution at global scale.

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 2 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-V2.0] 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 151-277962 of 7th 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 7th 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	Descriptor
CGLOPS1_SSD	Service Specifications Document of the Copernicus Global Land Service.
CGLOPS1_SVP	Service Validation Plan of the Copernicus Global Land Service.

CGLOPS1_URD_LC100m	User Requirements Document dynamic land cover 100m product
CGLOPS1_PSD_LC100m	Product Specifications Document of the dynamic land cover 100m product.
CGLOPS1_TrainingDataReport_LC100m	Report describing the training dataset used for Dynamic Land Cover 100m product

1.3.3 Output

Document ID	Descriptor
CGLOPS1_VR_LC100m-V2.0	Validation Report describing the results of the scientific quality assessment of the dynamic moderate land cover 100m version 2 product
CGLOPS1_PUM_LC100m-V2.0	Product User Manual Document of the dynamic moderate land cover 100m version 2 product
CGLOPS1_PUM_LCC100m-V2.1	Product User Manual of the land cover change detection maps over Africa (version 2.1)

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 http://www.vito-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 300m and/or 100m 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 100m and/or 300m and/or 1km.
 - The target baseline location accuracy shall be 1/3 of the at-nadir instantaneous field of view.
 - Pixel co-ordinates shall be given for centre of pixel.
- **Geographical coverage:**
 - geographic projection: lat long
 - geodetical datum: World Geodetic System 1984 (WGS84)
 - pixel size: minimum 1/112° - accuracy: min 10 digits
 - minimum global window coordinates:
 - Upper Left: 180°W-75°N
 - Bottom Right: 180°E, 56°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.

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 LC100m V2.0 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 land change processes	UNFCCC	UNCCD	OECD	SEEA/FAO	SDGs
1	Urban/built-up areas	Urbanization	✓	✓	✓	✓	✓
2	Cropland	Crop expansion	✓	✓	✓	✓	✓
3	Cropland and other vegetation	Land abandonment	✓	✓	✓	✓	✓
4	Forest	Deforestation	✓	✓	✓	✓	✓
5	Forest	Reforestation	✓	✓	✓	✓	✓
6	Wetland	Wetland degradation	✓	✓	✓	✓	✓
7	Water body	Expansion of water surface			✓	✓	✓
8	Water body	Reduction of water surface			✓	✓	✓
9	Bare areas	Desertification			✓	✓	✓

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

Code Level 1	Code Level 2	UN LCCS level	Land cover class	Forest modelling/REDD+	Crop monitoring	Biodiversity	Monitoring Environment and Security in Africa	Climate modelling	Class included in the product
10		A12A3A20B2	Forest/tree cover	X		X	X	X	Yes
	11	A12A3A20B2D 2E1	Evergreen Needleleaf forest	X			X	X	Yes
	12	A12A3A20B2D 1E1	Evergreen Broadleaf forest	X			X	X	Yes
	13	A12A3A20B2D 2E2	Deciduous Needleleaf forest	X			X	X	Yes
	14	A12A3A20B2D 1E2	Deciduous Broadleaf forest	X			X	X	Yes
	15	A12A3A20B2D 1D2	Mixed forest	X		X			Yes
	16	A12A3A10B2X	Intact forest	X		X		X	To map these classes,

Code Level 1	Code Level 2	UN LCCS level	Land cover class	Forest modelling/REDD+	Crop monitoring	Biodiversity	Monitoring Environment and Security in Africa	Climate modelling	Class included in the product
		XXX (assuming that an intact forest is a very dense forest)							addition research is required for methodology. Either we develop an expert rule based on other datasets such as: http://www.intactforest.org/data.ifl.html or new training and validation datasets. These classes could potentially be included in the next product evolutions.
	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		AA12A4A20B3(B9)	Shrub			X	X	X	Yes
	21	A12A4A20B(B9)XXE1	Evergreen shrubs			X			These classes could be potentially included in the next product evolutions. We will have to collect corresponding training and validation data.
	22	A12A4A20B3(B9)XXE2	Deciduous shrubs			X			
30		A12A2(A6)A20B4	Herbaceous vegetation			X	X	X	Yes
		A12A6A10 // A11A1A11B4X XXXXXF2F4F7 G4-F8	Pasture/managed grassland					X	To map these classes, addition research 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
		A122(A6)A10	Natural grassland			X		X	
		A12A2	Grass types for Western Africa			X			

Code Level 1	Code Level 2	UN LCCS level	Land cover class	Forest modelling/REDD+	Crop monitoring	Biodiversity	Monitoring Environment and Security in Africa	Climate modelling	Class included in the product
									evolutions.
		A12A3A11B2X XXXXXF2F4F7 G4-A12; A12A3A11B2- A13; A12A1A11	Savannas			X			The LC100 V2 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 100m x100m pixel may include less trees but still be considered as savanna. However, users are encouraged to use the fraction layers to produce their own savanna layer by applying specific thresholds for tree, shrub and grass cover.
40		A11A3	Cultivated and managed vegetation/agriculture		X	X	X	X	Yes
	41	A11A3XXXXXX D3(D9)	Irrigated cropland		X			X	To map these classes, addition research 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
	42	A11A3XXXXXX D1	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		X				

Code Level 1	Code Level 2	UN LCCS level	Land cover class	Forest modelling/REDD+	Crop monitoring	Biodiversity	Monitoring Environment and Security in Africa	Climate modelling	Class included in the product
			winter/summer crops						evolutions.
		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		A24A1(A2/A3/A4)	Wetland			X	X	X	Yes
		A24A3	Mangroves	X		X			To map this class, addition research 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.

Note that the change detection is tested (demonstrated) on the Africa continent only and, hence, is not included yet in this ATBD. Users are referred to the Product User Manual of the Africa Change maps to get more details on the current methodology used for change detection [CGLOPS1_PUM_LCC100m-V2.1]. With the release of the third version of product (planned mid 2020), this ATBD will include the final change detection methodology.

3 METHODOLOGY DESCRIPTION

3.1 OVERVIEW

The CGLS Dynamic Land Cover Map 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, and
5. Including established third party datasets via expert rules.

The workflow, shown in Figure 1, can be divided into the following sections:

1. PROBA-V UTM Analysis Ready Data (ARD) generation,
2. data cleaning & compositing,
3. quality indicator generation for PROBA-V UTM ARD+ input data,
4. data fusion,
5. metrics generation/extraction,
6. ancillary datasets products generation,
7. training data generation,
8. classification / regression,
9. cover fraction layers generation
10. land cover map generation plus final quality layer assembling.

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 (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 multi-spectral 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-daily time steps in the

100 m and 300 m PROBA-V EO time series. This is needed since the PROBA-V satellite provides a daily global coverage for the 300 m L3 product corresponding to 5-daily 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-daily time series a Data Density Indicator (DDI) is calculated which is used as an input quality indicator in the supervised learning process.

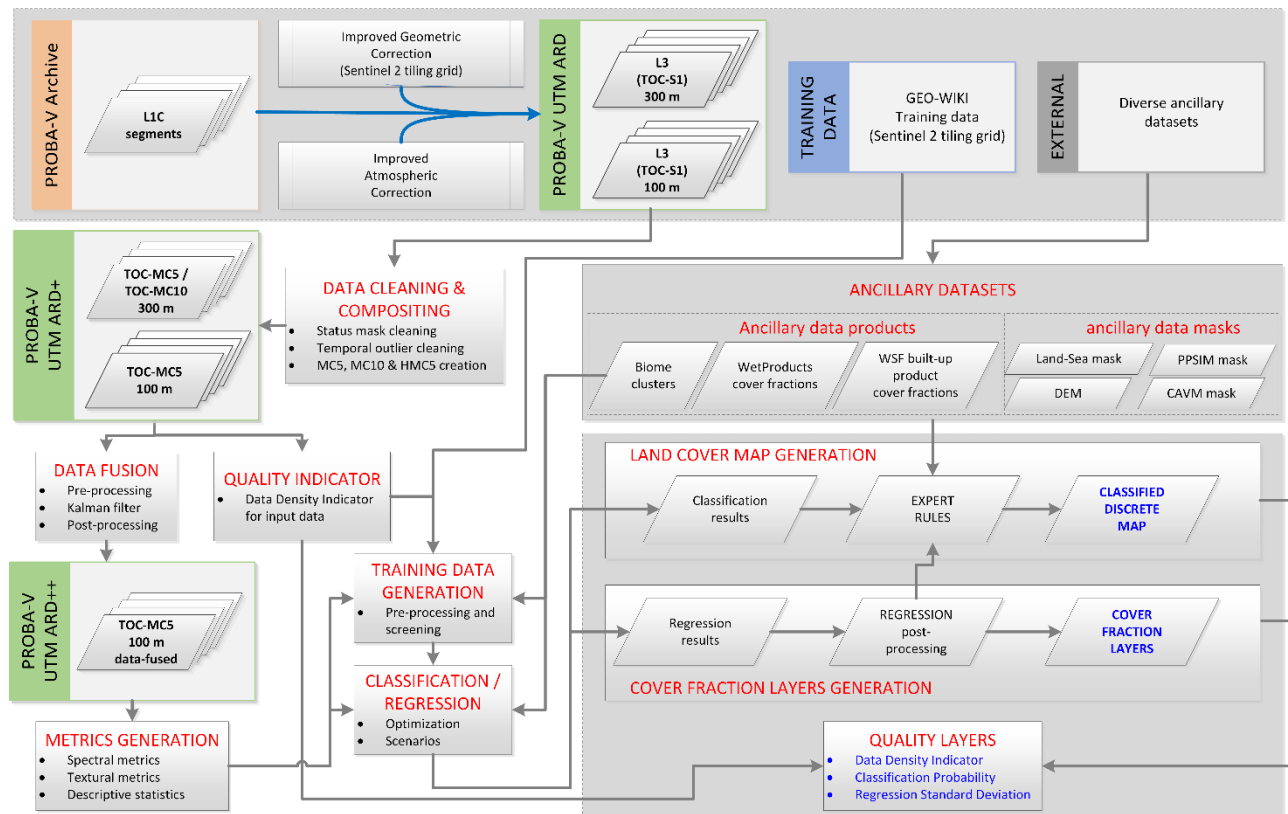


Figure 1: Workflow diagram for the CGLS Dynamic Land Cover 100 m (CGLS-LC100) products

Next, to improve the data density in the 5-daily 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 upgrades the PROBA-V UTM ARD+ archive into the PROBA-V UTM ARD++ archive, which gives temporal cleaned, consistent and dense 5-daily 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. Therefore, several metrics explaining the time series are extracted as input for the supervised classification and regression. Thus a harmonic model is fitted through the time steps of each of the reflectance bands of the time series data as well as each of the additional 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, descriptive statistics and textural metrics are generated. Overall, 270 metrics are extracted from the PROBA-V UTM ARD++ archive for the 2015 epoch.

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, retilled into the Sentinel-2 tiling grid, and post-processed to usable ancillary data products. 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. In the classification and regression preparation, the metrics of the training points 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) are input in the training of the supervised classifier/regressor using Random Forest (RF) techniques. The generated models are optimized via a 5-folded cross-validation in order to estimate the optimal classifier/regressor parameters.

Finally, we build upon the success of previous global mapping efforts and/or other ancillary datasets which are included via expert rules in the land cover map generation step. The produced land cover map uses 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. A novelty of this product is the generation of continuous cover fields that allow proportional estimates of cover fractions (also known as Percentage Vegetation Cover (PVC)) of all main land cover classes. The input are the cover fractions collected for all training points which are used in a Random Forest regression.

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 11 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 post-processing 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

Main EO input into the processing line is the PROBA-V level 1C (L1C) Collection 1 archive (May 2013 to present). 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 2.

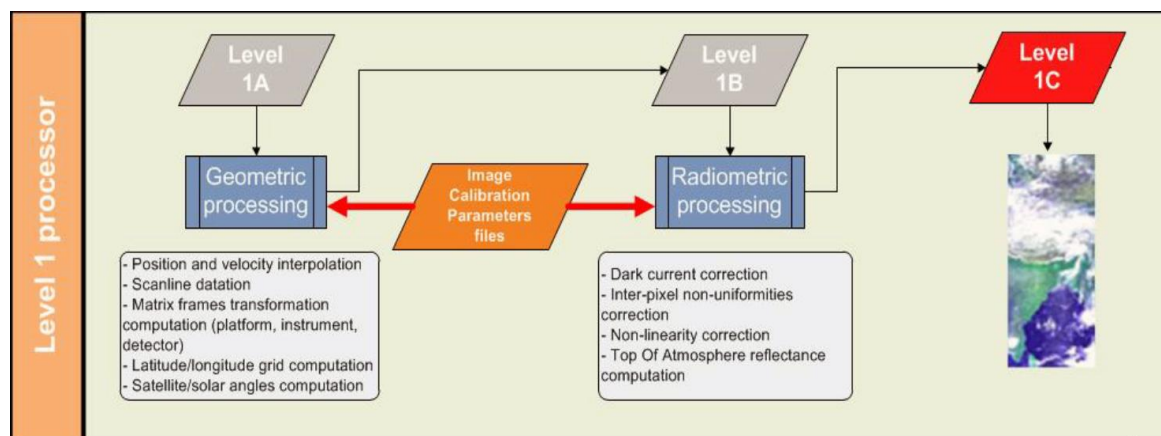


Figure 2: PROBA-V L1C processing chain flowchart (adapted from [PROBA-V User Manual])

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.

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 Maximum (FWHM)
BLUE	0.447 – 0.493 μm	0.460 μm	0.042 μm
RED	0.610 – 0.690 μm	0.658 μm	0.082 μm
NIR	0.770 – 0.893 μm	0.834 μm	0.121 μm
SWIR	1.570 – 1.650 μm	1.610 μm	0.089 μm

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 3 shows an example of the USGS shoreline vector layer for the Street of Gibraltar overlaid to a Google Earth image and the internal PROBA-V land/sea mask.

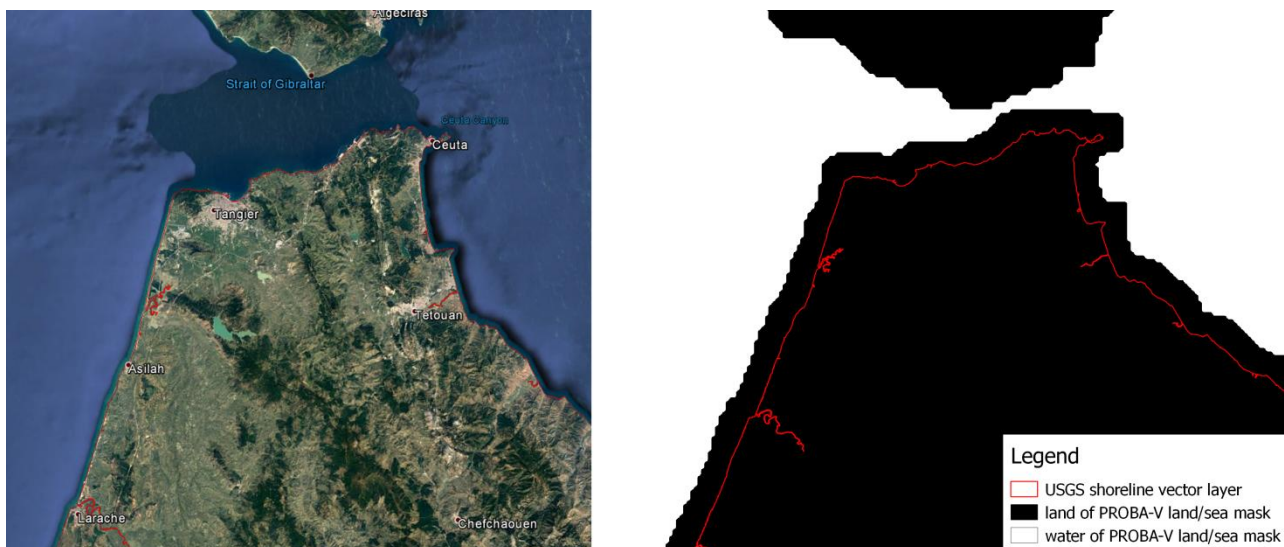


Figure 3: The USGS shoreline vector layer (Sayre et al., 2013). Zoom in to the Street of Gibraltar showing the shoreline vector layer (red) overlay to a Google Earth image (left) and to the PROBA-V land/sea mask (right). Note that the PROBA-V land/sea mask has a buffer around the land masses in order 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 as the GSHHG dataset, we fused the African continent shoreline of the Sayre et al. (2013) dataset into the GSHHG vector dataset (Figure 4).



Figure 4: 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 5 shows the derived vector file filtered to the global land masses using the shoreline layer described in section 3.2.2.

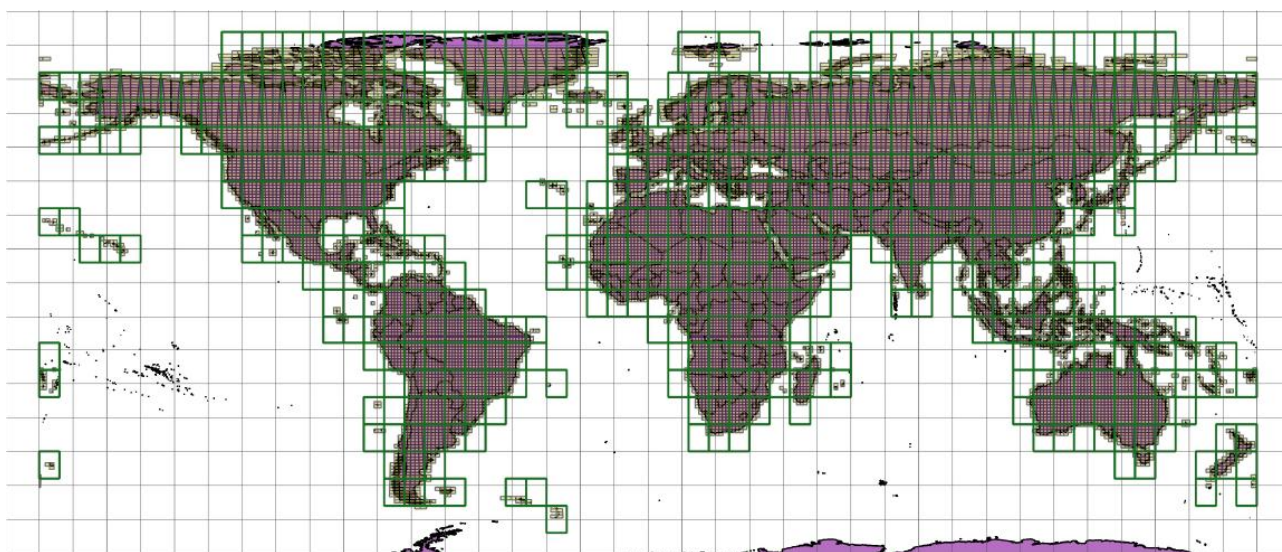


Figure 5: Sentinel-2 tiling grid overlaid on the global shoreline layer. Green rectangles show 6x8 deg UTM Grid Zone Designators (GZD's) overlaid on single 100,000-meter grid squares within the GZD's which form the single tiles.

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 x 0.125 degree spatial resolution and with a 6-hour temporal resolution:

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

3.2.5 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.

3.2.6 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.

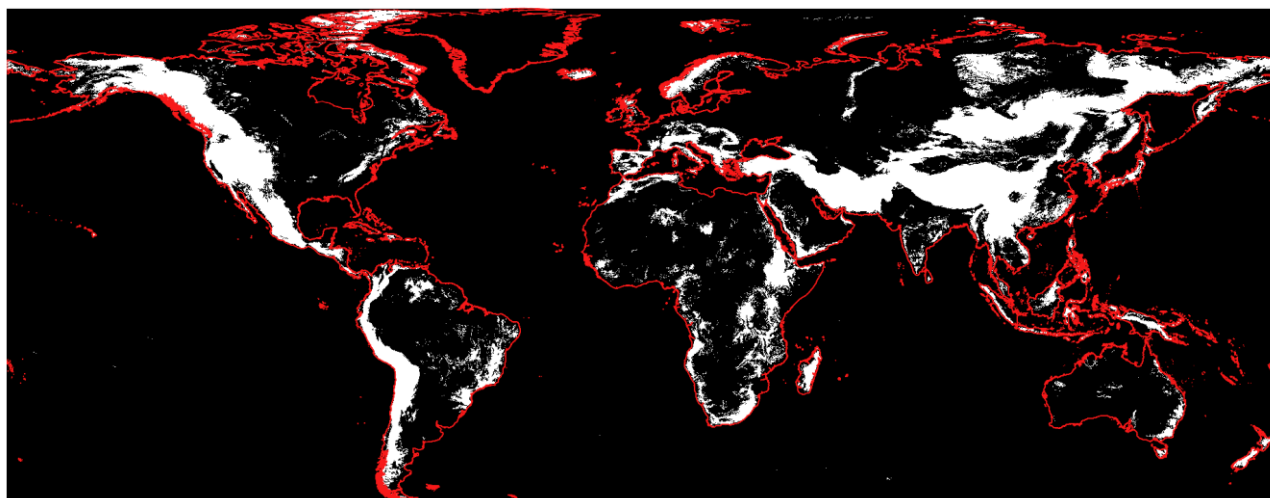


Figure 6: Global Mountainous Area Mask (MAM) showing mountain areas in white. Note: for better visualization the global shoreline layer is overlaid in red.

3.2.7 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 2017 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 3.2.11);
- Global ecological zones for Food and Agriculture Organization of the United Nations (FAO) forest reporting - update 2010 (FAO, 2012).

3.2.8 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 7).

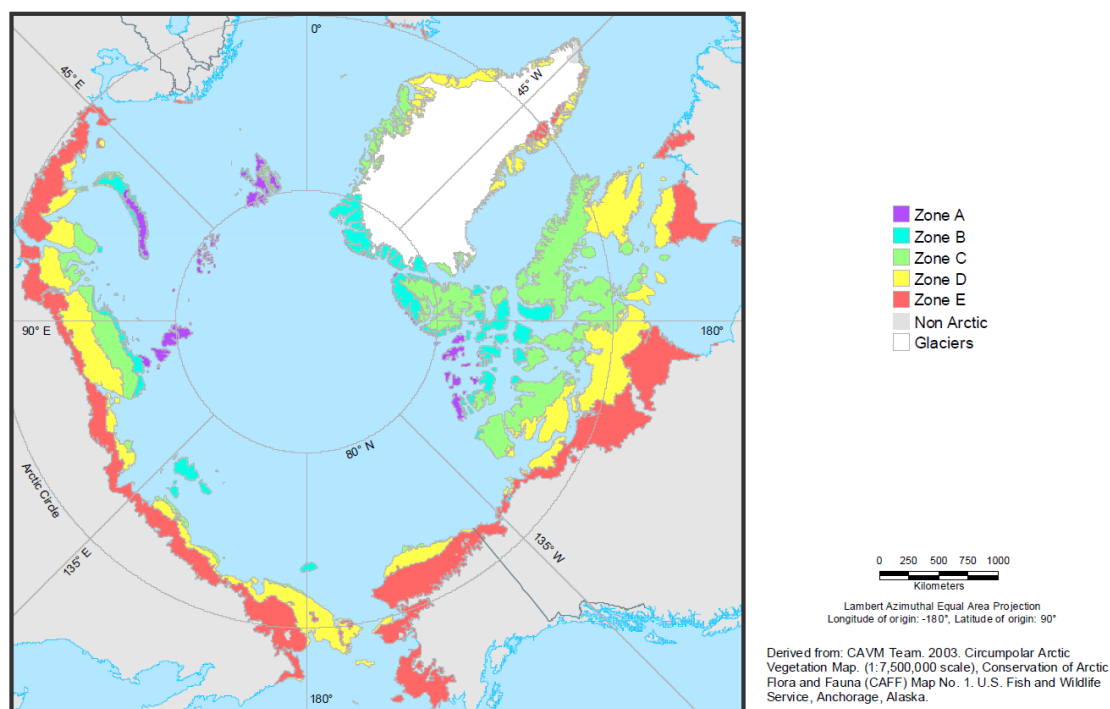


Figure 7: Circum-Polar Arctic Bioclimate Subzones. Image derived from <http://www.ArcticAtlas.org>

3.2.9 DLR's World Settlement Footprint dataset

In order to generate an urban mask in the CGLS-LC100 workflow, DLR's World 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 8 shows an example of the WSF 2015 layer for the Street of Gibraltar. More details regarding additional processing can be found in section 3.7.2.

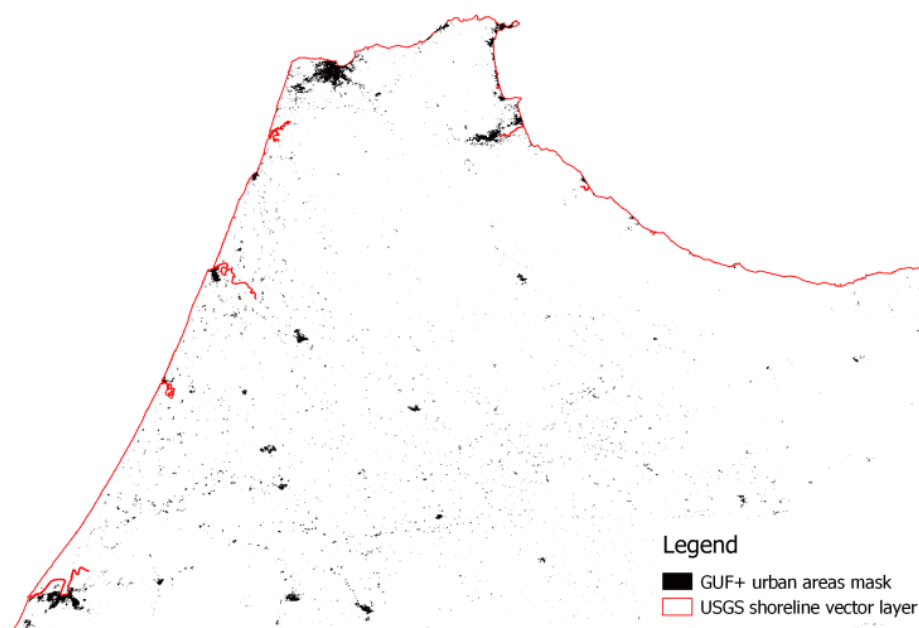


Figure 8: The DLR World Settlement Footprint layer (Marconcini et al., 2017b). Zoom in to Street of Gibraltar showing the urban areas in black. Note: in red the shoreline vector layer is shown.

3.2.10 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 maximum water extent and water seasonality 2015 layers. The maximum water extent provides information on all the locations ever detected as water over the Landsat data archive period (32 year period) and 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 9 shows an example of the GSW maximum water extent layer and the water seasonality 2015 layer for the Street of Gibraltar. More details regarding additional processing can be found in section 3.7.3.

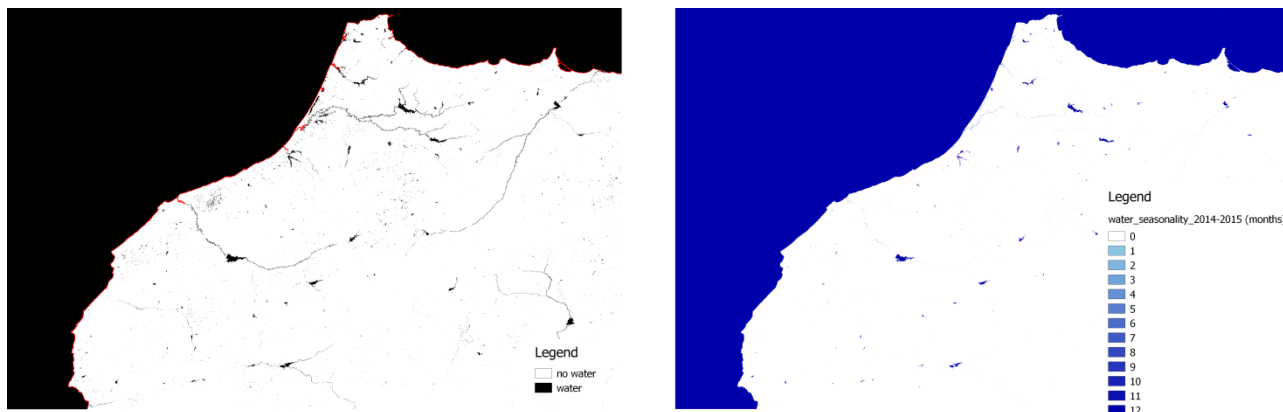


Figure 9: The JRC Global Surface Water product (Pekel et al., 2016). Zoom in to the Street of Gibraltar showing the maximum water extent layer (left) and the water seasonality mask for 2015 (right). Note: in red the USGS shoreline vector layer is shown in the maximum water extent layer (left image).

3.2.11 Digital Elevation Model datasets

To include topographic metrics (height, slope, aspect) in the CGLS-LC100 product workflow, the National Aeronautics and Space Administration (NASA) Shuttle Radar Topography Mission (SRTM) plus digital elevation model (DEM) data in 1 arc second resolution (SRTMGL1) is used as primary input (NASA, 2013). The SRTMGL1 dataset is provided on a near-global scale from 56° S to 60° N, and with a ~30 m spatial resolution. Figure 10 shows an example of the SRTMGL1 dataset for the Street of Gibraltar.

Since the high quality SRTMGL1 is not covering High Latitudes, the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) global DEM V2 was used (NASA/METI/AIST/Japan Spacesystems, and U.S./Japan ASTER Science Team, 2009) to fill missing areas.

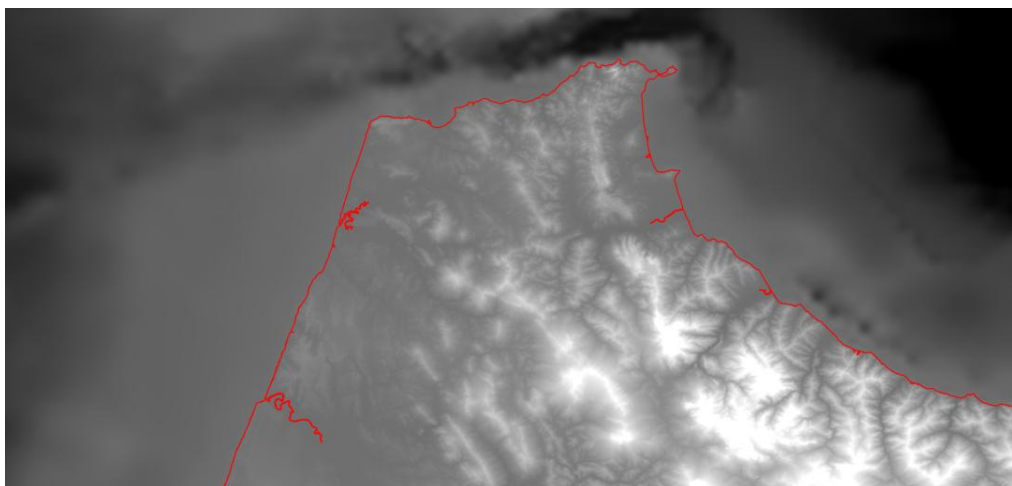


Figure 10: The NASA Shuttle Radar Topography Mission Global 1 arc second dataset. Zoom in to the Street of Gibraltar. Note: in red the USGS shoreline vector layer is shown.

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 11 shows the processing workflow developed to generate the ARD in a fully automatic mode in VITO's data centre.

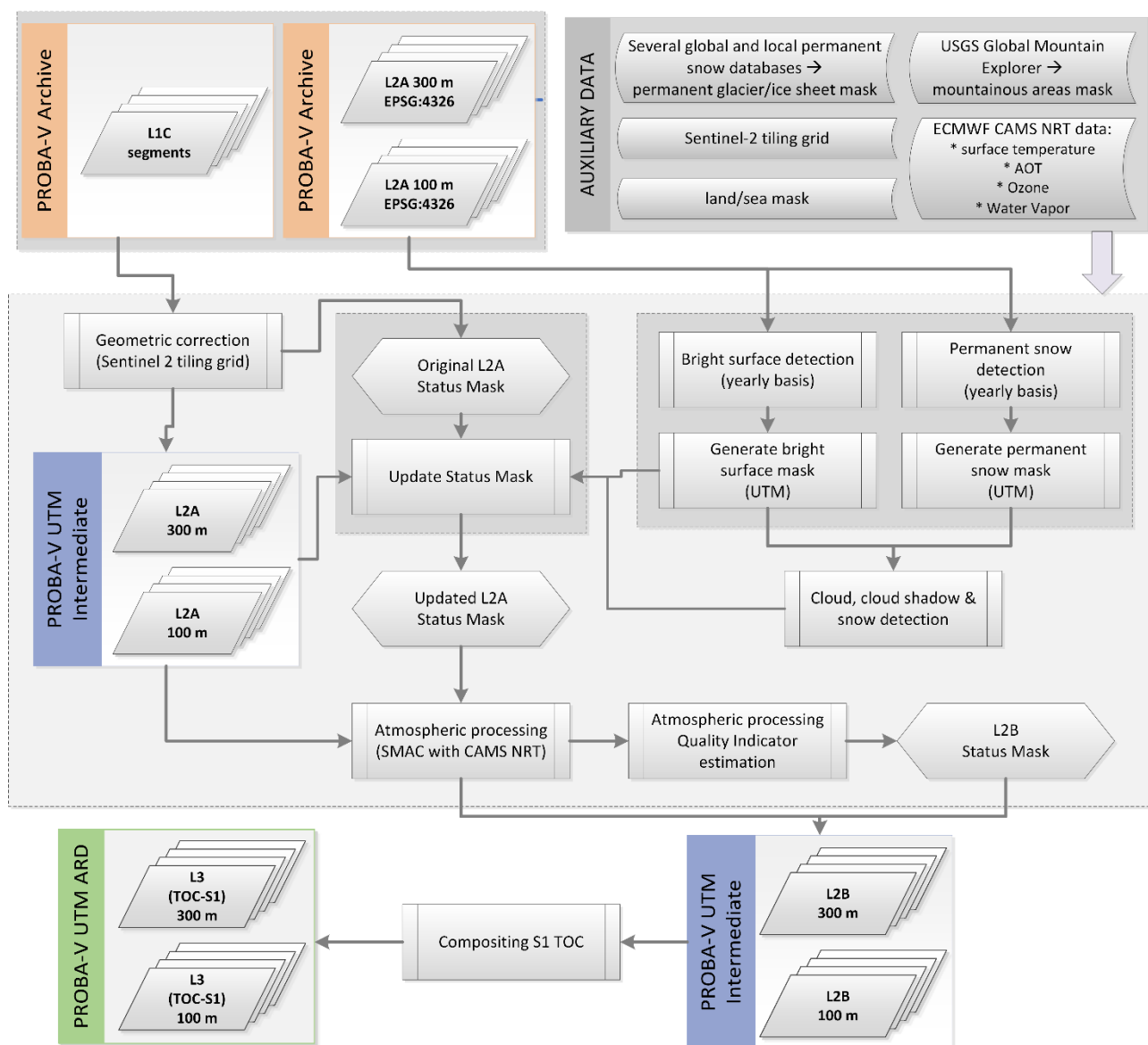


Figure 11: Workflow chart showing the generation of the PROBA-V UTM Analysis Ready Data (ARD) archive out of the PROBA-V L1C archive.

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-

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 11) 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° of longitude with an own designated central meridian. This allows a projection of the EO data with less over-/under-sampling 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 later also the continuation of the CGLS-LC100 mapping service when the operational phase of the PROBA-V mission has ended (estimated mid 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 planes, 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 11).

Afterwards, all single PROBA-V UTM L2B EO images were post-processed to generate 1-daily 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 reduce the false snow detection in bright areas (Figure 11). 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 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 12 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 13).

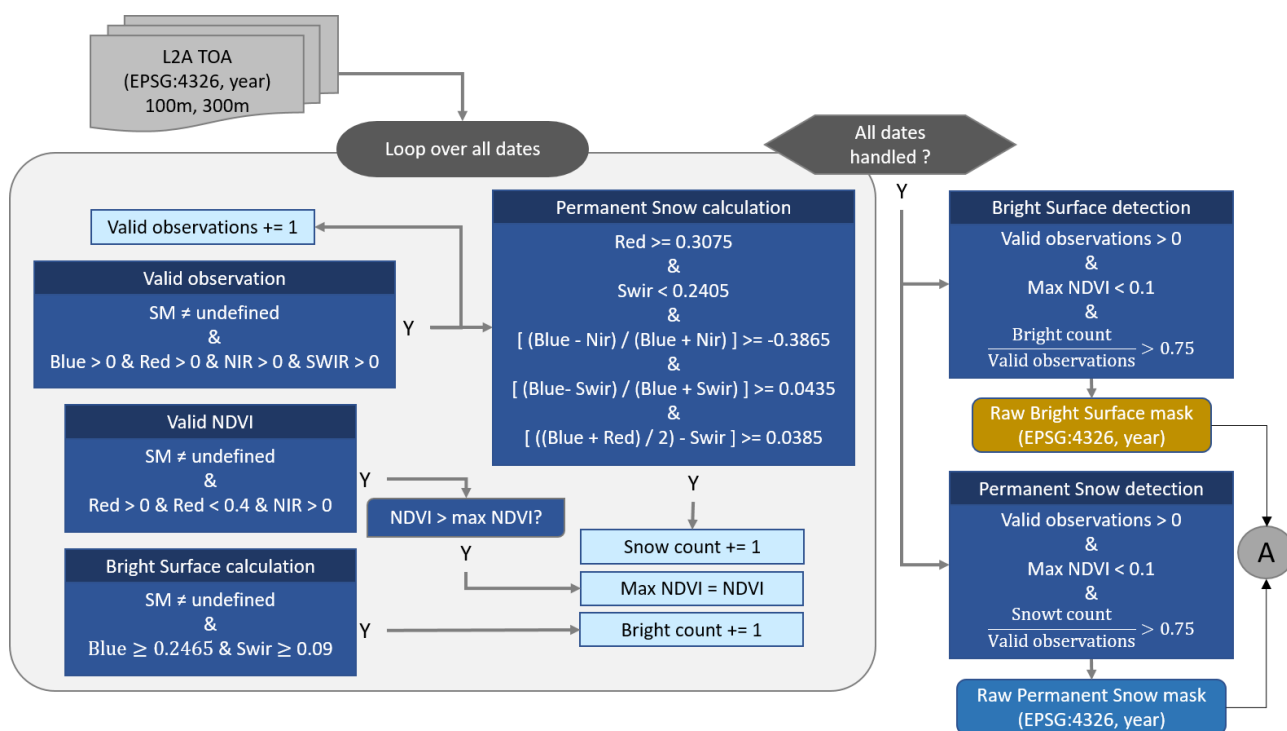


Figure 12: Yearly Bright Surface and Permanent Snow detection – Part A: Raw mask generation using the PROBA-V L2A archive.

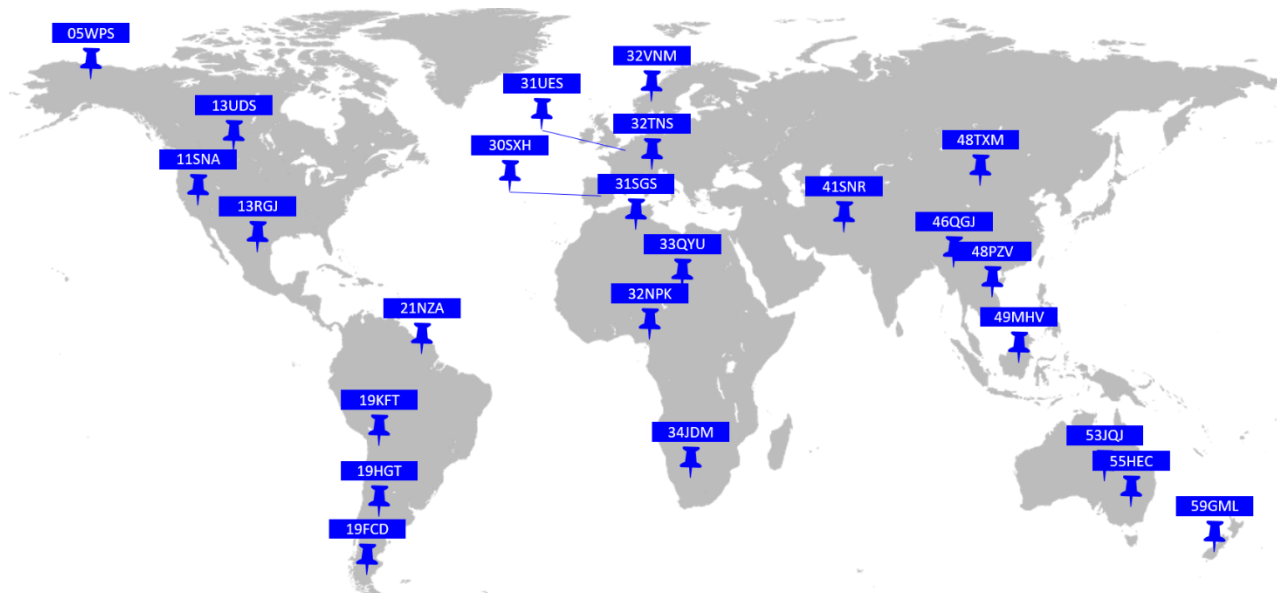


Figure 13: Distribution of global test sites used in the PROBA-V UTM ARD generation. Test sites following the Sentinel-2 naming convention.

Figure 14 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 3.2.5), 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 PROBA-V UTM ARD product (Figure 14). 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 15 shows some examples of the detected bright surfaces and permanent snow surfaces for the reference year 2015.

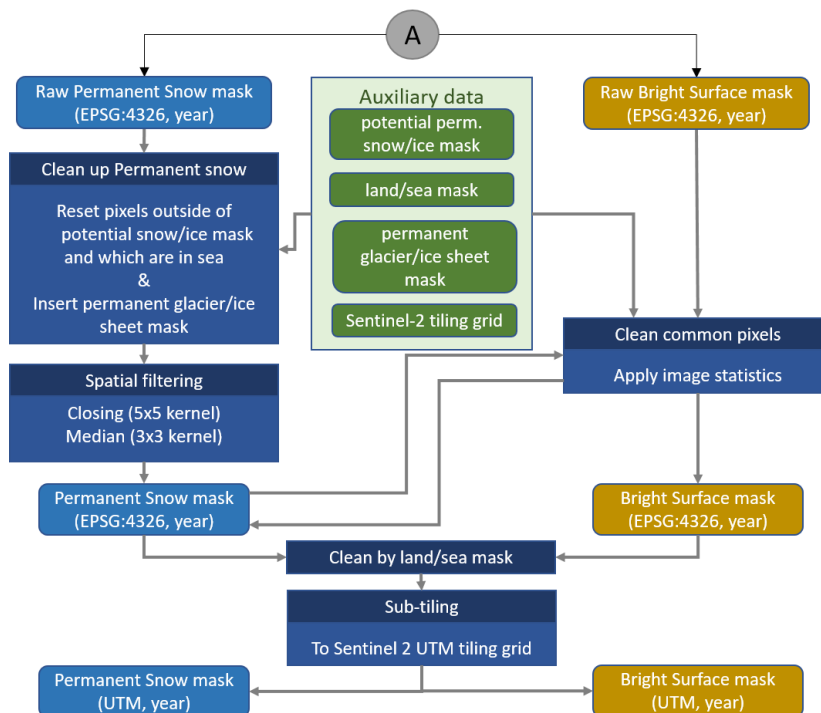


Figure 14: Yearly Bright Surface and Permanent Snow detection – Part B: Mask cleaning and final Sentinel-2 tile generation.



Figure 15: 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.

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
		001: shadow
		010: undefined
		011: cloud
		100: snow/ice
4	Land/sea mask	0: sea 1: land
5	SWIR quality flag	0: invalid data 1: valid data
6	NIR quality flag	0: invalid data 1: valid data
7	RED quality flag	0: invalid data 1: valid data
8	BLUE quality flag	0: invalid data 1: valid data
9	SWIR coverage	0: no 1: yes
10	NIR coverage	0: no 1: yes
11	RED coverage	0: no 1: yes
12	BLUE coverage	0: no 1: yes

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 11). 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 out of 18,595 tiles with a dimension of 110 x 100 km and covering all global land masses (Figure 16). Note: the tiling grid is still called “Sentinel-2 tiling grid”.

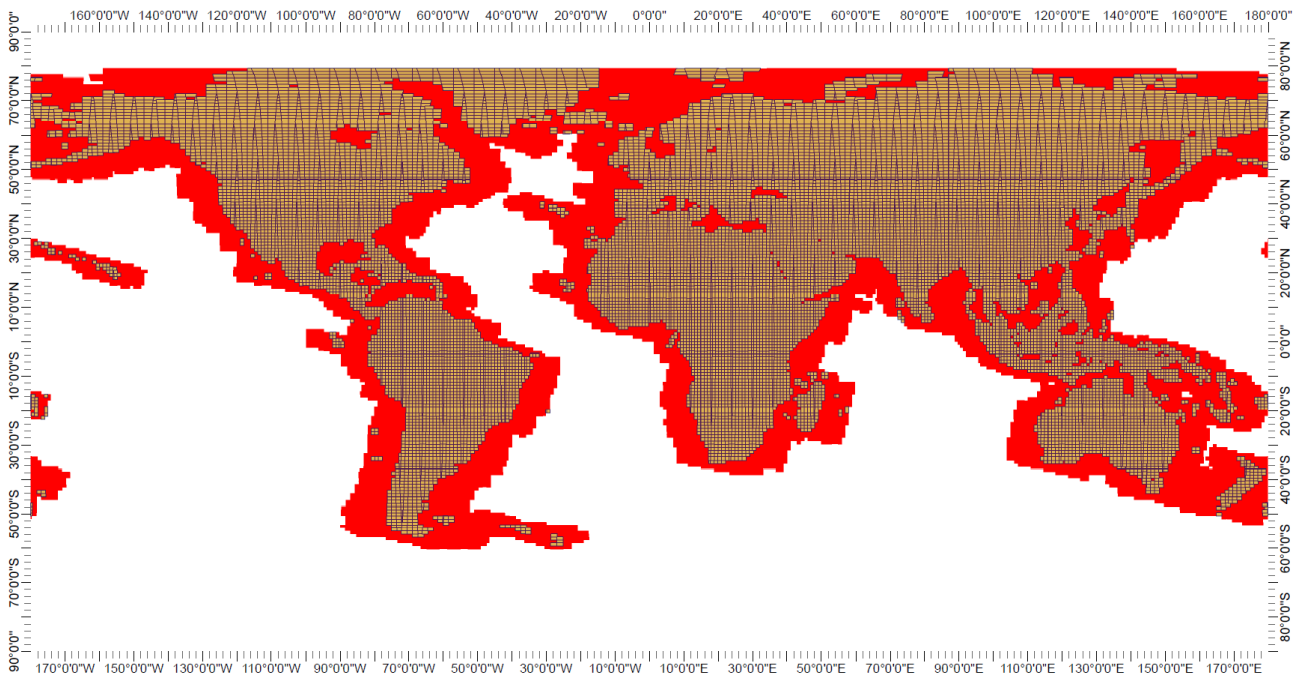


Figure 16: Sentinel-2 tiling grid overlaid on PROBA-V imaging area. Shown are the 18,595 tiles covering the land masses which form the PROBA-V UTM ARD archive.

The remaining PROBA-V L2A processor stayed untouched and provided for PROBA-V UTM images next to the status mask files also the needed Sun-Sensor-Object geometries as well as the overpass time files, which are input into the L2B processor. 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 11).

The PROBA-V UTM L2A post-processing includes an improved snow/ice detection as well as cloud detection module (Figure 17 and Figure 18), which results in an update of the PROBA-V UTM status mask and therefore replacement of the old one (Figure 19). 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 17). 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 3.2.6),

the land/sea 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 17). 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°C or a daily minimum temperature above -2°C in combination with a daily maximum temperature above 4 °C cannot contain snow.

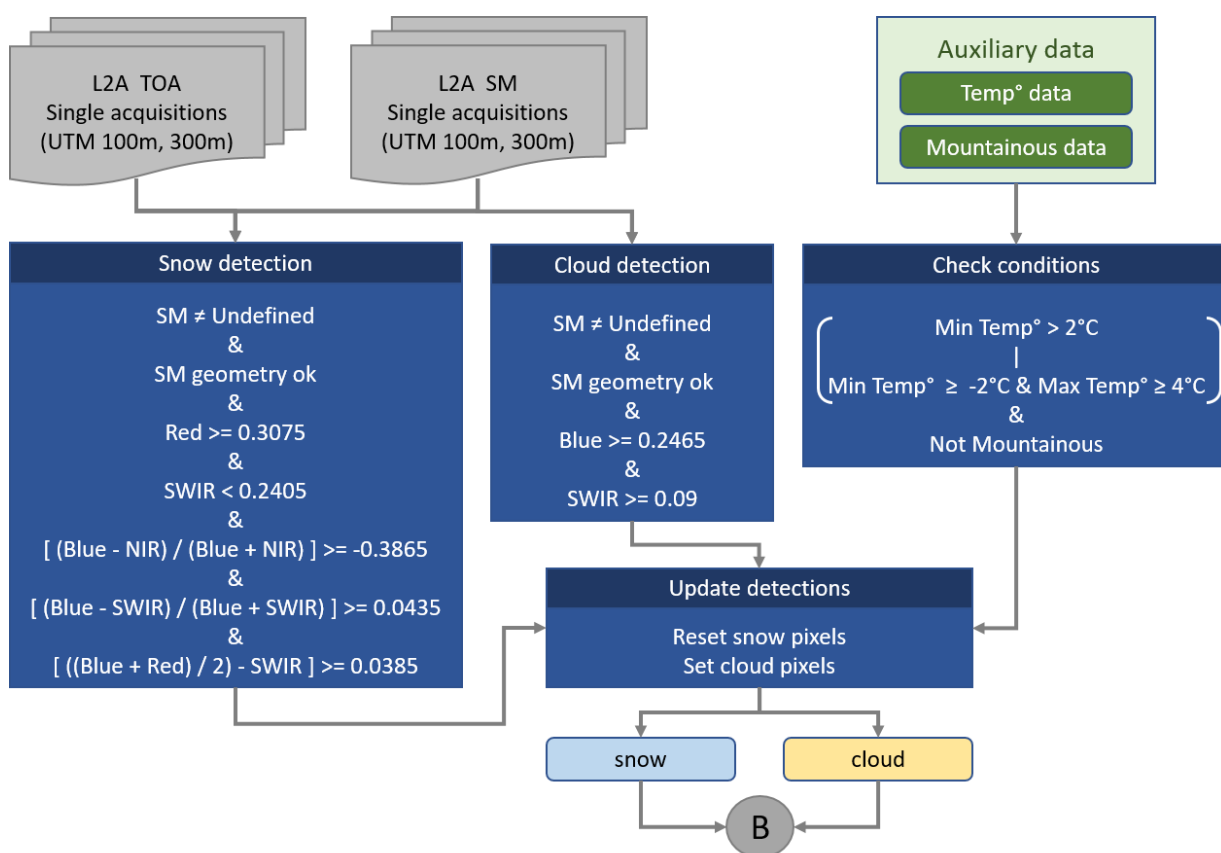


Figure 17: PROBA-V UTM L2A post-processing – Part A: improved snow and cloud detection – raw detection by thresholding.

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 18).

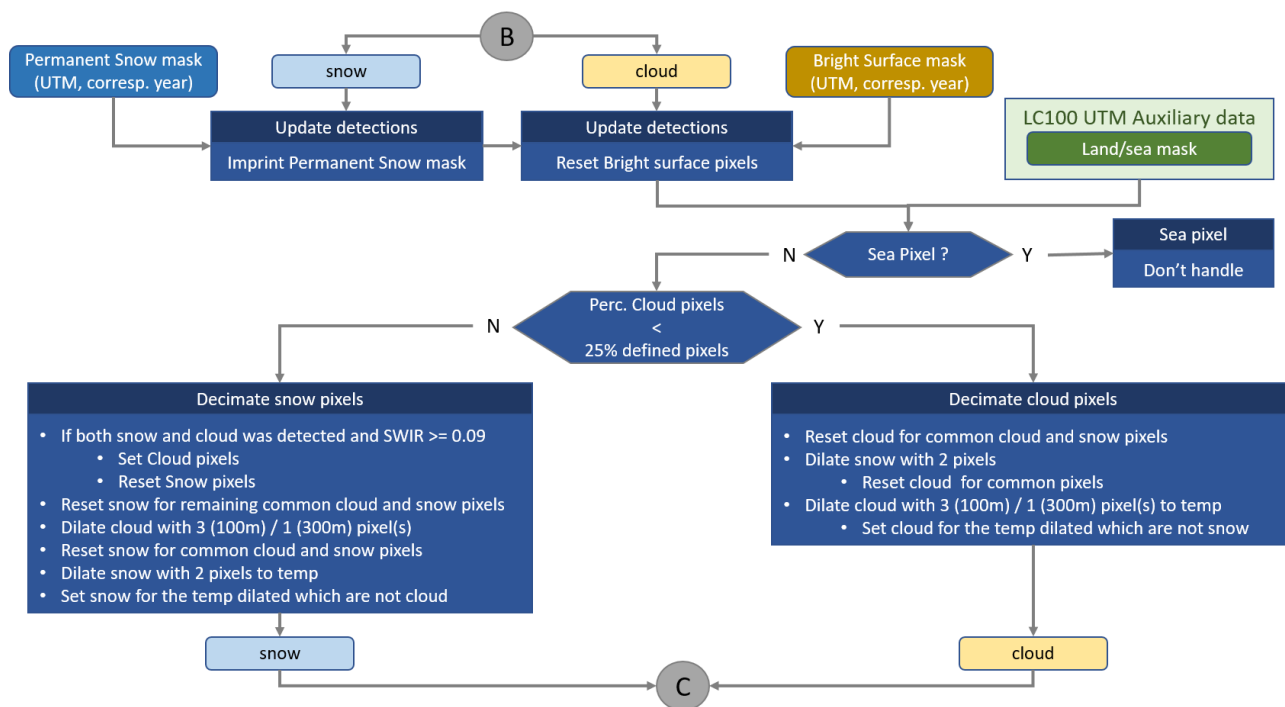


Figure 18: PROBA-V UTM L2A post-processing – Part B: improved snow and cloud detection – cleaning by masks and resolving confusions.

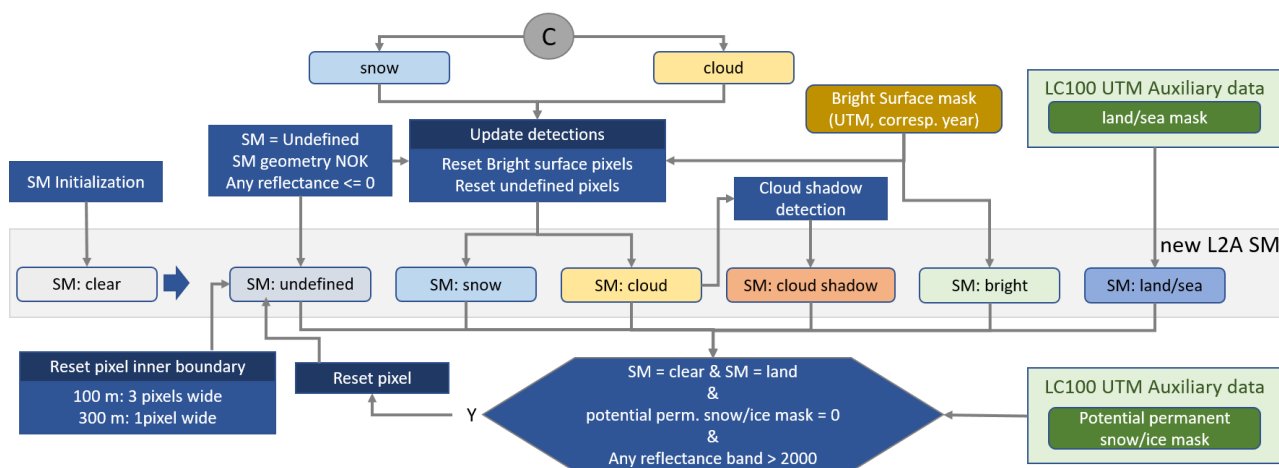


Figure 19: PROBA-V UTM L2A post-processing – Part C: updating of status mask.

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 19). 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

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 11, 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 20). The atmospheric correction is carried out pixel-based 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.

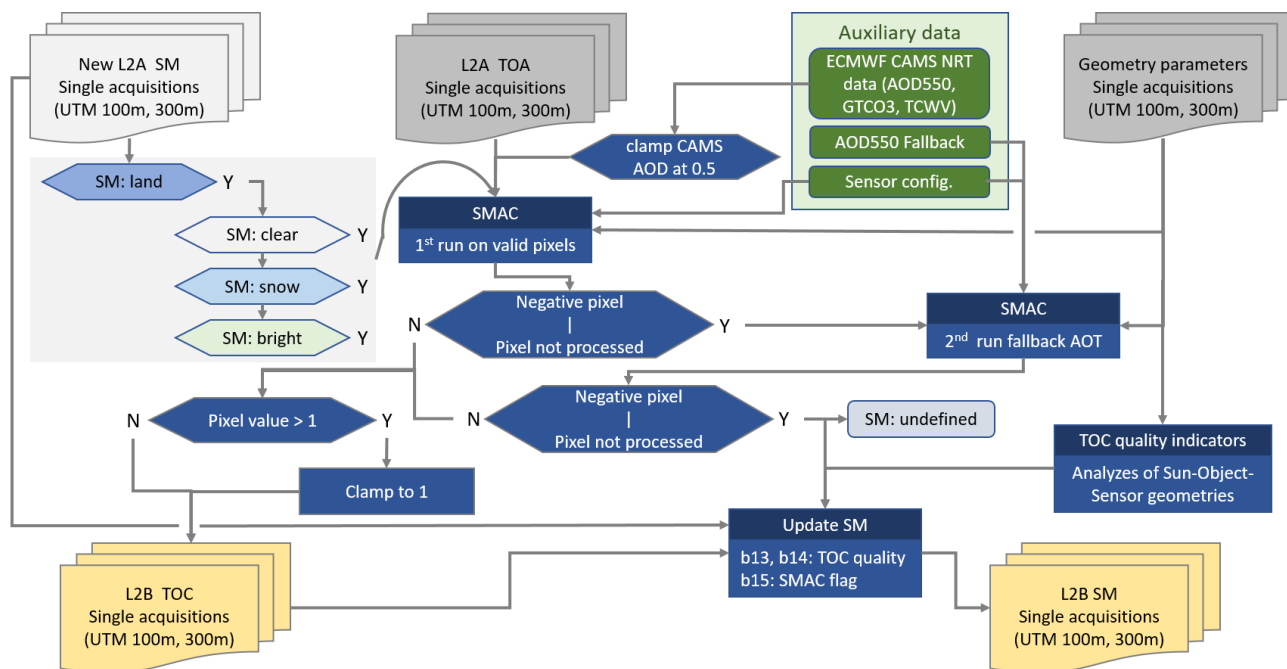


Figure 20: PROBA-V UTM L2B processor – improved SMAC routine and quality indicator generation for atmospheric correction.

Main input into the PROBA-V UTM L2B processor (Figure 20) 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);
- 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 21 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 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 20). 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.

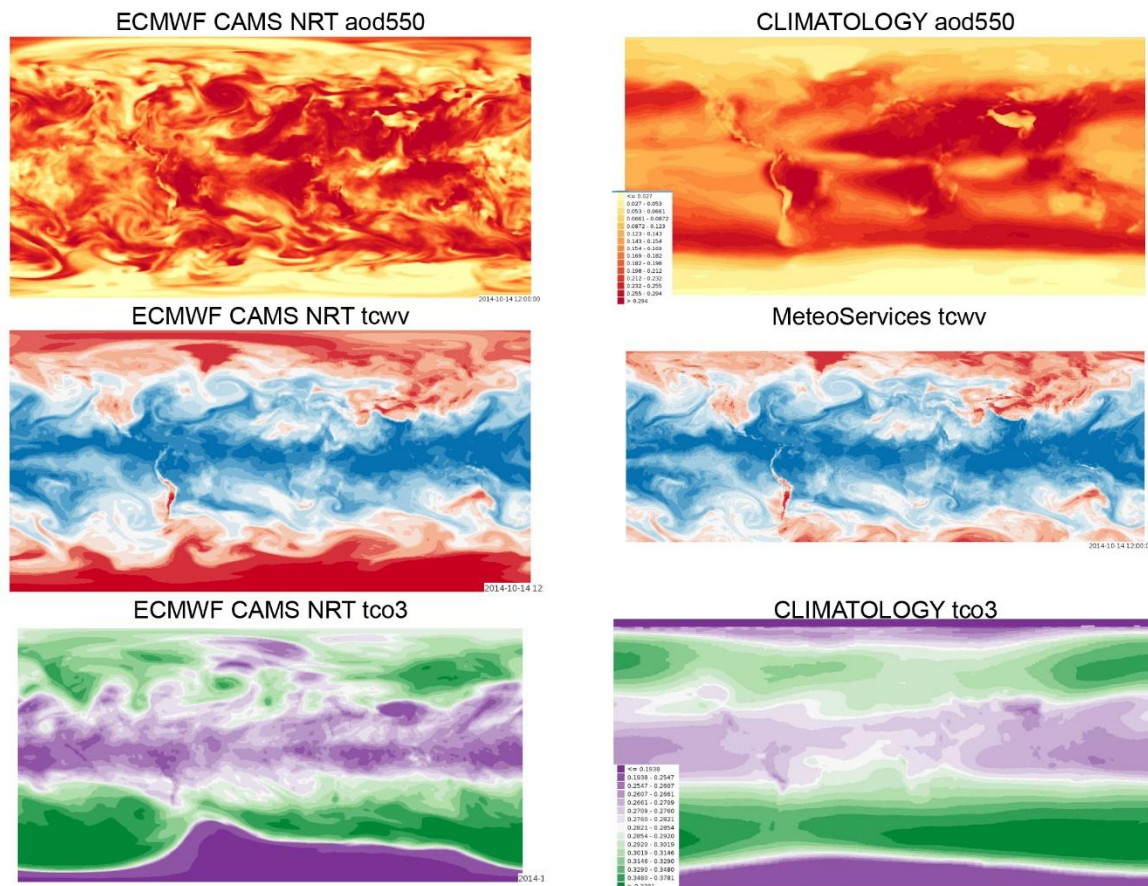


Figure 21: Comparison of used ancillary data used in the improved and original PROBA-V L2B processor. Left: ECMWF CAMS NRT data with 6-hour temporal resolution; right: long-term monthly data used in original L2B processor. Note: visualizations are shown for October 15, 2014.

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: $(\text{SZA} \leq 45^\circ) \text{ AND } (\text{VZA} \leq 45^\circ) \text{ AND } (\text{SZA} + \text{VZA} \leq 70^\circ)$;
- Acceptable: $((\text{SZA} > 45^\circ) \text{ AND } (\text{SZA} \leq 60^\circ)) \text{ OR } ((\text{VZA} > 45^\circ) \text{ AND } (\text{VZA} \leq 60^\circ)) \text{ OR } ((\text{SZA} + \text{VZA} > 70^\circ) \text{ AND } (\text{SZA} + \text{VZA} \leq 100^\circ))$;

- Unreliable: $((\text{SZA} > 60^\circ) \text{ AND } (\text{SZA} \leq 75^\circ)) \text{ OR } ((\text{VZA} > 60^\circ) \text{ AND } (\text{VZA} \leq 75^\circ)) \text{ OR } ((\text{SZA} + \text{VZA} > 100^\circ) \text{ AND } (\text{SZA} + \text{VZA} \leq 140^\circ))$;
- Bad: $(\text{SZA} > 75^\circ) \text{ OR } (\text{VZA} > 75^\circ) \text{ OR } (\text{SZA} + \text{VZA} > 140^\circ)$.

An extra rule applies on pixel already classified as “good”. In case the pixel values for gtco3 is higher than 0.35 atm-cm or the pixel value for tcwv is higher than 2.25 g/cm², 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 20 and Table 5).

3.3.5 Level 3 processing and final status mask

The last step in the PROBA-V UTM ARD generation (Figure 11) is the compositing – called L3 processing - of the single observations of all sensor overpasses per day into daily synthesis at 100 m and 300 m, respectively. The daily 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 22 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 daily 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 daily compositing. Due to the fact that in the original Sentinel-2 tiling grid a tile overlap of 10km to each side is defined, this processing steps 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 – now.

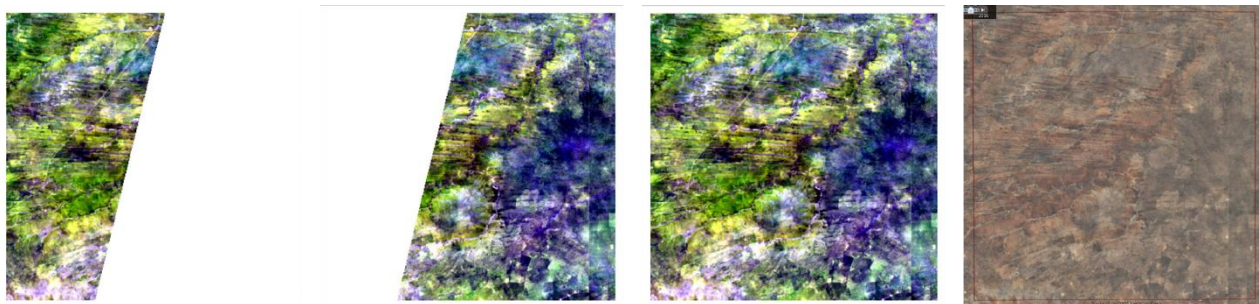


Figure 22: PROBA-V UTM L3 processing – generation of daily synthesis. From left to right: partly covered tile in first and second overpass, daily synthesis, and Google Earth image for comparison.

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

Bit	Name	Description
1 -3	Observation	000: clear
		001: shadow
		010: undefined
		011: cloud
		100: snow/ice
		111: bright surface
4	Land/sea mask	0: sea 1: land
5	SWIR quality flag	0: invalid data 1: valid data
6	NIR quality flag	0: invalid data 1: valid data
7	RED quality flag	0: invalid data 1: valid data
8	BLUE quality flag	0: invalid data 1: valid data
9	SWIR coverage	0: no 1: yes
10	NIR coverage	0: no 1: yes
11	RED coverage	0: no 1: yes
12	BLUE coverage	0: no 1: yes

13 - 14	TOC quality	00: bad 01: unreliable 10: acceptable 11: good
15	SMAC flag	0: not carried out 1: successful
16	ARD flag	0: not usable 1: all steps successful

3.3.6 Comparison PROBA-V versus PROBA-V UTM

Figure 23 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.

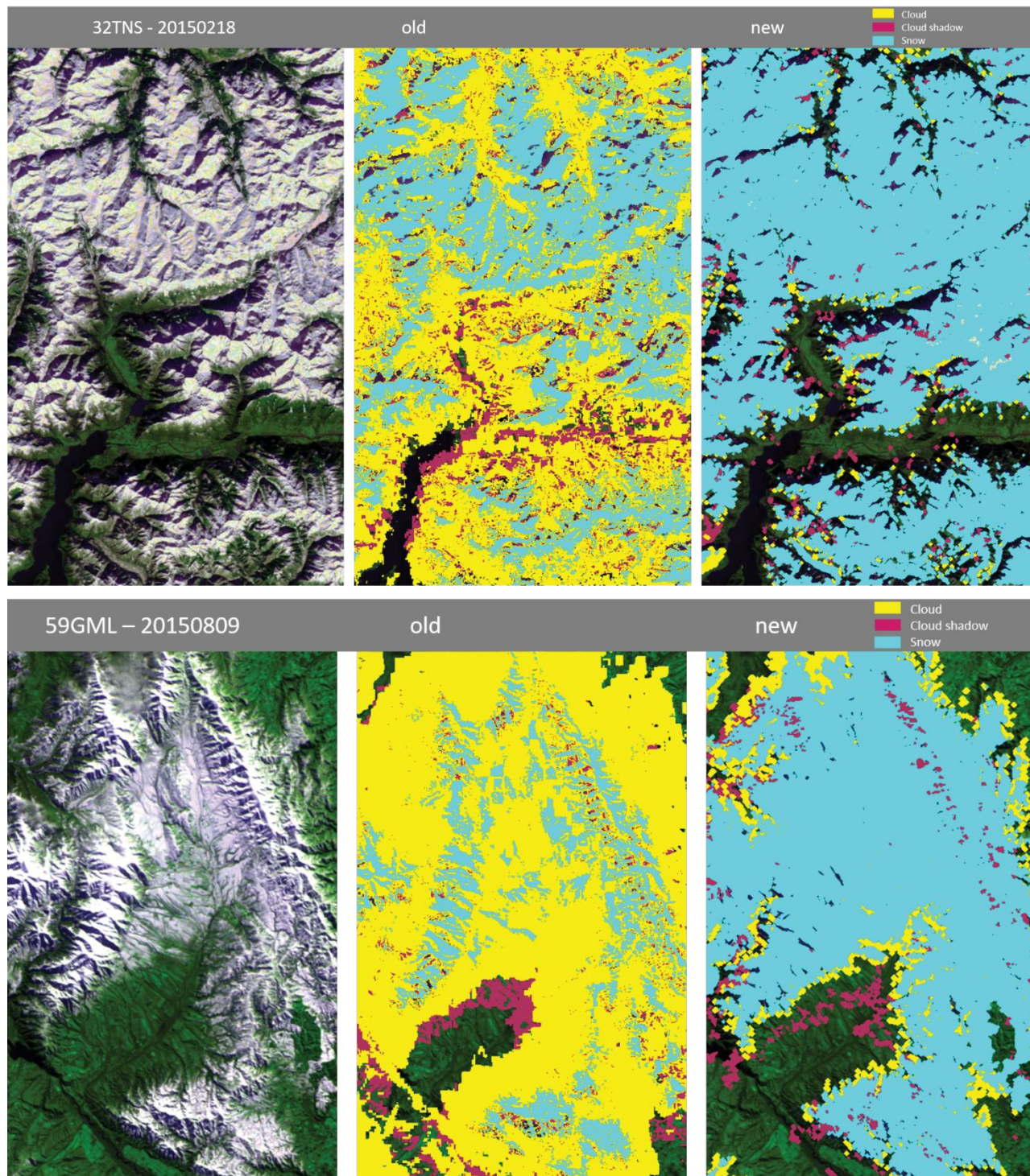


Figure 23: Comparison between PROBA-V and PROBA-V UTM cloud/snow detection.

Figure 24 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.

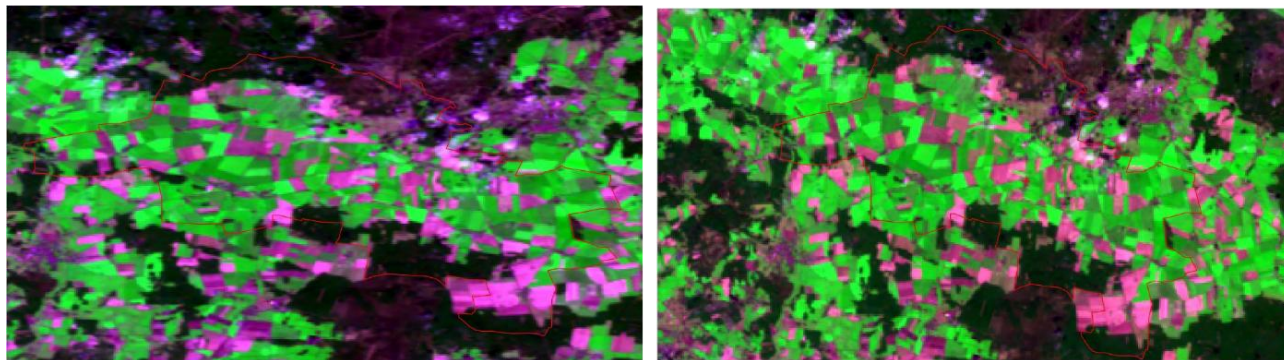


Figure 24: Comparison between PROBA-V (left) and PROBA-V UTM (right) geometric correction (L2A). Already in mid-latitudes a reduced distortion is viewable. Area shown: province Niedergoersdorf, Teltow-Flaeming, Germany.

3.4 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 1) is the data cleaning and compositing (Figure 25). Input data are the multi-spectral PROBA-V UTM TOC-S1 datasets as described in section 3.3. 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 daily synthesis products as well as to generate regular time steps, a median compositing algorithm is used to generate 5-daily and 10-daily median composites (MC5 and MC10, respectively). For the 300 m PROBA-V UTM ARD data also a Harmonized 5-daily Median Composite (HMC5) for the whole PROBA-V data archive (2013 – now) 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 1). 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 25 are only produced temporary as input for the next processing steps.

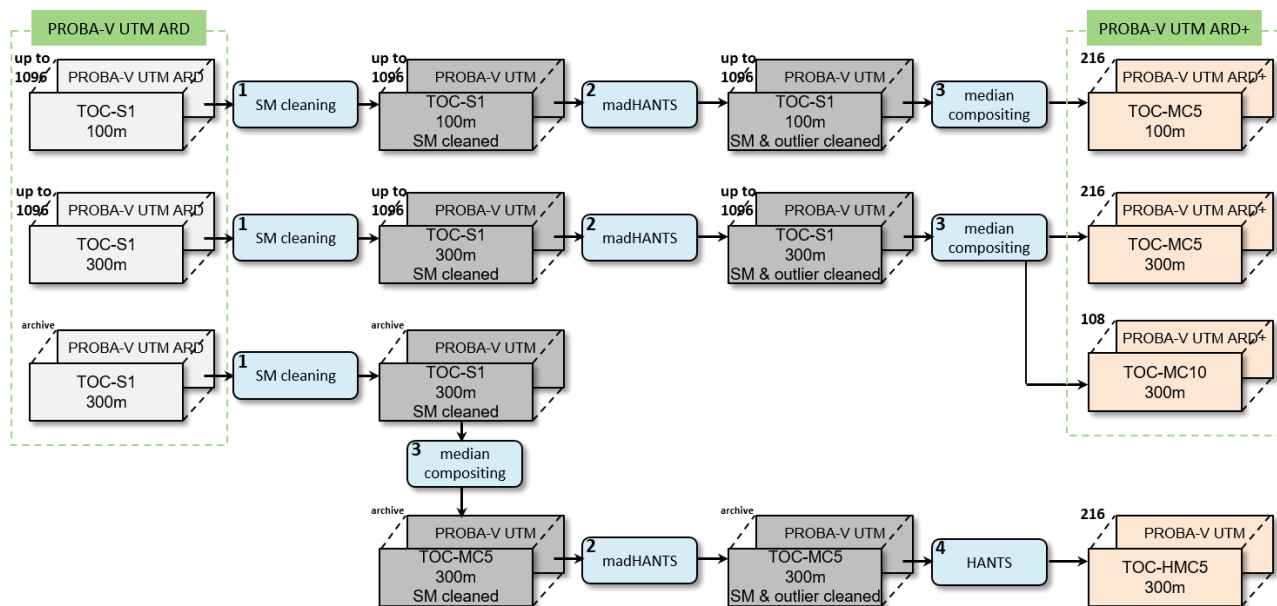


Figure 25: General overview of the data cleaning and compositing section in the CGLS-LC100 product workflow. Note: Numbers in the upper left corner of data container indicate the number of layers - here number of observations.

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 times 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 the corresponding harmonized pixel reflectance values in the numerator allows a better adaptation of the outlier detection algorithm to seasonality.

$$s = \frac{|x_i - x_{iHANTS}|}{\text{median}(|x_1 - x_{1HANTS}|, \dots, |x_n - x_{nHANTS}|)} \quad \text{Eq. 1}$$

, 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_{iHANTS} is the harmonized pixel reflectance value of the current time step using the HANTS algorithm.

A threshold of 3.5 standard deviations is used 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 26 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° lon, 6.562° lat.

Figure 26 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).

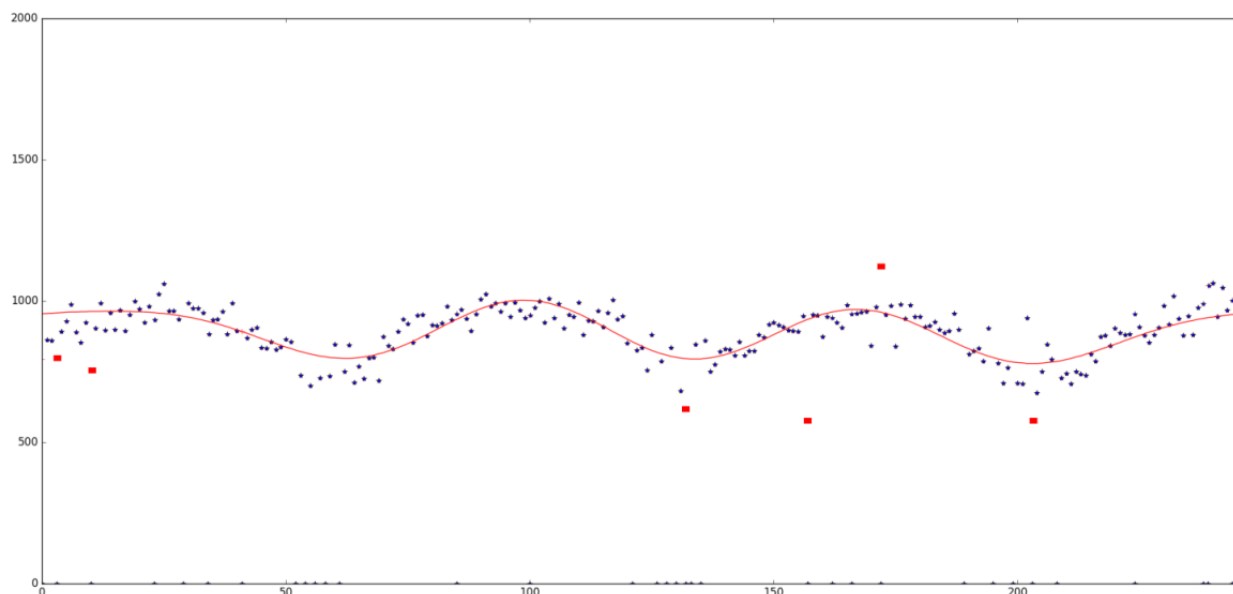


Figure 26: Example for the madHANTS temporal outlier detection algorithm. Note: blue stars mark all original pixel values in the time series (invalid observation values are set to -1), the red line shows the harmonized time series by applying the amplitudes and phases of the identified frequencies during the HANTS transformation, red squares show the detected outliers which are 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 25) can be split in two steps. The first is the generation of status mask cleaned and temporal outlier cleaned 5-daily and ten-daily 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 daily TOC products.

Overall three processing sub-steps are needed in order to generate the 5-daily (MC5) and 10-daily (MC10) composites from the PROBA-V UTM ARD archive (Figure 25). First, all multi-spectral 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. This can be 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-daily at 100 m), but also on the latitude (the higher the latitude the more overlapping orbits exists, but also the less 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 27 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 28 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.

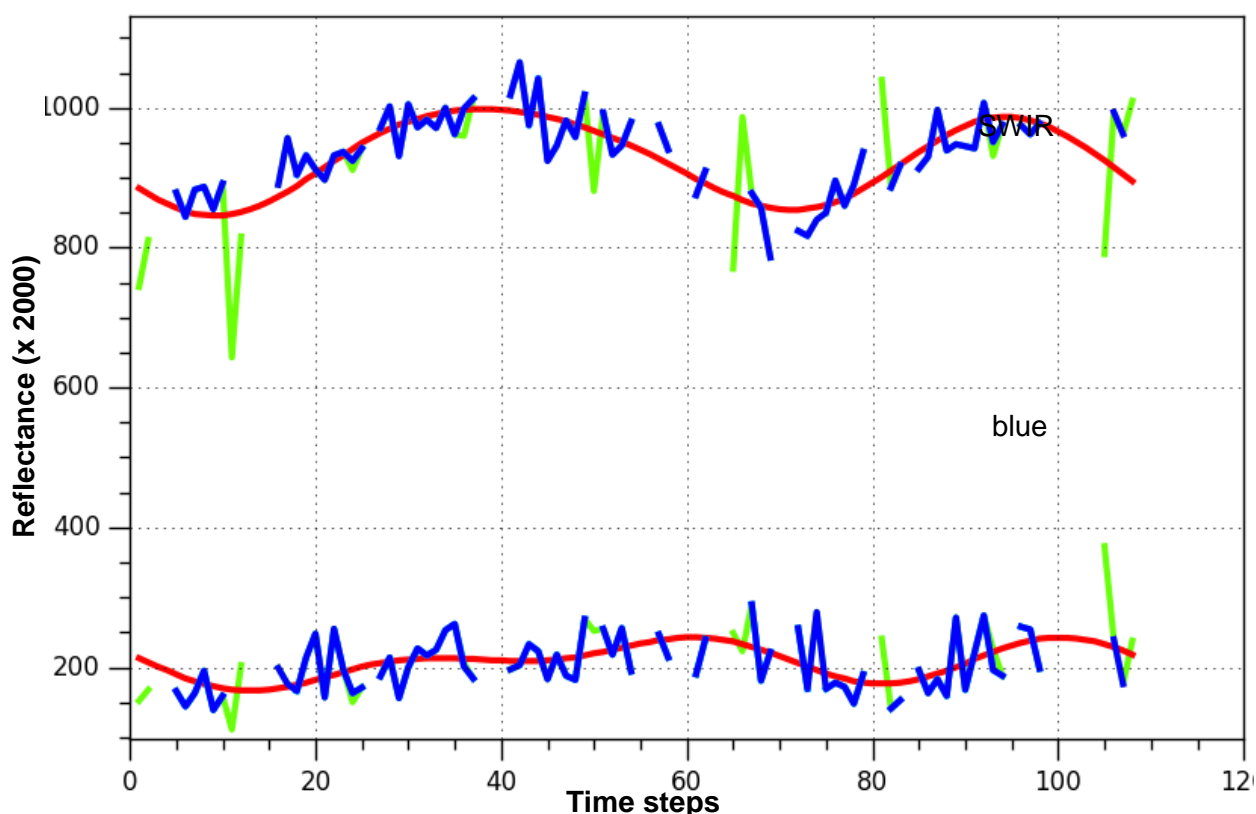


Figure 27: Example showing the temporal outlier detection using the madHANTS algorithm on the blue and SWIR reflectance values of a pixel at location 9.459° lon, 6.562° lat. Note: red curve shows the harmonized time series; blue shows the valid pixel reflectance values; green shows flagged outliers cumulative detected in the blue or SWIR reflectance band.

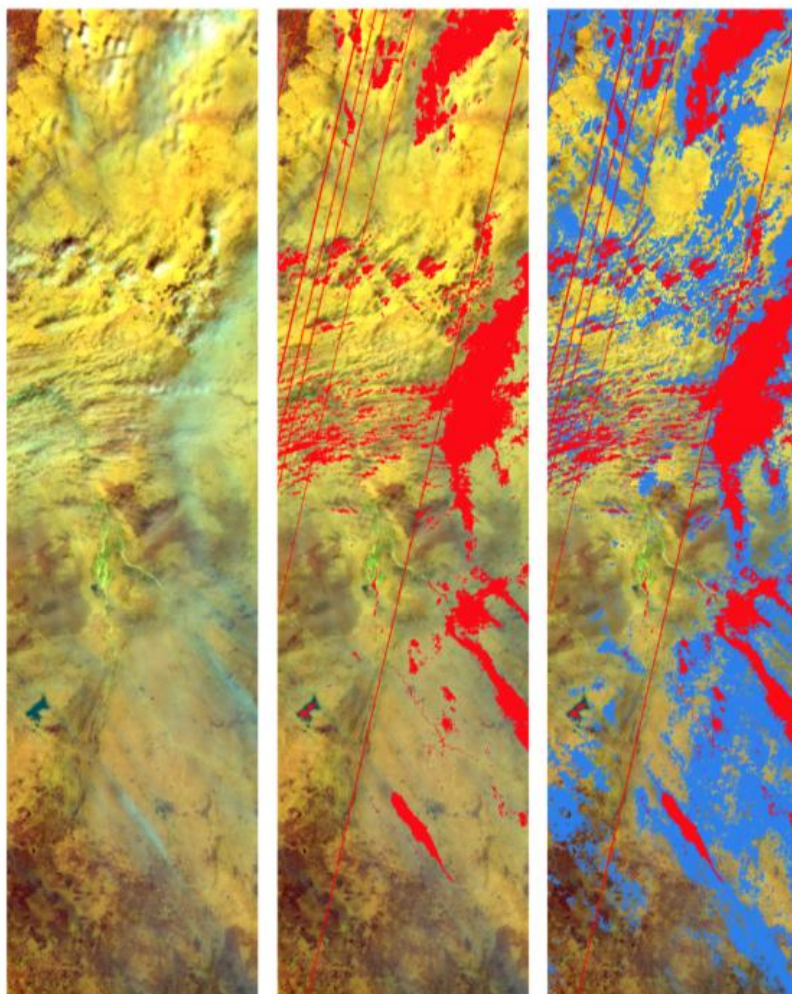


Figure 28: Example for status mask and temporal data cleaning for part of a PROBA-V UTM TOC-S1 100 m tile (observation taken 2016-03-06). Image is shown as false colour composite (RGB = SWIR, NIR, blue) for a sample area in tile in Nigeria: raw image (left), status masked cleaned image (status mask flagged 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-daily Median Composite (HMC5) generation

The second step in the general workflow for the data cleaning and compositing section (outlined in Figure 25) is the generation of the long-term 5-daily 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 – now), 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 generated first in the HMC5 workflow (Figure 25). This 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-daily compositing compresses the data amount by factor 5. Again, the MC5 composites for the four spectral bands are obtained by calculating the median of the PROBA-V 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. Final output is the following dataset:

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

3.5 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 30). 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. 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 1). 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 29 shows two representative temporal profiles for a pixel time series with a high DDI and low DDI, respectively.

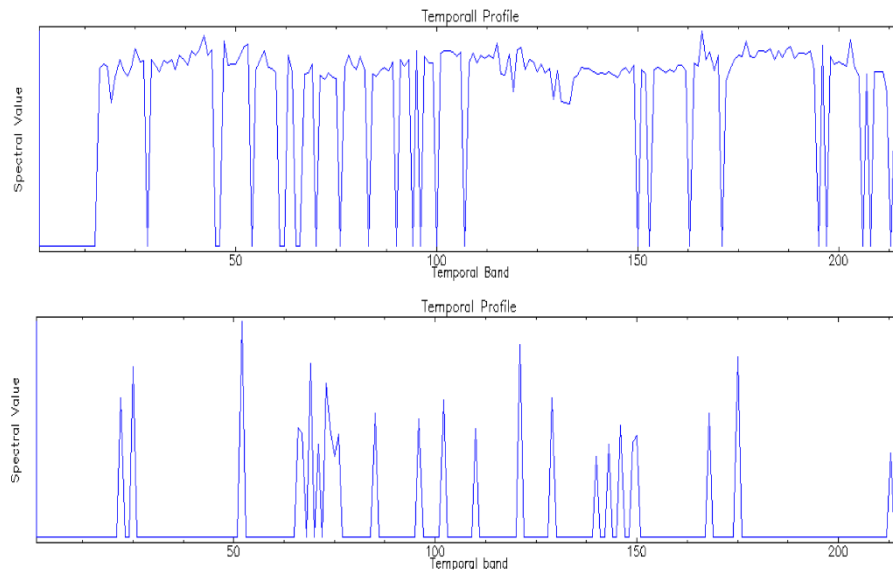


Figure 29: Exemplary temporal profiles for PROBA-V UTM ARD+ time series with a high Data Density Indicator (DDI) (upper graph) and low DDI (lower graph).

The overall processing workflow, shown in Figure 30, 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.

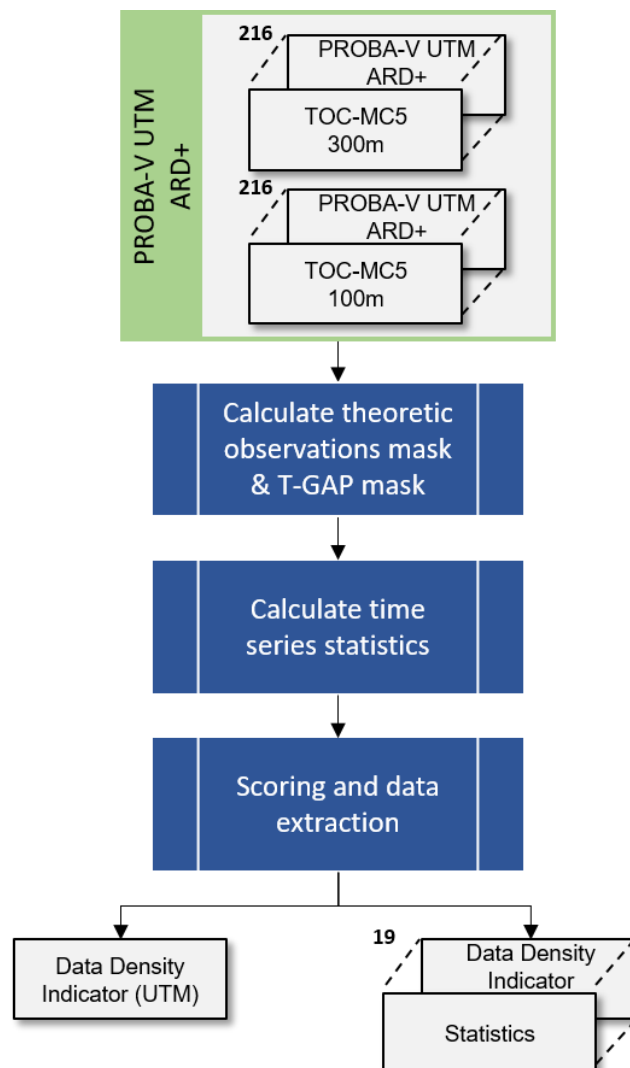


Figure 30: Data Density Indicator processing workflow.

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 31), and the generation of the t-GAP mask from the valid/invalid observations in the PROBA-V UTM ARD+ time series (Figure 31). Therefore 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.

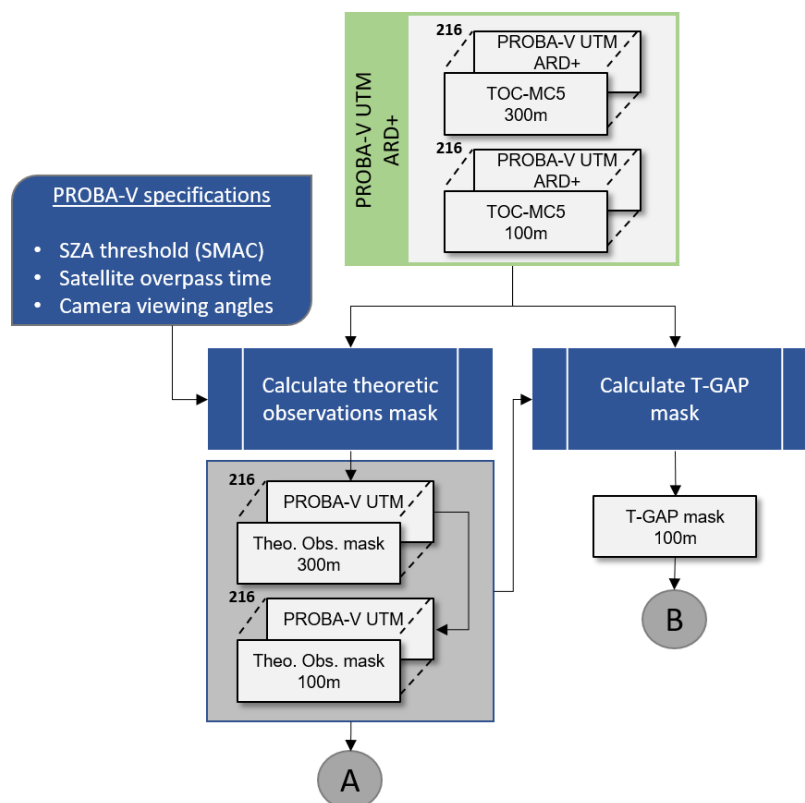


Figure 31: Calculating the theoretical maximum number of observations (Theo. Obs. mask) and the longest concurrent gap in data time series data (T-GAP) is the first step in calculating the DDI characteristics. Calculating the Theo. Obs. Mask requires three parameters: a threshold on the SZA as used by 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 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 uncertainty in the filled-in data. 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 experimental 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 PROBA-V 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 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.

Following 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 N° of observations** (Figure 32, equation 2). Calculated on the “raw” time series. This is the total number of five daily observations (valid or invalid) over the epoch (reference year +/- 1 year) and currently fixed to 216 observations.
- **N° of valid observations** (Figure 32, equation 3). Calculated on the “raw” time series. This is the number of valid observations over the epoch.
- **Observation ratio** (Figure 32, equation 4). This is the ratio of the *N° of valid observations* to the *Total N° of observations*.

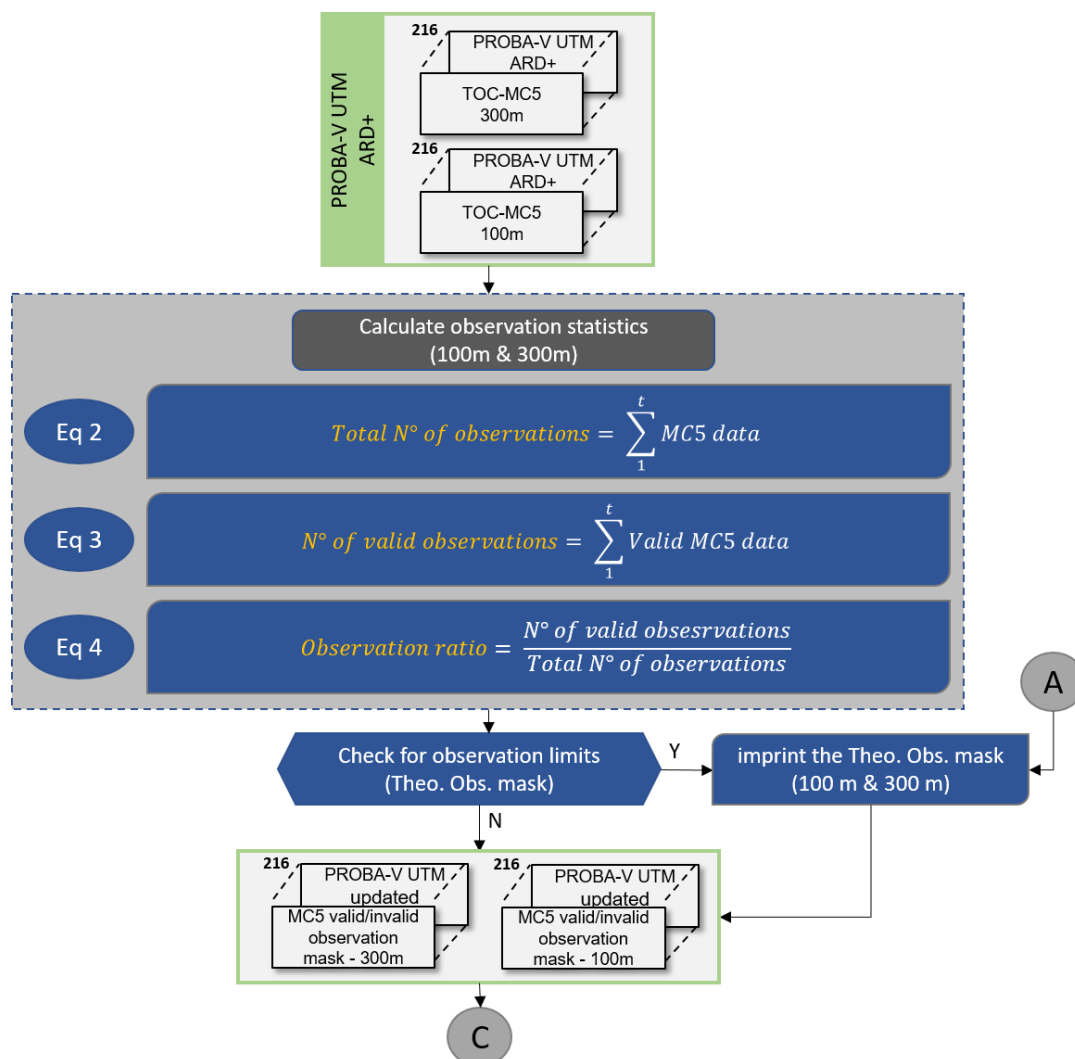


Figure 32: The raw time series statistics calculated on the PROBA-V UTM ARD+ data. After the raw time series statistics are calculated the PROBA-V UTM ARD+ data is checked for sensor specific observation limitations and updated in this case by imprinting the Theo. Obs. Mask.

- **N° of theoretical observations** (Figure 33, equation 5). Calculated on the “cleaned” time series. This is the total number of possible observations from the theoretical observation mask – indicating how many theoretical observation would be even possible given under the PROBA-V sensor specifications (maximum would equal the total N° of observations).
- **Theoretical observation vs maximum observation ratio** (Figure 33, equation 6). This is the ratio of the *N° of theoretical observations* to the *Total N° of observations*.
- **Theoretical observation ratio** (Figure 33, equation 7). This is the ratio of the *N° of valid observations* to the *N° of theoretical observations*.
- **Gap weight** (Figure 33, 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 33).

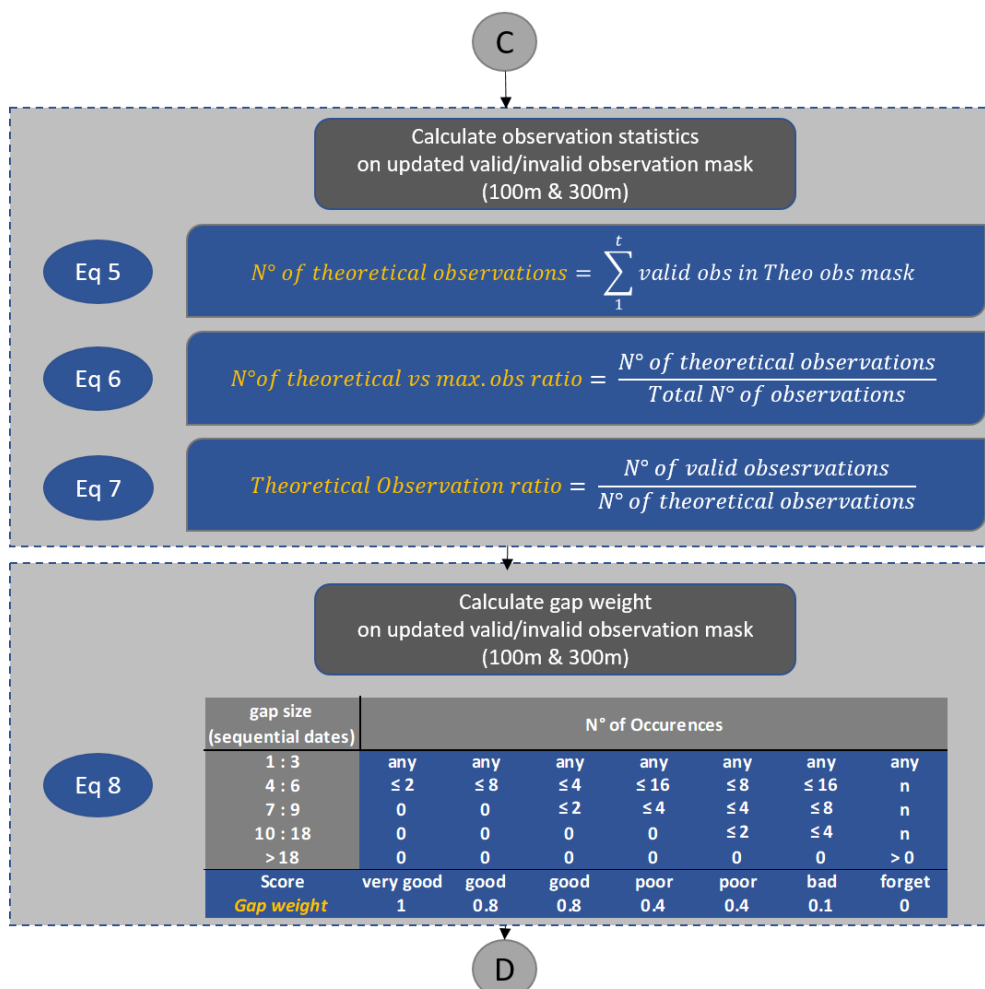


Figure 33: The cleaned time series statistics – Part A.

- **Observation frequency score** (Figure 34, 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 final observation frequency score is calculated depending on the relationship between valid observations and invalid observations.

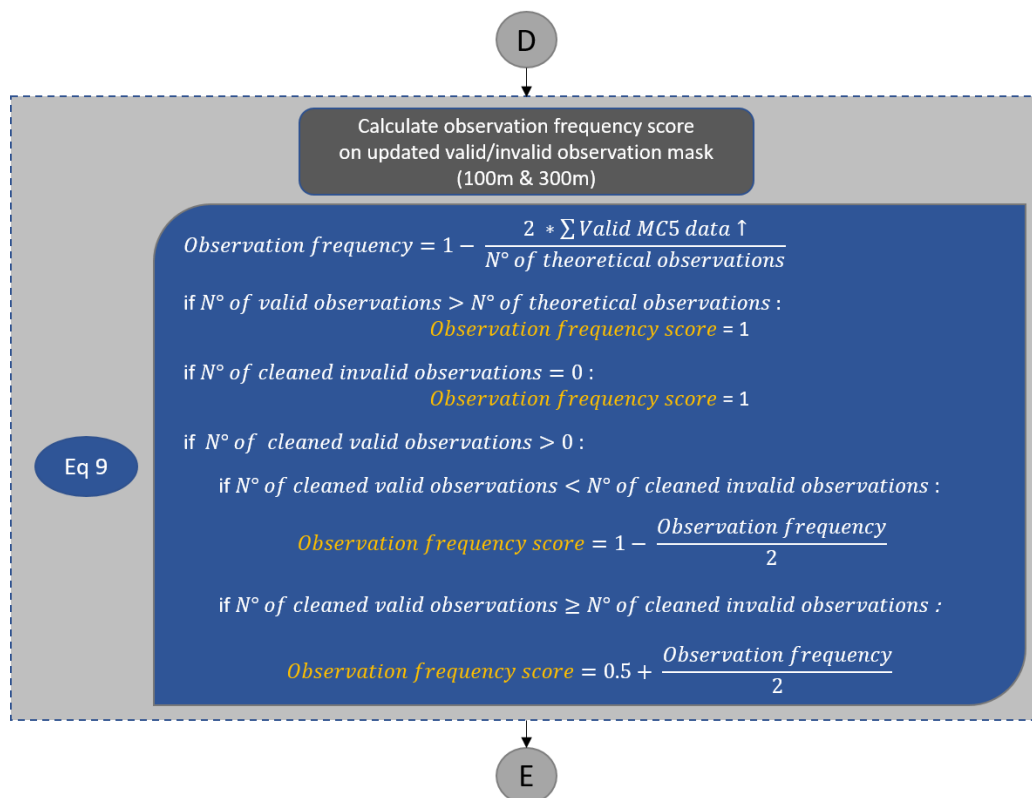


Figure 34: The cleaned time series statistics – Part B.

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 35). Overall 19 time series statistics are provided:

- N° of valid observations for 100 m and 300 m;
- Total N° 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;
- N° 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 35. Figure 36 shows besides a visual representation of the PROBA-V UTM ARD+ for tile 32NPM (over central Africa) also some of the statistical images calculated on the time series.

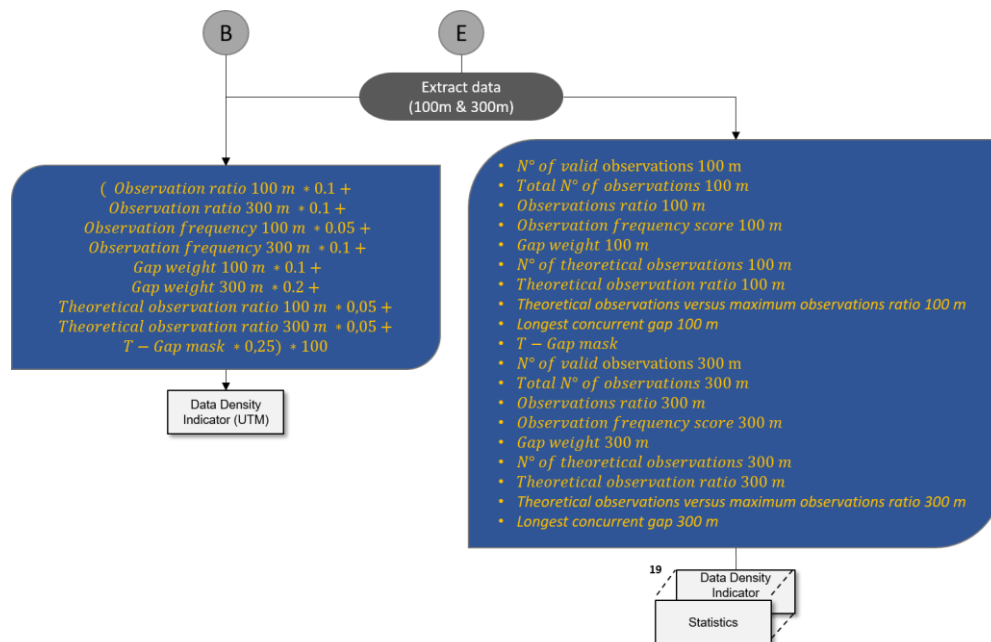


Figure 35: Final scoring of time series statistics to compute the Data Density Indicator (DDI) and output of the PROBA-V UTM ARD+ statistics.

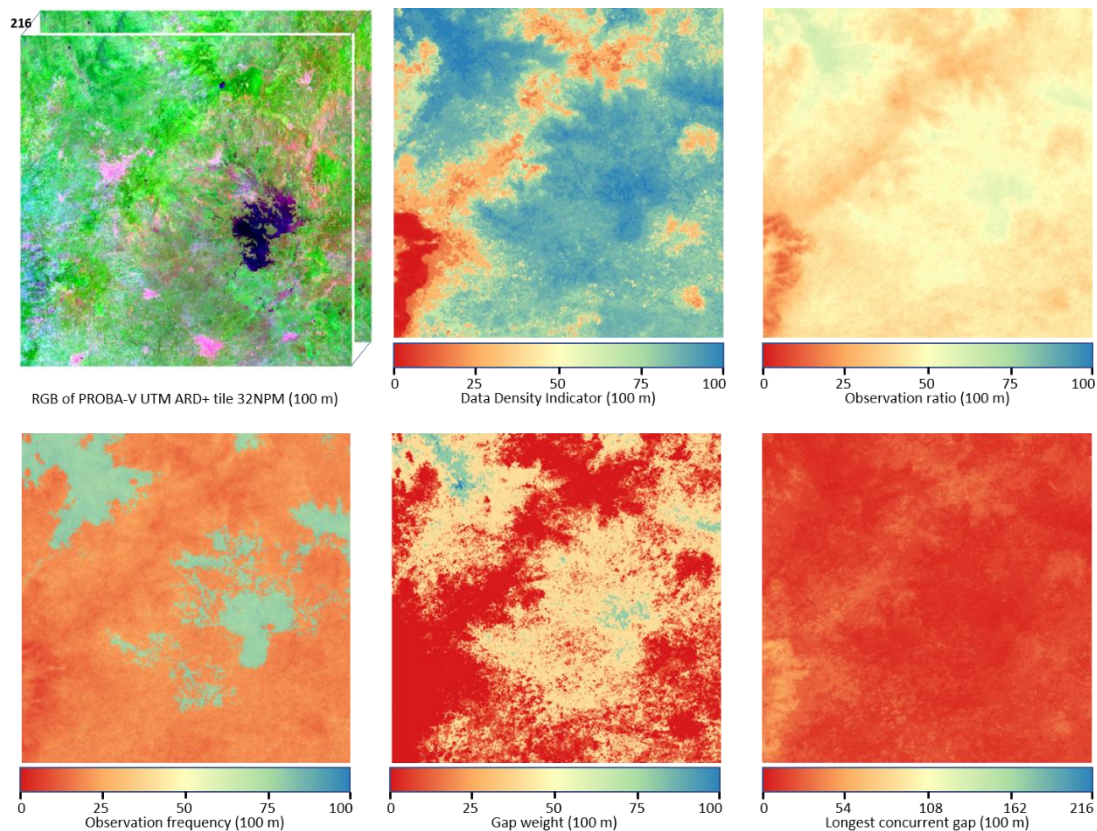


Figure 36: Visual RGB representation of UTM tile 32NPM (2016) with some of the calculated DDI statistics for the 100 m resolution, i.e. Data Density Indicator, Observation ratio, Observation frequency, Gap weight and Longest concurrent gap.

3.6 DATA FUSION

3.6.1 Overview

High seasonal cloud coverage in several regions of the globe, e.g. Sahel zone, Central America, are challenging for all land cover classification approaches depending on optical data solely (Figure 37).

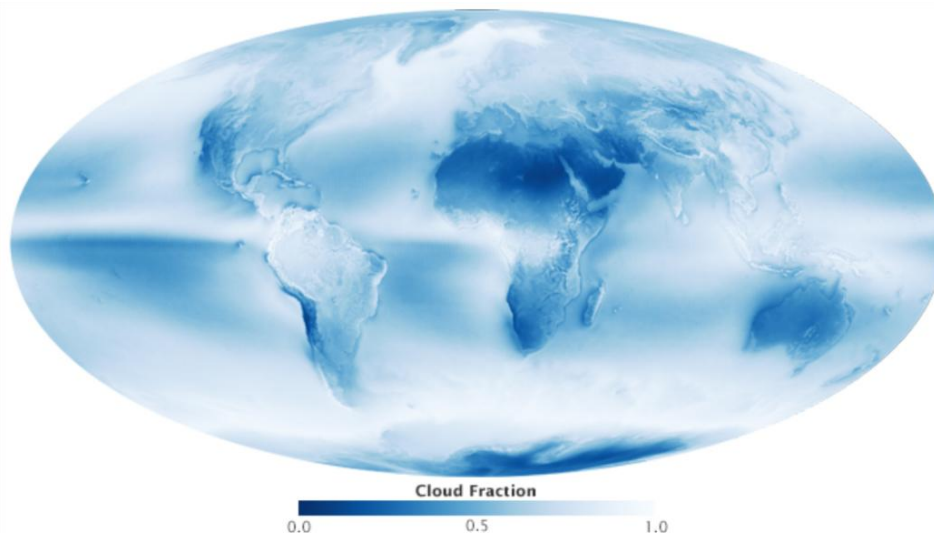


Figure 37: Average cloud cover over a 13 year period from July 2002 to April 2015 captured by the MODIS Aqua RS sensor. Note: Colours range from dark blue (no clouds) to light blue (some clouds) to white (frequent clouds). Image credits: NASA Earth Observatory, 2019.

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 38 illustrates the PROBA-V UTM 100 m and 300m data density for Africa.

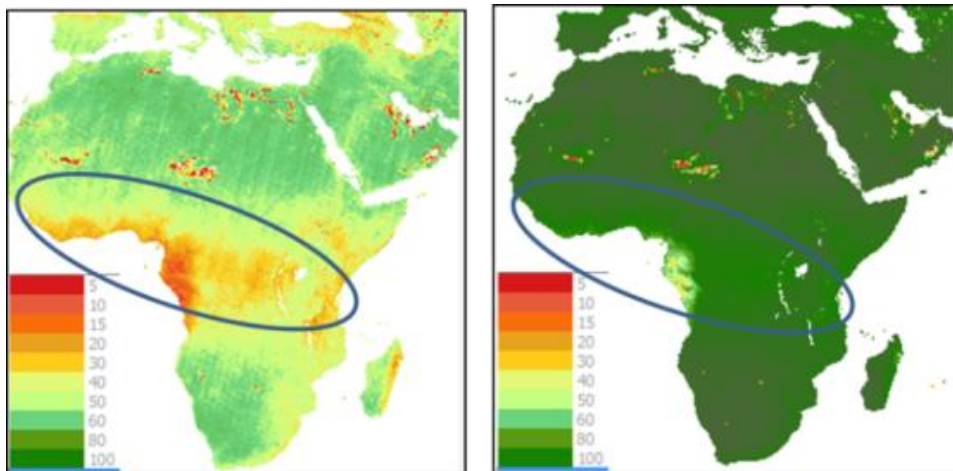


Figure 38: Number of cloud free data observations for PROBA-V UTM 100 m (left) and PROBA-V UTM 300 m (right) in the reference year 2015 for Africa.

Figure 39 illustrates an overview of the data fusion section in the CGLS-LC100 product workflow (Figure 1). 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 huge 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-daily image stacks for all global land masses at 100 m resolution (Note: 300 m data is not existing in the ARD++).

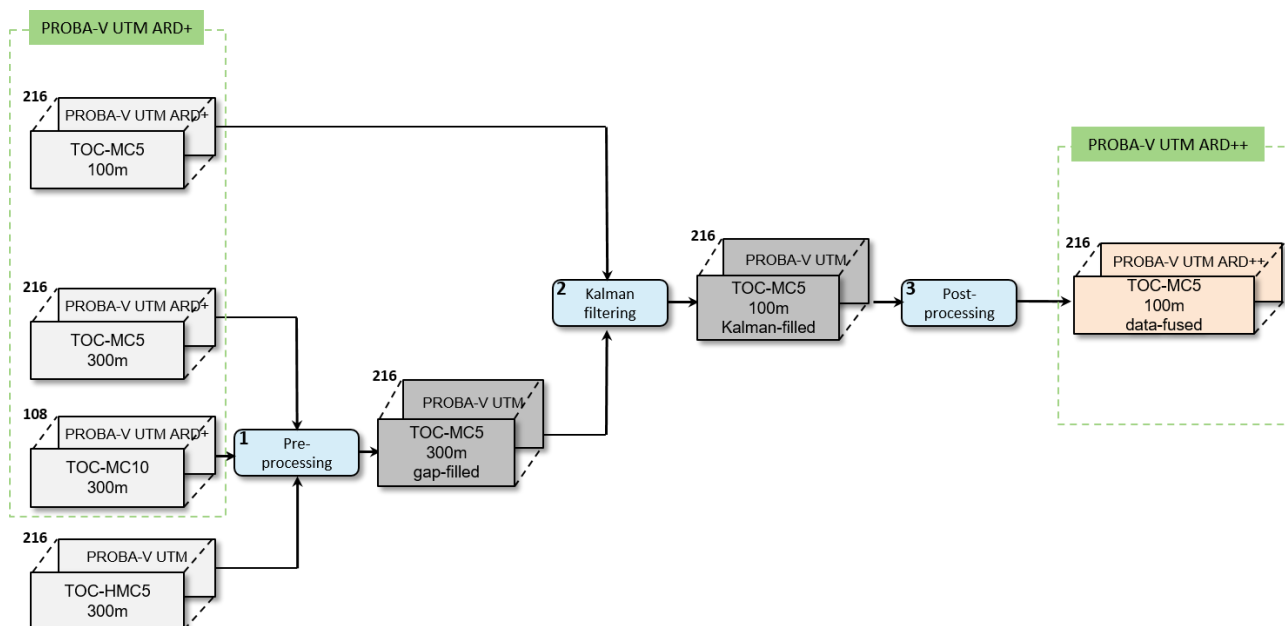


Figure 39: General overview of the data fusion section in the CGLS-LC100 product workflow. Note: Numbers in the upper left corner of data container indicate the number of layers - here number of 5-daily observations.

3.6.2 Data Fusion pre-processing

Input data in the data fusion pre-processing step are the generated PROBA-V UTM TOC-MC5, TOC-MC10 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-daily time series product are filled with the pixel values of the 10-daily time series product for the corresponding time steps. In a second step, bigger gaps (> 25 day gaps) in the 300 m 5-daily time series product are filled via interpolation with the HMC5 300 m long term trend product for the corresponding time series steps (see Figure 40). This is needed in order to guide the Kalman filtering approach in cases where no PROBA-V 100 m and 300 m TOC-MC5 data is available for more than 1 months in the 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 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 39).

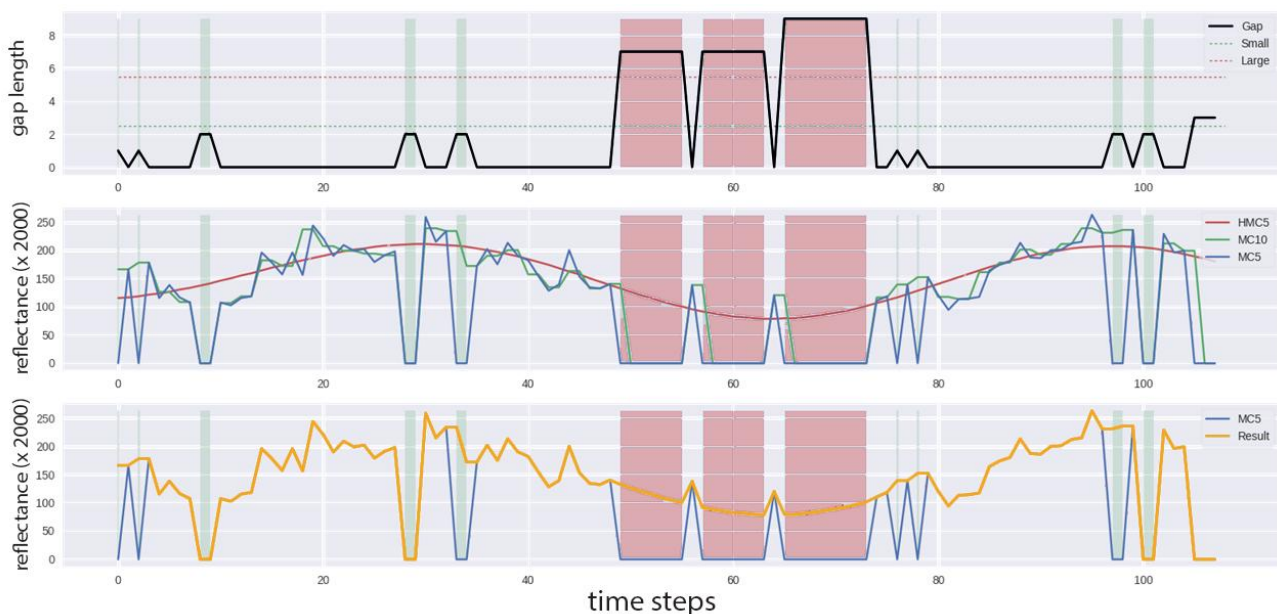


Figure 40: Example for data fusion pre-processing results for PROBA-V UTM TOC-MC5 300 m time series at pixel location 9.459° lon, 6.562° lat. top) continuous gap length in the 300 m time series (dashed green line indicates threshold for small gaps, dashed red line indicates threshold for big gaps), middle) 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 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 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 41 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.

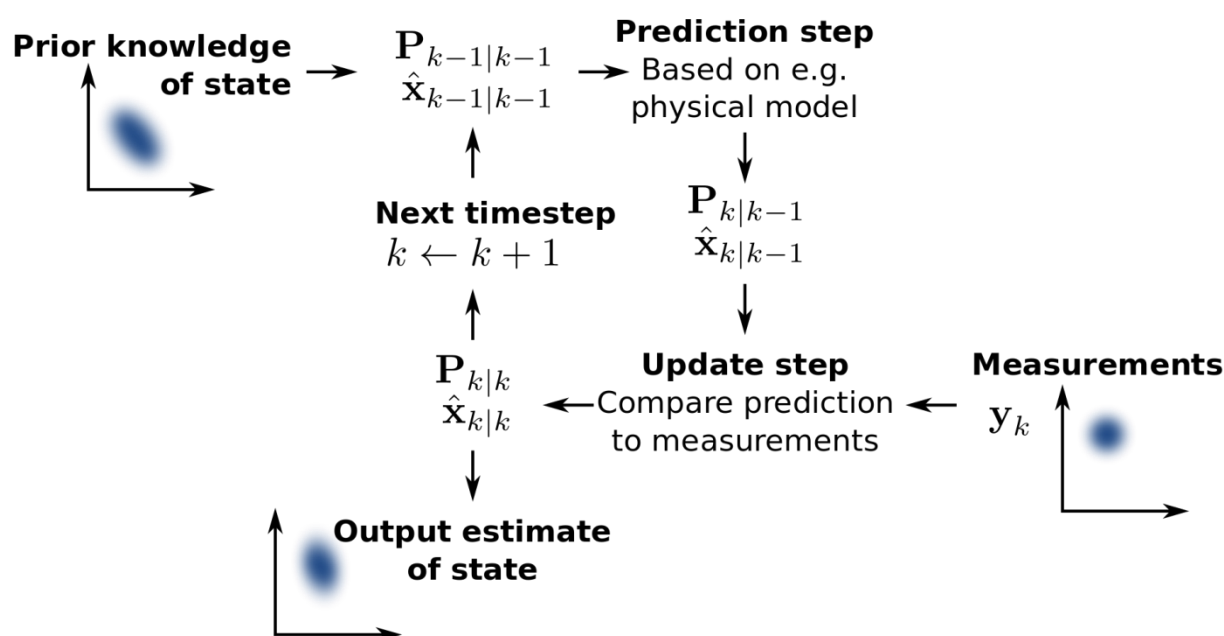


Figure 41. Basic concept of the Kalman filtering approach. Adapted from Aimonen (2011).

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 this 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 huge influence on the processing speed. In the first version of the CGLS-LC100 workflow this shortcoming 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 42 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).

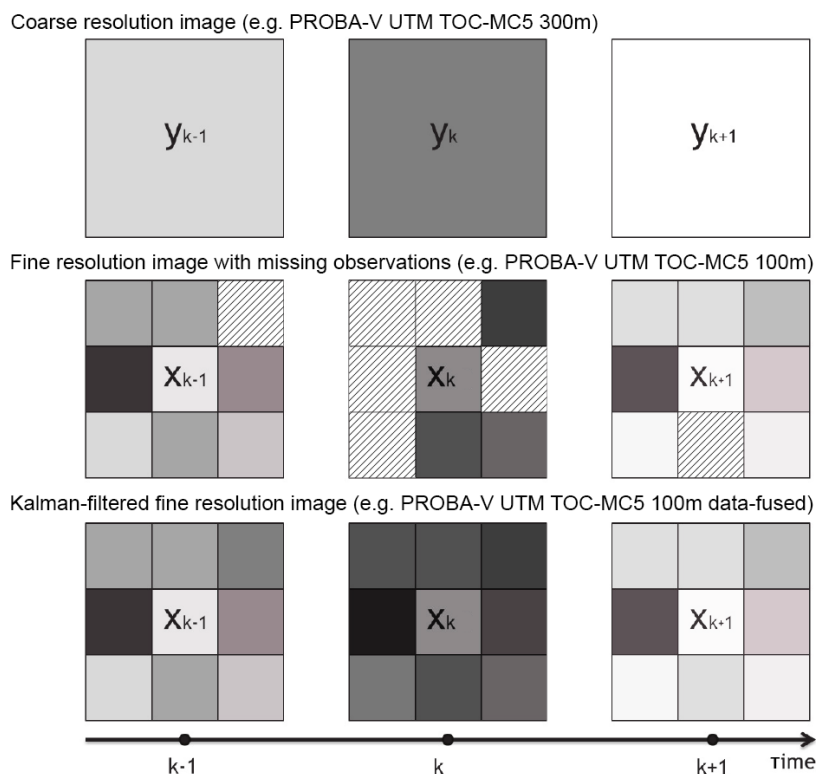


Figure 42: Principle of VITO's enhanced Kalman-filtering approach for EO data fusion based on the ptools algorithm. Note: Image adapted from Kempeneers et al. (2016).

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 43. The upper part in Figure 43 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.

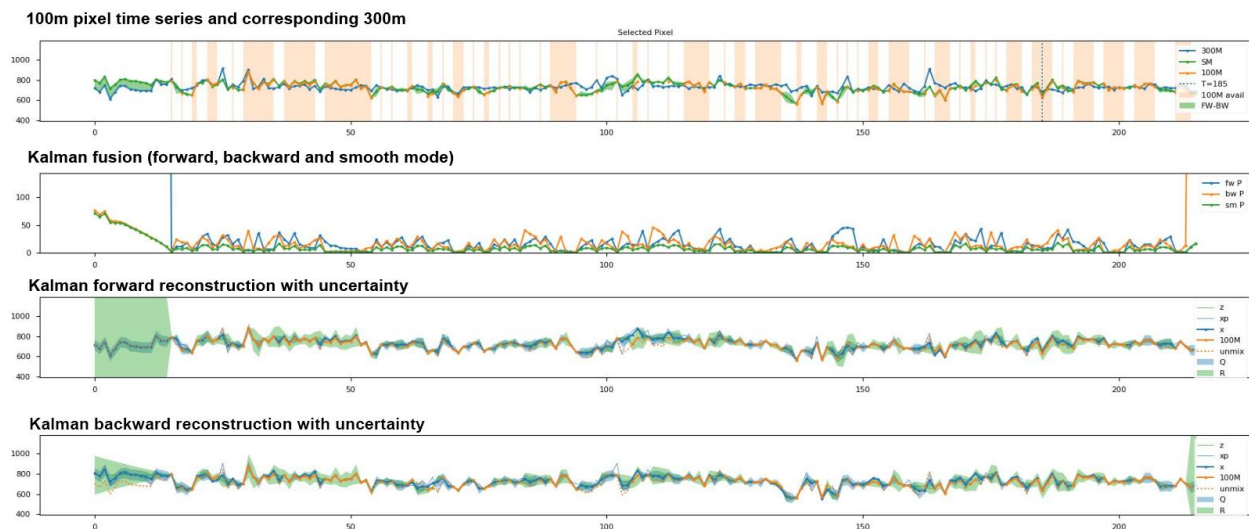


Figure 43: Example for 100 m time series reconstruction of one pixel using VITO's improved Kalman filter. Upper image shows original 100 m timeseries in orange (orange line shows existing values and orange shading shows missing time steps), the corresponding 300 m time series in blue and the final 100 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.

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 43 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 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 43 shows the Kalman backward reconstruction together with the calculated uncertainty, where the colour schema follows 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 43 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 43 where the final smooth time series are introduced in green (the uncertainty in light green, respectively) into the original TOC-MC5 100 m time series shown in orange where time steps were missing.

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 and the colour coding. The left side shows the full PROBA-V 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 43 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 gives a zoom in for the 100 m and 300 m tile as well as the 3x3 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 3x3 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 zoomed-in tile as well as for the 3x3 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-daily time series product in fine resolution – called PROBA-V UTM TOC-MC5 100 m Kalman-filled - for the four reflectance bands.

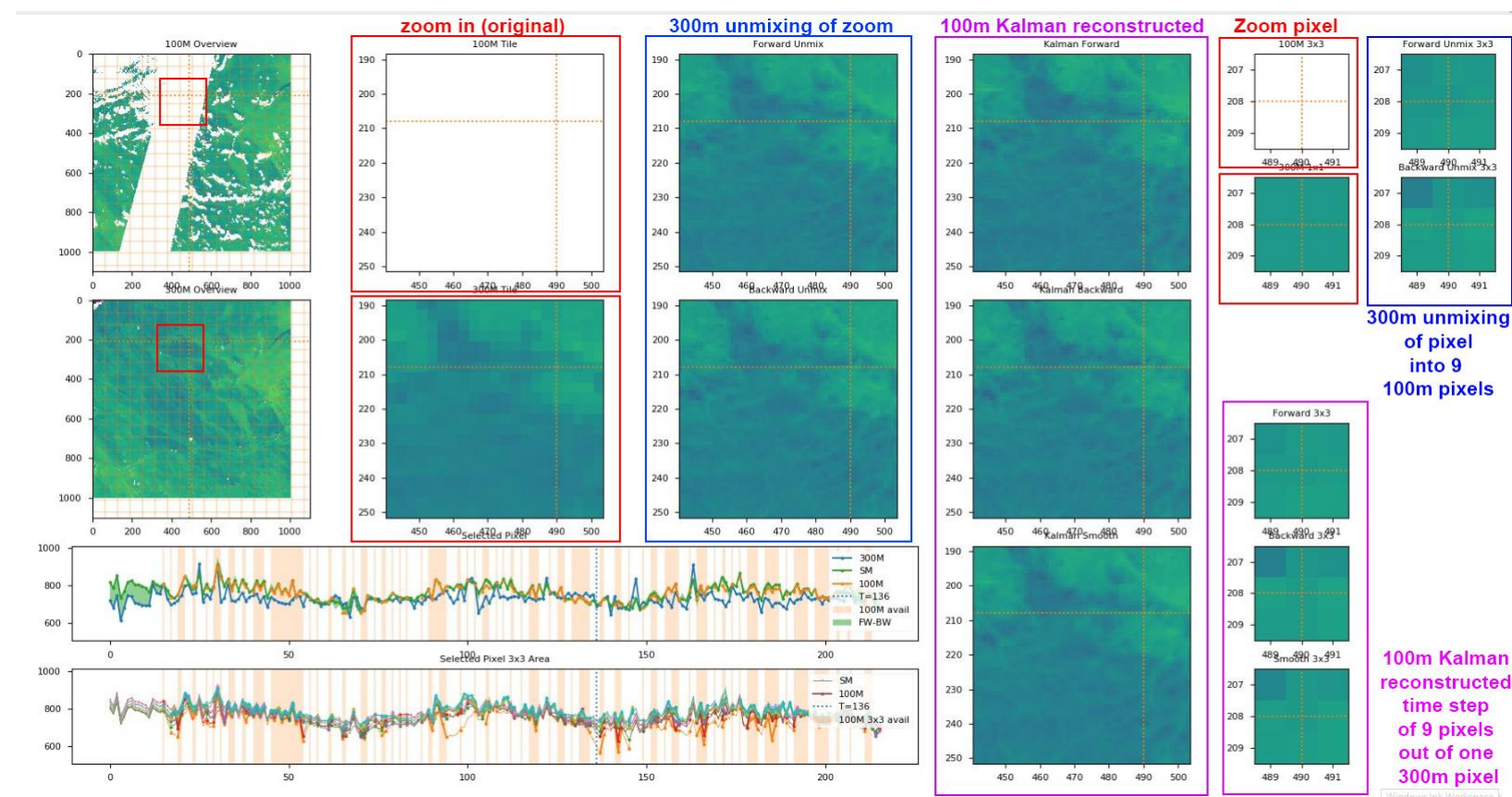


Figure 44: Visualization of data use with the improved Kalman filter for one 3x3 100 m pixel block in PROBA-V UTM ARD+ datasets. Note: red rectangles show the zoom into selected area of 100 m and corresponding 300 m tile, and the selected 300 m pixel for selected time series step (Note: 100 m has no 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 3x3 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 PROBA-V UTM TOC-MC5 100 m data is screened for consistency and introduced outliers (Figure 39). This 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 45.

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

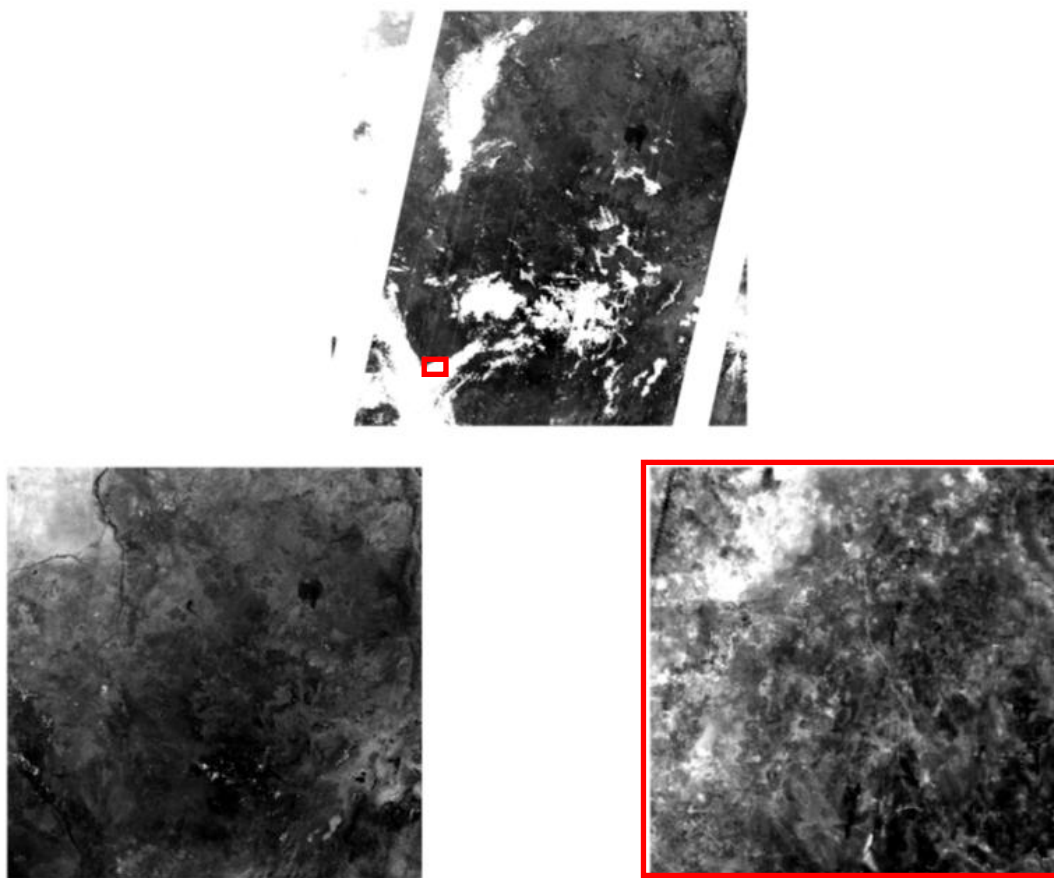


Figure 45: top) full PROBA-V UTM TOC-MC5 100 m tile for red reflectance channel (areas with missing data are shown in white), bottom-left) PROBA-V UTM TOC-MC5 100 m data fused image, bottom-right) zoom in to full PROBA-V UTM 100 m resolution of left image over red box shown in top image (the area in 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 1), the external datasets have to be pre-processed. This 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) create from the input datasets shown in section 3.2.5. 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 3.2.8 for more details).
- The combined DEM (see section 3.2.11 for more details).

Anyhow, some ancillary dataset had 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;
- Word Settlement Footprint (WSF) Built-up cover fraction product;
- WetProducts permanent and seasonal water cover fractions products which are generated out of the Global Surface Water (GSW) datasets.

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 sub-sections 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 46). The term “biome cluster” herby 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”*.

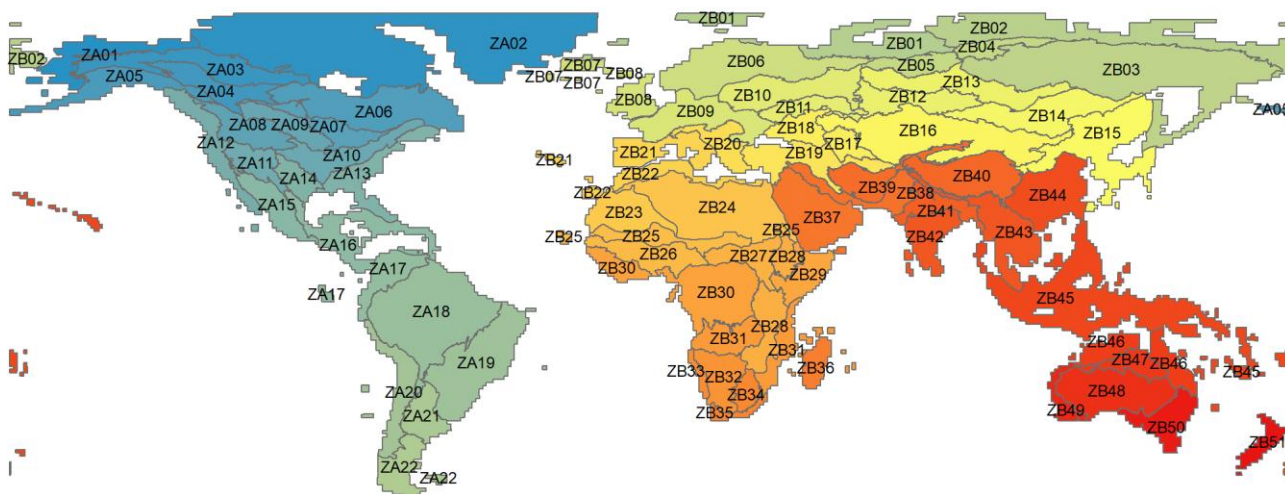


Figure 46: Global stratification into 73 biome clusters to improve regional classification/regression in the CGLS-LC100 workflow.

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. Input into the layer creation were well-known existing datasets (see section 3.2.7) 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 3.2.11) 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 46. The continental assignment of the biome clusters can be seen in Table 6. The final vector layer is directly used in several processing steps, but also a raster mask following the Sentinel-2 tiling and naming was generated for the final CGLS-LC100 discrete map generation.

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 WSF 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 Word Settlement Footprint (WSF) Built-up cover fraction product used in the CGLS-LC100 workflow was generated from the DLR's Word Settlement Footprint (WSF) dataset for 2015 (Marconcini et al., 2017, Marconcini et al., in review) (see section 3.2.9) showing an overall accuracy between 83.27 – 89.33 % depending on the “built-up” definition (Marconcini et al., in review).

Since the original built-up raster layer is provided in 10 m spatial resolution, we first resampled and projected the dataset to 10 m pixels in the UTM system and then applied a moving window approach in order to transform the binary mask into built-up cover fractions at 100 m spatial resolution. The final step was then the tiling and naming of the dataset following the Sentinel-2 tiling convention.

3.7.3 WetProducts permanent and seasonal cover fraction products 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 body detection (Figure 47). The main change compared to Bertels et al. (2016) algorithm is the incorporation of the GSW maximum water extent cover fraction layer, the GSW seasonal water cover fraction layer and the GSW permanent water cover fraction layer which are derived from JRC's GSW maximum water extent and water seasonality 2014-2015 layers (Pekel et al., 2016). The maximum water extent provides information on all the locations ever detected as water over the Landsat data archive period. The water seasonality layer gives information on the number of months water was detected for the

concerned year. In pre-processing step (Figure 47) the two external GSW layers (see section 3.2.10) are first projected to UTM and resampled to 10 m spatial resolution and subsequently these layers are used to calculate the three GSW cover fraction layers on 100 m resolution, i.e. the GSW maximum water extent cover fraction, GSW seasonal water cover fraction 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 maximum water extent and water seasonality 2014-2015 layers show 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 these accuracy could not be reached with an own algorithm, our water detection based on Bertels et al. (2016) algorithm is more an add-on. In detail, to the time of the CGLS-LC100 algorithm version 2 development, the JRC's GSW layers were only available up to 2014-2015 (not fully covering the year 2015). Therefore, we use the GSW results transferred into 100 m resolution and cover fractions, but had to implement our own water detection algorithm to close the temporal gap in the external dataset. As shown by Figure 47, the water body detection algorithm for the reference year 2015 uses the PROBA-V UTM ARD++ derived VI data layers HUE, VALUE and NDVI for a per pixel water body detection while taking into account the three GSW derived cover fraction layers. The three output layers, i.e. permanent water cover fraction, seasonal water cover fraction and the wetland indication mask, are calculated by applying specific thresholds (empirically defined, cross-checked with against reference points and Pekel et al., 2016) on HUE, VALUE and NDVI. Finally the Land/Sea mask (see section 3.2.2) is used to clean-up the permanent water cover fraction and seasonal water cover fraction layers.

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) the pixel must be indicated by the maximum water extent mask; and
- (ii) its 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
- (iii) 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. In combination with the GSW derived cover fractions, decisions can be made on the actual cover fraction. For this we need at least 10 valid observations and the GSW maximum water extent must be greater than 0. Subsequently we define a pixel as permanent water when the detected water occurrence is over 90 %. Otherwise, when at least 5% water was detected, the pixel is defined as seasonal water. 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) 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.

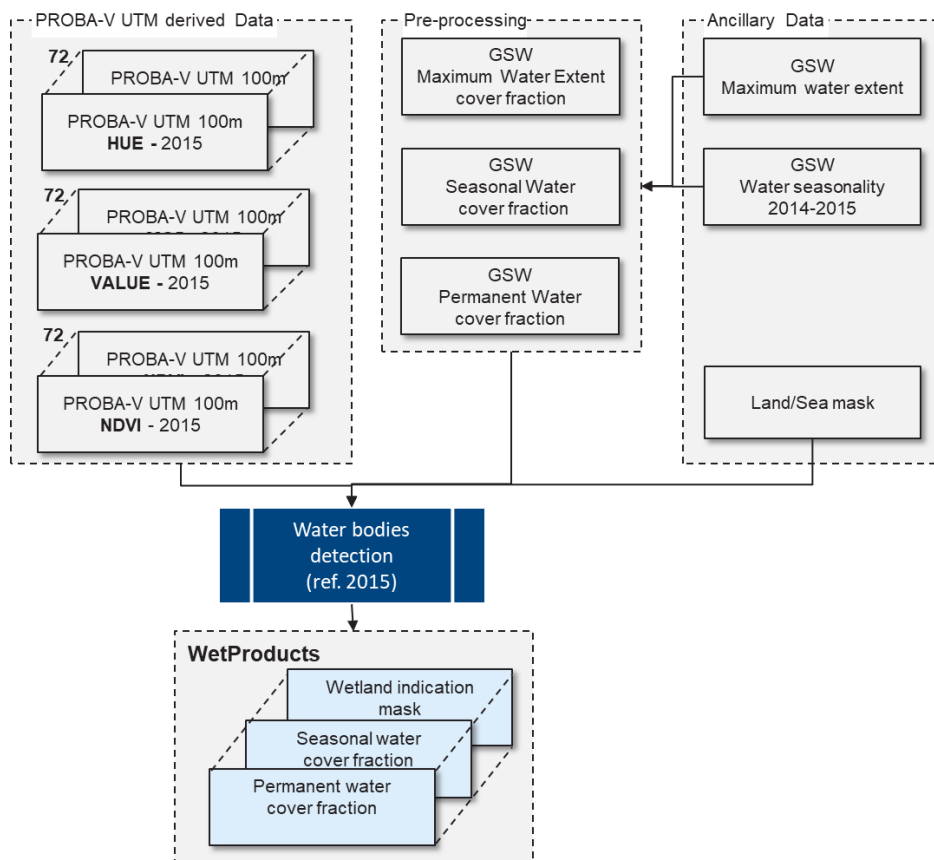


Figure 47: General overview of the water bodies detection algorithm. Besides the PROBA-V UTM derived VI data products, i.e. HUE, VALUE and NDVI, three ancillary layers are used as input for the WetProducts generation.

3.8 METRICS GENERATION

3.8.1 Overview

Figure 48 gives an overview about the metrics generation section in the CGLS-LC100 product workflow (Figure 1). The term “metrics” hereby refers in 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. 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. 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 14 layers (the four PROBA-V reflectance bands, 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 14 layers. Also a textural metric showing the uniformity of a pixel compared to its neighbouring pixels (3x3 box) is generated. In the last processing step, the post-processing, the overall 270 generated metrics for each PROBA-V UTM 100 m pixels are combined into one container.

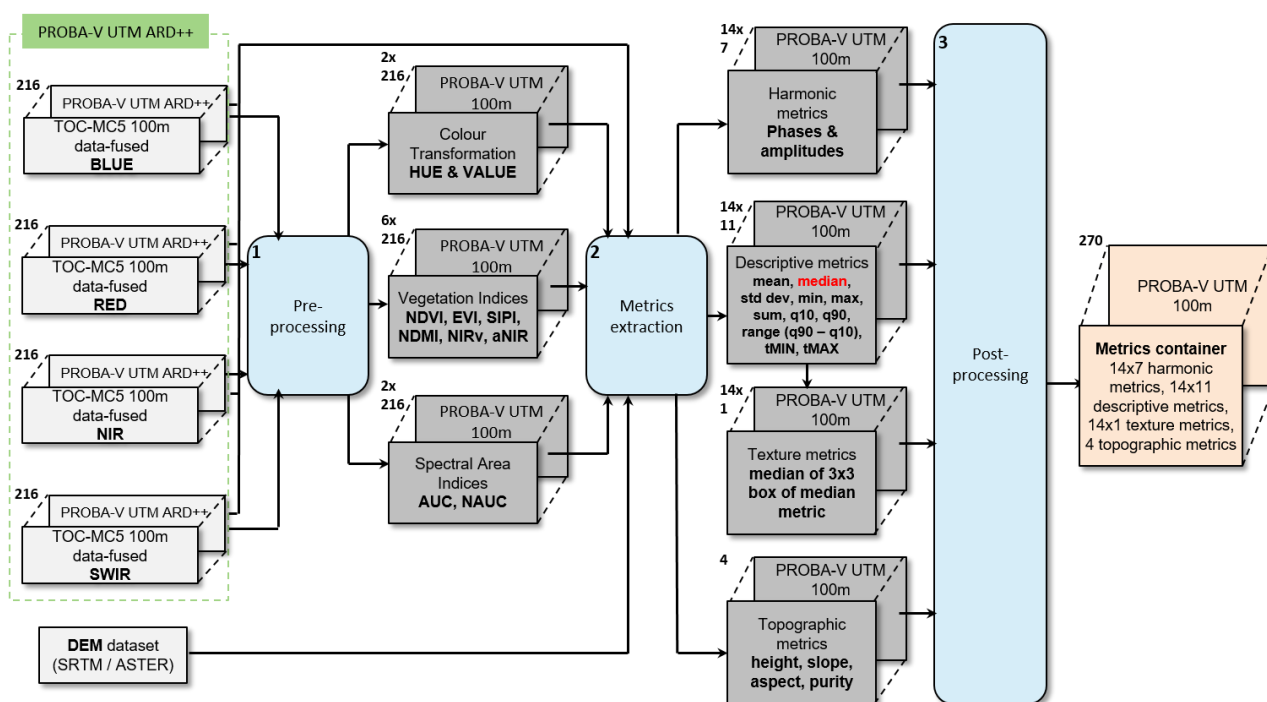


Figure 48: General overview of the metrics generation section in the CGLS-LC100 product workflow.
Note: datasets marked in red are re-used in following sub-steps. Numbers in the upper left corner of data containers indicate the number of layers.

3.8.2 Pre-Processing

The pre-processing step of the metrics generation section can be subdivided into three sub-steps (Figure 48): 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; 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:

$$NDVI = \frac{NIR - red}{NIR + red} \quad \text{Eq. 10}$$

, 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:

$$EVI = 2.5 * \frac{NIR - red}{(NIR + 6 * red - 7.5 * blue + 1)} \quad \text{Eq. 11}$$

, 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 -inf to inf. 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:

$$SIP I = \frac{NIR - blue}{NIR - red} \quad \text{Eq. 12}$$

, 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:

$$NBR = \frac{NIR - SWIR}{NIR + SWIR} \quad \text{Eq. 13}$$

, where NIR and SWIR refers to the reflectance value of the corresponding PROBA-V reflectance bands.

5. Near-Infrared reflectance of vegetation (NIR_v) using the red and the NIR reflectance bands (Badgley et al., 2017). The NIR_v is one of the newest VI'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, NIR_v 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:

$$NIR_v = \left(\frac{NIR - red}{NIR + red} - 0.08 \right) * red \quad \text{Eq. 14}$$

, 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 2π radians (Equation 15).

$$aNIR = \arccos \left(\frac{a^2 + b^2 + c^2}{2ab} \right) \quad \text{Eq. 15}$$

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:

$$a = \sqrt{(red - NIR)^2 + (WL_{red} - WL_{NIR})^2} \quad \text{Eq. 16}$$

$$b = \sqrt{(NIR - SWIR)^2 + (WL_{NIR} - WL_{SWIR})^2} \quad \text{Eq. 17}$$

$$c = \sqrt{(red - SWIR)^2 + (WL_{red} - WL_{SWIR})^2} \quad \text{Eq. 18}$$

,with WL_{Bx} 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 49.

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 PROBA-V 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 49). 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° to 360°. 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.

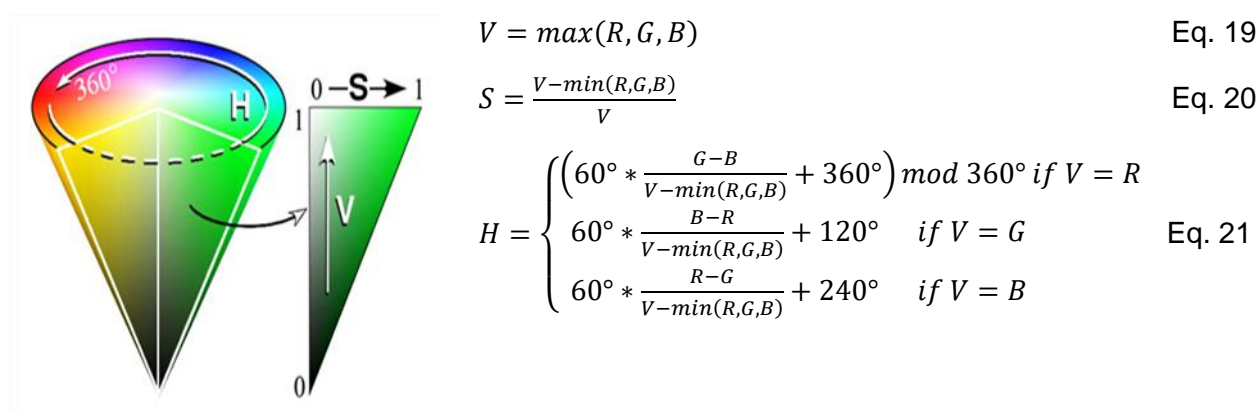


Figure 49: The HSV colour space and the formulas for transforming the RGB colour space into the HSV colour space. Note: V=Value, S= Saturation, H=Hue, R=red, G=green, B=blue.

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 50). 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.

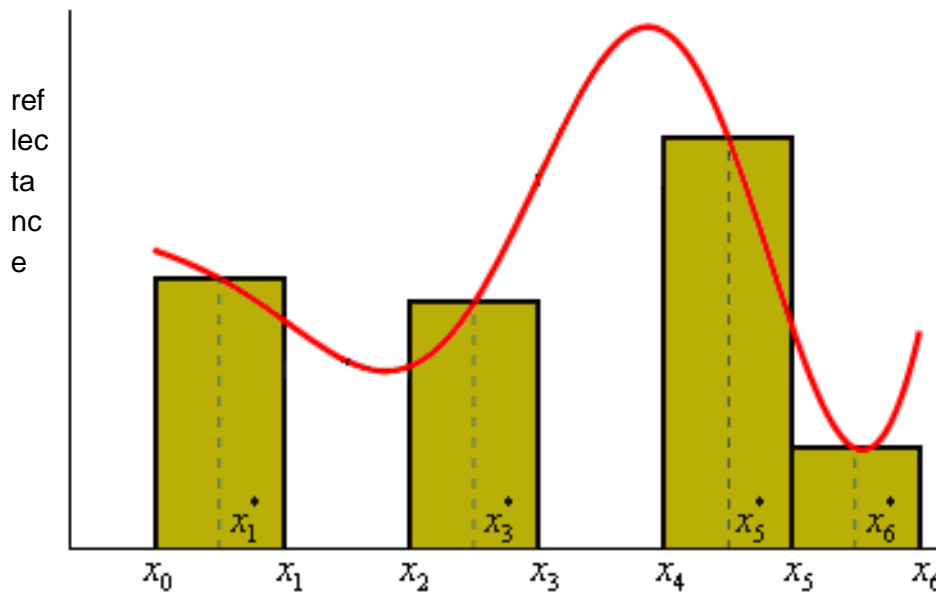


Figure 50: Spectral Area Indices explained. The sum of the FWHM reflectance (reflectance over the full band width) for all reflectance bands can be used as single scalar to compare spectral signatures.

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 50). Equation 22 shows the used formula:

$$AUC = \sum_{x_B}^n x_B * FWHM_{x_B} \quad \text{Eq. 22}$$

, where x_B is one of the PROBA-V bands blue, red, NIR and SWIR and $FWHM_{x_B}$ 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:

$$NAUC = \sum_{x_B}^n (x_B - M) * FWHM_{x_B} \quad \text{Eq. 23}$$

, where x_B is one of the PROBA-V bands blue, red, NIR and SWIR and $FWHM_{x_B}$ is the bandwidth of that band (Table 3), and $M = \text{argmin}(\text{blue}, \text{red}, \text{NIR}, \text{SWIR})$ 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 48), 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 and allow so the analysis of complex inter-annual behaviors of vegetated and non-vegetated land areas via easy to interpretable and comparable variables. Input in the metrics extraction step are the four reflectance bands of the PROBA-V UTM TOC-MC5 100 m data fused (PROBA-V UTM ARD++) time series profiles (see section 0), the six PROBA-V UTM MC5 VI's time series profiles at 100 m 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 14 input layer for which the harmonic, descriptive and textural metrics are separately extracted. In a last step, the topographic metrics of the DEM (see section 3.2.11) are 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 CGLS-LC100 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 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 14 input bands/layers simultaneously.

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 for each pixel for all 14 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 HANTS-modelling 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 98 harmonic metrics (7 model parameter for each of the 14 input bands/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-daily 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 51 shows an example for a harmonized time series profile of the NDVI.

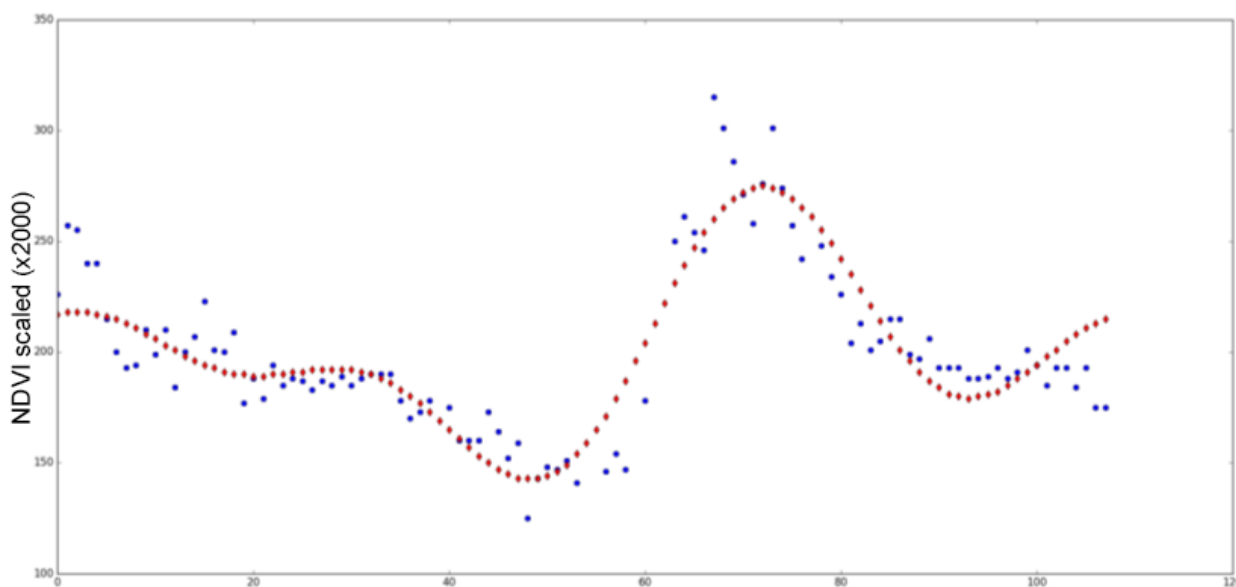


Figure 51: Example for a harmonized time series profile. The result of the HANTS model is shown in red dots, where the blue dots show the real observations.

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 11 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. minimum: to identify extrema in the time series;
4. maximum: to identify extrema in the time series;
5. sum: to quantify the overall throughput 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;
6. time step of the first appearance of the minimum value: to identify the day of the year when the extrema appeared;
7. time step of the first appearance of the maximum value: to identify the day of the year when the extrema appeared;
8. median: to identify the “middle” value which separates the higher half of the pixel values in the time series from the lower one;
9. 10th percentile: to identify the pixel value below which 10% of the pixel values of the time series can be found;
10. 90th percentile: identify the pixel value below which 90% of the pixel values of the time series can be found, and
11. 10th – 90th percentile range: to identify the interval in which 80 % of the pixel values of a time series are located.

The 11 descriptive statistics are independently calculated for each of time series profiles of the 14 input bands/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-daily 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 steps values and this can lead to false interpretation.

Output of this sub-step are overall 154 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 14 bands/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 3x3 moving window for each pixel. 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 14 textural metrics (one for each of the 14 input bands) which are directly used in the classification/regression approach.

3.8.3.4 Topographic metrics

The separation of classes only by metrics from spectral data is often imperfectly, 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 into the CGLS-LC100 classification/regression. Input data is the external DEM (see section 3.2.11) created out of the combination of the SRTMGL1 (NASA, 2013) and ASTER global DEM (NASA/METI/AIST/Japan Spacesystems, and U.S./Japan ASTER Science Team, 2009). 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 dataset. 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 be not confused with the common Topographic Position Index (TPI). The purity metric is calculated by identifying the highest absolute difference between the center pixel and its eight neighbors in a 3x3 box around this center 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.

Output of this sub-step are overall 4 topographic metrics (elevation, slope, aspect, and purity) which are directly used in the classification/regression approach.

3.8.4 Post-Processing

The last processing step in the metrics generation section (Figure 48), the post-processing, handles all calculated 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, 270 metrics - 7 harmonic metrics plus 11 descriptive metrics plus 1 textural metric for the 14 time series profiles (4 reflectance bands, 6 vegetation indices, 2 colour transformed bands, and 2 spectral area indices) plus 4 topographic metrics - are combined in the VRT file. This VRT file is then input in the classification/regression section of the automated processing chain. Figure 52 shows a visualization of some selected metrics.

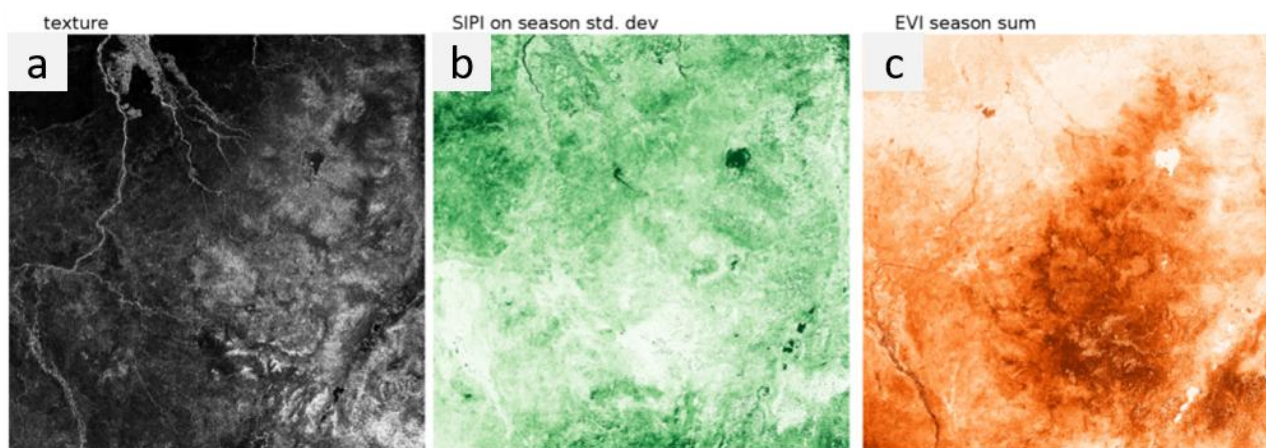


Figure 52: Examples of the 270 derived metrics for tile X18Y06. a) texture metric (the lighter the colour the more homogeneous is the pixel compared to its surrounding pixels), b) standard deviation of the Structure Intensive Pigment Index for the vegetation season (the greener the colour the higher 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

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 150,405 unique locations. 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 checked 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 141,000 sample sites (Figure 53).

Since the classification/regression is conducted in geographical subsets, 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 (see section 3.5) was applied to ensure that all training points have sufficient time series data. In detail, training points in locations with a DDI value of 0 meaning that no PROBA-V UTM time series data was available for this location was removed from further processing.

In the next step, the PROBA-V UTM time series metrics (see section 3.8.4) were 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.

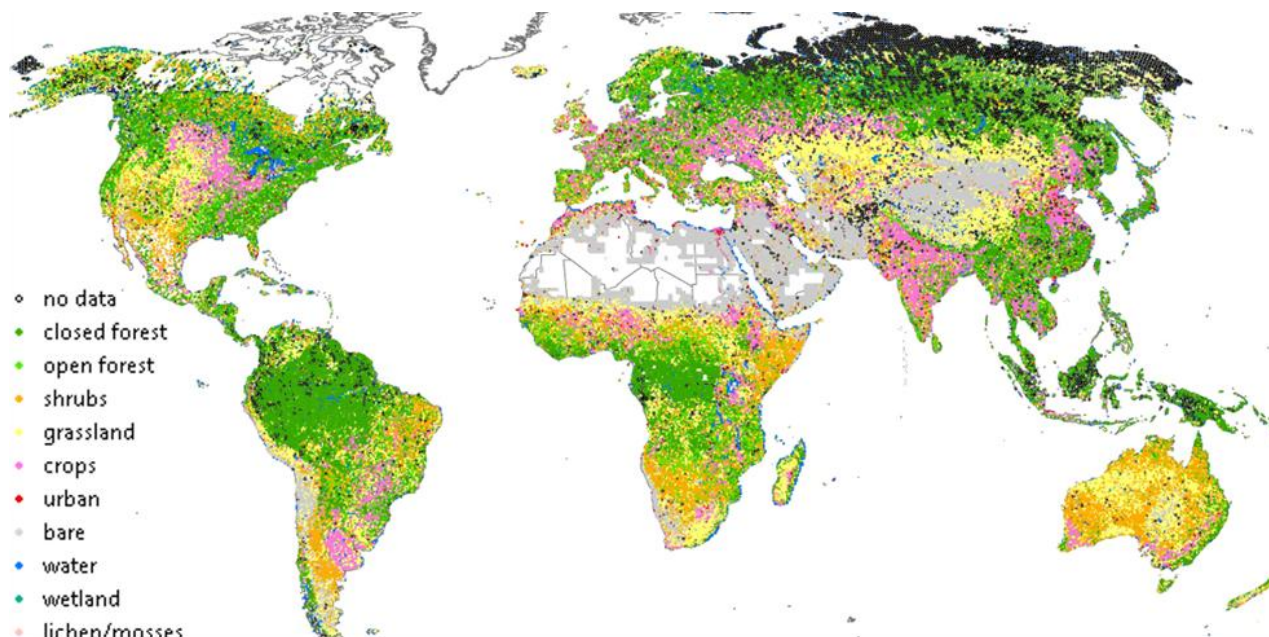


Figure 53: Training dataset used in the CGLS-LC100 workflow showing the ~141K training locations together with their sampled land cover class at 100 m.

3.10 CLASSIFICATION / REGRESSION

3.10.1 Overview

In order to adapt the classification/regressor algorithms to regional patterns, the classification/regression of the data is carried out in data subsets (73 biome cluster subsets) using the created global biome clusters layer (see section 3.7.1). Prior to the classification/regression of the EO data (Figure 1), additional pre-processing and optimization steps of the training data are needed. Therefore, the 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. Finally, the land cover classification and regression to estimate the cover fractions for each pixel is conducted scenario-based. 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 54). 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).

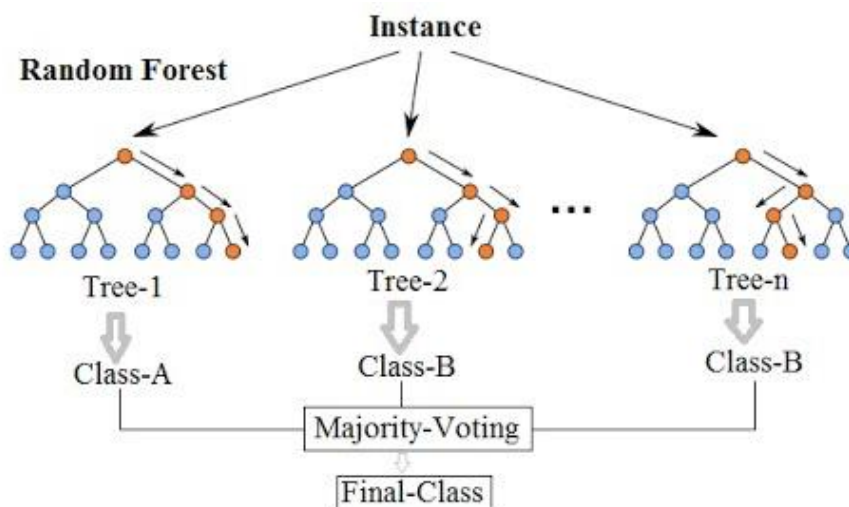


Figure 54: The Random Forest classifier principle. Image by Niraj (2016).

3.10.3 Training data and classifier/regressor optimization

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 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.

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 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 **N** by **N** matrix, where **N** is the total number of training point spectra in the biome cluster (Figure 55);

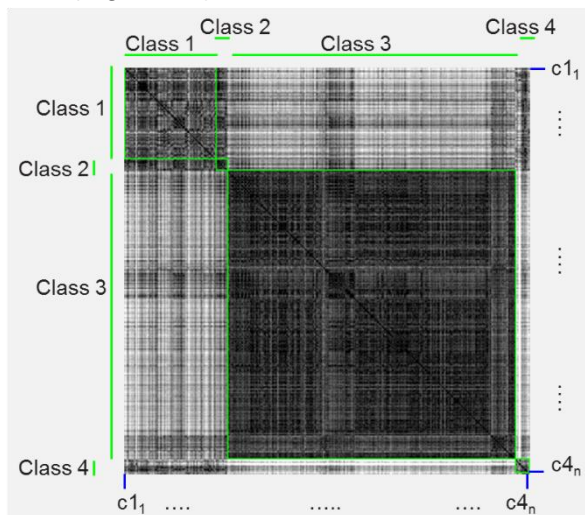


Figure 55: Rule matrix calculated for the total of N training points which belong to four different LC classes. For each training point the RMSE value is calculated against all the other training points, i.e. c11 with c11, c12, c13 till c4n, subsequently c12 with c11, c12, c13 till c4n and so on.

3. The rule matrix is used to exclude inter-class outliers. E.g. when the median RMSE value of “training point **c1₁** calculated with the remaining spectra of its own class (Class 1)” is greater than the median RMSE value of “training point **c1₁** calculated with all the spectra of the other classes” then training point **c1₁** 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 270 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. 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:

$$S_{AB} = \frac{nA + nB}{totA + totB} \quad \text{Eq. 24}$$

Here S_{AB} is the separability between LC class A and LC class B with nA and nB 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. 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 hyper-parameter. 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 Scenario-based Classification

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 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 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 with different settings are carried out:

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 Scenario-based Regression

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), and using the biome cluster-specific generated training datasets, selected best metrics and hyper-parameters. Again, the DDI (see section 3.5) for each training point is used as weight factor during the RF regression model training.

An important change compared to algorithm version 1 is the additional calculation of a RF regression quality indicator and the 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 preparation) 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. The RF quality indicator for each scenario is given 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.

Overall six regression scenarios for each biome cluster are carried out:

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 CGLS-LC100 COVER FRACTION 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), two additional processing steps are

needed in order to generate the 16 CGLS-LC100 cover fraction products (9 discrete base classes and the seasonal water cover fraction layer, six cover fraction quality indicator). The two processing steps are outlined in Figure 56. Main input data layers come from the scenario-based classification results, the scenario-based regression results together with their corresponding 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).

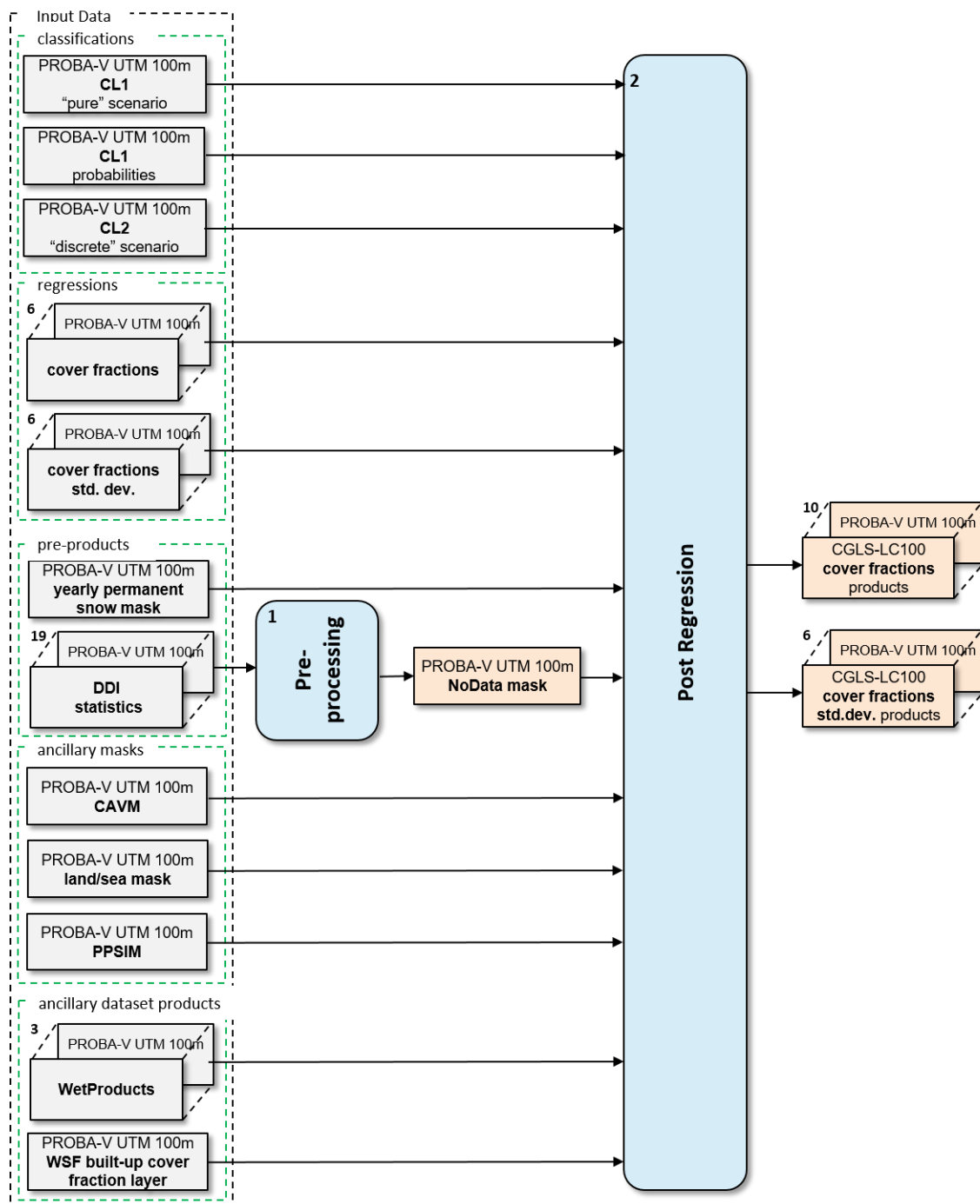


Figure 56: General overview of the cover fraction layers generation section in the CGLS-LC100 product workflow.

In the first processing step – the pre-processing-, the DDI statistics are analysed to generate the PROBA-V UTM 100 m NoData mask. Therefore, the “N° 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). 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 56). The next sub-section explains the post-regression in more detail (see section 3.11.2).

Output of the cover fraction layers generation section are the CGLS-LC100 cover fraction products consisting of ten thematic cover fraction layers, cover fraction quality layers for six of the thematic layers, and the PROBA-V UTM 100m NoData mask.

3.11.2 Cover fraction products generation

The final processing step in the generation of the CGLS-LC100 cover fraction products is the regression post-processing. Here, the 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 57 and Figure 58.

The five input dataset groups are in detail (see also Figure 57):

- Three products of the 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 layers for the classes forest, shrub, herbaceous vegetation, agriculture, moss & lichen, and bare (see section 3.10.5);
- The six regression quality indicator layers of the PROBA-V 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 WSF built-up cover fraction layer (see section 3.7.2) as part of the ancillary dataset products.

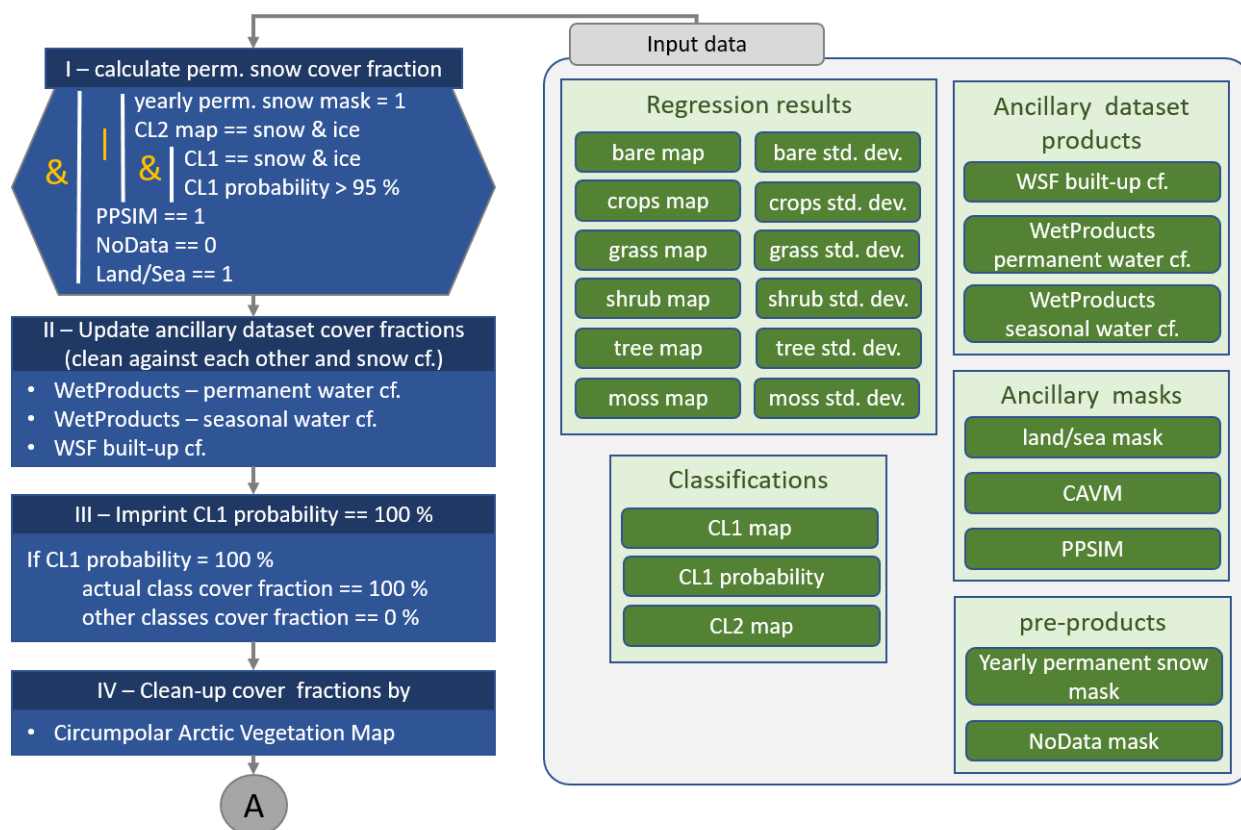


Figure 57: Rule block 1 to 4 of the CGLS-LC100 cover fraction layer generation – generation of snow, built-up, permanent and seasonal water cover fraction and cleaning by CL1 and CAVM.

The final cover fraction layers and their quality indicators are calculated in six sub-steps. First, the CGLS-LC100 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 57).

Second, the final 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 WSF built-up cover fraction layer are cleaned against the final permanent snow&ice cover fraction layer. This 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 is cleaned against the WSF built-up cover fraction layer. 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 WSF 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 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 warming).

Sub-step 5 is the most important step in the whole post-regression section (Figure 58). 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 factor 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).

$$100 \% - \text{builtup cover fraction} - \text{snow cover fraction} - \text{permanent water cover fraction}$$

=

$$\begin{aligned} & \text{tree cf.} * \text{tree conversion} + \text{shrub cf.} * \text{shrub conversion factor} + \\ & \text{grass cf.} * \text{grass conversion factor} + \text{bare cf.} * \text{bare conversion factor} + \text{Eq. 25} \\ & \text{crop cf.} * \text{crop conversion factor} + \text{moss cf.} * \text{moss conversion factor} \end{aligned}$$

$$\text{conversion factor}^{\textcircled{*}} = \frac{\text{regression std.dev.}^{\textcircled{*}}}{\sum \text{regression std.dev.}^{\textcircled{*}}} \quad \text{Eq. 26}$$

$\textcircled{*}$ 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.

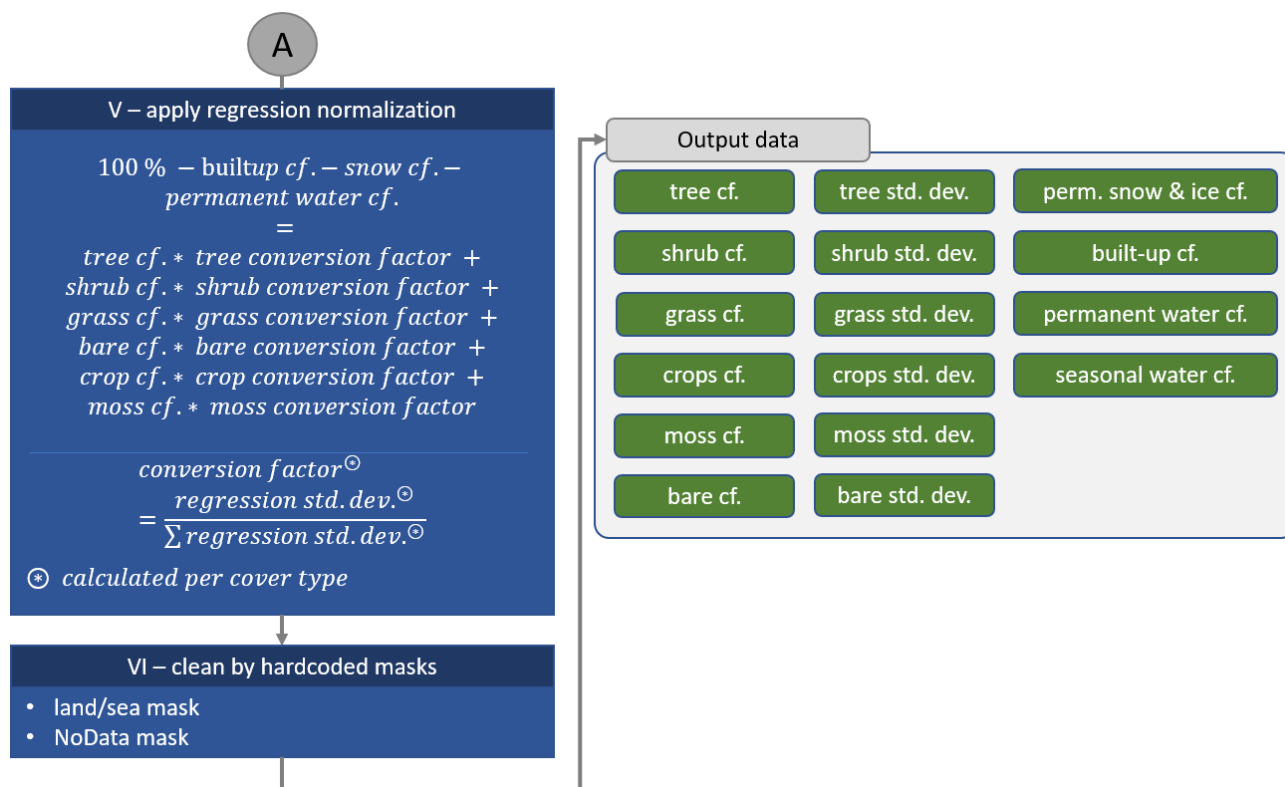


Figure 58: Rule block 5 and 6 of the CGLS-LC100 cover fraction layer generation – cover fraction normalization and cleaning by the land/sea mask.

3.12 CGLS-LC100 LAND COVER MAP GENERATION

3.12.1 Overview

In order to generate the CGLS-LC100 discrete map products, two processing steps are needed (Figure 59). Main input data layers come from the scenario-based classification results, the final CGLS-LC100 cover fraction products, some products from earlier processing steps, and the ancillary data masks as well as ancillary dataset products (see section 3.7).

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 out of the CGLS-LC100 cover fraction products as well as the final CGLS-LC100 forest type map. 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 60 shows the pre-processing for the new layers in more detailed. 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 CGLS-LC100 cover fraction layers (see section 3.11). 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 60. This discrete LC map is called Cover-2-discrete (CL5) during the expert rules.

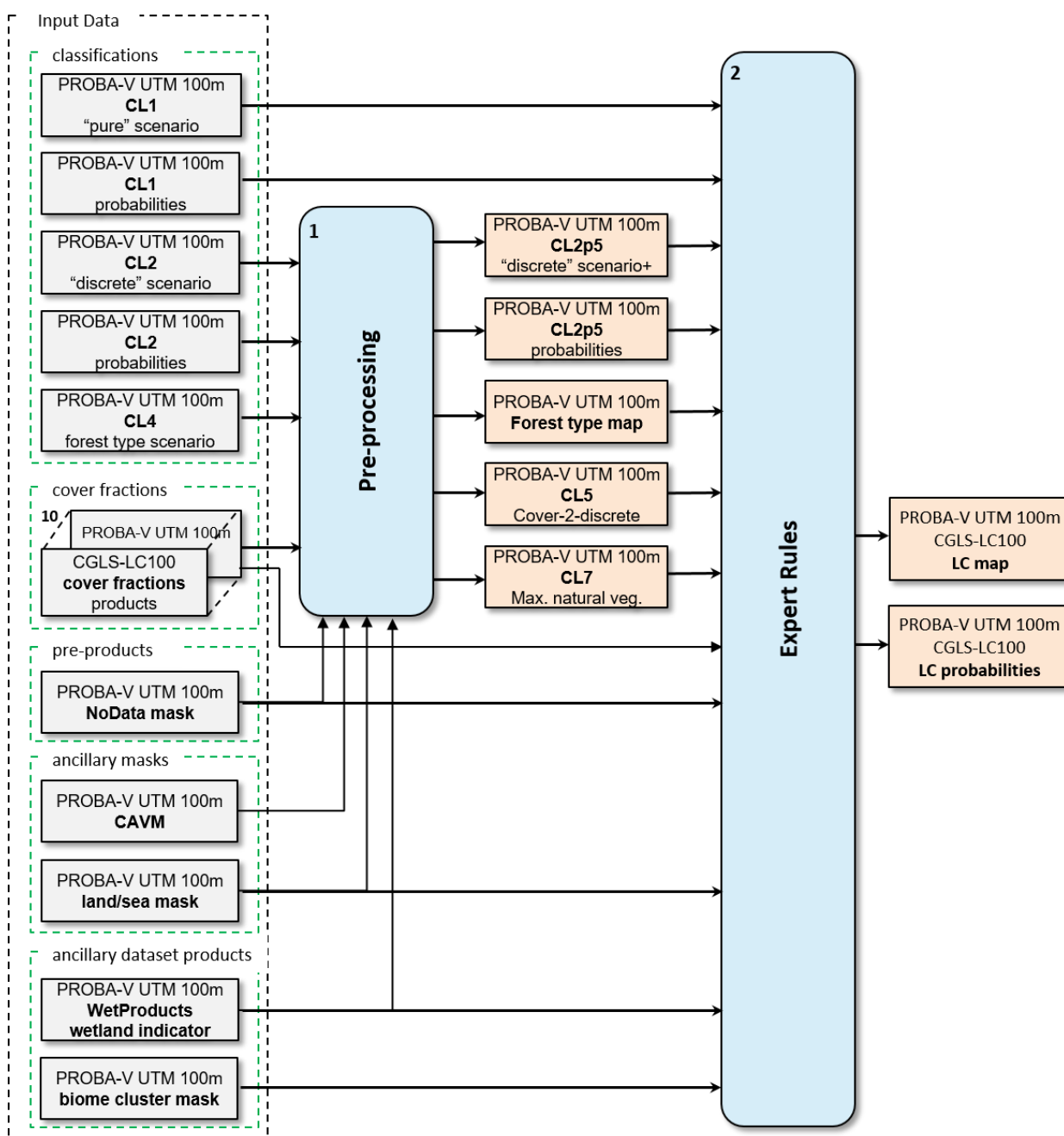


Figure 59: General overview of the Land Cover map generation section in the CGLS-LC100 product workflow.

Next, the maximum natural vegetation mask is generated. Therefore, the CGLS-LC100 cover fraction products 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 60. The last pre-processing step is the generation of the CGLS-LC100 forest type map, whereby the CL4 scenario-based classification map is only cleaned up by the CGLS-LC100 tree cover fraction layer. 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 60).

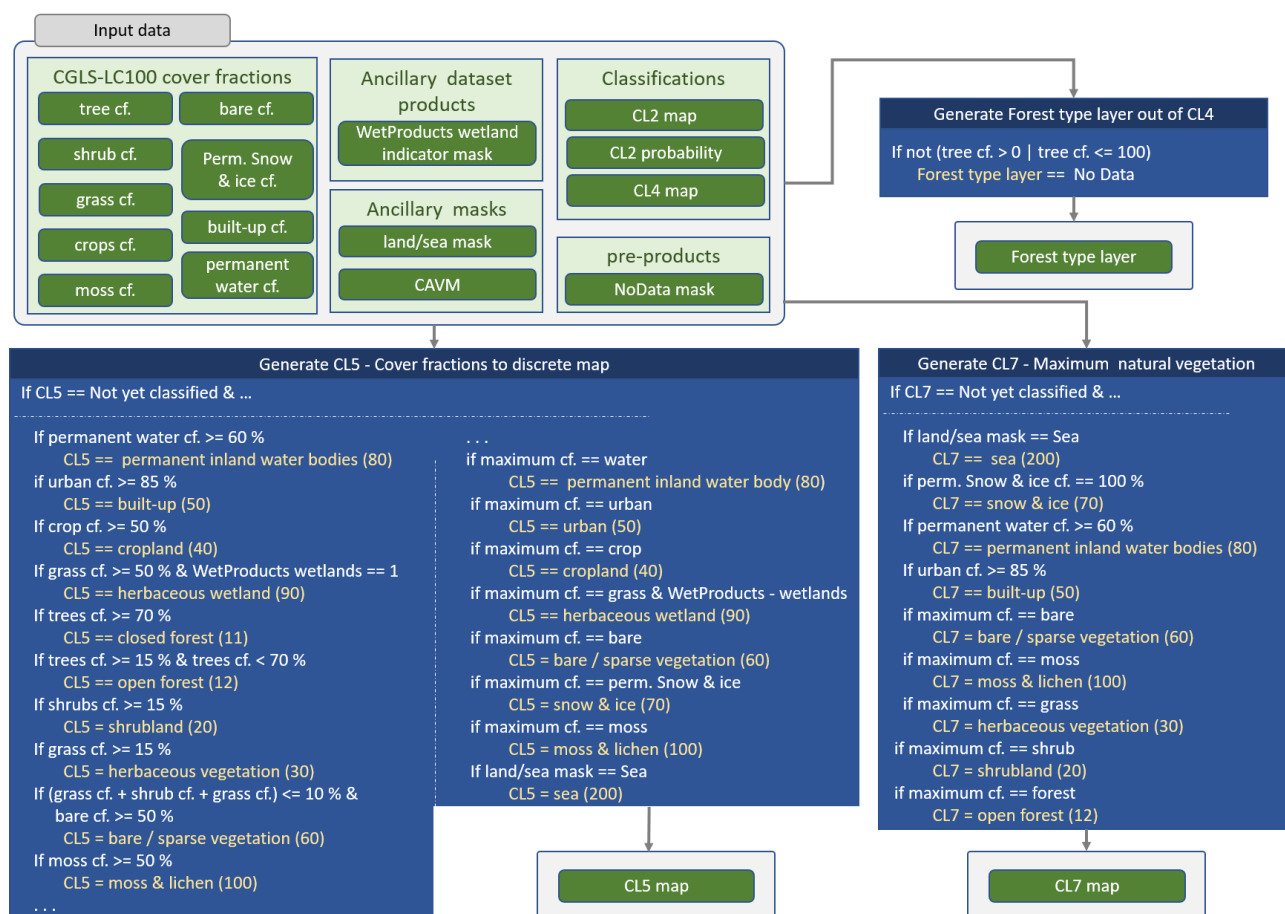


Figure 60: Flowchart of the classification post-processing to produce additional input layers for the CGLS-LC100 discrete map generation.

The next sub-section explains expert rules – the second processing step in the CGLS-LC100 discrete product generation – in more detail. There we combine existing knowledge represented by ancillary datasets with the classification and regression results (see section 3.12.2).

Output of the expert rule section are the CGLS-LC100 discrete map products consisting of 23-class discrete map, a classification probability layer serving as quality indicator, and the PROBA-V UTM 100 m forest type map.

3.12.2 Discrete map generation via expert rules

The final processing step in the generation of the CGLS-LC100 discrete map products is the combining of all classification scenarios and ancillary datasets into a final LC map via expert rules. 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 61 to Figure 65 illustrate the applied expert rules in detail.

Overall 23 discrete classes have to be compiled out of the input data. Table 7 gives an overview of the classes, class groups and the final assigned raster values in the CGLS-LC100 discrete map.

Table 7: The 23 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)	
		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)	
shrubs (20)			
herbaceous vegetation (30)			
cropland (40)			
build-up (50)			
moss & lichen (100)			
bare/sparse vegetation (60)			
water	permanent inland water bodies (80)		
	open sea (200)		
herbaceous wetland (90)			
permanent snow & ice (70)			
no satellite observations (0)			

Overall 16 input datasets which can be grouped into five input dataset groups are used in the expert rules. In detail (see also Figure 61):

- Two products of the 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 CGLS-LC100 cover fraction products (permanent snow&ice, built-up, tree, shrub, and permanent water) (see section 3.11.2);
- The PROBA-V UTM 100 m NoData mask generated in the pre-processing step of the CGLS-LC100 cover fraction products generation (see section 3.11.1);
- 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;
- 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.

Before the first expert rule (Figure 61) is applied on the input data, two output datasets are initialized with a “No Data” value – i.e. the CGLS-LC100 LC (set to -1) and CGLS-LC100 LC probabilities (set to 255). 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 61).

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 layers, i.e. the CGLS-LC100 permanent snow&ice cover fraction, CGLS-LC100 permanent water cover fraction and the CGLS-LC100 built-up cover fraction (Figure 62). 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 85 % 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 pixel come partly from external data and therefore a probability based on the classification training cannot be given.

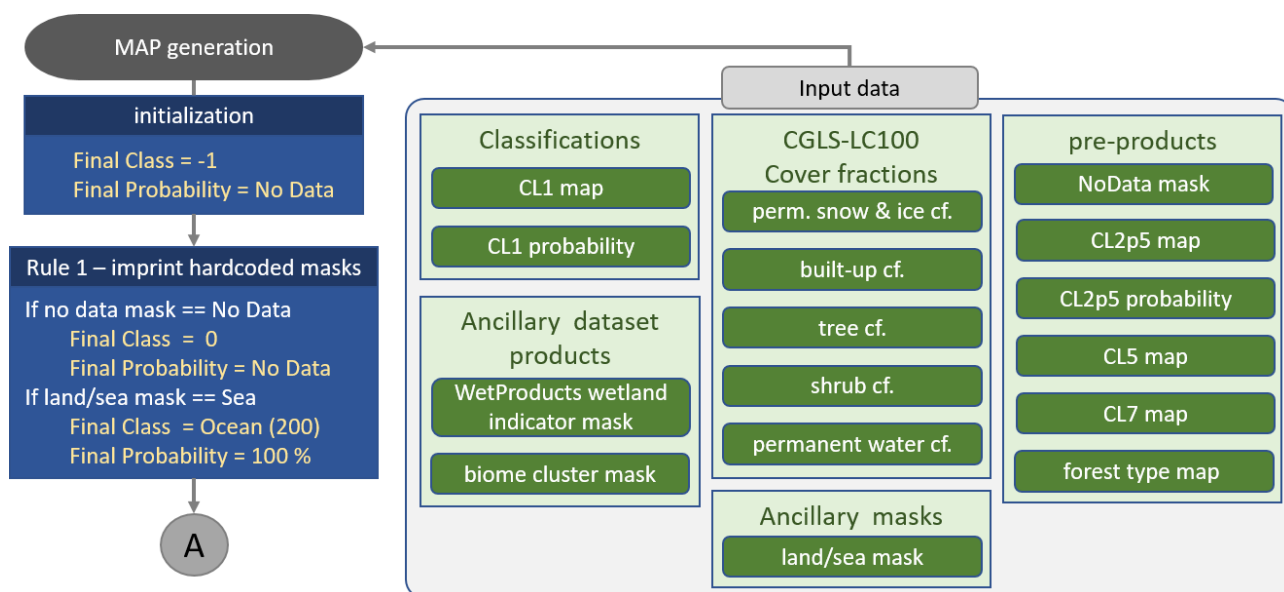


Figure 61: Initialization and Expert Rule block 1 of the CGLS-LC100 discrete map and map probability layer generation – imprinting hard masks.

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 62). 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 they are 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.

Subsequently, rule 4 resolves land pixels which are not yet classified in the final map (Figure 62). Most important in this filling step is that we have to guarantee that no 'permanent snow&ice' pixels from CL2p5 or CL5 are taken over 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 trigger 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.

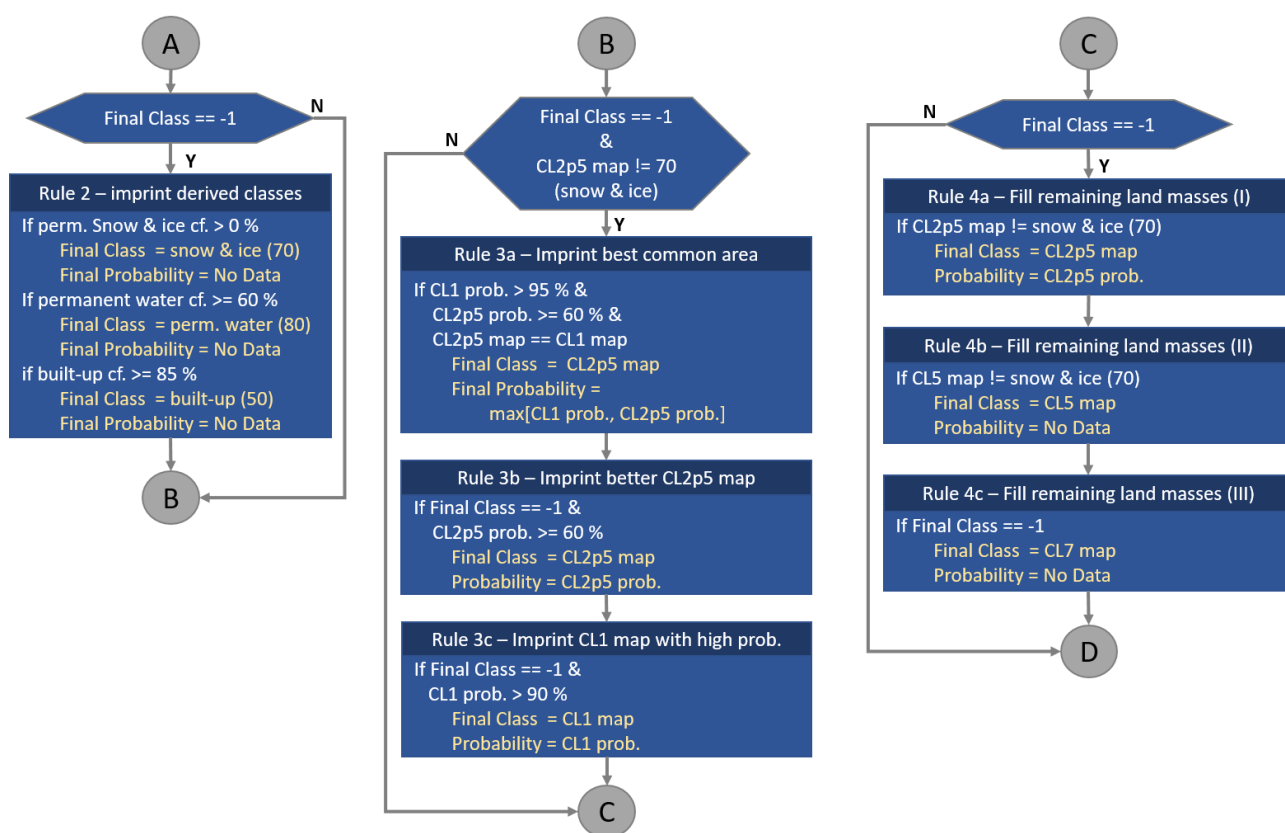


Figure 62: Expert Rule block 2 to 4 of the CGLS-LC100 discrete map and map probability layer generation – imprinting the derived cover fraction layers for classes, resolving classification scenarios, and filling of the remaining land pixels.

To minimize misclassifications between closed forest, open forest, shrub and herbaceous vegetation in the final discrete map, another set of expert rules is applied by rule 5 (Figure 63, Figure 64). These rules are continent as well as biome cluster specific. Figure 46 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'.

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 65).

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 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
 - 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 65).

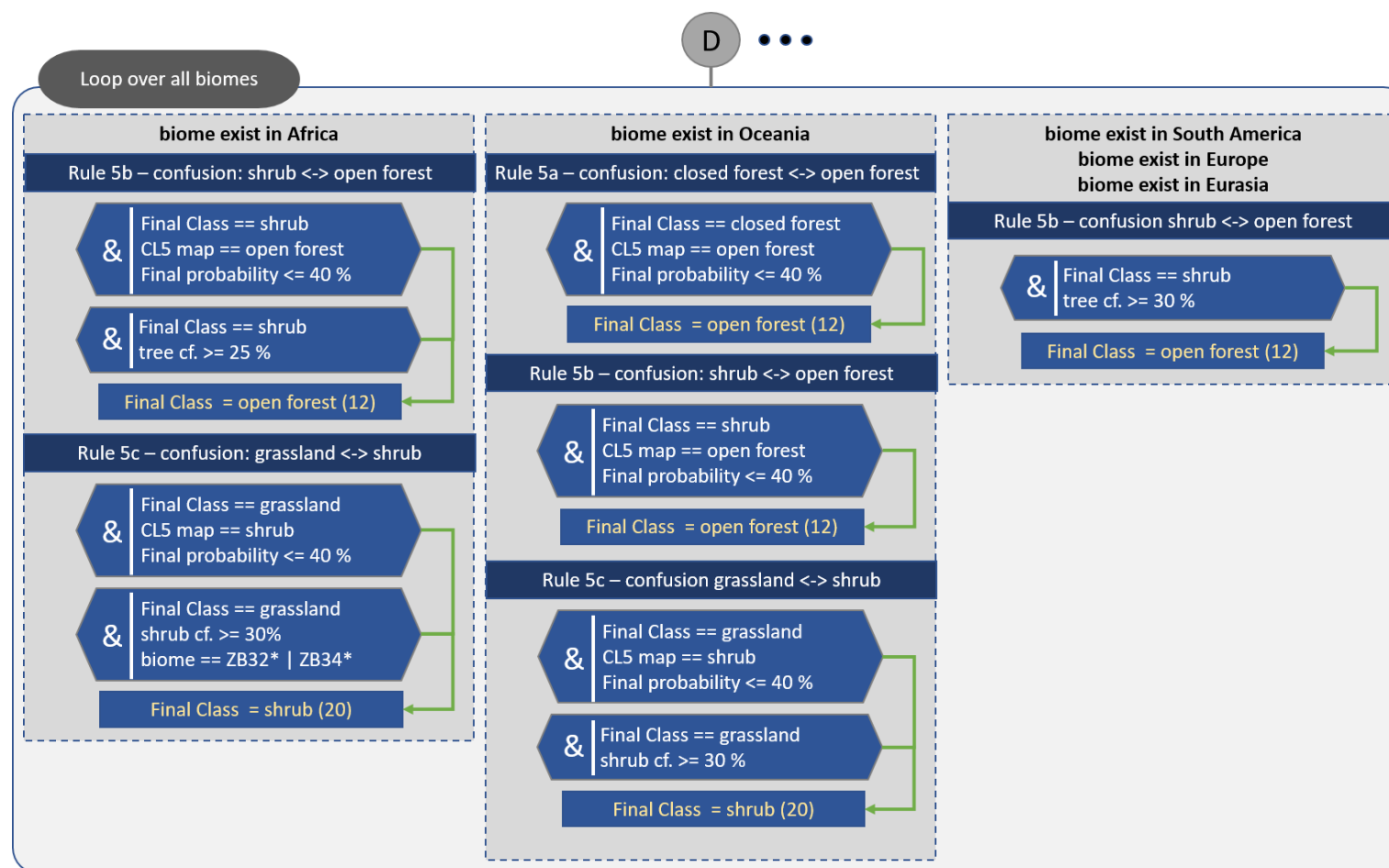


Figure 63: Expert Rule block 5 part A of the CGLS-LC100 discrete map and map probability layer generation – resolving confusion between closed forest, open forest, shrub and grassland on continental level. *See Figure 46 and Table 6 for biome cluster location and naming.

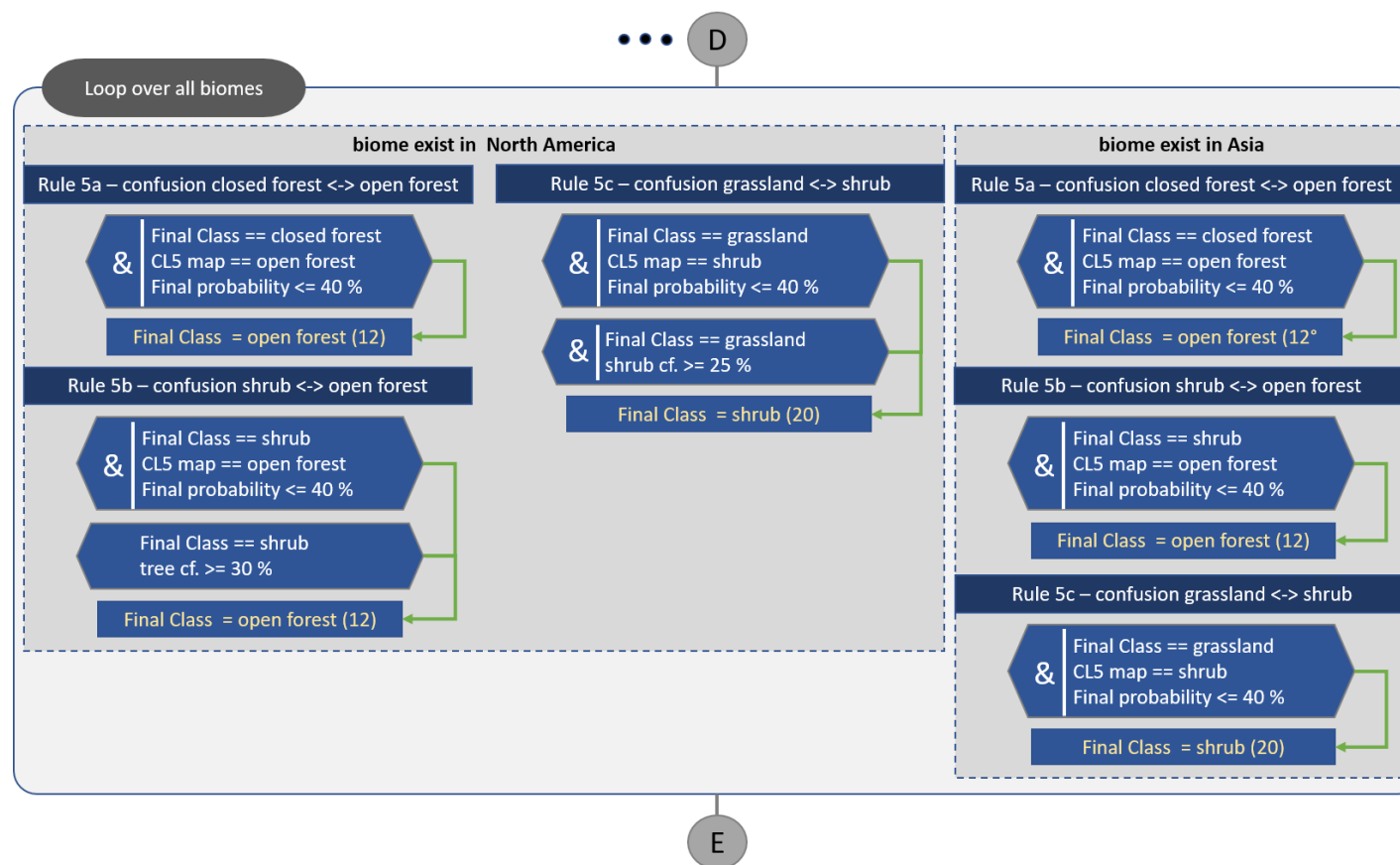


Figure 64: Expert Rule block 5 part B of the CGLS-LC100 discrete map and map probability layer generation – resolving confusion between closed forest, open forest, shrub and grassland on continental level.

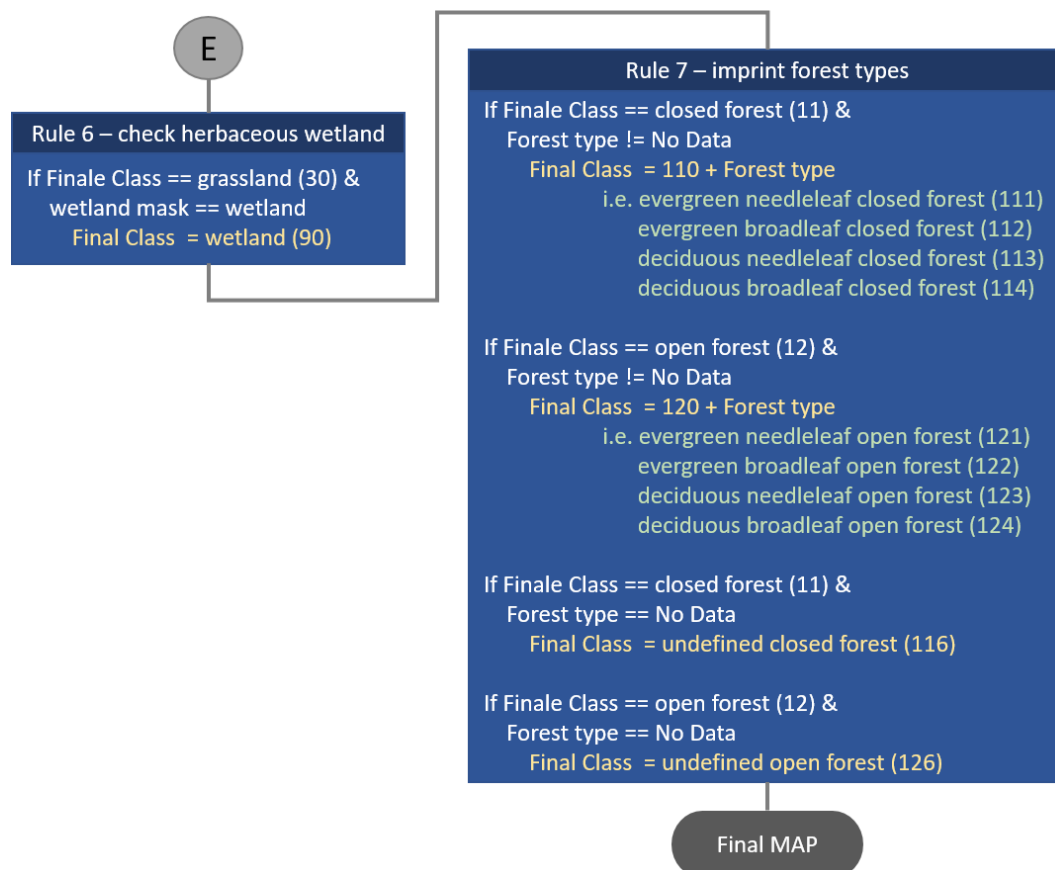


Figure 65: Expert Rule block 6 and 7 of the CGLS-LC100 discrete map and map probability layer generation – wetland class improvements and final separation of the open/close forest class by forest type.

3.13 CGLS-LC100 PRODUCTS

The final step in the 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 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-V2.0]).

Overall twenty layers are provided as the CGLS-LC100 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 bare/sparse vegetation regression standard deviation;
6. CGLS-LC100 cropland cover fraction layer;
7. CGLS-LC100 cropland regression standard deviation;
8. CGLS-LC100 herbaceous grassland cover fraction layer;
9. CGLS-LC100 herbaceous grassland regression standard deviation;
10. CGLS-LC100 moss & lichen cover fraction layer;
11. CGLS-LC100 moss & lichen regression standard deviation;
12. CGLS-LC100 shrubland cover fraction layer;
13. CGLS-LC100 shrubland regression standard deviation;
14. CGLS-LC100 tree cover fraction layer;
15. CGLS-LC100 tree regression standard deviation;
16. CGLS-LC100 permanent snow & ice cover fraction layer;
17. CGLS-LC100 built-up cover fraction layer;
18. CGLS-LC100 permanent water cover fraction layer;
19. CGLS-LC100 seasonal water cover fraction layer; and
20. CGLS-LC100 Data Density Indicator for PROBA-V UTM 100m input data.

Figure 66 shows an overview of the product on global scale, where Figure 67 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 67 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-V2.0]) and distribution of the product.

Figure 68 shows the 10 provided CGLS-LC100 cover fraction layers in a collage. The shown colours for the cover fractions are the ones integrated as RGB colour bars in the metadata of the cover fraction product layers.

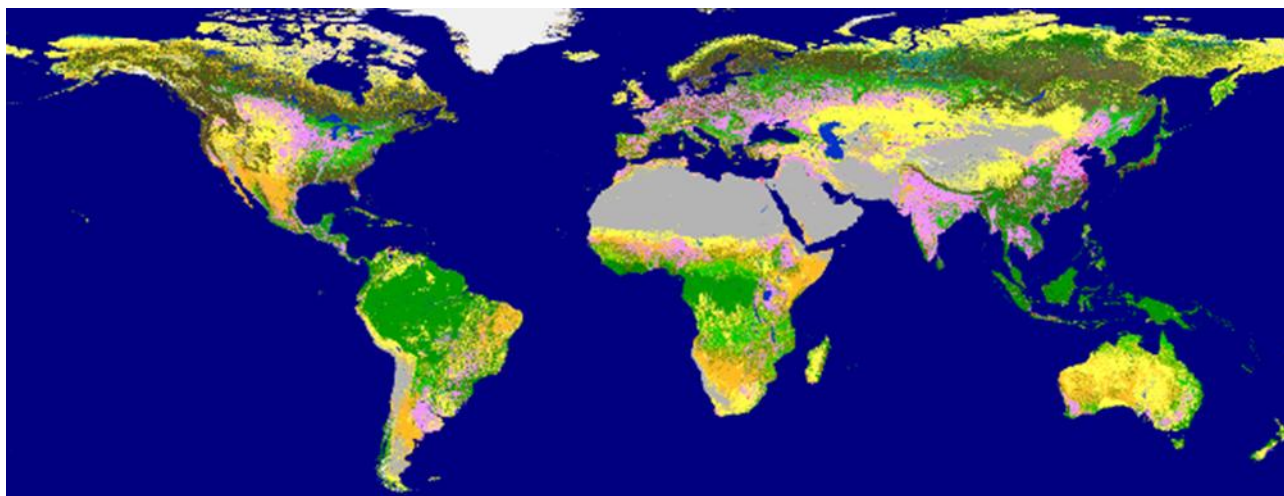


Figure 66: The CGLS-LC100 discrete map product for epoch 2015 with 23 discrete classes (detailed legend in Figure 67)

LEVEL 1						
value map	grouping	group code	R	G	B	class name
0		0	40	40	40	unknown (no input data available)
111	X	10	0	100	0	forest
113						
112						
114						
115						
116						
121						
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 2						
value map	grouping	group code	R	G	B	class name
0		0	40	40	40	unknown (no input data available)
111	X	11	0	120	0	closed forest
113						
112						
114						
115						
116	X	12	100	140	0	open forest
121						
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 67: 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.

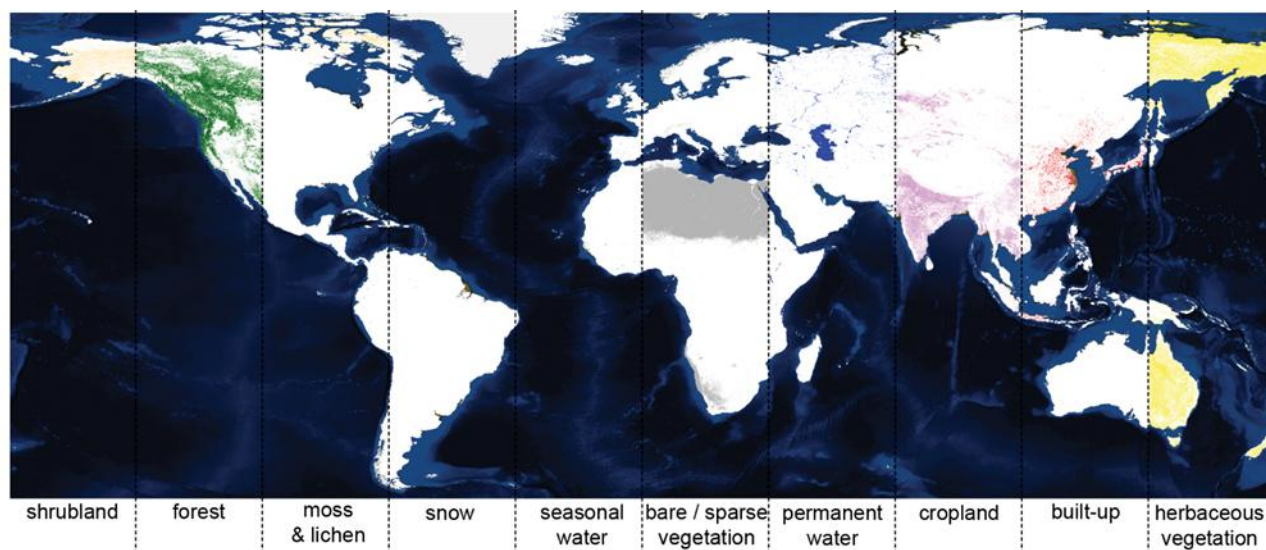


Figure 68: The cover fraction layers for the 9 base land cover classes and the seasonal inland water cover fraction of the CGLS-LC100 product for epoch 2015 (shown as a collage on global scale).

Finally, to identify all dataset versions of the yearly CGLS-LC100 product, we decided to assign unique Digital Object Identifier (DOI) to each yearly 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 (10.5281/zenodo.3243508) will resolve all the time into the newest version of the dataset where the version DOI can be used if a specific dataset wants to be accessed or cited. The DOI information for all dataset versions of the CGLS-LC100 product will be always updated in the Product User Manual [CGLOPS1_PUM_LC100m-V2.0]. Table 8 shows the provided information exemplary for Collection 1 and 2 for the epoch 2015.

Table 8: Overview of existing CGLS-LC100 products produces by algorithm version

Dataset name	Version 1	Version 2
Dataset version	V1.0.2	V2.0.2
Provided epochs	2015	2015
Used algorithm version for Dataset	V1.0	V2.2
Concept DOI	-	10.5281/zenodo.3243508
Version DOI	-	10.5281/zenodo.3243509
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. DOI:

		10.5281/zenodo.3243509
Direct access	-	http://doi.org/10.5281/zenodo.3243509
Data layers per epoch	8	20
File Size	6.6 GB	63.3 GB
Note	only Africa	global

4 LIMITATIONS

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

- Due to the used input data the product is limited to a latitudinal extent between 60° South and 78.25° North.
- 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 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, which could lead to an underestimate of built-up areas.
- 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, 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.
- Since the water body detection algorithm was adapted from Bertels et al. (2016), the limitations of this algorithm also have to be taken into account. Misclassifications of water bodies can happen in:
 - Dark areas caused by anthropogenic activity, e.g. heavy industry;
 - Dark areas caused by shadow, e.g. high buildings in large cities;
 - Anthropogenic structures with spectral signatures equal to Water Bodies (WB), e.g. some agricultural fields, built-up areas;
 - Natural surfaces with spectral signatures very close to WBs, e.g. volcanic soils;

- Areas where the spectral properties of WB fall outside the defined thresholds used in the water detection algorithm (empirical defined thresholds).
- 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 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.
- Since the PROBA-V time series is limited at the moment (2014 to 2019), misclassifications can happen especially in temporary ultra-dry and temporary ultra-wet areas. This is due to the fact that some regional effects with a low frequency (e.g. El Nino and La Nina) could not correctly captured with the training data reference base of 2015.
- 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.

5 RISK OF FAILURE AND MITIGATION MEASURES

In case the quality of the PROBA-V sensor degrades, the spectral response correction needs to be adapted and the defined thresholds for PROBA-V 100 m Water Body detection algorithm need to be updated if necessary. Moreover, the algorithm has to be adapted in case for a reference year the stated input data (PROBA-V 100 m EO sensor data or ancillary data sets) is not available (see section 3.2) and has to be replaced by EO data from other sensors such as Landsat, Sentinel-2, etc.. In such cases, the metric extraction procedure and also the expert rules have to be adapted to the new input data.

The PROBA-V mission will end in April 2020, and hence no input data will be available anymore. Continuation will be guaranteed using the Sentinel-2 dataset and using Sentinel-1 for gap filling. A preliminary test over a region in West-Africa has already shown that such continuation is possible, and a larger continental test is ongoing.

6 REFERENCES

- Aimonen, P. (2011). Basic concept of Kalman filtering.svg. Image published under the Creative Commons CC0 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-treeline-map>.
- 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](https://doi.org/10.1126/sciadv.1602244).
- 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](https://doi.org/10.1111/j.1365-2745.2006.01104.x).
- 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.
- 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/rs8121010](https://doi.org/10.3390/rs8121010).
- 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](https://doi.org/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-landcover-cci.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>.
- 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](https://doi.org/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., Pinte, 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., Vogt, 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](https://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., Tsensibazar, 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](https://doi.org/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_20151209T095117_V20150622T000000_21000101T000000_B00.kml. Retrieved: January 15, 2018.
- 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](https://doi.org/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.
- 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](https://doi.org/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](https://doi.org/10.1016/S0034-4257(02)00096-2).
- 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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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/jss.v036.i11](https://doi.org/10.18637/jss.v036.i11).
- Lesiv, M., Fritz, S., and See, L. (2017). A New Open Reference Global Dataset for Land Cover Mapping at a 100m 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/j.jesp.2013.03.013](https://doi.org/10.1016/j.jesp.2013.03.013).
- 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 preparation). Fractional land cover classification method assessment using PROBA-V satellite data. Remote Sensing of Environment.
- NASA Earth Observatory (2019): weblog „Cloudy Earth”. Retrieved: October 4th, 2019 from webpage: <https://earthobservatory.nasa.gov/images/85843/cloudy-earth>.
- NASA JPL. (2013). NASA Shuttle Radar Topography Mission Global 1 arc second [Data set]. NASA LP DAAC. DOI: [10.5067/MEaSURES/SRTM/SRTMGL1.003](https://doi.org/10.5067/MEaSURES/SRTM/SRTMGL1.003).
- NASA/METI/AIST/Japan Spacesystems, and U.S./Japan ASTER Science Team (2009). ASTER Global Digital Elevation Model [Data set]. NASA EOSDIS Land Processes DAAC. DOI: [10.5067/ASTER/ASTGTM.002](https://doi.org/10.5067/ASTER/ASTGTM.002).
- 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: Instituto 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](https://doi.org/10.1080/01431161.2012.674230).
- 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](https://doi.org/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](https://doi.org/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](https://doi.org/10.3189/2014JoG13J176).
- Proud, S. R., Fensholt, R., Rasmussen, M. O., and Sandholt, I. (2010b). A comparison of the effectiveness of 6S and SMAC in correcting for atmospheric interference of Meteosat Second Generation images, *J. Geophys. Res.*, 115, D17209, DOI: [10.1029/2009JD013693](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/10.1080/01431161.2014.883094).
- Tsendbazar, N. E., de Bruin, S., Fritz, S., & Herold, M. (2015). Spatial Accuracy Assessment and Integration of Global Land Cover Datasets. *Remote Sensing*, 7(12), 15804-15821. DOI: [10.3390/rs71215804](https://doi.org/10.3390/rs71215804).

- 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](https://doi.org/10.1016/0034-4257(79)90013-0).
- 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.
- 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](https://doi.org/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
- 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](https://doi.org/10.1029/96JB00104).