

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 1

Issue I1.00

Organization name of lead contractor for this deliverable:

Book Captain: 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 25.09.2017
Approval:	Roselyne Lacaze	Sign 	Date 10.10.2017
Endorsement:	Michael Cherlet	Sign	Date
Distribution:	Public		

Change Record

Issue/Rev	Date	Page(s)	Description of Change	Release
	25.09.2017	All	Initial Version	11.00

TABLE OF CONTENTS

Executive Summary	15
1 Background of the document	16
1.1 Scope and Objectives	16
1.2 Content of the document	16
1.3 Related documents	16
1.3.1 Applicable documents	16
1.3.2 Input.....	16
1.3.3 Output.....	17
1.3.4 External documents	17
2 Review of Users Requirements	18
3 Methodology Description	23
3.1 Overview	23
3.2 Input Data	24
3.2.1 PROBA-V TOC daily synthesis surface reflectances for 100 m and 300 m	25
3.2.2 PROBA-V land/sea mask	26
3.2.3 FAO global ecological zones dataset.....	26
3.2.4 Shoreline vector layer for Africa	27
3.2.5 DLR's Global Urban Footprint plus (GUF+) layer.....	28
3.2.6 JRC's Global Human Settlement (GHS) layer	29
3.2.7 JRC's Global Surface Water (GSW) product	30
3.2.8 NASA's Shuttle Radar Topography Mission Global 1 arc second dataset	30
3.3 Data Cleaning & Compositing	31
3.3.1 Overview	31
3.3.2 The HANTS and madHANTS algorithms	32
3.3.3 Median Composites (MC) generation.....	34
3.3.4 Long-term 5-daily Median Composite (HMC5) generation	37
3.4 Data Fusion	38
3.4.1 Overview	38
3.4.2 Data Fusion pre-processing	39
3.4.3 Data Fusion using the Kalman-Filtering approach	40
3.4.4 Data Fusion post-processing.....	42
3.5 Metrics Generation	43
3.5.1 Overview	43
3.5.2 Pre-Processing	45

3.5.3	Metrics extraction.....	54
3.5.4	Post-Processing.....	58
3.6	Ancillary Dataset Products.....	59
3.6.1	Ecozone Buffering product	59
3.6.2	Shoreline product	59
3.6.3	Urban product generation	59
3.6.4	Water Products Generation.....	60
3.7	Training Data Generation	62
3.8	Classification / Regression	64
3.8.1	Overview	64
3.8.2	The Random Forest Approach	64
3.8.3	Training data and classifier/regressor optimization	65
3.8.4	Scenario-based Classification.....	67
3.8.5	Scenario-based Regression	68
3.9	Cover Fraction Layers Generation	69
3.9.1	Regression post-processing	69
3.9.2	Metadata	69
3.10	Land Cover Map Generation	70
3.10.1	Overview.....	70
3.10.2	Assembling and generation of the input datasets for the expert rules.....	71
3.10.3	Map Generation	73
3.10.4	Metadata	81
4	Limitations	84
5	Risk of failure and Mitigation measures.....	85
6	References	86

List of Figures

Figure 1: Workflow diagram for the CGLS Dynamic Land Cover 100m product for Africa 2015	23
Figure 2: PROBA-V land/sea mask for the African continent.....	26
Figure 3: FAO global ecological zones dataset for the African continent. Image edited – original from http://foris.fao.org/static/data/fra2010/ecozones2010.jpg (FAO, 2012).....	27
Figure 4: 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.....	28
Figure 5: The DLR Global Urban Footprint plus layer (Marconcini et al., 2017b). Zoom in to the Street of Gibraltar showing the urban areas in black. Note: in red the USGS shoreline vector layer is shown.	29
Figure 6: The JRC Global Human Settlement layer (Pesaresi et al., 2015). Zoom in to the Street of Gibraltar showing the urban areas in black. Note: in red the USGS shoreline vector layer is shown.	29
Figure 7: 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 2014/2015 (right). Note: in red the USGS shoreline vector layer is shown in the maximum water extent layer (left image).....	30
Figure 8: 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.	31
Figure 9: 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.	32
Figure 10: 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.....	34
Figure 11: Example showing the temporal outlier detection using the madHANTS algorithm on the blue and SWIR reflectances of a pixel a location 9.459° Lon, 6.562° lat. Note: red curve shows the harmonized time series; blue shows the valid pixel values; green shows flagged outliers cumulative detected in the blue or SWIR reflectance band.	35
Figure 12: Example for data cleaning process of PROBA-V 100m image from 2016-03-06. Image is shown as false color composite (RGB = SWIR, NIR, blue) for a sample area in tile X18Y06 (Nigeria) – left) raw image, middle) status masked cleaned image (quality flagged areas are	

shown in red), right) madHANTS cleaned image (additional pixels which are flagged as outliers or clouds are shown in blue)..... 36

Figure 13: Number of cloud free data observations for PROBA-V 100 m and PROBA-V 300 m in the reference year 2015 for Africa. Note: Blue circled area indicates area with lowest observation density in Africa. 38

Figure 14: 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 observations. 39

Figure 15: Example for data fusion pre-processing results for PROBA-V 300m MC5 time series for pixel location 9.459° lon, 6.562° lat. Top: continuous gap length in the 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 5-daily median composite time series, green line shows original 10-daily median composite time series, red line shown the long term harmonized time series for the full PROBA-V archive), bottom: 300m time series after pre-processing (blue line shows the original 5-daily median composite time series, orange line shows the final pre-processed time series which will be used for the data fusion). 40

Figure 16: Principle of the Kalman-filtering approach for EO data fusion using the pktools algorithm. Adapted from Kempeneers et al. (2016)..... 41

Figure 17: top) PROBA-V 100 m MC5 pre-processed image (areas with missing data is shown in white), bottom-left) PROBA-V 100 m MC5 image after data fusion approach, bottom-right) zoom in to full PROBA-V 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 over a region located in Nigeria. 43

Figure 18: General overview of the metrics generation section in the CGLS LC100 product workflow. Note: datasets marked in red are re-used in sub-steps of a processing step. Numbers in the upper left corner of data container indicate the number of layers. 44

Figure 19: Overview of the processing sub-steps within the pre-processing step of the metrics generation section in the CGLS LC100 product workflow. Note: Numbers in the upper left corner of data container indicate the number of layers. 45

Figure 20: The HSV color space and the formulas for transforming the RGB color space into the HSV color space. Note: V=Value, S= Saturation, H=Hue, R=red, G=green, B=blue. 48

Figure 21: The Water Bodies Potential Mask generation algorithm – filtering and expanding the detected lowest points. Expanding the initially detected lowest point by systematically raising an imaginary water level in steps of 1 m. Note: the corresponding 30 m spatial resolution pixels are indicated by the dots at the bottom. 49

Figure 22: Example for the phenology product processing. Note: PROBA-V 100m NDVI values in red, smoothed curve in green, identified local maxima are shown as blue dots. Blue lines

indicate important phenological metrics (SoS = start of season, MoS = mid of season, EoS = end of season) and vertical black lines indicate the reference year)..... 51

Figure 23: Phenology for Africa 2015 in PROBA-V 100m spatial resolution. (top-left) Start of season 1, (top-right) end of season 1, (bottom-left) start of season 2, (bottom-right) end of season 2. Note: the phenology dates are shown in decades starting with the year before the reference year (first decade of reference year = 37). 52

Figure 24: Quality masks. (top-left) tGAPmask showing pixels with high change of uncertain filled data during the data fusion step, (top-right) number of vegetation seasons in reference year, (bottom-left) SEASONALITYmask indicating if a pixel has a seasonality, (bottom-right) length of the combined vegetation seasons in decades. 54

Figure 25: 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. 56

Figure 26: Example for splitting the time series profiles in reference year, vegetation season and off-vegetation season. Blue line show the identified reference year in the time series profile and red lines the time steps to separate vegetation from off-vegetation season. Note: shows an example with only one vegetation season in the reference year..... 57

Figure 27: Examples of the 392 derived metrics for tile X18Y06 (Nigeria). a) texture metric (the lighter the color 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 color the higher is the SIPI change within the vegetation season), c) sum of the Enhanced Vegetation Index for the reference year (the redder the colour the more vegetation). 58

Figure 28: General overview of the Water Bodies Detection Algorithm. 60

Figure 29: Decision tree of the water body detection algorithm. 62

Figure 30: Training points with discrete land cover class for the CGLS LC100 product workflow. . 63

Figure 31: The Random Forest classifier principle. Image by Niraj (2016). 65

Figure 32: 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 with 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..... 66

Figure 33: The cover fraction layers for the forest, shrub, herbaceous vegetation and bare land cover classes of the CGLS LC100 product for Africa 2015 (shown at continental scale). 70

Figure 34: General overview of the Land Cover map generation section in the CGLS LC100 product workflow. 71

Figure 35: Expert rule I of the CGLS LC100 discrete map generation process..... 74

Figure 36: Expert rules II and III of the CGLS LC100 discrete map generation process. 75

Figure 37: Expert rules IV of the CGLS LC100 discrete map generation process..... 76

Figure 38: Expert rule Va of the CGLS LC100 discrete map generation process.	77
Figure 39: Expert rules Vb and Vc of the CGLS LC100 discrete map generation process.....	78
Figure 40: Expert rules Vd of the CGLS LC100 discrete map generation process.....	78
Figure 41: Expert rules VIa – VI d of the CGLS LC100 discrete map generation process.	79
Figure 42: Expert rules VII and VII of the CGLS LC100 discrete map generation process.	80
Figure 43: Expert rules IX and X of the CGLS LC100 discrete map generation process.	81
Figure 44: Legend for the 18 discrete classes of the CGLS LC100 discrete map for Africa 2015. Note: the number in brackets represents the numerical code for a land cover class.	82
Figure 45: The CGLS LC100 discrete map for Africa 2015 with 18 discrete classes (shown at continental scale).....	83

List of Tables

Table 1: Summary of stakeholder requirements.....	19
Table 2: List of land cover classes requested by users	21
Table 3: Spectral characteristics of the PROBA-V bands	25
Table 4: Status Map bit mapping of the PROBA-V data	25
Table 5: The 18 discrete classes of the CGLS LC100 discrete product. Note: final classes are shown in bold.....	73

List of Acronyms

Acronym	Meaning
ASTER	Advanced Spaceborne Thermal Emission and Reflection Radiometer
ATBD	Algorithm Theoretical Basis Document
CF	Climate & Forecast conventions
CGLS	Copernicus Global Land service
DEM	Digital Elevation Model
DLR	Deutschen Zentrums für Luft- und Raumfahrt
EO	Earth Observation
EoS	End of Season
EPSG	European Petroleum Survey Group Geodesy
EVI	Enhanced Vegetation Index
EZ	Ecological Zone
FAO	Food and agriculture organization of the united nation
FAPAR	Fraction of Absorbed Photosynthetically Active Radiation
FCC	False Color Composite
FFT	Fast Fourier Transformation
GEZ	Global Ecological Zone
GHS	Global Human Settlement
GLIMPSE	GLobal IMage Processing SoftwarE
GSD	Ground Sampling Distance
GSW	Global Surface Water
GUF	Global Urban Footprint
HANTS	Harmonic Analysis of Time Series
HCM5	Harmonized 5-daily median composite
HSV	Hue Saturation Value color system
HUE	Chromaticity
IIASA	International Institute for Applied Systems Analysis
JRC	Joint Research Center
LAI	Leaf Area Index
LC	Land Cover
LC100	Land Cover map at 100m resolution
LCCS	Land Cover Classification System
LocDG	Longest concurrent Data Gap
LoVS	Length of Vegetation Season
MAD	Median Absolute Deviation
madHANTS	Median Absolute Deviations of HANTS
MC10	10-daily median composite
MC5	5-daily median composite
MESA	Monitoring Environment and Security in Africa
MODIS	Moderate Resolution Imaging Spectroradiometer
MOS	Mid of Season

MWEM	Maximum Water Extent Mask
NASA	National Aeronautics and Space Administration
NBR	Normalized Burn Ratio
NDVI	Normalized Difference Vegetation Index
NIR	Near Infra Red reflectance
NIRv	Near-Infrared reflectance of vegetation
NoS	Number of Seasons
NOVO	Number of valid observations
NPP	Net Primary production
QGIS	Quantum Geographic Information System
PFTs	Plant Function Types
PHENODef	Phenology Defined
PROBA-V	Vegetation instrument on board of PROBA satellite
PUM	Product User Manual
REDD+	Reducing Emissions from Deforestation and forest Degradation
RF	Random Forest classifier
RGB	Red-Green-Blue colour space
RMSE	Root Mean Square Error
RS	Remote Sensing
S1	1-daily synthesis
SIPI	Structure Intensive Pigment Index
SM	Status Map
SMAC	Simplified Model for Atmospheric Correction
SOS	Start of Season
SPIRITS	Software for the Processing and Interpretation of Remotely sensed Image Time Series
SRTM	Shuttle Radar Topography Mission
SRTMGL1	Shuttle Radar Topography Mission Global 1 arc second
STS-99	Space Transportation System #99
SVP	Service Validation Plan
SWIR	Short Wave Infra Red reflectance
tGAPmask	Time series GAP mask
TOC	Top Of Canopy
UN	United Nations
URD	User Requirements Document
USGS	United States Geological Survey
VITO	Flemish Institute for Technological Research
VRT	Virtual Raster Translator
WB	Water Bodies
WBPM	Water Body Potential Mask
WGS	World Geodetic System
WUR	Wageningen University & Research

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 reliable, automatic and 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 PAR absorbed by the vegetation (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.

The Dynamic Land Cover map at 100 m resolution is a new product in the portfolio of the CGLS and targets to deliver a yearly global land cover map at 100 m spatial resolution. 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 with 18 classes. Next to these classes, the product also provides a set of four vegetation continuous field layers that provide proportional estimates for vegetation cover for the land cover types forest, herbaceous vegetation, shrub and bare ground. 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, crop monitoring, biodiversity and conservation, monitoring environment and security in Africa, climate modelling, etc.)

This first Land Cover map (V1.0) is provided for the 2015 reference year over the African continent, derived from the PROBA-V 100 m time-series, a database of high quality land cover training sites and several ancillary datasets.

This Algorithm Theoretical Based Document (ATBD) describes the methods used for classifying the land cover on PROBA-V data at 100 m resolution on continental 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 1 provided in the Copernicus Global Land Service. It details the methodology to be applied on the PROBA-V data, together with a description of the limitations.

A thorough validation of the products is performed by a full quality assessment exercise [CGLOPS1_VR_LC100m-V1] 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 of the Copernicus Global Land Service.
CGLOPS1_SVP	Service Validation Plan of the Copernicus Global Land Service.
CGLOPS1_URD_LC100m	User Requirements Document of the Dynamic land

cover 100m product

CGLOPS1_TrainingDataReport_LC100m	Report describing the training dataset used for Dynamic Land Cover 100m product
GIOGL1_ATBD_WB1km-PROBAV-V2	Algorithm Theoretical Basis Document of Collection 1km water body detection version 2 from PROBA-V

1.3.3 Output

Document ID	Descriptor
CGLOPS1_VR_LC100m-V1	Validation Report describing the results of the scientific quality assessment of the Dynamic land cover 100m product
CGLOPS1_PUM_LC100m-V1	Product User Manual Document of Dynamic land cover 100m product

1.3.4 External documents

PROBA-V	http://proba-v.vgt.vito.be/
PROBA-V User Manual	User Guide of the PROBA-V data, available on http://www.vito-eodata.be/PDF/image/PROBAV-Products_User_Manual.pdf
SPOT-VGT Collection 3 User Manual	User Guide of the SPOT VEGETATION Collection 3 data, available on http://www.vgt.vito.be/pages/SPOT_VGT_PUM_v1.0.pdf

2 REVIEW OF USERS REQUIREMENTS

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

- **Definition:** Dynamic global land cover products at 300m and/or 100m resolution using 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: WGS84
 - pixel size: 1/112° - accuracy: min 10 digits
 - coordinate position: pixel centre
 - 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 LC100 map. More details can be found in [CGLOPS1_URD_LC100m].

Table 1 provides a summary of the major requirements from the stakeholders, while Table 2 shows an overview of the requested classes to be covered by the mapping.

Table 1: Summary of stakeholder requirements

Land cover change information	
Forest modelling/REDD+	Forest change information is needed for identifying areas of tree loss and gain.
Crop monitoring	Static land cover maps of a high accuracy are of high priority
Biodiversity and conservation	Reliable information on the extent, location and change of habitats is needed for integration in a change alert system.
Monitoring Environment and Security in Africa (MESA)	Depending on application, both types of maps are needed: change maps and static land cover map.
Climate modelling	Priority is given to stable land cover maps. Change maps are desirable as well, accompanied with a measure of reliability quantifying their statistical accuracy.
Resolution	
Forest modelling/REDD+	1-20 m – higher is better
Crop monitoring	100 m resolution is satisfactory for cropland mask
Biodiversity and conservation	1-20 m – higher is better
Monitoring Environment and Security in Africa	100 m is acceptable
Climate modelling	100 m resolution is very good to produce better PFT fraction estimations at coarser scales
Accuracy/error information	
All users	Overall thematic accuracy > 80% and should be based on stratified random sampling design, with a number of sample points per land cover class calculated (Wagner et al, 2015)
	Accuracy estimates should be not only overall, but also class specific.
	Accuracy has to be calculated at different geographical levels, e.g. regional, national, continental, global
	Minimum error has to be less than 15% or 20% at class level and at regional or national level (large country).
	Qualifying the error in a spatial manner is important, e.g. using covariance matrices, (Tsendbazar et al, 2015).
Thematic requirements	
Forest modelling/REDD+	Mapping human impact on forest: primary and secondary forests, intactness, core/edge, managed/unmanaged, as well as forest parameters such as tree height and carbon stock/biomass, NPP,

	etc.
Crop monitoring	More classes on managed land/cultivated areas: irrigation, big/small farming, permanent crops, fallow, grassland (artificial, natural), some plantations
Biodiversity and conservation	Savannah, wooded shrubs, wetlands, natural vs man-made; Abandoned land; Infrastructure such as mines, roads, built infrastructure, including settlements, roads, electric lighting, canals and water control structures.
Monitoring Environment and Security in Africa	Forestry, Inland Waters, Pastoral Resources, Land Cover Change Assessment (including urbanization), Land Degradation, Natural Habitat Conservation Assessment, Monitoring and Assessment of Environmental Impacts of Mineral Resources Exploitation
Climate modelling from vegetation	Classes related to PFTs: trees vs shrubs vs grasses, C3 crops vs C4 crops vs irrigated crops; leaf types; managed vs natural classes, change vs phenology, etc.
All users	More land cover classes of Level 2. More details in a section below. UN LCCS should be used by default.
Projection	
All users	Commonly used projection (e.g. WGS 1984, EPSG: 4326), eventually easy to convert.
Access	
All users	Easy and open access, options for countries with slow connections, options to choose between global and regional products
Other requirements	
All users	Yearly updates and consistency among consecutive products.
	Continuity on nomenclature of the land cover products. Reprocess operations should be performed whenever the nomenclature evolves.
	A clear distinction should be made between “date of issue” and the “data used” (period).

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
10		A12A3A20B2	Forest/tree cover	X		X	X	X
	11	A12A3A20B2D2 E1	Evergreen Needleleaf forest	X			X	X
	12	A12A3A20B2D1 E1	Evergreen Broadleaf forest	X			X	X
	13	A12A3A20B2D2 E2	Deciduous Needleleaf forest	X			X	X
	14	A12A3A20B2D1 E2	Deciduous Broadleaf forest	X			X	X
	15	A12A3A20B2D1 D2	Mixed forest	X		X		
	16	A12A3A10B2X XXX (assuming that an intact forest is a very dense forest)	Intact forest	X		X		X
	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
	21	A12A4A20B(B9)XXE1	Evergreen shrubs			X		
	22	A12A4A20B3(B9)XXE2	Deciduous shrubs			X		
30		A12A2(A6)A20B4	Herbaceous vegetation			X	X	X
		A12A6A10 // A11A1A11B4X XXXXXF2F4F7 G4-F8	Pasture/managed grassland					X
		A122(A6)A10	Natural grassland			X		X
		A12A2	Grass types for Western Africa			X		
		A12A3A11B2X XXXXXF2F4F7 G4-A12; A12A3A11B2-A13; A12A1A11	Savannas			X		
40		A11A3	Cultivated and managed vegetation/agriculture		X	X	X	X
	41	A11A3XXXXXX	Irrigated cropland		X			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
		D3(D9)						
	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 winter/summer crops		X			
		A11A2	Multiple crop cycles		X			
50		B15A1	Urban/built up			X	X	X
60		B16A1(A2)	Bare/sparse vegetation				X	X
70		B28A2(A3)	Snow and Ice				X	X
80		B28A1	Open water				X	X
		A24A1(A2/A3/A4)	Wetland			X	X	X
		A24A3	Mangroves	X		X		

Green colour indicate classes that can be included into the legend without any risk; yellow shows classes that will potentially be integrated into the legend dependent on the results of the risk assessment after some tests; and red is for classes that are to be analysed in a second evolution of the map, as the risk is too high and require additional effort.

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 cleaning and outlier detection techniques,
2. Applying data fusion techniques at multiple levels,
3. Supervised classification through collecting reference data, including crowdsourcing techniques,
4. Including established third party datasets via expert rules.

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

1. data cleaning & compositing,
2. data fusion,
3. metrics generation,
4. training data generation,
5. ancillary datasets products,
6. classification / regression,
7. cover fraction layers generation,
8. land cover map generation.

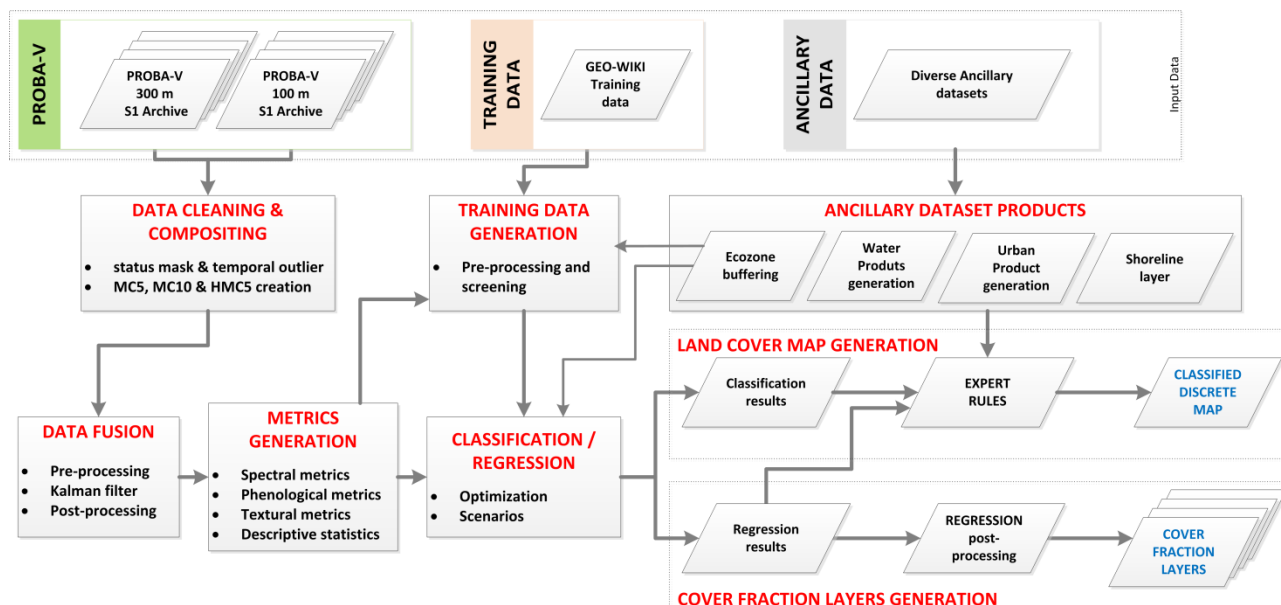


Figure 1: Workflow diagram for the CGLS Dynamic Land Cover 100m product for Africa 2015

To generate the product, 5-daily PROBA-V multi-spectral image data with a ground sampling distance (GSD) of ~ 0.001 degree (~ 100 m) is used as primary earth observation (EO) data, and PROBA-V daily multi-spectral image data with a GSD of ~ 0.003 degree (~ 300 m) secondarily. Next to a status mask cleaning using the internal quality flags of the PROBA-V EO data, a temporal cloud and outlier filter built on a Fourier transformation is applied to clean the data. Next, the 5-daily PROBA-V 100 m and daily 300 m datasets are fused using a Kalman filtering approach. The Kalman-filled 100 m data set is then automatically checked for consistency before extracting several metrics. Therefore, a harmonic model is fitted through each of the reflectance bands of the time series data as well as each of the additional derived vegetation indices for each time series step. 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 are extracted for the reference year as well as for the vegetation season and off-season within that reference year using phenological parameters (e.g. start- and end of season) extracted from the harmonic model itself. Overall, 392 metrics are extracted from the PROBA-V EO data.

The training data 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 in PROBA-V 100 m resolution. In the classification preparation, the metrics of the training points are analysed for intra- and inter- specific outliers, as well as screened for the best metrics combinations to run an optimized classification. The optimized training data is then used in a supervised classification using Random Forest techniques.

Finally, we build upon the success of previous global mapping efforts and/or other ancillary datasets. Therefore, the external datasets are resampled to PROBA-V 100 m spatial resolution and 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 and 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 vegetation continuous fields that provide proportional estimates for vegetation cover for trees, herbaceous vegetation, shrub and bare ground. The input are the cover fractions collected for all training points which are used in a Random Forest regression.

3.2 INPUT DATA

Eight different sets of input data are used in the CGLS LC100 product workflow, i.e. PROBA-V S1-TOC reflectance data for 100 m and 300 m, PROBA-V 100 m land/sea mask, FAO global ecological zones dataset, shoreline vector layer for Africa, DLR's Global Urban Footprint Plus mask, JRC's Global Human Settlement mask, JRC's Global Surface Water product, and NASA's Shuttle Radar Topography Mission dataset.

3.2.1 PROBA-V TOC daily synthesis surface reflectances for 100 m and 300 m

Global PROBA-V level 3 Top Of Canopy daily synthesis (S1-TOC) products are used as main input in the CGLS LC100 product workflow. More details about the generation of the geometrically and atmospherically corrected S1-TOC product can be found in the [PROBA-V User Manual].

The PROBA-V Level 3 Top of Canopy daily synthesis products used in the CGLS LC100 workflow are provided at a ground sampling distance (GSD) of ~0.001 degree (~100 m spatial resolution) and a GSD of ~0.003 degree (~300 m spatial resolution). The surface reflectance is available for four spectral bands corresponding to the selected measurement (Table 3). The atmospheric correction is performed using SMAC 4.0 (Rahman and Dedieu, 1994). Standard input data layers includes Normalized Difference Vegetation Index (NDVI), geometric viewing and illumination conditions, reference to date and time of observations for four reflectance bands (Table 3) and a status map containing identification of snow, ice, shadow, clouds, land/sea for every pixel (Table 4). The data is stored in a szip compressed hdf5 file.

Table 3: Spectral characteristics of the PROBA-V bands

Spectral band	Wavelength
BLUE	0.447 – 0.493 μm
RED	0.610 – 0.690 μm
NIR	0.770 – 0.893 μm
SWIR	1.570 – 1.650 μm

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

Bit	Name	Description
1 -3	Observation	000: clear 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 (Most significant)	BLUE quality flag	0: invalid data 1: valid data

The status mask bit mapping for the S1-TOC 300 m and S1-TOC 100 m products is the same. For each pixel of the product the following information can be extracted: the observation status (bits 1 to 3) is based upon the most optimal observation used for the daily synthesis; the land/sea mask bit (bit 4) is an exact copy of the PROBA-V land/sea mask; Bits 5, 6, 7 and 8 are quality flags for the daily synthesis spectral bands (quality flag is set to 1 when the radiometric quality is good).

3.2.2 PROBA-V land/sea mask

The PROBA-V land/sea mask was implemented in the PROBA-V spacecraft on-board algorithms in order to predict the land/sea transitions to reduce the amount of data generated. The PROBA-V land/sea mask was generated out of the SPOT VGT land/sea mask [SPOT-VGT Collection 3 User Manual].

Since the land/sea mask is implemented in the PROBA-V S1-TOC status mask, the land/sea mask can be easily extracted and is shown in Figure 2.

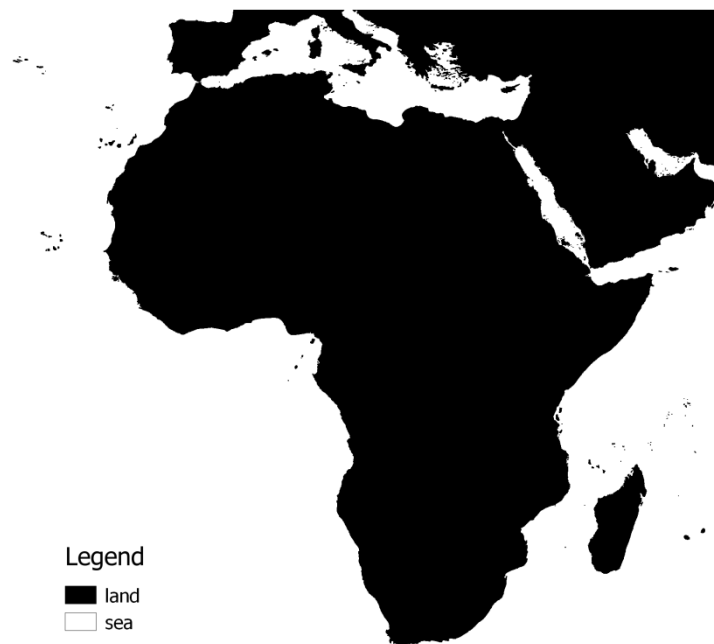


Figure 2: PROBA-V land/sea mask for the African continent

3.2.3 FAO global ecological zones dataset

In order to group EO data for faster processing or adaptation of algorithms to specific regions, we use the global ecological zone (GEZ) dataset for 2010 of the Food and Agriculture Organization of the United Nations (FAO) (FAO, 2012). FAO defines an ecological zone (EZ) as: “A zone or area with broad yet relatively homogeneous natural vegetation formations, similar (not necessarily identical) in physiognomy. Boundaries of the EZs approximately coincide with the map of Köppen-Trewartha climatic types, which was based on temperature and rainfall. An exception to this

definition are “Mountain systems”, classified as one separate EZ in each Domain and characterized by a high variation in both vegetation formations and climatic conditions caused by large altitude and topographic variation” (Simons, 2001).

Figure 3 shows the GEZ dataset for the African continent. Overall, 14 ecozones subdivide the African continent and are used to subset the EO data in several processing steps in the CGLS LC100 product workflow.

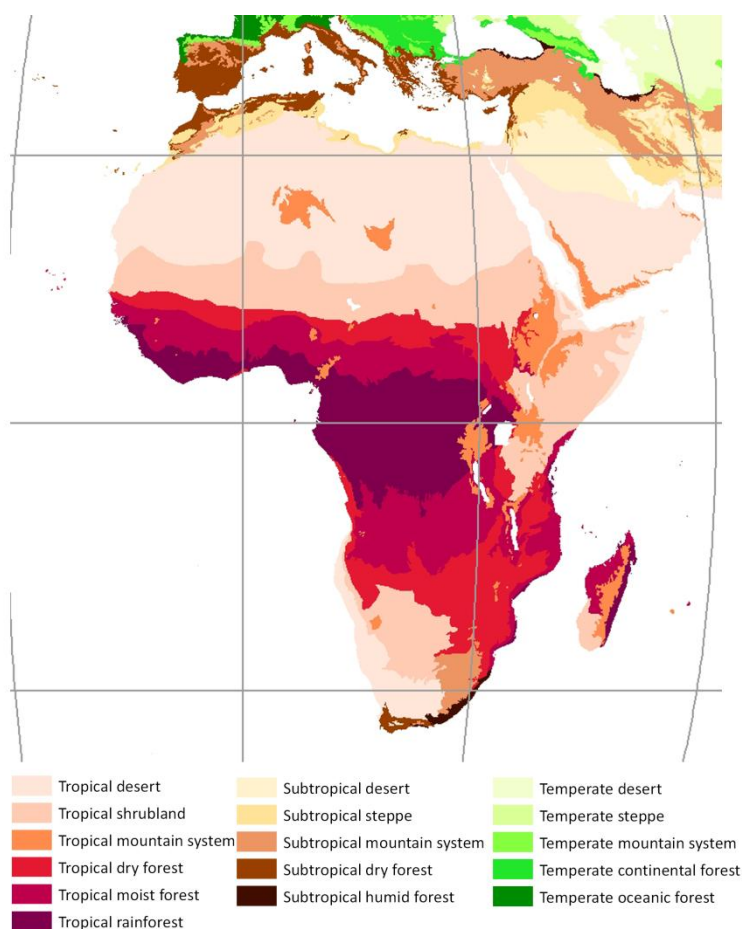


Figure 3: FAO global ecological zones dataset for the African continent. Image edited – original from <http://foris.fao.org/static/data/fra2010/ecozones2010.jpg> (FAO, 2012)

3.2.4 Shoreline vector layer for Africa

The shoreline layer is mainly used to distinguish between open land water and open sea water. 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 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. Therefore the PROBA-V land/sea mask is not usable for distinguishing open land water pixels from open sea water pixels. Figure 4 shows an example of the USGS shoreline vector layer for the Street of Gibraltar overlaid to a Google Earth image and the PROBA-V land/sea mask.

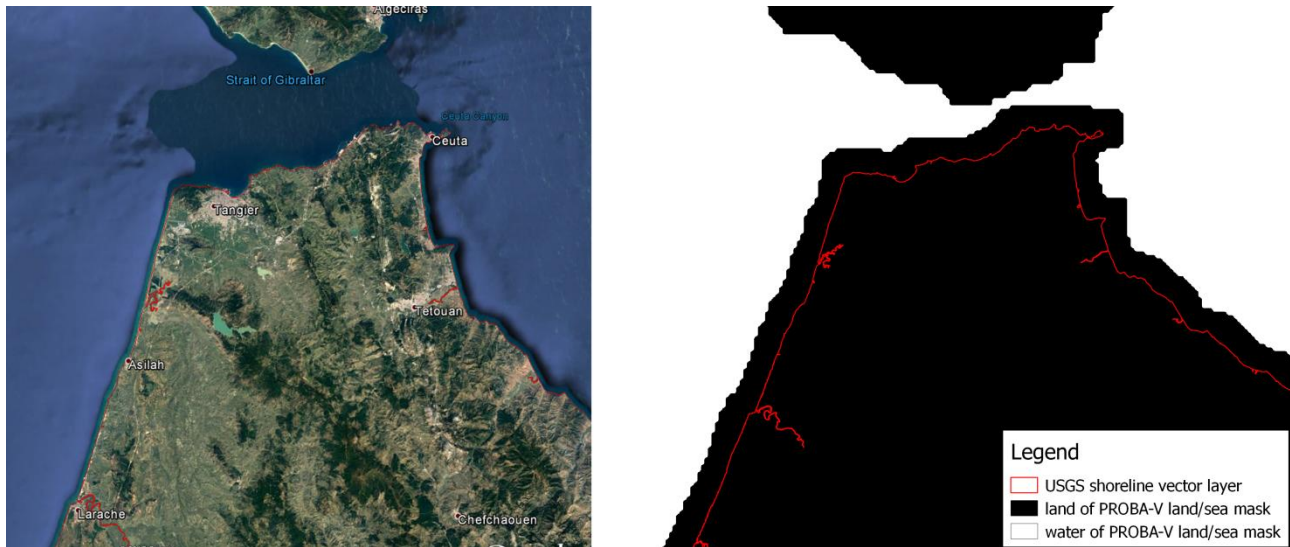


Figure 4: The USGS shoreline vector layer (Sayre et al., 2013). Zoom in to the Street of Gibraltar showing the shoreline vector layer (red) overlay to a Google Earth image (left) and to the PROBA-V land/sea mask (right). Note that the PROBA-V land/sea mask has a buffer around the land masses in order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

3.2.5 DLR's Global Urban Footprint plus (GUF+) layer

In order to generate an urban mask in the CGLS LC100 workflow, DLR's Global Urban Footprint Plus layer (GUF+) for 2015 (Marconcini et al., 2017a, Marconcini et al., 2017b) is used. The GUF+ 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. Compared to DLR's GUF layer which is using TerraSAR-X and TanDEM-X radar data, the GUF+ Sentinel-1 radar data is provided via the open data policy. Therefore the GUF+ layer is freely available. Figure 5 shows an example of the GUF+ 2015 layer for the Street of Gibraltar.

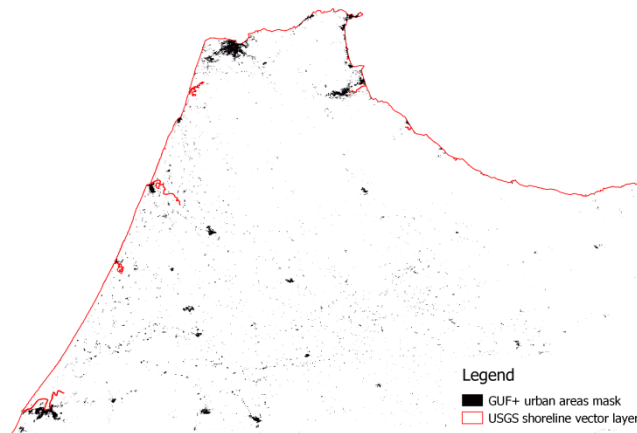


Figure 5: The DLR Global Urban Footprint plus layer (Marconcini et al., 2017b). Zoom in to the Street of Gibraltar showing the urban areas in black. Note: in red the USGS shoreline vector layer is shown.

3.2.6 JRC's Global Human Settlement (GHS) layer

A second external dataset needed to generate an urban mask in the CGLS LC100 workflow is JRC's Global Human Settlement Layer (GHS) for 2014 (Pesaresi et al., 2015). The GHS built-up grid used the 30 m Landsat EO data archive to generate a human settlement layer with a spatial resolution of 38 m. Compared to the GUF+ layer which uses solely EO data in the generation, the GHS built-up infrastructures mask is created by combining global archives of fine-scale satellite imagery, census data, and volunteered geographic information in a spatial data mining technology (model) (Pesaresi et al., 2015). The GHS built-up grid raster dataset data is provided via the open data policy. Figure 6 shows an example of the GUF+ 2015 layer for the Street of Gibraltar.

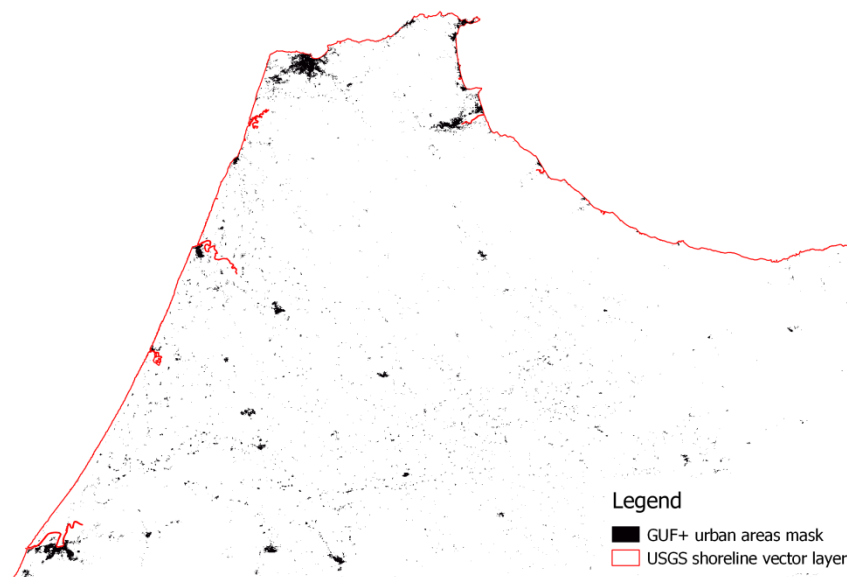


Figure 6: The JRC Global Human Settlement layer (Pesaresi et al., 2015). Zoom in to the Street of Gibraltar showing the urban areas in black. Note: in red the USGS shoreline vector layer is shown.

3.2.7 JRC's Global Surface Water (GSW) product

In order to generate a water product (permanent and temporary water bodies) in the CGLS LC100 workflow, JRC's Global Surface Water (GSW) product is used (Pekel et al., 2016). The GSW used the 30 m Landsat 5, 7 and 8 EO data archive to generate a mask showing the water surfaces that are 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 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 (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 7 shows an example of the GSW maximum water extent layer and the water seasonality 2014/2015 layer for the Street of Gibraltar.

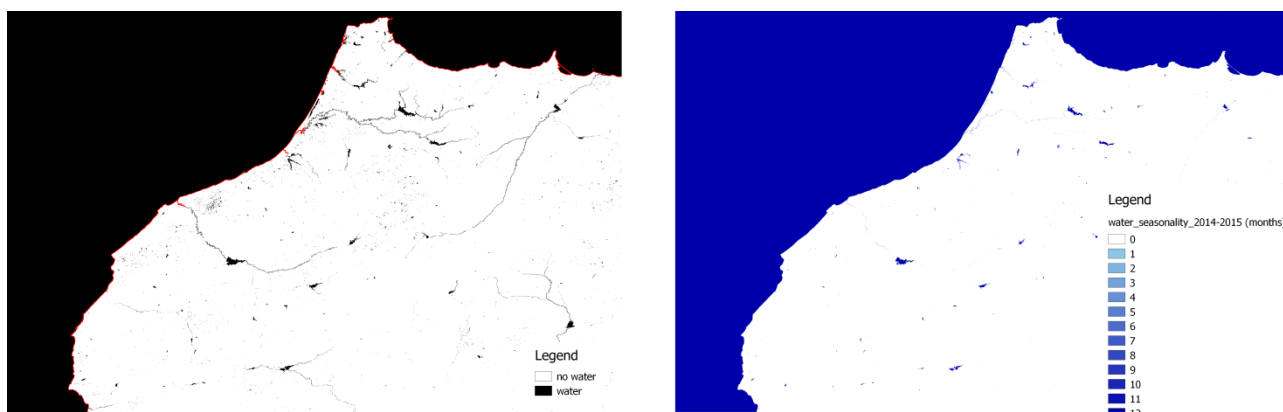


Figure 7: The JRC Global Surface Water product (Pekel et al., 2016). Zoom in to the Street of Gibraltar showing the maximum water extent layer (left) and the water seasonality mask for 2014/2015 (right). Note: in red the USGS shoreline vector layer is shown in the maximum water extent layer (left image).

3.2.8 NASA's Shuttle Radar Topography Mission Global 1 arc second dataset

For the generation of the water bodies potential mask (WBPM) within 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 input (NASA, 2013). SRTM was the primary payload on the STS-99 mission of the Space Shuttle Endeavour, and provided the first complete high-resolution digital elevation model (DEM) on a near-global scale from 56° S to 60° N. The global 1 arc second dataset provides a ~30 m spatial resolution and is the third version of this dataset which improved the DEM quality by filling void pixel with the ASTER Global Digital Elevation Model, the Global Multi-resolution Terrain

Elevation Data, and the National Elevation Dataset (NASA, 2013). Figure 8 shows an example of the SRTMGL1 dataset for the Street of Gibraltar.

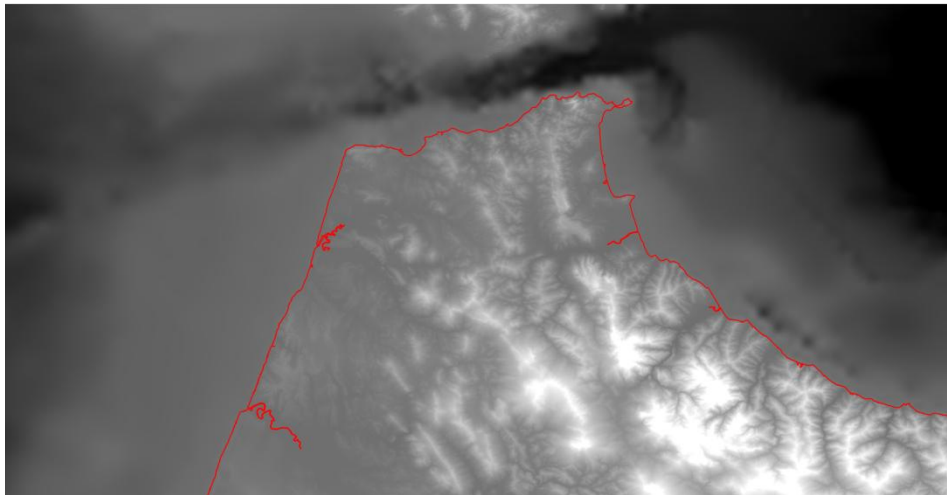


Figure 8: The NASA Shuttle Radar Topography Mission Global 1 arc second dataset.). Zoom in to the Street of Gibraltar. Note: in red the USGS shoreline vector layer is shown.

3.3 DATA CLEANING & COMPOSITING

3.3.1 Overview

The first section in the CGLS LC100 product workflow (Figure 1) is the data cleaning and compositing section (Figure 9). Input data is EO data from the PROBA-V multi-spectral satellite as described in 3.2.1. Next to the SM cleaning of the 100 m and 300 m PROBA-V TOC data for the reference year plus/minus 3 months (up to 548 daily observations)), a temporal outlier cleaning algorithm called madHANTS is applied. In order to reduce the noise in the S1 synthesis products as well as to generate regular time steps, a median compositing algorithm is used to generate 5-daily and 10-daily TOC composites. For the 300 m PROBA-V data also a harmonized 5-daily median composite (HMC5) for the whole PROBA-V data archive (2013 – 2017) is produced. Therefore, a harmonization algorithm called HANTS is used.



Figure 9: 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.

3.3.2 The HANTS and madHANTS algorithms

The Harmonic ANalysis of Time Series (HANTS) algorithm and derived madHANTS (median absolute deviations of the Harmonic ANalysis of the Time Series) 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 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 for the CGLS LC100 product workflow where the time series profiles for each pixel are 1 ½ years long (reference year plus/minus 3 months).

2. In the second step, the outlier detection is performed - the original pixel values along the time series are evaluated against the corresponding harmonized 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 values from the current pixel value x_i in the numerator of the division. The adaptation in Equation 1 using instead the median the corresponding harmonized pixel value in the numerator allows a better adaptation of the outlier detection algorithm to seasonality.

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

, where s is the evaluation score to determine if a pixel value along a time series is an outlier compared to the neighbouring time steps, x_i is the current pixel value of the to test time series step, x_{iHANTS} is the harmonized pixel value of the current time step using the HANTS algorithm.

A threshold of 3.5 standard deviations is used to determine if a pixel value has to be characterized as outliers compared to the neighbouring pixel values in the time series. 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 10 shows an example of the applied madHANTS algorithm on a PROBA-V 100 m time series of a pixel a location 9.459° lon, 6.562° lat.

Figure 10 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). Therefore, we decided to extent the EO time series data of the reference year for that the CGLS LC100 product is produced by plus/minus three months.

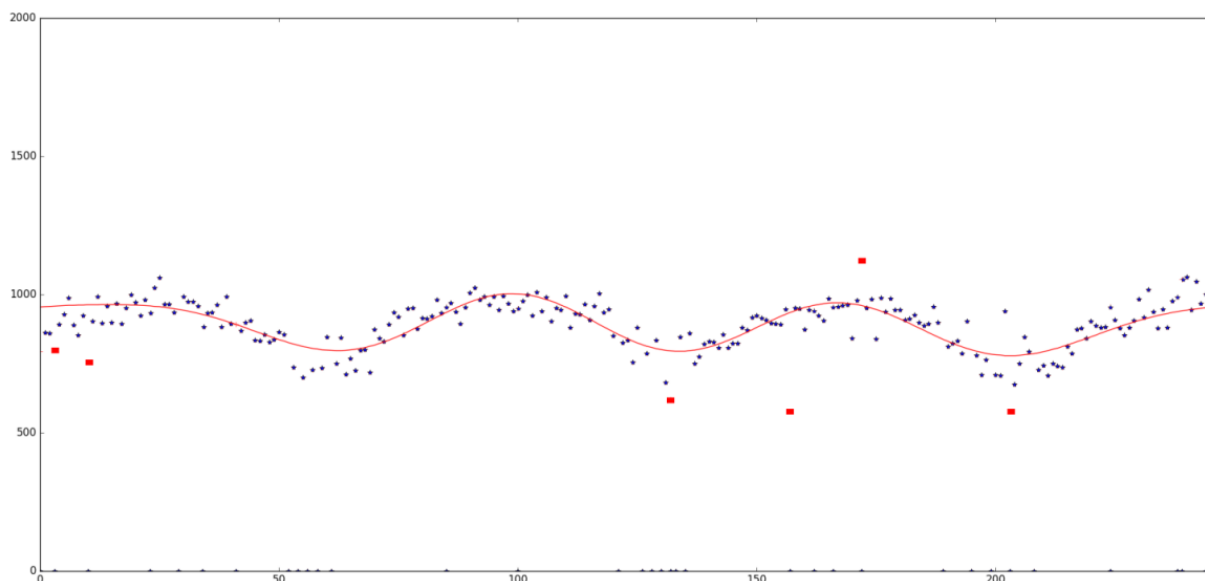


Figure 10: Example for the madHANTS temporal outlier detection algorithm. Note: blue stars mark all original pixel values in the time series (invalid observation values are set to -1), the red line shows the harmonized time series by applying the amplitudes and phases of the identified frequencies during the HANTS transformation, red squares show the detected outliers which are 3.5 standard deviations away from the harmonized pixel value.

3.3.3 Median Composites (MC) generation

The general workflow for the data cleaning and compositing section (outlined in Figure 9) can be split in two steps. The first is the generation of SM cleaned and temporal outlier cleaned 5-daily and ten-daily median composites (MC) for the 100 m and 300 m time series data. The median compositing reduces the random component of the noise compared to the S1-TOC.

Overall three processing sub-steps are needed in order to generate the 5-daily (MC5) and 10-daily (MC10) TOC composites for the S1-TOC PROBA-V EO data (Figure 9). First, the PROBA-V multi-spectral TOC image data with a ground sampling distance (GSD) of ~0.001 degree (~100 m) and PROBA-V multi-spectral TOC image data with a GSD of ~0.003 degree (~300 m) are retrieved from the S1 (daily) Collection 1 archive for the African continent for the reference year 2015 plus 3 months before and after the reference year. This can be up to 548 observations (time steps) for each single pixel for the 1 ½ years long time series for the PROBA-V 100 m archive as well as the PROBA-V 300 m archive. Each dataset of a time step consists of four spectral bands and the corresponding SM file (see chapter 3.2.1) which are handled simultaneously. The status mask cleaning sets pixel flagged as noise, cloud, or sea to a “no data” value.

In the next processing sub-step, the madHANTS algorithm (see 3.3.2) is applied to clean the time series from remaining haze and undetected clouds. Therefore, the madHANTS outlier test is conducted for the blue and SWIR reflectance bands of the PROBA-V S1-TOC 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 11 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 tile using multi-core computing as well as several PROBA-V tiles are processed simultaneously via cloud computing. Output of the madHANTS processing step are PROBA-V S1-TOC SM and outlier cleaned 100 m time series for the four reflectance bands (300 m respectively). Figure 12 shows an example of the data quality improvement by applying the SM cleaning and temporal outlier cleaning processing steps on a PROBA-V S1-TOC 100m image.

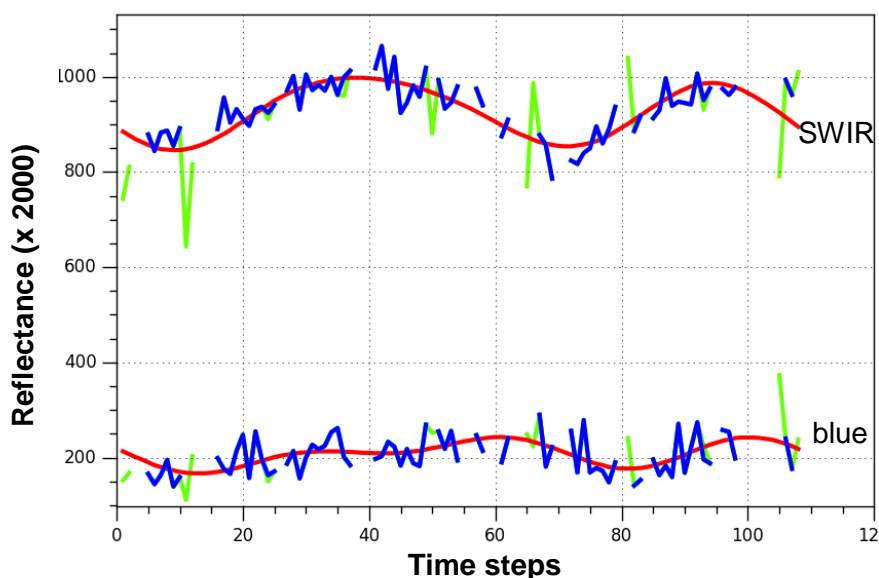


Figure 11: Example showing the temporal outlier detection using the madHANTS algorithm on the blue and SWIR reflectances of a pixel a location 9.459° lon, 6.562° lat. Note: red curve shows the harmonized time series; blue shows the valid pixel values; green shows flagged outliers cumulative detected in the blue or SWIR reflectance band.

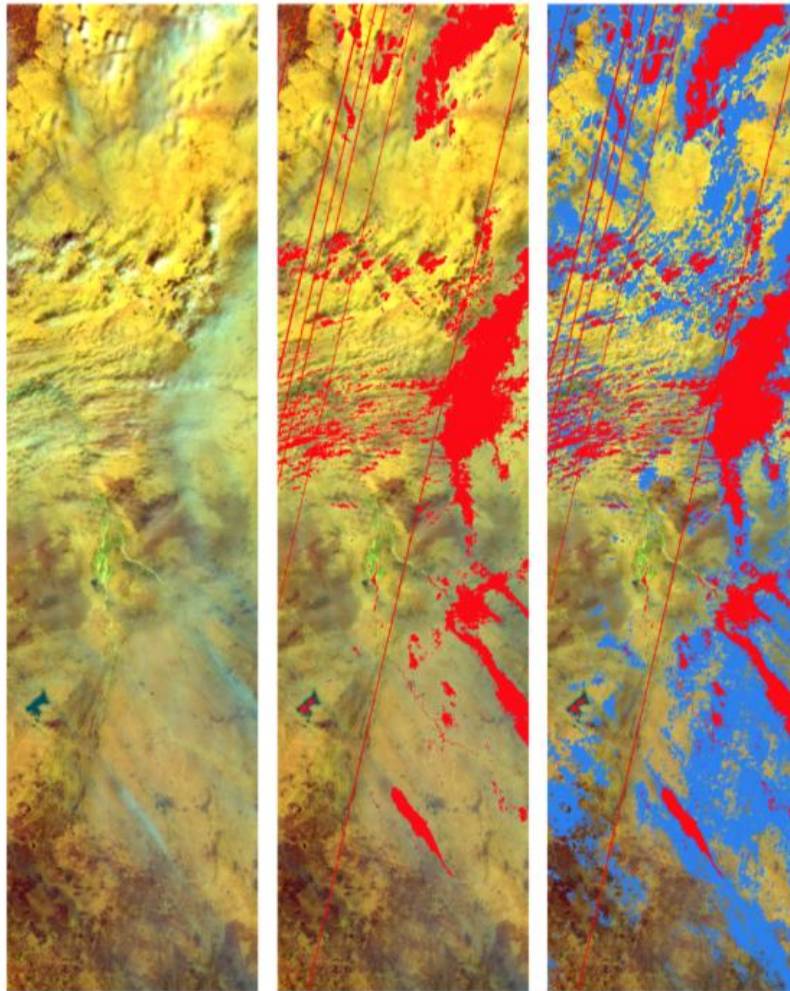


Figure 12: Example for data cleaning process of PROBA-V 100m image from 2016-03-06. Image is shown as false color composite (RGB = SWIR, NIR, blue) for a sample area in tile X18Y06 (Nigeria – left) raw image, middle) status masked cleaned image (quality flagged areas are shown in red), right) madHANTS cleaned image (additional pixels which are flagged as outliers or clouds are shown in blue).

The last processing sub-step is the generation of the MC5 and MC10 data sets for the 100 m and 300 m time series data. The MC5 (MC10 respectively) composites for the four spectral bands are obtained by calculating the median of the S1-TOC daily reflectance values over a 5-days window (10-days window respectively). The MC is calculated across the four spectral bands for each pixel and all pixels in a tiled TOC-S1 image simultaneously. Output of the MC generation step are PROBA-V MC5-TOC SM and outlier cleaned as well as PROBA-V MC10-TOC SM and outlier cleaned 100 m time series for the four reflectance bands (300 m respectively).

3.3.4 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 9) is the generation of the long-term 5-daily median composite (HMC5) for the PROBA-V S1-TOC 300 m time series. Since the HMC5 is produced out of the whole PROBA-V data archive (2013 – 2017), the HMC5 workflow had to be adapted to the data volume and differs slightly from the MC generation workflow (section 3.3.3).

First, all PROBA-V daily multi-spectral TOC image data with a GSD of ~0.003 degree (~300 m) is retrieved from the S1 (daily) Collection 1 archive for the African continent. This can be over 1500 observations (time steps) for each single pixel for the current 4 years (2013-2017) of PROBA-V. Next follows the status mask cleaning which flags pixel stated as noise, cloud, or sea to a “no data” value.

Instead of proceeding with the temporal outlier screening of the PROBA-V S1-TOC SM cleaned 300 m time series data, the HMC5 workflow generates the MC5 composites first (Figure 9). This was needed since the madHANTS algorithm gets slower the longer the time series gets, which made a time effective processing impossible. The 5-daily compositing compress the data amount by factor 5. Again, the MC5 composites for the four spectral bands are obtained by calculating the median of the PROBA-V S1-TOC SM cleaned 300m daily reflectance values over a 5-days window. Output are PROBA-V MC5-TOC SM cleaned 300 m time series for the four reflectance bands. Now follows the temporal outlier screening and removal using the madHANTS workflow as explained in section 3.3.3. Output of that are PROBA-V MC5-TOC SM and outlier cleaned 300 m time series for the four reflectance bands.

The final sub-step is the generation of the harmonized composites. Therefore, the HANTS algorithm is used (see section 3.3.2). 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 MC5-TOC SM and outlier cleaned 300 m data. 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 intervals as the PROBA-V MC5-TOC 300 m data set. Output are PROBA-V HMC5-TOC SM and outlier cleaned 300 m time series for the four reflectance bands.

3.4 DATA FUSION

3.4.1 Overview

High seasonal cloud coverage in several African regions are challenging for all optical based land cover classification approaches. In order to overcome the low data density and therefore data gaps in the PROBA-V 100 m MC5 time series product (see Figure 13), PROBA-V 300 m data which has a daily revisit time is fused in via a Kalman filtering, also known as linear quadratic estimation, approach (Kalman, 1960).

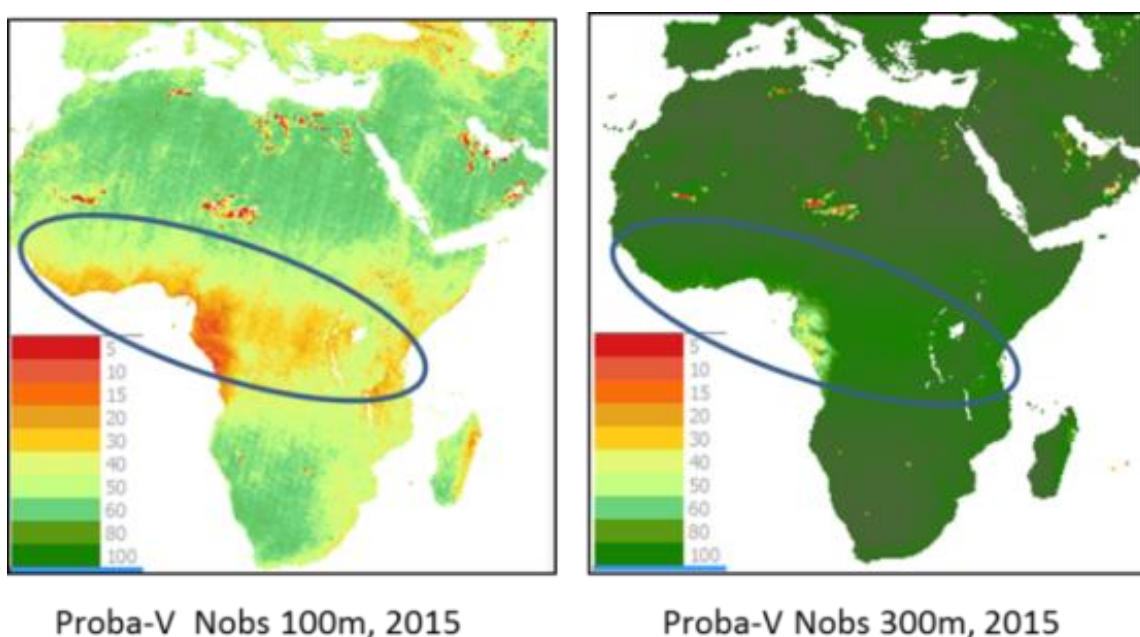


Figure 13: Number of cloud free data observations for PROBA-V 100 m and PROBA-V 300 m in the reference year 2015 for Africa. Note: Blue circled area indicates area with lowest observation density in Africa.

Figure 14 illustrates an overview for 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 are the PROBA-V 100 m and 300 m 5-daily and 10-daily median composites as well as the generated PROBA-V 300 m long-term 5-daily median composite for all four PROBA-V reflectance bands. In a first step, data gaps in the 100 m and 300 m MC5 time series are filled with corresponding existing observations in the MC10 time series, and huge data gaps (> 25 days) in the 300 m MC5 time series 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 of the fused 100 m and 300 m for outliers and the consistency of the data is checked.

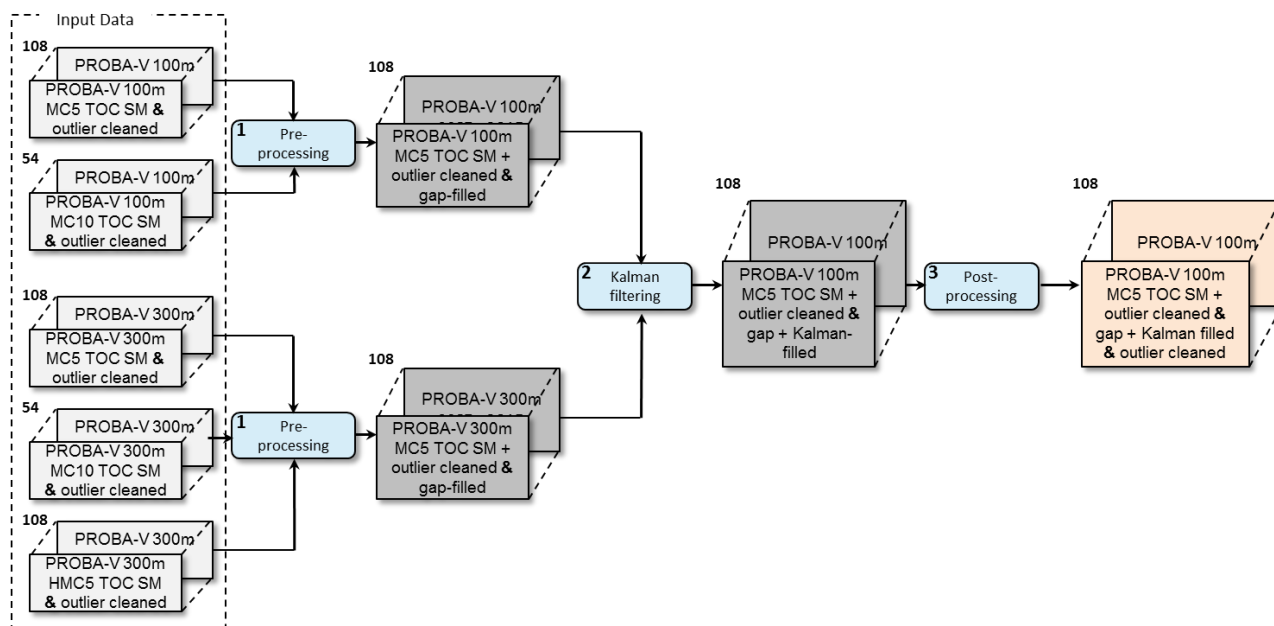


Figure 14: General overview of the data fusion section in the CGLS LC100 product workflow. Note: Numbers in the upper left corner of data container indicate the number of layers - here number of observations.

3.4.2 Data Fusion pre-processing

Input data in the data fusion pre-processing step are the generated PROBA-V MC5, MC10 and HMC5 time series cubes for all four reflectance bands (see section 3.3 and Figure 14). In detail:

- PROBA-V 100m MC5-TOC SM & outlier cleaned for the blue, red, NIR and SWIR band;
- PROBA-V 100m MC10-TOC SM & outlier cleaned for the blue, red, NIR and SWIR band;
- PROBA-V 300m MC5-TOC SM & outlier cleaned for the blue, red, NIR and SWIR band;
- PROBA-V 300m MC10-TOC SM & outlier cleaned for the blue, red, NIR and SWIR band;
- PROBA-V 300m HMC5-TOC SM & outlier cleaned for the blue, red, NIR and SWIR band.

In the data fusion pre-processing, small gaps (5 – 10 day gaps) in the 100 m and 300 m MC5 time series products are filled with the pixel values of the MC10 time series products for the corresponding time steps. In a second step, bigger gaps (> 25 day gaps) in the 300 m MC5 time series product are filled via interpolation with the HMC5 300 m long term trend product for the corresponding time series steps (see Figure 15). This is needed in order to guide the Kalman filtering approach in cases where no PROBA-V 100 m and 300 m MC5 data is available for more than 1 months in the row for a pixel. Note: big gaps are not filled in the 100 m data – therefore the Kalman-Filtering approach is used. The post-processing works tile based meaning all pixels and corresponding pixel time series are processed simultaneously for the 100 m (300 m respectively)

time series datasets. Moreover, all four PROBA-V reflectance bands are processed simultaneously.

Output of the pre-processing are two gap-filled datasets, the PROBA-V MC5-TOC SM & outlier cleaned and gap filled 100 m time series for the four reflectance bands, and the PROBA-V MC5-TOC SM & outlier cleaned and gap filled 300 m time series for the four reflectance bands.

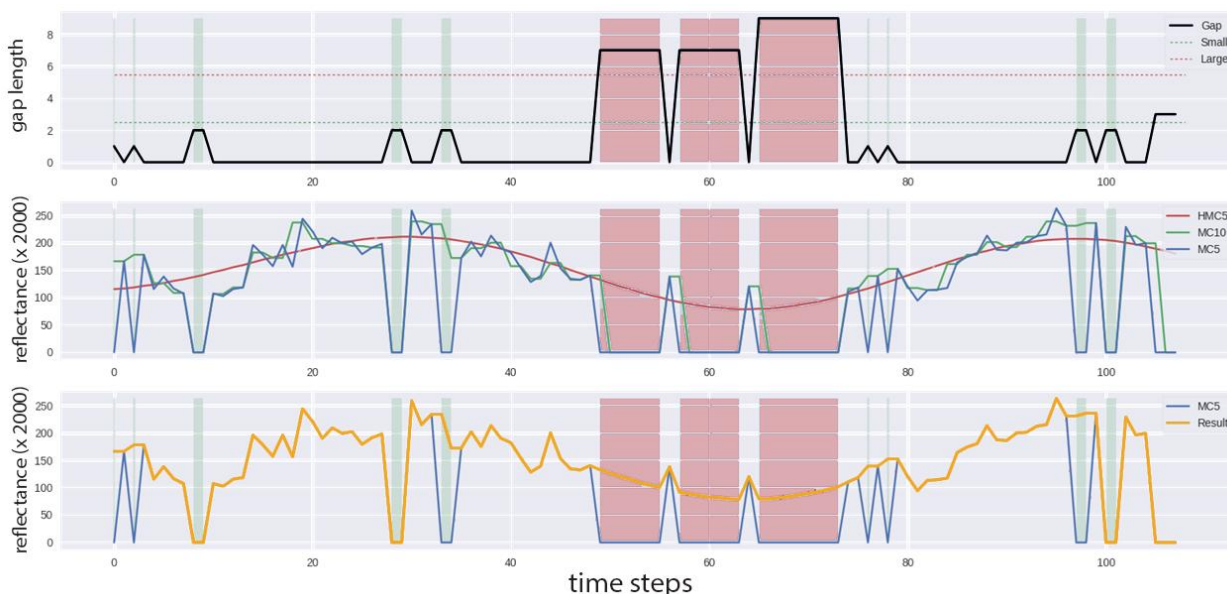


Figure 15: Example for data fusion pre-processing results for PROBA-V 300m MC5 time series for pixel location 9.459° lon, 6.562° lat. Top: continuous gap length in the time series (dashed green line indicates threshold for small gaps, dashed red line indicates threshold for big gaps), middle: time series before pre-processing (blue line shows original 5-daily median composite time series, green line shows original 10-daily median composite time series, red line shown the long term harmonized time series for the full PROBA-V archive), bottom: 300m time series after pre-processing (blue line shows the original 5-daily median composite time series, orange line shows the final pre-processed time series which will be used for the data fusion).

3.4.3 Data Fusion using the Kalman-Filtering approach

The Kalman-filtering approach 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.

The Kalman filter algorithm used in our workflow was introduced 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 100 m pixels are within one PROBA-V 300 m pixel at the same geographic location) as independent variables of the coarse pixel trend (spectral unmixing), all 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. Figure 16 illustrates this concept and shows that the pixel values of the fine resolution image between time step $k-1$, k , and $k+1$ are still independent for each other, but still all fine resolution pixel value trends follow the single trend of the pixel values of the coarse resolution image between the time stamps.

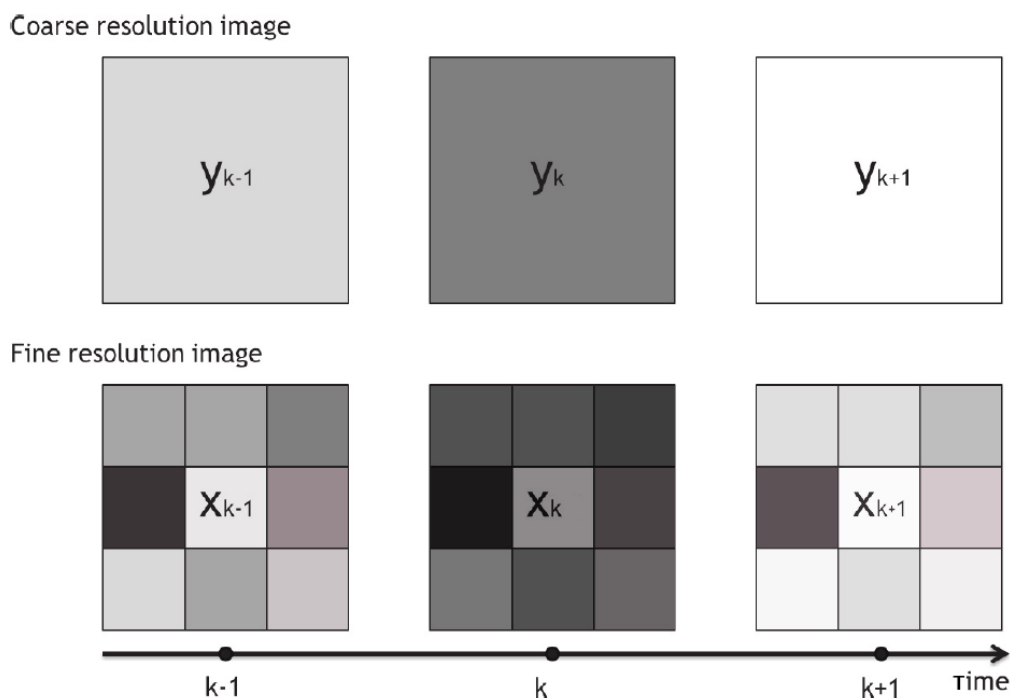


Figure 16: Principle of the Kalman-filtering approach for EO data fusion using the pktools algorithm. Adapted from Kempeneers et al. (2016).

Input data into the Kalman-Filtering processing step are the PROBA-V MC5-TOC SM & outlier cleaned and gap filled 100 m time series for the four reflectance bands, and the PROBA-V MC5-TOC SM & outlier cleaned and gap filled 300 m time series for the four reflectance bands. The four reflectance bands are processed independently from each other, but multi-core processing is used to process all data simultaneously. The pktools Kalman-filter algorithm is run in three modes which have to be processed sequentially. In the first run, the forward mode, the pixel values for all time

steps in the fine resolution data (PROBA-V 100m) are predicted for the current date meaning the prediction model is built up by walking forward through the time steps. 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. 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. Output of the data fusion step is a continuous, gap-free 5-daily time series in fine resolution - PROBA-V MC5-TOC SM & outlier cleaned and gap & kalman filled 100 m time series - for the four reflectance bands.

3.4.4 Data Fusion post-processing

The last step in the data fusion workflow is the post-processing in which the Kalman filled PROBA-V 100 m MC5-TOC data is screened for consistency and introduced outliers. 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. Also the mentioned simplification of the Kalman-filtering algorithm in the pktools (see section 3.4.3) by applying only a single trend to all fine pixels within one coarse pixel, can create artefacts in extreme heterogeneous landscapes.

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 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 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 MC5-TOC SM & outlier cleaned and gap filled 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 extreme values, 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 3.3.2). 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 100 m time series for the reference year in 5-days intervals (PROBA-V MC5-TOC SM & outlier cleaned and gap & kalman filled & outlier cleaned 100 m time series) 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 17.

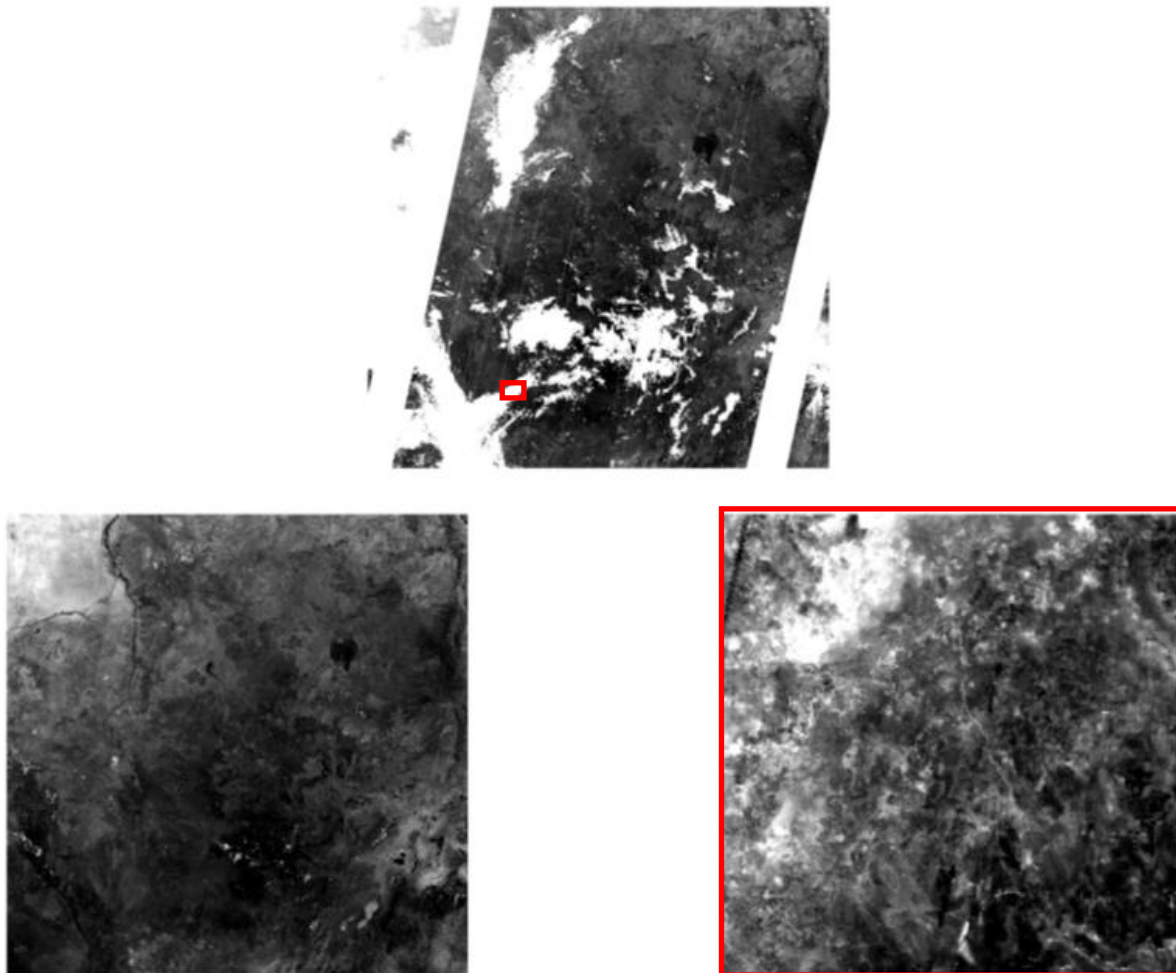


Figure 17: top) PROBA-V 100 m MC5 pre-processed image (areas with missing data is shown in white), bottom-left) PROBA-V 100 m MC5 image after data fusion approach, bottom-right) zoom in to full PROBA-V resolution of left image over red box shown in top image (the area in the red box had no observations at all for that time stamp and shows, after data fusion, consistent image data). Example over a test area in tile X18Y06 (Nigeria), on 2016-03-06.

3.5 METRICS GENERATION

3.5.1 Overview

Figure 18 illustrates an overview for the metrics generation section in the CGLS LC100 product workflow (Figure 1). The term “metrics” 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.

Overall three processing steps are needed in order to produce all metrics for the CGLS LC100 product. The first step is the calculation of time series of additional vegetation indices (NDVI, EVI, SIPI, NBR, NIRv) out of the four reflectance bands of the PROBA-V 100m data fused time series profiles. Moreover, a HSV colour transformation, phenology analysis, water bodies probability map, and additional masks generation is performed in the first processing step. During the second processing step, the metrics extraction, the descriptive statistics of the time series profiles of the four PROBA-V reflectance bands and five calculated vegetation indices (VI) and two color transformed bands 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 reflectance bands and VI's profiles. 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 392 generated metrics for each PROBA-V 100m pixels are combined into one container.

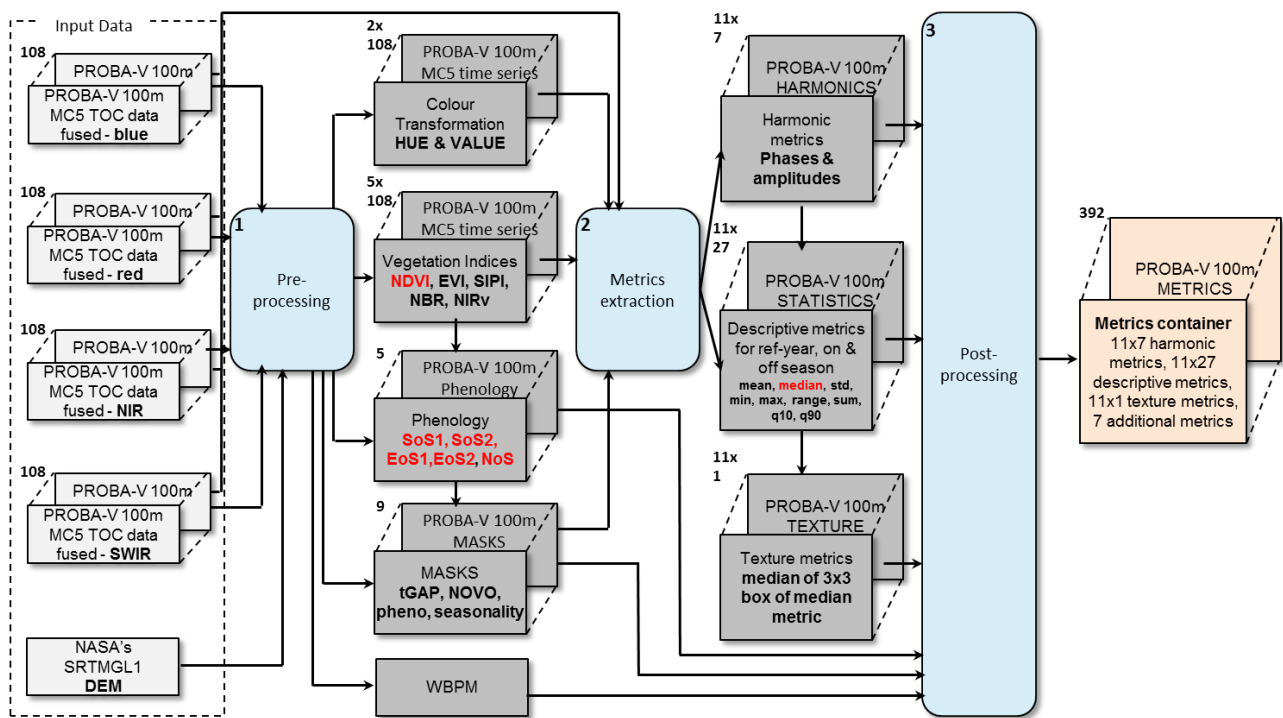


Figure 18: General overview of the metrics generation section in the CGLS LC100 product workflow.
Note: datasets marked in red are re-used in sub-steps of a processing step. Numbers in the upper left corner of data container indicate the number of layers.

3.5.2 Pre-Processing

The pre-processing step of the metrics generation section can be subdivided into five sub-steps (Figure 19): in the first, additional vegetation indices for each time step in the PROBA-V 100 m MC5-TOC data fused time series are generated; in the second, the HSV colour transformation is carried out; in the third, the water bodies probability map is generated out of the DEM; in the fourth, the phenology product is generated for the PROBA-V 100 m time series data; and in the fifth sub-step, additional masks are generated.

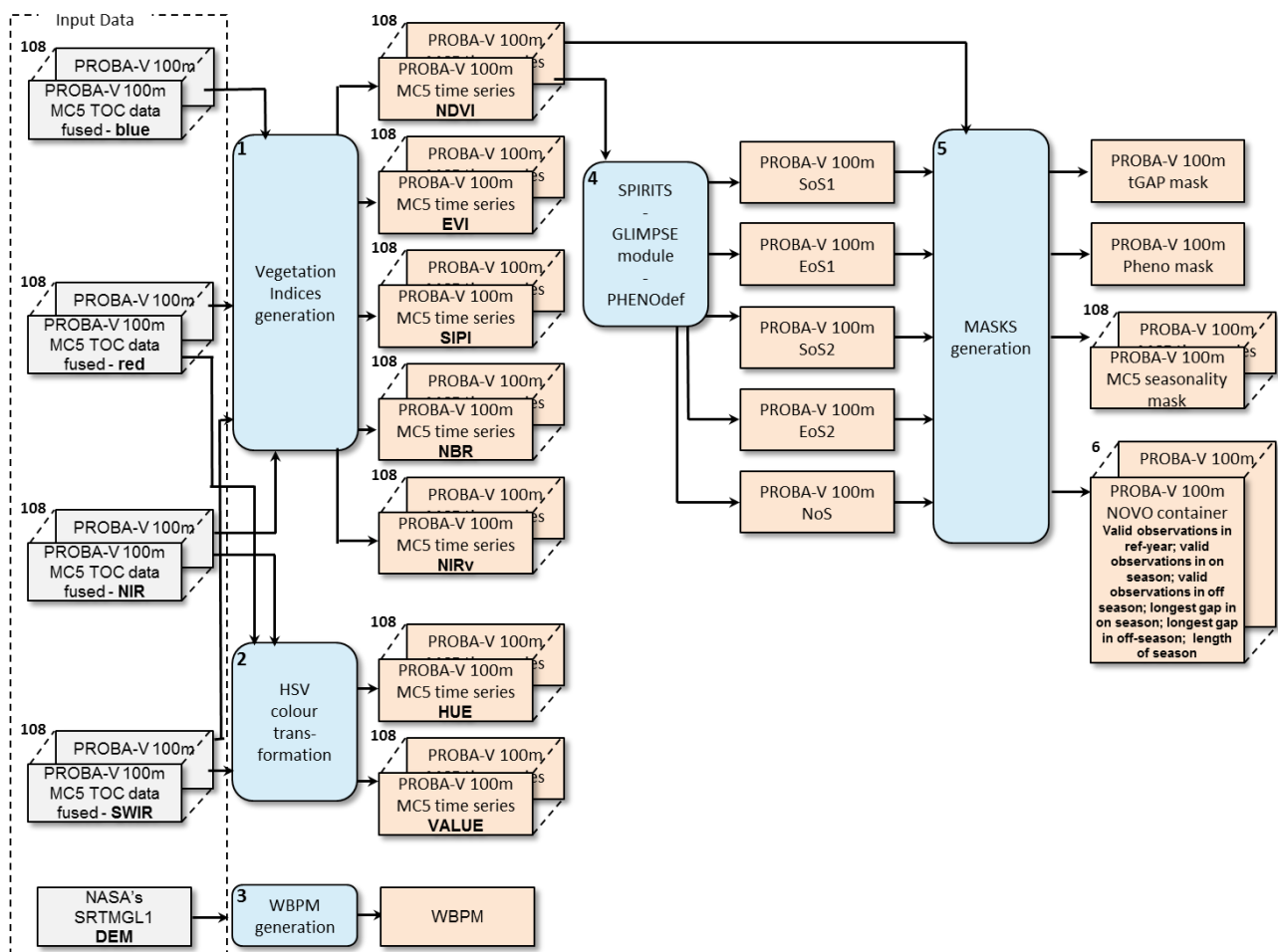


Figure 19: Overview of the processing sub-steps within the pre-processing step of the metrics generation section in the CGLS LC100 product workflow. Note: Numbers in the upper left corner of data container indicate the number of layers.

3.5.2.1 PROBA-V 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 five 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 2 shows the formula for the NDVI:

$$NDVI = \frac{NIR - red}{NIR + red} \quad \text{Equation 2}$$

, where NIR and red refers to 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 -1 to 1. The common range for green vegetation is 0.2 to 0.8. Equation 3 shows the formula to calculate the EVI for PROBA-V:

$$EVI = 2.5 * \frac{NIR - red}{(NIR + 6 * red - 7.5 * blue + 1)} \quad \text{Equation 3}$$

, where NIR, red and blue refers to 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 0 to 2. The common range for green vegetation is 0.8 to 1.8. Equation 4 shows the calculation formula for PROBA-V:

$$SIPI = \frac{NIR - blue}{NIR - red} \quad \text{Equation 4}$$

, where NIR, red and blue refers to the corresponding PROBA-V reflectance bands.

4. Normalized Burn Ratio (NBR) using the NIR and the SWIR reflectance bands (Key and Benson, 2005). The NBR is a commonly used index used to detect burned areas and in some cases to estimate the burn severity. Equation 5 shows the calculation formula:

$$NBR = \frac{NIR - SWIR}{NIR + SWIR} \quad \text{Equation 5}$$

, where NIR and SWIR refers to 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 6:

$$NIR_v = \left(\frac{NIR - red}{NIR + red} - 0.08 \right) * red \quad \text{Equation 6}$$

, where NIR and red refers to the corresponding PROBA-V reflectance bands.

Input data into the calculation of the VI's are the PROBA-V MC5-TOC SM & outlier cleaned and gap & kalman filled & outlier cleaned 100 m time series (for easier readability shorted to PROBA-V MC5-TOC data fused 100m, see Figure 19) 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 Equation 2 to Equation 6. Thus, all five VI's and time steps can be processed simultaneously. Output are VI time series profiles which are named:

- PROBA-V 100m MC5-NDVI,
- PROBA-V 100m MC5-EVI,
- PROBA-V 100m MC5-SIPI,
- PROBA-V 100m MC5-NBR,
- PROBA-V 100m MC5-NIR_v.

3.5.2.2 PROBA-V HSV colour transformation

The HSV colour transformation is an approach that transforms the RGB (Red, Green, Blue) color space into the HSV (Hue, Saturation, and Value) color space which decouples chromaticity and luminance. The HSV color space is commonly used in image processing. It is a nonlinear transformation of the RGB color space using equations 7, 8 and 9 presented in Figure 20.

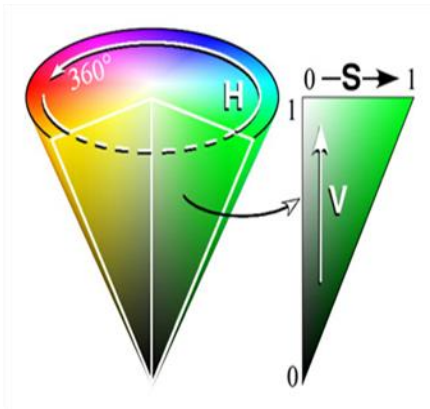
Current remote sensing applications which use the HSV color transformation are e.g. surface water detection algorithms. Therefore, by analyzing the distribution of the pixels in the Hue - Value space, the pixels can be classified as "water" and as "no water" based on thresholds (Pekel et al.,

2014; Bertels et al., 2016). Also the water product generation algorithm within the CGLS LC100 product workflow uses the HSV color transformation. Furthermore, we decided to use the time series of Hue and Value as indices in the metrics extraction process.

Input in the HSV color transformation are the red, NIR and SWIR reflectance bands of the PROBA-V 100m MC5-TOC data fused time series profiles with R mapped to SWIR band, G to NIR band and B to the RED band (Figure 20). The Hue (H) is defined as the dominant wavelength of the perceived color. 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 color point, going from 0° to 360°. The Saturation (S) is defined as the degree of purity of the color and may be intuitively considered as the amount of white mixed in a color. 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 color. 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 100m MC5-HUE,
- PROBA-V 100m MC5-VALUE.



$$V = \max(R, G, B) \quad E7$$

$$S = \frac{V - \min(R, G, B)}{V} \quad E8$$

$$H = \begin{cases} \left(60^\circ * \frac{G-B}{V - \min(R, G, B)} + 360^\circ\right) \text{ mod } 360^\circ & \text{if } V = R \\ 60^\circ * \frac{B-R}{V - \min(R, G, B)} + 120^\circ & \text{if } V = G \\ 60^\circ * \frac{R-G}{V - \min(R, G, B)} + 240^\circ & \text{if } V = B \end{cases} \quad E9$$

Figure 20: The HSV color space and the formulas for transforming the RGB color space into the HSV color space. Note: V=Value, S= Saturation, H=Hue, R=red, G=green, B=blue.

3.5.2.3 Water Bodies Potential Mask (WBPM) generation

Pixels in hilly terrain having low reflectance values due to shadow or dark vegetation are often misclassified as water bodies. To minimize these commission errors, a Water Bodies Potential Mask (WBPM) is generated which indicates if a location has the ability to hold a water body. The algorithm for the WBPM is based on Bertels et al. (2016), and uses the National Aeronautics and Space Administration (NASA) Shuttle Radar Topography Mission (SRTM) plus digital elevation model (DEM) data in 1 arc second resolution (SRTMGL1) (NASA, 2013) as input, which has 30 m

horizontal and 1 m vertical resolutions. For a detailed description of the algorithm see [GIOGL1_ATBD_WB1km-PROBAV-V2]. Main change to Bertels et al. (2016) original algorithm was the usage of a DEM with a higher resolution, since the original algorithm was developed for PROBA-V data with a GSD of 1 km.

The mask was constructed in three steps:

1. Search for the lowest points in the terrain: A pixel is a candidate for the lowest point and a potential water body (WB) when the pixel elevation is lower or equal to the pixel elevation of its eight surrounding neighbours. Therefore, the 8 pixels neighbourhood of each SRTMGL1 pixel is evaluated.
2. Filtering and expanding the detected lowest points: The next step in generating the WBPM is expanding the detected lowest points depending on the topography. For each detected lowest point, an imaginary water level is raised in steps of 1m till the maximum rise of 5m or the flooding condition is reached. As long as the edge of the potential WB is not flooded, its area is extended according to the raised level, i.e. all neighbouring pixels having the additional elevation are added to the potential WB area. This is schematically shown in a two dimensional representation in Figure 21.

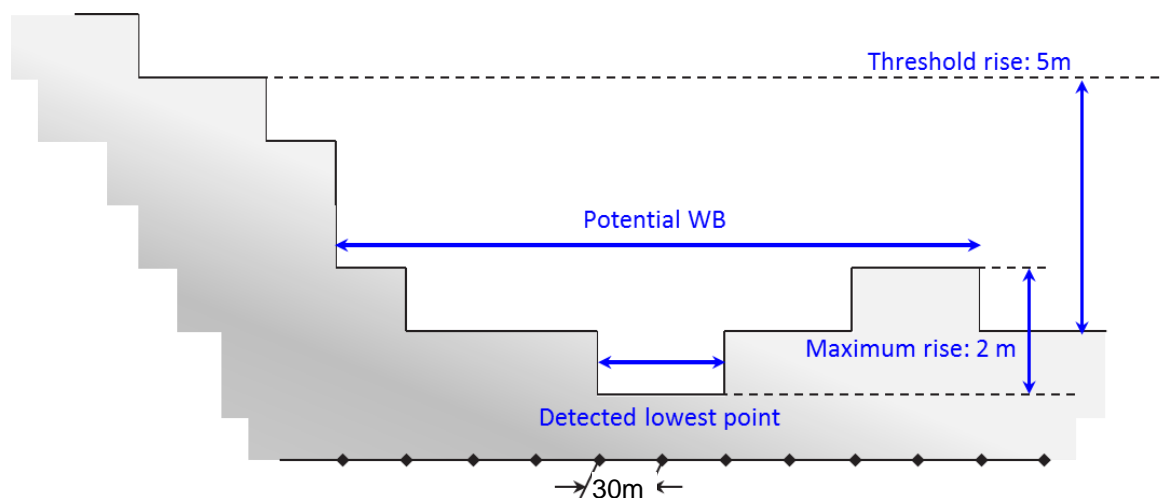


Figure 21: The Water Bodies Potential Mask generation algorithm – filtering and expanding the detected lowest points. Expanding the initially detected lowest point by systematically raising an imaginary water level in steps of 1 m. Note: the corresponding 30 m spatial resolution pixels are indicated by the dots at the bottom.

3. Deriving the WBPM: In the final step, the 30 m spatial resolution potential WBs map is re-sampled to the PROBA-V 100 m spatial resolution. For each pixel in the PROBA-V image, the corresponding pixels in the 30m potential WBs map are located.

Each pixel in the resulting 100 m WBPM can have one of two flags: 1="potential water body", 0="no water body possible". The WBPM is used as an extra input in the classification/regression step of the CGLS LC100 product workflow. The main intent of this "metric" is to guide the classifier in decisions like:

- Grassland vs. wetland
- Bare or dark soil vs. water

3.5.2.4 PROBA-V 100m Phenology Product generation

Vegetation and its live cycle are influenced by seasonal and inter-annual variations in climate and other factors (such as elevation). Phenology is the study of these changes in the timing of seasonal events. Of main importance from a RS point of view in classification approaches are the start and end date of a vegetation cycle as well as the number of vegetation cycles in a year. This can help in the distinguishing between different vegetation or agriculture classes.

Remote sensing is able to consistently generate estimates of the start, peak, duration, and end of the growing season over large areas. However, these phenological parameters are only an approximation of the true biological growth stages. Van Hoolst et al. (2016) has developed a procedure which derives such phenological information from time series of VI such as the NDVI. .

In general terms, Van Hoolst et al. (2016) algorithm inspects a time series of dekadal VI (e.g. NDVI) images and detects, for a given civil year (January to December), the number of green seasons (0, 1 or at most 2), and for each such cycle (green seasons) the start of the season (SOS), the end of the season (EOS) and, optionally, also some other variables such as MOS (date of maximum), the VI-values at SOS/MOS/EOS, the season length, etc. The algorithm is implemented in the GLIMPSE module PHEN0def which is part of the SPIRITS (Software for the Processing and Interpretation of Remotely sensed Image Time Series) software (Eerens et al., 2014). Input in the module is the generated PROBA-V 100m MC5-NDVI time series profile (see section 3.5.2.1) which covers the reference year plus/minus 3 months. The following processing steps are then carried out (Figure 22):

1. Smoothing of the NDVI profile with a running mean filter of length 5, which replaces each observation by the mean of itself plus the two neighbouring values at the left and right;
2. Flagging of exceptions for land pixels without apparent seasonality (equatorial forest, desert, water) by comparing the mean NDVI to a threshold;
3. Finding of all remaining local maxima (blue dots) and prune them via different tests/steps until at most two cycles remain (the most significant ones) within the reference year;
4. For each season (1 or 2) the SOS is defined as the moment when the rising NDVI-curve cuts a threshold, and EOS as the date when the descending curve crosses a second threshold. Often the time series of a pixel only contains one maximum (or cycle) which is labelled as "Season1" (SOS1/EOS1), but if the reference year contains two seasons, they are labelled chronologically: "Season 1" is the one with the earliest maximum in the concerned target year and "season 2" is the last one, regardless their mutual importance.

Output of the PHENOfdef sub-step in the pre-processing step of the metrics generation section are five raster datasets in PROBA-V 100 m resolution which are input in several other processing steps (Figure 23 and Figure 24):

1. Start date of Season 1 in the reference year (PROBA-V 100m SoS1);
2. End date of Season 1 in the reference year (PROBA-V 100m EoS1);
3. Start date of Season 2 in the reference year (PROBA-V 100m SoS2);
4. End date of Season 1 in the reference year (PROBA-V 100m EoS2);
5. A mask giving the number of Seasons in the reference year (PROBA-V 100m NoS).

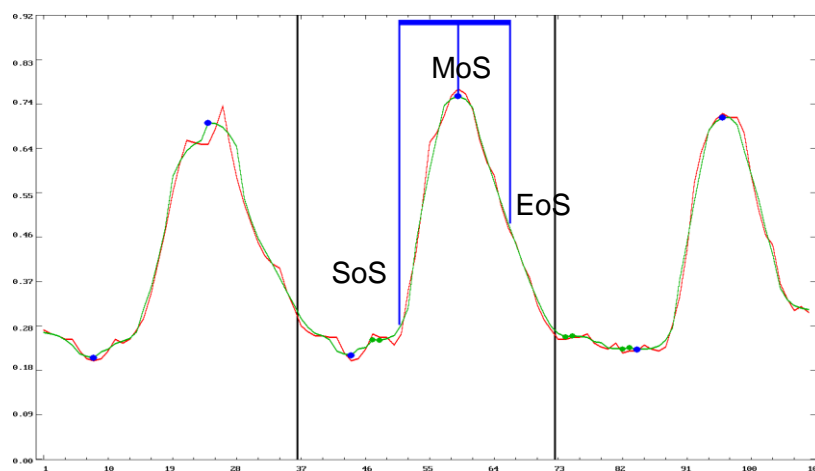


Figure 22: Example for the phenology product processing. Note: PROBA-V 100m NDVI values in red, smoothed curve in green, identified local maxima are shown as blue dots. Blue lines indicate important phenological metrics (SoS = start of season, MoS = mid of season, EoS = end of season) and vertical black lines indicate the reference year).

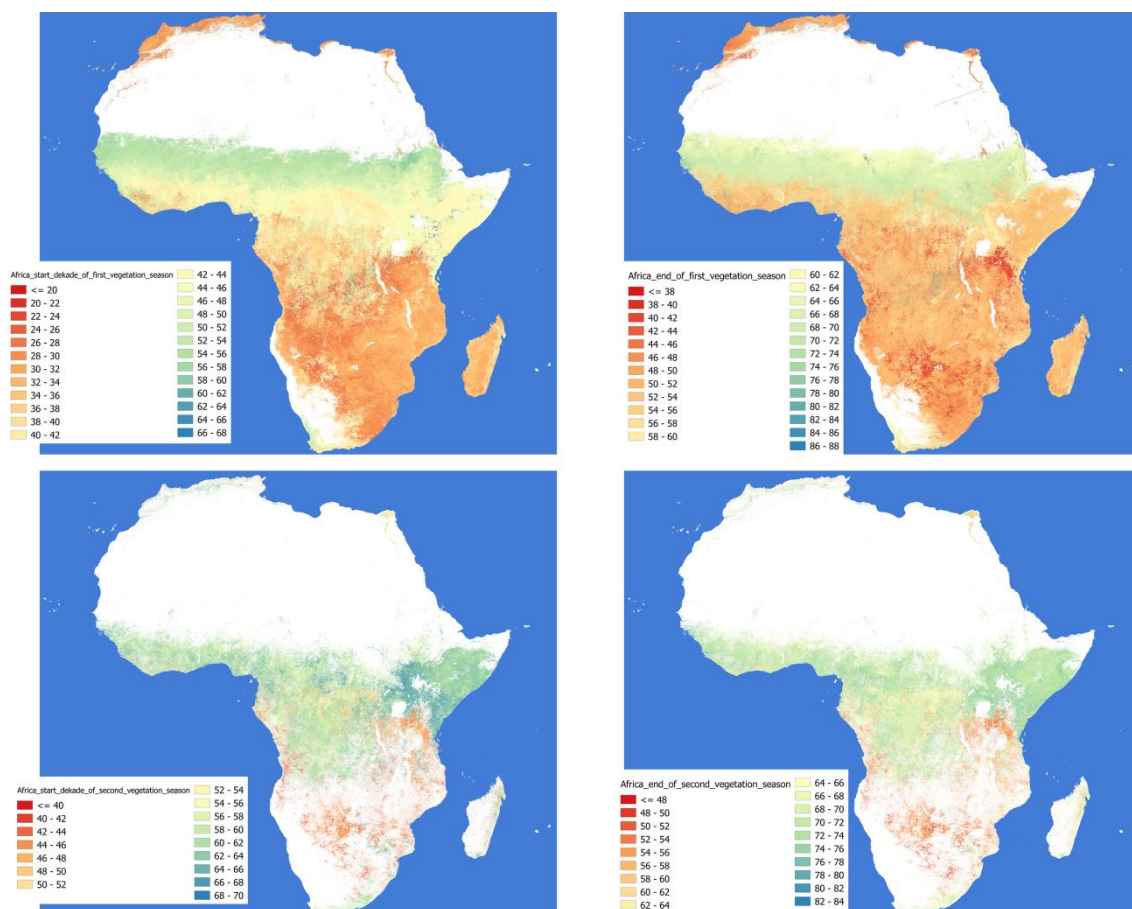


Figure 23: Phenology for Africa 2015 in PROBA-V 100m spatial resolution. (top-left) Start of season 1, (top-right) end of season 1, (bottom-left) start of season 2, (bottom-right) end of season 2. Note: the phenology dates are shown in decades starting with the year before the reference year (first decade of reference year = 37).

3.5.2.5 Generation of PROBA-V quality masks and phenology masks

Since we use data fusion algorithms in the CGLS LC100 workflow, we decided to generate quality flags for several processing steps which use the data fused time series profiles. Moreover, the metrics extraction algorithm needs phenology masks instead of derived dates. Overall, two quality layers and two phenology masks are generated in this processing sub-step.

The first quality layer is a binary mask and called time series gap mask (tGAPmask). The mask is used to evaluate the influence of the Kalman filling approach on the data quality by flagging these 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 time series of a pixel, and then to evaluate the length of this gap. Input into this approach is the red reflectance band of the PROBA-V MC5-TOC SM & outlier cleaned and gap filled 100 m time series as well as the five output layer of the PROBA-V phenology product (see section 3.5.2.4). The number of seasons and their start and end dates are used to calculate the longest concurrent data

gap in 5-daily increments for the reference year, the combined vegetation season (time of season 1 plus season 2), and the off-vegetation season (time in the reference year which didn't belong to one of the maximal 2 vegetation seasons). Then, the three gap lengths for each pixel are evaluated against adaptive thresholds. A pixel is flagged as a tGAP (or in other words, pixel with a higher change of uncertain filled data) when one of the following rules is fulfilled:

1. data gaps longer than 30 – 60 days in the combined vegetation season (threshold depends on the overall length of the vegetation season of a pixel divided by 3);
2. data gaps longer than 60 – 90 days in the off-vegetation season (threshold depends on the overall length of the off-vegetation season of a pixel divided by 3);
3. data gaps longer than 90 days in the reference year (important for pixel with no seasonality).

The second quality layer contains six parameters which were calculated during the tGAPmask generation, and which can be used as quality indicators and/or directly as a metric in the classification/regression process of the CGLS LC100 product workflow. The six raster layers are:

1. Number Of Valid Observations (NOVO) in the reference year
2. number of valid observations in the combined vegetation season
3. number of valid observations in the off-vegetation season
4. length of longest concurrent data gap in the combined vegetation season
5. length of longest concurrent data gap in the off-vegetation season
6. overall length of the vegetation season

Note: these metrics can reach a value up to 72. For pixels without seasonality, the vegetation season is set to the reference year but the overall length of the vegetation season is set to 0.

The two phenology masks were also already calculated during the tGAPmask generation. The first mask is a binary mask flagging all time stamps in the time series profile of a pixel which are within the combined vegetation season of the reference year. This mask is later used in the metrics extraction processing step. The second mask, also a binary mask, flags all pixels in the image which have one or two vegetation seasons. The names of all produced masks are:

- PROBA-V 100m tGAPmask (binary mask, 1 = pixel has high change of uncertain filled data)
- PROBA-V 100m NOVO container including:
 - PROBA-V 100m NOVOref_year (integer, max 72)
 - PROBA-V 100m NOVOon (integer, max 72)
 - PROBA-V 100m NOVOoff (integer, max 72)
 - PROBA-V 100m LocDGon (longest concurrent data gap)(integer, max 72)
 - PROBA-V 100m LocDGoff (longest concurrent data gap) (integer, max 72)
 - PROBA-V 100m LoVS (length of vegetation season) (integer, max 72)
- PROBA-V 100m MC5-PHENOMask (binary mask, 1 = time step is in vegetation season)
- PROBA-V 100m SEASONALITYmask (binary mask, 1 = pixel has vegetation season/s)

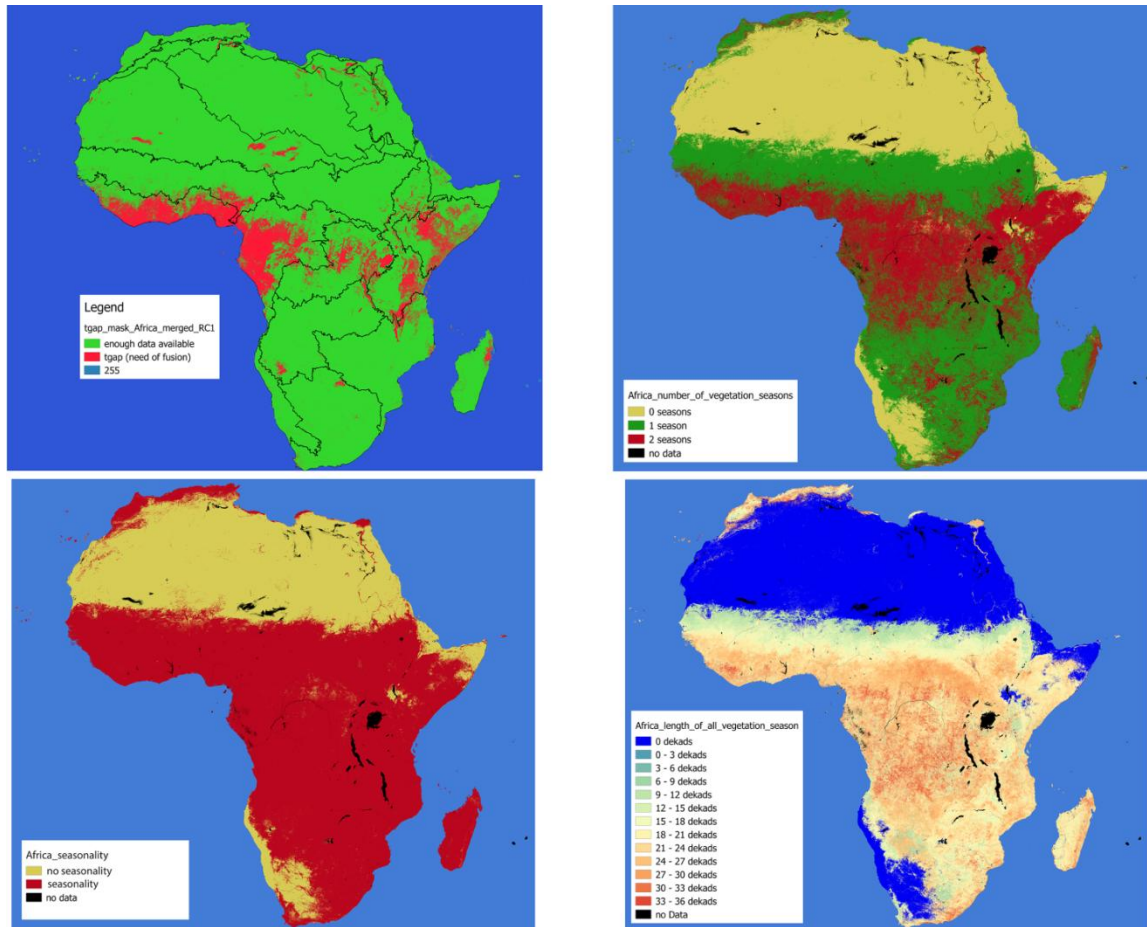


Figure 24: Quality masks. (top-left) tGAPmask showing pixels with high change of uncertain filled data during the data fusion step, (top-right) number of vegetation seasons in reference year, (bottom-left) SEASONALITYmask indicating if a pixel has a seasonality, (bottom-right) length of the combined vegetation seasons in decades.

3.5.3 Metrics extraction

The metrics extraction, the second processing step in the metrics generation section (Figure 18), 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 behaviours 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 MC5-TOC data fused 100m time series profiles (see section 3.4), the five PROBA-V 100m vegetation indices time series profiles generated in the pre-processing step (see section 3.5.2.1), the PROBA-V 100m HUE and VALUE time series profiles of the HSV colour transformation pre-processing step (see section 3.5.2.2), and the PROBA-V 100m MC5-PHENOMask showing for each pixel and time step in its profile if an observation belongs to the vegetation period (see section 3.5.2.5).

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 to choose. An easy solution to condense the information content is that statistical metrics are derived for the different 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 for the reference year of the time series profiles, then fit a harmonic model through the time series profiles and use the model parameters as attributes, and finally we calculate descriptive statistics for the vegetation and off-vegetation season from the harmonized time series profiles of the model. 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 input bands simultaneously.

3.5.3.1 Harmonic metrics

The first sub-step is the generation of the harmonic metrics. Thus, a harmonic model is fitted through each of the reflectance bands as well as the five additional generated VI's and the two colour transformed bands of the PROBA-V 100 m time series (overall 11 input time series profiles for each pixel in the image). The harmonic model is again based on the HANTS algorithm using a Fourier transformation (see section 3.3.2 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 77 harmonic metrics (7 model parameter for each of the 11 input bands) which are directly used in the classification/regression approach. Moreover, for each of the input bands, 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 25 shows an example for a harmonized time series profile of the NDVI.

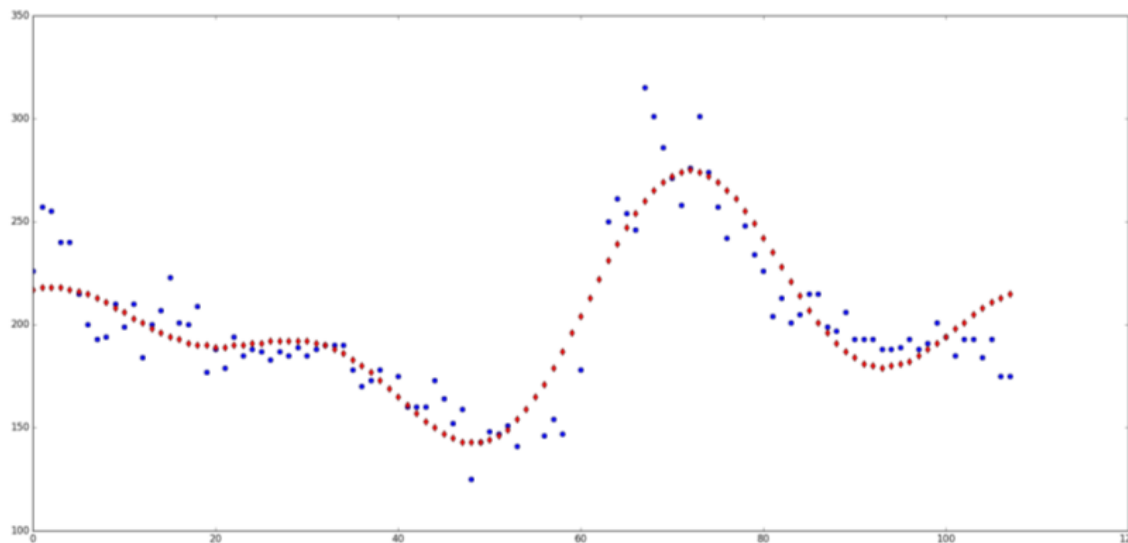


Figure 25: Example for a harmonized time series profile. The result of the HANTS model is shown in red dots, where the blue dots show the real observations.

3.5.3.2 Descriptive metrics

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

1. mean: for identifying the central value of a time series;
2. standard deviation: to quantify the amount of variation within the pixel values of a time series profile;
3. minimum: to identify extrema in the time series;
4. maximum: to identify extrema in the time series;
5. minimum-maximum range: to identify the interval in which the pixel values of a time series are located;
6. sum: to quantify the overall throughput of the time series in a given time length;
7. median: to identify the “middle” value which separates the higher half of the pixel values in the time series from the lower one;
8. 10th percentile: to identify the pixel value below which 10% of the pixel values of the time series can be found; and
9. 90th percentile: identify the pixel value below which 90% of the pixel values of the time series can be found.

The 9 descriptive statistics are independently calculated for each of time series profiles of the 11 input bands (4 reflectance bands, 5 VI's, and 2 colour transformed bands) for the reference year, the combined vegetation season and the off-vegetation season (Figure 26). Where for the reference year the “original” time series profiles (PROBA-V 100m MC5-TOC data fused reflectance bands and derived VI's plus HUE/VALUE) are used, the harmonised time series profiles (see section 3.5.3.1 – reconstructed 5-daily time series profiles using the phases and amplitudes of the

identified frequencies in the original time series) of the 11 input bands together with the PROBA-V PHENOMask are used to calculate the statistics for the vegetation and off-vegetation season. This is needed since the data fused PROBA-V 100m time series profiles (and therefore the derived products) can still contain data gaps which can have a low influence on the statistics of the reference year, but a high influence on the statistics of short vegetation seasons.

Output of this sub-step are overall 297 descriptive metrics (9 parameter for each of the 11 input bands multiplied by 3 time intervals) which are directly used in the classification/regression approach.

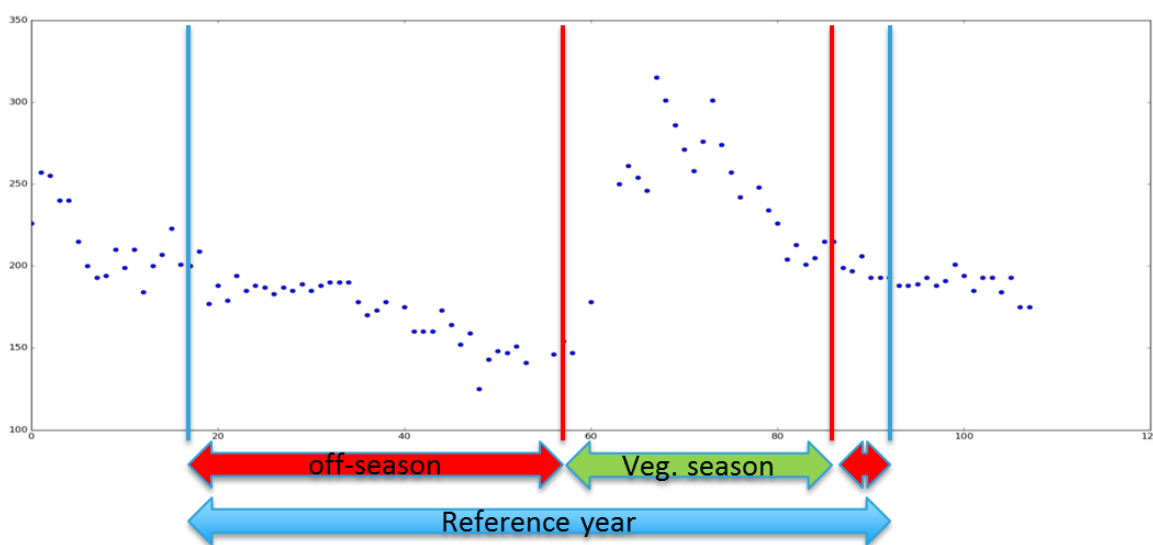


Figure 26: Example for splitting the time series profiles in reference year, vegetation season and off-vegetation season. Blue line show the identified reference year in the time series profile and red lines the time steps to separate vegetation from off-vegetation season. Note: shows an example with only one vegetation season in the reference year.

3.5.3.3 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 3.5.3.2) parameter for the reference year part of the time series profiles, where the textural parameter is calculated for each of the 11 input bands 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 11 textural metrics (one for each of the 11 input bands) which are directly used in the classification/regression approach. Figure 27a shows an example for textural metric generated for a whole tile.

3.5.4 Post-Processing

The last processing step in the metrics generation section (Figure 18), the post-processing, warps up all results of the metrics extraction step and adds additional metrics generated during the pre-processing step. The huge amount of metrics data is easier to handle when all metrics for each pixel in an image can be found in one file, therefore we decide to generate virtual raster datasets (VRT) out of the single metrics files. The VRT file is an 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.

Next to the metrics generated in the metrics extraction sub-step, the additional metrics include the phenological parameters for start and end of season (PROBA-V 100m SoS1, PROBA-V 100m SoS2, PROBA-V 100m EoS1, PROBA-V 100m EoS2), the seasonality mask (PROBA-V 100m SEASONALITYmask) index indicating if a pixel has a seasonality overall, and the length of the vegetation season (PROBA-V 100m LoVS). Moreover, the Water Bodies Potential Mask (PROBA-V 100m WBPM) is used as a topographic parameters/metric indicating if a pixel could be possible a water body.

Overall, 392 metrics (7 harmonic metrics plus 27 descriptive metrics plus 1 textural metric for the 11 time series profiles (4 reflectance bands, 5 vegetation indices, 2 colour transformed bands) plus 7 additional metrics) are combined in the VRT file and are input in the classification/regression section of the automated processing chain. Figure 27 shows a visualization of some selected metrics.

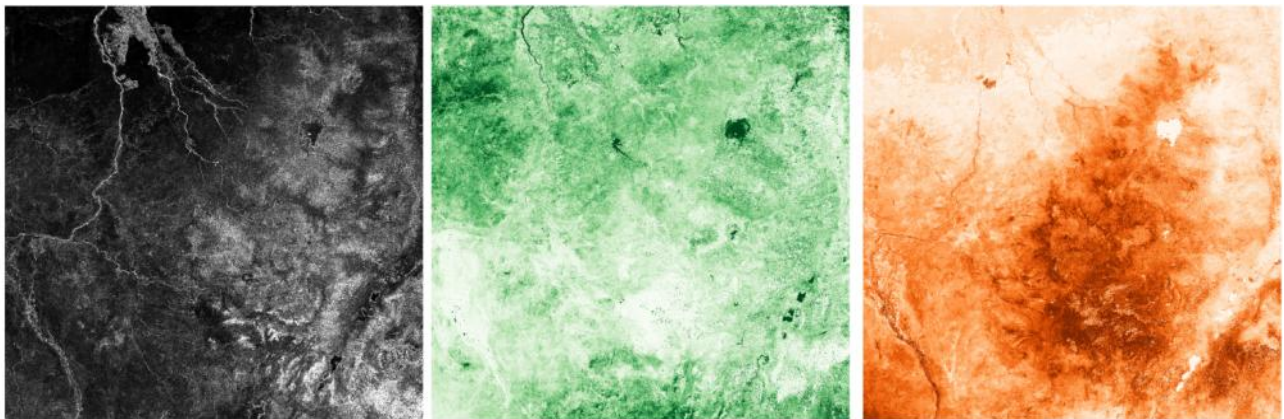


Figure 27: Examples of the 392 derived metrics for tile X18Y06 (Nigeria). Left: texture metric (the lighter the color the more homogeneous is the pixel compared to its surrounding pixels), middle: standard deviation of the Structure Intensive Pigment Index for the vegetation season (the greener the color the higher is the SIPI change within the vegetation season), right: sum of the Enhanced Vegetation Index for the reference year (the redder the colour the more vegetation).

3.6 ANCILLARY DATASET PRODUCTS

In order to include ancillary datasets in the CGLS LC100 product workflow (Figure 1), the external datasets have to be pre-processed. This includes the resampling to PROBA-V 100 m resolution and/or the complete generation of ancillary datasets by using third-party algorithms. The ancillary datasets are not only input in the LC map generation process, but also in the training data generation and the classification/regression section of the workflow. Overall, four ancillary products have to be generated: a buffered ecozone vector layer product, the African shoreline mask, the urban area mask, and the two water body products.

3.6.1 Ecozone Buffering product

In order to group EO data for faster processing or adaptation of algorithms to specific regions, we use the global ecological zone (GEZ) dataset for 2010 of the FAO (FAO, 2012) (see section 3.2.3). This vector layer can be directly used in several processing steps, but for the training data generation we need a buffered version in order to reduce border effects during the classification/regression. Thus, each ecozone in the GEZ 2010 layer was buffered with a 2 degree buffer using the QGIS software (open source Geographic Information Software package).

3.6.2 Shoreline product

In order to distinguishing open land water pixels from open sea water pixels in the post-classification process, we need a shoreline mask. This mask is generated out of the USGS African shoreline vector layer (Sayre et al., 2013) (see section 3.2.4). The USGS shoreline vector layer was rasterized and then resampled to the PROBA-V 100 m spatial resolution using the QGIS software.

3.6.3 Urban 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 PROBA-V urban mask was generated through the combination of DLR's Global Urban Footprint Plus layer (GUF+) for 2015 (Marconcini et al., 2017a, Marconcini et al., 2017b) and JRC's Global Human Settlement Layer (GHS) for 2014 (Pesaresi et al., 2015). Both datasets are explained in detail in section 3.2.5 and 3.2.6, respectively.

Both raster layers had to be resampled to the PROBA-V 100 m spatial resolution in a first step. And secondly, the GUF+ and GHS layers have been fused whereby missing urban areas in the GUF+ layer have been incorporated from the GHS layer (mainly needed for islands).

3.6.4 Water 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 100m water body detection algorithm (called WetProducts generation) (Figure 28). The main change compared to Bertels et al. (2016) algorithm is the incorporation of JRC's maximum water extent and water seasonality 2014-2015 layers (Pekel et al., 2016) (see section 3.2.7). The maximum water extent (provides information on all the locations ever detected as water over the Landsat data archive period) and the water seasonality layer had to be resampled from the 30 m Landsat resolution to the PROBA-V 100 m spatial resolution in a pre-processing step.

As shown by Figure 28, the water body detection algorithm consists of three major steps. The SWIR, NIR and RED bands of the 5-daily composites (MC5) for the reference year 2015 are transformed to HUE, SATURATION and VALUE using a RGB to HSV colorimetric transformation (see section 3.5.2.2). Subsequently, the application, per pixel, of specific threshold values on HUE, VALUE and NDVI while taking into account the Maximum Water Extent Mask (MWEM) allows water body detection. All threshold values were empirically defined. Subsequently the Land/Sea mask, the Shoreline Mask and the water seasonality layer are used to re-imprint the permanent water body pixels. Finally, a water occurrence layer is calculated based on the WB detection statistics.

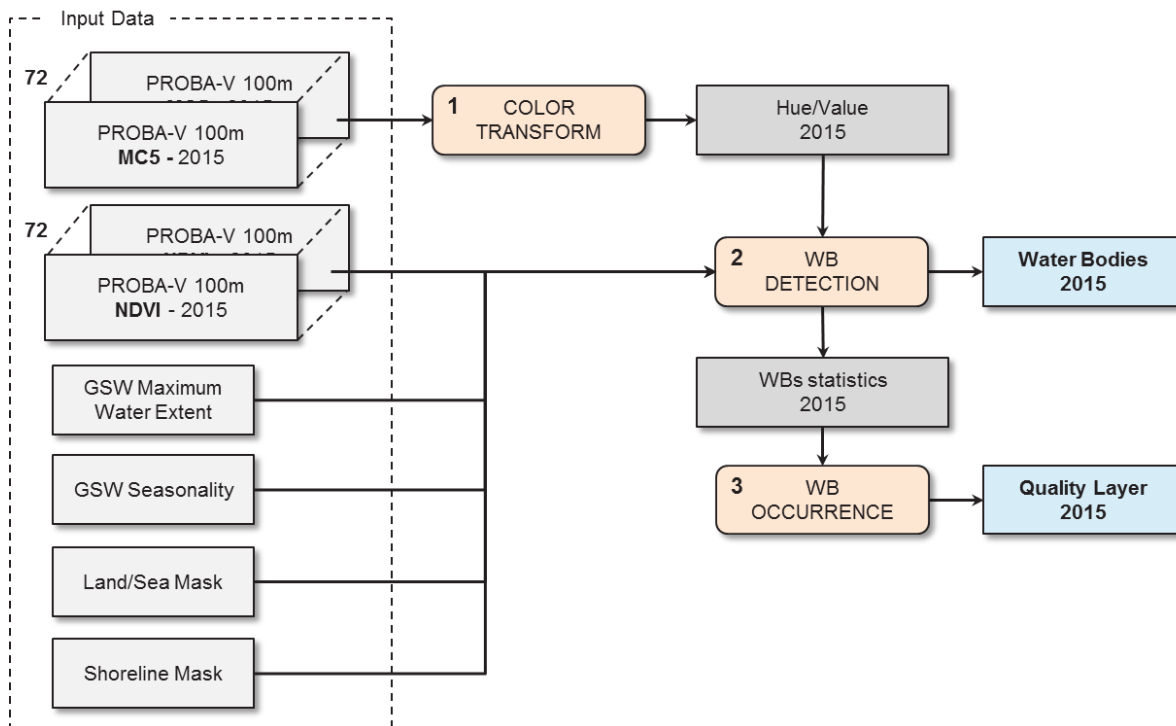


Figure 28: General overview of the Water Bodies Detection Algorithm.

The following datasets, generated in previous processing steps, are needed for the PROBA-V 100m water body detection:

- PROBA-V 100m MC5-NDVI time series profiles (see section 3.5.2.1),
- PROBA-V 100m MC5-HUE time series profiles (see section 3.5.2.2),
- PROBA-V 100m MC5-VALUE time series profiles (see section 3.5.2.2),
- Maximum Water Extent product from the JRC Global Surface Water (GSW) dataset (see section 3.2.7),
- Water Seasonality 2014/2015 product from the JRC GSW dataset (see section 3.2.7),
- the PROBA-V 100m Land/Sea Mask (see section 3.2.2), and
- the PROBA-V 100m Shoreline Mask (see section 3.6.2).

Figure 29 gives a schematic overview of the decision tree used for the detection of the three different water body classes, i.e. permanent WBs, wetlands and temporary WBs. First, a check on valid data pixels is performed. Both the HUE and VALUE pixels must have values greater than zero and the NDVI pixels may not have a 'No data' value. Subsequently, thresholds on HUE, VALUE and NDVI are applied to detect water body pixels for which their statistics are being updated, i.e.

- (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.

The 'total N° of valid observations' (TotalObs) is increased with each MC5 for each valid pixel; the 'total N° of water detections' (TotalDetect) on the other hand is increased only for each detected water body pixel. The 'maximum N° of consecutive detected water body' (MaxFreq) will hold the longest consecutive period that a pixel was detected as water body. The 'overall water occurrence' (OverallPerc) is calculated as the ratio TotalDetect/TotalObs. Once the statistics are calculated, the final water body detection is done. When the 'overall water occurrence' is greater than 90% and the pixel has at least eleven valid observations unless the shoreline or land/sea mask indicate them as water or the seasonality layer indicates water over the whole year, the pixel is indicated as 'permanent water body'. When the 'overall water occurrence' is greater than 5% and the 'total NDVI' (which is the sum of the NDVIs for all MC5s) is greater than 17.5, the pixel is indicated as 'Wetland', when the 'total NDVI' is less than 17.5 the pixel is indicated as a 'temporary water body'.

The obtained water classes, i.e. 'Permanent water body', 'Wetland' and 'Temporary water body' are subsequently used to imprint them in the final land cover maps. The water body statistics are outputted as a separate layer.

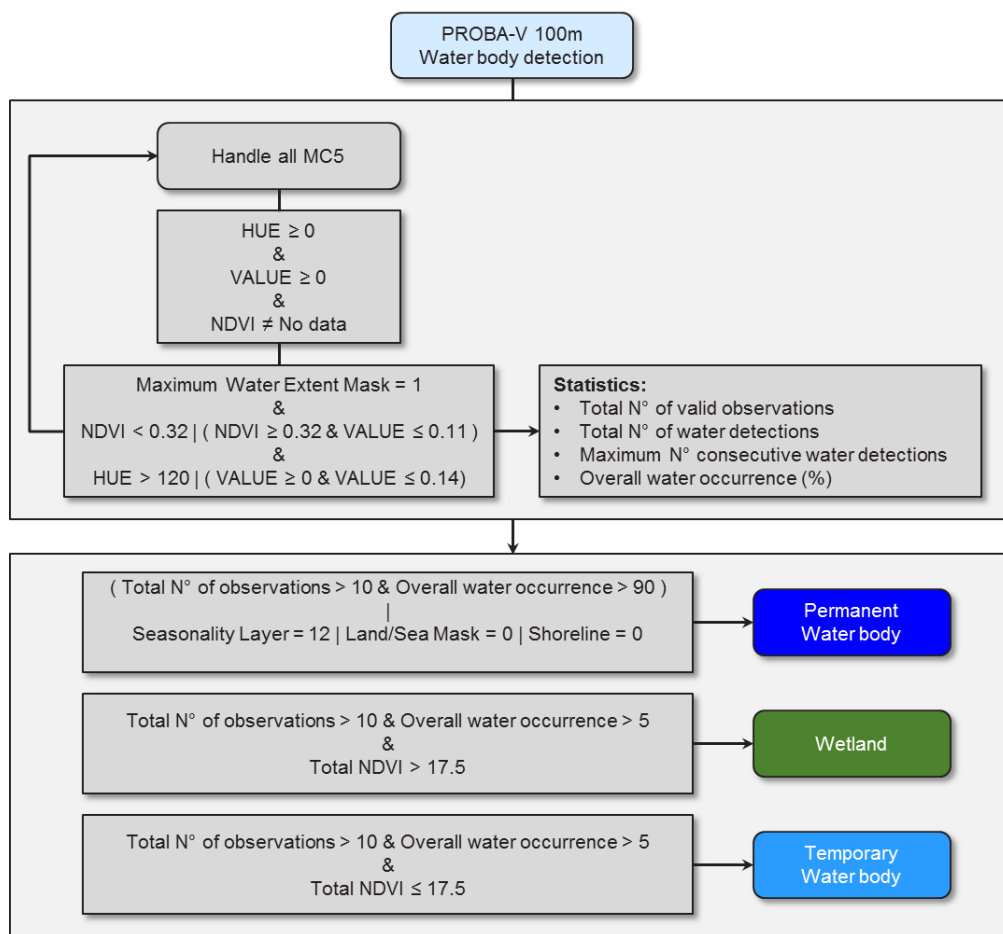


Figure 29: Decision tree of the water body detection algorithm.

3.7 TRAINING DATA GENERATION

Training data has been collected through the Geo-Wiki engagement platform. A new branch of Geo-Wiki (<http://geo-wiki.org/>) was developed for collecting reference data at the required resolution and grid (PROBAV-100m 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 10m x 10m each. Using visual interpretation of the underlying very high resolution imagery, experts (group of people trained by 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 [CGLOPS1_URD_LC100m].

The 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 100m imagery.

In total, the experts have classified almost 24,000 unique locations [CGLOPS1_TrainingDataReport_LC100m]. 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 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 3.8.3), but instead of directly removing suspicious training points, the training points were checked by visual interpretation. 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 circa 20,000 sample sites (Figure 30).

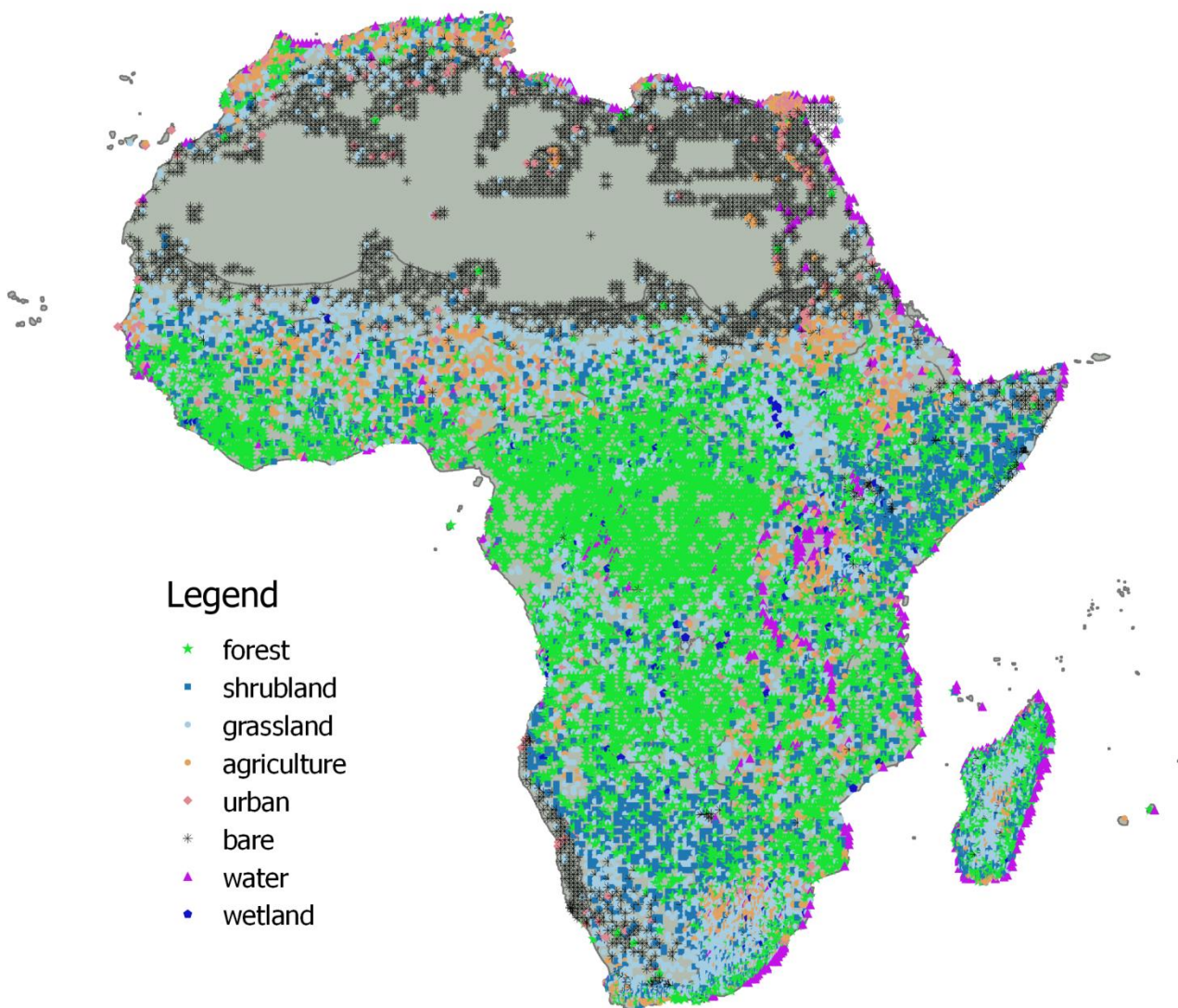


Figure 30: Training points with discrete land cover class for the CGLS LC100 product workflow.

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 buffered GEZ 2010 layer (see section 3.6.1), since the 2 degree buffer around each ecological zone is used to reduce border effects during the classification/regression processing step. Moreover, the tGAPmask (see section 3.5.2.5) was applied to ensure that all training points have sufficient time series data. In detail, training points in locations with long data series gaps (longer than 30 – 60 days in the vegetation season, or longer than 60 – 90 days in the off-vegetation season, or longer than 90 days in the full reference year; see section 3.5.2.5) are more prone to introduced noise, since the long gaps are filled by the data fusion approach (see section 3.4), and therefore removed from further processing.

In the next step, the PROBA-V time series metrics (see section 3.5.4) were extracted for the geographic location of each training point. Output is a complete training data set for each ecological zone in the GEZ 2010 layer containing the land cover ground reference information and the spectral information of the EO time series data condensed in the form of the metrics.

3.8 CLASSIFICATION / REGRESSION

3.8.1 Overview

In order to adapt the classification/regressor algorithms to continental patterns, the classification/regression of the data is carried out in data subsets (called ecozones) using the FAO ecological zones dataset (see section 3.2.3). 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 ecozone are screened for inter-class outliers for each ecozone independently. Next, the best bands (metrics) for each ecozone are identified via an all relevant feature selection process in order to optimize the classification/regression algorithms for each ecozone (see section 3.7). Moreover, during the optimization phase also the classification/regression algorithm parameter are optimized for each ecozone 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 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).

3.8.2 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 31). 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).

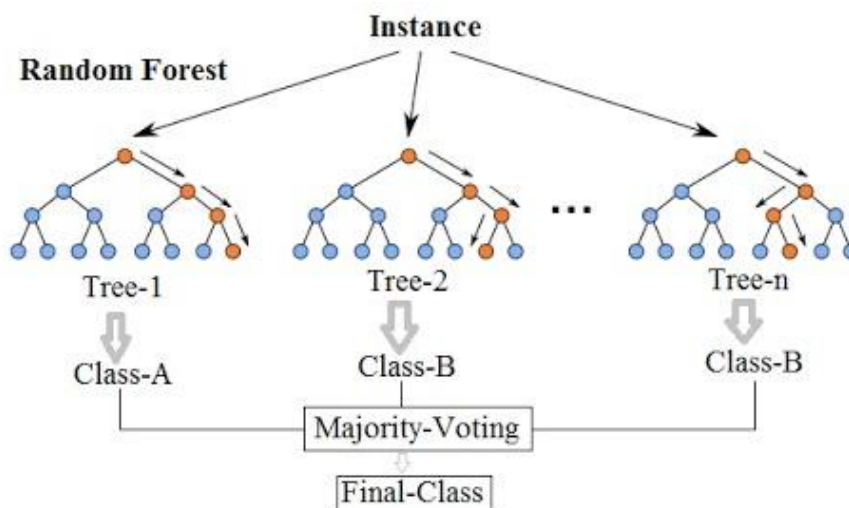


Figure 31: The Random Forest classifier principle. Image by Niraj (2016).

3.8.3 Training data and classifier/regressor optimization

Since we use ecozones to subset the EO data, the RF classifier/regressor algorithms should be optimized for each ecozone trainings 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 ecozone.

The training data used in the CGLS LC100 classification was selected by experts (see section 3.7). Nevertheless, inter class confusion might still be present between individual ground reference data (training points). In a first step, ground reference data analysis is performed individually for each ecozone-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 ecozone 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. This results in a square N by N matrix, where N is the total number of training point spectra (Figure 32).

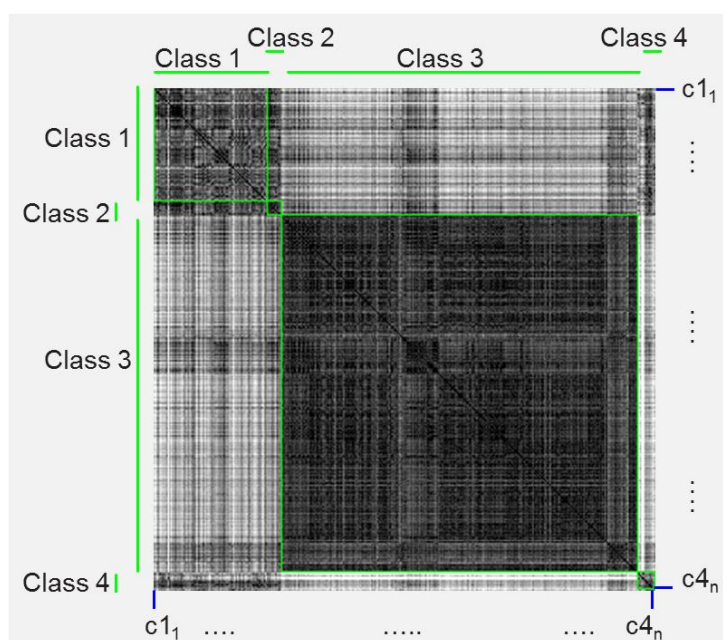


Figure 32: Rule matrix calculated for the total of N training points which belong to four different LC classes. For each training point the RMSE value is calculated with all the other training points, i.e. c_{11} with c_{11} , c_{12} , c_{13} till c_{4n} , subsequently c_{12} with c_{11} , c_{12} , c_{13} till c_{4n} and so on.

3. The rule matrix is used to exclude inter-class outliers. E.g. when the median RMSE value of “training point c_{11} calculated with the remaining spectra of its own class (Class 1)” is greater than the median RMSE value of “training point c_{11} calculated with all the spectra of the other classes” than training point c_{11} 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 with an impact score over 50% are removed from further processing.

In a second step, the best metrics to separate the LC classes are selected for each ecozone-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 10:

$$S_{AB} = \frac{nA + nB}{totA + totB} \quad \text{Equation 10}$$

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. The best metrics selection is used to sort the 392 metrics (see section 3.5.4) by their separability (highest to lowest) for each ecozone-based training dataset.

In order to overcome the Hughes phenomenon (Hughes, 1968) due to the fact that the 392 metrics include a high information redundancy, an all-relevant feature selection approach is used to select those best metrics for each ecozone-based training dataset which solve the classification/regression problem best. This step is needed since the identified best metrics can still be redundant (carry the same information content). The aim of this processing step is to identify those of the sorted best metrics which are non-redundant. 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 used 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 ecozone-based training dataset of the maximal 50 best metrics which are non-redundant.

The last optimization step is the hyper-parameter search for each ecozone-based training dataset. In detail, the optimal parameters to identify the best machine learning RF model created by the training data for each ecozone 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 ecozone-based training data set.

3.8.4 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 was conducted for each ecozone independently using the GEZ 2010 dataset to split up the input data (generated metrics for the PROBA-V time series), and

using the ecozone-specific generated training datasets and hyper-parameters. 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 ecozone with different settings have been carried out:

1. “pure class” scenario (CL1): in this scenario, only training points with a cover percentage over 95% in the bare, grassland, shrub, forest, 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 minimum of 95% of forest cover).
2. “discrete class” scenario (CL2p5): 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 scenarii.

3.8.5 Scenario-based Regression

A novelty of the CGLS LC100 product is the generation of vegetation continuous fields that provide proportional estimates for vegetation cover of trees, herbaceous vegetation, shrub and bare ground. The input data are the cover fractions collected for all training points which are used in a Random Forest regression. The RF regression was conducted for each of the main land cover types (forest, shrub, grassland, agriculture, bare) and ecozones independently using the GEZ 2010 dataset and using the ecozone-specific generated training datasets and hyper-parameters. Overall five regression scenarios for each ecozone have been carried out:

1. Forest (R1): the forest cover percentages of training points are used in the regression model.
2. Shrub (R2): the shrub cover percentages of training points are used in the regression model.
3. Herbaceous vegetation (R3): the grassland cover percentages of training points are used in the regression model.
4. Bare (R4): the bare cover percentages of training points are used in the regression model.

5. Agriculture (R5): the agriculture cover percentages of training points are used in the regression model. **Note:** the agriculture cover fraction map is only used to create an agriculture mask and will be not delivered as a cover fraction layer in the final product.

3.9 COVER FRACTION LAYERS GENERATION

3.9.1 Regression post-processing

The last processing step in the generation of the cover fraction layers for forest, shrubs, herbaceous vegetation, and bare as part of the CGLS LC100 product is the regression post-processing (Figure 1). Input are the cover fraction layers, indicating the proportional estimates of land cover for the specific land cover type, which were generated by the scenario-based RF regression step (see section 3.8.5). The main processing step is a linear normalization for pixels with a cover fraction sum of more than 100 % in the combined five regression results (R1-R5). In detail, pixels with an overall percentage over 100 % in the combined cover fractions result of forest, shrub, grassland, agriculture and bare are proportional scaled that their sum is 100 %. Moreover, the permanent water body mask as part of the PROBA-V 100m WetProducts (see section 3.6.4) and PROBA-V 100 m urban mask (see section 3.6.3) are incorporated by setting the pixel values for all cover fractions to 0 % in as permanent water body or urban area identified pixel locations.

3.9.2 Metadata

Finally, metadata attributes compliant with version 1.6 of the Climate & Forecast conventions (CF V1.6) and the colorbars translating the vegetation continuous fields code into the legend are injected. Overall four final cover fraction layers are provided:

1. LC100-COV-FOREST
2. LC100-COV-SHRUB
3. LC100-COV-GRASSLAND
4. LC100-COV-BARE

As already mentioned in section 3.8.5, a cover fraction layer for the agriculture class is not provided and only used to generate an agriculture mask for the LC map generation process. Main reason for this decision is the high confusion of the agriculture LC class with the grassland and shrubland LC class in the lower cover percentage value range (0 – 25 %).

Figure 33 shows the four provided CGLS LC100 cover fraction layers for the land cover classes forest, shrub, herbaceous vegetation and bare. The shown colours for the cover fractions are the ones integrated as RGB colour bars in the metadata of the products.

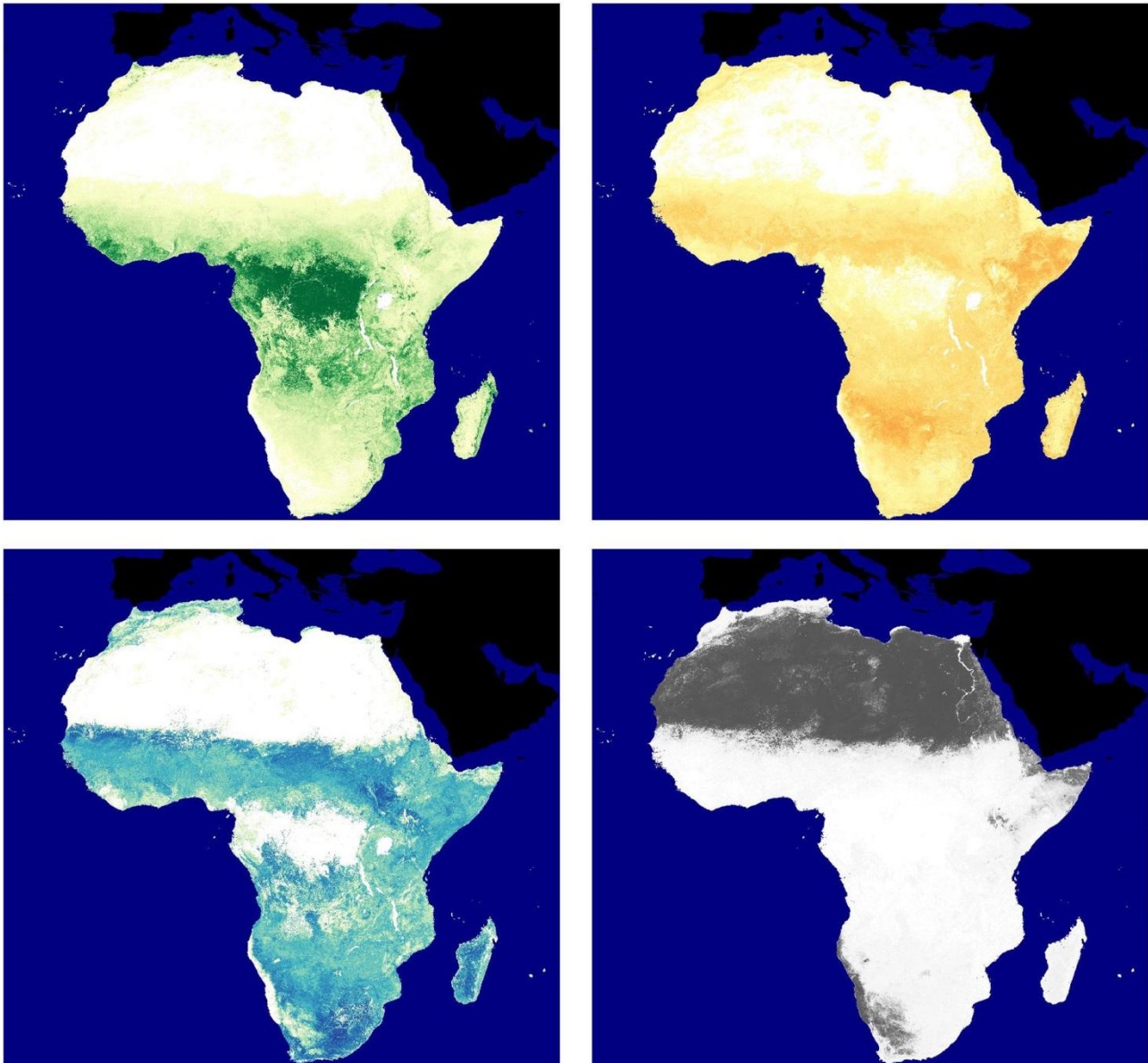


Figure 33: The cover fraction layers for the forest, shrub, herbaceous vegetation and bare land cover classes of the CGLS LC100 product for Africa 2015 (shown at continental scale).

3.10 LAND COVER MAP GENERATION

3.10.1 Overview

In order to generate the CGLS LC100 discrete map product, three processing steps are needed (Figure 34). Input are the results of the PROBA-V 100m time series classification and regression results, the PROBA-V 100m NOVO container holding the number of valid observations in the reference year (NOVOref_year) layer, the PROBA-V 100m WetProducts layer, and the external urban and shoreline masks. In the first step, the pre-processing, the input dataset are assembled

and additional information layers are generated. Next, expert rules are applied to combine the existing knowledge represented by the ancillary datasets with the classification and regression results. The last step in the discrete map product generation is the infusion of metadata.

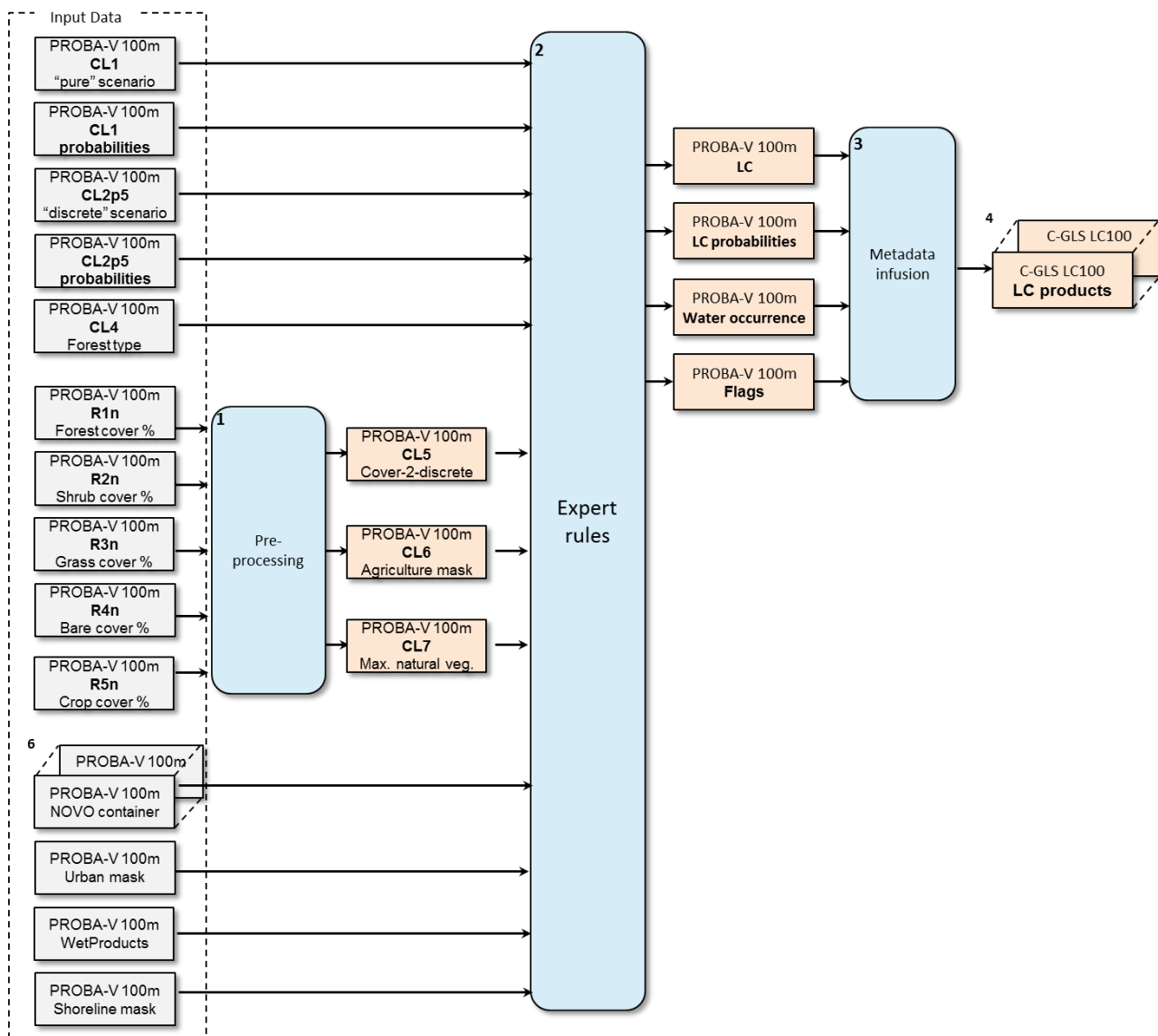


Figure 34: General overview of the Land Cover map generation section in the CGLS LC100 product workflow.

3.10.2 Assembling and generation of the input datasets for the expert rules

Overall 12 datasets are input in the expert rules step which combines the scenario-based classifications and regressions results with the ancillary datasets. Most of the input datasets were

already generated in the prior CGLS LC100 product workflow sections. Only three input layer have to be generated in this pre-processing step.

First, in order to incorporate the vegetation cover fraction layers (see section 3.9) in the LC map generation process, a discrete map is generated by applying the training data rules [CGLOPS1_TrainingDataReport_LC100m] on the normalized forest, herbaceous vegetation, shrub, bare ground and agriculture cover fraction layers. In detail, during the training data collection (see section 3.7 and [CGLOPS1_TrainingDataReport_LC100m]), 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) [CGLOPS1_TrainingDataReport_LC100m]. 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. Moreover, since only the cover fraction layers for vegetation are produced during the regression step, the LC map only show natural vegetation and agriculture. This discrete LC map is called Cover-2-discrete (CL5) during the expert rules.

Next, the normalized agriculture cover fraction layer (R5n) was used to generate an agriculture mask by applying a threshold. The threshold was empirically defined and set to 30 %. The masked is called CL6 (agriculture mask) in the expert rules. The last input layer which is generated is a maximum natural vegetation mask. Therefore, the normalized cover fraction layers for forest, herbaceous vegetation, shrub and bare ground are processed and, for each pixel, the LC class with the maximum cover fraction is assigned. This input layer is called CL7 (max. natural vegetation) in the LC map generation process (Figure 34).

The following datasets are the final input in the expert rules to generate the CGLS LC100 discrete map:

1. Random Forest classification result of the “pure class” scenario (CL1);
2. Predicted class probability layer of the Random Forest classification result of the “pure class” scenario (CL1_probabilities);
3. Random Forest classification result of the “discrete class” scenario (CL2p5);
4. Predicted class probability layer of the Random Forest classification result of the “discrete class” scenario (CL2p5_probabilities);
5. Random Forest classification result of the “forest type” scenario (CL4);
6. Discrete map generated from the five normalized cover fraction layers (CL5);
7. Agriculture mask generated from the normalized agriculture cover fraction layer (CL6);
8. Maximum natural vegetation mask generated from the normalized cover fraction layers for forest, herbaceous vegetation, shrub and bare ground (CL7);
9. “number of valid observations” mask showing pixels with no PROBA-V 100 m observations in the whole reference year (novo mask) (see section 3.5.2.5);
10. PROBA-V 100m urban mask (urban mask) (see section 3.6.3);
11. WetProduct layer including the permanent water body mask, temporary water body mask, and herbaceous wetland mask (WetProducts) (see section 3.6.4); and
12. PROBA-V 100m Shoreline mask (Shoreline).

3.10.3 Map Generation

For combining all classification scenarii and ancillary datasets into a final LC map, expert rules implemented as a decision tree are applied pixel-based on the input data. Overall 10 expert rules have been identified and established (Figure 35 to Figure 43). The predicted class probabilities are used as thresholds in the decision tree in order to generate the 18 class discrete map product, but are not the only deciding factor. The 18 discrete classes in the CGLS LC100 product are shown in Table 5.

Table 5: The 18 discrete classes of the CGLS LC100 discrete product. Note: final classes are shown in bold.

Forest	Closed forest	evergreen needleleaf closed forest
		evergreen broadleaf closed forest
		deciduous needleleaf closed forest
		deciduous broadleaf closed forest
	Open forest	evergreen needleleaf open forest
		evergreen broadleaf open forest
		deciduous needleleaf open forest
		deciduous broadleaf open forest
shrubs		
herbaceous vegetation		
cropland		
urban		
bare/sparse vegetation		
snow/ice		
water	permanent water bodies	
	temporary water bodies	
	open sea	
herbaceous wetland		

Before the first expert rule, shown in Figure 35, is applied on the input data, two output dataset are initialized with a “no data” value – the PROBA-V 100m LC and PROBA-V 100m LC probabilities

datasets. Afterwards the decision tree is applied pixel-based on all 12 input datasets. Expert rule I (imprint agriculture mask rule) imprints the agriculture mask within the African land masses, as well as sets the class probabilities for all agriculture pixels. Since the agriculture mask was generated from the normalized agriculture cover fraction layer, no class probabilities are directly available. Thus, the class probabilities are extracted from the CL1 and CL2p5 classification scenarios – with higher priority on the CL1 scenario – and where no probability can be extracted is set to the “no data” value (Figure 35).

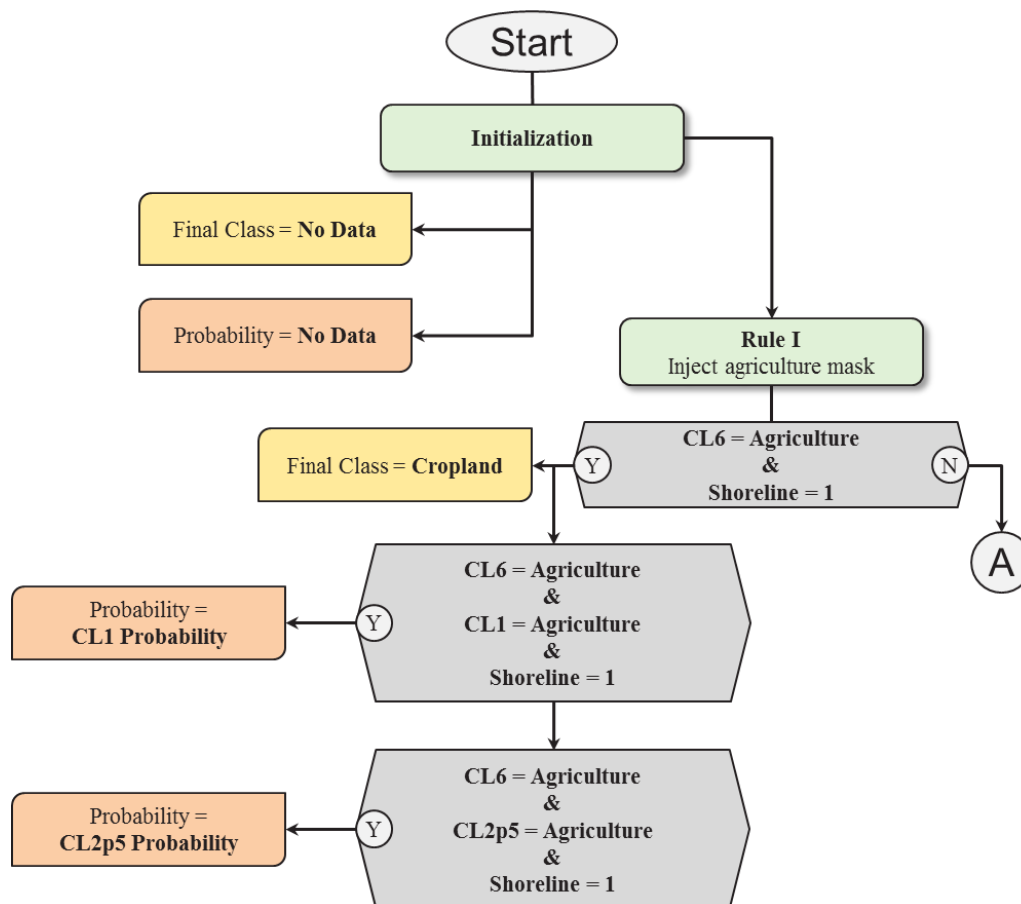


Figure 35: Expert rule I of the CGLS LC100 discrete map generation process.

Remaining “nodata” pixels in the PROBA-V 100m LC map are input in expert rule II (pure class rule) (Figure 36). Therefore, only pixels with a predicted class probability over 90 % and with no agriculture LC class are used from the CL1 classification scenario. These rules can be seen as the pure endmember classification of the natural vegetation. The CL1 class and probability value is copied to the final dataset (Figure 36). Next, the 50% rule (expert rule III) is applied on the remaining “nodata” pixels. Pixel locations of the CL2p5 classification scenario with a predicted class probability smaller 90 % in CL1 scenario and a predicted class probability bigger equal 50 %

in CL2p5 scenario and not being an agriculture class are written to the output dataset. Moreover the corresponding class probability for these pixels is copied (Figure 36).

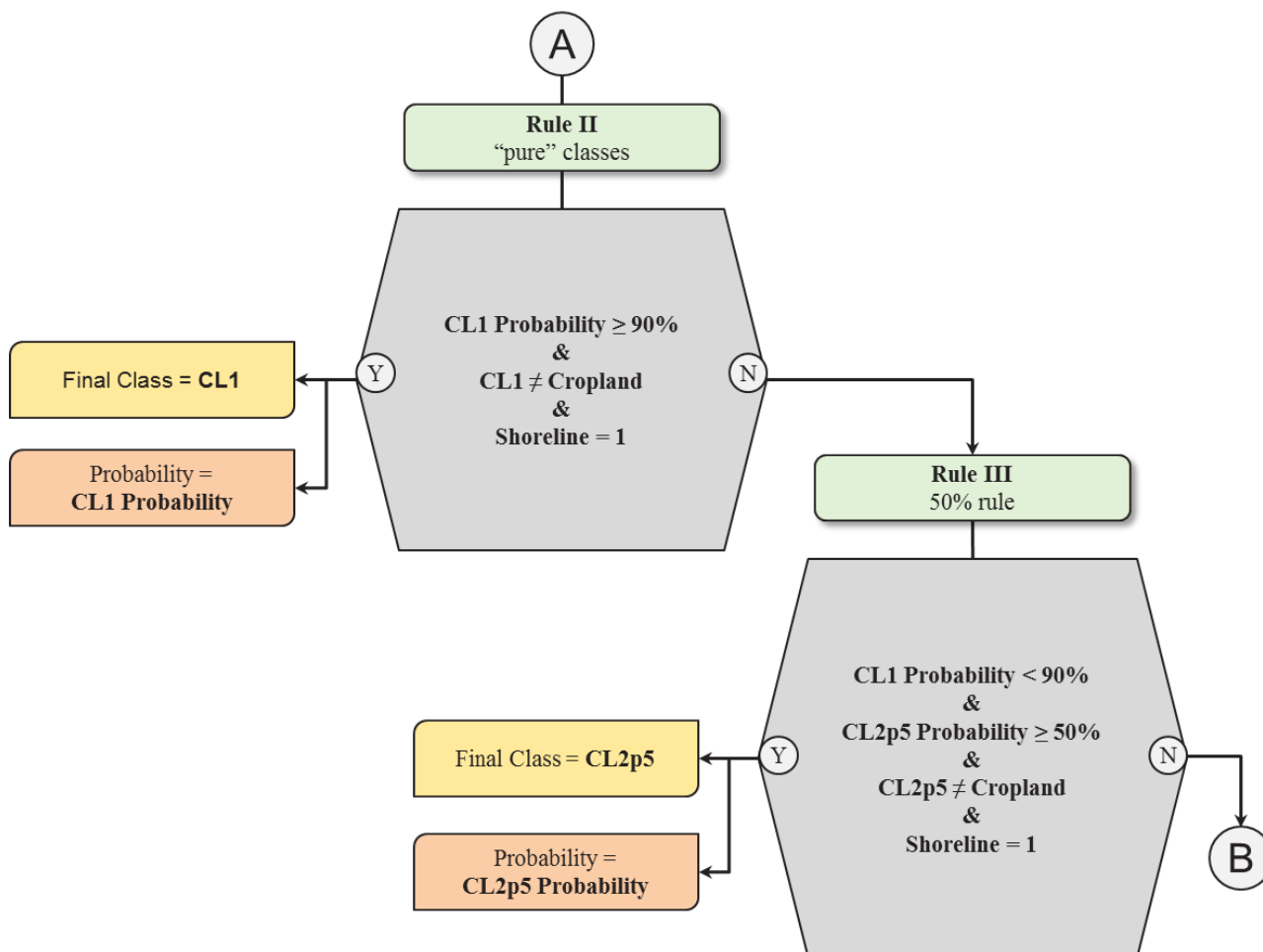


Figure 36: Expert rules II and III of the CGLS LC100 discrete map generation process.

Expert rule IV deals with the CL5 – the discrete LC map generated from the normalized cover fraction layers by applying the training data rules - and is split into two parts (Figure 37). In part a, remaining “no data” pixels in the output dataset which have a predicted class probability smaller 90 % in the CL1 scenario and a predicted class probability smaller 50 % in the CL2p5 scenario and where the CL5 class equals the pixel class in the classification scenario CL2p5 plus are not agriculture in CL2p5, are copied to the output dataset together with the corresponding class probability. Part b of the rule, applies rule IVa respectively to CL1 (Figure 37). Thus, the CL5 LC map was mainly used as an extra decision rule for pixels with a low class probability – a therefore high chance of a misclassification - in the classification scenarios CL1 and CL2p5.

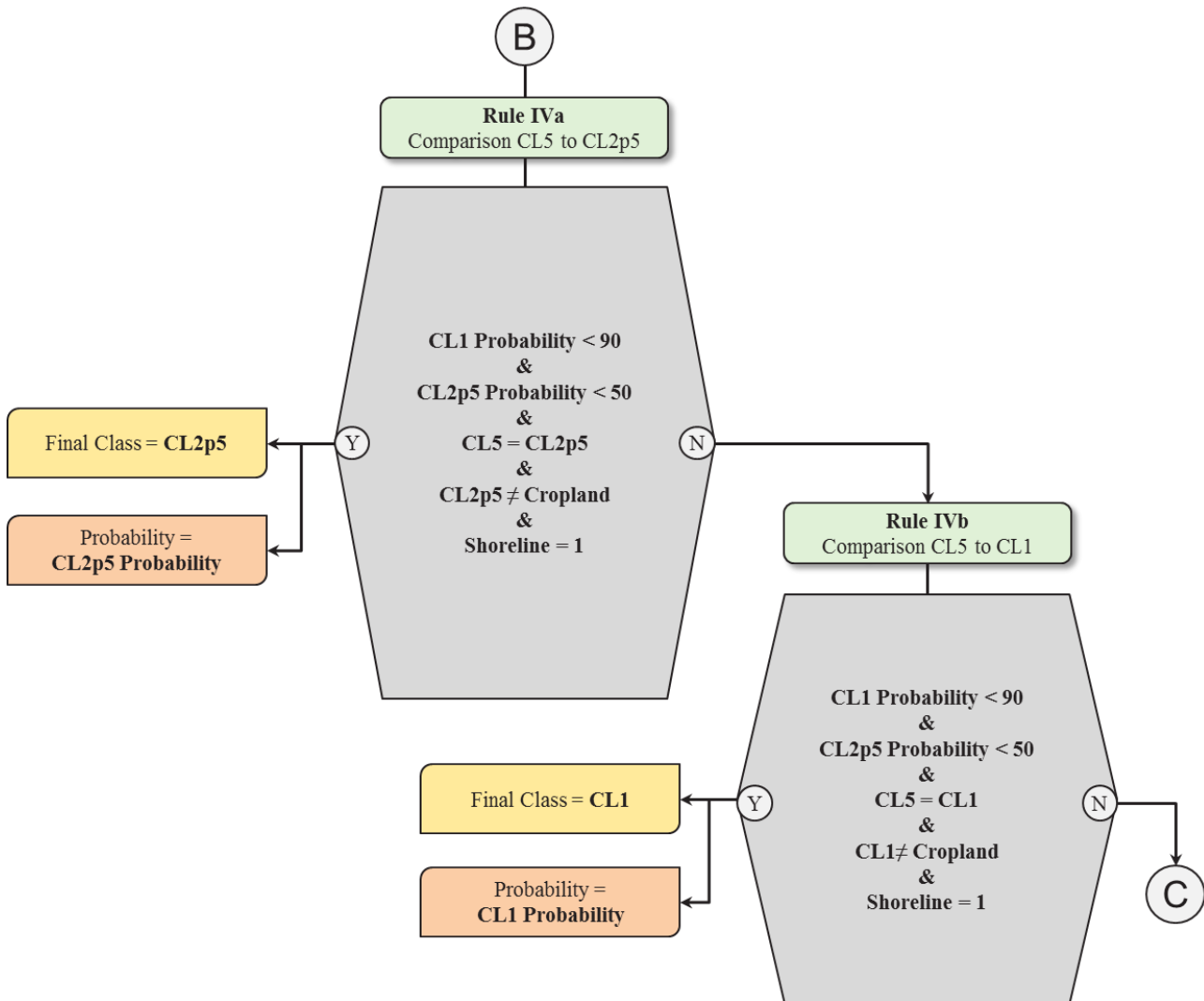


Figure 37: Expert rules IV of the CGLS LC100 discrete map generation process.

After applying expert rule IV, most of the land masses pixels which are not urban or inland water bodies got their final LC class already assigned. Expert rule V was developed to fill all remaining land masses pixels which don't belong to the water or urban classes and are still "no data" values in the output dataset. This has to be done in four steps (Figure 38, Figure 39, and Figure 40). In step one, Expert rule Va, remaining pixels which are grassland in classification scenario CL2p5 and not agriculture in CL2p5 and were classified as shrub or open forest in CL5 are identified and the LC class of CL5 is written to the output dataset. For these identified pixels, no class probability can be assigned (Figure 38). The remaining output pixel locations with "no data" values are then filled up with the results of the classification scenario CL2p5 if the LC class is not agriculture and also the corresponding class probabilities are taken over (Figure 38).

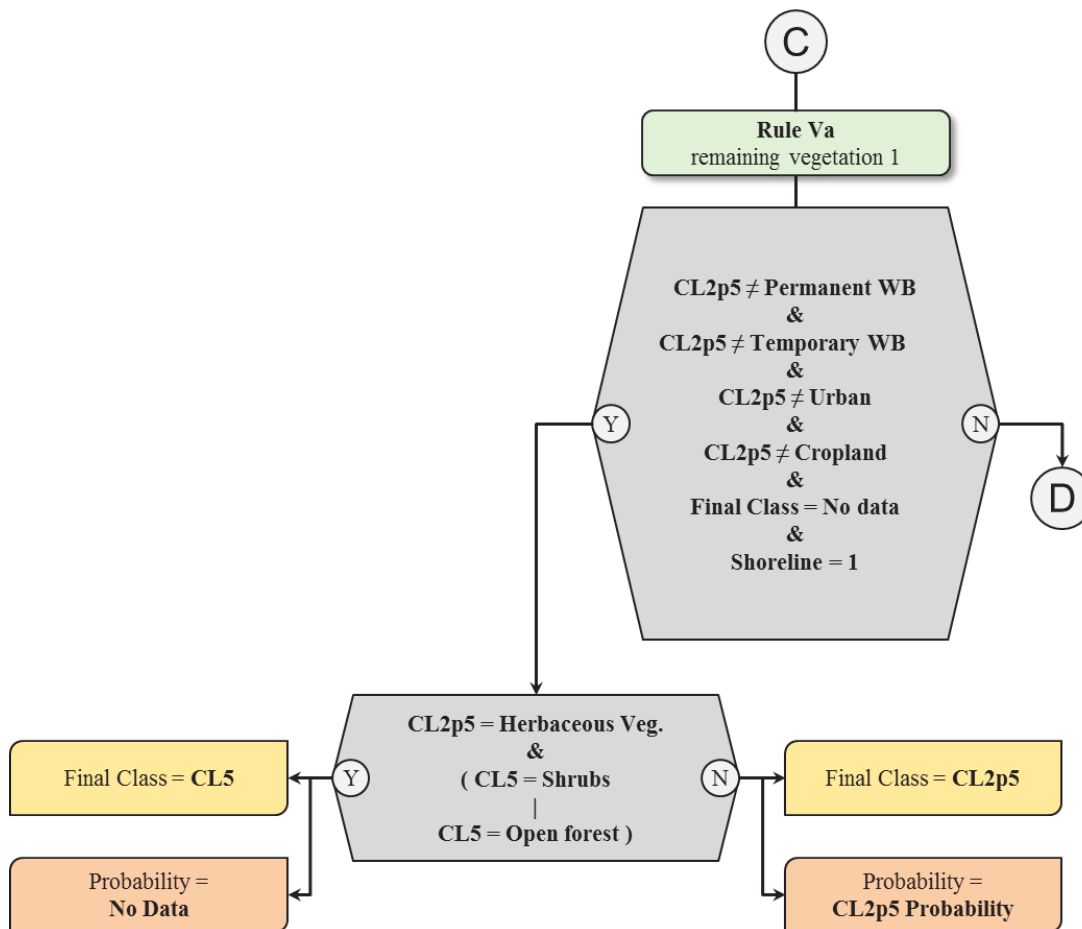


Figure 38: Expert rule Va of the CGLS LC100 discrete map generation process.

Step two of expert rule V identifies those of the remaining “no data” pixels in the output dataset which are not agriculture in the CL5 discrete map and copies them to the output dataset (again no class probability can be assigned for these pixels) (Figure 39). Remaining land masses pixels with no assigned LC class (other than water or urban) in the output dataset are subject to expert rule Vc. This rule selects those of the remaining pixels which pixel locations have not an agriculture LC class in the CL1 classification scenario and copies the LC class of CL1 together with the class probability (Figure 39). If then still “no data” pixels have no assigned LC class, expert rule Vd was implemented to fill those. Thus, the maximum natural vegetation mask generated from the normalized cover fraction layers for forest, herbaceous vegetation, shrub and bare ground (CL7) is used. This mask gives for each pixel location the LC class which had the highest cover fraction percentage in the inputted cover fraction layers. All remaining land masses pixels with no assigned LC class (other than water or urban) in the output dataset are filled with the corresponding LC class from CL7 (again no class probability can be assigned for these pixels) (Figure 40).

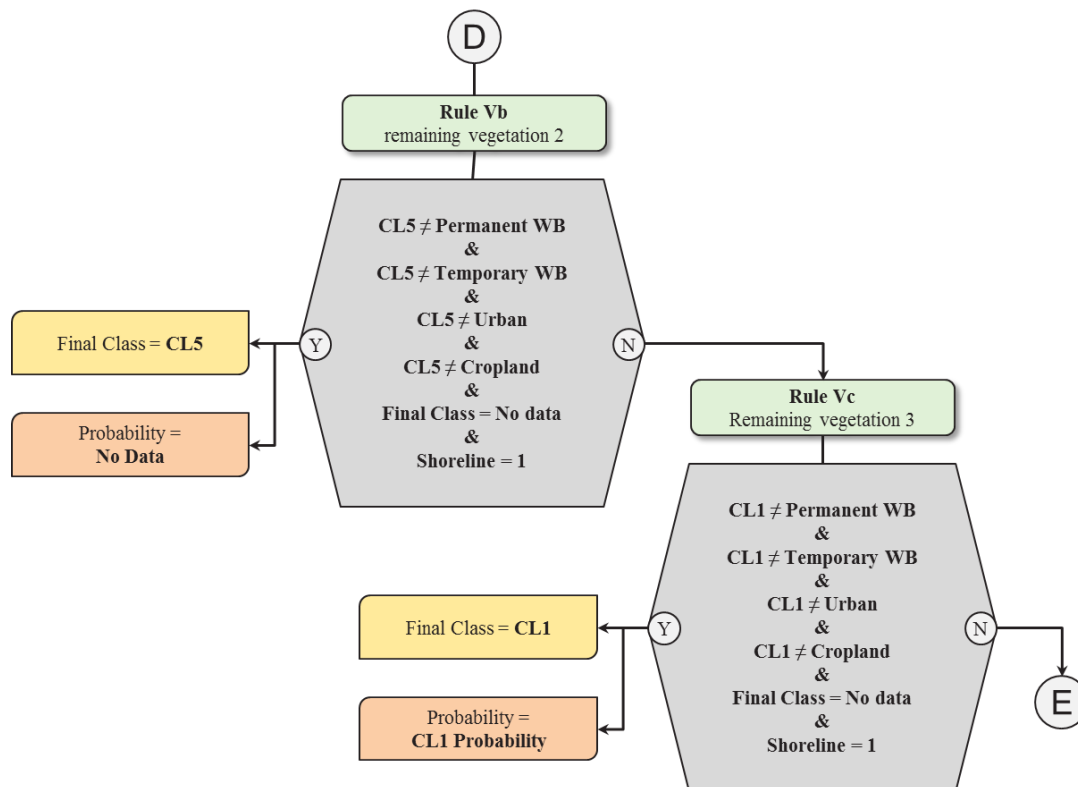


Figure 39: Expert rules Vb and Vc of the CGLS LC100 discrete map generation process.

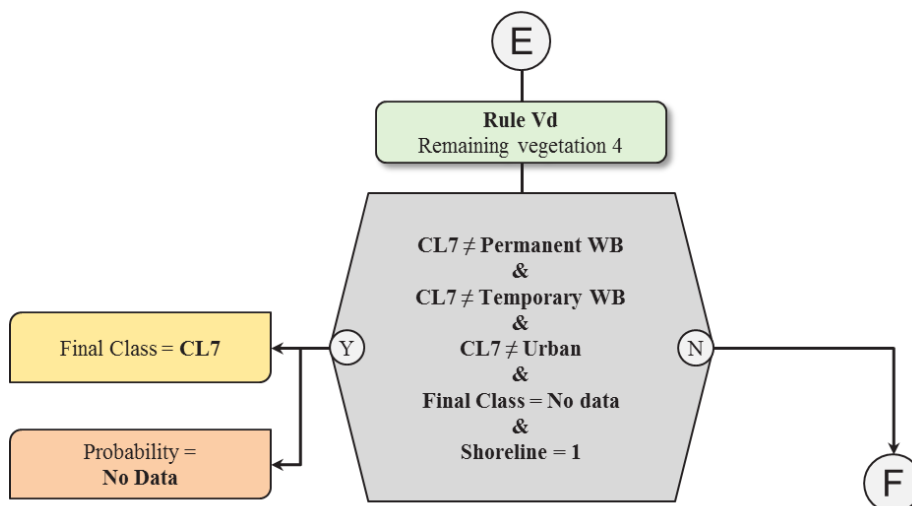


Figure 40: Expert rules Vd of the CGLS LC100 discrete map generation process.

Expert rule VI is now implementing the remaining hard masks: the permanent water body mask, the temporary water body mask, the herbaceous wetland mask and the urban mask. These hard masks can be overruling, meaning they can even overwrite a pixel with an already assigned LC

class in the output dataset. First, in expert rule VIa, the herbaceous wetland mask is compared to the classification results of scenario CL2p5. If a wetland mask pixel is also classified as grassland in the CL2p5 classification map, then pixel location is set as herbaceous wetland in the output dataset (Figure 41). This kind of check is also applied to the temporary water body mask (expert rule VIc). Only those temporary water body locations which are not classified as agriculture in the output dataset are taken to the output dataset (again no class probability can be assigned for these pixels) (Figure 41). In contrast to that, the permanent water body mask and urban mask are truly overruling masks and are directly written to the output dataset (expert rule VIb and VIId in Figure 41). Moreover the class probabilities are set to 100 % for pixels in these two masks.

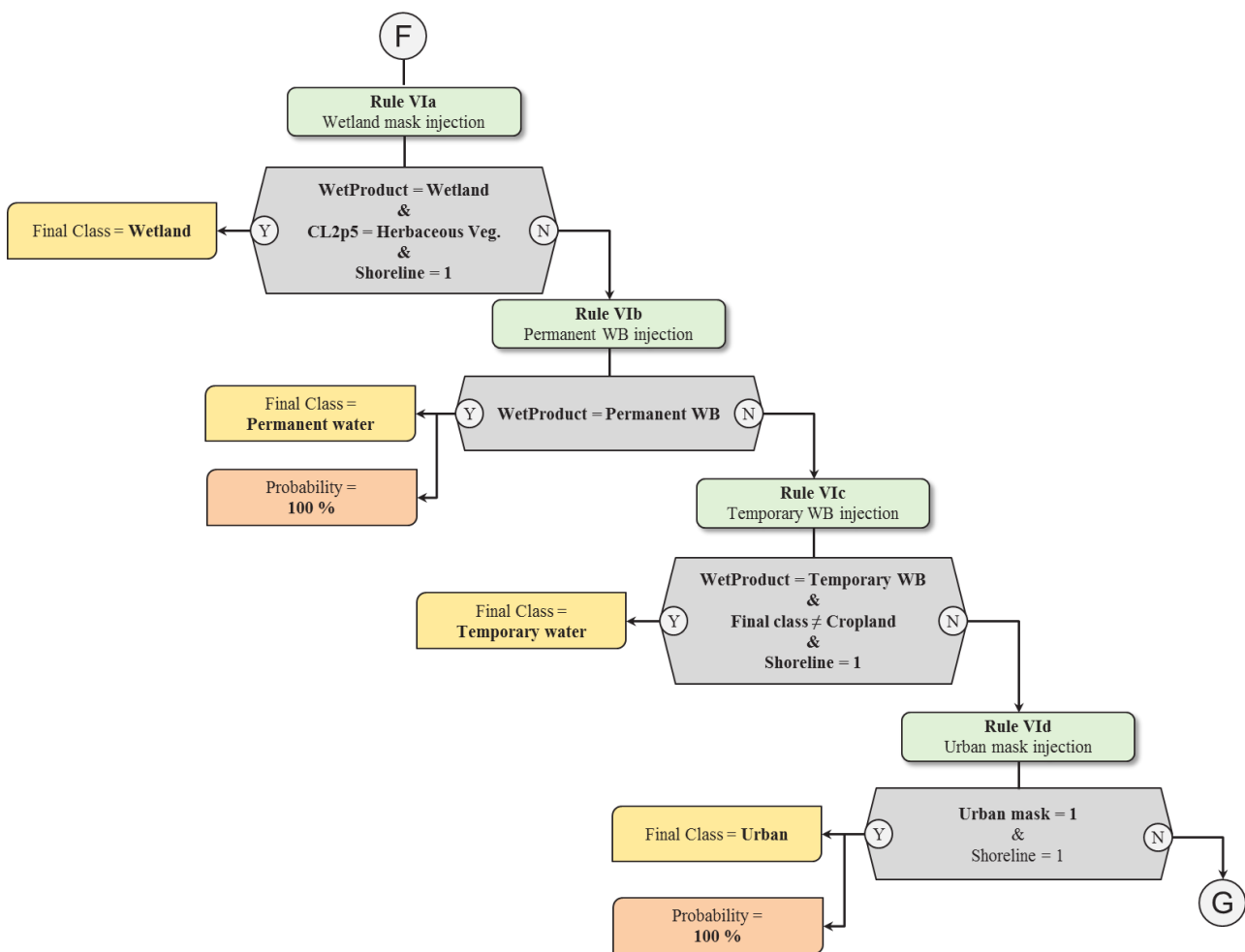


Figure 41: Expert rules VIa – VIId of the CGLS LC100 discrete map generation process.

Expert rules VII, VII and IX are clean up rules, where rule VII sets “no data” pixels outside the African land masses to the LC class permanent water, checks expert rule VIII the NOVO mask for pixel with no PROBA-V 100m observations in the reference year. This was implemented to make sure that pixels with absolutely no EO observation data are set to the “no data” value even when

the data fusion has maybe generated artificial time series data (Figure 42). Expert rule IX applies a shrub to open forest conversion (Figure 43). We noticed a slightly overestimation of shrubland in the classification CL2p5 scenario. This is mainly due to confusion between shrubland and open forest in the RF classification algorithm, since training points with low forest cover fractions can have similar metric profiles than shrubland training points. The CL5 map which was created by applying the training data rules on the normalized cover fraction layers don't show this shortcoming. Thus, expert rule IX checks all in the output dataset as shrub classified pixels against the CL5 classification, and overwrites those pixels were the LC class values is open forest in CL5.

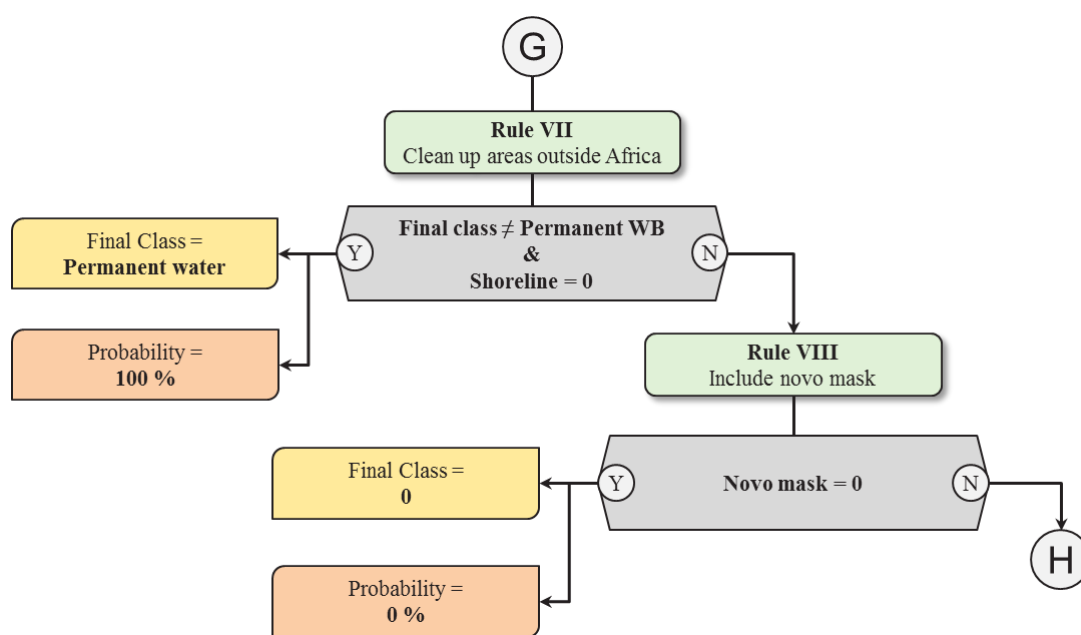


Figure 42: Expert rules VII and VII of the CGLS LC100 discrete map generation process.

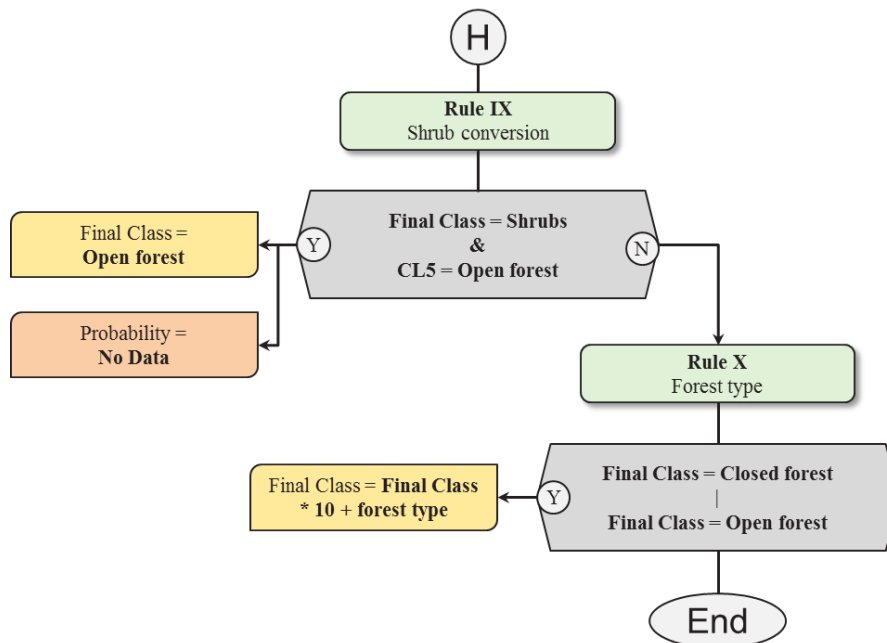


Figure 43: Expert rules IX and X of the CGLS LC100 discrete map generation process.

The results of “forest type” classification scenario (CL4) are used in the final expert rule number X (Figure 43). The CL4 layer is used to separate the discrete classes “closed forest” and “open forest” in the output dataset into the different forest type classes.

3.10.4 Metadata

The final step in the CGLS LC100 discrete map generation is the injection of metadata attributes compliant with version 1.6 of the Climate & Forecast conventions (CF V1.6) and the color bars translating the discrete class code into the legend. Furthermore, the water occurrence layer, showing where and how long surface water appeared over the reference year, is generated out of the WetProducts statistic layer (see section 3.6.4). Also the tGAPmask (see section 3.5.2.5) is extracted and provided as a quality layer showing pixel locations with a higher uncertainty in the data quality/classification accuracy due to errors during the data fusion process.

Overall four layers are provided as the CGLS LC100 discrete product:

1. LC100-LCCS
2. LC100-LCCS-PROB
3. LC100-LCCS-QFLAG
4. LC100-OCCUR-WB

Figure 44 shows the legend for the CGLS LC100 discrete map with 18 classes together with the colour codes and assigned byte values for each LC class, where Figure 45 shows an overview of the product on continental scale.





















 no ProbaV 100m data available (0)	 herbaceous vegetation (30)
 evergreen needleleaf closed forest (111)	 cropland (40)
 evergreen broadleaf closed forest (112)	 urban (50)
 deciduous needleleaf closed forest (113)	 bare / sparse vegetation (60)
 deciduous broadleaf closed forest (114)	 snow & ice (70)
 evergreen needleleaf open forest (121)	 permanent water bodies (80)
 evergreen broadleaf open forest (122)	 temporary water bodies (81)
 deciduous needleleaf open forest (123)	 herbaceous wetland (90)
 deciduous broadleaf open forest (124)	 open sea (200)
 shrubs (20)	 continental land mass not classified (255)

Figure 44: Legend for the 18 discrete classes of the CGLS LC100 discrete map for Africa 2015. Note: the number in brackets represents the numerical code for a land cover class.

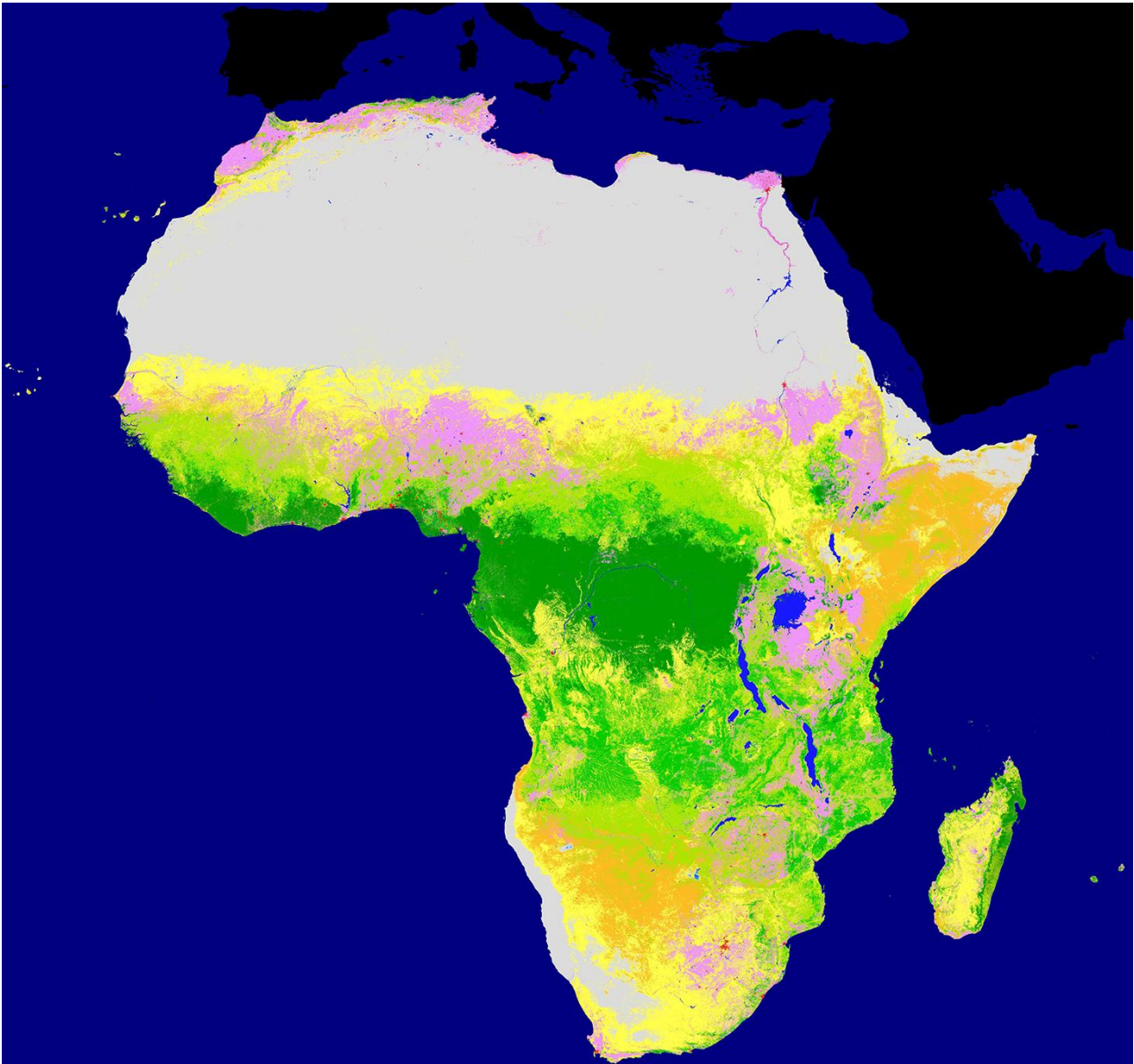


Figure 45: The CGLS LC100 discrete map for Africa 2015 with 18 discrete classes (shown at continental scale).

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:

- Artefacts at boundaries of ecozones can appear due to the used ecozone vector layer as well as the ecozone-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 (burned areas) were not yet taken into account and therefore could lead to misclassifications.
- Artefacts in the phenological cycle detection can 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 100m (i.e. Nigeria, Ghana). This could lead to overestimate the croplands.
- Areas with low cropland fragmentation (very sparse cropland fields of a very small size) are difficult to map because of the resolution of 100m. This could lead to underestimate the croplands.
- Very small African villages are difficult to map, especially when not detected by the GUF+ layer at 12 m resolution, which could lead to an underestimate of urban.
- Some limitations are due to the legend or class definition:
 - 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 WBs, e.g. some agricultural fields, build-up areas;
 - Natural surfaces with spectral signatures very close to WBs, e.g. salt lakes;
 - Areas where the spectral properties of WB fall outside the defined thresholds.

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 100m Water Body detection algorithm need to be updated if necessary. Moreover, the algorithm has to be adapted in cases for a reference year the stated input data (PROBA-V 100m 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.

6 REFERENCES

- Badgley, G., Field, C. B., & Berry, J. A. (2017). Canopy near-infrared reflectance and terrestrial photosynthesis. *Science Advances*, 3(3), e1602244.
- Bertels, L., Smets, B., Wolfs, D. (2016). Dynamic Water Surface Detection Algorithm Applied on PROBA-V Multispectral Data. *Remote Sens.*, 8, 1010.
- 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.
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32.
- Eberenz, J., Verbesselt, J., Herold, M., Tsendbazar, N.-E., Sabatino, G., Rivolta, G. (2016). Evaluating the Potential of PROBA-V Satellite Image Time Series for Improving LC Classification in Semi-Arid African Landscapes. *Remote Sens.*, 8, 987.
- Eerens, H., Haesen, D., Rembold, F., Urbano, F., Tote, C., and Bydekerke, L. (2014). Image time series processing for agriculture monitoring. *Environ. Model. Soft.* 53, 154–162. doi: 10.1016/j.envsoft.2013.10.021
- FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS (FAO) (2012). Global ecological zones for FAO forest reporting: 2010 Update. Forest Resources Assessment Working Paper 179, FAO, Rome, Italy.
- Huete, A., et al. (2002). Overview of the Radiometric and Biophysical Performance of the MODIS Vegetation Indices. *Remote Sensing of Environment*, 83,195–213.
- Hughes, G. F. (1968). On the Mean Accuracy of Statistical Pattern Recognizers. *IEEE Transactions on Information Theory*, IT-14:55-63.
- Kalman, R. E. (1960). A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*. 82: 35. doi:10.1115/1.3662552.
- 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
- Key, C., Benson, N. (2005). Landscape Assessment: Remote Sensing of Severity, the Normalized Burn Ratio; and Ground Measure of Severity, the Composite Burn Index. In *FIREMON: Fire Effects Monitoring and Inventory System*, RMRS-GTR, Ogden, UT: USDA Forest Service, Rocky Mountain Research Station.
- 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.

- 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.
- Marconcini, M., Ureyen, S., Esch, Y., Metz, A., Zeidler, J. (2017a). 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., Üreyen, S., Esch, T., Metz, A., Zeidler, J. and Palacios-Lopez, D. (2017b). “Outlining the urban side of the Earth – the GUF+2015”, *Scientific Data* (in preparation).
- NASA JPL. (2013). NASA Shuttle Radar Topography Mission Global 1 arc second [Data set]. NASA LP DAAC. <https://doi.org/10.5067/MEaSURES/SRTM/SRTMGL1.003>
- Niraj (2016): Data Science, ML & AI Tutorials – Machine Learning. Webpage: <https://sites.google.com/site/nirajatweb/>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, p. 2825-2830.
- Pekel, J.F., Cottam, A., Gorelick, N., Belward, A.S. (2016). High-resolution mapping of global surface water and its long-term changes. *Nature* 540, 418-422. (doi:10.1038/nature20584)
- Pekel, J.-F., Vancutsem, C., Bastin, L., Clerici, M., Vanbogaert, E., Bartholomé, E., & Defourny, P. (2014). A near real-time water surface detection method based on HSV transformation of MODIS multi-spectral time series data. *Remote Sensing of Environment*, 140, 704–716. doi:10.1016/j.rse.2013.10.008
- Pesaresi, M., Ehrlich, D., Florczyk, A. J., Freire, S., Julea, A., Kemper, T.P., Syrris, V. (2015). GHS built-up grid, derived from Landsat, multitemporal (1975, 1990, 2000, 2014). European Commission, Joint Research Centre (JRC) [Dataset] PID: http://data.europa.eu/89h/jrc-ghsl-ghs_built_ldsmt_globe_r2015b
- 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.
- 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
- 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.
- Simons, H. (2001). FRA 2000. Global Ecological Zoning for the Global Forest Resources Assessment 2000. FRA Working Paper 56. FAO, Rome.
- 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.
- Tucker, C.J. (1979). 'Red and Photographic Infrared Linear Combinations for Monitoring Vegetation. *Remote Sensing of Environment*, 8(2),127-150.
- Van Hoolst, R.; Eerens, H.; Haesen, D.; Royer, A.; Bydekerke, L.; Rojas, O.; Li, Y.; Racionzer, P. (2016). FAO's AVHRR-based Agricultural Stress Index System (ASIS) for global drought monitoring. *International Journal of Remote Sensing* Vol. 37 , Iss. 2.
- 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.
- 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.