Climate Change Initiative Extension (CCI+) Phase 2
New Essential Climate Variables (NEW ECVS)
High Resolution Land Cover ECV (HR_LandCover_cci)

Product Validation and Algorithm Selection Report

(PVASR)

Prepared by:

Università degli Studi di Trento
Fondazione Bruno Kessler
Università degli Studi di Pavia
Università degli Studi di Genova
Université Catholique de Louvain
Politecnico di Milano
LSCE
CREAF
University of Exeter
e-GEOS s.p.a.
Planetek Italia

Changelog

Issue	Changes	Date
1.0	First issue	21/01/2025
1.1	Updated based on RIDs	21/02/2025

Detailed Change Record

Issue	RID	Description of discrepancy	Sections	Change
1	ESA-01	GlobCover Land Cover map 300m 2019, to which map are you referring to? GlobCover or C3S MRLC 2019?	Table 5, Table 6, Table 7	The reference to the 'GlobCover Land Cover map 300m 2019' is incorrect. The validation dataset was extracted from the CCI Medium Resolution Land Cover (MRLC) Map 300m 2019. Tables and text in the document have been updated accordingly.
2	ESA-02	GlobCover validation points are from 2008, this dataset has been updated to 2019?	3.2.1.1 and 3.2.1.2.1	
3	ESA-03	Please remove the space from the link after "download/" http://maps.elie.ucl.ac.be/CCI/viewer/download/ESA CCI-LC-Ph2-PUGv2_2.0.pdf	Reference [11]	The document is updated accordingly.
4	ESA-04	For the comparison maybe it is better to use the same scale for the values on the Y axis for both, the weekly and monthly plots.	Fig. 37 and 38	The document is updated accordingly.
5	ESA-05	Did you investigate the moisture index? NDMI = (NIR – SWIR) / (NIR + SWIR). Maybe it could provide some additional information on the LC	3.4.2	We will consider NDMI for the following experiments.

Contents

1 Introduction

1.1 Executive summary

By the end of the first year, the EOS team conducted extensive comparative testing and performance analyses to evaluate the effectiveness of different algorithms for specific components of the processing chain designed to produce high-resolution land cover (HR LC) products. The results of these experiments, along with recommendations for the best-performing techniques, are detailed in section 3. The analysis is ongoing, as the team continues to assess various approaches to select the final algorithms that will ensure optimal performance for both static and historical land cover (LC) and land cover change (LCC) products.

1.2 Purpose and scope

The Product Validation and Algorithm Selection Report (PVASR) provides an in-depth overview of the comparative tasks conducted to assess the best-performing algorithms and techniques for inclusion in the classification blocks of the overall processing chain. The current version presents the activities carried out during the first year, with a particular focus on the classification of optical and SAR imagery, decision fusion and change detection. Key areas of emphasis include:

1. Testing optical pre-processing and evaluating its performance in terms of accuracy, computational efficiency and composite quality.
2. Testing classifiers and evaluating their performance in terms of accuracy, computational efficiency, and adaptability for model/code modifications to meet specific requirements and implementation needs.
3. Exploring methods for creating reliable training datasets from existing products, which may be sub-optimal in terms of spatial resolution (coarse to medium) and legend detail (less comprehensive compared to HR LC products as outlined in the ATBD).
4. Assessing sets of multitemporal features used as inputs for classifiers.
5. Evaluating multisensory decision fusion methods in terms of both accuracy and computational time.
6. Evaluating optical composite generation to enhance detection accuracy and analyzing different feature spaces to ensure reliable change detection maps.

1.3 Applicable documents

Ref.	Title, Issue/Rev, Date, ID
[AD1]	CCI HR Technical Proposal phase 1
[AD2]	CCI HR Technical Proposal phase 2
[AD3]	CCI Extension (CCI+) Phase 1 – New ECVs – Statement of Work, v1.3, 22/08/2017, ESA-CCI-PRGM-EOPS-SW-17-0032
[AD4]	CCI_HRLC_Ph2-D1.1_URD, latest version
[AD5]	CCI_HRLC_Ph1-D2.2_ATDB, latest version
[AD6]	CCI_HRLC_Ph2-D2.2_ATDB, latest version

1.4 Reference documents

Ref.	Title, Issue/Rev, Date, ID
[RD1]	The Global Climate Observing System: Implementation Needs, 01/10/2016, GCOS200

1.5 Acronyms and abbreviations

3D-FCN	3-Dimensional - Fully Convolutional Network
AC	Atmospheric Correction
ATBD	Algorithm Theoretical Basis Document
BSI	Bare Soil Index

CCI+	Climate Change Initiative Extension
dB	Decibel
DEM	Digital Elevation Model
DL	Deep Learning
ER	Ecoregion
GEE	Google Earth Engine
GRD	Ground Range Detected
HR	High Resolution
HRLC10	CCI High Resolution Land Cover Map at 10m resolution of 2019
HRLC30	CCI High Resolution Land Cover Map at 30m resolution from 1990 onwards every 5 years
HRLCC30	CCI High Resolution Land Cover Change Map at 30m resolution from 1990 onwards
IW	Interferometric Wide Swath
L-5/7/8/9	Landsat-5/7/8/9
LC	Land Cover
LCC	Land Cover Change
MOLCA	Map Of LC Agreement
MRLC	Medium Resolution Land Cover
NDVI	Normalized Difference Vegetation Index
NDWI	Normalized Difference Water Index
OA	Overall Accuracy
PA	Producer Accuracy
PCC	Post Classification Comparison
RF	Random Forest
S1/2	Sentinel-1/2
SAR	Synthetic Aperture Radar
SCL	Sen2Cor Scene Classification Layer
SITS	Satellite Image Time Series
SNAP	Sentinel Application Platform
SR	Surface Reflectance
UEXT	Urban EXTent
VH	Vertical-Horizontal polarization
VV	Vertical-Vertical polarization

2 Selection procedure

The overall procedure for the selection of best performing algorithms and methods is performed according to a three-step procedure. The algorithms presented in the Technical Proposals [AD1] [AD2] and ATBD [AD6] are considered for the comparisons together with a set of proposed solutions for each task such as generating training samples and building multitemporal features. The evaluation-selection procedure is devised in such a way that the selected algorithms/techniques are the most suitable to satisfy project requirements.

The three steps of the procedure are the following:

- **Step 1: Qualitative pre-screening of algorithms**
  A pre-screening of the algorithms and methods from a State-of-the-art pool of competitors is carried out in order to identify the most relevant methodologies with respect to the project objectives. This preliminary analysis is driven by the selection criteria described in Section 2.1. In this first step, a high-level qualitative evaluation of these criteria is conducted in order to identify techniques that clearly cannot reach a satisfactory ranking on several categories of parameters. These techniques are discarded and not considered in the next steps. Algorithms and methods that passed the pre-screening are reported in the Technical Proposal [AD2] and more in detail in the ATBD [AD6]. In this report only the methods that passed the pre-screening are considered explicitly.
- **Step 2: Quantitative evaluation of algorithms**
  Algorithms that pass the pre-screening in step 1 are analyzed in greater detail with a quantitative evaluation. This analysis is based on different parameters, ranging from a scientific and technical analysis to possible impacts on the application and users. For each investigated item (algorithm, method, technique, etc.) details on the quantitative evaluation of the comparison activities can be found in a dedicated section of this document.
- **Step 3: Final decision**
  According to the analysis carried out for each individual comparison task, a **final decision** is taken according to the best performer and its relevance with respect to project objectives. Final decision is reported.

It is worth noting that the pre-processing algorithms are not included in the evaluation and ranking procedure because we expect to import in the project basic pre-processing chains already developed for both multispectral and SAR data.

2.1 Criteria

In this section the criteria adopted for evaluating the relevance of methods and algorithms with respect to project requirements are listed. Up to seven categories of parameters are considered divided in different issues.

1. **Scientific Background and Technical Soundness** – The scientific validity of the algorithms and of the methodologies on which the algorithms are based is considered as an important parameter. The rationale is that selected algorithms should be based on a solid theoretical background that guarantees the accuracy of its results also at an operational level. The guidelines for rating are as follows:
   - The methodology is solid;
   - The methodology is technical convincing;
   - The methodology is at the state-of-the-art;
   - The methodology is published in high quality journals;
   - The methodology is included in several other scientific publications or project technical reports.

2. **Robustness and Generality** – In order to obtain a reasonable estimation for the robustness and generality of the investigated algorithms, different parameters are considered, such as:
   - The method is suitable to be used with different kinds of images (e.g., S2, Landsat, SAR, etc.);
   - The method shows high performance on different images (Sentinel, Landsat, etc.) and over the three test areas as described in URD [AD4];
   - There are software implementations or examples for the implementation available;
   - The algorithm can be used in combination with other methodologies.

3. **Novelty** – An appropriate candidate algorithm should have been published or reported for the first time relatively recently in the literature. It is not required that algorithms are completely innovative; the novelty may consist in both combining well established methodologies or applying well-known techniques in a novel way. As a main guideline, a tested method should be already applied in literature to solve existing problems.

4. **Operational Requirements** – The expected operational requirements (in terms of computational complexity, time effort, cost, etc.) for the final implementation of an algorithm/technique are evaluated. Although no actual constraints are fixed on the algorithm computational complexity, the most optimized implementations available in literature are preferred. Other crucial aspects are:
   - The algorithm is prone to architectural modifications;
   - The processing time scaling is likely to be linear with image size;
   - The hardware and disk-storage requirements are appropriate.

   Algorithm/method consistency with project requirements is also extremely relevant, following guidelines from GCOS [RD1] and SoW [AD3][AD3] :
   - Algorithms and methodologies must be effective for high resolution images (e.g., optical data at 10-30m).
   - Documented accuracy must be within the boundaries imposed by GCOS (see[RD1]) and as reported in SoW [AD3].

5. **Accuracy** – An algorithm is positively evaluated if able to provide a high absolute accuracy in all test areas, especially keeping into account the different climatological conditions and possibly different data availability conditions. Accordingly, the following guidelines are used for evaluating accuracy characteristics:
   - Accuracy/uncertainty to be in line with GCOS [RD1] requirements as reported in SoW [AD3].
   - The algorithm matches the end-user (climatologist and other users from the community) requirements;
   - For unsupervised tasks the accuracy should not depend on the availability/quality of prior information.
   - For supervised tasks the accuracy should be robust to the availability/quality of prior information.

6. **Level of Automation** – From an operational point of view, it is mandatory that an algorithm runs in a completely automatic way. Algorithms requiring any amount of manual work, strong interaction with the final users are negatively evaluated.

7. **Specific End-users Requirements** – From an operational point of view, capability of an algorithm to satisfy and meet possible end-user requirements is another important parameter of evaluation. The main guidelines for driving this ranking are:
   - The algorithm is robust to the use in several climatological regions;
   - The algorithm can be reasonably included in an operational procedure.

2.2 Evaluation

The evaluation procedure of each comparative task aimed at deciding on a specific algorithm/technique is carried out by considering all criteria listed before. To each reported activity, a thorough discussion is given regarding how these criteria are weighted in the overall evaluation, which aspects are given strong emphasis and which ones are considered less relevant. The evaluation activity provides answers about best performing algorithms/techniques that are included in the processing chain of the current version of HR LC products.

3 Algorithms and procedures (year 1)

3.1 Optical data processing

The optical processing chain is designed to primarily work with images at 10/30m resolution, producing outputs at the same resolution. It leverages multitemporal, multispectral data from recent years, including Sentinel-2 (S2) and Landsat-8/9 (L-8/9), alongside legacy data from Landsat-5/7/8 (L-5/7/8). The pre-processing methodology follows the same logical framework as Phase 1, with planned adjustments aimed at enhancing both the quality of the output (e.g., optical composites) and the computational efficiency of the process.

In the classification stage, Phase 1 faced challenges including: (i) limited availability of photo-interpreted data suitable for mapping large areas; (ii) constraints related to input features; and (iii) the need to optimize the classification algorithm for efficiency. Phase 2 focuses on refining the classification pipeline to address these challenges.

The pre-processing phase addresses radiometric and geometric distortions specific to sensors and platforms, as well as harmonization tasks. Radiometric corrections are applied to mitigate issues such as variations in illumination, viewing geometry, atmospheric conditions, and sensor-specific noise or response variations. These factors depend on the sensor, platform, and acquisition conditions. A major challenge with optical imagery is cloud coverage, which requires targeted processing to accurately identify cloud and cloud shadow pixels, potentially incorporating restoration techniques to recover spectral information for occluded areas. Figure 1 shows the main blocks of the optical pre-processing chain.

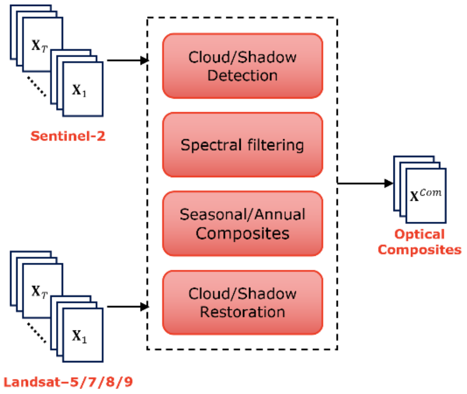


Figure 1. Optical pre-processing chain.

Figure 2 illustrates the optical data processing chain used to produce both static and historical high-resolution land cover (HRLC) maps. The workflow involves pre-processing the images to create optical composites, integrating these with ancillary data (e.g., Copernicus Digital Elevation Model [DEM]) to extract features for classification. The classifiers are trained on available training data points and subsequently generate pixel-wise class-posterior probabilities. These probabilities feed into the decision fusion processing chain, resulting in the final land cover products.

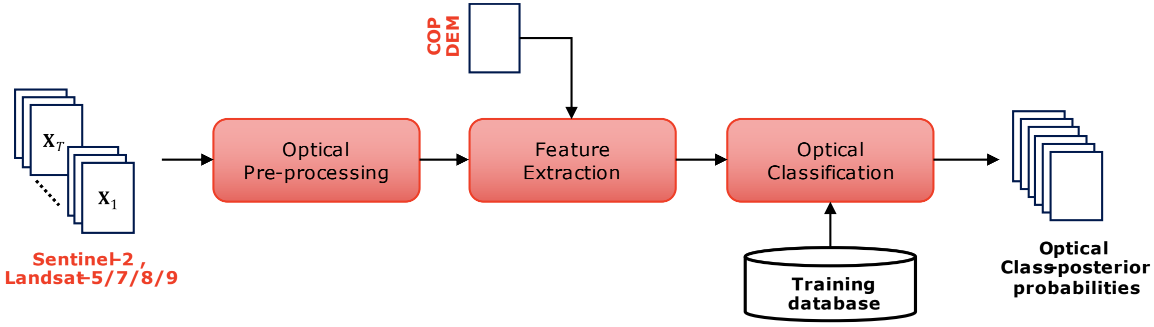


Figure 2. Optical data processing chain for the prototype production of both the static and the historical HRLC maps obtained by classifying the time series of HR optical data.

The following sections provide the results of the activities on the optical processing chain during year 1 of the project. The results here presented are focused on three main activities:

1. Cloud detection: validation of the S2 Sen2Cor Scene Classification Layer (SCL) Cloud and Cloud shadow mask enhancement;
2. Composite generation: code optimization and alternative compositing strategies;
3. Optical classification: preliminary evaluation of deep learning architectures for HRLC mapping with optical Satellite Image Time Series (SITS).

3.1.1 Cloud and Cloud shadow Detection

The improvements foreseen for the pre-processing chain of optical data require both quality and speed in the computation. This last requirement is essential for reducing costs and production times, especially considering the scaling of the production to large areas. For this reason, during Phase 1, Surface Reflectance (SR) products were not generated as part of the processing chain, and instead they were retained from the providers, thus relying on the providers Atmospheric Correction (AC) algorithms. This allowed both to use consolidated AC strategies and to save processing time. On top of that, cloud detection is often coupled with AC during the generation of SR products. Therefore, clouds and cloud shadows masks are usually provided alongside the SR products. The availability of precomputed cloud and cloud shadows masks means that we have the possibility of saving computation time by utilizing the available masks. During Phase 1, Fmask [1] cloud and cloud shadows masks, provided with the SR Landsat products, have shown to be effective in the cloud detection in Landsat imagery with satisfying performance levels [AD5]. Therefore, it has been used for masking clouds in the historical processing chain for HRLC30 products. On the other hand, Sen2Cor has shown some limitations in the cloud detection capabilities with S2 SR products during Phase 1. Therefore, we developed a cloud and cloud shadow enhancement module. This approach is being re-evaluated in Phase 2, considering the recent improvements in the Sen2Cor cloud detection module. Indeed, the new Collection 1 of reprocessed S2 products (the data that will be used in the Phase 2) made available by ESA consider also upgraded cloud detection strategies that leverage the parallax effect of S2 MSI sensors.

Therefore, we are considering whether to use the new Collection 1 masks as they are or whether to apply the Phase 1 cloud enhancement module to them. The selection will weigh the performance improvements against the necessity of additional processing time, which would undermine the production time for HRLC10 products. Indeed, the cloud and cloud shadow enhancement module require the computation of a seasonal background image, used to better discriminate clear sky observations from clouds. This means that the cloud detection of each acquisition relies on the time series of observations of the current season, which is computationally expensive (similar to the composite generation). Therefore, this increase in computation time requires a sufficient improvement in cloud detection performance to be acceptable.

In this analysis, we compare the recent Sen2Cor and the Ph1 Cloud Enhancement module using the multi-temporal global benchmark dataset CloudSEN12+ [2], [3], for cloud and cloud shadow detection with S2. This dataset provides 49,250 S2 image patches (IPs) with different annotation types: (i) 10,000 IPs with high-quality pixel-level annotation, (ii) 10,000 IPs with scribble annotation, and (iii) 29,250 unlabelled IPs. The labelling phase was conducted by 14 domain experts using a supervised active learning system. A rigorous four-step quality control was designed to guarantee high quality in the manual annotation phase. Furthermore, CloudSEN12+ ensures that for the same geographical location, users can obtain multiple IPs with different cloud coverage: cloud-free (0%), almost-clear (0–25%), low-cloudy (25–45%), mid-cloudy (45–65%), and cloudy (>65%), which ensures scene variability in the temporal domain. Therefore, CloudSEN12+ provides a reliable benchmark for precisely evaluating different cloud detection algorithms. For this experiment, we consider only IPs of 2000×2000 pixels for which high quality annotations are available. Each pixel is labelled as “Clear”, “Thick Cloud”, “Thin Cloud”, or “Cloud Shadow”. The main difference between thick and thin clouds is that thin clouds are semi-transparent, while thick clouds are opaque and highly reflective in the visible bands. However, we don’t make a difference between thick and thin clouds, thus they are merged into the same class “Cloud” in our analysis. While CloudSEN12+ is a very large dataset, it does not contain all the adjacent acquisitions of each labelled scene, making it difficult to use with the Phase 1 Sen2Cor enhancement module, which requires the computation of a seasonal background by aggregating all the acquisition of the same season as the target acquisition. Therefore, we selected a subset of CloudSEN12+. Specifically, we selected three S2 tiles for which we have multiples IPs during the same season and a fair representation of each class:

- 18NWL with 4 IPs in Winter 2019;
- 21HUA with 3 IPs in Spring 2019;
- 36SWH with 4 IPs in Winter 2017.

For each of these tiles, we collected all the S2 L2A acquisitions of the corresponding season, and computed the enhanced masks. Figure 3 shows the visual comparison of the cloud and cloud shadow masks obtained by Sen2Cor and the enhancement module. From the images, it is clear that Sen2Cor overestimates clouds, and the enhancement module further accentuates this phenomenon. The quantitative results in Table 1 confirm the observations, reporting the User’s Accuracy (UA), Producer’s Accuracy (PA) and F1 score against the CloudSEN12+ subset. Indeed, the enhancement module reduces the omission errors at the cost of highly increasing the commission errors of the Cloud category over the Clear category. Cloud shadow performance remain unchanged.

Given the limited improvements provided by the enhancement module applied to the new Sen2Cor masks, the decision is to remove the enhancement step from the processing chain, reducing the overall computation time.

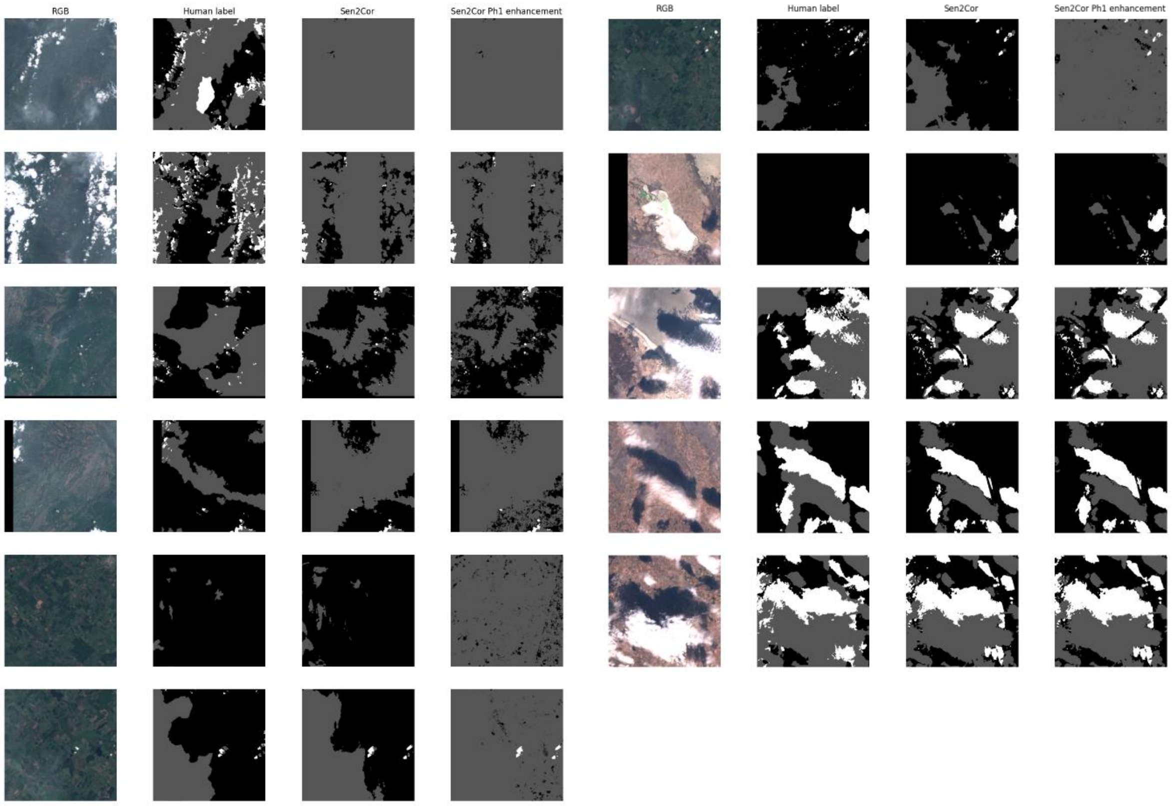


Figure 3. Comparison of the expert annotation with the Sen2Cor and Ph1 Improved Sen2Cor cloud and cloud shadow masks. (black = clear, gray = cloud, white = cloud shadow)

Table 1 Quantitative accuracy metrics from the benchmark analysis on the CloudSEN12+ dataset.

	Sen2Cor			Ph1 Sen2Cor Enhancement		
	UA	PA	F1	UA	PA	F1
Clear	0.91	0.85	0.88	0.86	0.44	0.58
Cloud	0.68	0.84	0.75	0.38	0.85	0.52
Cloud Shadow	0.91	0.75	0.82	0.91	0.75	0.82

3.1.2 Composite Generation

The composite generation is the most expensive step in the optical pre-processing chain. The activities of the first year focused on the improvement of the processing time while maintaining or improving the composite quality. During Phase 1, optical composites were generated by using a processor written in Python, using GDAL, numpy and basic multiprocessing capabilities. The approach is here compared to two different alternatives: FORCE and an improved Python processor adopting more advanced libraries such as xarray and dask.

FORCE has the advantage of being compiled, thus it allows generally fast processing time. Moreover, FORCE is developed with a focus on time series analysis: it implements the concepts of data cube and allows for parallel processing of long Satellite Image Time Series (SITS). However, it has some restrictions, as it requires the data to be saved on disk in a specific format, achievable only by processing Top-of-Atmosphere (TOA) data directly with FORCE AC, which adds an additional step to the processing chain. Also, FORCE does not provide a strategy for generating optical composites exactly as they were designed during Phase 1. Nevertheless, workaround strategies are under analysis both to avoid the AC processing of FORCE and to use instead the providers SR products, and to generate products equivalent to Phase 1 monthly/seasonal composites by exploiting specific parametrizations of FORCE time series analysis (TSA) tool.

The improved Python processor has been designed by keeping in mind two main aspects: the advantage of modelling data cubes for processing SITS data and the capability of scaling the processor in a distributed scenario. The improved Python processor implements two approaches to composite generation: band-wise median aggregation and medoid approach to most-representative-image selection (see ATBD [AD6]). The advantage of using the medoid approach is the generation of a composite whose spectral signature actually matches the spectral signature of on the observations that has been aggregated. This allows to have more consistent representation of the spectral signature of the land cover. Indeed, the band-wise median aggregation creates a composite whose reflectance values can come from different acquisitions, thus distorting the spectral signature of the land cover by mixing the spectral signature of potentially very different acquisitions. In the current implementation, the medoid is computed for each pixel as the image that minimizes the Euclidean distance in the spectral domain from the median composite.

The test area considered for the analysis is the S2 tile 21KUQ in the Amazonia area. From this area, we collected a time series of 73 S2 acquisitions over the year 2023, with acquisitions every 5 days. For the experiment, we considered both monthly and bimonthly composites:

- Legacy Phase 1 processor: test only with 12 bimonthly composites (pixel-wise median of each band of all the acquisition in a given month +- 15 days);
- Improved Python processor: test both monthly and bimonthly composites;
- FORCE: a custom parametrisation of TSA is used to generate monthly composites.

Note that FORCE does not only compute the median for each band by month, but it estimates a cloud-free SITS by cloud masking and interpolation beforehand. First, all clouds, cloud shadows and defective pixels are discarded. Then, the acquisitions are reprojected on a regular time grid by interpolation, generating a full time series of multispectral acquisitions. Only after this, the pixel-wise and band-wise median is computed by month. This approach effectively fills in missing data in the SITS, which is partly equivalent to Phase 1 cloud/shadow restoration step. However, the Phase 1 approach does this after the composite generation. Therefore, FORCE is able to better exploit the temporal information in the original SITS to fill in the missing data. In order to efficiently interpolate the time series, FORCE uses temporal convolutions with Radial Basis Function (RBF) kernels [4]. This ensures fast processing exploiting low-level implementations of the convolution operation. In the experiments, we consider an interpolation step (defining the spacing in the temporal grid) of 14 days, using three RBF kernel with sigma values of 14, 28, and 42 days. The sigma values define the width of the gaussian bell of each filter. Using different filters allows to more precisely follow the SITS when frequent acquisitions are available with the smaller kernels, whereas it allows for interpolating more distant time steps with larger kernels when few acquisition area available. The three filters are combined by weighting them by the acquisition's density within each filter. Since gaussian kernels have in principle infinite width, a cut-off density of 0.95 is set, reducing the effective width of each kernel. The maximum width allowed for each filter is also limited to ± 54 days around the centre of the kernel. After interpolation, the FORCE folding capabilities is used to generate temporal aggregates on a monthly basis. FORCE does not allow bimonthly folds, thus we only tested monthly composites. Nonetheless, the interpolation capabilities of FORCE actually address the same issue addressed by the use of bimonthly composites over monthly composites, i.e., reduced data availability. Therefore, FORCE monthly composites are actually comparable to both the monthly and bimonthly composites generated by both the legacy Phase 1 processor and the improved processor.

Table 2. Processing time of the different processors for composite generation

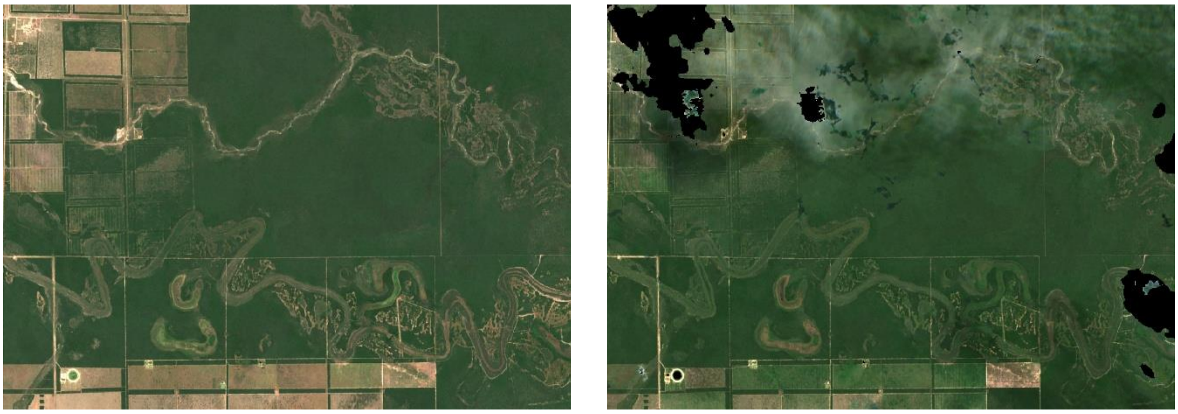
	12 Monthly composites	12 Bimonthly composites
Legacy Phase 1 processor	–	$\geq 3h$
Upgraded Python processor (median)	$33m$	$1h\ 26m$
Upgraded Python processor (medoid)	$1h\ 45m$	–
FORCE (interpolation + median)	$33m\ (+1h\ 18m\ L2\ gen)$	–

Table 2 reports the processing times of the different processors. These results were achieved on a workstation with 64GB of RAM (2x32GB DDR4 3200MHz CL16 non-ECC), a 8C/16T Ryzen 7 5800X CPU, and the data was stored on a M.2 NVMe v1.4 SSD Samsung 980. From these results, we can observe a clear 2x improvement in processing time over the legacy processors by the upgraded Python processor. If we consider the interpolation capabilities of FORCE, we achieve a ~6x improvement over the legacy processor. Considering monthly composites, the processing time of the upgraded processor with median aggregation and FORCE are similar. However, there are few aspects to keep into consideration. On one hand, FORCE performs both interpolation and median aggregation taking the same time as the upgraded processor, which performs only the median aggregation operation. On the other hand, FORCE needs the Level-2 imagery to be already stored following the specific FORCE requirements for tiling, projections, and file formats. Currently, the only way to achieve this is to use the Level-2 generation capabilities of FORCE, which adds an additional step to the processing ($\sim 1h\ 18m$ for generating all the Level-2 images of the 73 S2 acquisition considered). This makes the use of the upgraded processor more appealing. Approaches for avoiding this time consuming step are under analysis. If we can utilize the pre-computed Level-2 data from the providers, FORCE would become the most efficient approach to composite generation, also including an interpolation step before temporal aggregation, which further reduce gaps in the data. In the case this is not possible, the upgraded Python processor will be updated to integrate the interpolation procedure before computing the median composite, thus aligning it with FORCE. The processing time of the upgraded processor with medoid aggregation takes 3x the median aggregation. This large increase in processing undermines the applicability of the medoid operator. The results for computing the bimonthly composites with the medoid operator are missing due to the high memory requirements of this approach. However, we can expect worse processing time than the legacy processor.

Figure 4 shows some examples of optical composites for January (Figure 4(a)-(c)-(e)-(g)) and February (Figure 4(b)-(d)-(f)-(h)) 2023 in tile 21KUQ. While January is populated with mostly clear-sky observations for the whole tile, February consists of almost completely clouded acquisitions, with the exception of a single acquisition (Feb. 5th 2023) with sparse cloud cover. This challenge provides a good benchmark for understanding the most proper composite generation approach. First, we can observe that the January composites in Figure 4(a)-(c)-(e)-(g) are all very similar and of good quality. This is expected given the good number of cloud-free observations for all the pixels. Instead, the February composites in Figure 4(b)-(d)-(f)-(h) show some differences. First, Figure 4(b)-(d), generated by monthly median and medoid composites, respectively, show some gaps in the scene (the black regions). This is expected given the availability of a single acquisition with low cloud cover (Feb. 5th 2023), whereas all the other acquisitions have a cloud cover $> 90\%$. However, the other acquisitions still contribute to the final composite for the few available cloud-free pixels (from the computed cloud masks). Note that the cloud masks in these scenes are underestimating the cloud coverage, thus cloudy pixels are actually used in the composites. In these areas of the scene, we can observe a difference in the median and medoid composites (otherwise the same). In particular, the medoid composite seems noisier than the median (Figure 4(b)-(d)), as it tends to select the cloudy pixels, failing in filtering the available observations. If we consider the bimonthly composite of February 2023 (Figure 4(f)), we can see improvements related to absence of gaps in the scene. Thanks to the temporal window ranging from January 15th to March 15th, the bimonthly composite is able to exploit the cleaner observations of January and February to fill the gaps. However, we can observe artifacts due to the large temporal distance in the selected pixel values, which reflects the gaps due to clouds. The FORCE monthly composite, instead, is the one producing the cleanest February composite, despite using a median aggregation by month. This is possible thanks to the interpolation computed before aggregation, which not only allowed to avoid gaps in the scene by combining January and March acquisitions but also produced an artifact-free composite.

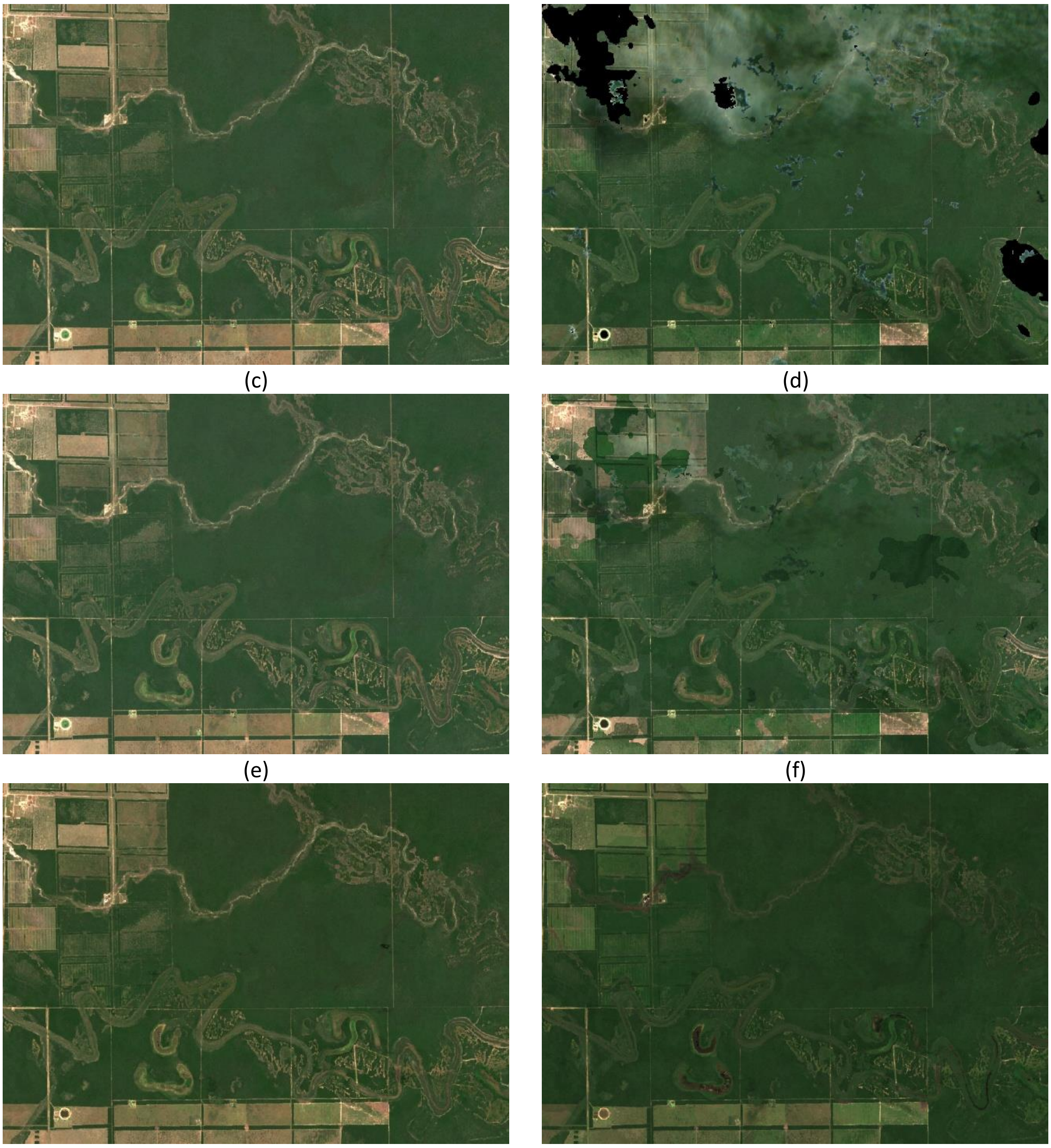
In conclusion, we improved the efficiency of the composite generation step, reaching good processing times. Among the considered approaches, medoid is the most expensive with no composite quality improvements, whereas FORCE has shown to be both efficient and to produce high-quality composites. The key aspect to the improved quality of FORCE can be identified in the interpolation procedure performed before median aggregation. However, we need to take into account that FORCE requires the input L2 data to be consistent to its internal representation. Therefore, we are considering two directions:

- The alignment of the upgraded processor to the FORCE TSA pipeline;
- The introduction of converter that is able to satisfy FORCE internal requirement, such that to remove the necessity of performing AC, which adds unavoidable processing time.



(a)

(b)



(c) (d)

(e) (f)

(g) (h)

Figure 4. Examples of generated composites for S2 tile 21KUQ in 2019 for January [firs column, i.e., (a) (c) (e) (g)] and February [second column, i.e., (b) (d) (f) (h)]. (a) (b) Monthly median composites. (c) (d) Monthly medoid composites. (e) (f) Bimonthly median composites. (g) (h) Monthly FORCE median composites. Black areas are pixels with not clear-sky observations in the considered period.

3.1.3 Optical Classification

During Phase 1, the optical classification was performed by using Support Vector Machines (SVMs). The main challenges in using SVMs were related to features selection and accurate pixel-wise class-posterior probabilities estimation. To solve both these problems in Phase 2, we consider deep learning strategies. Indeed, a deep learning architecture is able to learn a good feature representation starting from the raw input features, reducing time in tuning the optimal feature set, and it can directly estimate the pixel-wise class-posterior probabilities, instead of relying on external calibration approaches such as in the case of SVMs.

In our preliminary analysis, we compared three different architectures, all of which incorporate temporal information. To further examine the effectiveness of the methodologies in utilizing contextual information, we also included pixel-level networks. The considered architectures are the following:

1. Transformer: we employed the encoder architecture of a Transformer [5]. Positional encoding was applied to the SITS, focusing on classifying single pixels without incorporating contextual information.
2. Swin Transformer: we adapted a Vision Transformer (ViT), designed to capture both local and global

dependencies in the image. Instead of using global attention across the entire image like traditional ViTs, the Swin Transformer divides the image into non-overlapping windows, performing self-attention within each window. Our implementation considers an adaptation to the RS scenario of [6], initially designed for super-resolution of videos, here adapted to jointly model the spatial and temporal information of the SITS.

3. Convolutional Long Short-Term Memory (ConvLSTM): this hybrid model integrates convolutional neural networks (CNNs) with long short-term memory (LSTM) networks. It is tailored for spatio-temporal data, maintaining spatial information within individual time steps while capturing temporal dependencies across time steps [7].

Given the complexity of training deep network with the full HRLC legend on large areas, we performed a benchmarking exercise on restricted area in the Amazonia area. In order to provide a sufficiently large training dataset, we employed the HRLC10 map generated during Phase 1 as reference. Therefore, the accuracy metrics provide information on the level of agreement with the existing products. On one hand, a high level of agreement does not provide an actual accuracy metric. On the other hand, a high agreement means that the considered deep network agrees with the whole Phase 1 processing chain, including the optical-SAR decision fusion, spatial harmonization and post processing results, which improved the overall classification of the optical SVM classifier. Therefore, we consider this benchmark sufficient for comparing the different models and selecting the most suitable one for Phase 2. For our analysis, we selected the dataset corresponding to the MGRS S2 tile 22KGV in Brazil, chosen for its heterogeneous distribution of land cover classes, including tree cover (evergreen broadleaf, deciduous broadleaf, evergreen needleleaf), shrub (evergreen, deciduous), grassland, aquatic vegetation, cropland, bare areas, built-up areas, seasonal open water, and permanent open water.

We utilized all S2 images available for 2019, generating 12 monthly median composites. To ensure a balanced training dataset, we adopted a heterogeneity-based patch extraction approach. Specifically, we divided the S2 composite images into non-overlapping 5×5 pixel patches. For each patch, the dominant land cover class was identified. These patches were then randomly divided into training, validation, and test sets with a minimum requirement of N representative patches per class, where N was fixed at 1000.

For minority classes with fewer than N patches, 50% of available patches were allocated to the training set, 25% to the validation set, and 25% to the test set. This process resulted in a training dataset representing approximately 0.22% of the area. Validation set sizes were determined as 25% of the training samples per class, while the remaining patches in the tile, not included in the training or validation sets, comprised the test set. This methodology ensured a small but representative dataset for the full classification scheme.

To ensure the robustness of the evaluation, we conducted all tests five times, using distinct samplings for the training, validation, and test sets in each iteration. For reproducibility, we utilized predefined seed values for the random number generator (RNG). The evaluation metrics reported include Overall Accuracy (OA), mean F1 Score (mF1), and mean Intersection over Union (mIoU) on the test sets. Additionally, the standard deviation of these metrics was calculated as an unbiased estimator to provide insight into the consistency of the results. Table 3 shows the overall performance of the three deep architectures. In general, ConvLSTM and Transformer Encoder perform similarly, with no significant differences in overall performance metrics. On the other hand, Swin Transformer performs significantly better than the other, achieving the best performance overall.

Table 3. Comparison of the accuracies obtained by different architectures on the considered S2 tile. The best performance are in bold. The related standard deviation is reported in brackets.

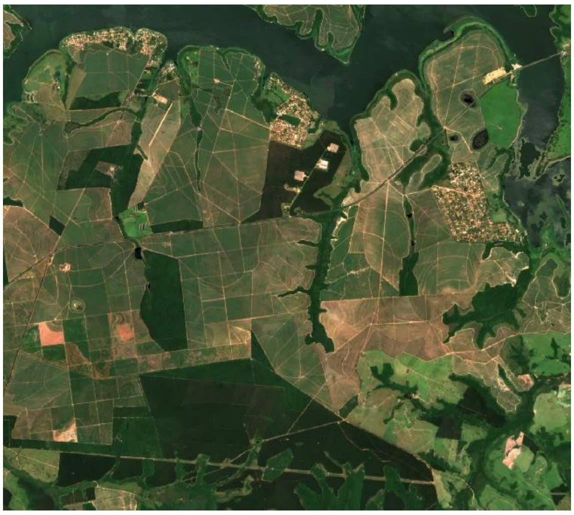
	mF1 (%)	OA (%)	mIoU (%)
Transformer Encoder	60.80 (0.59)	80.21 (1.06)	48.66 (0.55)
Swin Transformer	**65.42 (1.10)**	**83.53 (0.95)**	**53.14 (1.10)**
ConvLSTM	61.81 (1.38)	81.65 (1.03)	49.73 (1.35)

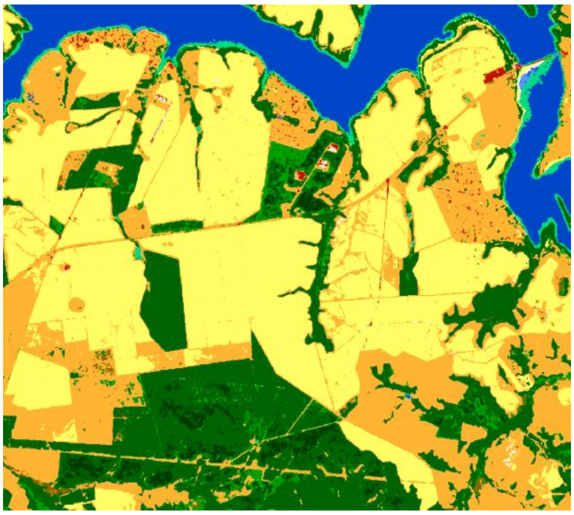
In order to validate class-wise accuracies, we selected a random seed to initialize the networks' weights before training for all the considered architectures. For each class, we computed the User's Accuracy (UA), the Producers Accuracy (PA) and the F1 Score. Table 4 shows the accuracies metrics of the different architectures. Swin Transformer is again the best performing architecture in terms of F1 scores, achieving the best performance on most classes. However, Transformer Encoder is the one achieving the best PA scores on most classes. These results highlight that the best performance is achieved by architectures employing the attention mechanisms to model the temporal information (i.e., Transformer Encoder and Swin Transformer), and that the best results are achieved by combining it with the spatial context (Swin Transformer). Indeed, the proper modelling of the temporal information helps in identifying the phenology of different land covers, while the spatial context provides more control over commission errors on categories with spatial features such as shrubs and build-up areas.

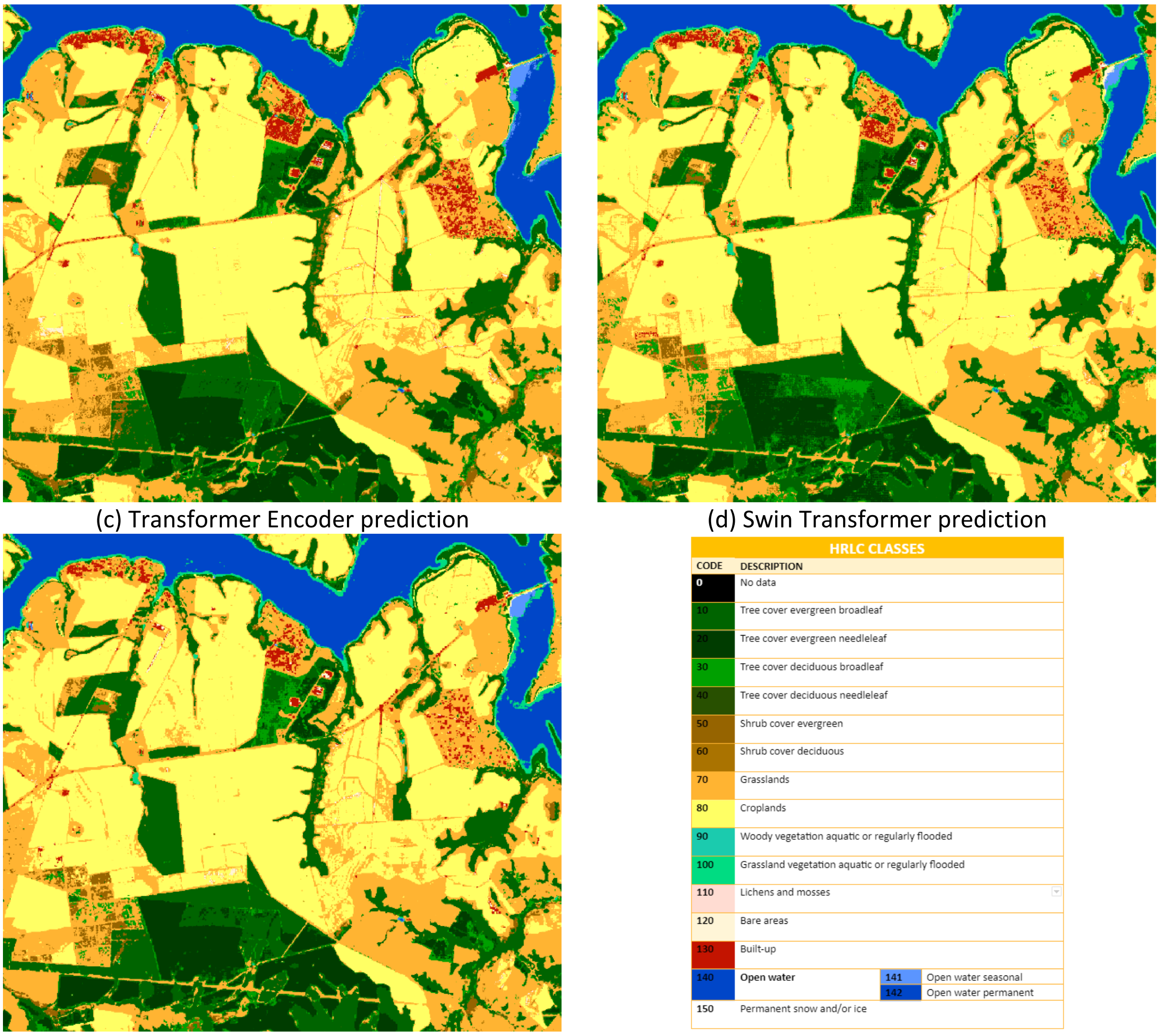
Table 4. Class-wise accuracy metrics obtained by different architectures on the considered S2 tile. The best performance are in bold.

	Transformer Encoder			ConvLSTM			Swin Transformer		
	UA	PA	F1	UA	PA	F1	UA	PA	F1
Tree cover evergreen broadleaf	**91.18%**	65.89%	76.50%	86.48%	74.50%	80.04%	85.33%	**76.93%**	**80.91%**
Tree cover evergreen needleleaf	46.95%	**96.47%**	63.16%	49.57%	91.90%	64.40%	**54.53%**	86.53%	**66.90%**
Tree cover deciduous broadleaf	43.66%	41.91%	42.77%	**45.67%**	44.54%	45.10%	43.38%	**57.03%**	**49.28%**
Shrub cover evergreen	25.10%	**81.68%**	38.40%	**38.58%**	56.37%	45.81%	35.62%	73.72%	**48.04%**
Shrub cover deciduous	**5.70%**	38.22%	**9.92%**	2.90%	37.48%	5.38%	5.10%	**41.98%**	9.09%
Grasslands	89.52%	**89.79%**	89.65%	89.72%	87.85%	88.77%	**92.95%**	87.05%	**89.90%**
Croplands	**94.35%**	**90.95%**	**92.62%**	91.66%	89.91%	90.77%	94.25%	90.88%	92.53%
Grassland vegetation aquatic	74.02%	74.29%	74.15%	72.18%	80.84%	76.27%	**76.01%**	**84.24%**	**79.91%**
Bare areas	23.30%	63.30%	34.06%	26.26%	55.74%	35.70%	**27.75%**	**73.38%**	**40.27%**
Built-up	42.49%	**84.32%**	56.51%	57.15%	78.33%	66.09%	**66.20%**	75.26%	**70.44%**
Open water seasonal	51.03%	**81.18%**	62.66%	57.19%	66.31%	61.41%	**58.98%**	74.52%	**65.84%**
Open water permanent	97.88%	**98.40%**	**98.14%**	**98.89%**	96.70%	97.78%	98.78%	97.22%	97.99%

Figure 5 shows a qualitative comparison of the predictions of the three deep architectures with the reference HRLC10 map. Notably, we can see that all the deep methods seem to better delineate the boundaries between the different land covers, especially for tree cover categories. Additionally, the deep learning methods seem to better capture the urban fabric, providing better results for the built-up class (especially the Transformer model). The Swin Transformer seems to provide noisy predictions in the example. However, the inference algorithm utilized with Swin Transformer was not optimized to perform proper semantic segmentation of the scene. Indeed, it process the image in a block-wise manner instead of using a rolling window approach. This leads to some checkboard artifacts, especially where the model is uncertain. Future analysis will include the qualitative results with proper inference algorithms, which will most likely improve the visual fidelity of the prediction maps of Swin Transformer.


(a) January median composite


(b) HRLC10 map



(c) Transformer Encoder prediction

(d) Swin Transformer prediction

(e) ConvLSTM prediction

HRLC CLASSES			
CODE	DESCRIPTION		
0	No data		
10	Tree cover evergreen broadleaf		
20	Tree cover evergreen needleleaf		
30	Tree cover deciduous broadleaf		
40	Tree cover deciduous needleleaf		
50	Shrub cover evergreen		
60	Shrub cover deciduous		
70	Grasslands		
80	Croplands		
90	Woody vegetation aquatic or regularly flooded		
100	Grassland vegetation aquatic or regularly flooded		
110	Lichens and mosses		
120	Bare areas		
130	Built-up		
140	**Open water**	141	Open water seasonal
		142	Open water permanent
150	Permanent snow and/or ice		

Figure 5. Qualitative comparison of the predictions from the different deep architectures considered. (a) RGB image of the January median composite of 2019. (b) Reference HRLC10 map. (c) Transformer Encoder prediction. (d) Swin Transformer prediction. (e) ConvLSTM prediction.

The preliminary analysis for the optical classification suggests that the use of deep learning architectures is promising, with results similar to the output of the full processing chain of HRLC10 (and in some cases qualitatively better). In particular, Swin Transformer and Transformer Encoder networks are promising options for improving the optical classification in Phase 2. Further analysis will focus on defining proper training strategies for defining deep networks able to operate at a large scale. This will include self-supervision approaches, as they have affirmed as the state-of-the-art approaches for model pre-training, and weak supervision approaches, in order to be able to augment the training dataset. Indeed, while the available photo interpreted points are fundamental for defining models that properly map the target land cover classes, the size of the training database is still scarce for training deep learning models.

3.2 SAR data processing

Figure 6 illustrates the processing workflow for generating the high-resolution (HR) land cover (LC) map through the classification of Sentinel-1 (S1) time series. Initially, the SAR images undergo preprocessing to convert the backscattered signal into σ_0, expressed in decibels (dB).

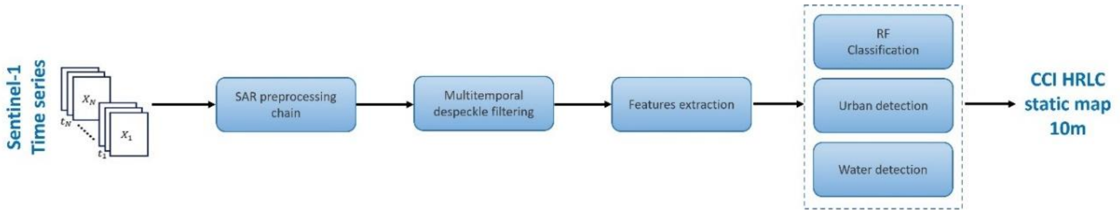


Figure 6. Block diagram of the SAR classification chain.

To determine the most suitable algorithm for the final SAR land cover classification process to be applied in Phase 2 of the project, a series of carefully designed experiments were conducted. These experiments aimed to assess the performance and accuracy of different approaches, with each outcome meticulously evaluated to ensure the optimal selection.

In more detail, the initial phase of testing employed the SAR classification chain, as illustrated in Figure 6, which had been developed during Phase 1 of the CCI+ project. This chain was used to generate preliminary classification outputs, providing an initial benchmark for performance. The Random Forest (RF) classifier was trained using the ground truth data collected by the team in Phase 1 through the hierarchical approach. Training points for the urban and water classes were excluded from the dataset.

However, recognizing the inherent limitations and challenges associated with traditional machine learning techniques for complex classification tasks, additional methods were considered. Specifically, alternative approaches leveraging Deep Learning (DL) techniques, Attention U-Net, Swin-Unet, and 3D FCN, were explored to address the limitations of the initial framework and potentially improve classification accuracy, particularly in dealing with the high-dimensional nature of SAR data. Figure 7 presents a simplified block diagram that outlines the key steps and processes involved in the system, providing a clear overview of the workflow and the interconnections between each stage of the procedure.

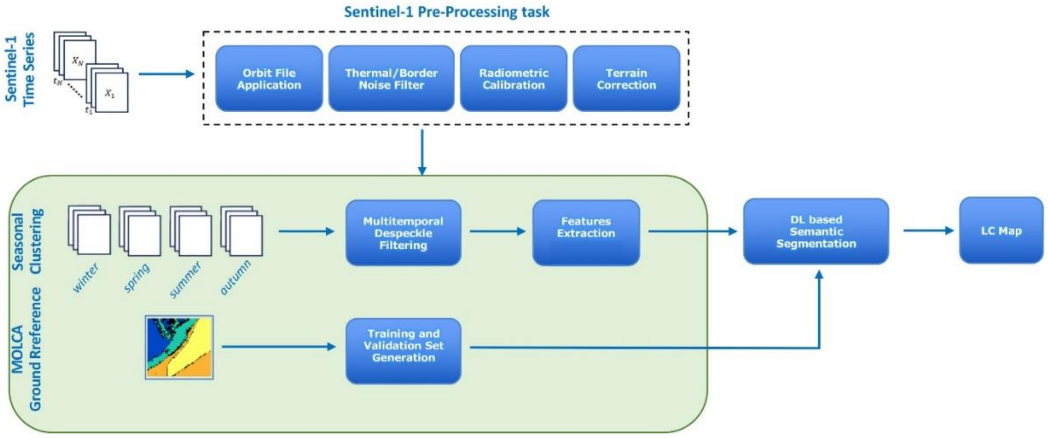


Figure 7. A simplified workflow illustrating the DL mapping procedure applied to SAR time series.

3.2.1 Data and Methods

A first test was carried out in the Amazonian region, specifically focusing on the 21KUQ tile, shown in Figure 8, which was selected as one of the benchmark areas during Phase 1 of the project.

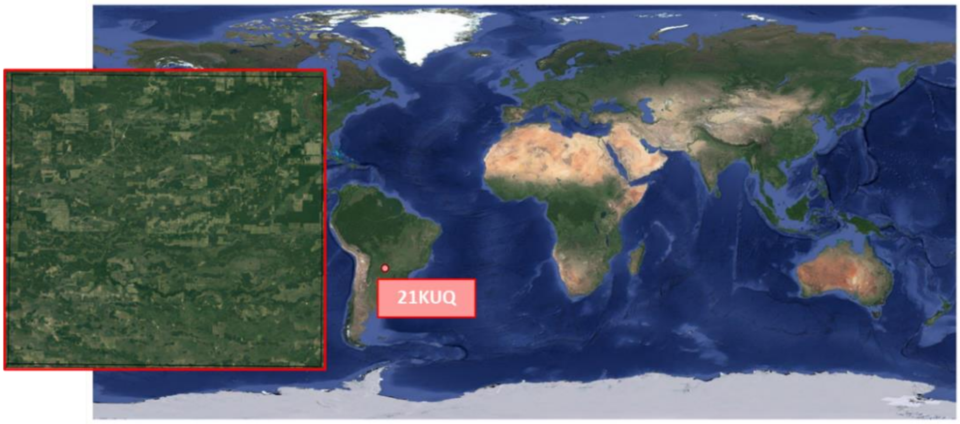


Figure 8. S2 tile 21KUQ in Amazonia, selected to test the enhancement of the RF classification.

This tile is particularly significant due to its diverse land cover types and ecological importance, making it an ideal test case for assessing the classification methods.

For this test region, the performance of the classification procedure based on the use of the RF algorithm was specifically evaluated, considering an expansion of the feature set compared to what was implemented during Phase 1 of the project.

S1 Level-1 Ground Range Detected (GRD) products, acquired in Interferometric Wide Swath (IW) mode, were employed for the analysis. The dataset spans the entire year 2019, offering a comprehensive temporal representation of land cover dynamics. The preprocessing stage constitutes a foundational step in generating a time series specifically tailored for land cover map production. S1 products are first subjected to radiometric calibration to ensure consistency in backscatter values, followed by terrain correction to account for topographical distortions. These steps are executed using the Sentinel Application Platform (SNAP) software provided by ESA, ensuring standardised and reliable data preparation.

For the multitemporal analysis, the images are stratified into the four annual seasons, winter, spring, summer, and autumn, to capture seasonal variations in land cover. For each seasonal subset, a multitemporal despeckling filter is applied to reduce noise while preserving significant spatial and temporal details. This process results in a "super image" for each season, representing an aggregated and noise-reduced view of the tile.

From each super image, a set of spatial features is extracted, including statistical metrics such as Lee, Min, Max, Max-Min, Mean, and Median, which provide valuable information about texture and spatial variability. These spatial features, combined with the super images themselves, form the inputs for training and executing a RF classification.

3.2.1.1 Random Forest classifier

Following Phase 1, a RF classifier was applied in this area. Just to quickly recall it, the RF algorithm has become a cornerstone in land cover mapping due to its robustness and adaptability, particularly when dealing with complex datasets such as those derived from S1 time series.

The classical RF approach, used as a baseline in Phase 1, relied on 28 features derived from four seasonal composites of S1 VH polarisation data collected in 2019. These features captured a range of spatial and statistical characteristics, including local edge enhancement, minimum and maximum values, the difference between them, as well as mean and median measures. While effective in providing a basic understanding of land cover patterns, this classical method was limited in its ability to distinguish subtle variations in the landscape. The need for more detailed classification motivated the development of an expanded feature set.

This initial test investigates potential improvements to the Phase 1 RF classification framework as per the previous paragraphs by incorporating an expanded feature set and implementing advanced preprocessing techniques, following the methodology outlined in [8]. These innovations aim to address the limitations of traditional approaches and provide more accurate and detailed maps, contributing to a better understanding of land cover dynamics in challenging environments like the Amazonian region. Specifically, the enhanced approach introduced data from both VH and VV polarisations, broadening the scope of information captured by the model. S1 images were grouped into 24-day intervals, and multitemporal despeckling was applied to reduce noise. For each period, an arithmetic mean composite was generated, adding 15 new features to the original set. This process resulted in a comprehensive 43-feature set, significantly increasing the algorithm’s potential to discern finer details in land cover. These additional features offered improved temporal resolution and better representation of surface dynamics, particularly in areas with rapid environmental changes. A simplified block scheme is depicted in Figure 9.

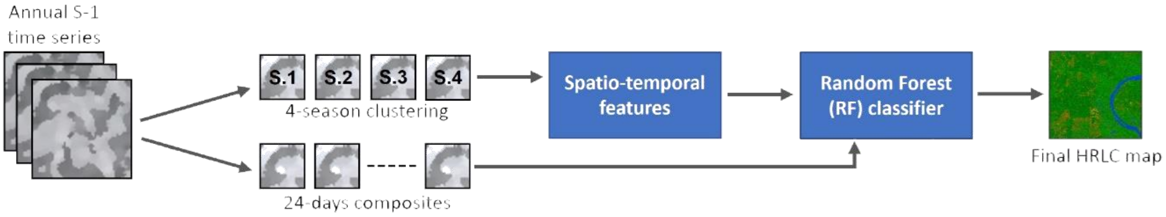
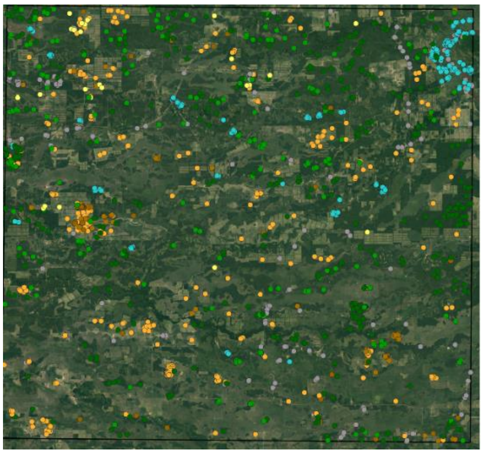


Figure 9. A simplified block diagram illustrating the land cover mapping procedure, highlighting enhancements to the feature pool.

Another critical aspect of the study was the application of an angular-based radiometric slope correction. This preprocessing step, inspired by the method proposed by Vollrath et [9], addressed terrain-induced distortions in S1 images. These distortions, often present in rugged terrains, can skew classification results by altering the apparent radiometric values of the imagery. By correcting for these angular effects, the study sought to enhance the reliability of the data, especially in regions like the Amazon, where topographic variability can significantly impact radiometric measurements.

As mentioned, this first experimental setup was centred on the Amazonian region, specifically the S2 21KUQ tile. This area presented a diverse range of land cover types, making it an ideal testbed for the proposed enhancements. The training data used consisted of points manually collected through a hierarchical approach during Phase 1 within the region, covering a spectrum of land cover classes, including evergreen forests, grasslands, croplands, and bare soil. Figure 10 shows the spatial distribution of tile 21KUQ in the Amazon region, along with the corresponding legend indicating the number of points for each class.



Value	Classes	Number of points
1	Tree cover everg. broadl.	439
3	Tree cover decid. broad.	199
5	Shrubland cover evergreen	38
6	Shrubland cover deciduous	82
7	Grassland	344
8	Cropland	27
10	Grassland vegetation aquatic or regularly flood	237
12	Bare areas	165

Figure 10. Spatial distribution and legend of the training points collected over the benchmark area 21KUQ in the Amazon region.

Three distinct classification experiments were conducted to assess the proposed enhancement under evaluation. The first experiment utilised the classical 28-feature set without any slope correction, serving as the control. The second incorporated the expanded 43-feature set, also without slope correction, to assess the impact of the additional features alone. The final experiment applied the expanded feature set with slope correction, examining the combined effect of the two enhancements. All classifications were implemented within Google Earth Engine, a platform that allowed efficient processing of large datasets.

Validation datasets for the quantitative assessment were sourced from well-established global products, including the Copernicus Global Land Cover (CGLC) at 100m 2019 [10], and the CCI Medium Resolution Land Cover (MRLC) Map 2019 [11], as well as a validation set provided by PoliMi and verified by UNITN, ensuring a robust comparison against external benchmarks. The validation points from existing global land cover products were randomly extracted to create a balanced and representative dataset for each land cover class. This approach ensures that the distribution of points adequately reflects the diversity of the classes, avoiding biases that could skew the classification accuracy assessment. Care was taken to select points from various geographic regions within the study area, capturing variations in land cover types due to environmental and climatic factors. The balanced dataset enables a robust comparison across classes, providing a comprehensive evaluation of the classification model's performance.

Table 5. List of validation datasets used to evaluate the classification performance based on the RF approach.

Validation set description	Class	Number of Points
PoliMi/UNITN Validation Set	Tree cover (evergreen broadleaf)	658
	Tree cover (deciduous broadleaf)	338
	Grassland	308
	Cropland	137
	Built-up	250
CGLC Map 100m 2019	Tree cover (evergreen broadleaf)	1000
	Tree cover (deciduous broadleaf)	1000
	Shrubland (evergreen)	1000
	Grassland	1000
	Cropland	1000
	Grassland veg. aq. or reg. flooded	1000
	Bareland	198
	Built-up	1000
	Open water (permanent)	1000
CCI MRLC Map 300m 2019	Tree cover (evergreen broadleaf)	1000
	Tree cover (deciduous broadleaf)	1000
	Grassland	1000
	Cropland	1000
	Grassland veg. aq. or reg. flooded	1000
	Open water (permanent)	1000

3.2.1.2 *Assessing the Impact of Enhancements on Random Forest Classification*

The results of the improved RF provide significant insights into the effectiveness of the proposed enhancements. Qualitative assessments of the classification maps show that the expanded feature set improves the visual clarity and spatial detail of the outputs. Features such as roads and bare areas within the study region are more distinctly identified, highlighting the value of the additional temporal composites. However, the application of slope correction introduces unexpected challenges. While it aimed to standardise radiometric values, it also increased noise in the classification process, particularly in homogeneous regions, leading to less consistent maps, as highlighted in Figure 11.

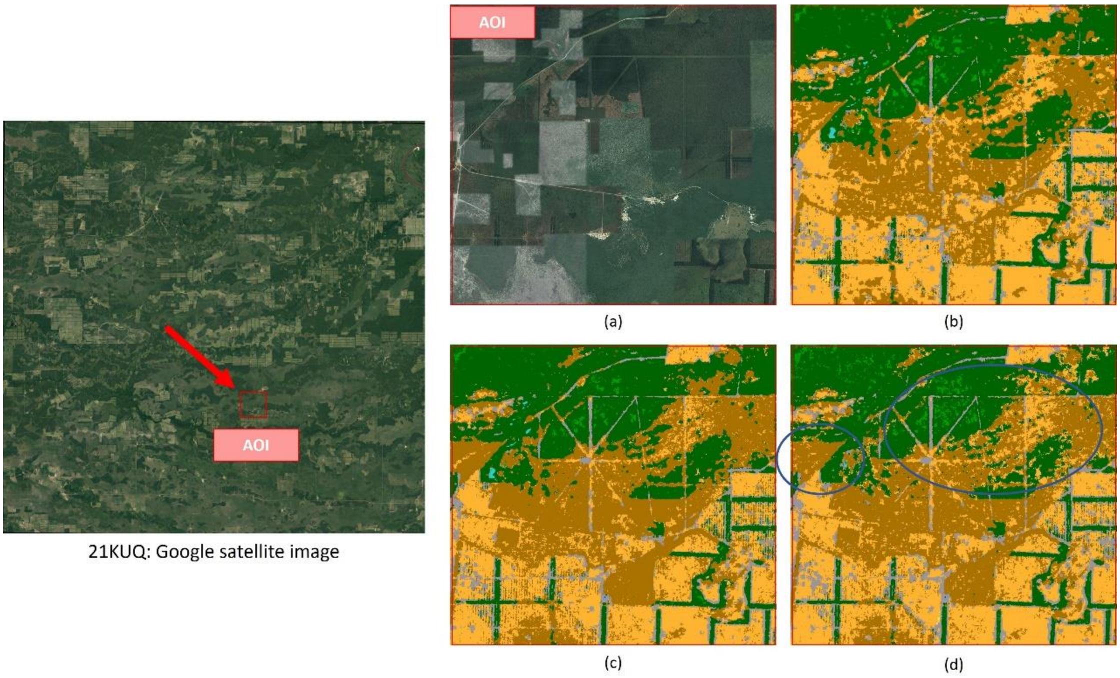


Figure 11. Visual comparison of the 21KUQ AOI tile between: (a) Google satellite image, (b) the RF map produced by the 'classical' approach, (c) S1 RF map with an extended feature set and no slope correction, and (d) S1 RF map with an extended feature set and slope correction.

A visual assessment in Figure 11 suggests that extending the feature space enhances classification performance, making spatial details more apparent. For instance, within the AOI, routes are more distinctly identified and classified as bare land. However, the slope correction appears ineffective in resolving class recognition confusion, resulting in a classification map that appears noisier. Notable examples of this issue are highlighted in the blue areas.

Quantitative analyses supported these findings, with OA metrics indicating a clear improvement when the expanded feature set was used without slope correction. This approach achieved higher differentiation between land cover classes and better alignment with validation datasets. Conversely, the slope-corrected classifications exhibited lower OA, underscoring the need for further refinement of the correction algorithm to avoid introducing artefacts into the data.

Table 6 compares the performance of different RF classification experiments across three validation datasets. The results are expressed in terms of Producer's Accuracy (PA$_X$, percentage accuracy per class x, where *x* indicates the class number) and OA, overall classification accuracy across all classes).

Table 6. Comparison of Producer's Accuracy (PA) for individual classes and Overall Accuracy (OA) across three validation datasets using different Random Forest (RF) experimental approaches: Classical Approach, Features Expanded, and Features Expanded and Slope Corrected.

Validation set	RF experiment	PA$_1$ (%)	PA$_2$ (%)	PA$_3$ (%)	PA$_4$ (%)	PA$_5$ (%)	PA$_6$ (%)	PA$_7$ (%)	PA$_8$(%)	PA$_9$(%)	PA$_{10}$(%)	PA$_{11}$(%)	OA(%)
PoliMi/UNITN Validation Set	Classical approach	56	75	-	-	51	87	-	-	-	-	-	54
	Features expandend	56	80	-	50	93	-	-	-	-	-	-	53
	Features expandend and Slope corrected	55	79	-	53	92	-	-	-	-	-	-	53
CGLC Map 100m 2019	Classical approach	42	37	17	-	29	40	25	8	-	-	91	29
	Features expandend	43	35	-	-	31	40	20	8	-	-	91	29
	Features expandend and Slope corrected	43	41	-	-	30	46	25	7	-	-	92	28
CCI MRLC Map 300m 2019	Classical approach	47	27	-	-	18	36	44	-	-	-	91	28
	Features expandend	47	27	-	-	18	31	40	-	-	-	91	27,6
	Features expandend and Slope corrected	47	22	-	-	18	42	48	-	-	-	91	27,5

*__PA__: *Producer Accuracy;* **1**: *Tree cover (evergreen broadleaf);* **2**: *Tree cover (deciduous broadleaf);* **3**: *Shrubland (evergreen);* **4**: *Shrubland (deciduous);* **5**: *Grassland;* **6**: *Cropland;* **7**: *Grassland vegetation aquatic or regularly flooded;* **8**: *Bareland;* **9**: *Built-up;* **10**: *Open water (seasonal);* **11**: *Open water (permanent).*
Note: The symbol '-' indicates the absence of validation points for a specific class.

Concerning the PoliMi/UNITN Validation Set, expanding features improves class accuracy, especially for PA5 (51% → 93%) compared to the classical approach. The slope-corrected approach achieves similar accuracy for PA5

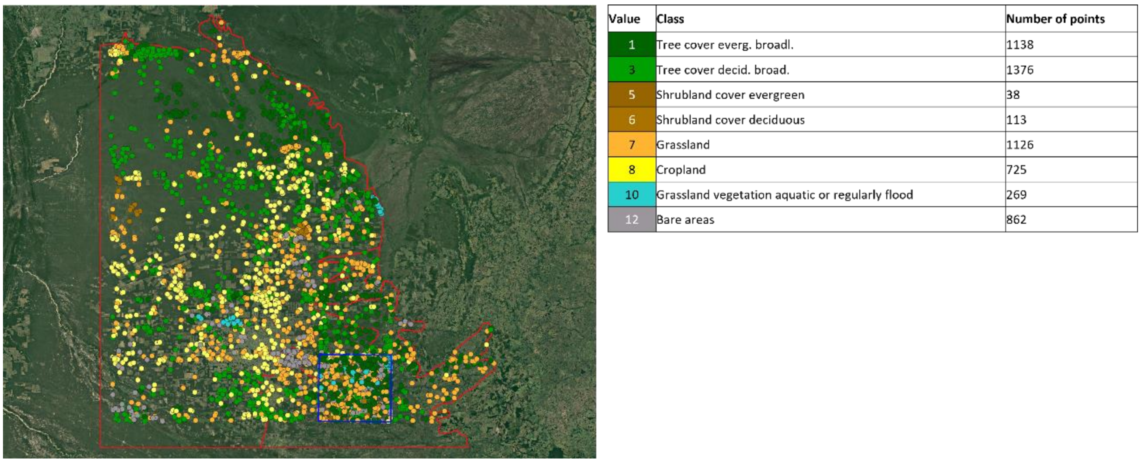
(92%). All approaches yield similar OA (53%-54%), suggesting that improvements in specific classes do not significantly affect global performance.

For the CGLC validation set, the slope-corrected approach slightly improves PA_6 (46%) compared to the classical approach (40%). Low accuracy is observed for classes such as PA_1 (42%-43%) and PA_2 (35%-41%). Incorporating slope corrections results in a marginal decrease in OA (28% vs. 29%).

For the CCI MRLC map, the slope-corrected approach enhances PA_7 (48%) compared to other methods. Other classes remain stable or decline slightly (e.g., PA_2 drops from 27% to 22%). The OA remains almost unchanged (27.5%-28%). In conclusion, expanding features or adding slope corrections improves the accuracy of certain classes (e.g., PA_5 and PA_7). However, OA remains relatively consistent across experimental approaches, indicating that these refinements do not translate into significant global improvements. Performance varies across validation datasets, with the PoliMi/UNITN dataset achieving the highest OA (53%-54%) compared to the other datasets (27.5%-29%).

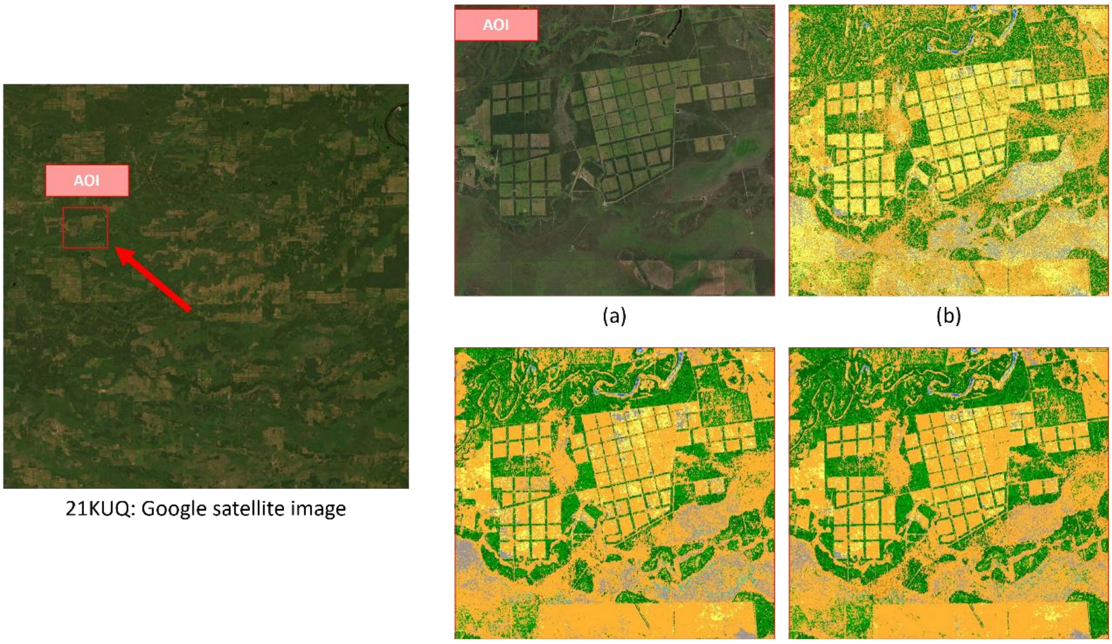
3.2.1.2.1 RF Performance Assessment with Ecoregion Classification

To improve classification performance, the RF model was trained using training points manually collected by the EO team and organised by Ecoregions, as shown in Figure 12. This approach takes advantage of the ecological distinctions specific to each Ecoregion, allowing for a more tailored and accurate land cover classification. The model was applied across all three scenarios under consideration, ensuring comprehensive coverage of varying ecological contexts. The performance of this Ecoregion-based RF model was evaluated using the same three validation datasets employed in prior experiments. The Ecoregion (ER)-based training and classification process was carried out using Google Earth Engine (GEE), and the resulting classification maps are in Figure 13.



Value	Class	Number of points
1	Tree cover everg. broadl.	1138
3	Tree cover decid. broad.	1376
5	Shrubland cover evergreen	38
6	Shrubland cover deciduous	113
7	Grassland	1126
8	Cropland	725
10	Grassland vegetation aquatic or regularly flood	269
12	Bare areas	862

Figure 12. Spatial distribution and legend of the training points collected across the corresponding Ecoregions in the Amazon region.



21KUQ: Google satellite image

(a)

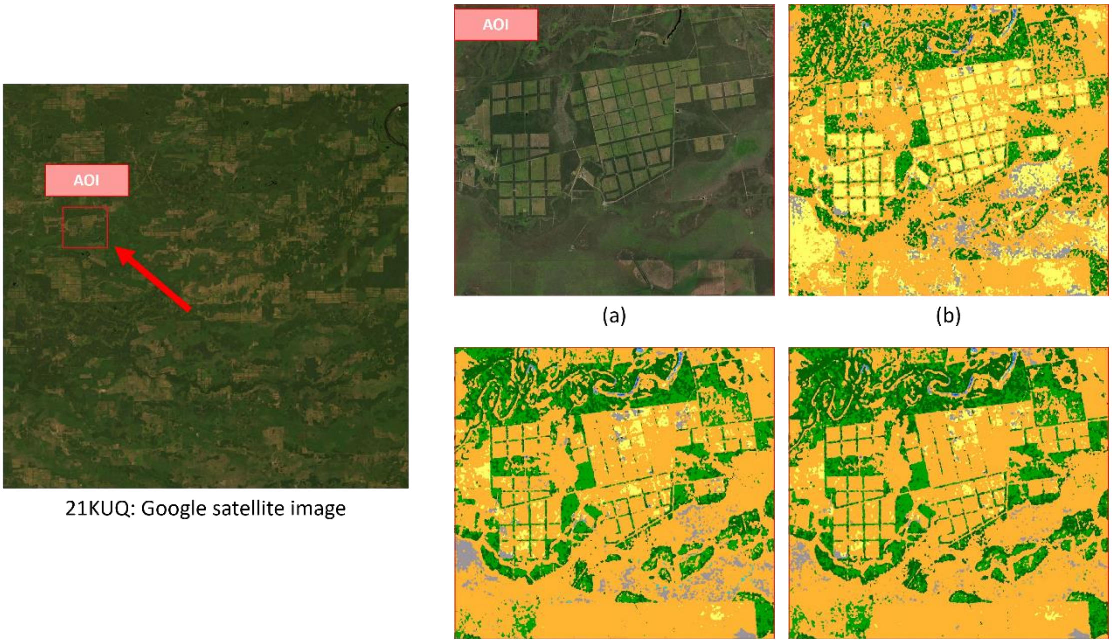
(b)

(c)

(d)

Figure 13. Visual comparison of the 21KUQ AOI tile between: (a) Google satellite image, (b) the RF map produced by the 'classical' approach ER-based, (c) the S1 RF map with an extended feature set and no slope correction ER-based, and (d) S1 RF map with an extended feature set and slope correction ER-based.

Furthermore, to mitigate residual misclassification errors and refine the output, a neighborhood reduction filter was applied as a post-processing step, with the resulting classification maps shown in Figure 14. This filter aims to smooth classification outputs by addressing isolated misclassified pixels, thereby enhancing spatial coherence and improving the overall map quality.



21KUQ: Google satellite image

(a)

(b)

(c)

(d)

Figure 14. Visual comparison of the 21KUQ AOI tile between: (a) Google satellite image, (b) the RF map produced by the 'classical' approach ER-based, neighbourhood reduced (c) the S1 RF map with an extended feature set and no slope correction ER-based, neighbourhood reduced, and (d) S1 RF map with an extended feature set and slope correction ER-based, neighbourhood reduced.

This additional step is particularly effective in reducing noise in areas where class transitions are abrupt or where small misclassified patches might distort the final results. This technique significantly improved map homogeneity, reducing the 'salt and pepper' effect often observed in RF outputs. The filter proved particularly effective for the extended feature sets, further enhancing the usability of the classification results for practical applications.

Table 7 presents the results of classification performance for the different RF experiments, where training was based on Ecoregion (ER) data. The experiments were evaluated using the three validation sets used previously. The performance of the RF model is measured by PA for each class (PA₁, PA₂, PA₃, etc.) and OA.

Table 7. Comparison of Producer's Accuracy (PA) for individual classes and Overall Accuracy (OA) across three validation datasets using different Random Forest (RF) experimental approaches ER-based: Classical Approach, Features Expanded, and Features Expanded and Slope Corrected. The performances have been also compared with their corresponding neighborhood reduced filtering version.

Validation set	RF experiment	PA_1 (%)	PA_2 (%)	PA_3 (%)	PA_4 (%)	PA_5 (%)	PA_6 (%)	PA_7 (%)	PA_8(%)	PA_9(%)	PA_{10}(%)	PA_{11}(%)	OA(%)
PoliMi/UNITN Validation Set	Classical approach ER-based	46	48	-	-	40	47	-	-	-	-	-	42,6
	Classical approach ER-based, neigh. reduced	52	55	-	-	68	39	-	-	-	-	-	51,1
	Features expandend ER-based	56	59	-	-	55	55	-	-	-	-	-	55,4
	Features expandend ER-based neigh. reduced	61	64	-	-	56	55	-	-	-	-	-	59
	Features expandend ER-based, Slope corrected	53	55	-	-	56	60	-	-	-	-	-	53
	Features expandend ER-based, Slope corrected, neigh. reduced	55	59	-	-	59	56	-	-	-	-	-	54,1
CGLC Map 100m 2019	Classical approach ER-based	36	26	5	-	18	33	12	3	-	-	91	24,6
	Classical approach ER-based, neigh. reduced	48	35	-	-	17	48	17	12	-	-	91	28
	Features expandend ER-based	40	26	-	-	21	34	15	6	-	-	91	25,8
	Features expandend ER-based neigh. reduced	47	29	-	-	20	36	15	11	-	-	91	26,6
	Features expandend ER-based, Slope corrected	37	27	-	-	27	35	15	7	-	-	92	25,9
	Features expandend ER-based, Slope corrected, neigh. reduced	47	29	-	-	20	36	15	11	-	-	91	26,6
CCI MRLC Map 300m 2019	Classical approach ER-based	40	22	-	-	23	33	28	-	-	-	91	27,9
	Classical approach ER-based, neigh. reduced	47	25	-	-	25	35	30	-	-	-	91	30,4
	Features expandend ER-based	46	29	-	-	27	43	36	-	-	-	91	31,9
	Features expandend ER-based neigh. reduced	51	32	-	-	26	54	36	-	-	-	91	32,9
	Features expandend ER-based, Slope corrected	44	28	-	-	27	36	36	-	-	-	91	31
	Features expandend ER-based, Slope corrected, neigh. reduced	50	33	-	-	27	50	36	-	-	-	91	33,2

*__PA__: *Producer Accuracy;* **1**: *Tree cover (evergreen broadleaf);* **2**: *Tree cover (deciduous broadleaf);* **3**: *Shrubland (evergreen);* **4**: *Shrubland (deciduous);* **5**: *Grassland;* **6**: *Cropland;* **7**: *Grassland vegetation aquatic or regularly flooded;* **8**: *Bareland;* **9**: *Built-up;* **10**: *Open water (seasonal);* **11**: *Open water (permanent).*
Note: The symbol '-' indicates the absence of validation points for a specific class.

For the PoliMi/UNITN Validation Set, the highest OA (OA = 59%) is achieved by the "Features expanded ER-based neighbourhood reduced" experiment, indicating that adding feature expansion and applying the neighbourhood reduction filter leads to improved classification performance. The "Classical approach ER-based" experiment shows lower OA (OA = 42.6%), suggesting that using a classical approach without additional post-processing or feature expansion does not perform as well. "Features expanded ER-based" and "Slope corrected" methods provide relatively better results (OA = 55.4% and OA = 53%, respectively).

Using the CGLC set, the "Classical approach ER-based, neigh. reduced" experiment shows the best performance with an OA of 28%, followed by "Features expanded ER-based neigh. reduced" with an OA of 26.6%. Adding the neighbourhood reduction filter improves the classification, as seen in experiments like "Classical approach ER-based, neigh. reduced" (OA = 28%) and "Features expanded ER-based neigh. reduced" (OA = 26.6%). The "Features expanded ER-based" experiment without neighbourhood reduction or slope correction results in lower accuracy (OA = 25.8%).

With the CCI MRLC validation points, the best performance is observed in the "Features expanded ER-based neigh. Reduced" experiment, which achieves an OA of 32.9%, indicating that this combination of methods (feature expansion with the neighbourhood reduction filter) delivers the most accurate classification results. "Classical approach ER-based, neigh. reduced" (OA = 30.4%) and "Features expanded ER-based" (OA = 31.9%) also perform well, showing that adding neighbourhood reduction and expanding features consistently improves results. The lowest OA (OA = 27.9%) is obtained from the "Classical approach ER-based" experiment, underlining the importance of additional processing steps like feature expansion and neighbourhood reduction for improved accuracy.

The application of feature expansion and neighbourhood reduction significantly enhances classification accuracy, particularly when combined with Ecoregion-based training. The Slope corrected method provides some improvements in accuracy, but not as consistently as the neighbourhood reduction filter.

The Classical approach without feature expansion or post-processing methods yields the lowest accuracy, indicating that relying solely on basic RF techniques may not be sufficient for accurate classification across diverse land cover types. This analysis highlights the importance of data preprocessing and model refinement (e.g., feature expansion, slope correction, and neighbourhood reduction) in improving the performance of land cover classification using RF models.

Some classes consistently exhibit low PA values, which may indicate underlying issues related to the distribution of training samples or confusion between similar classes. These low PA values, coupled with equally low OA scores, suggest that the current RF approach may not be capturing the complexities of certain land cover types effectively. Such results point to the necessity of exploring alternative, more powerful methods for land cover classification. In particular, DL approaches, with their ability to model complex spatial relationships and learn hierarchical features directly from raw data, may offer significant improvements in performance. DL methods.

3.2.1.3 Deep Learning Models as an alternative to RF for LC Classification

The performance of the 'classical' RF classification, as used in Phase 1 and based on 28 features per season, was also evaluated in comparison with three advanced DL algorithms: Attention U-Net, Swin-Unet, and 3D Fully Convolutional Network (3D FCN).

For the training of the three DL algorithms the channel axis of the input tensors was utilised to encode the spatio-temporal information derived from the sequence of seasonal features extracted from the original SAR images. This approach allowed the networks to process the temporal dynamics effectively. The input tensors were structured as $B \times T \times W \times H$, where:

- B represents the batch size,
- T corresponds to the temporal dimension (seasonal features),
- W and H denote the width and height of the input images, respectively.

Training was conducted over 30 epochs, ensuring sufficient iterations for the models to converge. The remaining parameters were configured as described in reference [12]. Specifically:

- **Learning rate**: 10^{-4}, optimised for gradual and stable learning,
- **Batch size**: 1, chosen to efficiently handle the 28-feature temporal sequence while managing memory constraints.
- **Optimizer**: *Adam*, used for its effectiveness in handling sparse gradients and adaptive learning rates.
- **Loss function**: *Categorical Cross-Entropy*, applied to optimise for multiclass classification.

This configuration ensured the models could capture the intricate spatio-temporal relationships inherent in the seasonal SAR features, thereby enhancing their performance in classifying diverse land cover types.

The DL training set was extracted from the MOLCA dataset [13], which provides a comprehensive collection of labeled land cover data. This dataset includes various land cover classes with high spatial resolution, and it was specifically selected to ensure the availability of accurate training samples. The methodology for selecting and processing this training data is described in detail in the ATDB (Algorithm Technical Description Book) [AD6], where the steps for data extraction, sample refinement, and class assignment are outlined to ensure high-quality input for the DL model.

These algorithms were applied not only to the previously mentioned Amazonia tile, but to three static and geographically diverse areas representing the Amazon, Africa, and Siberia regions. These regions, selected during Phase 1 of the project for their distinct ecological and land cover characteristics, are illustrated in Figure 15.

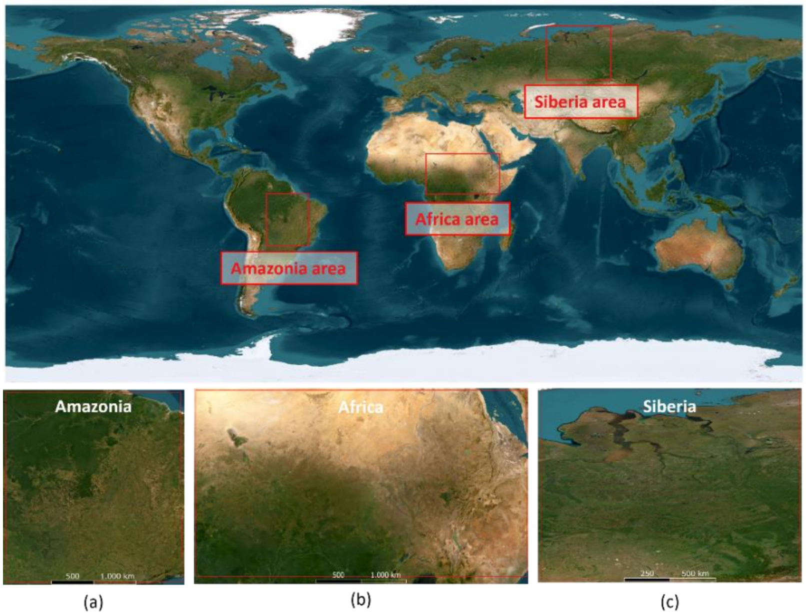


Figure 15. Views of the static areas identified in Phase 1: (a) Amazonia (62.1014° W, 23.5983° S to 42.9441° W, 0° N, WGS 84), (b) Africa (9.8986° E, 0.0885° S to 43.2908° E, 18.0891° N, WGS 84), and (c) Siberia (64.4361° E, 51.2789° N to 93.4017° E, 75.6847° N, WGS 84).

The performance of the classification models was rigorously compared across the three static areas., each represented by a subset of S2 tiles listed in Table 8 and shown in Figure 16.

Table 8. List of tiles selected to assess the DL classification performance for each static area: Amazonia (86 S2 tiles), Africa (103 S2 tiles), and Siberia (64 S2 tiles).

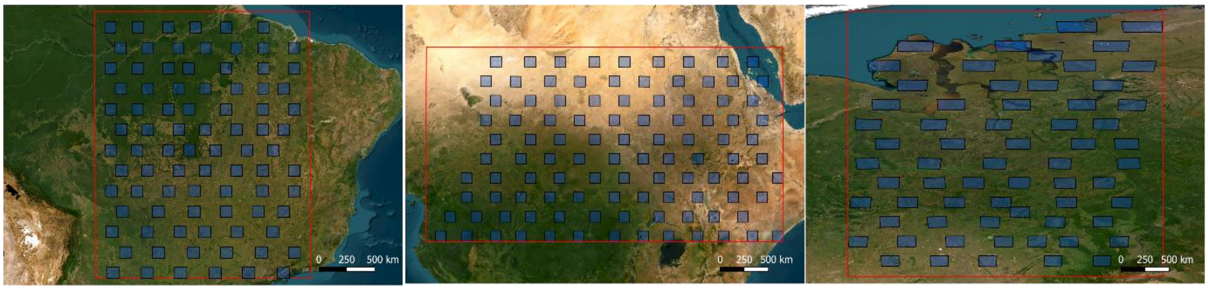
Amazonia	Africa	Siberia
20KQD	32NQF	41UPA
20KQV	32NRH	41VPE
20KRF	33NUA	41WPN
20LQH	33NUE	42UVD
20LQM	33NUG	42UWF
20LQR	33NVC	42VUP
20LRK	33NWJ	42VVH
20LRP	33NXA	42VVM
20MQD	33NXE	42VVR
20MQV	33NXG	42VWK
20MRB	33NYC	42VXP
20MRT	33PWN	42WVD
21KTS	33PWS	42WVV
21KVQ	33PXL	42WWB
21KVU	33PXQ	42WXT
21KWA	33PZN	42XWF
21KWS	33PZS	43UCU
21KYQ	33QXU	43UDA
21KYU	34NBF	43UFU
21LUC	34NBK	43VCC
21LUG	34NBP	43VCG
21LUL	34NCH	43VDE

21LVE
21LVJ
21LXC
21LXG
21LXL
21LYE
21LYJ
21LZG
21LZL
21MUQ
21MUU
21MVN
21MVS
21MXQ
21MXU
21MYN
21MYS
21MZQ
22KBB
22KBF
22KCD
22KDV
22KEB
22KEF
22KFD
22KGV
22KHF
22LBH
22LCK
22LCP
22LDM
22LEH
22LER
22LFK
22LFP
22LGM
22MBD
22MCB
22MCT
22MED
22MEV
22MFB
22MFT
23KKS
23KLU
23KMA
23KMQ
23KNS
23LKC

34NCM
34NEF
34NEK
34NEP
34NFH
34NFM
34NHF
34NHK
34NHP
34PCR
34PCV
34PEB
34PET
34PFR
34PFV
34PHB
34PHT
34QCD
34QFD
35NLC
35NLG
35NNA
35NNE
35NNJ
35NPC
35NPG
35NRA
35NRE
35NRJ
35PLL
35PLQ
35PNN
35PNS
35PPL
35PPQ
35PRN
35PRS
35QLU
35QPU
36NUH
36NUM
36NVF
36NVK
36NVP
36NXH
36NXM
36NYF
36NYK
36NYP

43VDL
43VEG
43VFC
43VFJ
43WDQ
43WEU
43WFN
43WFS
44ULF
44UNF
44UPD
44VLK
44VMJ
44VMM
44VMR
44VPH
44VPP
44WMV
44WNB
44WNE
44WPT
44XMF
45UUA
45UWU
45UXA
45VUE
45VUG
45VVL
45VWC
45VWG
45VXE
45WVQ
45WVU
45WWN
45WWS
45WXQ
45WXU
45XVC
45XWA
46VCP
46VCR
46XDH

23LKE	36PUR	
23LKJ	36PUV	
23LKL	36PWB	
23LLG	36PWT	
23LMJ	36PXR	
23LNC	36PXV	
23LNE	36PZB	
23LNL	36PZT	
23MKN	36QUD	
23MKQ	36QXD	
23MKS	37NBC	
23MKU	37NBG	
23MMN	37NCA	
23MNQ	37NDE	
23MNS	37NDJ	
	37NEC	
	37NEG	
	37NFA	
	37NGE	
	37NGJ	
	37PCL	
	37PCQ	
	37PDN	
	37PDS	
	37PFL	
	37PFQ	
	37PGN	
	37PGS	
	37QCU	
	37QFU	
	38NKF	
	38NKM	



(a) (b) (c)

Figure 16. Visual representation of the spatial distribution of the tiles selected for (a) Amazonia, (b) Africa, and (c) Siberia, to assess the classification performance of the DL networks.

For the comparison, the S1 Level-1 GRD dataset from 2021 was utilised. The dataset was segmented into four subsequences corresponding to the seasons (winter, spring, summer, and autumn). This seasonal division provided a comprehensive temporal perspective for analysing land cover dynamics across diverse regions.
The resulting dataset included:

- **Amazonia**: for 86 S2 tiles, a total of 5105 SAR images, distributed as 1264 for winter, 1310 for spring, 1338 for summer, and 1193 for autumn.
- **Africa**: A collection of 5827 SAR images for 103 S2 tiles, comprising 1470 for winter, 1480 for spring, 1654 for summer, and 1405 for autumn.
- **Siberia**: For 64 S2 tiles, a dataset of 3396 SAR images, with 768 for winter, 744 for spring, 984 for summer, and 900 for autumn.

All images were acquired in IW mode with VH polarisation and descending orbit configuration.

Each algorithm was evaluated in terms of its ability to accurately classify land cover types, with particular attention to how well each model handled the unique characteristics of the SAR data in these diverse regions. This evaluation aimed at determining the most effective DL model for the classification in the Phase 2 of the project, ensuring robust and reliable land cover maps across the selected global regions.

3.2.1.4 Comparative Performance Analysis of DL Models and RF for LC Classification

Table 9 displays the performance of different three DL models for the three regions: Amazonia, Africa, and Siberia. As a general statement, Swin-Unet consistently outperforms the benchmark CNN-based models (3D-FCN and Attention U-Net) and the RF model in terms of OA, kappa, and F1-score across all three regions. While RF remains robust in simpler scenarios, DL models outperform it by leveraging their ability to learn hierarchical and contextual features [14].

Table 9. Overall Accuracy (OA), Kappa Coefficient, F1-Score, and Producer Accuracy (PA) for the evaluated models across different regions.

Region	Model	O.A.	kappa	F1-Score	pa_1	pa_2	pa_3	pa_4	pa_5	pa_6	pa_7	pa_8
Amazonia	Random Forest	0.622	0.328	0.565	0.51	0.09	0.25	0.07	0.18	0	0.02	0.18
	3D-FCN	0.818	0.709	0.795	0.89	0.08	0.49	0.74	0.24	0	0.06	0.75
	Attention U-Net	0.843	0.760	0.832	0.91	0	0.61	0.6	0	0	0	0.96
	Swin-Unet	**0.933**	0.898	0.924	0.97	0.5	0.88	0.85	0.03	0	0.85	0.99
Africa	Random Forest	0.745	0.544	0.712	0.64	0.11	0.13	0.32	0.07	0.69	0	0.64
	3D-FCN	0.735	0.563	0.723	0.46	0.27	0.44	0.53	0.11	0.36	0.2	0.77
	Attention U-Net	0.738	0.596	0.751	0.66	0	0.55	0.18	0	0.02	0	0.89
	Swin-Unet	**0.936**	0.900	0.932	0.97	0.52	0.83	0.73	0.16	0.93	0.09	0.77
Siberia	Random Forest	0.677	0.362	0.634	0.31	0	0.34	0.18	0.06	0	0.18	0.42
	3D-FCN	0.854	0.760	0.842	0.87	0	0.58	0.75	0.21	0.16	0.01	0.87
	Attention U-Net	0.903	0.847	0.885	0.97	0	0.42	0.61	0.56	0.06	0.01	0.99
	Swin-Unet	**0.974**	0.959	0.974	0.99	0	0.86	0.95	0.79	0.75	0.92	0.98

*__pa__: *Producer Accuracy;* **1**: *Forest;* **2**: *Shrubland;* **3**: *Grassland;* **4**: *Cropland;* **5**: *Wetland;* **6**: *Bareland;* **7**: *Built-up;* **8**: *Water.*

A visual comparison of the results from the Swin-Unet model, including random S1 super images from the validation dataset, MOLCA reference data (Ground Truth), and the corresponding predictions, is illustrated in the Figure 17, Figure 18 and Figure 19, for Amazonia, Africa and Siberia, respectively.

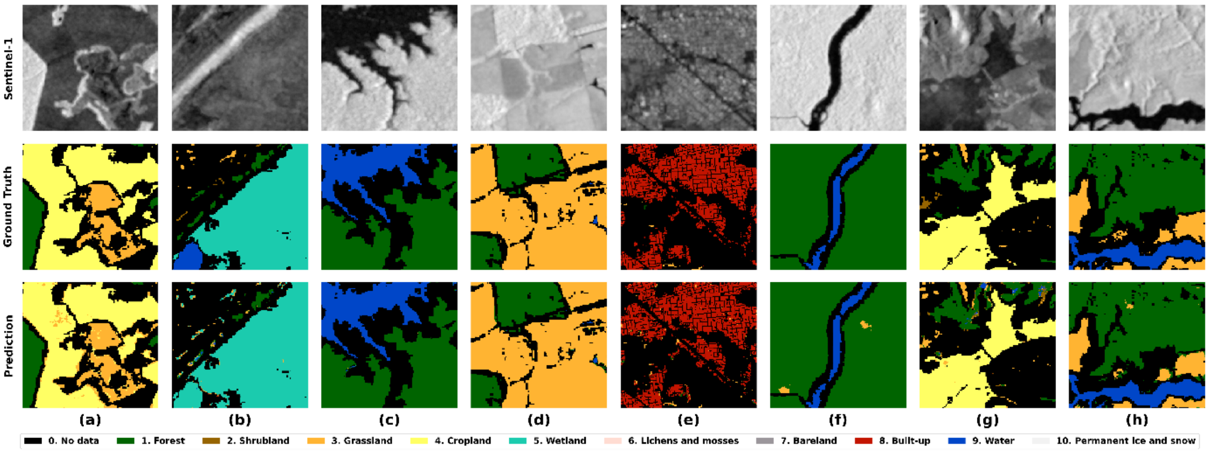


Figure 17. Visual comparison for the Amazonia region, showing Sentinel-1 super image data (top row), Ground Truth

(middle row), and predicted patches (bottom row).

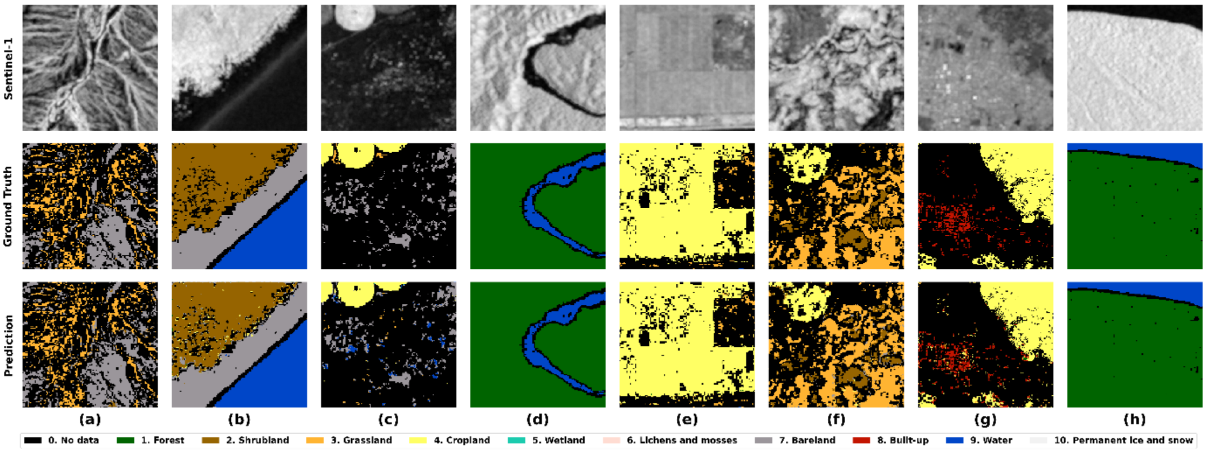


Figure 18. Visual comparison for the Africa region, showing Sentinel-1 super image data (top row), Ground Truth (middle row), and predicted patches (bottom row).

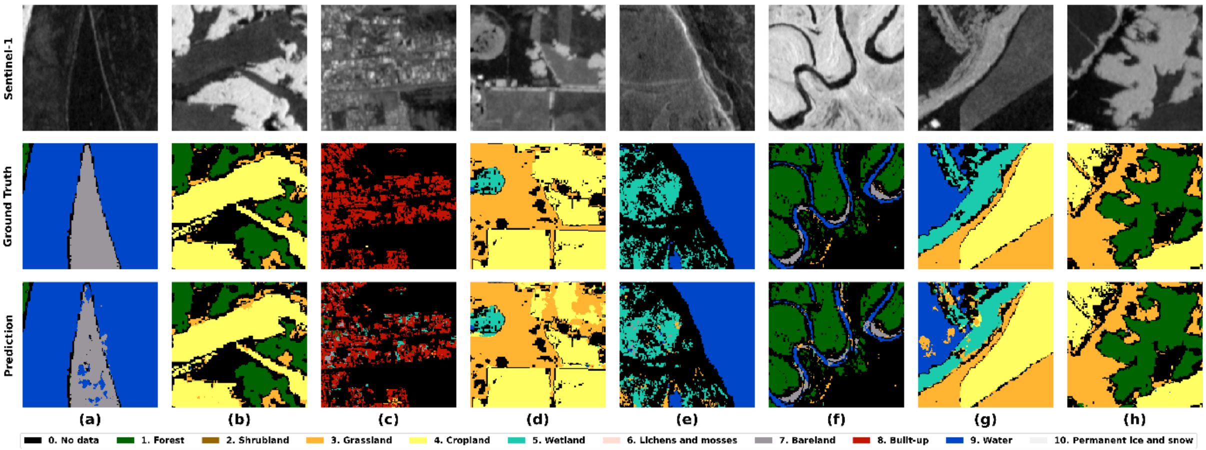
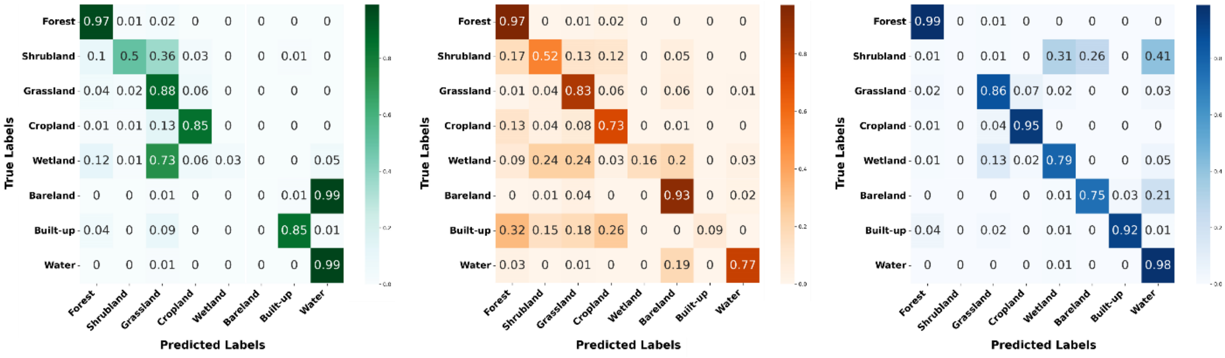


Figure 19. Visual comparison for the Siberia region, showing Sentinel-1 super image data (top row), Ground Truth (middle row), and predicted patches (bottom row).

In these regions, the Swin-Unet model achieves OAs of 93.3%, 93.6%, and 97.4%, respectively. The corresponding confusion matrices are presented in the Figure 20, with each diagonal element indicating the PA for the respective class. Notably, the results table indicates that the Forest, Grassland, Cropland, and Water classes are effectively extracted using the leading model across all three study areas.



(a)

	Forest	Shrubland	Grassland	Cropland	Wetland	Bareland	Built-up	Water
Forest	0.97	0.01	0.02	0	0	0	0	0
Shrubland	0.1	0.5	0.36	0.03	0	0	0.01	0
Grassland	0.04	0.02	0.88	0.06	0	0	0	0
Cropland	0.01	0.01	0.13	0.85	0	0	0	0
Wetland	0.12	0.01	0.73	0.06	0.03	0	0	0.05
Bareland	0	0	0.01	0	0	0	0.01	0.99
Built-up	0.04	0	0.09	0	0	0	0.85	0.01
Water	0	0	0.01	0	0	0	0	0.99

(b)

	Forest	Shrubland	Grassland	Cropland	Wetland	Bareland	Built-up	Water
Forest	0.97	0	0.01	0.02	0	0	0	0
Shrubland	0.17	0.52	0.13	0.12	0	0.05	0	0
Grassland	0.01	0.04	0.83	0.06	0	0.06	0	0.01
Cropland	0.13	0.04	0.08	0.73	0	0.01	0	0
Wetland	0.09	0.24	0.24	0.03	0.16	0.2	0	0.03
Bareland	0	0.01	0.04	0	0	0.93	0	0.02
Built-up	0.32	0.15	0.18	0.26	0	0	0.09	0
Water	0.03	0	0.01	0	0	0.19	0	0.77

(c)

	Forest	Shrubland	Grassland	Cropland	Wetland	Bareland	Built-up	Water
Forest	0.99	0	0.01	0	0	0	0	0
Shrubland	0.01	0	0.01	0	0.31	0.26	0	0.41
Grassland	0.02	0	0.86	0.07	0.02	0	0	0.03
Cropland	0.01	0	0.04	0.95	0	0	0	0
Wetland	0.01	0	0.13	0.02	0.79	0	0	0.05
Bareland	0	0	0	0	0.01	0.75	0.03	0.21
Built-up	0.04	0	0.02	0	0.01	0	0.92	0.01
Water	0	0	0	0	0.01	0	0	0.98

Figure 20. Normalized confusion matrices derived from the validation set for the regions: (a) Amazonia, (b) Africa and (c) Siberia.

The evaluation of the models on the African dataset indicates that CNN-based approaches may have difficulty differentiating Bareland from classes such as Water or Grassland within SAR imagery. In Africa, the 3D-FCN model scores 0.77 for Water, but only 0.36 for Bareland, 0.53 for Cropland, and 0.44 for Grassland. The Attention U-Net framework, on the other hand, achieves a PA of 0.89 for Water, 0.66 for Forest, and 0.55 for Grassland, but only 0.02 for Bareland. In contrast, the Swin-Unet achieves a final PA of 0.97 for Forest, 0.83 for Grassland, and 0.93 for Bareland, surpassing the other two models.

The improvement in recognizing Bareland with the Swin-Unet model is evident not only in Africa but also in the Siberian dataset. Despite representing only 1.04% of the available tiles, Bareland achieves a PA of 0.75 with this model, marking a substantial enhancement from the initial values of 0.16 and 0.06 recorded with the 3D-FCN and Attention U-Net, respectively. These outcomes suggest that the transformer model is capable of recognizing various LC classes and tends to make more balanced classifications compared to the other two CNN-based models, which generally show high PA only for specific classes like Water. This difference may stem from the Attention U-Net model's limited ability to generalize, which results in strong classification performance primarily for classes with easily identifiable spatial patterns or unique brightness values, such as Water or Forest. However, it struggles with more complex morphological relationships of specific LC classes like Bareland. Consequently, the Attention U-Net, which depends on locality-based attention mechanisms, may find it challenging to classify classes that share similar pixel value distributions. Regarding the 3D-FCN model, its three-dimensional logic-based structure does not provide additional benefits when the input data lacks a dense temporal sequence and instead consists of seasonal synthetic images, referred to as features. In this context, a 2D-CNN model, like the Attention U-Net, appears to be sufficient. In contrast, transformer-based models like Swin-Unet consider global pixel relationships, leading to a more accurate assessment of context and environment. The global attention mechanisms in these models can help recover this context, which CNNs cannot achieve due to the local nature of the convolutional kernel. Furthermore, the Built-up class showed significant improvement through the Swin-Unet model in Amazonia and Siberia with final values of 0.85 and 0.92. For the African region, the value for this specific class remains poor, likely due to the lower number of representative labels for the this class in this area (only 0.5%). Another contributing factor could be the small and fragmented nature of urban areas, particularly when compared to more spatially uniform LC classes like Forest, Grassland, or Cropland. In favor of these latter classes, the highest confusion is observed when predicting Built-up, indicating a significant challenge in visually distinguishing small urban areas from surrounding classes due to their reduced size and scattered appearance. For Wetland, a high PA of 0.79 is obtained in the Siberian region, while for the other two areas, the accuracy remains

For both Amazonia and Siberia, RF achieves lower OA than all three DL approaches. In Africa, RF reaches an OA of 0.745, slightly exceeding the OA of 3D-FCN and Attention U-Net. However, its Kappa coefficient (0.544) and F1-Score (0.712) remain lower than those of these DL models. The Swin-Unet achieves the highest metrics overall, with an OA of 0.936, a Kappa coefficient of 0.900, and an F1-Score of 0.932.

Finally, it is important to note that both the DL models and RF, show moderate performance in classifying the *Forest* and *Water* classes. RF records producer accuracies (PA) of 0.51 and 0.18 for these classes in Amazonia, 0.64 for both in Africa, and 0.31 and 0.41 in Siberia, respectively. However, RF struggles significantly with other classes. In Amazonia, it performs poorly for *Shrubland* (PA = 0.09), *Cropland* (PA = 0.07), *Bareland* (PA = 0), and *Built-up* (PA = 0.02). Similar issues arise in Africa for *Shrubland* (PA = 0.11), *Wetland* (PA = 0.07), and *Built-up* (PA = 0), as well as in Siberia for *Shrubland* (PA = 0), *Wetland* (PA = 0.06), and *Bareland* (PA = 0). These low PAs suggest that RF struggles with complex and highly variable LC classes.

3.2.1.4.1 DL Swin-Unet Performance Assessment with Ecoregion Classification

Additional experiments were conducted for the best DL model using ground truth data organized by ecoregions. Figure 21 shows the MOLCA training patches for the Amazon, Africa, and Siberia, divided by climatological zones. The legend includes a color map and numerical codes for each region.

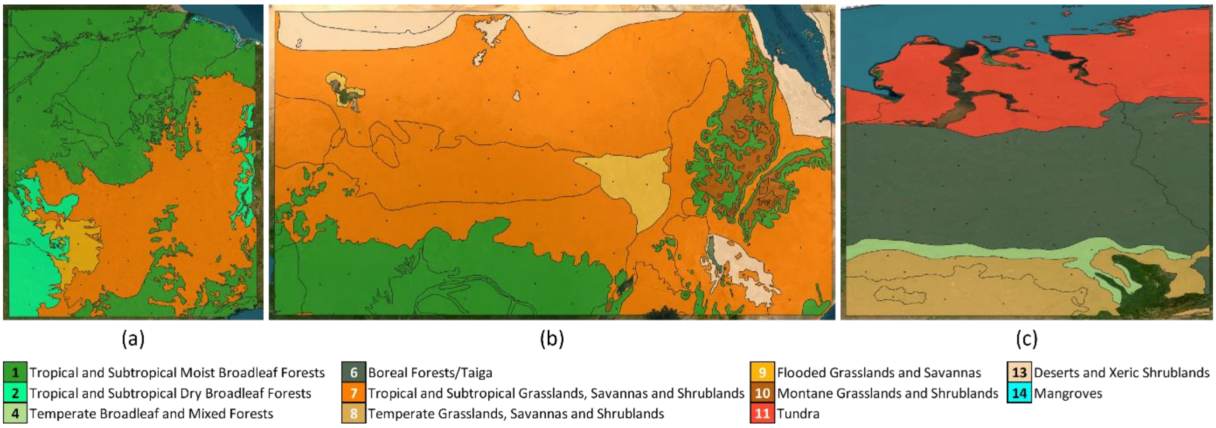

(a)
(b)
(c)
1 Tropical and Subtropical Moist Broadleaf Forests
2 Tropical and Subtropical Dry Broadleaf Forests
4 Temperate Broadleaf and Mixed Forests
6 Boreal Forests/Taiga
7 Tropical and Subtropical Grasslands, Savannas and Shrublands
8 Temperate Grasslands, Savannas and Shrublands
9 Flooded Grasslands and Savannas
10 Montane Grasslands and Shrublands
11 Tundra
13 Deserts and Xeric Shrublands
14 Mangroves

Figure 21. Division of the MOLCA patches in the training set based on ecoregions for (a) Amazonia, (b) Africa, and (c) Siberia.

The experiments were designed to incorporate the ecoregions within the three study areas. This additional analysis aims to evaluate the performance of the model when training and testing are done within the same ecoregion compared to when they are conducted across different ecoregions, with the goal of providing a deeper assessment. The results obtained using the Swin-Unet model are displayed in Figure 22 , and are presented in terms of OA. Diagonal elements indicate cases where training and testing are done within the same ecoregion, while off-diagonal elements represent cross-validation across different ecoregions.

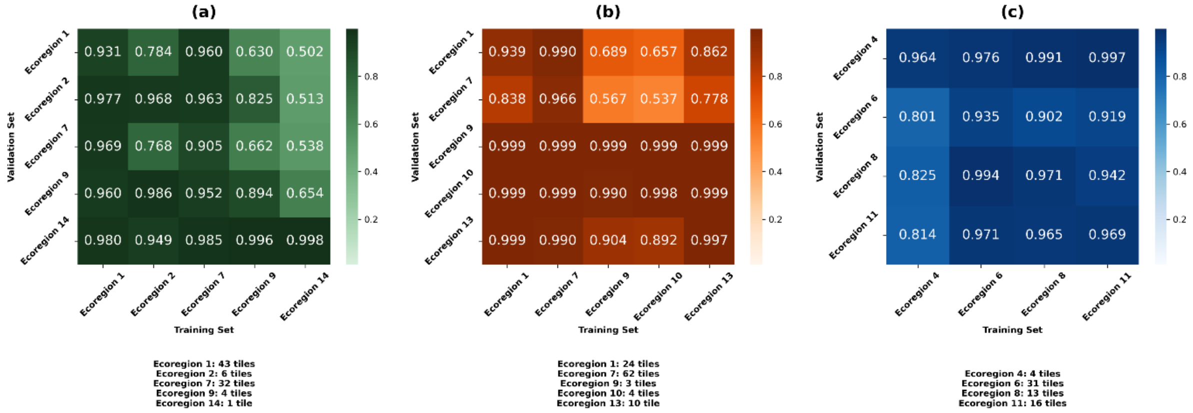


Figure 22. Overall Accuracy (OA) matrices derived from various training/test ecoregion pairings for (a) Amazonia, (b) Africa, and (c) Siberia.

In the Amazonia region in Figure 22(a), the OA. is generally higher when both training and testing occur within the same ecoregion. However, a notable exception arises when Ecoregion 14 and Ecoregion 9 are used for training and tested on other ecoregions. The poorer performance in these cases is likely due to the limited number of tiles in Ecoregion 14 (1 tile) and Ecoregion 9 (4 tiles), restricting the diversity and quantity of training data. A similar trend is observed in the African region (Figure 22(b)), where performance is lower when training on Ecoregion 9 and Ecoregion 10, which contain only 3 and 4 tiles, respectively. This small dataset may be insufficient for the model to generalize well across more diverse ecoregions. In the Siberian region (Figure 22(c)), OA. values remain high, surpassing 80% across all combinations. However, the lowest performance is noted when training on Ecoregion 4 and testing on other ecoregions, even though it contains just 4 tiles. Despite this, the model's accuracy remains relatively stable, demonstrating the strong generalization capability of the model, even when trained with limited data from other ecoregions.

3.2.1.4.2 Comparison with Conventional SAR Time Series Analysis

The following analysis presents a comparative evaluation classifications obtained in the static areas identified in Phase 1 for Amazonia, Africa, and Siberia either with the full time-series data or just the seasonal. The objective is to show that the approach based on seasonal SAR features surpasses the conventional method of using multiple temporal SAR images for each season.
For this experiment, twenty temporal images per tile were collected for 2021, i.e., five images per season. This

process generated datasets containing 1620 S1 images for Amazonia, 1860 for Africa, and 960 for Siberia. Strict selection criteria ensured that only tiles with complete spatial and temporal coverage were included. As a result, the final dataset was refined to 81 MOLCA tiles for Amazonia (reduced from 86), 93 for Africa (down from 103), and 48 for Siberia (from 64), preserving the data quality for all analyses. The seasonal distribution of the collected S1 time series is presented in Table 10.

Table 10. Seasonal count of S1 images from the 2021 time series for Amazonia, Africa and Siberia static area.

Region	Platform	Acquisition mode	Product format	Polarization	Resolution	Season*	Number of images
Amazonia	S1A/S1B	IW	GRDH	VH	10 m	Winter	405
	S1A/S1B	IW	GRDH	VH	10 m	Spring	405
	S1A/S1B	IW	GRDH	VH	10 m	Summer	405
	S1A/S1B	IW	GRDH	VH	10 m	Autumn	405
Africa	S1A/S1B	IW	GRDH	VH	10 m	Winter	465
	S1A/S1B	IW	GRDH	VH	10 m	Spring	465
	S1A/S1B	IW	GRDH	VH	10 m	Summer	465
	S1A/S1B	IW	GRDH	VH	10 m	Autumn	465
Siberia	S1B	IW	GRDH	VH	10 m	Winter	240
	S1B	IW	GRDH	VH	10 m	Spring	240
	S1B	IW	GRDH	VH	10 m	Summer	240
	S1B	IW	GRDH	VH	10 m	Autumn	240

***Winter**: 01.01 to 03.31, **Spring**: 04.01 to 06.30, **Summer**: 07.01 to 09.30, **Autumn**: 10.01 to 12.31

The comparative study employed the Swin-Unet DL model, which demonstrated superior performance among the three analyzed DL approaches, alongside the RF classifier, ensuring a fair comparison by using identical training and validation sets. The results, reported in Table 9 and Table 11, reveal a significant reduction in accuracy when standard time-series images are used compared to the seasonal features series. These findings highlight the efficiency and reliability of the synthesized spatial information derived from the features, establishing it as a superior approach for multitemporal data analysis.

Table 11. Results for standard input time series for Amazonia, Africa and Siberia.

Region	Model	O.A.	kappa	F1-Score	pa_1	pa_2	pa_3	pa_4	pa_5	pa_6	pa_7	pa_8
Amazonia	Random Forest	0.612	0.309	0.553	0.49	0.09	0.21	0.08	0.18	0	0.02	0.16
	Swin-Unet	0.903	0.855	0.901	0.92	0.75	0.68	0.67	0.14	0	0.75	0.94
Africa	Random Forest	0.743	0.535	0.707	0.61	0.09	0.13	0.17	0.07	0.70	0	0.44
	Swin-Unet	0.872	0.795	0.854	0.95	0.03	0.68	0.16	0	0.89	0.27	0.77
Siberia	Random Forest	0.670	0.315	0.619	0.25	0.00	0.34	0.19	0.08	0	0.19	0.30
	Swin-Unet	0.896	0.817	0.893	0.93	0	0.44	0.73	0.65	0.57	0.28	0.79

***pa**: *Producer Accuracy;* **1**: *Forest;* **2**: *Shrubland;* **3**: *Grassland;* **4**: *Cropland;* **5**: *Wetland;* **6**: *Bareland;* **7**: *Built-up;* **8**: *Water.*

Specifically, two significant drawbacks of the traditional time-series approach emerged:

- **Increased computational demand**: Handling numerous temporal images requires extensive storage and processing resources.
- **Data limitations**: Some regions lack sufficient temporal data consistency, hampering detailed studies.

The features series approach provides a significant benefit by condensing temporal dynamics into a single composite image, known as the super image. This image captures seasonal variations, offering a consistent representation while eliminating the need for extra spatial constraints. This method simplifies the workflow, enhancing coherence and reducing complexity.

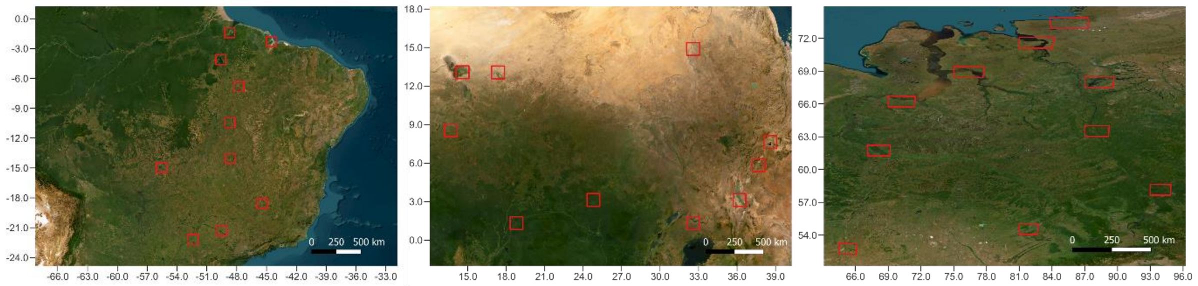
3.2.1.5 *Comparative Analysis for Water LC Class Recognition between the Swin-Unet DL Approach and SAR Water Detector*

To assess the performance of water recognition by the dedicated detector selected in Phase 1 in comparison with the DL Swin-UNet approach, 10 benchmark areas were selected within the three static regions of interest. Each of these ten test areas were chosen to represent water bodies that are most commonly found in these regions, including rivers, streams, basins, lakes, seas, and other aquatic features. The tiles selected as benchmarks for each area are listed in Table 12, and are shown in Figure 23, to offer a clear geographical context for each benchmark location.

Table 12. List of benchmark tiles selected for Amazonia, Africa, and Siberia to assess the performance of the Phase 1 water extractor against the DL Swin-Unet approach.

Amazonia	Africa	Siberia
21LXD	33PUK	41UPU
22KCA	33PVQ	42VVP
22KFB	33PYQ	42WWU
22LGK	34NBG	43WES
22LGP	35NKD	44UNF
22MFA	36NVG	44WNE
22MGD	36NZJ	45VWL
22MHT	36PVB	45WWR
23KMV	37NCG	45XVB
23MNT	37NDJ	46VEK

The selection ensures a diverse and comprehensive evaluation of the detection capabilities in varying geographical and environmental contexts. Each region presents unique challenges related to the scale and type of water bodies, enabling a detailed performance assessment.



(a) (b) (c)

Figure 23. S2 tiles selected for evaluating water detection performance in (a) the Amazon, (b) Africa, and (c) Siberia. These tiles highlight regions with diverse hydrological features, providing a robust basis for performance comparison**.**

For each tile in each area, a representative patch of 549x549 pixels was selected through visual inspection of the optical imagery to identify potential water basins of interest. This approach allowed for a targeted evaluation by focusing on specific, clearly defined regions within the tiles, enabling a detailed understanding of the detector's robustness in real-world conditions, particularly in regions with diverse hydrological features.

For water extraction using the dedicated detector, the annual series of S1 data from 2021 was considered for each tile. The series was subsequently divided into monthly temporal subsets, for which the following temporal features were calculated: mean, minimum, maximum, and variance. These monthly feature sets served as inputs to the water extractor, which generates binary monthly water maps (where pixel values are 0 for non-water and 1 for water).

The water extraction is an unsupervised routine introduced in [15], utilizing a K-means clustering approach. This procedure differentiates areas with low backscatter and low variance along the temporal series (characteristic of water bodies) from other potential regions of interest. The final water map is derived from these monthly water maps, following the methodology described in the deliverable ATDB [AD1], particularly in the section titled *"8.2.2.2 Water Dynamics Analysis: Seasonal vs. Permanent Water Identification."*

The monthly aggregation of statistical measures such as mean and variance helps capture the temporal dynamics of water bodies, which may fluctuate due to seasonal changes or hydrological events. The reliance on K-means clustering makes this method adaptable to diverse regions without requiring labeled data, enhancing its applicability for large-scale studies. Concerning the final water map, b synthesizing monthly maps, the process distinguishes between seasonal and permanent water, providing a comprehensive view of water dynamics over the year.

To refine the final water map, a masking operation was applied to address desert areas or regions with strong similarities to water due to their dominant backscattering mechanisms. Specifically, water areas identified by the detector that belong to desert or bare soil classes, according to coarser resolution land cover maps, such as the Copernicus Land Cover Map at 100 m spatial resolution, were excluded. This refinement step enhances the accuracy of the water map by reducing false positives, particularly in arid or semi-arid regions where surface characteristics can mimic water signatures in satellite imagery.

The classification of water bodies using the Swin-Unet approach is instead performed by utilizing the models trained in the previous experiments. Seasonal spatial feature sets were computed and used as input for the Swin-Unet DL framework. These features capture the temporal variability and spatial characteristics of water bodies, enabling the model to account for seasonal changes and better distinguish water from other land cover types.

To compare the water recognition performance of the two methodologies, the permanent and seasonal water classes from the water map generated by the dedicated extractor were merged into a single "water" class. This aggregation resulted in a binary map where pixel values of 0 represent "non-water" areas, and pixel values of 1 represent "water" areas. This simplification was necessary because the MOLCA dataset, used for training the DL approach, does not include seasonal information in its LC classes. Concerning the DL maps, regions encompassing various land surface types, such as forests, urban areas, agricultural fields, deserts, and other terrains that do not exhibit hydrological characteristics, such as low backscatter or high reflectivity, were grouped into the non-water land cover category. These areas are uniformly assigned a value of 0, distinctly separating them from water classes. This aggregation simplifies the classification process by consolidating diverse non-water types into a singular, cohesive category. Such an approach streamlines data analysis, enhancing the ability to compare and evaluate water-detection results against these non-water regions.

The classification results and the visual comparison are presented in Figure 24, Figure 25 and Figure 26, corresponding to the Amazonia, Africa, and Siberia, respectively. For each region, four representative S2 tiles were selected as examples. The water maps derived from the dedicated extractor and the DL Swin-Unet approach were evaluated by comparing them against two reference datasets: the ESRI optical basemap, which provides high-resolution visual reference imagery, and the super image calculated from the seasonal S1 temporal series, which aggregates backscatter data to highlight hydrological patterns over time.

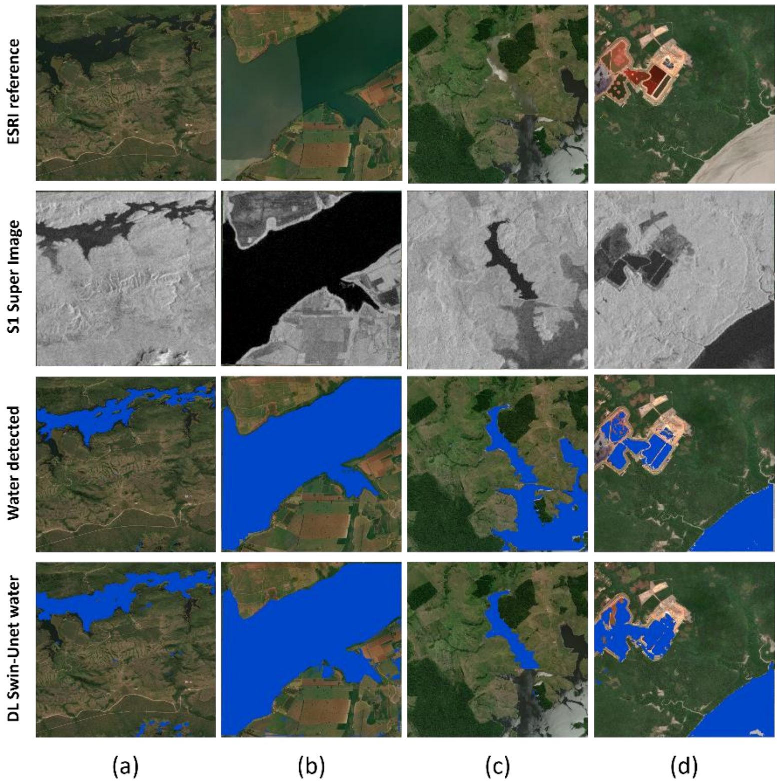


Figure 24. Visual assessment for the Amazonia region comparing (first row) the ESRI reference image, (second row) the S1 seasonal super image, (third row) the water map derived from the dedicated extractor, and (fourth row) the water land cover classes from the DL Swin-Unet classification. The comparison is presented for patches of 549x549 pixel size, corresponding to the following S2 tiles: (a) 21LXD, (b) 22KCA, (c) 22MFA, and (d) 23MNT.

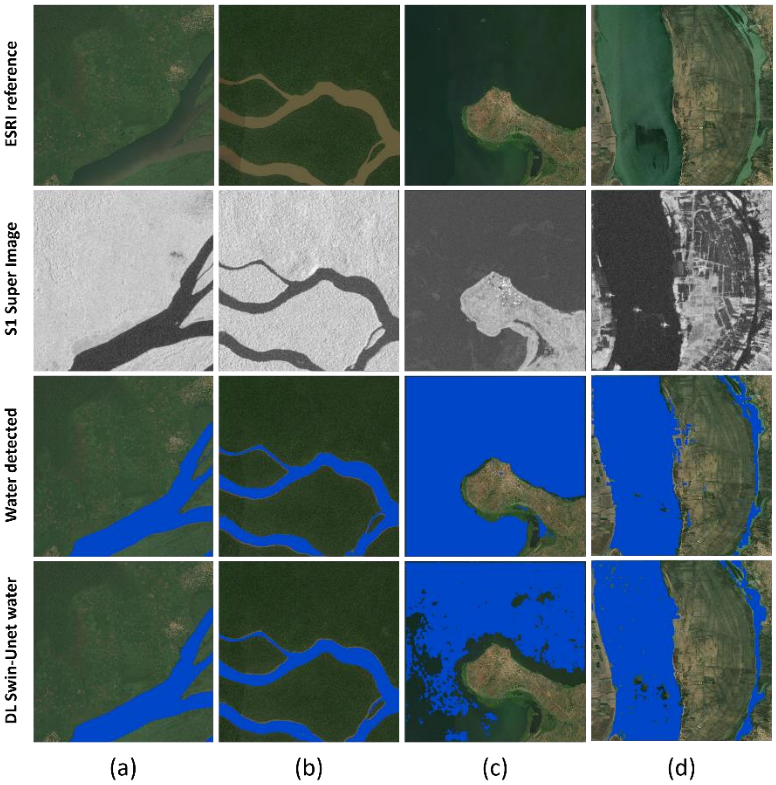


Figure 25. Visual assessment for the Africa region comparing (first row) the ESRI reference image, (second row) the S1 seasonal super image, (third row) the water map derived from the dedicated extractor, and (fourth row) the water land cover classes from the DL Swin-Unet classification. The comparison is presented for patches of 549x549 pixel size, corresponding to the following S2 tiles: (a) 34NBG, (b) 35NKD, (c) 36NVG, and (d) 36PVB.

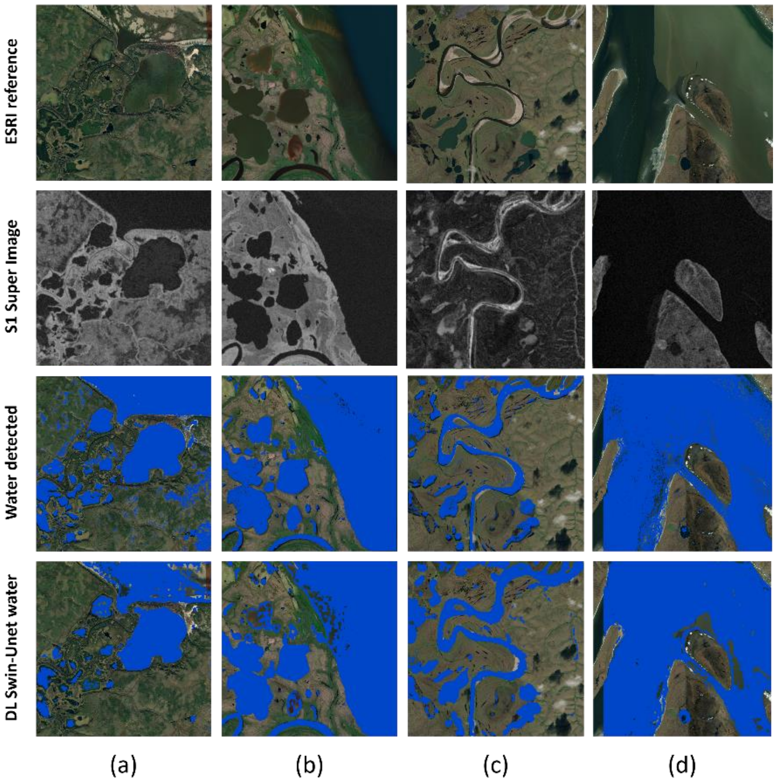


Figure 26. Visual assessment for the Siberia region comparing (first row) the ESRI reference image, (second row) the S1 seasonal super image, (third row) the water map derived from the dedicated extractor, and (fourth row) the water land cover classes from the DL Swin-Unet classification. The comparison is presented for patches of 549x549 pixel size, corresponding to the following S2 tiles: (a) 42WWU, (b) 43WES, (c) 44WNE, and (d) 45XVB.

This multi-faceted comparison highlights the differences in water detection performance, accounting for the strengths and limitations of the two methodologies. In the Amazonia region, for example, examining Figure 24 and column (a), it is evident that the DL approach can identify water bodies not detected by the dedicated extractor. Conversely, column (c) demonstrates a limitation of the Swin-Unet network: half of a lake within the examined patch is not classified as water, whereas the dedicated extractor's output map successfully identifies the entire lake. Similar observations can be made for the Africa and Siberia regions, as shown in the respective figures (Figure 25 and Figure 26). In these cases, the DL approach and the dedicated extractor exhibit complementary strengths, with the former excelling in identifying smaller or less prominent water features, and the latter showing greater reliability in capturing larger, more stable water bodies.

To complement the evaluation, a thorough quantitative analysis is introduced, providing a numerical framework to validate and expand upon the insights gained from the earlier qualitative assessment. Several metrics are considered, including OA Precision, Recall, F1-Score, and Producer Accuracy (pa) for both non-water (pa_0) and water classes (pa_1). These metrics provide complementary insights into different aspects of model performance:

- **OA** is used to assess the proportion of correctly classified pixels across the entire dataset, giving a general view of model effectiveness.
- **Precision** measures the accuracy of water detection, indicating the proportion of true water pixels among all those classified as water. This helps to understand how well false positives are minimized.
- **Recall** evaluates the ability of the model to identify all water pixels, reflecting how many true water features are correctly detected.
- **F1-Score** offers a balanced measure of Precision and Recall, particularly useful when there is an imbalance between water and non-water classes in the dataset.

These metrics were computed for each test area (Amazonia, Africa, Siberia) to allow both regional and aggregated comparisons between the water detection methodologies. The analysis, reported in Table 13, highlights not only the overall reliability of each method but also their strengths and weaknesses in capturing diverse hydrological patterns under varying environmental conditions. The ESA WorldCover map, with its 10-meter spatial resolution, was utilized as the validation dataset to assess the performance of the water detection approaches. This dataset provides a globally consistent, high-resolution land cover classification, enabling a reliable comparison between detected water bodies and the ground truth information.

Table 13. Overall Accuracy (OA), Precision, Recall, F1-Score, and Producer Accuracy (pa) for the evaluated approaches across different regions.

Region	Model	OA	Precision	Recall	F1-Score	pa_0	pa_1
Amazonia	Water extractor	0.96	0.97	0.96	0.96	98.06	90.40
	Swin-Unet DL	0.97	0.97	0.97	0.97	96.56	91.81
Africa	Water extractor	0.94	0.95	0.94	0.94	94.62	89.35
	Swin-Unet DL	0.80	0.90	0.80	0.80	99.47	63.65
Siberia	Water extractor	0.85	0.91	0.85	0.87	87.77	66.86
	Swin-Unet DL	0.92	0.95	0.92	0.93	93.31	90.51

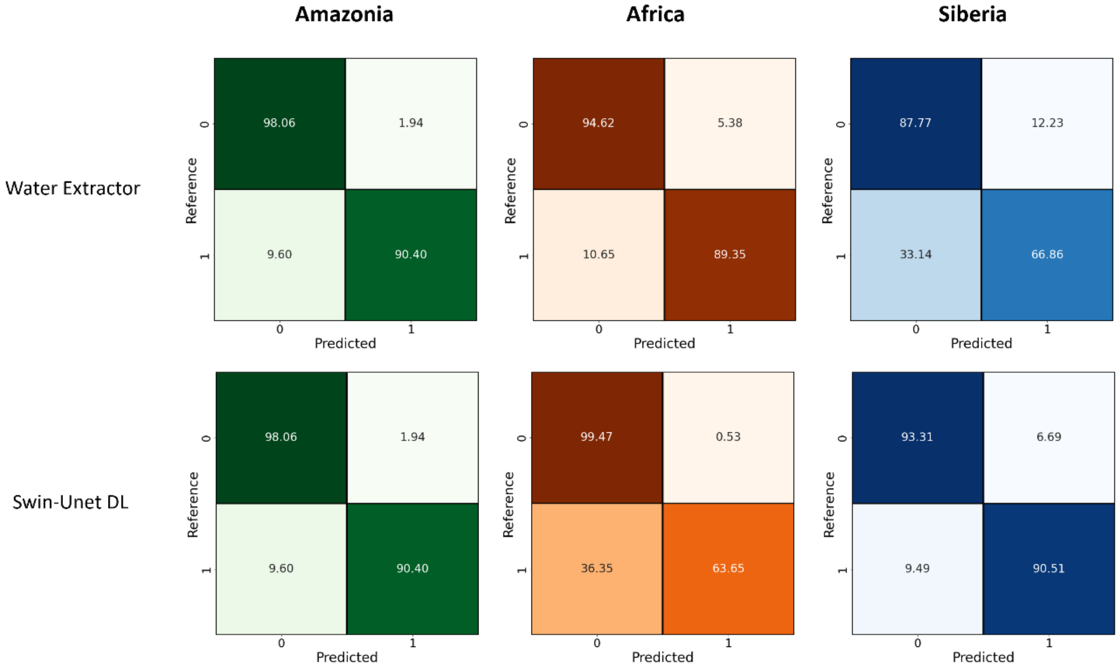
***pa**: *Producer Accuracy;* **0**: *Non-Water;* **1**: *Water.*

In Table 13, the performance in Amazonia shows that both the water extractor and the Swin-Unet DL model exhibit high accuracy. The OA values are 0.96% for the water extractor and 0.97% for Swin-Unet DL, indicating strong performance for both methods. The Precision, Recall, and F1-Score metrics are also very close, ranging from 0.96% to 0.97% for both models. For Producer Accuracy (pa), the performance for non-water areas (pa_0) is higher with the water extractor at 98.06%, compared to 96.56% for Swin-Unet DL. However, for the water class (pa_1), Swin-Unet DL outperforms the water extractor, achieving 91.81%, compared to 90.40% for the latter. This suggests that while both models perform similarly in most aspects, the water extractor has a slight edge in correctly identifying non-water areas, while Swin-Unet DL slightly excels in detecting water bodies in Amazonia.

In Africa, as seen in Table 8, the water extractor outperforms Swin-Unet DL in several key metrics. The OA is significantly higher for the water extractor (0.94%) compared to Swin-Unet DL (0.80%). Similarly, the F1-Score of the water extractor (0.94) is considerably better than Swin-Unet DL (0.80%). This suggests that the water extractor is more effective in overall classification performance in Africa. Looking at the Producer Accuracy (pa) values, Swin-Unet DL performs better in detecting non-water areas (pa_0), with a value of 99.47%, compared to 94.62% for the water extractor. However, for water areas (pa_1), the water extractor significantly outperforms Swin-Unet DL, with a pa_1 of 89.35% compared to 63.65% for Swin-Unet DL. This disparity indicates that Swin-Unet DL struggles more with identifying water bodies in the complex and diverse African landscape, while the water extractor is more proficient in detecting water features in this region.

For the Siberia region, Swin-Unet DL shows better overall performance compared to the water extractor. The OA for Swin-Unet DL is 0.92%, surpassing the water extractor's OA of 0.85%. Similarly, the F1-Score for Swin-Unet DL (0.93%) outperforms the water extractor (0.87%). These metrics highlight Swin-Unet DL's stronger overall classification performance in Siberia. In terms of water class recognition, Swin-Unet DL also significantly outperforms the water extractor in terms of Producer Accuracy for water (pa_1), with a value of 90.51% compared to 66.86%. This indicates that Swin-Unet DL is more adept at recognizing water bodies in the Siberian landscape. However, for non-water class recognition (pa_0), the water extractor has a slight edge with a value of 87.77%, while Swin-Unet DL achieves 93.31%. This suggests that while Swin-Unet DL excels in identifying water bodies, the water extractor performs slightly better at distinguishing non-water areas.

The corresponding confusion matrices for the models' performance in the Amazonia, Africa, and Siberia regions are presented in the Figure 27.



+

Figure 27. Confusion matrices for the (first row) water extractor and (second row) Swin-Unet DL approach across regions of (first column) Amazonia, (second column) Africa, and (third column) Siberia.

In conclusion, the water extractor typically offers a more consistent and balanced performance across all regions. It performs particularly well in Africa, where it demonstrates strong water detection capabilities, even within complex landscapes. In contrast, Swin-Unet DL shows more variability, excelling in Amazonia and Siberia, but struggling in Africa due to the region's complex land cover and seasonal dynamics. This indicates that while DL models such as Swin-Unet DL perform well in simpler or less dynamic environments, they may face challenges in regions with significant seasonal changes and diverse land cover types.

Furthermore, it is important to emphasize that the water extractor-based approach allows for the identification of seasonal variations in water bodies, unlike the DL network, which requires a reference dataset where the water land cover class is already separated based on seasonality. This makes the water extractor more feasible for use in the final SAR production chain, as it is capable of handling seasonal variations and land cover complexities more effectively than models like Swin-Unet DL. Given that the water extractor can operate independently of pre-labeled seasonal datasets, it offers a more flexible and robust solution, particularly in environments where seasonal dynamics play a crucial role in water detection. This adaptability makes it a practical choice for large-scale operational applications, where data consistency and continuous performance are key.

3.2.1.6 Comparative Analysis for Built-up LC Class Recognition between the Swin-Unet DL Approach and UEXT Detector

A similar analysis was conducted to compare the performance of built-up land class recognition using the Urban EXTent (UEXT) module and DL classification based on the Swin-Unet network. For this analysis, a test area shown in Figure 28, was identified in Siberia, specifically in the S2 45UVB tile, where open-pit mines were incorrectly classified as built-up areas in the final product of Phase 1 of the CCI+ project. The aim is to assess whether the DL approach can resolve this issue and improve classification accuracy in such cases. Given the lack of an accurate urban reference dataset for this region, particularly one that is sufficiently distributed to cover the unique landscape features and urban patterns, only a visual comparison is presented in this section. This visual comparison helps identify the potential advantages of DL in distinguishing between built-up areas and non-urban land uses, such as open-pit mining, which may be misclassified due to their structural similarities to built-up regions. The analysis also highlights how well Swin-Unet can adapt to these complex, mixed landscapes compared to traditional methods like the UEXT module.

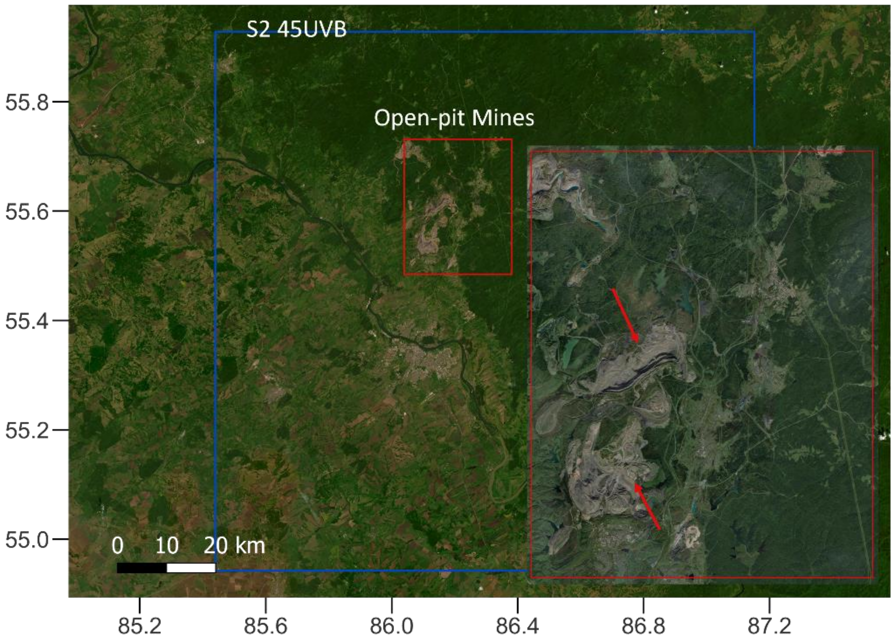


Figure 28- Open-pit mining site (in red) located in Siberia, within the S2 tile 45UVB (in blue), used for comparing built-up land class recognition performance between the UEXT module and the Swin-Unet DL classification approach.

The urban map was generated using the dedicated UEXT module, as described in the Phase 1 ATDB [AD1], leveraging the S1 time series for the year 2021. Similarly, for the same year, the LC map was derived using the DL Swin-Unet approach, from which the built-up land class of interest was extracted. Regarding the DL approach, the classification was executed by segmenting the input data into patches with dimensions of 549x549 pixels. To address potential inaccuracies at the patch boundaries, an overlay of 10% was applied between adjacent patches. This overlap strategy reduces edge effects, ensuring smoother transitions and improving classification coherence.

The generated patch-wise outputs were subsequently merged using a mosaicking process to create a continuous classification map covering the entire area of interest. This integration step not only consolidates the individual outputs but also addresses inconsistencies that may arise due to variations in local feature distributions. By combining overlapping regions, the approach enhances the overall spatial accuracy and ensures that all regions, including boundaries, are classified with minimal discrepancies.

Additionally, the methodology leverages the advantages of a patch-based workflow, such as scalability for large datasets and computational efficiency. This structured approach is particularly beneficial for extensive or heterogeneous regions, where variations in land cover types or topographic features might otherwise challenge the classification process. The resulting final map provides a detailed, high-resolution representation of the area, suitable for downstream analysis and validation.

The figure presents a performance comparison of these two results: one obtained via the UEXT module and the other using the DL approach. To ensure a comprehensive evaluation, the optical ESRI reference map and the S1 super image are also included in the analysis. The inclusion of these references provides additional context for assessing the consistency and accuracy of the built-up class identification.

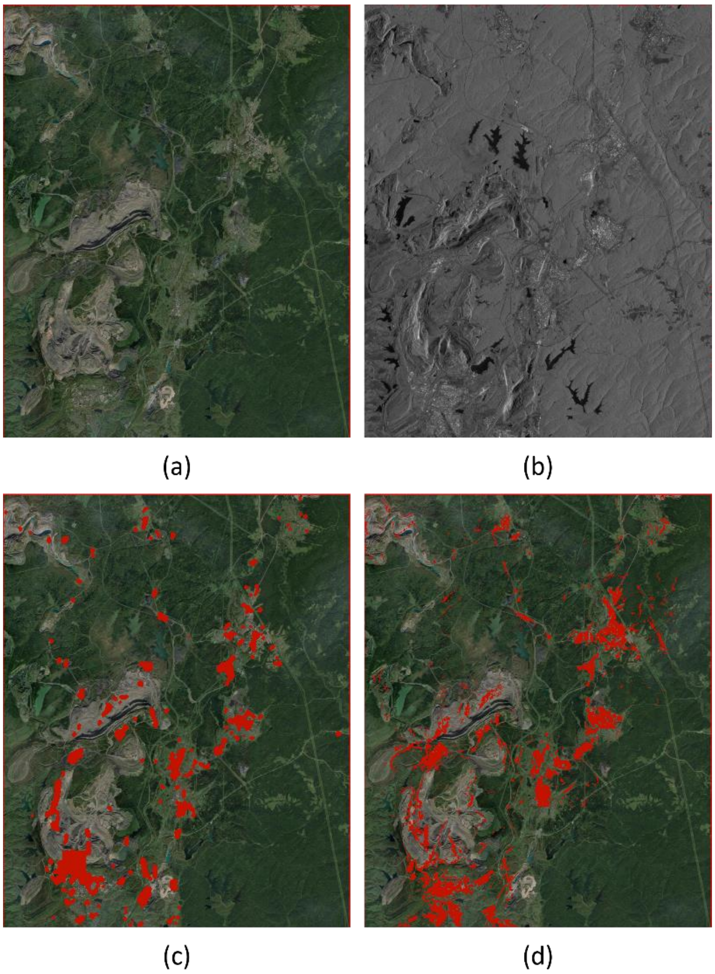


Figure 29. Visual assessment of an area of interest within the S2 tile 45UVB. The analysis was conducted to evaluate the performance of the UEXT detector and the Swin-Unet DL classification chain. The comparison includes: (a) the ESRI optical reference, (b) the 2021 S1 super image, (c) UEXT output, and (d) DL Swin-Unet classification network output.

Through visual inspection, the recognition of the built-up class by both methods appears to show comparable performance. Specifically, both the UEXT module output map and the Swin-Unet DL network's classification reveal misclassification of open-pit mines as "built-up" areas. This highlights the challenges presented by the spectral and spatial characteristics of open-pit mines, which closely resemble built-up areas in the feature space utilized by both methodologies.

This shows that there is no “free lunch” using DL approaches. To enhance the urban extraction performances for these models, a more accurate training dataset is required. This is what will be explored in the next step of the project, aiming at a reduction of open-mines misclassifications as urban areas, a limitation observed in both DL models and traditional methods like UEXT.

3.2.2 Final decision

The SAR classification chain identified for Phase 2 of the CCI+ project will involve the Swin-Unet DL model and the specialized water and urban area extractors. According to the previously described analysis, it combines innovative algorithms and robust validation processes to address the unique complexities of SAR data. The final framework represents a careful balance between DL model and specialized detectors, ensuring optimal results for diverse land cover applications.

The Swin-Unet DL network has been selected as the primary model for the SAR classification chain. Extensive testing demonstrated its superior performance in classifying complex and heterogeneous landscapes across regions such as Amazonia, Africa, and Siberia. The model's global attention mechanism enables precise feature extraction and contextual understanding, making it the most reliable and accurate solution for SAR-based land cover classification.

For the water class, the SAR classification chain will continue to use the dedicated water detector. This detector excels in capturing the seasonal dynamics of water bodies, a critical requirement for accurate hydrological mapping. Its ability to differentiate between permanent and seasonal water ensures it remains the most effective tool for this purpose, particularly in regions with significant hydrological variability.

Regarding the built-up class, the choice of methodology will depend on future developments. While the Swin-Unet network has shown potential in detecting urban areas, challenges such as the misclassification of open-pit mines highlight the need for an enhanced training dataset. If improvements to the training set yield the desired results, the Swin-Unet could become the preferred choice for this class, given its advanced feature extraction capabilities. For the moment, however, the dedicated UEXT extractor remains a viable option, and there are plans to enhance its performance using spatial indicators to better address current limitations. This dual-path strategy ensures flexibility, allowing the project to adapt and adopt the most effective solution based on further testing and refinement.

In conclusion, the SAR classification chain for the HRLC ECV project aims at improving the performance of the Phase 1 processing chain. The adoption of Swin-Unet for general classification, the dedicated water detector for hydrological features, and a flexible approach for built-up areas demonstrates a commitment to achieving accuracy, reliability, and adaptability.

3.3 Decision Fusion

3.3.1 Decision fusion methods

3.3.1.1 Multi-temporal fusion method

In the context of historical land cover mapping where the availability of the data is different for each year, the classification process which is done independently in each historical year inevitably causes temporal inconsistencies in the classification product. To address this problem, a multi-temporal model called cascade model [16] was introduced in Phase 1. The idea behind that model was to propagate the information from the 2019 static map, whose generation benefits from more data availability, to other historical maps backward. The cascade model reduced incorrect land cover transitions significantly, but the use of the temporal information was limited because its processing considered only pairs of historical years i.e., 2019 coupled with each of other years. Taking into account this limitation, we adopt an approach based on the theory of Hidden Markov Models (HMMs) in Phase 2, to consider the information from posteriors coming from all of the years in which land cover map is produced. We refer to the ATBD for the general methodological formulation, and here, we only specify algorithmic details. Let T be the total number of observation times (years), ℓ_t be the class label of a generic pixel in time t ($t = 1,2,\dots,T$), $\boldsymbol{x}_t$ be the corresponding feature vector (resulting from the optical and SAR observations), and $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_T)$ be the vector collecting the observations of that pixel on *all* dates across the whole time series. The inference of the label at each time step $P(\ell_t|\boldsymbol{x})$ is accomplished using the forward-backward algorithm. Indeed, it is possible to prove that [17]:

$$P(\ell_t|\boldsymbol{x}) = \frac{\alpha(\ell_t)\beta(\ell_t)}{\sum_{\ell_t} \alpha(\ell_t)\beta(\ell_t)}.$$

Here, $\alpha(\ell_t)$ defines a forward step i.e., the joint probability of observing all feature vectors up to time t and the label at time t, evaluated through a sequential recursive procedure along the forward time direction:

$$\alpha(\ell_t) \equiv P(\boldsymbol{x}_1, \dots, \boldsymbol{x}_t, \ell_t)$$
$$\alpha(\ell_t) = \frac{P_F(\ell_t|\boldsymbol{x}_t)}{P(\ell_t)} \sum_{\ell_{t-1}} \alpha(\ell_{t-1})\, P(\ell_t|\ell_{t-1}),$$

where $P_F(\ell_t|\boldsymbol{x}_t)$ (single-time pixelwise posterior) derives from the logarithmic opinion pool (LOGP) for the decision fusion of the posteriors received from the optical and SAR processing chains. $P(\ell_t)$ is the prior probability of the label, and $P(\ell_t|\ell_{t-1})$ is the transition probability stating the probability of one class in t changing to another class in $t-1$.

In dual fashion, a backward recursive sequential procedure calculates the conditional probability $\beta(\ell_t)$ of all observations from time $t+1$ up to T given the value of ℓ_t:

$$\beta(\ell_t) \equiv P(\boldsymbol{x}_{t+1}, \dots, \boldsymbol{x}_T|\ell_t),$$

$$\beta(\ell_t) = \sum_{\ell_{t+1}} \beta(\ell_{t+1}) \frac{P_F(\ell_{t+1}|\boldsymbol{x}_{t+1})}{P(\ell_{t+1})} P(\ell_{t+1}|\ell_t).$$

The calculation of $P(\ell_t|\boldsymbol{x})$ is done sequentially starting from 2019 backward to 1990. It is implied by the equations above that the algorithm formulated in Phase 2 enables the information propagation from the whole available years, because in each year, the forward procedure will include the information from the future years, and the backward process considers the information from the past years. Furthermore, the sequential inference approach allows the algorithm to run efficiently from the computational point of view.

3.3.1.2 Spatial fusion method

In Phase 1, a Markov Random Field (MRF) model [18] was introduced to consider the spatial context during the fusion process. This enabled the label regularization to be imposed on the fusion maps. As described in the latest Phase 2 ATBD [AD2], the energy function of an MRF model can be written as:

$$U(\mathbf{L}|\mathbf{X}) = -\sum_{s\in S} \alpha \log P_F(\ell_s|\boldsymbol{x}_s) + \sum_{\substack{s\in S \\ r\in\partial s}} V(\ell_s, \ell_r),$$

where $\mathbf{L}$ and $\mathbf{X}$ refer to the output label map to be generated and the input image data, respectively, S is the pixel lattice and s is a shorthand notation for a generic pixel location (i, k). $\boldsymbol{x}_s$ collects the input optical and SAR data, and ℓ_s indicates the sample on pixel s of the random field $\{\ell_s\}_{s\in S}$ of class labels. $\partial s \subset S$ is the set of neighbouring pixels of pixel s, which can be typically in the form of four connected adjacent pixels (first order connectivity) or the surrounding eight pixels (second order connectivity). The energy function consists of: the estimated fused posterior probability $P_F(\ell_s|\boldsymbol{x}_s)$, and a positive weight α in its first term, which is associated with the likelihood of the class at the pixel level; as well as the second term $V(\ell_s, \ell_r)$ that is related to the spatial regularization process. While in Phase 1, we guided the degree of regularization through only one parameter γ, i.e., the weight of the pairwise potential in the energy function of the Potts MRF model:

$$V(\ell_s, \ell_r) = \gamma[1 - \delta(\ell_s, \ell_r)],$$

where $\delta(\cdot)$ is the Kronecker delta function, in Phase 2, we increase our control over the regularization by framing the approach within a more general family of probabilistic graphical models called Conditional Random Fields (CRFs). The extension of this spatial model includes two main points: Firstly, we change the one-for-all-class weight parameter γ for the label regularization to class-based weights through a function $\gamma(\ell_s, \ell_r)$:

$$V_{sr}(\ell_s, \ell_r|\mathbf{X}) = \gamma(\ell_s, \ell_r)\ [1 - \delta(\ell_s, \ell_r)].$$

This is to give us more freedom in compromising between regularization and detail preservation, considering that some classes may need stronger regularization while other classes should be preserved as they are. Secondly, we introduce a kernel function that measures the similarity associated with the feature vectors of the neighbouring pixels $\mathcal{K}\big(\boldsymbol{P}_s(\mathbf{X}), \boldsymbol{P}_r(\mathbf{X})\big)$, which makes the second term of the CRF energy function be written as:

$$V_{sr}(\ell_s, \ell_r|\mathbf{X}) = \gamma(\ell_s, \ell_r)[1 - \delta(\ell_s, \ell_r)]\mathcal{K}\big(\boldsymbol{P}_s(\mathbf{X}), \boldsymbol{P}_r(\mathbf{X})\big),$$

where $\boldsymbol{P}_s(\mathbf{X})$ is the vector collecting the fused posterior probabilities $P_F(\ell_s = \omega_k|\boldsymbol{x}_s)$, of all classes $\omega_k, k = 1,2,..,\mathcal{C}$, where $\mathcal{C}$ is the number of classes. In the current formulation, we chose a radial basis function as the similarity kernel:

$$\mathcal{K}\big(\boldsymbol{P}_s(\mathbf{X}), \boldsymbol{P}_r(\mathbf{X})\big) = e^{-\varphi\|\boldsymbol{P}_s(\mathbf{X}) - \boldsymbol{P}_r(\mathbf{X})\|^2},$$

where φ is a positive parameter. This enables us to put restraints on the regularization process when the two neighbouring pixels are very different in terms of their posterior probabilities, while simultaneously pushing for more regularization over homogeneous image regions.

The optimization method of the CRF energy function is based on the Iterated Conditional Mode (ICM) algorithm because of the good tradeoff it usually favors between computational burden and accuracy [19] . ICM is applied iteratively and in each iteration, the label of each pixel is updated according to:

$$\ell_s \leftarrow arg \min_{\omega_k} U(\ell_s = \omega_k|\boldsymbol{x}_s).$$

The initialization makes use of the non-contextual map resulting from the pixelwise fused posteriors, i.e., $\operatorname{argmax}_{\ell_s} P_F(\ell_s|\boldsymbol{x}_s)$. From the implementation point of view, we aim to keep the efficient execution of the ICM algorithm during Phase 1 by keeping the strategy of formulating the ICM processing through convolution-like operations, which are designed to be feasible even when applied to the large-scale data of the CCI+ HRLC project. In this respect, the choice of ICM is confirmed as an appropriate tradeoff, as compared to more computationally intensive graph-theoretic energy minimization approaches.

3.3.2 Qualitative evaluation

3.3.2.1 Experimental Results – Multitemporal Fusion

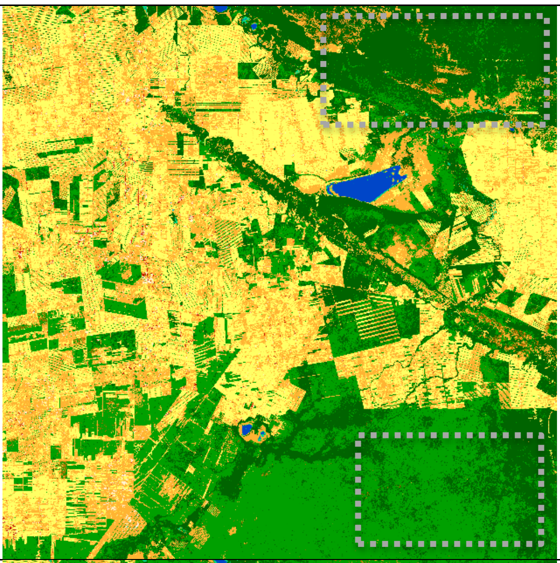
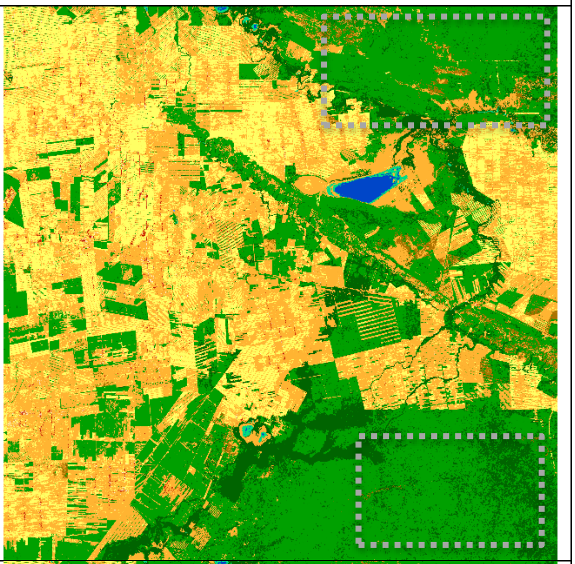
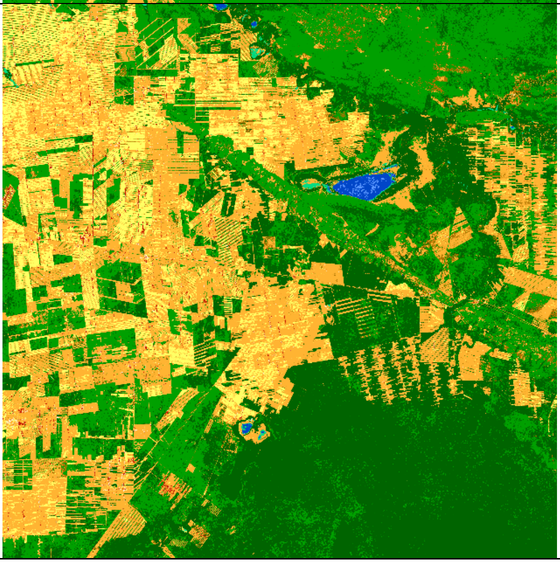
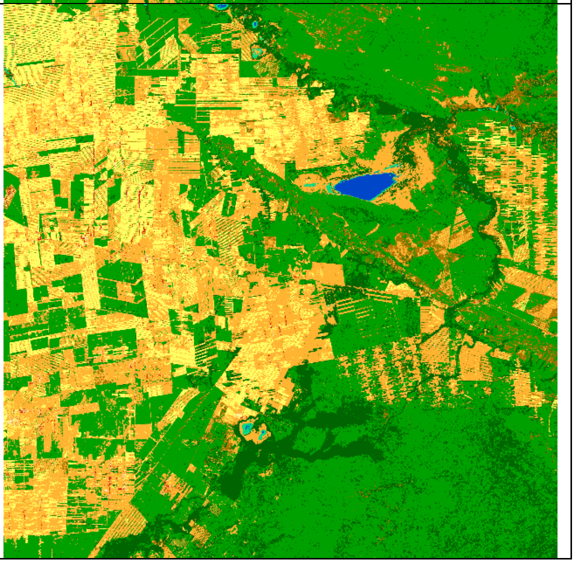
The experiments on the multi-temporal model using HMM were done on representative tiles of each region, selected either from the benchmark tiles of the first production or because they were affected by temporal artifacts during the Phase 1. Here, we will present experimental results on tiles 20KPF, 42WWD, and 37PHL. Tile 20KPF (Amazon) is one of the tiles that showed temporal inconsistency that could not be solved even after applying the cascade model in Phase 1. The same holds true for tile 42WWD of Siberia, while tile 37PHL in Africa was drawn from the list of benchmark tiles for Phase 2. Moreover, for a representative analysis of the model performance, we also made sure that all classes defined in the CCI+ HRLC project exist in those tiles, except for the snow/ice which almost never shows inconsistency naturally. In the following, we show the results of the comparison between the cascade model from Phase 1 and the HMM-based algorithm of Phase 2. For a fair comparison, we applied the same spatial fusion process by MRF used in Phase 1 after both procedures. The production will make use of both approaches simultaneously, and the next subsection will focus on CRF-based results for spatial fusion, but here, we deemed more appropriate not to apply them jointly to disentangle their possible impacts.

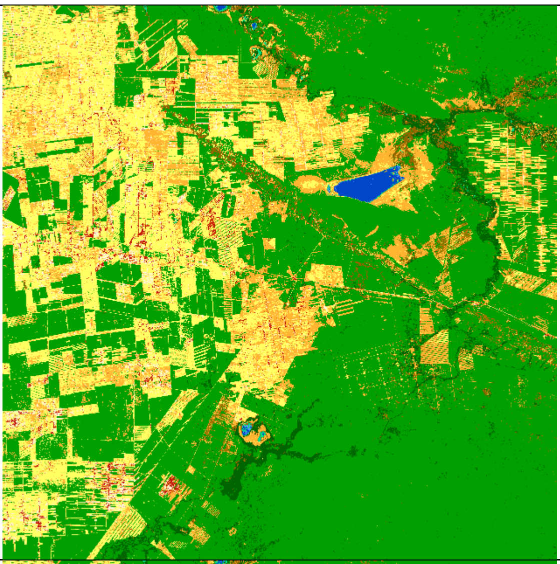
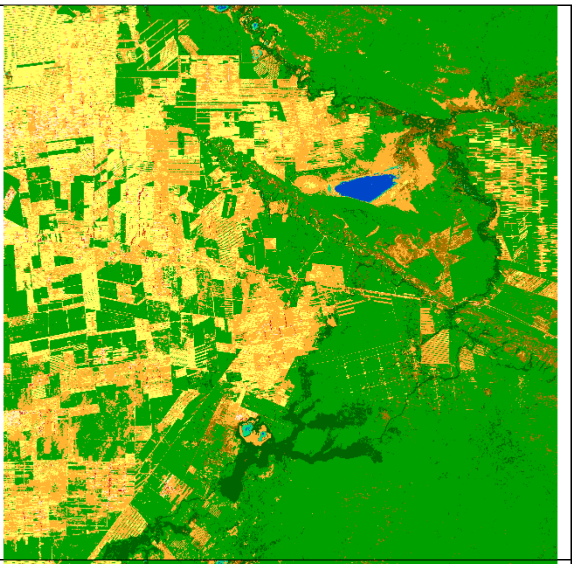
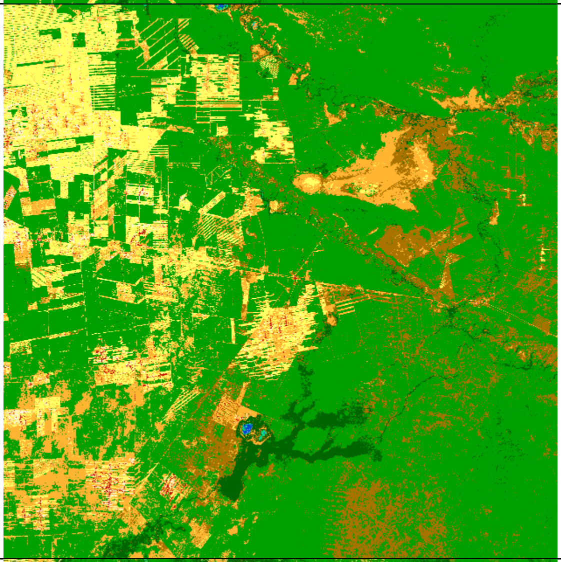
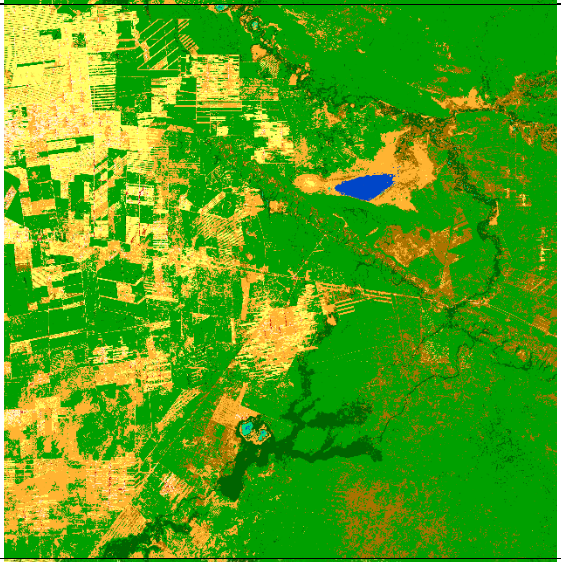
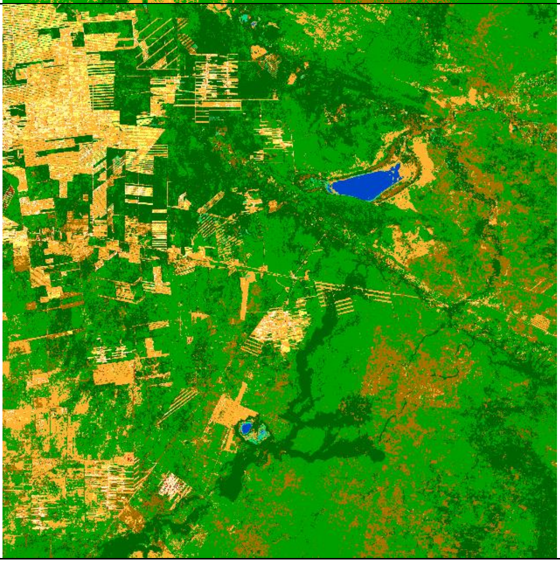
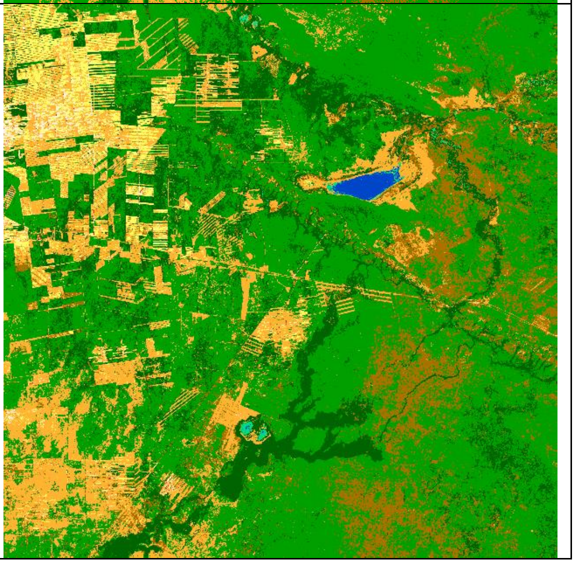
Figure 30 shows the comparison between the results of the cascade model and of HMM in the Amazonia area. The results using HMM provide more temporally consistent maps, especially in the areas marked by the grey rectangles placed on the 2019 maps. When the same areas are traced along the year, the cascade map produced inconsistencies between the evergreen broadleaf tree class (dark green) and the deciduous broadleaf tree (light green). For example, the bottom right rectangle on the cascade maps of 2019-2015-2010 shows an alternating pattern of deciduous-evergreen-deciduous, while the HMM maps indicate the pattern of deciduous with evergreen slowly disappearing as it goes backward. This artefact involving the evergreen class was indeed a major inconsistency problem pointed out during Phase 1. Furthermore, the information propagation by HMM is also beneficial in the case where the open water class (dark blue) disappears in the 2005 cascade map, while it is correctly revived in the 2005 HMM map because the class exists in every year except for that year. The favourable temporal behaviour can also be observed in the urban area (red) on the left side of the maps, where the HMM maps indicate more temporal consistency compared to the cascade maps, which for example exhibit an abrupt abundance in the 2010 instance.

This consistent temporal behaviour is also confirmed in the Siberia area shown in Figure 31. For example, focusing on the grey rectangle area on the 2019 map, HMM exhibits more stability regarding the shrub class (brown) along the time, while the class appears less in the 2000 cascade map although it exists relatively abundantly in 2005 and 1995. In Africa's tile, the urban class (red), which is highly inconsistent and sometimes excessively abundant in the cascade maps shown in Figure 32, displays a more natural and progressive transformation along the time in the HMM results. The deciduous broadleaf tree class (green) in the region indicated by the grey rectangle on the 2019 maps of the same tile, which diminishes on the 2005 cascade maps, seems also to appear in a favourable regular pattern on the HMM maps.

The conducted experiments demonstrate that the multi-temporal model based on the developed HMM formulation is advantageous in favouring the consistency across the land cover maps over time, and in reducing unwanted temporal patterns that were observable from the maps of three or more consecutive years within the Phase 1. Those temporal pattern could not, even theoretically, be addressed by the cascade model because the information propagation only happens in this model between one pair of years. On the contrary, the HMM formulation proposed here addresses multitemporal fusion at the level of the whole time series, which allows taking into account the temporal information of the entire sequence in the labelling of each pixel.

On one hand, this is beneficial in decreasing the inconsistent pattern which is one of the main problems encountered during Phase 1. On the other hand, it is worth noticing that generally there is a risk of propagating information from maps that are less reliable (e.g., those with problems of lacking acquisition data), and of correspondingly introducing noise in the classification results of the years which we have more confidence at. In the experiments conducted so far, this risk has not been observed. Nevertheless, in this regard, we are taking into account possible ways to further improve the model to minimize this risk. Moreover, in the next project period, we will also put our focus on tuning the transition probability matrix (TPM) which is one of the hyperparameters of the HMM that translates to the trade-off between the information propagation (which is associated to consistency) and the changes of classes in time. The goal is that the maps exhibit consistency over time but also do not end up censoring changes that actually happened.

Year	Cascade + MRF	HMM + MRF
2019		
2015		

2010		
2005		
2000		

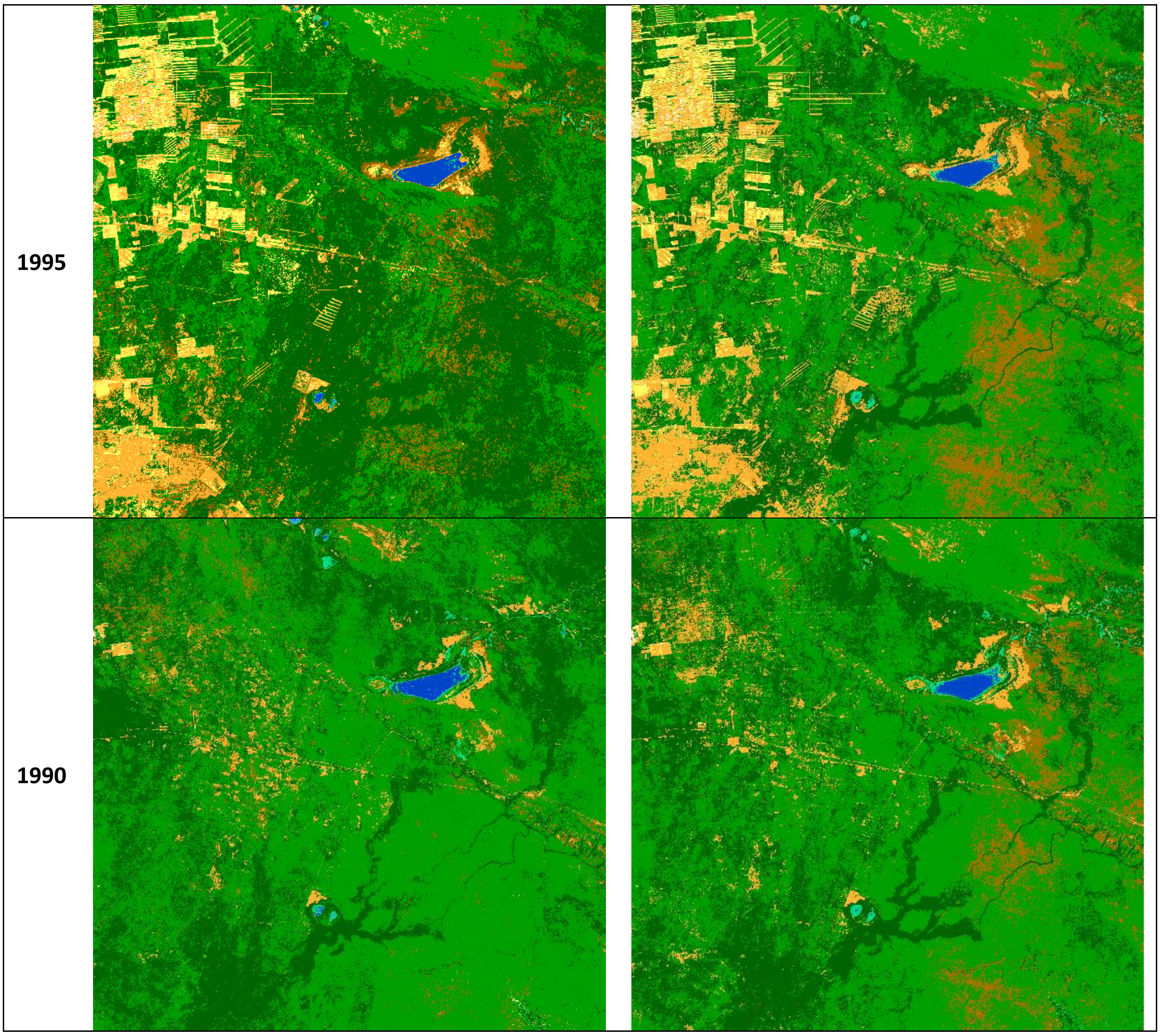
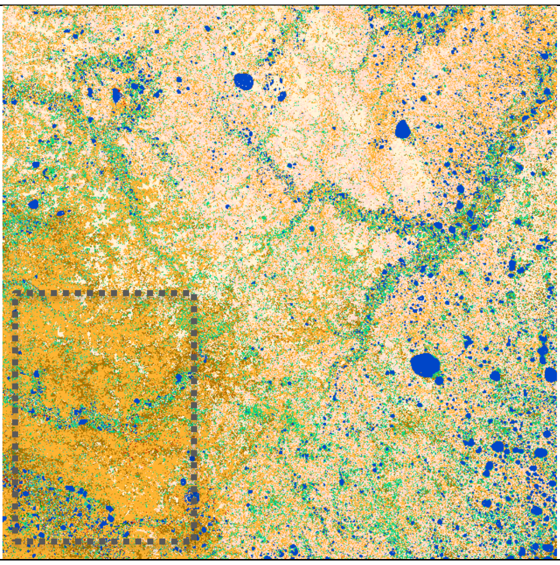
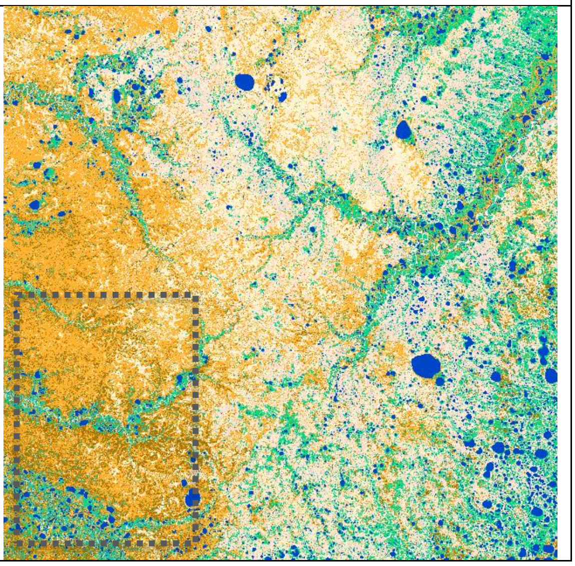
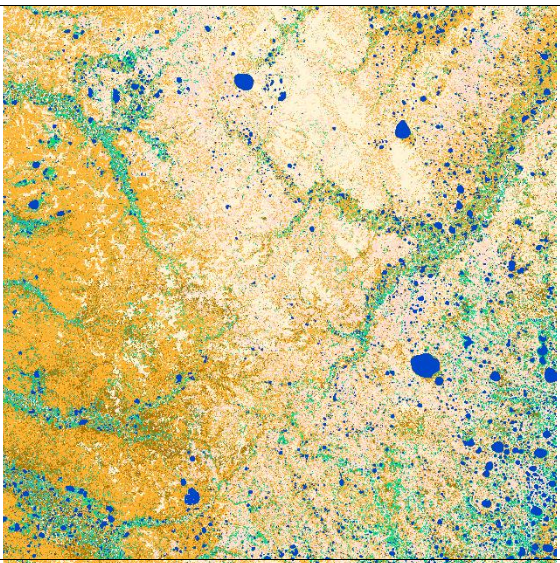
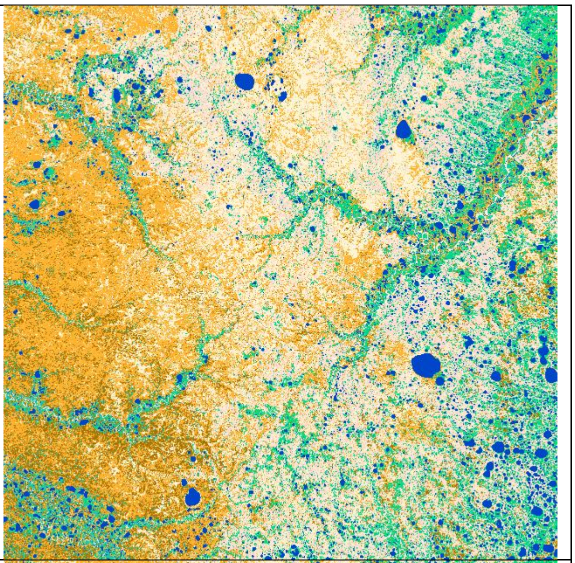
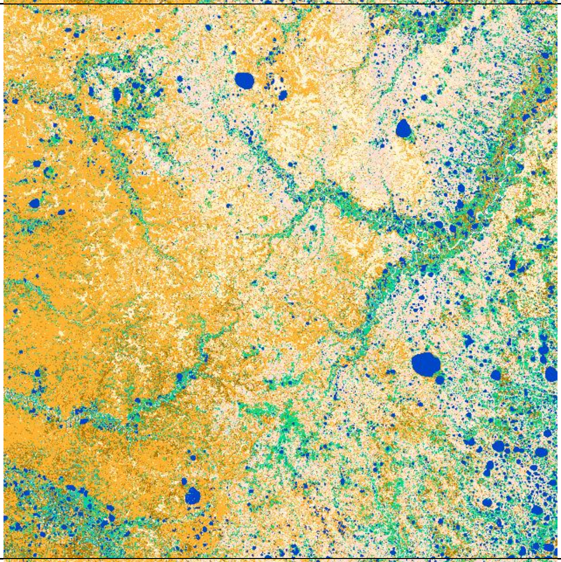
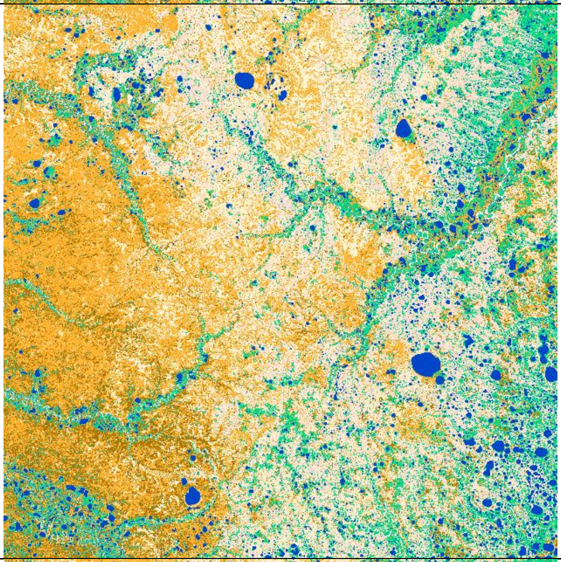
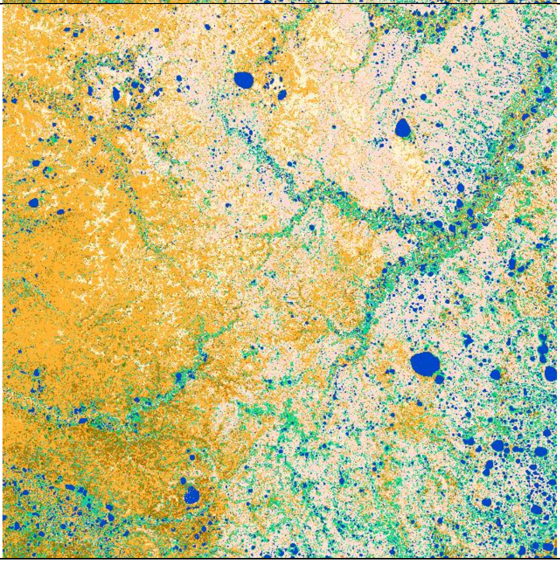
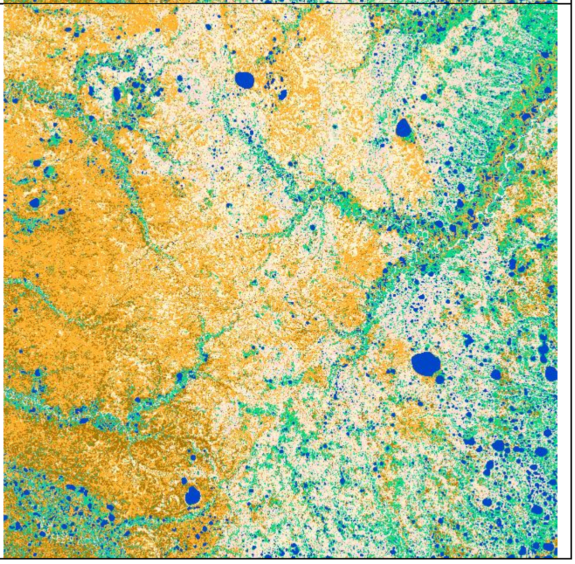


Figure 30. Comparison between the cascade model of Phase 1 (left) and the HMM of Phase 2 (right) for each historical year on tile 20KPF in Amazon

Year	Cascade + MRF	HMM + MRF
2019		

2015		
2010		
2005		

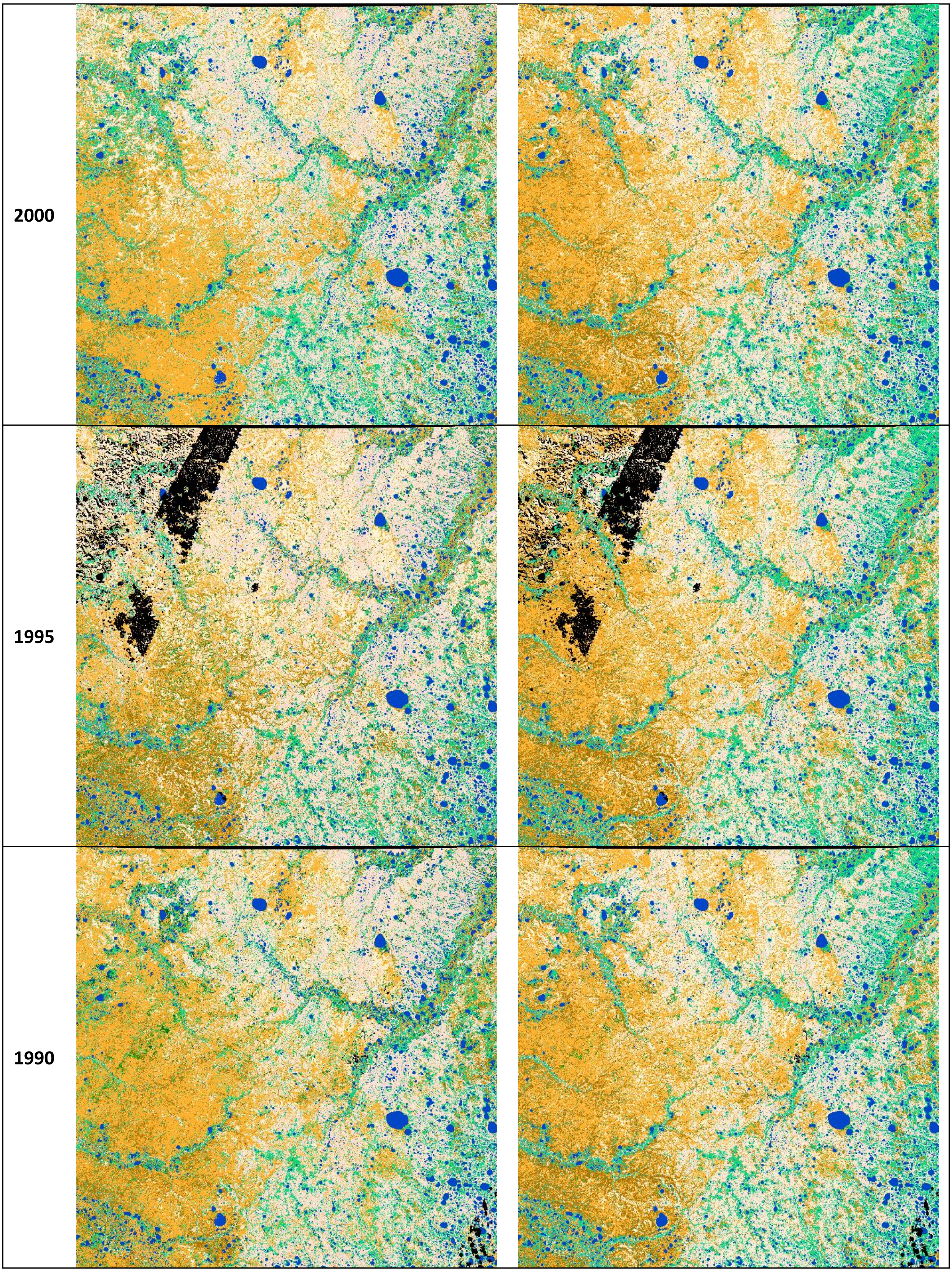
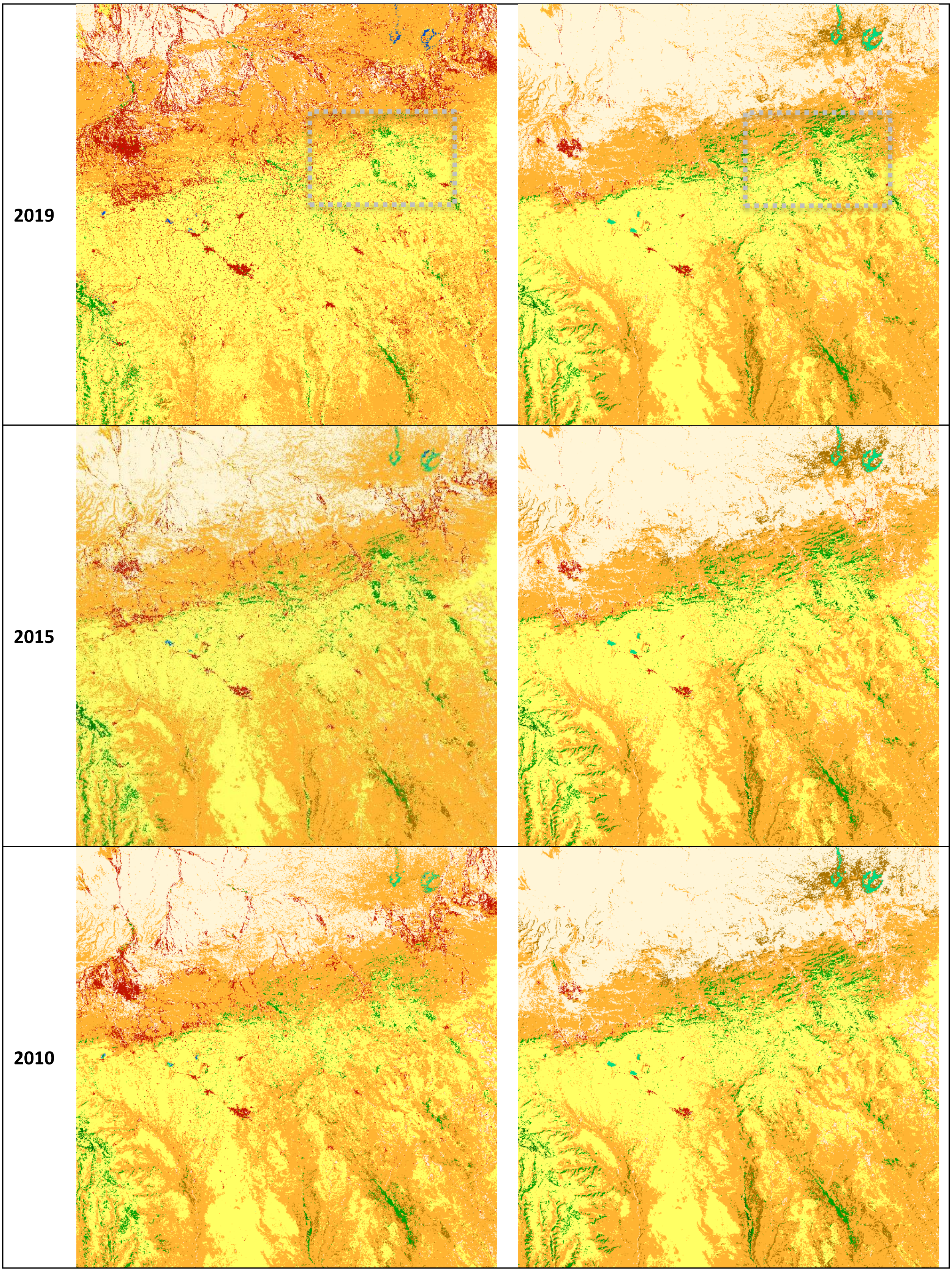
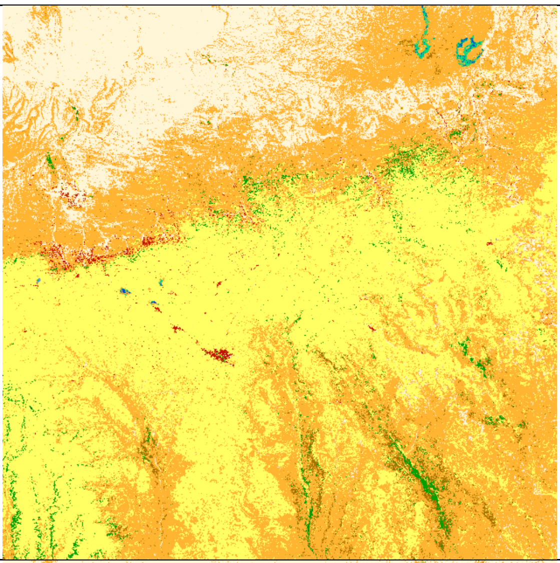
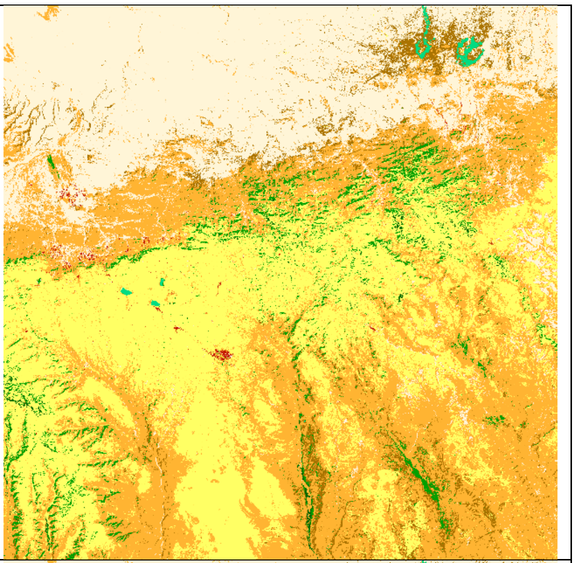
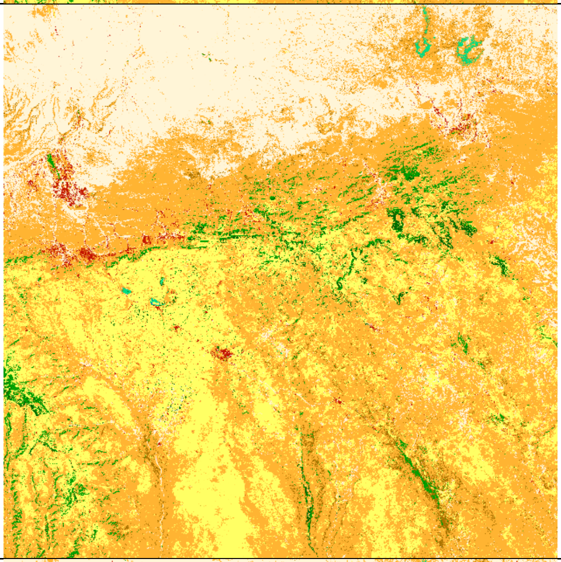
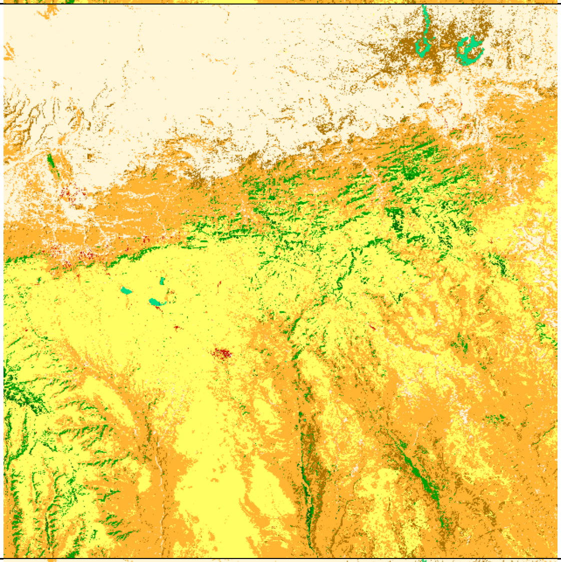
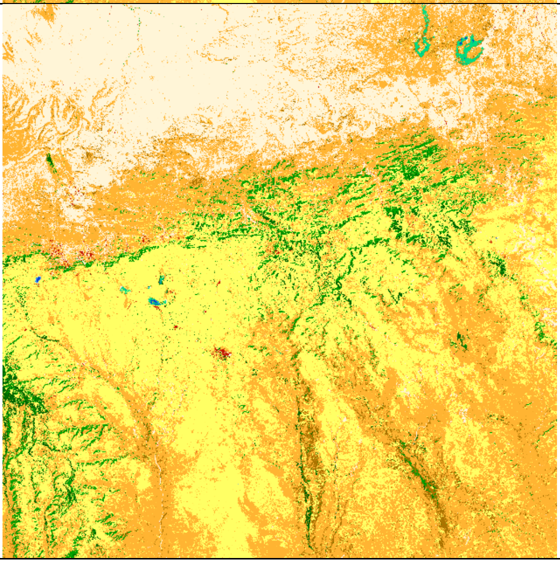
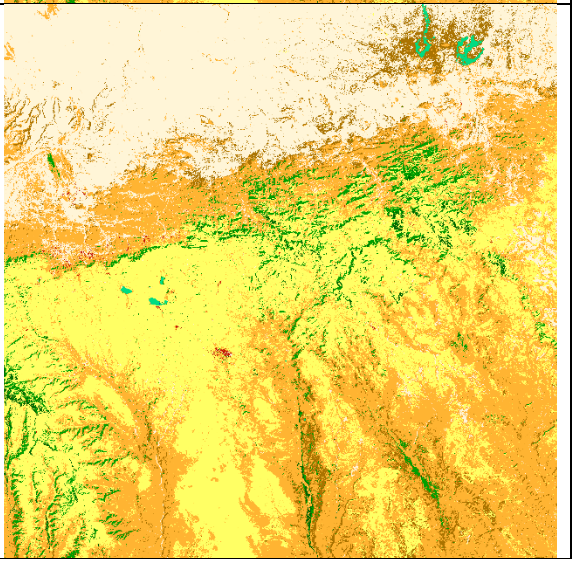


Figure 31. Comparison between the cascade model of Phase 1 (left) and the HMM of Phase 2 (right) for each historical year on tile 42WWD in Siberia. The black pixels in 1995 and 1990 correspond to areas with too limited data availability.

Year	Cascade + MRF	HMM + MRF


2019
2015
2010

2005		
2000		
1995		

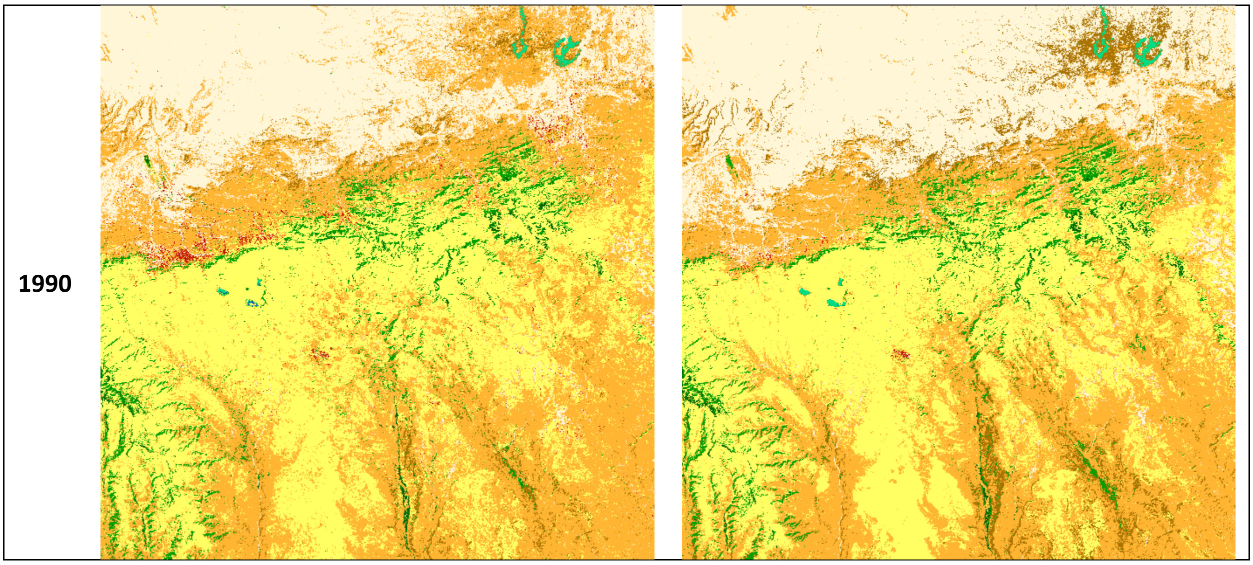


Figure 32. Comparison between the cascade model of Phase 1 (left) and the HMM of Phase 2 (right) for each historical year on tile 37PHL in Africa

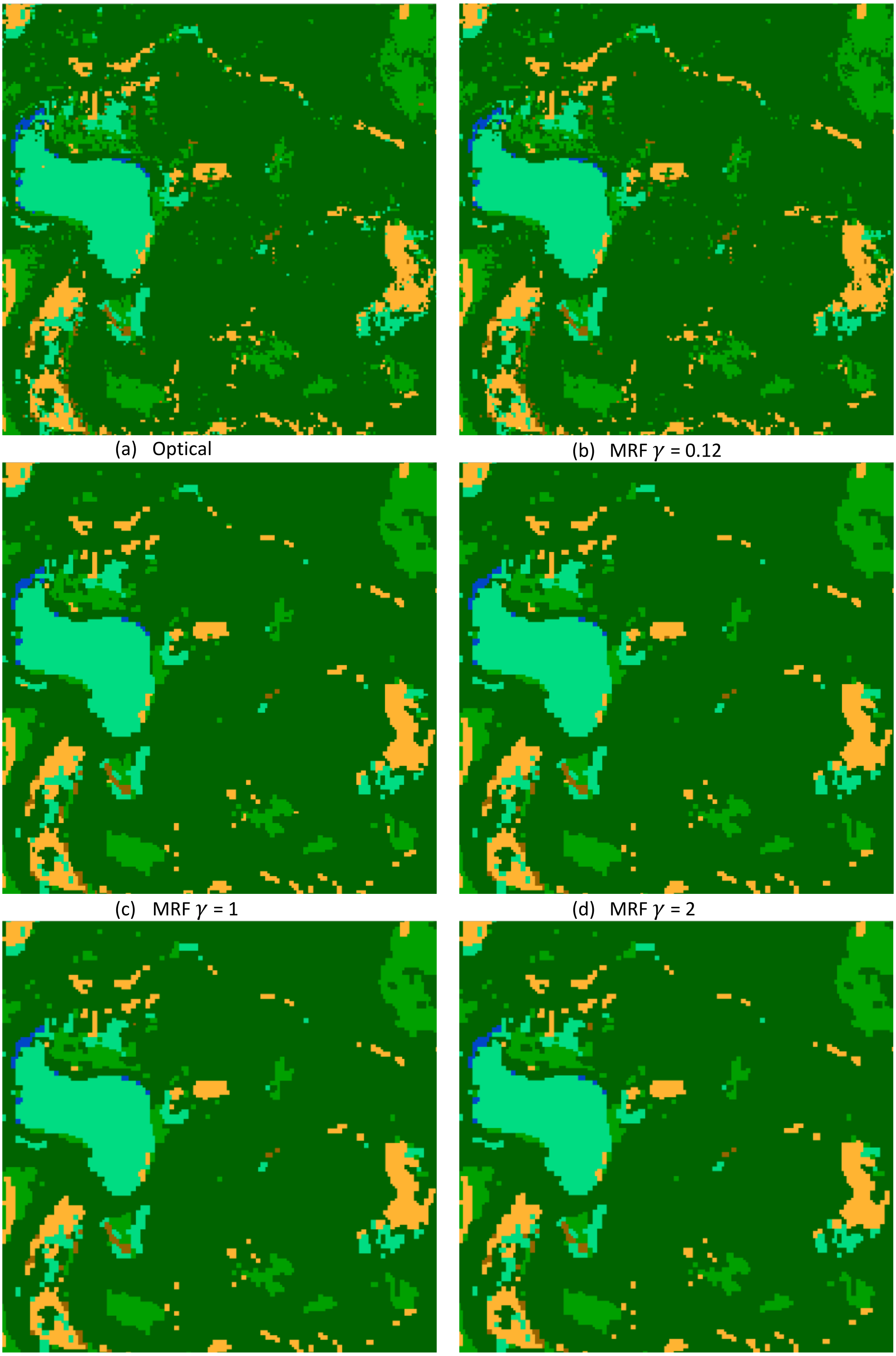
3.3.2.2 Experimental Results – Spatial-Contextual Fusion

As mentioned in the methodology section above, the class-specific weights and a kernel function were introduced to give more flexibility in controlling the degree of spatial regularization. Preliminary class-specific weighting has been considered, but in general, the assignment of class-specific weights in the production will take into account the opinion from the Climate Team regarding which classes should be regularized more than other classes. Here, we will focus the discussion of the results especially on the behaviour allowed by the contrast-sensitive CRF model with its kernel function. The experiments were done on the case of 10-meter resolution static maps, and we tried several values of parameter φ in the Gaussian radial basis function kernel of the CRF model, to appreciate the changes in regularization behaviour. To illustrate this behaviour, we show in Figure 33, Figure 34, and Figure 35, spatial details of the resulting spatial-contextual classification maps. For visual comparison purposes, the non-contextual maps that would be obtained using only the optical pixelwise posteriors are also shown.

Figure 33 shows the zoom-in of an area in tile 20KPF in Amazon. First, let us observe how the regularization parameter in the MRF model, γ, affects the regularization of the pixels. When γ is set to be very small (b), the regularization on the image source (a) is very little, and as it is increased to a bigger value (c), the regularization becomes stronger, as it can be observed that some spatial details are gone. Interestingly, we can see that when γ is enlarged even further, the regularization effect does not take place in this specific area anymore. This is an expected behaviour when the best class has much higher posterior probability than the other remaining classes. Next, with the CRF model, we intentionally set γ to be very large to see a clearer effect of the contrast-sensitive CRF kernel. Theoretically, when φ is close to zero, we will see that the CRF behaves similarly to the MRF. This similarity can be seen in (e) and (f). With a fixed big value of γ, as the value of φ is risen (f, g, h, i), we can see that spatial details are gradually back, while the impact of noise remains limited and without reviving unwanted fragmentation that exists in (b). This effect is expected because the kernel acts, to a certain extent, as an inhibitor of the regularization but only when the probability of the neighbouring pixels are not too similar.

Figure 34 and Figure 35 illustrate the case in tile 37PGN in Africa and in tile 42WWD in Siberia, respectively. With the same γ parameter setting, these figures show the behaviour of the classification results as a function of the tuning of the φ parameter of the CRF model. While in the African tile, we can observe a behaviour similar to that described in the case of the Amazonian tile, it is interesting to see that, in the Siberian tile, the comeback of the details is not as apparent as in the other two tiles as the φ parameter increases. This is interpreted as due to the uncertainty in this area being higher than in the other two areas, i.e., the posterior probabilities of neighbouring pixels appear to be relatively similar. In turn, this implies that the increasing of the φ value brings back only spatial details that have dissimilarity to their neighbourhoods up to a certain point. Indeed, this is a desired behaviour since it confirms the capability of the proposed approach to take into account and propagate uncertainty from the input data (and their possible data availability issues) to the output land cover product.

These experimental results confirm that using the CRF model gives us improved control over the regularization level, compared to the MRF model. This yields an increased flexibility in taking into account the feedback and needs from the Climate Team. Indeed, the spatial regularization of the output map in the production will be tuned according to the indications from the Climate Team about the desired trade-off between preserving spatial resolution and favouring spatial homogeneous labelling and about how this trade-off may change in the case of different classes.



(a) Optical

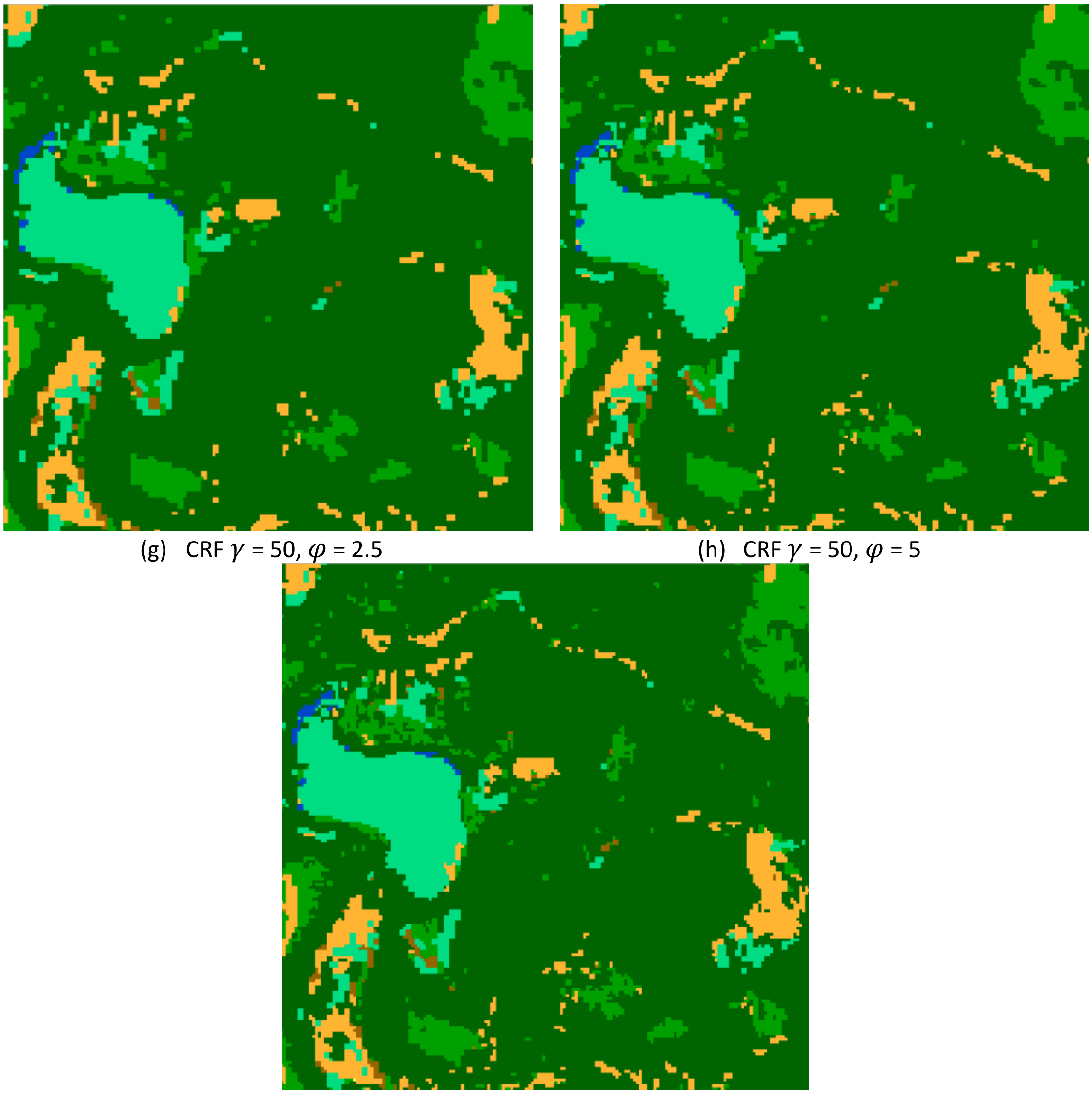
(b) MRF γ = 0.12

(c) MRF γ = 1

(d) MRF γ = 2

(e) MRF γ = 50

(f) CRF γ = 50, φ = 0.5

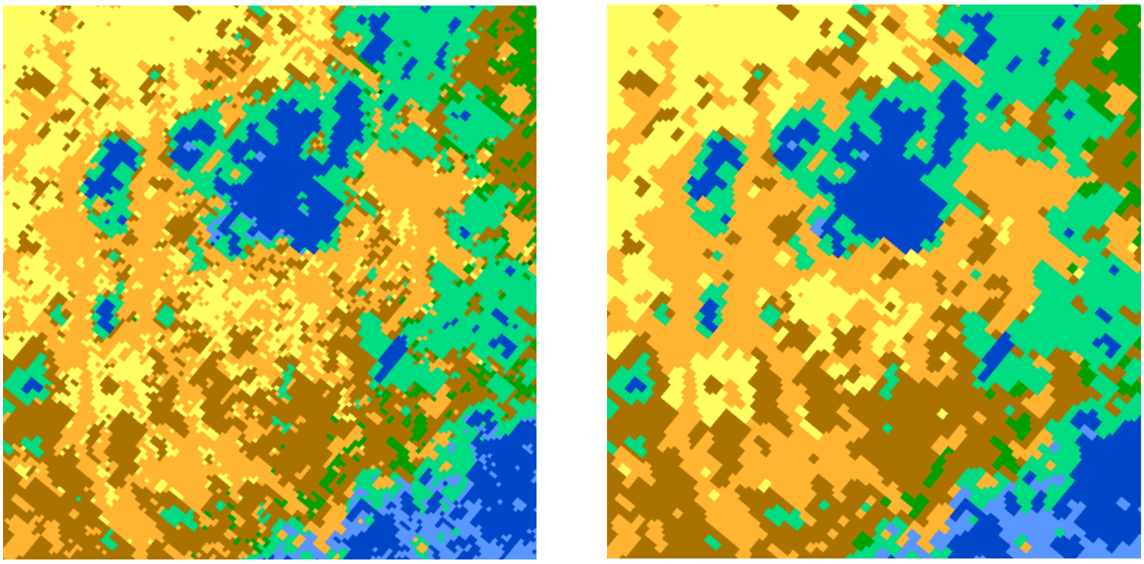


(g) CRF γ = 50, φ = 2.5

(h) CRF γ = 50, φ = 5

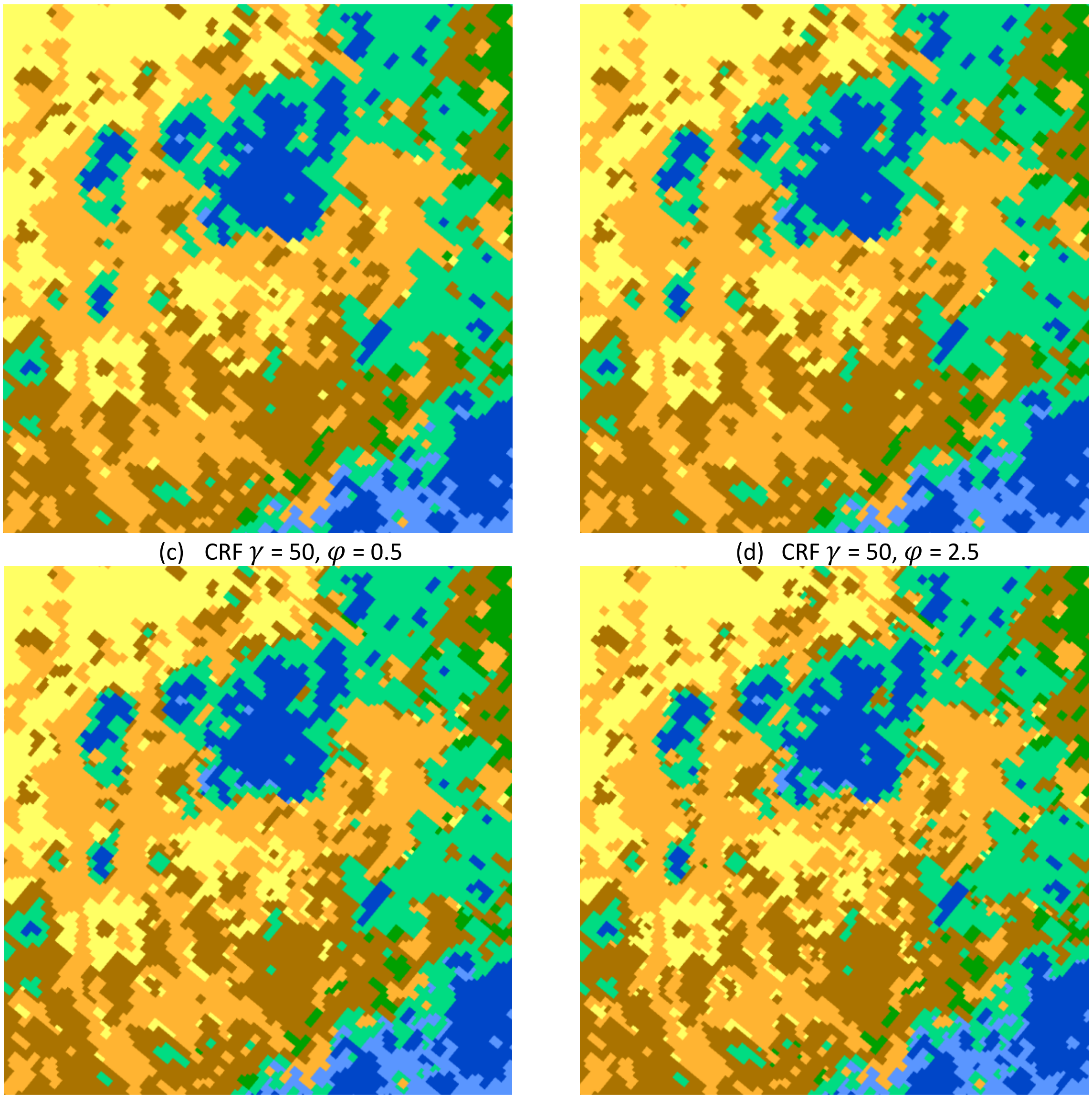
(i) CRF γ = 50, φ = 10

Figure 33. The comparison among (a) the map resulting from the original optical posteriors, (b,c,d,e) MRF with several regularization parameters, and (f,g,h,i) CRF with several kernel parameters



(a) Optical

(b) MRF γ = 50



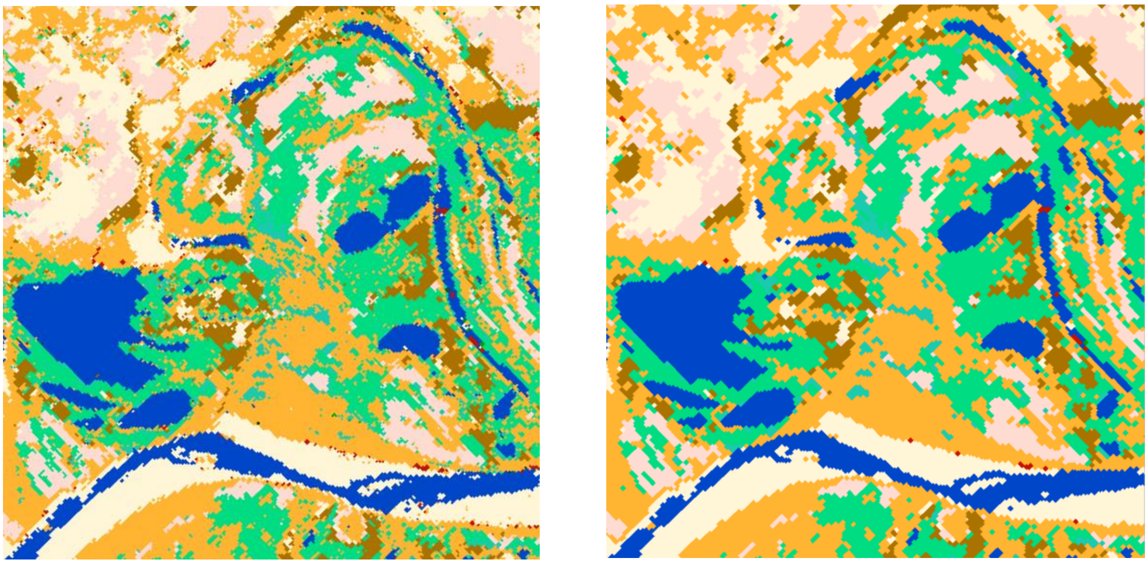
(c) CRF $\gamma = 50$, $\varphi = 0.5$

(d) CRF $\gamma = 50$, $\varphi = 2.5$

(e) CRF $\gamma = 50$, $\varphi = 5$

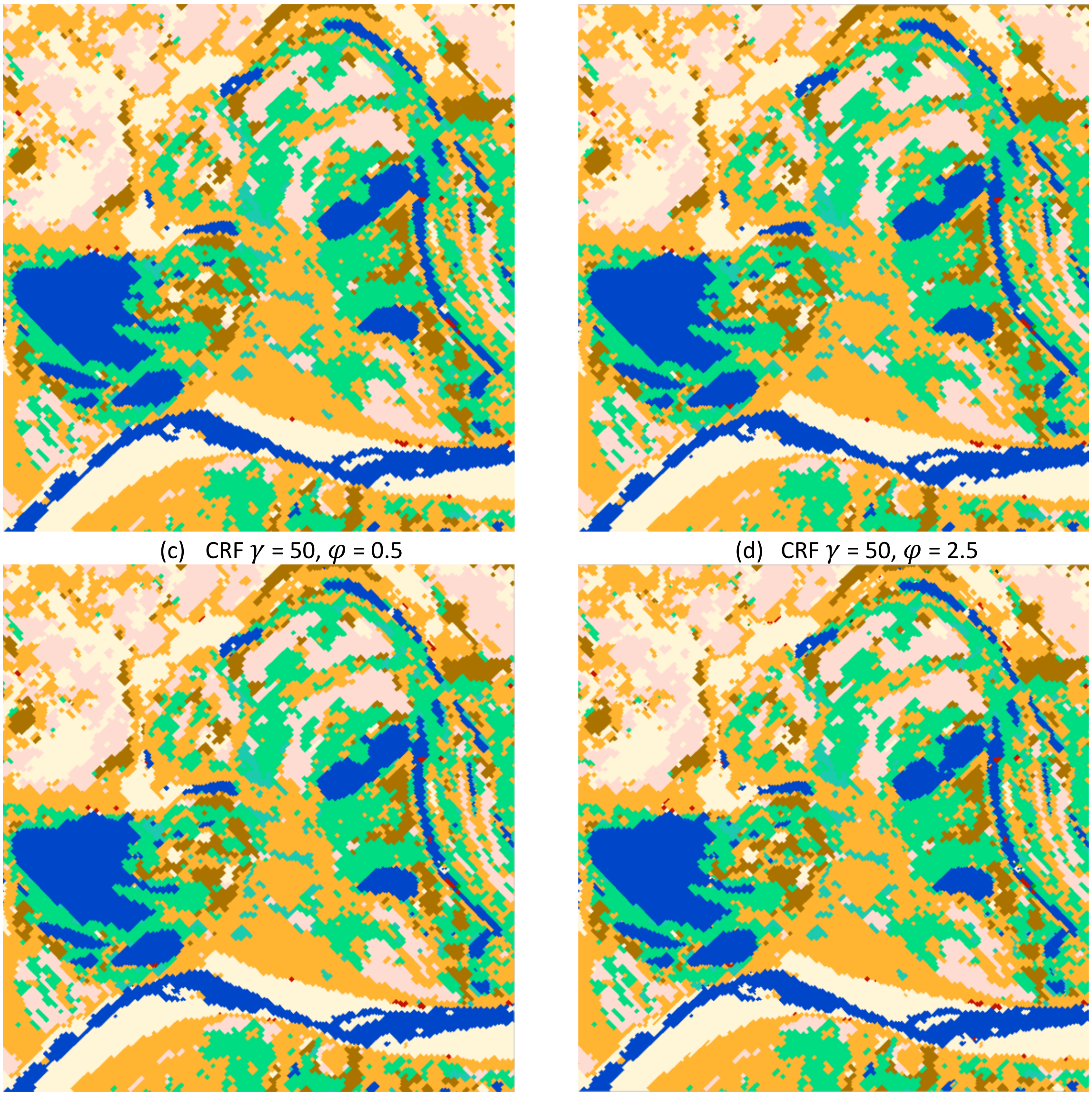
(f) CRF $\gamma = 50$, $\varphi = 10$

Figure 34. The comparison among (a) the map resulting from the original optical posteriors, (b) MRF with a big regularization parameter, and (c,d,e,f) CRF with several kernel parameters



(a) Optical

(b) MRF $\gamma = 50$



(c) CRF γ = 50, φ = 0.5

(d) CRF γ = 50, φ = 2.5

(e) CRF γ = 50, φ = 5

(f) CRF γ = 50, φ = 10

Figure 35. The comparison among (a) the map resulting from the original optical posteriors, (b) MRF with a big regularization parameter, and (c,d,e,f) CRF with several kernel parameters

3.3.3 Final decision

In the context of multi-temporal fusion, HMM has demonstrated to provide more consistent results temporally compared to the cascade model. Similarly, the CRF model gives enhanced flexibility in guarding the spatial regularization process than the MRF model can do and allows for stronger spatial smoothing while preserving edges and small-scale details. Therefore, HMM and CRF-ICM are identified as appropriate solutions towards the first production.

Further development of both models may regard the tuning of the transition probability matrix of HMM, and the evaluation of alternate kernel choices in the CRF energy function, which may consider introducing similarity operators between the two probability distributions $P_s(\mathbf{X})$ and $P_r(\mathbf{X})$ instead of calculating the distance between them within a radial basis function kernel. These possible variations will be considered also in relation to the point of view of computational cost during the production.

3.4 Landcover Change Detection

In phase 2, the analysis and change detection process expands to include new regions and incorporates S2 datasets. The CD processing chain continues to operate at the pixel level on a yearly basis from 1990 to 2024, with a focus on three subcontinental areas: Amazonia, Africa, and Siberia (see Figure 36). This phase introduces several updates to the methodologies, accounting for the higher spatial resolution of S2 data, the challenge of harmonizing multi-sensor data across these diverse regions, and the need to incorporate advanced techniques for managing both temporal and spatial variability in land cover dynamics. The regional differences in land cover patterns, environmental conditions, and spatial distribution of vegetation types require careful adaptation of the methodology to ensure accurate change detection outcomes.

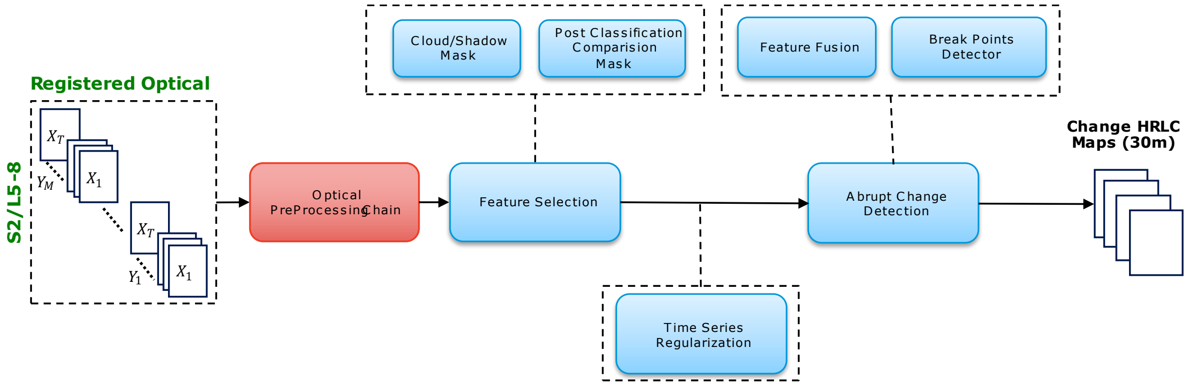


Figure 36. Multi-annual Multi-feature LC change detection.

To address these challenges, the preprocessing stages from optical data analysis were revisited, aiming to create an integrated and unified workflow for both land cover classification and land cover change detection. The goal is to streamline the preprocessing process across different data types (optical and radar), reducing the number of steps required while ensuring consistency and robustness in the data preparation pipeline. This approach is particularly important given the increasing complexity of data sources and the need for efficient processing across large datasets, such as those from S2, which provide more frequent, and higher-resolution imagery compared to other sources.

In terms of feature space design, an analysis was conducted to identify the most suitable features for the change detection task, considering the specific characteristics of the regions under study. The feature space will be regionally optimized, with distinct feature sets developed for Amazonia, Africa, and Siberia to capture the unique land cover dynamics of each area. This customization is essential for improving the accuracy of change detection. Furthermore, the integration of advanced techniques for managing spatio-temporal variability, such as those derived from machine learning models and time-series analysis, will allow for more accurate detection of abrupt changes and better handling of temporal variations.

3.4.1 Time Series Reconstruction

An analysis was conducted to evaluate the trend of the time series using the monthly composites generated in phase 1 for one of the tiles in Amazonia. The primary goal of this analysis was to assess whether the monthly composites, derived from the optical processing chain, could effectively capture the underlying trends of land cover changes over time without the need for time-consuming and computationally heavy time series reconstruction. The reconstruction process typically involves interpolating or filling in missing data points, which can be resource-intensive, especially when working with large datasets. By considering only one composite per month, the analysis aimed to test if this approach could maintain the temporal dynamics of the land cover changes and still yield accurate results.

The results of the analysis, displayed in the Figure 37 and Figure 38, show the comparisons between the weekly time series reconstruction from phase 1 and the monthly composites. This comparison indicates that the monthly composites are capable of handling the irregularities in the time series and adequately reflecting the temporal trends, even without the traditional reconstruction step. This suggests that it is possible to use monthly composites to capture the variation and dynamics in land cover changes, making the process more efficient. Consequently, this approach offers a promising alternative to time series reconstruction, especially when dealing with high-resolution S2 data in phase 2, as it reduces computational costs while still ensuring reliable breakpoint detection.

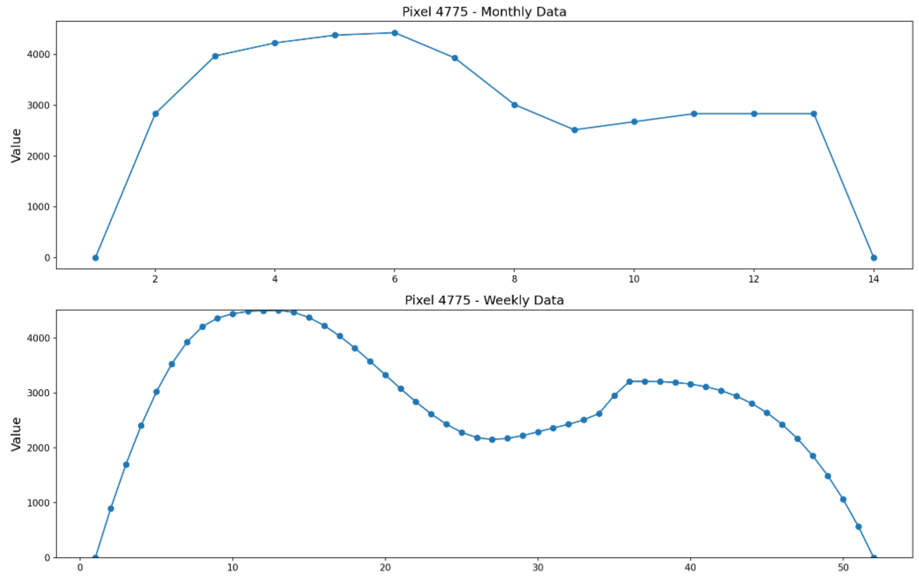


Figure 37. Comparison of time series data of the weekly time series reconstruction method and time series of the monthly composites for 2019.

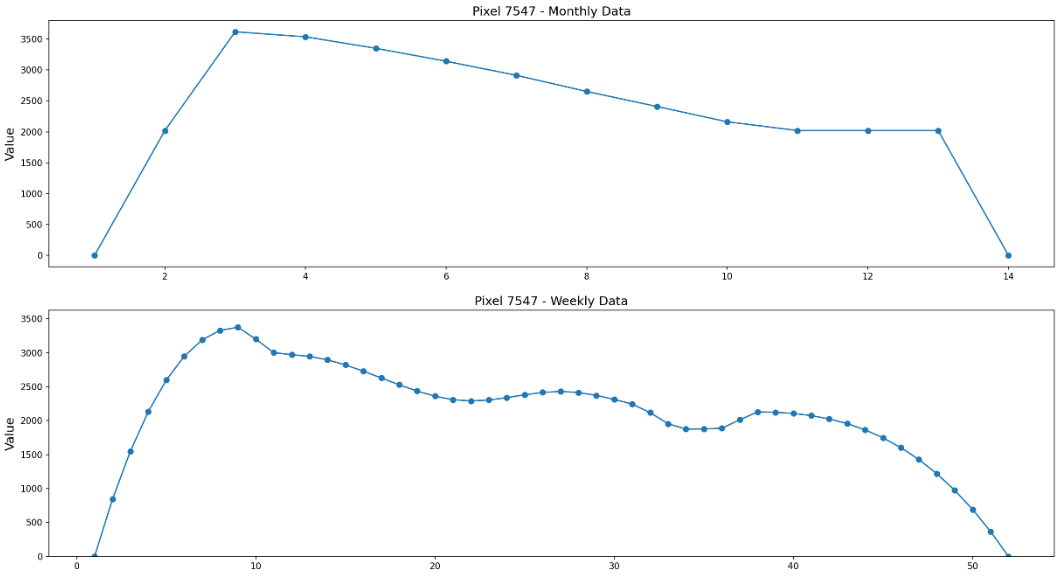


Figure 38. Comparison of time series data of the weekly time series reconstruction method and time series of the monthly composites for 2019.

3.4.2 Feature Selection

Another analysis was conducted to optimize feature selection for the developments in phase 2, focusing on several indices that are under evaluation for each of the areas of study. Each index provides valuable information about land cover, aiding in the analysis of changes. The analysis demonstrates that, depending on the specific area of study, certain indices may be more effective than others. In this context, temporal features are particularly useful for capturing how spectral information evolves over time, facilitating the identification of abrupt land cover changes. Time series analysis of indices such as Normalized Difference Vegetation Index (NDVI), Normalized Difference Water Index (NDWI), and Bare Soil Index (BSI) provides insights into patterns like deforestation, regrowth, and seasonal fluctuations in vegetation or water bodies. Additionally, feature selection will consider features extracted from the optical processing chain, such as deep learning-derived features and textural metrics like GLCM. Based on their relevance to change detection, these features will be incorporated into the feature selection and feature fusion process to enhance the accuracy and reliability of the analysis.

So, for a vegetation area affected by deforestation, the indices NDVI, NDWI, and BSI were combined for further analysis. The fusion of NDVI and BSI was done using the following equation:

$$fused_{index} = \sqrt{\frac{NDVI^2 + NDWI^2 + BSI^2}{3}}$$

This equation assumes that all three indices (NDVI, NDWI, and BSI) are equally weighted. Different weights could be assigned to each index, by considering the modified equation below. For example, if you wanted to give more importance to NDVI, the equation might look like this:

$$fused_{index} = \sqrt{\frac{\omega_1 . NDVI^2 + \omega_2 . NDWI^2 + \omega_2 . BSI^2}{\omega_1 + \omega_2 + \omega_3}}$$

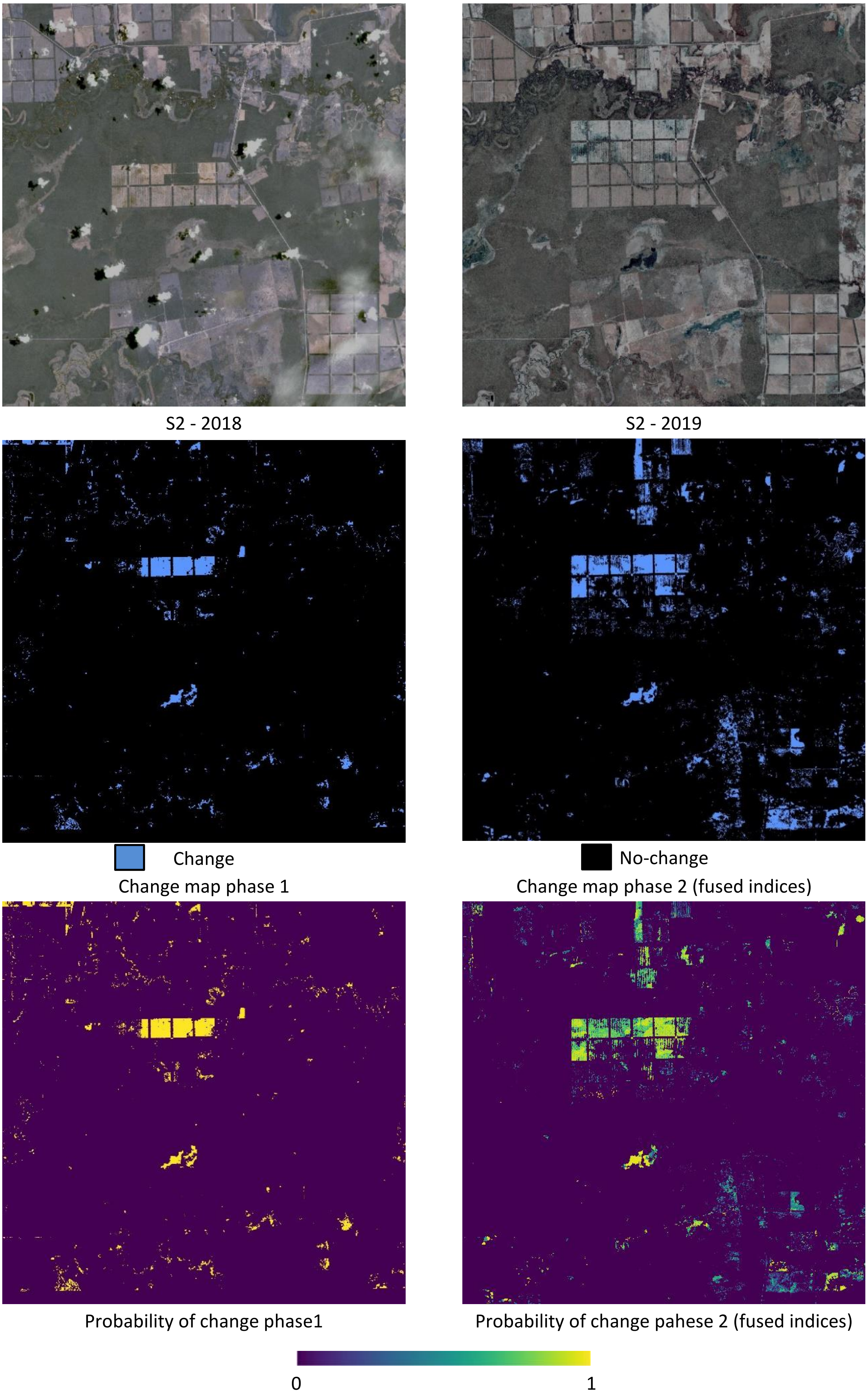
The weights for combining indices can be defined in several ways, depending on the specific application and what aspects of the indices are most important for the analysis. An analysis is currently ongoing to select an efficient and effective method for defining the weights using state-of-the-art methods [20], [21].

The outcome of this analysis, performed for the years 2018 and 2019 in the Amazonia area that experienced deforestation because of the crop construction, is shown in Figure 39. In this analysis, monthly composites were generated using high-resolution Sentinel-2 data and compared with the change map produced in Phase 1 using the Landsat dataset. The analysis was conducted without using the PCC mask, to assess the effectiveness of feature representations and the breakpoint detector in achieving accurate change detection. For the latter analysis, the PCC will be regenerated using Phase 2 products to align with the land cover (LC) maps and reduce computational burden.

By considering the sample S2 images provided in first row of the Figure 39, the changes that occurred from 2018 to 2019 reveal that the crop field has continued to be constructed, indicating ongoing agricultural development in the area. The change map generated using fused indices demonstrates following aspects:

- Improved change detection through fused indices: The change map generated using fused indices (NDVI, NDWI, and BSI) demonstrates enhanced change detection compared to the Phase 1 map. This improvement is due to the complementary information provided by these indices, which collectively increase the accuracy of identifying changes specific to the study area.
- Higher resolution insights with Sentinel-2 data: The use of high-resolution Sentinel-2 data at 10m resolution reveals finer details of the changes, offering a more precise representation of the landscape (see Figure 40). Notably, even the stripes within the crop field are distinctly detected, underscoring the capability of Sentinel-2 data to capture fine spatial details.
- Effective temporal change detection: The breakpoint detector effectively identified the year of change, showcasing its ability to handle high-resolution data for accurate temporal change detection.
- Efficient analysis without time series reconstruction: The analysis demonstrates that it is possible to bypass the computationally intensive time series reconstruction process by using Sentinel-2 composites. These composites provide sufficient information to capture temporal trends and detect changes accurately.

The updated approach enhances change detection by both in space and time while offering an efficient alternative to time series reconstruction, thereby reducing the computational burden on the processing chain. Furthermore, the analysis continues to focus on generating a comprehensive feature space for each area of interest, using the features generated from optical classification chain to further refine and enhance the results.



S2 - 2018

S2 - 2019

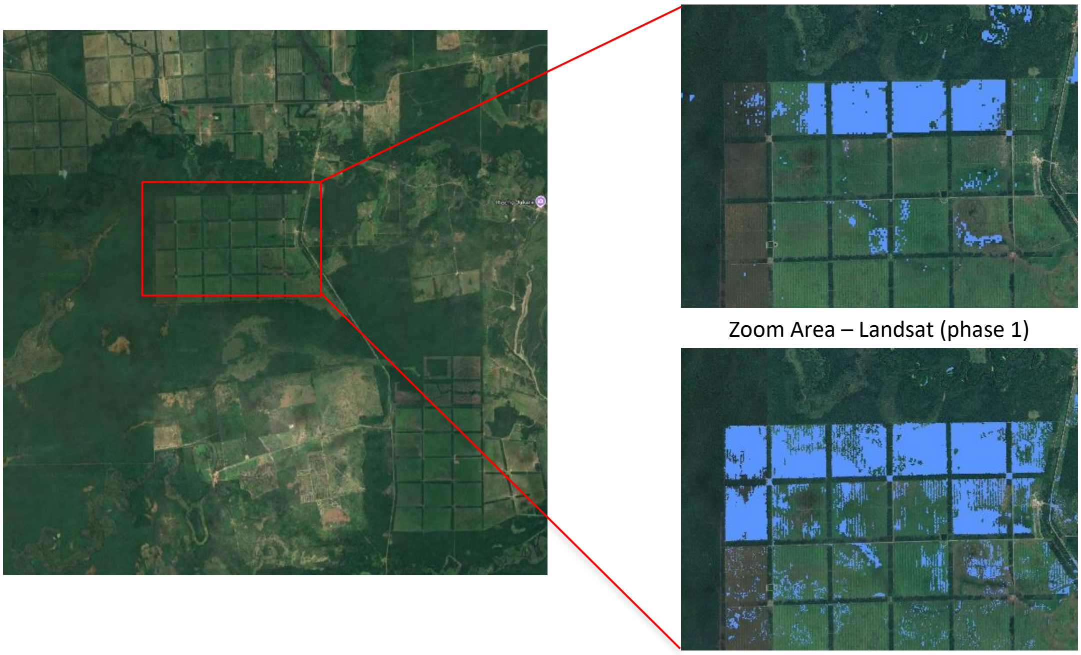
Change map phase 1

Change map phase 2 (fused indices)

Probability of change phase1

Probability of change pahese 2 (fused indices)

Figure 39. A comparison of phase 1 change map and the change map using fused feature space.



Zoom Area – Landsat (phase 1)

Zoom Area – Sentinel-2 (phase 2)

Figure 4041. The details of the generated change maps using Sentinel-2 fused features and phase1.

4 References

[1] S. Qiu, Z. Zhu, and B. He, "Fmask 4.0: Improved cloud and cloud shadow detection in Landsats 4–8 and Sentinel-2 imagery," *Remote Sensing of Environment*, vol. 231, p. 111205, Sep. 2019, doi: 10.1016/j.rse.2019.05.024.

[2] C. Aybar *et al.*, "CloudSEN12, a global dataset for semantic understanding of cloud and cloud shadow in Sentinel-2," *Sci Data*, vol. 9, no. 1, p. 782, Dec. 2022, doi: 10.1038/s41597-022-01878-2.

[3] C. Aybar *et al.*, "CloudSEN12+: The largest dataset of expert-labeled pixels for cloud and cloud shadow detection in Sentinel-2," *Data in Brief*, vol. 56, p. 110852, Oct. 2024, doi: 10.1016/j.dib.2024.110852.

[4] D. Frantz, "FORCE—Landsat + Sentinel-2 Analysis Ready Data and Beyond," *Remote Sensing*, vol. 11, no. 9, Art. no. 9, Jan. 2019, doi: 10.3390/rs11091124.

[5] M. Rußwurm and M. Körner, "Self-attention for raw optical Satellite Time Series Classification," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 169, pp. 421–435, Nov. 2020, doi: 10.1016/j.isprsjprs.2020.06.006.

[6] Z. Geng, L. Liang, T. Ding, and I. Zharkov, "RSTT: Real-time Spatial Temporal Transformer for Space-Time Video Super-Resolution," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 17420–17430. doi: 10.1109/CVPR52688.2022.01692.

[7] X. SHI, Z. Chen, H. Wang, D.-Y. Yeung, W. Wong, and W. WOO, "Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2015. Accessed: Jan. 07, 2025. [Online]. Available: https://proceedings.neurips.cc/paper/2015/hash/07563a3fe3bbe7e3ba84431ad9d055af-Abstract.html

[8] D. Marzi, A. Sorriso, and P. Gamba, "Automatic wide area land cover mapping using Sentinel-1 multitemporal data," *Front. Remote Sens.*, vol. 4, Dec. 2023, doi: 10.3389/frsen.2023.1148328.

[9] A. Vollrath, A. Mullissa, and J. Reiche, "Angular-Based Radiometric Slope Correction for Sentinel-1 on Google Earth Engine," *Remote Sensing*, vol. 12, no. 11, Art. no. 11, Jan. 2020, doi: 10.3390/rs12111867.

[10] M. Buchhorn, M. Lesiv, N.-E. Tsendbazar, M. Herold, L. Bertels, and B. Smets, "Copernicus Global Land Cover Layers—Collection 2," *Remote Sensing*, vol. 12, no. 6, Art. no. 6, Jan. 2020, doi: 10.3390/rs12061044.

[11] ESA, "Land Cover CCI Product User Guide Version 2.0." [Online]. Available: http://maps.elie.ucl.ac.be/CCI/viewer/download/ESACCI-LC-Ph2-PUGv2_2.0.pdf

[12] D. Marzi, J. I. S. Jara, and P. Gamba, "A 3-D Fully Convolutional Network Approach for Land Cover Mapping Using Multitemporal Sentinel-1 SAR Data," *IEEE Geoscience and Remote Sensing Letters*, vol. 21, pp. 1–5, 2024, doi: 10.1109/LGRS.2023.3332765.

[13] Gorica Bratic and Maria Antonia Virelli, "Map Of Land Cover Agreement - MOLCA." Accessed: Jun. 06, 2024. [Online]. Available: https://zenodo.org/records/8071675

[14] M. Pashaei, H. Kamangir, M. J. Starek, and P. Tissot, "Review and Evaluation of Deep Learning Architectures for Efficient Land Cover Mapping with UAS Hyper-Spatial Imagery: A Case Study Over a Wetland," *Remote Sensing*, vol. 12, no. 6, Art. no. 6, Jan. 2020, doi: 10.3390/rs12060959.

[15] D. Marzi and P. Gamba, "Inland Water Body Mapping Using Multi-temporal Sentinel-1 SAR Data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. PP, pp. 1–1, Nov. 2021, doi: 10.1109/JSTARS.2021.3127748.

[16] Swain, "Bayesian Classification in a Time-Varying Environment," *IEEE Trans. Syst., Man, Cybern.*, vol. 8, no. 12, pp. 879–883, 1978, doi: 10.1109/TSMC.1978.4309889.

[17] C. M. Bishop, *Pattern recognition and machine learning*. in Information science and statistics. New York: Springer, 2006.

[18] S. Z. Li, *Markov Random Field Modeling in Image Analysis*. in Advances in Pattern Recognition. London: Springer London, 2009. doi: 10.1007/978-1-84800-279-1.

[19] R. Szeliski *et al.*, "A Comparative Study of Energy Minimization Methods for Markov Random Fields with Smoothness-Based Priors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 6, pp. 1068–1080, Jun. 2008, doi: 10.1109/TPAMI.2007.70844.

[20] B. Desclée, P. Bogaert, and P. Defourny, "Forest change detection by statistical object-based method," *Remote sensing of environment*, vol. 102, no. 1–2, pp. 1–11, 2006.

[21] U. Paquet, "Empirical Bayesian change point detection," *Graphical Models*, vol. 1995, pp. 1–20, 2007.