Ref. Ares(2025)2896103 - 09/04/2025

Evoland
LAND MONITORING EVOLUTION

D2.16 Methods Compendium - final version

Work Package 2 – Method development

Authors:

Ruben Van De Kerchove, Wanda De Keersmaecker, Kasper Bonte, Astrid Vannoppen, Mathilde De Vroey, Nick Gutkin(VITO), Conrad M. Albrecht, Nassim Ait Ali Braham (DLR), Jordi Inglada, Ekaterina Kalinicheva, Julien Michel (UT3), Janik Deutscher, Martin Puhm, Heinz Gallaun, Felix Bernsteiner (JR), Alexandre Bouvet (IRD)

Date: 31.03.2025

Funded by the European Union

Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

Full Title	**Evolution of the Copernicus Land Service portfolio** (integrating novel EO data and latest machine learning algorithms to continuously monitor the status, dynamics and biomass of the land surface)			
Project number	101082130		**Acronym**	EVOLAND
Start date	01.01.2023		**Duration**	36 months
Granting authority	European Health and Digital Executive Agency (HaDEA)			
Project Coordinator	Vlaamse Instelling voor Technologisch Onderzoek N.V. (VITO)			
Date of delivery	**Contractual**	M27	**Actual**	M27
Type	R - Document, report		**Dissemination level**	PU - Public
Lead beneficiary	VITO			
Lead author	Ruben Van De Kerchove		**Email**	Ruben.vandekerchove@vito.be
Other authors	Wanda De Keersmaecker, Kasper Bonte, Nick Gutkin, Mathilde De Vroey (VITO), Conrad Albrecht, Nassim Ait Ali Braham (DLR), Jordi Inglada, Ekaterina Kalinicheva, Julien Michel (UT3), Janik Deutscher, Martin Puhm, Heinz Gallaun, Felix Bernsteiner (JR), Alexandre Bouvet (IRD)			
Reviewer(s)	Tim Ng (VITO)			
Keywords	Methodology, state-of-the-art, phase 2			

Document Revision History				
Version	**Issue date**	**Stage**	**Changes**	**Contributor**
0.1	10.02.2025	Draft	Review D2.16	VITO
0.2	15.03.2025	Draft	Consolidation partner input	ALL
1.0	31.03.2025	Final	Review and finalization	VITO

Disclaimer

Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

Copyright message

© EvoLand consortium, 2025

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgment of previously published material and of the work of others has been made through appropriate citation, quotation or both. Reproduction is authorised provided the source is acknowledged.

Acronyms

Acronym	Description
AGB	Above Ground Biomass
AGBD	Above Ground Biomass Density
ALOS	Advanced Land Observing Satellite
ALISE	Aligned SITS Encoder
AVOCADO	Anomaly Vegetation Change Detection
AD	Applicable Document
API	Application Programming Interface
AI	Artificial Intelligence
AE	AutoEncoders
BEAST	Bayesian Estimator of Abrupt change, Seasonal change, and Trend
BCE	Binary Cross Entropy
BRISQUE	Blind Referenceless Image QUality Evaluator
BFAST	Breaks for Additive Season and Trend
BNETD	Bureau National d'Études Techniques et de Développement
CARN	Cascading Residual Network
CCDC	Change Detection and Classification algorithm
CCI	Climate Change Initiative
CIGN	Comité Interministériel de la Géomatique Nationale
COTS	Commercial Off-The-Shelf
CEOS	Committee on Earth Observation Satellites
CNN	Convolutional Neural Network
COPDEM	Copernicus Digital Elevation Model
CGLS	Copernicus Global Land Service
CHIME	Copernicus Hyperspectral Imaging Mission for the Environment
CLMS	Copernicus Land Monitoring Service
CDL	Cropland Data Layer
CE	Cross Entropy
CROMA	Cross-Resolution Object Matching Algorithm
CUSUM	Cumulative Sum Control Chart
DEM	Digital Elevation Model
DESIS	DLR Earth Sensing Imaging Spectrometer Mission
DLT	Dominant Leaf Type
EO	Earth Observation
ESRGAN	Enhanced Super-Resolution Generative Adversarial Network
EnMAP	Environmental Mapping and Analysis Program
ECMWF	European Centre for Medium-Range Weather Forecasts
EFFIS	European Forest Fire Information System
EPSG	European Petroleum Survey Group
ESA	European Space Agency

EWMA	Exponentially Weighted Moving Average
FAO	Food and Agriculture Organization
FCH	Forest Canopy Height
FAPAR	Fraction of Absorbed Photosynthetic Active Radiation
FCOVER	Fraction of Vegetation Cover
FCN	Full Convolutional Network
GAN	Generative adversarial network
GEDI	Global Ecosystem Dynamics Investigation
GPU	Graphics Processing Unit
GPP	Gross Primary Production
GRD	Ground Range Detected
HRL	High-Resolution Layer
HOG	Histogram of Oriented Gradients
HIS	Hue, Intensity, Saturation
IT	Information Technology
IGN	Institut Géographique National
IEEE	Institute of Electrical and Electronics Engineers
ICOS	Integrated Carbon Observation System
JAXA	Japan Aerospace Exploration Agency
JRC	Joint Research Centre
KF	Kalman Filter
KLD	Kullback-Leibler Divergence
LSC	Land Surface Characteristics
LAI	Leaf Area Index
LPIPS	Learned Perceptual Image Patch Similarity
LUE	Light Use Efficiency
LR	Low Resolution
MAE	Masked Autoencoder
MGRS	Military Grid Reference System
MTF	Modulation Transfer Function
MoCo	Momentum Contrast
MS	Multi Spectral
MLP	Multi-Layer Perception
MM	Multi-Modal
MALICE	Multimodal ALIse Cross-Encoder
MMDC	Multi-Modal DataCube
MMVAE	Multi-Modal Variational AutoEncoder
MIPERS	Multi-Platform Earth Remote Sensing
NISAR	NASA-ISRO Synthetic Aperture Radar
NASA	National Aeronautics and Space Administration
NFI	National Forest Inventory
NLCD	National Land Cover Database
NLP	Natural Language Processing

NDSI	Normalized Difference Snow Index
NDVI	Normalized Difference Vegetation Index
NDWI	Normalized Difference Water Index
NICFI	Norway's International Climate and Forest Initiative
PASTIS	Panoptic Segmentation of Satellite image Time Series
PSNR	Peak Signal to Noise Ratio
PALSAR	Phased Array L-band Synthetic Aperture Radar
PSF	Point Spread Function
PE	Positional Encoder
PA	Producer Accuracy
RGB	Red Green Blue
RGBDEM	Red Green Blue Digital Elevation Model
ROI	Region of Interest
RDDB	Residual-in-Residual Dense Block
RMSE	Root Mean Square Error
RMSELR	Root Mean Square Error of Logarithmic Residuals
SITS	Satellite Image Time Series
SCL	Scene Classification Layer
SSTE	Sea Surface Temperature Estimation
DINO	self-DIstillation with NO labels
SSL	Self-Supervised Learning
SNAP	Sentinel Application Platform
SWIR	Short-wave infrared
SNR	Signal-to-Noise Ratio
SIM	Similarity Index Measure
SISR	Single Image Super-Resolution
SWI	Snow Water Index
SMAP	Soil Moisture Active Passive
SMOS	Soil Moisture and Ocean Salinity
STARFM	Spatial and Temporal Adaptive Reflectance Fusion Model
STF	Spatio-Temporal Fusion
SOTA	State-Of-The-Art
SGD	Stochastic Gradient Descent
SSIM	Structural Similarity Image Metric
SRCNN	Super-Resolution Convolutional Neural Network
SVM	Support Vector Machine
SVR	Support Vector Machine Regressor
SGKLD	Symmetric Gaussian Kullback–Leibler divergence
SAR	Synthetic Aperture Radar
TCW	Tasselled Cap Wetness
TAE	Temporal Attention Encoder
TAMRFSITS	Time Attention Multi-Resolution Fusion of Satellite Image Time Series
TSA	Time Series Analysis

TCD	Tree Cover Density
USGS	U.S. Geological Survey
USFS	United States Forest Service
UTM	Universal Transverse Mercator
UBARN	Unsupervised Bayesian Residual Network
UA	User Accuracy
VAE	Variational AutoEncoder
VLCC	Vegetated Land Cover Characteristics
VCT	Vegetation Change Tracker
VDSR	Very Deep Super-Resolution
VHR	Very High Resolution
WCM	Water Cloud Model
WP	Work package

Table of Content

Table of Tables....10
Table of Figures....14
Applicable Documents....28
Executive Summary....29
1 Introduction....31
1.1 About EvoLand....31
1.2 Purpose and objectives of WP2 Method development....31
1.3 Document structure....32
1.4 Link to WP1 requirements....33
1.5 Demonstration sites....33
2 Weakly supervised learning....34
2.1 State of the art....34
2.1.1 Introduction....34
2.1.2 Review of existing methods....35
2.1.3 Strengths and weakness of existing methods....41
2.2 Method description....42
2.2.1 Input data and study sites....43
2.2.2 Description of the candidate method....58
2.2.3 Benchmarking & evaluation....81
2.3 Implementation and results....89
2.3.1 SSL4EO-S12....89
2.3.2 DeCUR....109
2.3.3 SoftCon....113
2.3.4 CromSS....116
2.3.5 SpectralEarth....119
2.3.6 MMDC, pVAE, ALISE, MALICE, MALICE Aux....125
2.4 Summary and conclusions....131
3 Improved resolution....133
3.1 Single Image Super-Resolution of Sentinel-2 images....133
3.1.1 State of the art....133
3.1.2 Method description....140
3.1.3 Implementation and results....152

3.2 Fusion of Satellite Image Time Series including Thermal Infrared 165
3.2.1 Introduction 165
3.2.2 Proposed method 169
3.2.3 Experiments 178
3.3 Summary and conclusions 188
4 Continuous monitoring 190
4.1 Change detection – Time series modelling 190
4.1.1 State of the art 190
4.1.2 Method description 196
4.1.3 Implementation and results 200
4.1.4 Summary and conclusions 202
4.2 Disturbance classification 203
4.2.1 State of the art 203
4.2.2 Method description 204
4.2.3 Implementation and results 207
4.2.4 Summary and conclusions 211
4.3 Change detection – AI methods 211
4.3.1 State of the art 211
4.3.2 Method description 218
4.3.3 Implementation and results 223
4.3.4 Summary and conclusions 223
4.4 Continuous mapping of Land Surface Categories 224
4.4.1 State of the art 224
4.4.2 Method description 226
4.4.3 Implementation and results 231
4.4.4 Summary and conclusions 232
5 Biomass mapping 233
5.1 Introduction 233
5.2 Forest biomass 233
5.2.1 State of the art 233
5.2.2 Method description 238
5.2.3 Implementation and results 256
5.2.4 Validation and comparison with other results 274
5.2.5 Summary and conclusions 282

5.3 Grassland/cropland GPP biomass mapping 282
5.3.1 State of the art 282
5.3.2 An indicator based on vapour pressure deficit. Method description 285
5.3.3 Implementation and results 286
5.3.4 Summary and conclusions 288
6 Conclusion and outlook 289
References 290

Table of Tables

Table 1. Document structure. 32
Table 2. Summary of dataset utilized and assembled for EvoLand method development experiments on SSL. 43
Table 3. Summary of modalities included in SSL4EO-S12 v1.1. The range includes the typical value range, including the offset, but values below and above the range are present as well. 45
Table 4. Linear probing results for EuroSAT, BigEarthNet (BE) with 10% and 100% labels. We report overall accuracy for EuroSAT and mean average precision for BigEarthNet. 89
Table 5. Fine-tuning results for EuroSAT and BigEarthNet. 90
Table 6. Summary of benchmark results for forest semantic segmentation. 92
Table 7. ResNet-18 and Custom-ResNet-18 results. 96
Table 8. Forest benchmark. 97
Table 9. Comparison of forest-segmentation results from the first phase of the project (Custom-ResNet18) and the refined results from the second phase (Custom_ResNet18-WCE-Filtered). 105
Table 10. Benchmark of DeCUR on BigEarthNet-MM when compared to other models. 109
Table 11. Benchmark of DeCUR utilizing the Sentinel-1 encoder for the downstream classification task on BigEarthNet, only. 109
Table 12. DeCUR evaluated for semantic segmentation on the GeoNRW dataset. 110
Table 13. DeCUR evaluated for semantic segmentation on the GeoNRW dataset, but without feeding digital elevation (DEM) information when compared to Table 12. 110
Table 14. Linear probing results on BigEarthNet-SAR-10% and EuroSAT-SAR. 114
Table 15. Linear probing/fine-tuning accuracy on fMoW-sentinel-10%. 115
Table 16. Fine-tuning mIoU results on DFC2020. 115
Table 17. Transfer learning results (%) on the DFC2020 dataset from PSPNet with ResNet-50 backbones and UPerNet with ViT-Large backbones. 3B/9B/10B/13B in brackets indicate the number of S2 bands as inputs for each model. Frozen and

fine-tuned represent no updates of encoders and optimizing encoders along with decoders in the transfer learning setting. The best results are highlighted in bold. The annotations are the same in the following tables. 116
Table 18. Transfer learning results (%) on the SSL4EO-S12@DW dataset from DeepLabV3+ with ResNet-50 and UPerNet with ViT-Large. 117
Table 19. Performance of SpectralEarth hyperspectral model on 7 downstream tasks. Best performance per model architecture is highlighted in bold, and the second best is underlined 119
Table 20. Performance of SpectralEarth hyperspectral ViT models pre-trained with MAE on 7 downstream tasks. Best performance per model architecture is highlighted in bold, and the second best is underlined. 121
Table 21. Cross-sensor transferability of the Spec. ViT-L model pre-trained on SpectralEarth. 122
Table 22. Quality scores of reconstructing EnMAP from EnMAP. 125
Table 23. Quality scores of reconstructing EnMAP from Sentinel-2 125
Table 24. Regression performance with raw image input and MMDC embeddings. 127
Table 25. Performance of different embedding inputs on crop classification task. For S1, A/D denote ascending and descending orbits. Experiments were conducted using 5 folds; however, if an experiment was performed only with 1-fold, the results are labelled with "(1f)" 129
Table 26. TCD regression performance with different embedding inputs. 130
Table 27. Tree Height Regression performance with MALICE S1 ASC embeddings. 131
Table 28. BioMass Regression performance with MALICE S1 ASC embeddings 131
Table 29. Corresponding Sentinel-2 and VENμS bands used in the SEN2VENμS dataset. Note that B11 is used as high resolution band for both B8 and B8A Sentinel-2 bands 140
Table 30. Country and province of each of the selected VENμS sites. The usage column indicates if the site is used for model training and validation, or for testing (model performances evaluation). 143
Table 31. Models investigated in the study. Naming of configuration for carn is nb_cascading_blocks x nb_eres_blocks_per_cascading_blocks x nb_features. g1

indicates that grouping is set to 1 in convolution layers, and sw indicates if weights are shared between ERes blocks of a given cascading block. 153
Table 32. Order of magnitude of inference time of the carn_3x3x64g4sw model on full Sentinel-2 products for different processing configurations. 165
Table 33. Capabilities and limitations of the different categories of methods found in solving the general problem, (x) means that some methods in the literature partially achieve the capability or solve the limitation or weak assumption. 169
Table 34. Pool of strategies for MAE training. 175
Table 35. Number of 9.9x9.9 km2 images in LS2S2 dataset for each sensor and each split. 178
Table 36. Spectral bands included in the LS2S2 dataset for each sensor. The table layout allows to highlight which bands are similar between sensors and which bands are only available on one sensor. Despite the variable spatial resolution across bands for both Landsat and Sentinel-2, for each sensor all bands are up-sampled to the maximum resolution (10 meter for Sentinel-2 and 30-meter for Landsat) 179
Table 37. Landsat prediction metrics for TAMRFISTS when applying a random masking strategy of 50% of the dates, yielding 365 clear Landsat dates and 389 masked Landsat dates. RMSE is expressed in surface reflectances except for LST, where it is expressed in Kelvins. The 50% column shows median value across all dates, while the 95% shows the 95% quantile across all dates. 184
Table 38. Sentinel2 prediction metrics for TAMRFISTS when applying a random masking strategy of 50% of the dates, yielding 365 clear Landsat dates and 389 masked Landsat dates. RMSE is expressed in surface reflectances (unitless). The 50% column shows median value across all dates, while the 95% shows the 95% quantile across all dates. 186
Table 39. Median values of RMSE, BRISQUE and FR compared between TAMRFSITS and DeepHarmo [Sambandham et al., 2024], over the full testing set, for 176 same-day acquisitions where Sentinel-2 images have been kept aside for validation. Best values for each metric are highlighted in bold 188
Table 40. Sentinel-2 and derived input data sets. 198
Table 41. demonstration cites and their respective undisturbed and disturbed areas. 200

Table 42. Study sites and their respective disturbance event. 200
Table 43. Accuracy assessment. 201
Table 44. F1 scores for each model combination. 207
Table 45. Overview on data utilized for change detection experiments with GAF. 218
Table 46. input datasets. 226
Table 47. Overview of the LSC legend classes. 227
Table 48. Translation of the WorldCover labels into LSC labels. 228
Table 49. Overview of the benchmarked models. 230
Table 50. f1-score of the LSC classes and overall accuracy (OA) for each of the models. 230
Table 51. Input datasets. 239
Table 52. Input datasets used in the boreal study site. 244
Table 53. Filtering criteria for GEDI data. 246
Table 54. Bin size and maximum number of plots per bin for data balancing for different GEDI datasets. 247
Table 55. Benchmarking of different sensor combinations for the random forest based estimation of AGB in the boreal test site. 259
Table 56. Benchmarking summary table with validation results of forest canopy height. 262
Table 57. RMSE and Pearson correlation coefficient values for FCH 2020 for different combinations of UNet model type and training dataset processing. 268
Table 58. RMSE and Pearson correlation coefficient values for AGBD 2020 for different combinations of UNet model type and training dataset processing. 268
Table 59. Performance indicators for the 2020 and 2023 FCH and AGBD UNet models. 269
Table 60. Validation summary table for above ground woody biomass density for the boreal test site. 277
Table 61. Validation summary for forest canopy height for the boreal test site. 280
Table 62. Overview of different parameters used in the current Monteith approach of CLMS GPP. 284
Table 63. List of currently used datasets for the method development. 285

Table of Figures

Figure 1. EvoLand test sites. 33
Figure 2. Schematic workflow of Weakly-Supervised Learning with Self-Supervision pre-training and downstream task adaptation. 35
Figure 3. A taxonomy of self-supervised learning. 35
Figure 4. Example pretext tasks from the self-supervised learning literature. 36
Figure 5. An illustration of Masked Auto Encoders (He et al., 2022) 38
Figure 6. A comparison of existing contrastive methods image: X, Encoder: E, student: E', teacher: E". 39
Figure 7. Illustration of the two backbone architectures employed for AI models in EvoLand – left: a ResNet block with skip connection, right: a ViT for image classification. 40
Figure 8. Sample visualization of one patch from SSL4EO-S12 dataset. The rows top-down present grayscale and false-colour imagery based on the Sentinel-1 GRD product, Sentinel-2 level-1C, and Sentinel-2 level-2A multispectral data with corresponding columns representing the four seasons spring, summer, fall, and winter from left to right. 44
Figure 9. Geographical distribution of SSL4EO-S12 dataset. 45
Figure 10. Geospatial sampling distribution of SSL4EO-EU-Forest (left) from which we separated the samples of the EvoLand test sites (right). 46
Figure 11. Geospatial sampling distribution of SSL4EO-EU-Forest-V2 dataset 47
Figure 12. Geographical Distribution of the SpectralEarth dataset. 48
Figure 13. Geographical Distribution of the SpectralEarth dataset. 49
Figure 14. Illustration of various downstream tasks for SpectralEarth, from multi-label image classification to pixel-level semantic segmentation. 51
Figure 15. Samples from the SpectralEarth-MM dataset displaying EnMAP/S2/L8/S1 images for the same areas. 53
Figure 16. Tiles selected to train MMDC model. Note that we use only blue tiles present on this figure. 54
Figure 17. The general structure of DeCUR: M1 and M2 represent two modalities from which an embedding feature vector Z is distilled. For EvoLand, Z identifies

common and unique features in pairs of spatio-temporally aligned remote sensing imagery for Sentinel-1 & -2. 60
Figure 18. The general framework of SoftCon. We load vision foundation models for both the trainable and the momentum encoder when initializing the model. EMA represents exponential moving average. Siamese masking is only used with ViT backbones. 62
Figure 19. Two losses drive the optimization of a CromSS model: A semantic segmentation loss (top row) and a cross-modal consistency loss (bottom row, separated by a dashed line). Both are generated by three logical steps: pixel-wise confidence mask generation, cross-modal confidence enhancement, and thresholding for pixel weighting. 63
Figure 20. Comparison of single-modal training and multi-modal (middle/late) fusion strategies. 64
Figure 21. Illustration of SpectralEarth model architectures adapting ResNet and ViT to hyperspectral data input through a Spectral Adapter. 66
Figure 22. Schematic overview of the pre-training/fine-tuning pipeline. An encoder is pre-trained on SpectralEarth and adapted to the several downstream tasks with task-specific heads. 67
Figure 23. Visualization of masking strategies for spatial-spectral data cubes with C channels and WxH spatial pixels (left). From left to right most: (a) all channels of a given spatial pixel get masked out, (b) all spatial pixels of a (hyper)spectral channel gets masked, and (c) a (random) mix of spatial pixels get masked for each channel 67
Figure 24. Masked Autoencoder (MAE) with Vision Transformer (ViT) for hyperspectral imagery as utilized in our approach. The hyperspectral data cube gets patchified into spatial-spectral data cubes that get spectrally masked for reconstruction. 68
Figure 25. Encoder-decoder model. The input image is passed to the Encoder to produce the embeddings, which are then passer to the decoder to reconstruct the input image. 69
Figure 26. MMDC baseline. The model consists of two modality-specific encoders, as well as an encoder for weather data. Each modality-specific encoder first encodes the angles and DEM model, then this encoded data is concatenated with

encoded weather data and image data and passed to the final encoding model to extract a modality-specific embedding. The decoding part consists of two modality-specific decoders, which use the produced embeddings to reconstruct the input data, in mono-modal and cross-modal way. 70
Figure 27. MMDC experts' model. The encoder is designed in the same way as in the baseline model and produces two modality-specific embeddings that are then fused using a chosen expert strategy. The decoder part, as in the baseline model, is composed of two modality-specific decoder which reconstructs the input data from the expert embeddings. 71
Figure 28. MALICE encoder architecture. MALICE is composed of two branches each associated to one modality. The branches have different SSTE, but they share the same temporal projector. 76
Figure 29. MALICE pre-training overview. Discriminative losses as well as reconstruction losses are considered for pre-training. 77
Figure 30. Description of the reconstruction SSL losses employed to pre-train MALICE. Reconstruction losses are either mono-modal or cross-modal. The metadecoder performs a cross-attention between the latent representations Y (used as keys and values) and the queries Q which contain the information about the modality as well as the dates to be reconstructed. Superscripts refer to the modality S1 or S2 and subscripts to the sub-series A, B. 77
Figure 31. SSTE Aux module. Each module is modality-specific, except for the DEM encoder that is shared by the modalities. In red we show the newly added data and operations, compared to the original SSTE module. We start by 1/encoding the AGERA5 data. Then 2/we concatenate it with the SITS data that is passed to the SS module. 3/The positional encoding of the acquisition dates is done. 4/DEM features are encoded by a DEM UNet module. Finally, 5/those features are concatenated in temporal dimension with the sum of encoded SITS and PE, and passed to the temporal encoder. 80
Figure 32. Sample of reconstructed spectrum of a single spatial pixel. While gray bands mark masked bands (e.g., not covered by Sentinel-2), the Band Index labels correspond to EnMAP channels. 85
Figure 33. UTAE model (taken from (Garnot & Landrieu, 2021)). 88

Figure 34. Visualization of four geolocated image-patch-level SSL4EO-S12 embeddings (red dots) with two out of the four Sentinel-2 input images, background map: OpenStreetMap.....91
Figure 35. Comparison between SSL and Supervised Learning. From left to right: Sentinel-2, HRL forest mask (ground truth), Supervised Learning prediction, SSL & fine-tuning (UNet with ResNet-18 backbones for the encoder and decoder).....93
Figure 36. Comparison of pre-trained backbones with a UNet model. From left to right: Sentinel-2, HRL forest mask (ground truth), ResNet-18 prediction, ResNet-50 prediction, ViT-S prediction.....94
Figure 37. Two illustrative examples where the 2018 HRL forest masks (right column) do not well match the ground truth Sentinel-2 imagery for a random season that same year (left column): Forest seem to have converted into agriculture land (top) and forests might have undergone clearance by logging, etc.....95
Figure 38. From left to right: Sentinel-2, HRL forest mask (ground truth), ResNet-18 prediction, Custom-ResNet-18 prediction.....97
Figure 39. Illustration of forest prediction performance with FCN-ResNet-18. From left to right: Sentinel-2, HRL forest mask (ground truth), FCN-ResNet-18 prediction All layers, FCN-ResNet-18 prediction 3, 2, and 1 layer(s).....98
Figure 40. Visual demonstration of the single-pixel convolution model ResNet-18/1. From left to right: Sentinel-2, HRL ground truth, ResNet-18/1 prediction.....99
Figure 41. Visual demonstration of the single-pixel convolution model ResNet-18-Segment. From left to right: Sentinel-2, HRL ground truth, ResNet-18-Segment prediction.....100
Figure 42. Samples of cloudy images detected by the deep learning cloud segmentation model.....101
Figure 43. Samples illustrating the weighted cross entropy loss – Left: ground truth mask and Right: corresponding weight mask.....102
Figure 44. Algorithm used to filter out noisy samples from the SSL4EO-EU-Forests dataset.....103
Figure 45. Noisy samples filtered out from SSL4EO-EU-Forests - Left: S2 image, Middle: target mask and Right: model prediction.....105
Figure 46. Samples demonstrating our model's predictions– Left: Sentinel 2 patch, Middle: model prediction and Right: HRL mask.....107

Figure 47. Qualitative results for DLT classification – Left: S2 patch, middle: HRL 2018 mask, and right: model predictions. 108
Figure 48. Analysis of embedding alignments for DeCUR (red) vs. Barlow Twins (blue). 111
Figure 49. t-SNE analysis of unique (blue/green) vs. common (red) representations for Sentinel-1 & -2 joint embeddings generated by DeCUR. 112
Figure 50. GradCAM analysis of common and unique embeddings with DeCUR (top). Integrated Gradient analysis of Sentinel-2 spectral bands (bottom). 113
Figure 51. A detailed comparison of transfer learning performances on BigEarthNet-10%. S/B/L represents ViT-small/base/large. SoftCon (ours) achieves SOTA results with lighter backbones on both linear probing and fine-tuning, and both multispectral and SAR. 114
Figure 52. The segmentation maps obtained on the SSL4EO-S12@DW downstream dataset with S2 as inputs from UperNet with ViT-large pretrained by scaleMAE and satMAE, and the rest from PSPNets with ResNet-50 pretrained by different methods (upper: with frozen encoders, lower: with fine-tuned encoders). 118
Figure 53. Evaluation of SpectralEarth pre-trained model performance on downstream tasks when compared to training the model from scratch. 122
Figure 54. Performance of Spec. ResNet-50 on downstream tasks based on the fine-tuning setup. None: linear probing, Adapter: fine-tuning of Spectral Adapter (cf. Figure 13), Layer 1: fine-tune Spectral Adapter and 1st layer of ResNet-50 backbone, Layer 2: fine-tune Spectral Adapter, 1st and 2nd layers of ResNet-50 backbone, All: full fine-tuning. 123
Figure 55. Impact of the size of the downstream dataset evaluated on EnMAP-CORINE and EnMAP-NLCD for the Spec. ViT-L model. 124
Figure 56. Impact of the size of the downstream dataset evaluated on EnMAP-CORINE and EnMAP-NLCD for the Spec. ViT-L model. 124
Figure 57. Latent spaces obtained with MMDC baseline. First row: the original input data. Second row: data reconstructed from its proper latent space (S1 and S2 images reconstructed from S1 and S2 embeddings respectively). Third row: cross-reconstructed data (S1 and S2 images reconstructed from S2 and S1 embeddings respectively). Fourth and Sixth rows – mean and logvar of S1 embeddings. Fifth and Seventh rows – mean and logvar of S2 embeddings. 126

Figure 58. Latent spaces obtained with MMDC baseline. First row: the original input data. Second row: data reconstructed from its proper latent space (S1 and S2 images reconstructed from S1 and S2 embeddings respectively). Third row: cross-reconstructed data (S1 and S2 images reconstructed from S2 and S1 embeddings respectively). Fourth and Sixth rows – mean and logvar of S1 embeddings. Fifth and Seventh rows – mean and logvar of S2 embeddings. 127
Figure 59. Scatterplot of predicted LAI values vs ground truth values..... 128
Figure 60. LAI regression predictions with MMDC embeddings inputs. Note that in "Pred 0 2" 0 and 2 stand for minimum and maximum normalization values for the visualization of the results. 129
Figure 61. Scatterplot of predicted TCD values vs ground truth values with MALICE Aux embeddings. We do not present scatterplots for other embeddings, as they do not visually show any notable difference. 130
Figure 62. Scatterplot of predicted values vs ground truth values with MALICE S1 embeddings. 131
Figure 63. Image Restoration versus Image Super-Resolution..... 134
Figure 64. SRCNN network architecture (figure from [Dong et al., 2015]). 135
Figure 65. Understanding the PixelShuffle operation (figure from [Qin et al., 2020]). 136
Figure 66. Residual and cascading blocks of CARN architecture (figure from [Ahn et al., 2018]). 137
Figure 67. Overview of CARN architecture (figure from [Ahn et al., 2018]). 137
Figure 68. Spectral Sensitivity Response of Sentinel-2 and Venμs sensors..... 141
Figure 69. Location of the 29 sites composing the Sen2Venμs dataset. 141
Figure 70. Examples of patches. From left to right: columns 1-8 and 9-16 show rendering of two different patches. Columns 1 and 9: B4, B3, B7 (RGB natural) at 10m, columns 2 and 10: B4, B3, B7 (RGB natural) at 5m, columns 3 and 11: colour-mapped B8 at 10m (Wide Near Infra-Red), columns 4 and 12 : colour-mapped B8 at 5m (Wide Near Infra-Red), columns 5 and 13 : B7, B6, B5 colour composition (Red Edge 3 to 1) at 20m, columns 6 and 14: B7, B6, B5 colour composition (Red Edge 3 to 1) at 5m, columns 7 and 15: colour-mapped B8A at 20m (Narrow Near Infra-Red), columns 8 and 16: colour-mapped B8A at 5m (Narrow Near Infra-Red). 29 patches are displayed, one random patch for each site. Only 64 × 64 pixels crops of the

patches are displayed to improve readability. High resolution and low resolution patches radiometry was scaled to 8 bits with the same scaling factors. 142
Figure 71. Achieved PSNR with respect to model complexity (figures from [Anwar et al., 2020]) (a) PSNR vs. number of mult-add operations (b) PSNR vs. number of parameters. 144
Figure 72. Workflow of the joint 10m-20m SISR model, using the CARN backbone. Tensor shapes are indicated, where b stands for the batch dimension. 145
Figure 73. Example of 25% Low Gradient Magnitude Strata (blue) and 25% High Gradient 150
Figure 74. From left to right : FFT of bicubic up-sampled Sentinel-2 image, FFT of superresolved image, difference between both FFT (red means high FFT on SISR), for band B7. 152
Figure 75. PSNR achieved by the carn_3x3x64g4sw trained with different training losses, during inference on real input data over the testing set. 155
Figure 76. RGB Rendering of predictions from the carn_3x3x64g4sw trained with different training losses, on real patches from the testing set. 155
Figure 77. LR RMSE on real data for the carn_3x3x64g4sw trained with different losses functions. 156
Figure 78. Red Edges bands (B7,B6,B5) rendering of predictions from the carn_3x3x64g4sw trained with different training losses, on real patches from the testing set. 156
Figure 79. NIR and SWIR bands (B8,B9,B10) rendering of predictions from the carn_3x3x64g4sw trained with different training losses, on real patches from the testing set. 157
Figure 80. HR RMSE on real images for the carn_3x3x64g4sw trained with different losses functions. 157
Figure 81. HR RMSE on simulated images for the carn_3x3x64g4sw trained with different losses functions. 157
Figure 82. HR RMSE metric on simulated data for different CARN parameters. 158
Figure 83. BRISQUE score on simulated data for different CARN parameters. 158
Figure 84. RGB Rendering of prediction starting from carn_3x3x64g4sw, and disabling shared weights and using a grouping of 1, on real patches from the testing set. 159

Figure 85. HR RMSE metric on simulated data, for varying model depths.....160
Figure 86. BRISQUE score on real data, for varying model depths.....160
Figure 87. HF power variation with respect to bicubic on real data, for varying model depths. 161
Figure 88. TV variation with respect to target on real data, for varying model depths. 161
Figure 89. RGB Rendering of predictions of models with different depths, on real patches from the testing set.....161
Figure 90. BRISQUE score on real data, for pretrained vs. fine-tuned models.....162
Figure 91. HF power variation with respect to bicubic on real data, for pretrained vs. finetuned models.....163
Figure 92. Total Variation with respect to target on real data, for pretrained vs. fine-tuned models. 163
Figure 93. Red Edges bands (B7,B6,B5) rendering of predictions of fine-tuned vs. pre-trained models, form carn_1x3x64g4sw and carn_2x3x64g4sw, on real patches from the testing set. 164
Figure 94. Red Edges bands (B7,B6,B5) rendering of predictions of fine-tuned vs. pre-trained models, form carn_3x3x64g4sw with and without HPF in discriminator, on real patches from the testing set. 164
Figure 95. Red Edges bands (B7, B6, B5) rendering of predictions of fine-tuned vs. pre-trained models, form carn_4x3x64g4sw and esrgan_3x64x32, on real patches from the testing set. 165
Figure 96. Overview of capabilities of methods from the literature, as compared to ours. The LR SITS is represented in green, while the HR SITS is represented in blue. 166
Figure 97. Overview of the proposed method.SLR (in green) and (in blue) are fed into the TAMRFSITS encoder, which outputs latent representation $Slatent$ (in red). The latent representation is fed into the TAMRFSITS decoder, along with $Tquery$, which outputs $SLR\star$ (in green) and $SHR\star$ (in red).....171
Figure 98. Details of TAMRFSITS encoder. The input SITS SLR (in green) and (in blue) are first passed through separate date-wise spatial encoders based on the RRDB architecture, including spatial up-sampling of SLR. The resulting series $SLRlatent$ and $SHRlatent$ have matching dimensions, except for the temporal dimension. After

being informed by temporal positional encoding and a learnable sensor token, each pixel is converted to a sequence of tokens imputed to a pixel-wise transformer encoder, forming the latent SITS $Slatent$. 172
Figure 99. Details of TAMRFSITS decoder. The query acquisition times $Tquery$ are transformed into a sequence of tokens by means of temporal positional encoding and a learnable placeholder. The query sequence and the latent SITS $Slatent$ are processed by a pixel-wise temporal decoder, starting with a cross-attention block, which interpolates $Slatent$ at query acquisition times. The resulting SITS is then passed through a simple spatial decoder to predict the resulting SITS $SLR*$ and $SHR*$ 174
Figure 100. Geographical coverage of the LS2S2 dataset, with training AOIs in green and testing AOIs in red. 180
Figure 101. Landsat 8 and 9 time series. 180
Figure 102. Example of SITS from LS2S2 dataset training slice, for UTM tile 31UFR. Pixels masked according to the quality mask included in the product are highlighted in red. 181
Figure 103. Illustration of the TAMRFSITS model capabilities over Area Of Interest 31TFJ for year 2022 (from the test set). Landsat SITS has 51 dates, of which 26 dates are removed from the model input and kept for validation. Sentinel-2 has 81 dates, of which 41 are removed from the model input and kept for validation. In this figure, images that are part of the model inputs are highlighted in green, model predictions are highlighted in red, and reference images are highlighted in blue. From top to bottom: rows 1 and 2 shows prediction for an observed Landsat date, row 3 shows prediction for a cloudy Landsat date, row 4 shows prediction for an observed Sentinel-2 date, row 5 shows prediction for a cloudy Sentinel-2 date, and row 6 shows predictions for an unobserved date. 185
Figure 104. Distribution of average RMSE of the output of TAMRFSITS, over 10 meter, 3022 meter and 90 meter bands. 188
Figure 105. Benchmarking study sites and their locations within the demonstration sites. 197
Figure 106. Number of available clear-sky observations for the year 2019. 197
Figure 107. Proposed forest change detection methodology. 202

Figure 108. Spatial test and training data distribution (left) and number of unique labelled events per class (right). 205
Figure 109. F1 scores for all classes and each model and feature combination. 207
Figure 110. Confusion matrices for the best performing models for the two different input feature types. 208
Figure 111. Examples for inference using best performing LightGBM model and spectral input. 209
Figure 112. Examples for inference using best performing SVM model and embeddings input. 210
Figure 113. Examples for inference using Gaussian Mixture model and spectral input. 210
Figure 114. Typical workflow of change detection as depicted in (Shafique et al., 2022) 212
Figure 115. Change Detection fusion strategies as presented in (Cheng et al., 2023b). 216
Figure 116. General Artificial Neural Network design of the SSL4Change methodology separately convolving two remote sensing modalities (SAR, Optical) to a common embedding ($Conv5^{common}$). Our approach adapts the architecture for two Optical Sentinel-2 input at timestamps T1 and T2, i.e. “SAR input” is replaced by a 2^{nd} “Optical input”. 219
Figure 117. Sentinel-2 inputs (before, T1 and after, T2) to SSL4change and corresponding change maps for an area in mountain range Taunus west of the German city of Wiesbaden. 220
Figure 118. Sentinel-2 sample snapshots of a forest area in Germany (left: before, right: after) as demonstrator of our unsupervised change detection methodology. The forest that has been cleared from left to right is enclosed by white borders close to the image centre. 221
Figure 119. Detection of change (red) on demo scene based on contrastive unsupervised AI model SSL4ForestChange depending on threshold hyperparameter C. The background is filled by the Sentinel-2 data after the change. Known areas without forest are visually blocked in grey. 222
Figure 120. Example of annotations generated over a patch located in a cropland area. 229

Figure 121. Example of LSC predictions over an area in Spain, with fluctuations in herbaceous cover, and minor changes in water level. 231
Figure 122. Example of LSC predictions over a cropland area, where harvest, crop growth, and planting/sowing activities cause changes between bare and herbaceous cover over time. 231
Figure 123. Mediterranean study site used for testing of non-parametric models. 239
Figure 124. The filtered HV backscatter of 2020 over the 16 PALSAR-2 tiles of the Mediterranean prototype site. 239
Figure 125. GEDI waveform. A near-infrared pulse of laser energy is fired towards the surface where it is reflected by leaves and branches within a nominal 25 m diameter footprint (shown at right). The returned waveform (shown on left) is processed to find ground topography, canopy height, and various relative height (RH) metrics. From these metrics a variety of other products are derived, including leaf density profile, canopy cover and aboveground biomass (Dubayah et al., 2020). 240
Figure 126. Footprint-level relationships between predicted and observed AGBD. The predictions are from generalized linear models developed using waveform relative height metrics. (A) Evergreen Broadleaf Trees in Australia; (B) Deciduous Broadleaf Trees in North America. The percentage RMSE is 30.7%, and 38.2%, respectively (Dubayah et al., 2020). 241
Figure 127. Effect of the multi-temporal filter: unfiltered (left) vs filtered (right) image. 242
Figure 128, Extended study area used for training of the non-parametric models. Sentinel-2 NIR band overlaid with the location of the NFI plots in blue. In the center, the location of the demosite is drawn in Sentinel-2 RGB band combination, which cover 4 Sentinel-2 granules. 243
Figure 129. Sub-test site which is covered by LIDAR data to a large extend, which was used for non-parametric model testing. 244
Figure 130. The filtered HV backscatter of 2020 over the 8 PALSAR-2 tiles of the boreal prototype site. 244
Figure 131. French Guiana border and processing extent (red). Processing is split up into overlapping tiles (grey). 245

Figure 132. The two available AGB maps derived from Lidar flights during the TropiSAR campaign. 245
Figure 133. 1x1 km testing and training grid with French Guiana outline. 247
Figure 134. AGBD training data distribution for unbalanced data (blue), moderately balanced data (orange) and completely balanced data (green) 248
Figure 135. HV backscatter of 2020 for an example site at the border of French Guiana. 249
Figure 136. RGB Sentinel-2 image of 2020 for an example site at the border of French Guiana. 249
Figure 137. RGB Planet image of 2020 for an example site at the border of French Guiana. 250
Figure 138. VV backscatter of 2020 for an example site at the border of French Guiana. 250
Figure 139. Forest mask based on CLCplus Backbone 2021 reclassification for an example site at the border of French Guiana. Forest areas are green, non-forest areas are grey. The large grey areas are caused by their location outside the French Guiana border 250
Figure 140. Slope (on top) and Aspect (on bottom) derived from the Copernicus DEM for an example site at the border of French Guiana. 251
Figure 141. Calibration of the Water Cloud Model for HH (left) and HV (right) polarisations over the Sweden (top), Spain (middle) and French Guiana (bottom) sites with CCI AGB as a refence. The red curve is the regression curve obtained from the fitting. The pink area represents the standard deviation of the backscatter as modelled by MIPERS. 255
Figure 142. Calibration of the Water Cloud Model for HH (left) and HV (right) polarisations over the Sweden sites with the NFI as a reference. The red curve is the regression curve obtained from the fitting. 255
Figure 143. Boreal site, feature importance for estimation of AGB for the year 2022 based on Sentinel-2 optical bands, medoid from 1st June to 15th September 2022, the median of Sentinel1-VH and VV monthly mosaics from May to September 2022, and height and slope derived from the Copernicus-DSM. 257
Figure 144. Boreal site, estimation of AGB for the year 2020 (left) and 2022 (right). The scattergrams show the testing plots, which were not used in the training nor

in the calibration. The AGBD estimates are drawn in blue, and the estimation results after calibration are drawn in orange. 257
Figure 145. Selected RF model for estimating AGB in the Spain test site, based on Sentinel-2 optical features. 258
Figure 146. Boreal site, feature importance for estimation of FCH for the year 2022 based on Sentinel-2 optical bands, medoid from 1st June to 15th September 2022, the median of Sentinel1-VH and VV monthly mosaics from May to September 2022, the ALOS-PALSAR annual mosaics, and height and slope derived from the Copernicus-DSM. 260
Figure 147. Boreal site, estimation of FCH for the year 2020 (left) and 2022 (right). The scattergrams show the testing plots, which were not used in the training nor in the calibration. The forest canopy height estimates are drawn in blue, and the estimation results after calibration are drawn in orange. 260
Figure 148. Selected RF model for estimating FCH in the Spain test site, based on Sentinel-2 optical features. 262
Figure 149. Result of the Forest Canopy Height estimation for the test site in Sweden. 263
Figure 150. Detail zoom-in of the Forest Canopy Height estimation within the test site in Sweden (the crosses and labels are from NFI Sweden). 264
Figure 151. Result of the Above Ground Biomass estimation for the test site in Sweden. 264
Figure 152. Detail zoom-in of the Above Ground Biomass estimation within the test site in Sweden (the crosses and labels are from NFI Sweden). 265
Figure 153. Result of the Forest Canopy Height estimation for the test site in Spain. 266
Figure 154. Result of the Above Ground Biomass estimation for the test site in Spain. 267
Figure 155. Testing Plots for all four trained UNet models (2020 and 2023, FCH and AGBD). Note that units for FCH are cm. 269
Figure 156. Example (52.91°W 5.29°N) map of AGBD for 2020 (left) and 2023 (right). 270
Figure 157. Detailed example (52.95°W 5.28°N) map of AGBD for 2020 (left) and 2023 (right). 271

Figure 158. Comparison of TropiSAR Lidar data (2009) and UNet (2020 and 2023) maps at the Paracou site. 271
Figure 159. Example (52.91°W 5.29°N) map of FCH for 2020 (left) and 2023 (right). 272
Figure 160. AGB map of the boreal site from the parametric approach for year 2020, calibration with CCI AGB 5.0. 272
Figure 161. AGB map of the boreal site from the parametric approach for year 2020, calibration with NFI. 273
Figure 162. AGB map of the Mediterranean site from the parametric approach for year 2020, calibration with CCI AGB 5.0. 273
Figure 163. AGB map of the tropical site from the parametric approach for year 2020, calibration with CCI AGB 5.0. 274
Figure 164. Validation result: AGB 2022 (above) and 2020 (below) for the Sweden demonstration site. Approach by JR with random forest estimation and calibration based on temporary NFI plots. The validation is based on the permanent NFI plots. 275
Figure 165. Validation result: AGB for the year 2020, above with calibration based on NFI temporary plots, and below with calibration based on CCI data, for the Sweden demonstration site, based on permanent NFI plots. Approach by UT3, parametric estimation with PALSAR-2. 276
Figure 166. Validation result: AGB for the year 2020, CCI AGBD map 2020, with spatial resolution of 100m per pixel. The statistics are for information purposes only, as the observation units differ. The permanent NFI plots represent a diameter of 20m and the CCI map has a pixel resolution of 100m by 100m. 277
Figure 167. Qualitative comparison of AGB. 278
Figure 168. Validation result: FCH 2022 (above) and 2020 (below), for the Sweden site. Approach by JR with random forest estimation and calibration based on temporary NFI plots. The validation is based on the permanent NFI plots. 279
Figure 169. Validation result: Global Canopy Height 2020 dataset (Lang, Jetz, & Wegner, 2023), for the Sweden demonstration site, based on permanent NFI plots. 280
Figure 170. Qualitative comparison forest canopy height. 281

Figure 171. Left: GPP target compared to prediction for the optimized LUE for cropland. Right: GPP target compared to prediction for the optimized LUE for grassland. 287

Applicable Documents

ID	Applicable document
[AD1]	Grant Agreement
[AD2]	D1.2 Service requirements report - final version
[AD3]	D2.19 Test datasets (Weakly Supervised Learning, Improved Spatial/Temporal/Spectral resolution, Continuous monitoring, Biomass mapping - final version
[AD4]	D2.21 Source code (Weakly Supervised Learning, Improved Spatial/Temporal/Spectral resolution, Continuous monitoring, Biomass mapping - final version

Executive Summary

The Methods compendium – final version summarizes the results of the EvoLand project within Work Package (WP) 2 – Method Development. The goal of this WP is to develop the algorithms and methods required for the candidate CLMS services in WP3. The document provides a comprehensive summary on the methods in i) weakly supervised learning (task 2.4), ii) improved resolution (task 2.5), iii) continuous monitoring (task 2.6), and iv) biomass mapping (task 2.7). For each of these methods, first, the state of the art is described, followed by a description of the developed methods, the implementation, and results.

Within task 2.4, we introduced and benchmarked a suite of innovative **weakly-supervised deep learning methods** for Sentinel-2, Sentinel-1 & -2 data fusion along with the development of a large-scale hyperspectral foundation model using EnMAP data. Key contributions include the release of curated datasets such as SSL4EO-S12 and its EU-Forest extension, designed for evaluating tasks like semantic forest segmentation. Several new methodologies were introduced such as SoftCon (Enhances contrastive self-supervised learning for Sentinel-2 using existing landcover annotations); DeCUR (Enables explainable multi-modal SSL by fusing Sentinel-1/2 through modality-specific embeddings); CromSS (Uses weak supervision and confidence validation across modalities to improve segmentation quality); SpectralEarth (Establishes a hyperspectral foundation model with an energy-efficient fine-tuning strategy based on EnMAP data); MMDC (pixel-wise embeddings preserving spatial resolution for future evaluation across multiple downstream tasks); and MALICE and MALICE Auxiliary (Spatio-Temporal Sentinel-1 and Sentinel-2 models)

Next, within task 2.5 we focussed on **improving the resolution**. First by developing a lightweight Single Image Super-Resolution (SISR) tool, enhancing Sentinel-2 bands to 5-meter resolution. Given the need for efficiency in processing large volumes of data, relatively shallow networks were favoured over those found in existing literature. The trained model demonstrates strong SISR performance with minimal spectral distortion, making SISR products viable substitutes for traditional Sentinel-2 images in downstream applications without radiometry parameter adjustments. To accommodate the specificities of the Sen2Venμs dataset, a tailored training methodology was proposed, addressing input resolution discrepancies, multi-sensor dataset challenges, and the absence of SWIR bands in the target data. Besides the single image Super Resolution, task 2.5 also focussed on the fusion of multi-modal Satellite Image Time Series and application to Thermal Infra-Red. A new problem formulation has been proposed that allows to handle both temporal interpolation and super-resolution in a unified framework. This formulation led to a new Deep-Learning architecture and training strategy, that allows to achieve good performances on the fusion task with respect to the literature, and can tackle several fusion tasks (sensor translation, cloud interpolation, super-resolution, temporal extrapolation) where models from the literature are dedicated to a single task

Within task 2.6 EvoLand addressed key challenges in **continuous monitoring** by developing methods for i) change detection - time series modelling and forest disturbance detection; AI methods for change detection; and iii) Continuous Mapping of Land Surface Categories. Regarding the forest disturbance detection and classification, several ML models were developed and compared based on spectral features and geospatial foundation model embeddings. Regarding the AI methods for change detection, SSL4Change was developed which offers an unsupervised, multi-modal, contrastive change detection method that is cost-effective and executable on standard hardware. Initial experiments on forest change detection however revealed the model's sensitivity to training data and thresholding choices for binary change maps. It was determined that SSL4ForestChange may not be the most promising approach for production-ready forest change detection. Urban areas show change detection accuracies unlikely to exceed 50%.Despite limitations, the contrastive change product may serve as a proxy for forest change likelihood when combined with ground truth masks. Additionally, for continuous land surface category (LSC) mapping, early models showed that incorporating temporal statistics improved vegetation classification but reduced sensitivity to change. In the later phase, a hybrid deep learning model combining U-Net and MLP was introduced to capture both spatial context and fine-scale detail. Enhanced training data, including OpenStreetMap integration and the use of Sentinel-1/Sentinel-2 fusion, further improved model performance and classification of both physical land classes and occlusion attributes.

Finally, for **biomass mapping** (task 2.7), we successfully developed models to estimate above-ground woody biomass and forest canopy height using satellite time series and ground-based data. In addition, models for generating 10 meter resolution time series of cropland and grassland gross primary production (GPP) were also created. The recalibration of the light use efficiency (LUE) factor for these specific land cover types has already led to improved GPP estimates compared to using a global constant.

1 Introduction

1.1 About EvoLand

The Horizon Europe EvoLand (Evolution of the Copernicus Land Service portfolio) project (Project nr: 101082130) [AD1] aims at developing innovative methods, algorithms and candidate Copernicus Land Monitoring Service (CLMS) prototypes monitoring the status and changes of land use/land cover and various land surface characteristics at high spatial and temporal resolution.

Specifically, EvoLand addresses eleven next-generation CLMS product candidates by integrating innovative approaches in data fusion, Machine Learning, continuous monitoring and biomass mapping, as well as through integration of novel EO and in-situ data. EvoLand analyses policy, data and infrastructure requirements for the prototype services by interacting with the relevant Entrusted Entities as well as the main Copernicus Land stakeholders and users. These requirements will steer all methodological developments.

Within EvoLand, algorithms are developed, tested and implemented in an efficient, modular and scalable way on open-source infrastructure on various representative test sites, located both within and outside Europe. In a subsequent demonstration phase, the candidate services will be prototypically implemented over larger regions, focusing on key thematic domains such as agriculture, forest, water, urban and general land cover.

Finally, all candidate services are being regularly analysed in view of their innovation and policy support potential, technical excellence, and operational readiness, and a strategy to transferring these services to operations will be proposed.

The ambition of EvoLand is thus to support the Entrusted Entities by dedicated research, providing tangible proof of the evolution potential of the CLMS in terms of improved information content, quality and timeliness, and enable well-informed and facts-based decision-making on the future of the CLMS.

1.2 Purpose and objectives of WP2 Method development

The goal of WP2 is to develop the algorithms and methods required for the candidate CLMS services. This work package is shaped by the requirements analysis of WP1 [AD2] and is closely interlinked with WP3 - prototype demonstration. The method development is done by using existing and novel EO data, in-situ data and infrastructure.

First the consortium reviews, compiles and prepares all the required existing and upcoming EO datasets as well as all required in-situ datasets for model

development, calibration and validation. Additionally, a novel, modular, scalable and open-source IT infrastructure is established, suited to address the project needs. Next, these datasets and infrastructure are used to develop new algorithms for weakly supervised learning, improved resolution, continuous monitoring and biomass mapping.

All method and algorithm development are performed on test sites, located both within and outside Europe. The methods feed into the implementation of the eleven pre-operational CLMS prototypes, which will be demonstrated on larger prototypes sites.

WP2 encompass the following methodological developments:

- the integration of relevant novel EO datasets (T2.1)
- the acquisition of fit for purpose in-situ and training data (T2.2)
- set-up of a modular, scalable and open-source infrastructure (T2.3)
- Weakly Supervised Learning (T2.4)
- improved spatial, temporal and thematic resolution (T2.5)
- continuous monitoring (i.e. change detection and continuous land cover mapping) (T2.6)
- biomass mapping (T2.7)

1.3 Document structure

This deliverable is the **final version** (D2.17) of the **methods compendium** for the EvoLand project. Table 1 presents the structure of the document, highlighting the different themes and methods tested.

Table 1. Document structure.

Theme	Novel Method	Chapter
Weakly supervised learning	SSL4EO-S12	2.2.1.1, 2.2.2.1, 2.2.3.1, 2.3.1
Weakly supervised learning	DeCUR	2.2.2.2, 2.2.3.2, 2.3.2
Weakly supervised learning	SoftCon	2.2.2.3, 2.2.3.3, 2.3.3
Weakly supervised learning	CromSS	2.2.2.4, 2.2.3.4, 2.3.4
Weakly supervised learning	SpectralEarth	2.2.1.5, 2.2.2.5, 2.2.3.5, 2.3.5
Weakly supervised learning	MMDC, pVAE,	2.2.1.7, 2.2.2.6, 2.2.3.6, 2.3.6
Weakly supervised learning	ALICE, MALICE, MALICE aux	2.2.1.7, 2.2.2.7, 2.2.3.6, 2.3.6
Improved resolution	Single Image Super-Resolution	3.1
Improved resolution	Fusion of Satellite Image Time Series including Thermal Infrared	3.2
Continuous monitoring	Change detection – Time series modelling	4.1
Continuous monitoring	Disturbance classification	4.2
Continuous monitoring	Change detection – AI methods	4.3
Continuous monitoring	Continuous mapping of Land Surface Categories	4.4
Biomass mapping	Forest Biomass	5.2
Biomass mapping	Grass-/cropland GPP	5.3

This document is accompanied by two deliverables D2.19 Test datasets (Weakly Supervised Learning, Improved Spatial/Temporal/Spectral resolution, Continuous monitoring, Biomass mapping - final version [AD3] and D2.21 Source code (Weakly Supervised Learning, Improved Spatial/Temporal/Spectral resolution, Continuous monitoring, Biomass mapping - final version [AD4].

1.4 Link to WP1 requirements

Work Package 1 collects stakeholder and key end-user requirements for an enhanced CLMS portfolio in the Copernicus 2.0 period to 2027 and beyond, capturing key requirements how to respond to current and upcoming EU policy reporting objectives whilst, building on latest technological developments. This is performed through an analysis of existing studies and continuous screening of latest political & technological developments, complemented by interviews with key stakeholders and end users. These user requirements [AD2] were consolidated to clearly define the data, infrastructure and operational requirements for the envisaged services as well as envisaged KPIs for the different tasks in WP2 and WP3.

1.5 Demonstration sites

The prototypes (WP3) are implemented in eight representative demonstration sites mainly in the EEA-38 + UK territory (six in continental Europe and one in French Guyana) plus one in Africa (Côte d'Ivoire), covering various bio-geographic regions and landscapes.

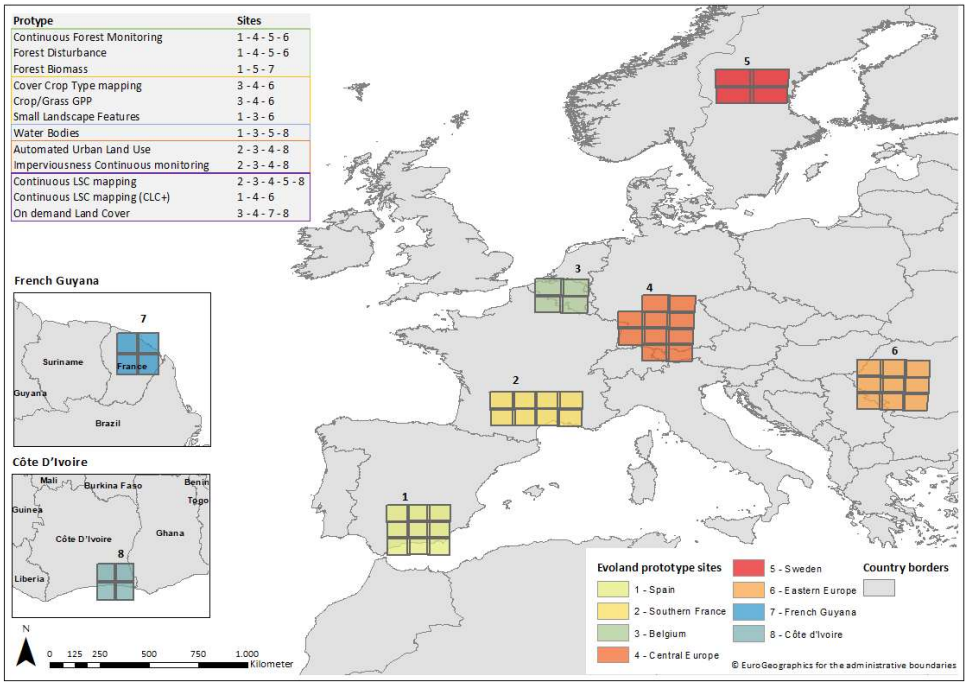


Figure 1. EvoLand test sites.

2 Weakly supervised learning

2.1 State of the art

2.1.1 Introduction

For the past 10 years advances in deep learning have led to major breakthroughs in computer vision, natural language processing, automatic speech recognition, and time series analysis (LeCun et al., 2015). Many of the early successes were achieved following the supervised learning paradigm, thanks to significant efforts to assemble large-scale, labelled datasets such as ImageNet (Deng et al., 2009). Unfortunately, annotating large-scale datasets is an extremely laborious, time-consuming, and expensive procedure. This limitation of the supervised learning paradigm strongly impedes the applicability of deep learning in real-world scenarios.

Significant efforts have been devoted to tackle the challenge of data annotation in the machine learning literature. In fact, numerous alternatives to supervised learning have been studied, such as unsupervised learning (Dike et al., 2018), semi-supervised learning (X. Yang et al., 2023), and meta-learning (Vanschoren, 2019). Recently, ***self-supervised learning*** (SSL) (Y. Wang et al., 2022a) has raised considerable attention in Computer Vision and Remote Sensing and achieved significant milestones towards the reduction of human supervision. These methods enable learning generic image representations that can be used for diverse downstream tasks by leveraging large amounts of unlabelled data.

Self-supervised methods follow a two-stages pipeline: pre-training followed by fine-tuning, as depicted in Figure 2. Based on a self-produced objective O, the so-called *self-supervision*, a large amount of unlabelled data X is exploited to train an artificial neural network model Z=E(X) by optimizing this objective O without requiring any manual annotation of X. Given a carefully designed self-supervised task defined through objective O, the model E learns high-level representations Z of the input X. Given labelled data, E can be further fine-tuned to supervised *downstream* tasks for applications such as the EvoLand prototypes of Work Package 3. From this perspective, the below review presents SSL as weakly supervised methodology where associated downstream tasks require annotated inputs X for supervised fine-tuning.

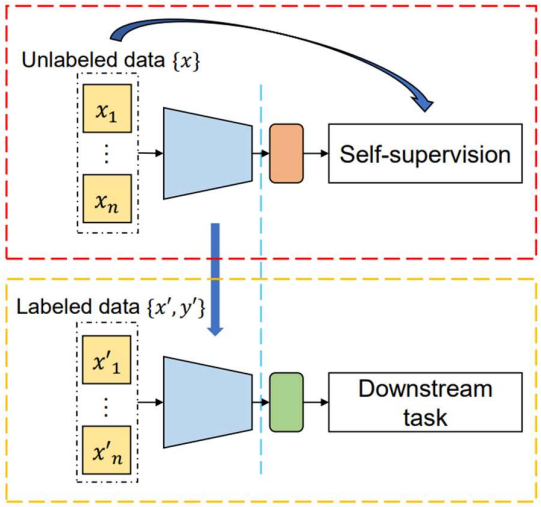


Figure 2. Schematic workflow of Weakly-Supervised Learning with Self-Supervision pre-training and downstream task adaptation.

2.1.2 Review of existing methods

Following the taxonomy defined in (Y. Wang et al., 2022a) (including an exhaustive list of references on SSL for remote sensing) and as summarized in Figure 3, self-supervised learning may get categorized into generative, contrastive, and predictive approaches. Generative methods learn to reconstruct or generate input data, predictive methods learn to predict self-generated labels, and contrastive methods learn to maximize the similarity between semantically identical inputs.

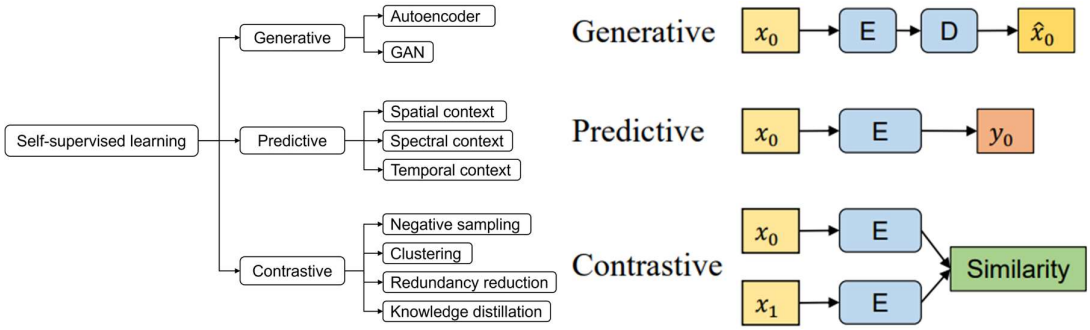


Figure 3. A taxonomy of self-supervised learning.

2.1.2.1 Predictive methods

Predictive self-supervised methods train an artificial neural network encoder E to predict specific properties of the data. To this end, the so-called *pretext tasks* get utilized. The rationale behind this approach is that, by predicting certain attributes of the input, the model will learn useful information about the structure of the data. In general, the design of different pretext tasks harnesses various context information of the input data (Figure 4). According to distinct context attributes,

we categorize pretext tasks as follows: spatial context, temporal context, and other semantic contexts:

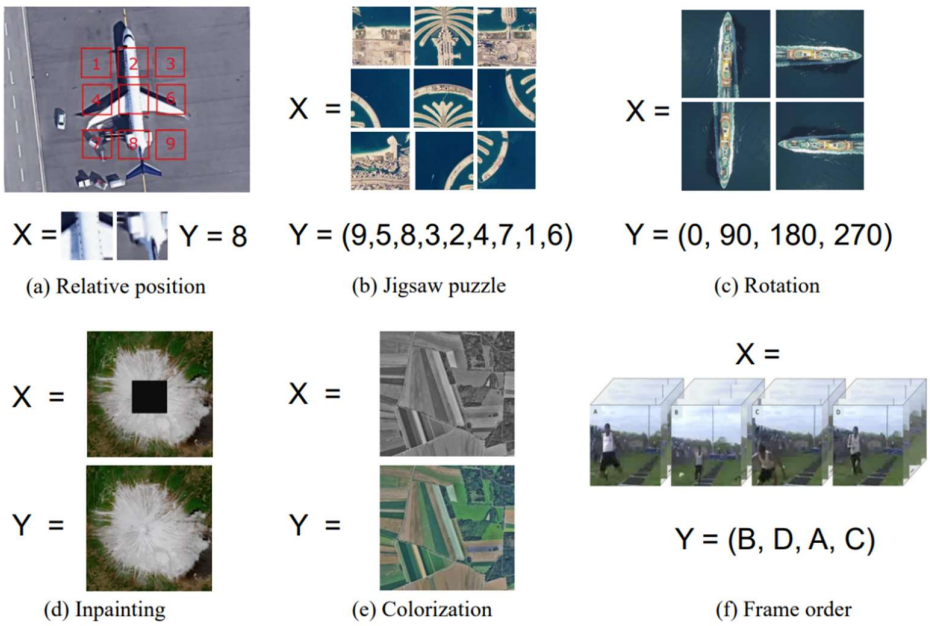


Figure 4. Example pretext tasks from the self-supervised learning literature.

1. **Spatial context:** Images contain rich spatial information for designing pretext tasks. (Doersch et al., 2015) proposed the first example of such methods, predicting the relative position Y=E(X) of pairs of randomly cropped image patches X drawn from a given input sample. It is assumed that doing well on this task requires a global understanding of the scene and its objects contained. Following this paradigm, the literature revealed a multitude of methods to learn image features solving an increasingly complex set of spatial puzzles Y=E(X) with some illustrated in Figure 4 (a) position, (b) order of tiles, (c) image rotation, (d) fill-in blanked image patches, and (e) image colorization. Further details and references in (Y. Wang et al., 2022b) .
2. **Temporal context:** Temporal context is key in domains such as video understanding. Footage clearly carries temporal correlation from frame to frame. Like spatial context-based methods, several works have exploited the order of the frames to design pretext tasks. Two main approaches have been investigated in the literature.
    a. temporal order verification (e.g., (Wei et al., 2018)), which amounts to verifying whether a sequence of frames is temporally sorted, and
    b. temporal order recognition, Figure 4(f), (e.g., (D. Kim et al., 2019)) where the network explicitly predicts timestamps to a set of given frames.
3. **Other semantic contexts:** Apart from spatial and temporal contexts, other semantic contexts are, for example: counting visual primitives from an input image (Noroozi et al., 2017), prediction of image artifacts (Jenni & Favaro,

2018). Multi-sensor-modal self-supervised learning provides another set of semantic contexts: (Owens et al., 2016) proposed to predict a statistical summary of the sound associated with a video frame. (Ren & Lee, 2018) trained an encoder-decoder network to predict depth maps, surface normal maps, and instance contour maps from a synthetic image, and a discriminator network is trained to distinguish between features from the synthetic image and the real image.

2.1.2.2 Generative methods

Generative self-supervised methods learn representations Z=E(X) from input data X by reconstruction or generation: X'=D(Z) ~ X. A prominent set of methods in this category are autoencoders (AE) with optimization criterion O that minimizes |X – D(E(X))|. However, one can easily imagine an autoencoder to fail when D·E=1, i.e. the joint application of E and D reduces to the trivial identity mapping. Thus, demanding the latent vector Z to retain information on X is insufficient for stable and efficient representation learning. Constraints are needed to prevent such scenarios. For example, in early practices, the encoder E typically resembles a bottleneck-like structure such that the dimension of the input X gets compressed down compared to the embedding Z. Other popular approaches include sparse auto-encoders (Shao et al., 2019), denoising autoencoder (Vincent et al., 2008), and variational autoencoder (VAE)(Vincent et al., 2008).

Autoencoders have been widely used to learn representation from various remote sensing data like multispectral images, hyperspectral images, and SAR images (for a comprehensive list of references, see (Y. Wang et al., 2022c)). For example, (Lu et al., 2017) proposes a combination of a shallowly weighted de-convolution network with a spatial pyramid model to learn multi-layer feature maps. In a subsequent step, these get classified by a support vector machine (SVM) for scene classification (downstream task). In hyperspectral image analysis, autoencoders are either exploited for pre-training (Mou et al., 2018). Furthers, downstream tasks like hyperspectral image classification (Kemker & Kanan, 2017) and hyperspectral image unmixing (B. Palsson et al., 2022) benefit from generative methods.

More recently, masked autoencoders (MAEs, see He et al., 2022 and its illustration replicated in Figure 5) raised great attention in the computer vision community by a breakthrough in self-supervised pre-training of vision transformers (ViT, Khan et al., 2022). Inspired by denoising autoencoder, an MAE masks out random patches m of the input image X=(x,m), sends visible patches x to the encoder E, and reconstructs the missing patches m from the latent representation z=E(x) and masked tokens n=0 by a decoder D(Z)=X'=(x',m') through an objective such as to minimize |X-X'|.

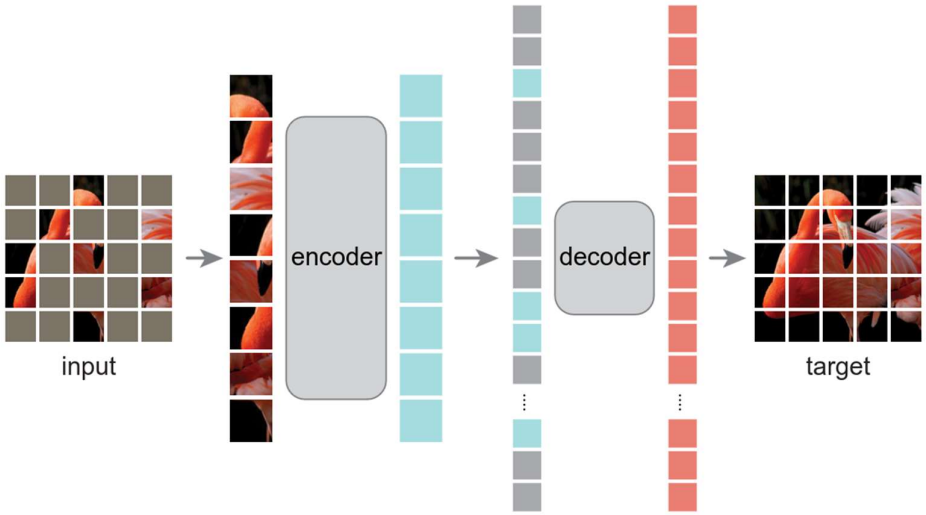


Figure 5. An illustration of Masked Auto Encoders (He et al., 2022).

- **Reconstruction target:** one can distinguish two groups of reconstruction targets:
  - low-level features: can be the raw pixel values in the simplest case such as MAE and ConvMAE (P. Gao et al., 2022) or hand-crafted features (e.g., HOG descriptor) such as MaskFeat (Y. Wang, Hernández, et al., 2023)
  - high-level features: BEiT (Bao et al., 2021) predicts visual token generated by a pre-trained model whereas Data2vec (Baevski et al., 2022a) and SIM (Fu et al., 2024) regress features from the momentum encoder during pre-training.
- **Loss function:** the most common optimization criterion with image reconstruction Y=D(E(X)) or feature Z=E(X) targets are L1-, L2-, and smooth L1-norm losses (Q. Wang et al., 2022) . When the target is a visual token, cross-entropy is often used.
- **Network architectures:** most works rely on attention-based backbones such as ViT and Swin Transformers (Z. Liu et al., 2021). Light-weight decoders are generally preferred over complex architectures.

2.1.2.3 Contrastive methods

Contrastive methods follow a Siamese network approach (Chicco, 2021;Chen & He, 2021) by taking as input X two views of the same image, X' and X" to train on generating similar representations Z'=E1(X') ~ Z"=E2(X") for these views. E1 and E2 are coupled by a student-teacher relationship (L. Wang & Yoon, 2022) or simply E=E1=E2. In practice, the views are stochastically generated using data augmentations, X->X' and X->X", thereby implicitly forcing the model(s) E(1/2) to become invariant to these transformations. However, only enforcing similarity between pairs of input can easily lead to a trivial solution. Indeed, a constant

mapping with E(X)=const. would be a valid solution to the problem since every pair of input would have identical representations optimizing for O: minimize |E(X')-E(X")|. This phenomenon is often referred to as *model collapse* and many solutions have been proposed in the SSL literature to mitigate. Correspondingly, the following classes of algorithms developed, cf. Figure 6:

- **Contrastive Methods with Negative Samples**, Figure 6(a) (E1=E2)**:** These methods explicitly prevent model collapse by training the network to pull apart the representations of *different* images in the dataset (the so-called *negative* pairs). Among the most popular contrastive learning algorithms: SimCLR (T. Chen et al., 2020) and MoCo (He et al., 2020)
- **Clustering Methods**, Figure 6 (b) (E1≠E2)**:** These methods cluster the representations Z of the inputs based on their similarity. One popular method is SwAV(Caron et al., 2020).
- **Distillation Methods**, Figure 6 (c) (E1≠E2)**:** These methods directly optimize for the similarity between positive pairs, X' and X". The mechanism on why these approaches avoid model collapse are not fully understood, but they all rely on a set of tricks that introduce an asymmetry in the Siamese architecture: momentum encoder, predictor head and stop gradient operations. Among the most effective algorithms are BYOL (Grill et al., 2020a) and SimSiam (X. Chen & He, 2021b)
- **Redundancy Reduction**, Figure 6 (d) (E1=E2)**:** These techniques introduce a regularization loss term to the optimization criterion O to decorrelate the embedding dimensions of Z from each other. Recent examples from this family of methods are Barlow Twins (Zbontar et al., 2021) and VICReg (Bardes et al., 2021) .

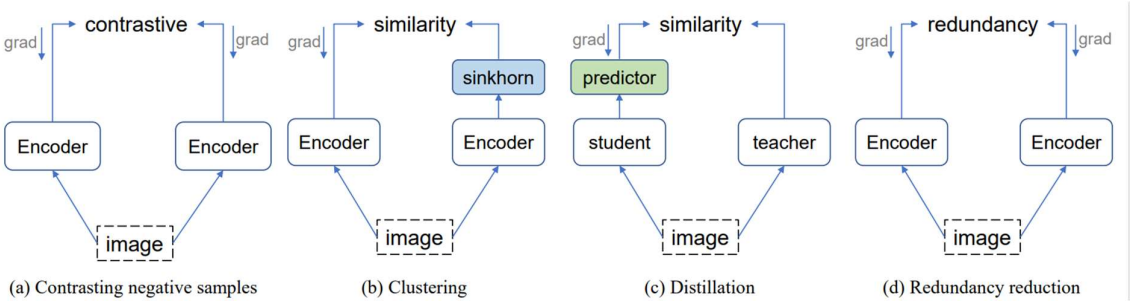


Figure 6. A comparison of existing contrastive methods image: X, Encoder: E, student: E', teacher: E".

2.1.2.4 Backbone architectures

Deep learning in general, and SSL in particular, consists of two main elements: an optimization protocol O for model (pre-)training (cf. Sect. 2.1.2.1-3) and a corresponding neural network architecture model (backbone). As alluded in chapter 2.1.3 for EvoLand we adopt the generative and contrastive training paradigm building on the most popular architectures currently available in computer vision: Vision Transformers (ViT) and Residual Networks (ResNet, (He et al., 2016a)), respectively. Figure 7 illustrates key elements of the backbones:

- **ResNet** (top, replicate from (He et al., 2016b)) with 208k+ citations, and
- the **ViT** (bottom, replicate from (Dosovitskiy et al., 2020) with ~32k citations as per google scholar[1] in March 2024.

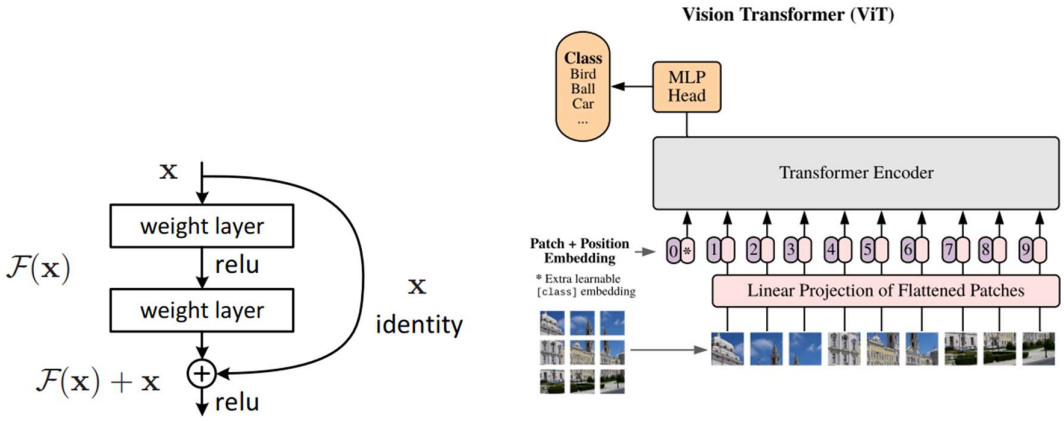


Figure 7. Illustration of the two backbone architectures employed for AI models in EvoLand – left: a ResNet block with skip connection, right: a ViT for image classification.

ResNet: Deep convolutional neural networks (CNNs) (Rawat & Wang, 2017)significantly boosted the performance of image classification tasks in computer vision. Given N pixels of a feature map fN and M pixels of a new feature map fM: While a fully connecting (linear) layer requires the training of NxM parameters, for a convolutional layer L the computational effort reduces to the number of K<<NxM sliding window kernel parameters connecting the feature maps, fM=L(fN). However, to convert an input image I (the initial feature map I=fN) into a stack of representative feature maps, it requires tens to hundreds of (two-dimensional) convolutions. When numerically optimizing by backpropagation through tens of layers bears the risk of a vanishing gradient (H. H. Tan & Lim, 2019). It is a hallmark of the chain rule for derivatives on the network layer parameters. ResNets solve this challenge through skip connections that provide backpropagation routes through the network without funnelling through all convolutional layers. Effectively, skip connections bypass convolutional layers modelling fM=fN+L(fN) such that L(fN) represents just a variation of converting fN into fM. The nomenclature for a ResNet with 18 or 50 convolutional layers reads: ResNet-18 and ResNet-50.

ViT: With the success of Transformers in Natural Language Processing (NLP) (Tunstall et al., 2022) emerged research lines to follow these trends in computer vision - with one such architecture the ViT backbone. While ResNets gradually build feature maps by increasing their field of view, transformers model correlations of distant locations in an image starting with layer one of the Transformer encoder. They do so, by breaking up an image into NxN patches (N=3 in Figure 7) and converting the patches by a linear projection into a sequence of vectors W (*word*

[1] https://scholar.google.com

embeddings in terms of NLP) feeding the W into the Transformer encoder in parallel (computational speed-up). To preserve information on the spatial location of a patch in the image, a (static) position embedding vector P is added, i.e. W -> W+P before it enters the Transformer encoder. For image classification, a placeholder input patch is prepended to the linear sequence which serves as an embedding for classification through a fully connected (linear) layer after running through the Transformer encoder.

The Transformer encoder itself implements an *attention mechanism* that models the global correlation between image patches/word embeddings. In fact, the Transformer encoder may include multiple attention and transformer layers which are chained with skip connections in analogy to the ResNet architecture. Based on the number of transformer layers: 12, 24, and 32, ViT defines ViT-B, ViT-L, and ViT-H. The size of a patch NxN is indicated by a trailing "/N", e.g., ViT-L/16 corresponds to a Vision Transformer with 24 layers (632 million parameters) and an input patch size of 16x16=256 pixels.

2.1.3 Strengths and weakness of existing methods

In summary: By designing an optimization criterion O on unlabelled data alone, self-supervised learning utilizes artificial neural networks to distil representations Z from data X through an encoder E(X)=Z. Downstream tasks for EvoLand prototypes in Work Package 3 may employ these representations when paired with a set of labelled data in a weakly-supervised setup.

Depending on the specific architecture of E and its corresponding training objective, downstream applications D(E(X))=Y may be tailored for either (multi-label) image classification or semantic segmentation on pixel-level. Although SSL promises generalizability of data X representations Z=E(X), to efficiently work for a wide range of downstream tasks, in practice there exists limits to this bold assumption which we will explore in Sect. 2.2. Downstream tasks implemented as *linear probing* are relatively inexpensive in computing resources when the downstream model D(Z)=Y represents a lightweight (linear) head with fixed pretrained model E=E(X)=Z. However, for the best accuracy possible, full fine-tuning includes adjusting the end-to-end model D∘E which may require larger amounts of labelled data.

Recent trends in SSL moved away from predictive methods since pretext tasks introduce potential bias to the representations Z when optimizing for Y=E(X) in pre-training. Thus, our candidate methods focus on generative (MAE) and contrastive (e.g., MoCo) SSL learning schemes utilizing ViT and ResNet backbones, respectively. The field of SSL is rapidly evolving due to the recent success of ChatGPT (Roumeliotis & Tselikas, 2023) that builds on Transformer models from which ViT derive.

The notion of *Foundation Models* (C. Zhou et al., 2023) is closely linked to SSL. In fact, geospatial foundation models have been released recently (Jakubik et al., 2023) exploiting the benefits of developments in large-language models (LLMs, (W.

X. Zhao et al., 2023)). Although ChatGPT remains closed source, a vital community around multi-modal models developed with openly available codes and datasets through the platform HuggingFace[2].

2.2 Method description

In this section we introduce candidate methods and corresponding datasets based on the following publications:

- **SSL4EO-S12** (Sentinel-2): (Y. Wang, Braham, et al., 2023)
  - dataset (pre-training & downstream benchmarks) in section 2.2.1.3
  - models and benchmarks in section 2.2.1.3 and sections 2.2.3.1 and 2.3.1
- **DeCUR** (Sentinel-1/2): (Y. Wang, Albrecht, et al., 2023)
  - data is same as SSL4EO-S12
  - models and benchmarks in section 2.2.2.2, 2.2.3.2 and 2.3.2
- **SoftCon** (Sentinel- 1/2): (Y. Wang et al., 2024)
  - data is same as SSL4EO-S12
  - Models and benchmarks in section 2.2.2.3, 2.2.3.3 and 2.3.3
- **CromSS** (Sentinel-1/2): (C. Liu et al., 2025)
  - data is based on SSL4EO-S12
  - Models and benchmarks in section 2.2.2.4, 2.2.3.4 and 2.3.4
- **SpectralEarth** (EnMAP): (Ait Ali Braham et al., 2024)
  - dataset (pre-training & downstream benchmarks) in section 2.2.1.5 2.2.1.5
  - models and benchmarks in sections 2.2.2.5, 2.2.3.5, 2.3.5, 2.2.2.5, 2.2.3.5 and 2.3.5
  - (Gonzalez et al., 2025) models to fuse SpectralEarth with multi-spectral data such as Sentinel-2 in sections 2.2.2.5.3, 2.2.3.5.4 and 2.3.5.5

As we demonstrate, image-level classification utilizing Z=E(X) works best for these models. However, DLR in collaboration with GAF explored a potential prototype on forest image semantic segmentation on pixel-level utilizing pretrained SSL4EO-S12 models by assembling a targeted fine-tuning dataset **SSL4EO-EU-Forest** (chapter 2.2.1.3) with corresponding benchmark & evaluation reported in chapter 2.3.1

We equally propose a Multi-Modal DataCube (MMDC) model able to extract common representations between Sentinel-1 and Sentinel-2 images exploiting the auxiliary data, such as meteorological variables and digital elevation model.

- **MMDC** (Sentinel-1/2):
  - dataset (pre-training & downstream benchmarks) in section 2.2.1.7
  - models and benchmarks in section 2.2.2.6 and 2.2.3.6

[2] https://huggingface.co

While the current model only exploits single timestamp data to produce latent representations, we are working on the integration of MMDC model into different multi-modal and multi-temporal self-supervised algorithms, such as U-Barn (Dumeur et al., 2024a), able to extract spatio-temporal features from satellite image time series.

Moreover, the latent representations Z=E(X) produced by current MMDC model are not physically interpretable, therefore, the physical models will be integrated into MMDC model (Zérah et al., 2023), so the produced latent variables represent different phenological parameters, such as Leaf-Area-Index of Leaf Water Content.

2.2.1 Input data and study sites

It follows a detailed description of datasets curated for pre-training and benchmarking the SSL models developed for EvoLand with summary in Table 2.

Table 2. Summary of dataset utilized and assembled for EvoLand method development experiments on SSL.

Dataset	Description	Spatial Resolution	Temporal resolution
SSL4EO (pretraining, existing)	spatio-temporally co-registered Sentinel-½ (radar/optical) tiles globally distributed	10m	4 seasons (4 timestamps)
SpectralEarth (pretraining, newly assembled)	hyperspectral EnMAP tiles	30m	single timestamp
SSL4EO-EU-Forest (downstream for GAF, newly assembled by DLR)	spatio-temporally co-registered Sentinel-½ (radar/optical) tiles over forest areas in Europe	10m	4 seasons (4 timestamps)
MMDC Dataset	12 tiles acquired over Europe from 2017 to 2020 containing S1, S2, Meteorological data, DEM	10 m	Irregular, depending on data availability
Pastis Dataset	2433 patches containing S1, S2, Meteorological data, DEM; acquired over French metropolitan area from oct.2018 to nov.2019, GT with 20 classes of crop cultures	10 m	Irregular, depending on data availability
TCD dataset	1 tile containing S1, S2, Meteorological data, DEM; acquired over Europe from March 2018 to October 2018, GT in form of points with TCD values	10m	Irregular, depending on data availability
Biomass and Tree Height Dataset	Sentinel-1 Ascending 1-year patches spread all over Sweden, GT in form of points with Tree Height and Biomass values. Data from 2015 to 2021. With 1-year of data for each GT point.	10m	Irregular, depending on data availability

2.2.1.1 SSL4EO-S12

SSL4EO-S12 (as in *Self-Supervised Learning for Earth Observation - Sentinel-1/2*) entails a large-scale, global, multimodal, and multi-seasonal corpus of satellite imagery from the ESA Sentinel-1 (radar) & -2 (multi-spectral) satellite missions. It comprises of over 3 million patches sampled all around the globe over more than 250,000 locations - each containing 4 timestamps (one per season) using imagery from the years 2020 and 2021 covering 3 products: Sentinel-2 L1C, Sentinel-2 L2A and Sentinel-1 GRD. Sampling locations are based on a protocol proposed by "SeCo - Seasonal Contrast" (Manas et al., 2021), where patches are selected from the top-10K most populated cities in the world using Gaussian distribution peaking at the city centre with standard deviation of 50km. It is assumed that most of the variability of imagery is cast by scenes from downtown to suburbs, to rural areas, and oceans around the major cities on planet Earth. Contrarily to SeCo, SSL4EO-S12 included a filtering constraint to guarantee minimal overlap of sample patches to reduce the redundancy in the data. A visualization of the spatial distribution of SSL4EO-S12 provides Figure 8 and Figure 9 illustrates the 264x264 pixel patches for all modalities over 4 seasons.

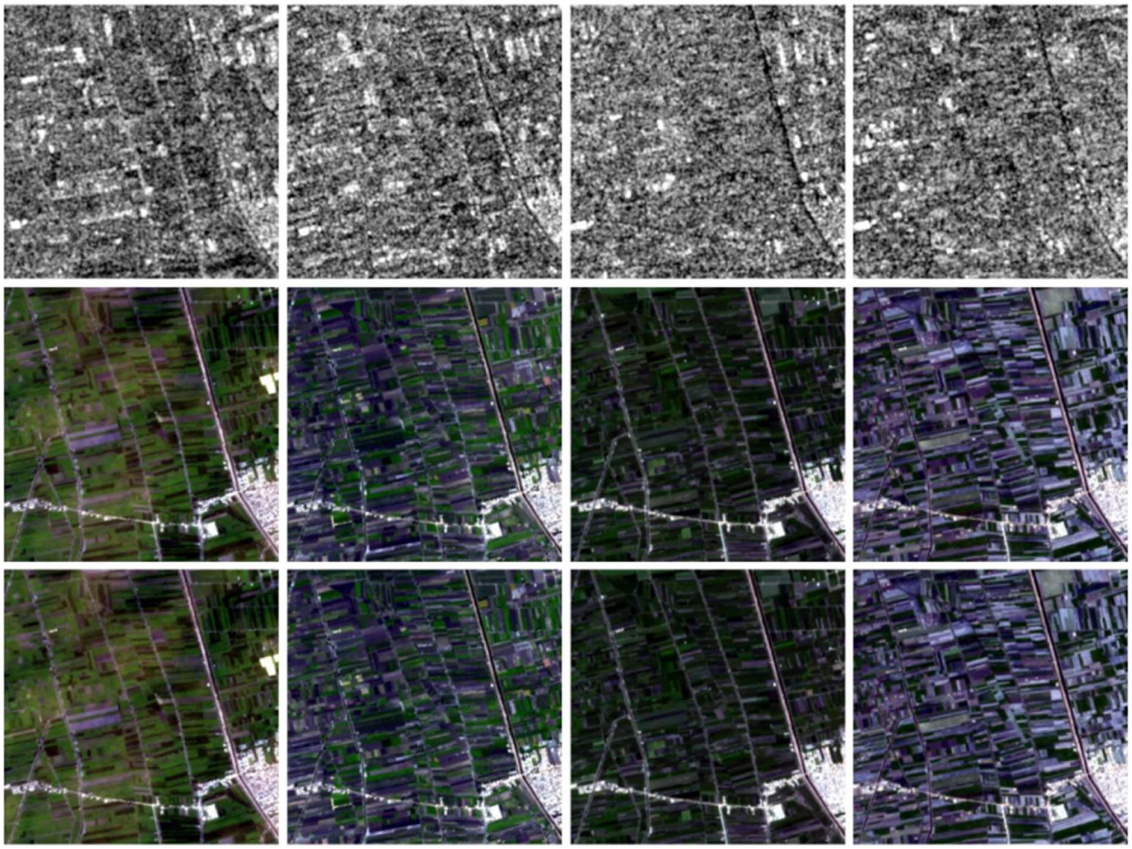


Figure 8. Sample visualization of one patch from SSL4EO-S12 dataset. The rows top-down present grayscale and false-colour imagery based on the Sentinel-1 GRD product, Sentinel-2 level-1C, and Sentinel-2 level-2A multispectral data with corresponding columns representing the four seasons spring, summer, fall, and winter from left to right.

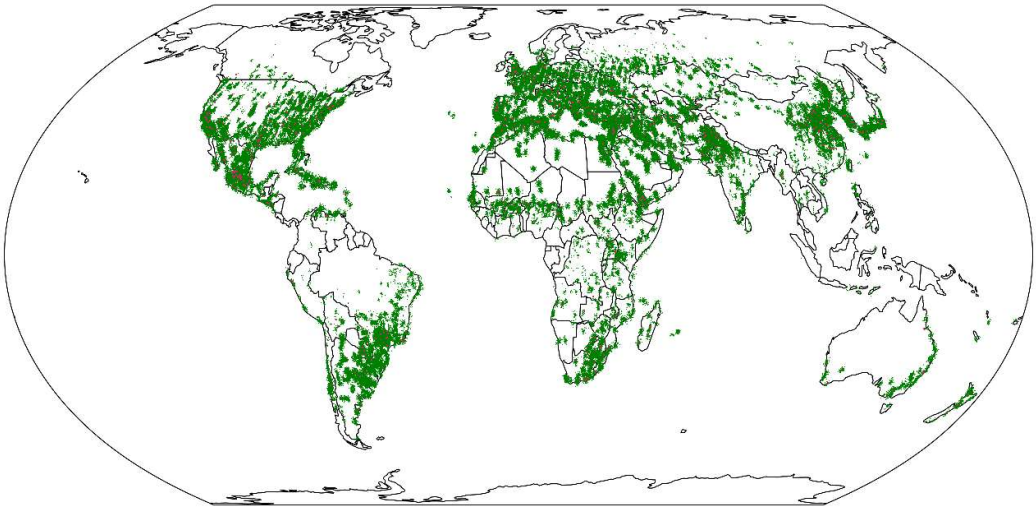


Figure 9. Geographical distribution of SSL4EO-S12 dataset.

2.2.1.2 SSL4EO-S12 V1.1

This update to the SSL4EO-S12 dataset introduces several improvements summarized below:

- **Spatial Alignment** – Sentinel-1 (S-1) data is now reprojected to match Sentinel-2 (S-2) L2A coordinates, reducing misalignment issues from the previous version. To preserve spatial accuracy. patches are retained in their native UTM zones, unlike the original version of SSL4EO-S12 where data has been reprojected to EPSG:4326.
- **Data Preprocessing** – Samples are now filtered more strictly to ensure better consistency across timestamps. Missing values beyond a 1% threshold are removed, and remaining gaps are filled using nearest-neighbour interpolation.
- **Cloud Mask Refinement** – The original Sentinel-2 cloud mask, known for inaccuracies, has been replaced with a refined mask generated using the SEnSeI v2 model.
- **Standardized Data Offsets** – A numerical offset of +1000 is applied to S-2 data in accordance with the latest ESA processing standards.

Table 3 summarizes the modalities included.

Table 3. Summary of modalities included in SSL4EO-S12 v1.1. The range includes the typical value range, including the offset, but values below and above the range are present as well.

Modality	DType	Range	Units	# Bands
S-1 GRD	float16	-50 – +1	dB	2
S-2 L1C	int16	1k – 11k	DN	13
S-2 L2A	int16	1k – 11k	DN	12
S-2 RGB	uint8	0 – 255	–	3

2.2.1.3 SSL4EO-EU-Forest

As an additional downstream task for pretraining SSL4EO-S12 models, DLR with support by GAF assembled a tree-cover mapping dataset pairing Sentinel-2 imagery with the High-Resolution Layer (HRL) 2018[3]. To increase the diversity of the training data, in accordance with the procedure of data collection for the SSL4EO-S12 dataset, we collect four timestamps per forest geolocation as indicated by the 2018 HRL dataset. The resulting (labelled) dataset consists of approx. 16,000 locations distributed across Europe, with a total of about 48,000 Sentinel-2 L2A patches as shown in Figure 10 (left).

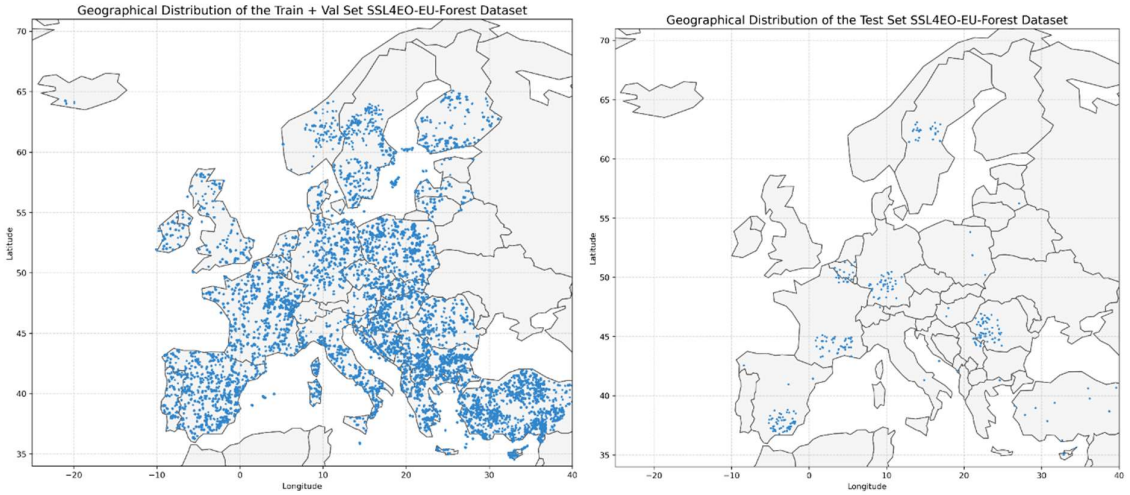


Figure 10. Geospatial sampling distribution of SSL4EO-EU-Forest (left) from which we separated the samples of the EvoLand test sites (right).

To evaluate the generalizability of our models, we split the dataset into training, validation, and test set. Training and validation are randomly distributed in Europe, while the test set contains images from the test areas (Figure 10, right).

2.2.1.4 SSL4EO-EU-Forest-V2

For the second version of the SSL4EO-EU-Forest dataset, we increased the initial pool to approximately 20,000 locations, expecting that some would be removed during quality filtering. As in SSL4EO-S12 v1.1, we used UTM zone projections to improve the spatial accuracy of image tiles. In addition, we applied more rigorous filtering, including cloud coverage thresholds and mask quality checks (see Section 2.3.1.3 for more details). After these steps, the final dataset includes around 17,000 locations with reduced label noise compared to the first version of SSL4EO-EU-Forest. The geographical distribution of the SSL4EO-EU-Forest-V2 dataset is displayed in Figure 11**.**

[3] https://doi.org/10.2909/486f77da-d605-423e-93a9-680760ab6791

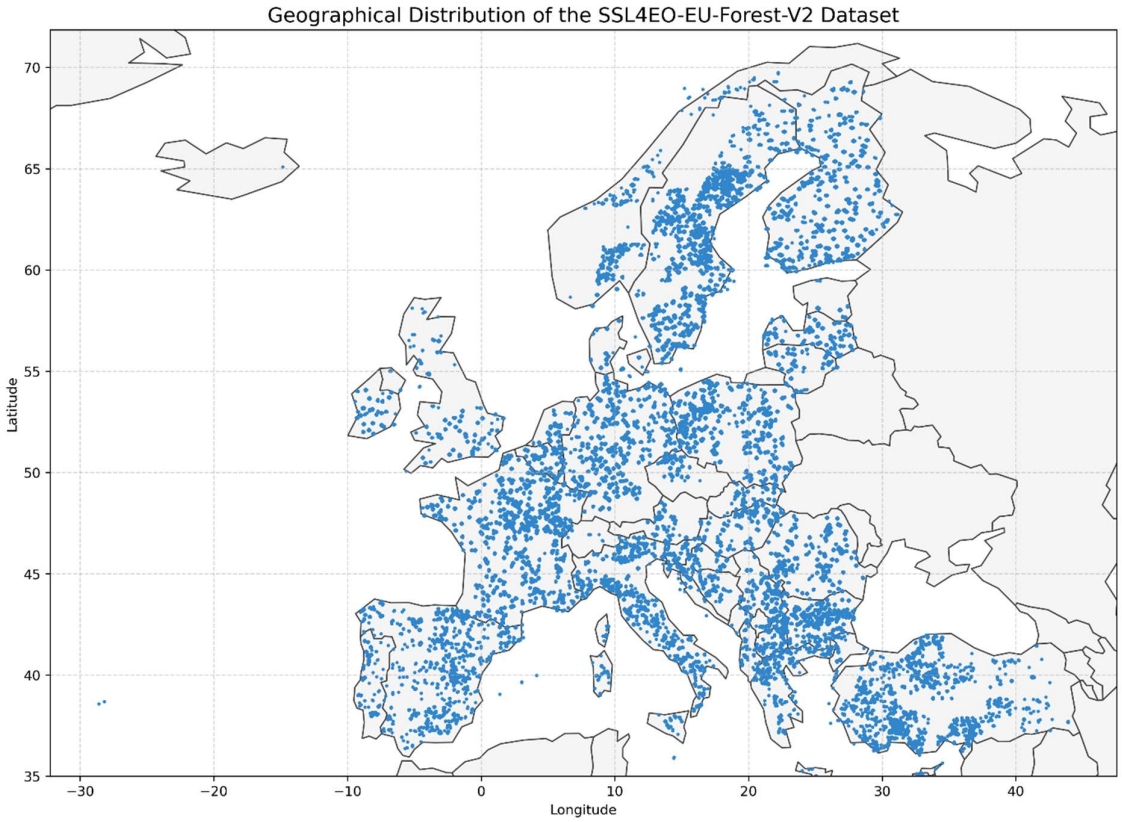


Figure 11. Geospatial sampling distribution of SSL4EO-EU-Forest-V2 dataset.

2.2.1.5 SpectralEarth

SpectralEarth, a large-scale dataset derived from the DLR EnMAP satellite mission, featuring over 250,000 hyperspectral 128x128 pixel image patches of 202 spectral channels with global spatial distribution and cloud coverage below 10%. In 2024, we have extended the spatial and temporal coverage of the SpectralEarth dataset. The dataset now comprises more than 538,000 image patches covering about 415,000 unique locations from 11,636 globally distributed EnMAP scenes spanning two years of archive. Additionally, 17.5% of these locations include multiple timestamps, enabling multi-temporal HSI analysis.

2.2.1.5.1 Pre-training data

The EnMAP data was sourced from the DLR GeoPortal[4]. A total of 11,636 tiles were retrieved via the portal's graphical interface. All tiles were carefully selected based on human visual inspection to guarantee cloud coverage below 10%. When assembled, the selection process covered the entire EnMAP satellite archive from April 2022 to April 2024. The EnMAP tiles have been broken up into geo-spatial patches of size 128x128 pixels, with each pixel encapsulating the full spectral

[4] https://eoweb.dlr.de/egp

resolution of 224 bands. To prevent identical or highly overlapping images which may adversely affect model performance for contrastive SSL due to redundant data, we engineered all SpectralEarth patches free from mutual overlap in space. To exploit all available tiles, we utilized overlaps between EnMAP data acquisitions to generate time series of EnMAP patches. This process required (a) the identification of tiles that overlap spatially and (b) extracting patches from the intersection of those tiles. Time series include valuable information on landscape evolution and provide characteristic signals for seasonal variation---an asset for contrastive learning. Details are provided in Figure 12.

Algorithm 1 Temporal Views Extraction

```
procedure MAIN PROCEDURE
    tiles ← EnMAPData
    overlap_graph ← GETOVERLAPS(tiles)
    R_tree ← K_0        ▷ empty tree, for SpectralEarth patches
    for tile in tiles do
        combs ← BUILDCOMBINATIONS(tile, overlap_graph)
        for tile_subset in combs do
            intersection ← INTERSECTION(tile_subset)
            patches ← PATCHIFY(intersection)
            UPDATE(R_tree, patches)
        end for
    end for
end procedure

function BUILDCOMBINATIONS(tile, overlap_graph)
    combinations ← GETEDGES(tile, overlap_graph)[2]
    for subset size n in [3, 4, ...] do
        get n-tuples from (n-1)-tuples in combinations
        compute intersections of all n-tuples
        keep largest n-tuples by area
        if no valid n-tuple found then
            break
        end if
        add n-tuples to combinations
    end for
    return combinations
end function
```

Figure 12. Geographical Distribution of the SpectralEarth dataset.

Bands dominated by water absorption were excluded due to frequent occurrences of missing values - a common practice in hyperspectral dataset creation (Fuchs & Demir, 2023). As a result, the total number of bands per image is reduced to 202. The geographical distribution of these 538,974 patches from 415,153 unique locations with 73,307 including multiple timestamps, each of size 128x128x202, showcases the dataset's extensive global coverage, as illustrated in Figure 13.

Notably, the data density varies across different areas, reflecting the flight request-based archive composition of EnMAP.

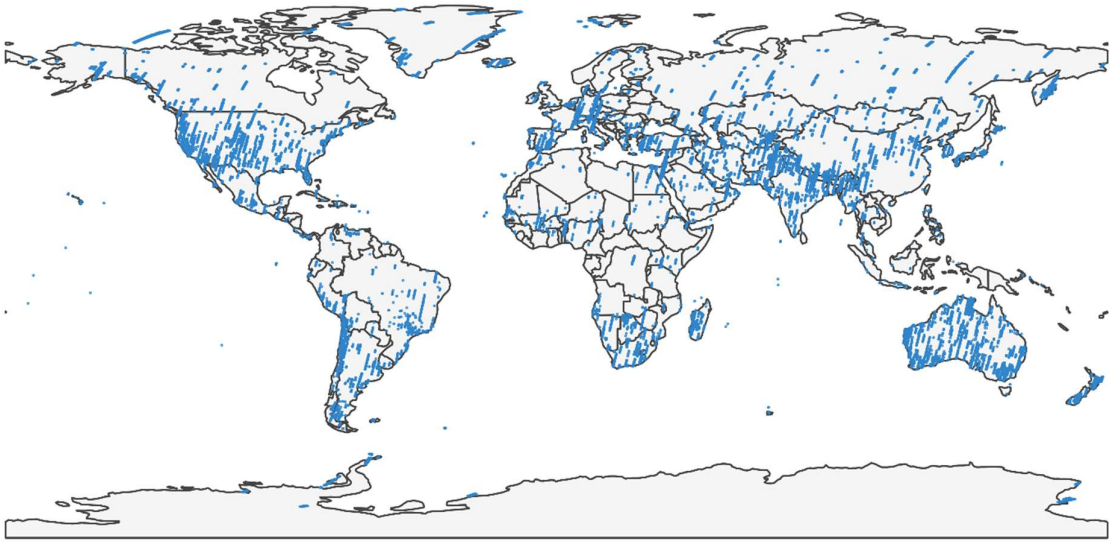


Figure 13. Geographical Distribution of the SpectralEarth dataset.

2.2.1.5.2 Downstream datasets

Since the last version of the report, we have updated existing datasets with improved filtering of clouds. Moreover, we have assembled additional downstream tasks to cover a wider range of applications, particularly forest related datasets. An overview of the downstream tasks is provided in Figure 14.

Figure 14. Illustration of various downstream tasks for SpectralEarth, from multi-label image classification to pixel-level semantic segmentation.

EnMAP-CORINE: This dataset pairs EnMAP imagery with the CORINE land cover database for Europe. The CORINE land cover map (Büttner, 2014) ships at 100m spatial resolution. This compares more coarse-grained w.r.t. EnMap's 30m pixel resolution. Our SpectralEarth image patches correspond to approximately 38x38 CORINE land cover pixels. Consequently, we create a multi-label classification benchmark where each SpectralEarth patch is annotated with the set of CORINE classes geospatially covering an area of 15 square kilometers. The resulting subset, called EnMAP-CORINE, includes 16,250 images, each multi-labeled with 19 distinct classes, aggregated from the 44 categories defined in CORINE following the taxonomy proposed by BigEarthNet-MM, (Sumbul et al., 2021a).

EnMAP-CDL: This dataset aligns EnMAP images with the Cropland Data Layer (CDL, (Boryan et al., 2011)) product for the United States, which provides extensive crop type annotations over the United States. The CDL map has been published annually by USDA's National Agricultural Statistics Service since 2008. It targets

agricultural crop classification. In addition to classes for cultivated land such as Soybeans and Winter Wheat, it encompasses non-cultivated areas like Grassland and Shrub-land. To minimize the impact of seasonal variations on crop classification, the dataset does focus on SpectralEarth patches exclusively from the summer of 2022 and 2023 - a period relevant for crop analysis compared to winter and fall. Given the strong imbalance among CDL classes, we resampled the class distribution to ensure a more balanced representation. The resulting dataset comprises of 1,600 hyperspectral images with a total of 14 classes.

EnMAP-NLCD: This downstream dataset leverages the National Land Cover Database (NLCD, (L. Yang et al., 2018)) from the U.S. Geological Survey (USGS) to provide pixel-level land cover annotations. The NLCD map, a comprehensive land cover product, has been published every 2–3 years since 2001. The data is available up to 2019. The resulting dataset contains 13,500 hyperspectral images, each annotated with pixel-wise land cover class labels from one out of the 15 land cover classes.

EnMAP-Eurocrops: The EnMAP-EuroCrops dataset aligns EnMAP imagery with crop type labels derived from the EuroCrops dataset, which is sourced from national agricultural inventories across European countries. This dataset focuses on crop-type classification for four countries—France, Germany (Brandenburg), Czechia, and Spain—for the year 2023, with additional coverage for France in 2022. The dataset consists of 1,800 patches categorized into 15 classes.

EnMAP-TreeMap: The EnMAP-TreeMap dataset maps EnMAP imagery from 2022 to TreeMap, a tree species classification product for the United States. Since TreeMap is based on data from 2016, we account for land cover changes by masking regions affected by deforestation or afforestation using the Hansen Global Forest Change dataset. The resulting dataset includes 10,000 patches covering 24 tree species, enabling fine-grained classification of forest types.

EnMAP-BDForet: The EnMAP-BDForet dataset combines EnMAP hyperspectral imagery with the BDForêt V2 dataset from the Institut Géographique National (IGN) France. This dataset focuses on classifying tree types using EnMAP patches information from 2022 to 2025. To mitigate land cover changes affecting classification accuracy, the latest IGN forest mask is applied to remove areas of deforestation. The dataset comprises 2,550 patches spanning 12 tree species, supporting forest ecosystem analysis.

EnMAP-BNETD: To explore the geographical generalizability of hyperspectral models beyond pre-training regions, we introduce the EnMAP-BNETD dataset, covering land cover classification in Côte d'Ivoire. This dataset utilizes the BNETD 2020 Land Cover Map, produced by the BNETD-CIGN with support from the European Union, and validated through extensive field campaigns in 2022 and 2023. EnMAP hyperspectral imagery is paired with land cover labels from this product, resulting in a dataset of 2,100 patches across 10 land cover classes.

DESIS-CDL: The DESIS-CDL dataset contains images captured in summer 2021-2023 from the DESIS (DLR Earth Sensing Imaging Spectrometer Mission (Eckardt et al., 2015)) instrument, matched with CDL for crop segmentation. DESIS images have 30m spatial resolution with 235 bands covering wavelengths between 400nm and 1000nm. The resulting dataset assembles 1000 images with 14 classes.

EO1-CDL: The EO1-CDL dataset investigates cross-sensor generalizability by pairing EO-1 Hyperion hyperspectral data with crop type labels from the Cropland Data Layer (CDL). EO-1 Hyperion provides 220 spectral bands covering 0.357 to 2.576 micrometres, offering high spectral resolution for vegetation analysis. The dataset includes Hyperion scenes from 2002 to 2016, focusing on agricultural regions over the US. Due to the limited availability of Hyperion imagery in Google Earth Engine, the dataset consists of 550 patches distributed across 14 crop classes.

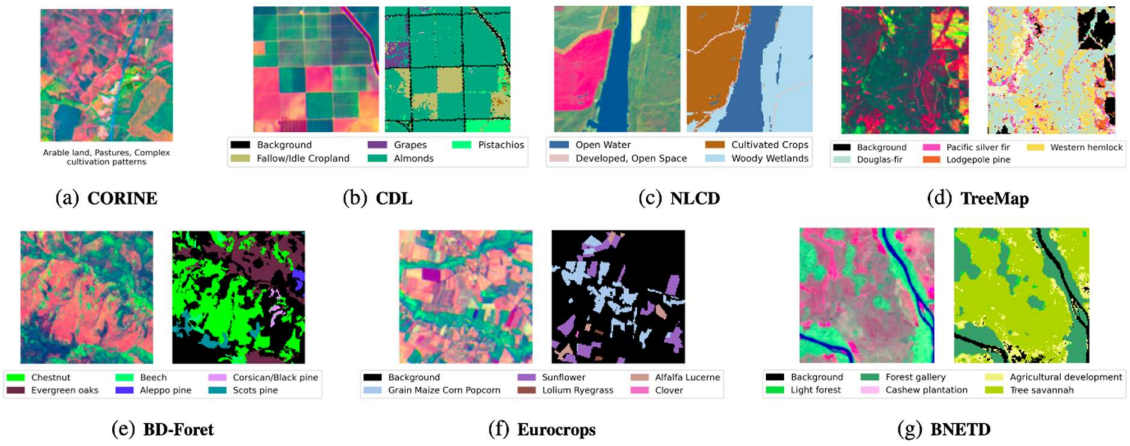


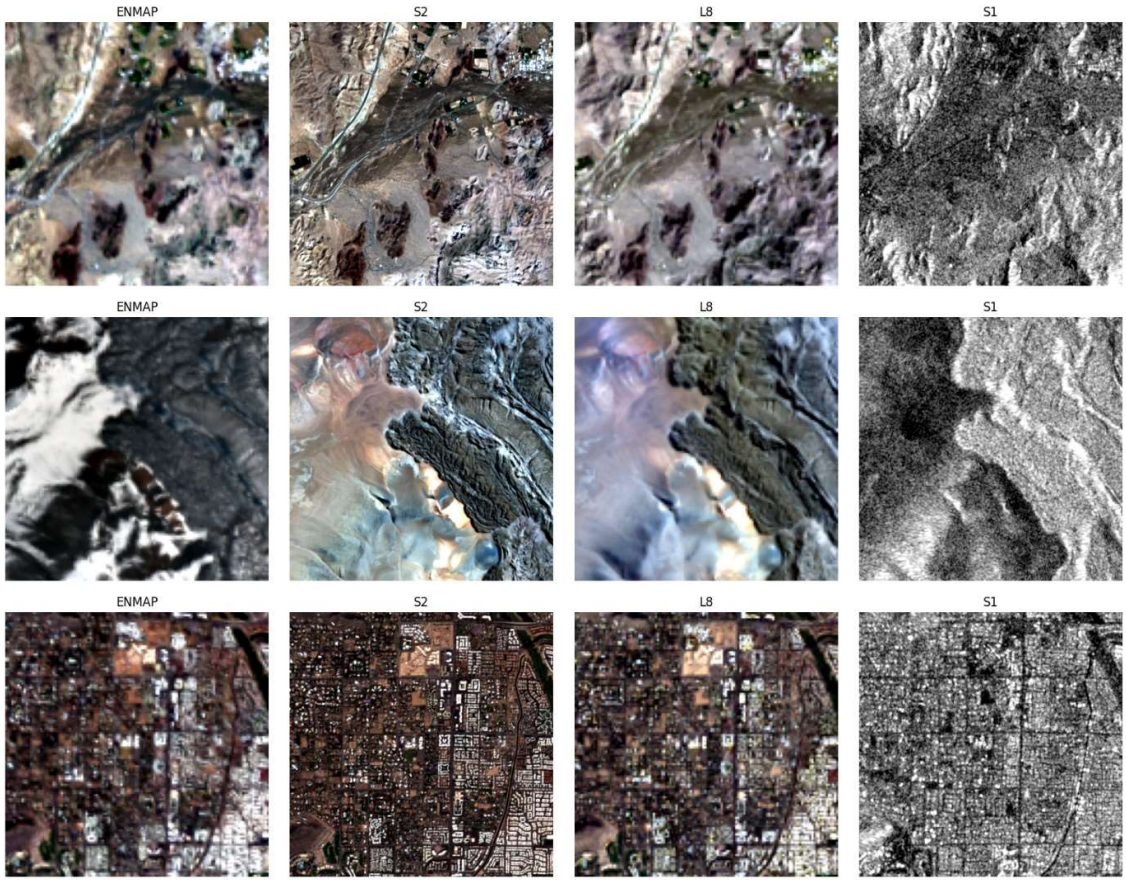
Figure 14. Illustration of various downstream tasks for SpectralEarth, from multi-label image classification to pixel-level semantic segmentation.

2.2.1.6 SpectralEarth-MM

SpectralEarth-MM extends SpectralEarth by pairing its hyperspectral EnMAP data with Sentinel-2 (S2), Landsat 8 (L8), and Sentinel-1 (S1) imagery. This dataset is designed to facilitate multi-sensor pretraining and assess cross-sensor transferability in remote sensing foundation models. A key objective is to investigate whether the higher spectral resolution of EnMAP can enhance representations learned from coarser multispectral sensors like Sentinel-2 and Landsat 8.

To construct SpectralEarth-MM, we matched EnMAP patches with S1, S2, and L8 images within a ±45-day temporal window, applying a cloud coverage threshold of 10% for optical sensors. This approach enabled the co-registration of 90% of EnMAP patches across all three additional sensors. The dataset provides collocated multi-sensor samples, allowing models to learn joint spectral-spatial representations and improve domain adaptation capabilities. Example collocated patches from different sensors are illustrated in Figure 15.

For the final phase of the project, we plan to explore multi-sensor pre-training using SpectralEarth-MM and assess the performance of the resulting models for a wide range of downstream tasks, along with their integration in WP3 prototypes.



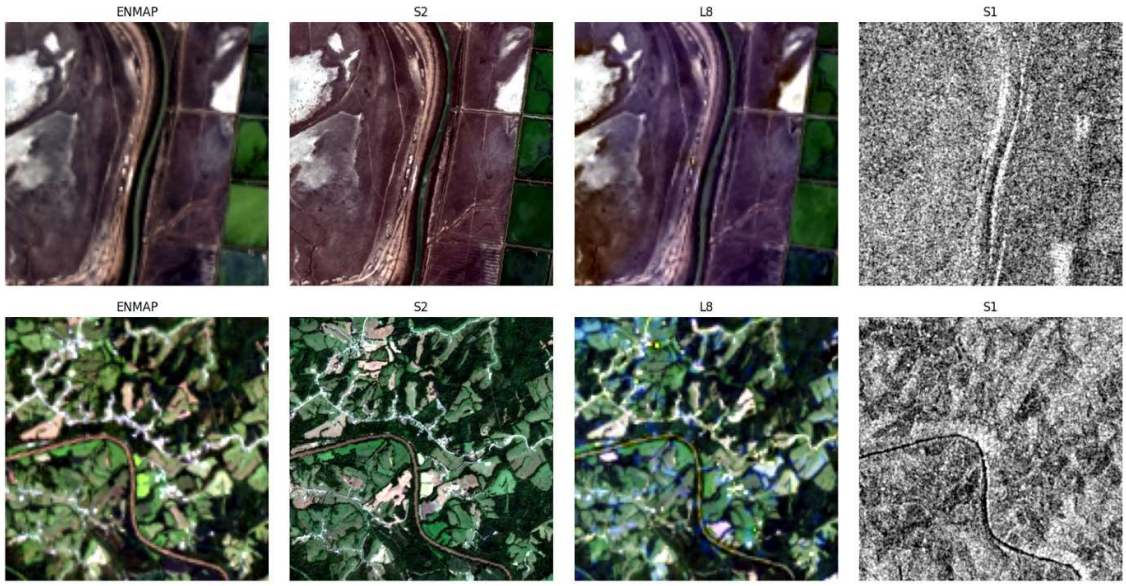


Figure 15. Samples from the SpectralEarth-MM dataset displaying EnMAP/S2/L8/S1 images for the same areas.

2.2.1.7 MMDC and Malice

To train and evaluate MMDC and Malice embedding algorithms, we have created several datasets. We divide them into pre-training (used to train embedding models) and DST (used to evaluate embedding models on DST) datasets.

2.2.1.7.1 Pre-training data

We have prepared two pre-training datasets named:

MMDC Single date (hereafter referred to as MMDC dataset),
Multi-temporal MMDC MALICE Auxiliary dataset (hereafter referred to as MMDC MALICE Aux dataset).

2.2.1.7.1.1 Base dataset MMDC EU

Both datasets were generated from MMDC-EU dataset. It consists of 12 training and 6 validation tiles spread over European Union with 4 years of continuous multi-modal data - from January 2017 to December 2020.

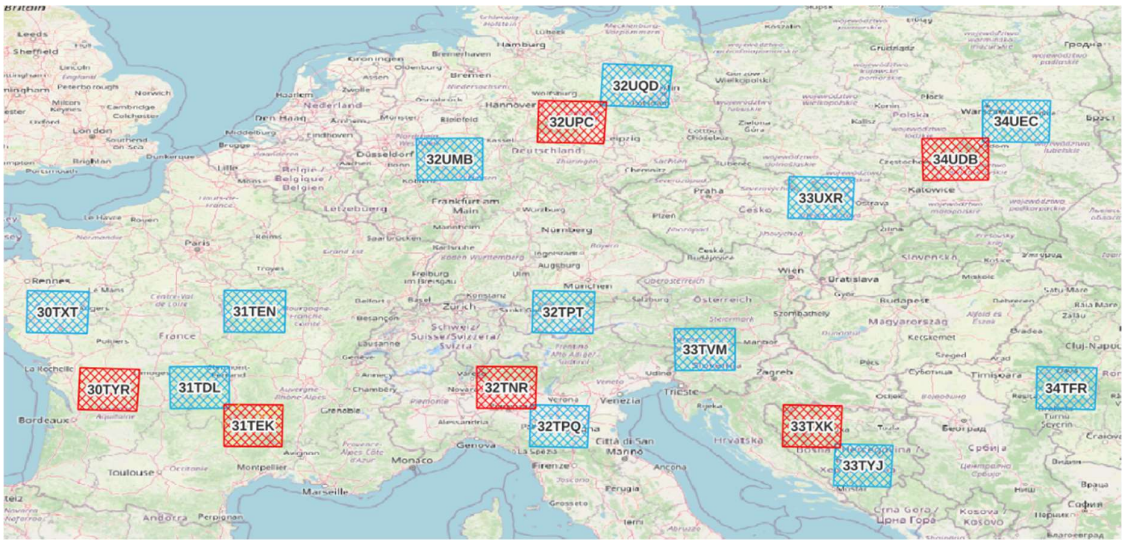


Figure 16. Tiles selected to train MMDC model. Note that we use only blue tiles present on this figure.

From each training tile, we randomly sample 10 regions of interest (ROIs), each measuring 512x512 pixels. In turn, 26 ROIs were sampled from each validation tile, each measuring 128x128 pixels. All the data is downloaded from OpenEO platform, the following modalities are present in the dataset:

Sentinel-2 images (Level 2A product, collection SENTINEL2_L2A). We select 10 spectral bands (B02, B03, B04, B05, B06, B07, B08, B8A, B11, B12), Cloud Mask (CLM) and Scene Classification Layer (SCL), as well as acquisition angles: Sun azimuth and zenith angles - θ_{S_a} and θ_{S_z} respectively, satellite mean view azimuth θ_{V_a} and zenith angles θ_{V_z}. All the bands and angles are resampled to 10m.
Sentinel-1 images (SENTINEL1_GRD product with noise removal corrected to γ_0-terrain). We download the images from both ascending and descending orbits (hereafter referred as S1 ASC and S1 DESC). We select both VV, VH polarization bands and local incidence angle θ_{inc} at 10m resolution.
Digital elevation model (DEM) (Copernicus Global 30 meter Digital Elevation Model dataset, COPERNICUS_30 collection) resampled to 10m.
AGERA5 Daily Meteorological data, originally at $0.1° \times 0.1°$ spatial resolution and resampled to 10m. We choose 8 variables:
dewpoint-temperature (2-meter dewpoint temperature, day average, Kelvin)
precipitation-flux (total precipitation over one day, mm)
solar-radiation-flux (Sum of Surface solar radiation downwards, J m-2)
temperature-max (Maximum air temperature at 2-meter, Kelvin)
temperature-mean (Average air temperature at 2-meter, Kelvin)
temperature-min (Minimum air temperature at 2-meter, Kelvin)
vapour-pressure (Average Vapour pressure, hPa)
wind-speed (average 10-meter wind component, m s-1)

2.2.1.7.1.2 MMDC Single Date

To train the MMDC single-date model, we generate a dataset named MMDC. We only use 12 training tiles from MMDC-EU dataset, as the size of validation tiles is insufficient for the task.

Dates selection: The main idea of MMDC model consists of finding common representations between Sentinel-1 and Sentinel-2 data. Therefore, to train the model, we need to select the dates where both Sentinel-1 and Sentinel-2 acquisitions are available. The dates are selected individually for each ROI. Sentinel-2 dates are used as reference, and S1 ASC and S1 DESC dates are matched to it. First, we pre-filter Sentinel-2 ROIs and keep only those with less than 90% of cloud coverage. Second, for each available Sentinel-2 date, we check if at least one of the two Sentinel-1 orbits is acquired on the same day (within ± 1 day of tolerance). All non-matched dates are discarded.

Finally, for each reference date t_i, we select 6-days weather time series $T_{w_i} \in [t_i - 4 : t_i + 1]$. We consider that the current weather series length is sufficient to represent a local climatic trend and is a good compromise for memory consumption. Note that the series length may be modified to assess the performance of different model configurations. Obviously, only one DEM image per ROI is available.

Data pre-processing: For each available date, each ROI is divided into four 256x256 non-overlapping patches. The following processing steps are applied to the patch data:

Sentinel-2: We keep 10 spectral bands and 4 acquisition angles from MMDC-EU dataset and transform the angles as follow:

$\cos(\theta_{S_z})$, $\cos(\theta_{S_a})$, $\sin(\theta_{S_a})$, $\cos(\theta_{V_z})$, $\cos(\theta_{V_a})$, $\sin(\theta_{V_a})$.

We create binary non-validity masks using Cloud Mask (CLM) and Scene Classification Layer (SCL) bands.

Sentinel-1: We keep both VV, VH polarization bands, and equally compute their ratio VH/VV as the supplementary third band. For each patch, we process ascending (S1 ASC) and descending (S1 DESC) orbits separately. The binary mask is created to indicate if only one orbit is present. All pixels are converted to the base-10 logarithm of their values. We select local incidence angle θ_{inc} as auxiliary data and convert it to $\cos(\theta_{inc})$. The binary validity masks are completed with non-valid pixels (saturated, NaN values, etc).

Note that, contrary to MALICE model with Sentinel-1 bands' order VH, VV, VH/VV, order in MMDC and MMDC MALICE Aux is VV, VH, VH/VV.

AGERA5 data: We keep raw values of 8 chosen variables. If NaN values are present and do not exceed two rows (or columns), we interpolate the neighbouring pixel values to replace them, otherwise, all modalities of the corresponding date are discarded.

DEM data: The original DEM data provides heights values $h_{i,j}$ per pixel. We compute slope and aspect as in (Horn, 1981).

For our dataset, we use 4 variables per pixel, such as
$h, \sin(slope), \cos(aspect), \sin(slope)$.

After all the modalities are processed, the patches with ratio of non-valid pixels (clouds, saturated and noisy pixels, NaN values, ...) superior to 15% within at least one modality are discarded.

2.2.1.7.1.3 MALICE and MALICE Aux

MMDC MALICE Aux dataset was produced to train MALICE Aux algorithm and based on the training dataset for Alise model and MMDC dataset. We discard S1~DESC modality and keep the rest of the modalities at their downloaded resolution.

We divide each training ROI of size 512x512 pixels into 16 sub-patches of 128x128 pixels, then, for each sub-patch, we randomly choose a timespan varying from 12 to 18 months and extract S1~ASC and Sentinel-2 SITS using all available images within this timespan. For Sentinel-2, we filter out images with more than 50% of clouds. For each day t_i of the SITS, we extract 6 days of weather data $T_{w_i} \in [t_i - 4 : t_i + 1]$, as in MMDC. To overcome the memory problem (both of data storage and dataloader during model training), we stock AGERA5 data in vector shape computing spatial mean per variable per day. Considering the coarse resolution of the original AGERA5 data of $0.1°$, we presume that the model performance will not be affected. A DEM is associated to each sub-patch.

Similar steps are done for the validation ROI tiles, except that the validation ROIs are smaller and, therefore, do not need to be clipped in sub-patches. However, to increase the variability of the validation data, we generate three different timespans of 12-18 months length and generate three different sets of time series for each validation ROI, contrary to only one for training dataset.

We select the same bands for each modality, as in MMDC dataset and pre-process the auxiliary data in the same way.

2.2.1.7.2 Downstream Datasets

2.2.1.7.2.1 Leaf-Area-Index (LAI) regression

To train LAI regression model, we use the same data as for model training. To avoid data imbalance, we select patches with abundant vegetation and discard urban

areas. The ground truth is produced from S2 images by a using a Sentinel Application Platform (SNAP) workflow.

2.2.1.7.2.2 Crop Classification (PASTIS)

To assess the performance of our embedding algorithms in a multi-temporal classification task, we utilize the Pastis-R dataset PASTIS-R dataset[5].

This dataset comprises 2,433 Sentinel-1 and Sentinel-2 SITS patches from the French metropolitan area (tiles 30UXV, 31TFJ, 31TFM, 32ULU), spanning from October 2018 to November 2019. Each patch includes pixel-wise annotations for 18 crop classes, along with one background and one void class. All the data is divided in 5 equally distributed folds.

In our work, we pre-train our embedding algorithms with the data downloaded from OpenEO. Its pre-processing level differ from Pastis-R dataset (we use γ_0-terrain values for Sentinel-1 data and Sen2Cor Level 2A correction). Therefore, we re-download Sentinel-2 and Sentinel-1 patches of Pastis-R dataset from OpenEO, keeping the same patch footprints, acquisition dates, and GT annotations as the original Pastis-R dataset.

While some of our embedding algorithms only require image data as an input, others require some auxiliary data. Therefore, we complete the dataset patches with AGERA5 6-days weather series and DEM data, as well as acquisition angles for Sentinel data.

The data pre-processing of the dataset patches (transformation of the acquisition angles to their $\cos$ and $\sin$, DEM slope computation, etc) is done in the same way as for the MMDC dataset.

2.2.1.7.2.3 Tree Cover Density (TCD) Regression

The tree cover density (TCD) is a part of the existing High Resolution Layer (HRL) Forest and HRL Vegetated Land Cover Characteristics (VLCC) of the Copernicus Land Service portfolio. To create a preliminary model, we start by working with a single tile. For a prototype, we choose 32TNT tile. We create our dataset by downloading all the modalities and auxiliary data from OpenEO platform. As in the original benchmark (VLCC 2018 TCD), we choose all S2 images available from March to October 2018 (leaf season) with cloud coverage less than 60%, plus four less cloudy images per month. However, contrary to the original benchmark, we also use Sentinel-1 data, exploiting all the available images for the given timestamp. The ground truth was provided by GAF partners in form of data-points, derived from HRL Forest TCD 2018. It is of high interest to further continue and improve the TCD product as part of the CLMS portfolio.

[5] https://github.com/VSainteuf/pastis-benchmark

2.2.1.7.2.4 Above Ground Biomass (AGB) and Forest Canopy Height (FCH) Regression

Due to important data cost, we only make the exhaustive analysis of MALICE S1 ASC embeddings on the above ground biomass (AGB) ant forest canopy height (FCH) regression task and, therefore, download only S1 ASC image data for the whole study area.

We process the image data based on the provided ground truth, downloading small patches corresponding to areas clipped around the available GT data.

The GT data is sourced from Sweden national forest inventory (NFI) with in situ measurements acquired from 2007 to 2021 all over Sweden. Each point of ground truth data is associated to a center of a circular plot of 154 m^2 and contains the variables such as mean plot tree height (in m), above ground biomass (t/ha) with or without foliage, mean stand age, mean stand density, and other forest descriptors.

We choose GT data from 2015 to 2021 year. Then we select corresponding points for each acquisition year, and we draw squared axis-oriented buffers of 50 pixels width. Usually, several GT points are located in the same area, therefore, we merge intersected buffers, transforming them into patches bounds of rectangular shape. The Corine Land Cover plus (CLC+) 2021 mask is applied to isolate forest GT points from other land cover areas. We choose the following classes: 2/Woody needle leaved trees, 3/Woody Broadleaved deciduous trees, 4/Woody Broadleaved evergreen trees, 5/Low-growing woody. Finally, for each patch boundary, we download a one-year SITS of Sentinel-1 ascending data.

2.2.2 Description of the candidate method

As mentioned earlier our Sentinel-2 models focus on image-level representations with single timestamps. We build our methodologies on well-established SSL algorithms and neural network architectures backbones to pre-train models for EvoLand. All developments are available as open-source code[6] as an open-source Python package utilizing PyTorch Lightning. Models have been defined in `evoland.model` and pretrained weights are either publicly available on HuggingFace through TorchGeo (Stewart et al., 2022), or will be shipped as serialized pickle files (Pilgrim, 2009) with the EvoLand deliverables by the end of the project.

[6] https://github.com/Evoland-Land-Monitoring-Evolution/evoland-software/tree/main/src/evoland

2.2.2.1 SSL4EO-S12: SSL on Sentinel-2

We perform pre-training using four well-established self-supervised methodologies (cf. section 2.1.2), each representing one popular type in self-supervision:

Contrastive:

- MoCo and its successors MoCo-v2 (X. Chen et al., 2020) and MoCo-v3 (Z. Xie et al., 2021) are contrastive method with explicit negative samples of embeddings Z optimizing the InfoNCE loss (Oord et al., 2018).
- DINO employs knowledge distillation without explicit negative sampling of embeddings Z.

Generative:

- MAE is a ViT-based autoencoder generating embeddings Z from image reconstruction.
- Data2vec (Baevski et al., 2022b) utilizes images with masked patches in a student-teacher Siamese network setup to generate contextualized embeddings Z.

We pre-train ResNet backbones with MoCo(-v2) and DINO, and ViT backbones for all four self-supervised methods listed above. To utilize multi-temporal information, we employ a *random season* data augmentation strategy, i.e., for MoCo and DINO, the input views are randomly picked from two seasons. For MAE and data2vec, one random season is assigned for each patch in every training epoch.

The SSL4EO-S12 models are integrated into TorchGeo:

- dataset[7]
- Sentinel-2 models[8]

2.2.2.2 DeCUR: Multi-Modal Contrastive Learning on Sentinel-1 & -2

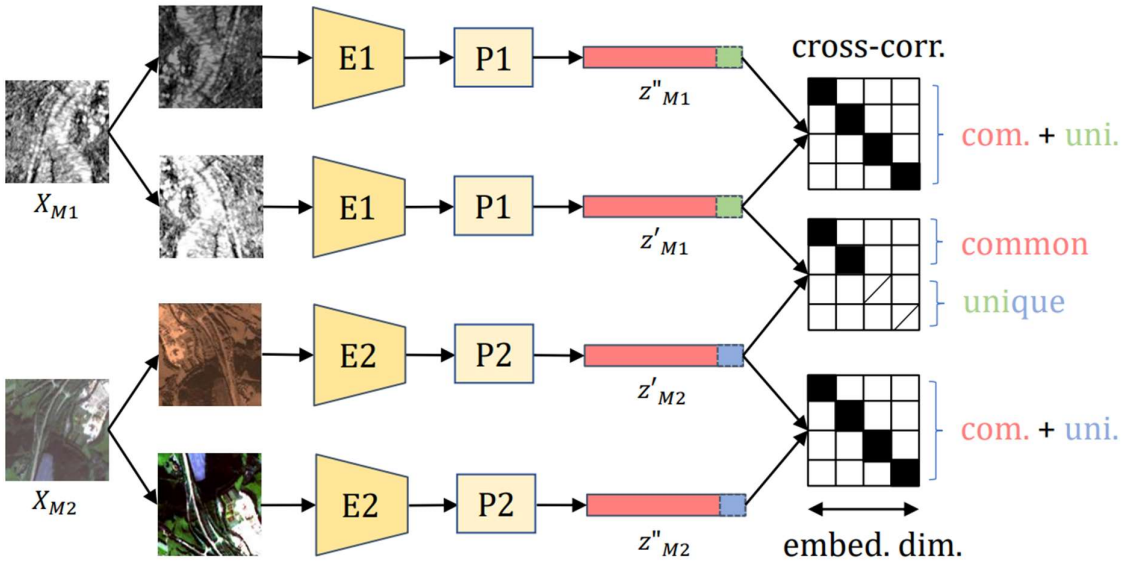
Most existing multimodal studies focus on common representations across modalities (Girdhar et al., 2023; Jain et al., 2022; Radford et al., 2021; Scheibenreif et al., 2022; L. Wang et al., 2021), while ignoring intra-modal and modality-unique representations. This forces the model to put potentially orthogonal representations into a common embedding space. Such approaches limit the model's capacity to better understand the information content contributed by each modality.

To tackle this challenge, DeCUR (*Decoupling Common and Unique Representations*) introduces the idea of decoupling common and unique

[7] https://torchgeo.readthedocs.io/en/stable/api/datasets.html#ssl4eo
[8] https://torchgeo.readthedocs.io/en/stable/api/models.html

representations. We maximize the similarity between common dimensions and decorrelate the unique dimensions across modalities as a model training objective. We also introduce intra-modal training on all dimensions, which optimizes for meaningful modality-unique dimensions. Furthers, the procedure enhances the model's ability to learn intra-modal knowledge.

=



.

Figure 17. The general structure of DeCUR: M1 and M2 represent two modalities from which an embedding feature vector Z is distilled. For EvoLand, Z identifies common and unique features in pairs of spatio-temporally aligned remote sensing imagery for Sentinel-1 & -2.

Technically, DeCUR serves as a multimodal extension of Barlow Twins. It performs self-supervised learning by redundancy reduction in the joint embedding space of augmented views from intra-/inter-modalities, see the illustration in Figure 17. Given a batch of multimodal input pairs X1 and X2, two batches of augmented views X1' and X1" (X2' and X2") are generated from each modality. Each of the four batches is then fed to a modality-specific encoder (E1 and E2) and projector (P1 and P2), producing batches of embeddings

- `Z1′ = P1(E1(X1′))` and `Z1″ = P1(E1(X1″))`
- `Z2′ = P2(E2(X2′))` and `Z2″ = P2(E2(X2″))`

Multimodal redundancy reduction is performed on the cross-correlation matrices of the embedding vectors, which is: Per definition (model hyperparameter), a subset of embedding dimensions serves as features ***common*** to both remote sensing modalities imaging the same, spatially aligned geospatial area. The remaining fraction of the embedding dimensions acts as ***unique*** set of features that are modality-specific and can be distilled from either encoder E1 or E2, only. The below details on the objective O to model such embeddings given a batch of spatially aligned images (cf. SSL4EO-S12 dataset, section 2.2.1.1).

Barlow Twins builds on the concept of cross-correlating embedding dimensions when augmented inputs X' and X" get encoded by E, i.e., it is computed the correlation matrix C from embeddings Z'=E(X') and Z"=E(X") with the training objective O to minimize |1-C| which forces E to diagonalize C=corr(Z',Z"). DeCUR extends this idea to multiple modalities: While the common dimensions of Z1' and Z2' have to correlate by definition, the unique embedding dimensions should decorrelate, i.e., we demand a novel training objective O1 where:

- `min |1 - C'|` for common embedding dimensions (correlation)
- `min |C'|` for unique embedding dimensions (decorrelation)

with `C'=corr(Z1', Z2')`.

Both objectives in O1 each carry a hyperparameter that determines the importance of (de)correlation (diagonal elements) and dimensional redundancy (off-diagonal elements). Training on the above objectives optimizes training parameters for both, E1 and E2. To avoid that the unique dimensions of the embeddings Z decorrelate entirely, we add the single-modality, standard BarlowTwins optimization objectives O2:

- `min |1 - corr(Z1', Z1")|` for modality X1 running through E1
- `min |1 - corr(Z2', Z2")|` for modality X2 running through E2

The source code of DeCUR[9] is publicly available as open-source with Apache-2.0 license.

2.2.2.3 SoftCon: Semantically Guided Contrastive Learning on Sentinel-2

Self-supervised learning methods typically leverage large-scale unlabelled imagery for representation learning. However, incorporating weak or noisy semantic labels—such as those from land cover products—can provide additional information to guide pretraining. In SoftCon, we rely on global landcover products in a self-weakly supervised fashion. The approach is built on two main components: (1) *soft contrastive learning* guided by land cover-derived multi-label supervision, and (2) *cross-domain continual pretraining* using strong vision foundation models. An overview of the method is illustrated in Figure 18**.**

[9] https://github.com/zhu-xlab/DeCUR

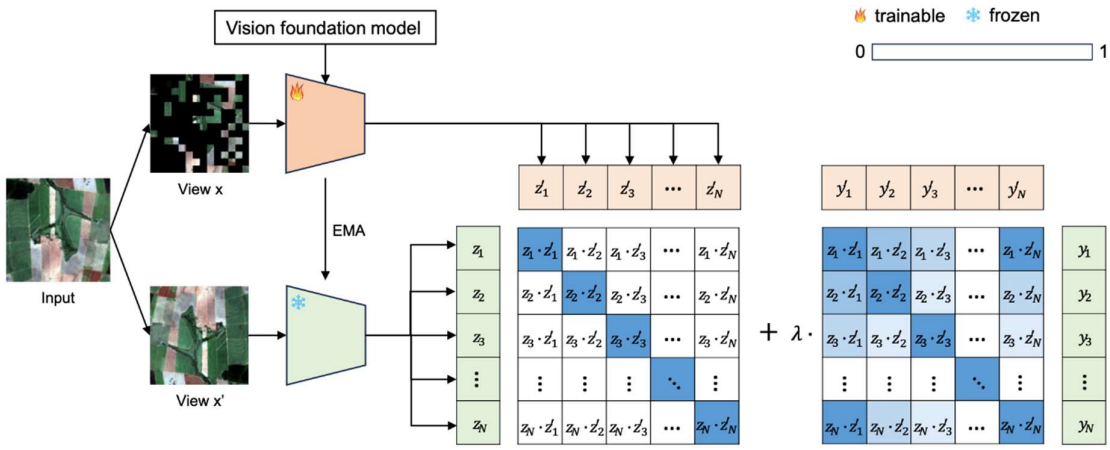


Figure 18. The general framework of SoftCon. We load vision foundation models for both the trainable and the momentum encoder when initializing the model. EMA represents exponential moving average. Siamese masking is only used with ViT backbones.

2.2.2.3.1 Multi-label Dataset

To enable label-informed SSL, it is available the global scene-level multi-label dataset *SSL4EO-S12-ML*[10]. It aligns seasonal Sentinel-1/2 SAR and multispectral images from SSL4EO-S12 with Dynamic World land cover maps. Matching is based on geolocation and timestamp metadata, yielding 780,371 images annotated with 9-class multi-labels at 10 m resolution.

2.2.2.3.2 Soft Contrastive Loss

SoftCon extends contrastive learning to accommodate multi-label supervision by introducing soft positive relationships based on label similarity. Cosine similarity between multi-hot label vectors defines a soft similarity matrix Y, which is used to guide the feature similarity matrix X through a binary cross-entropy loss. Based on semantic labels, this approach allows imagery in contrastive learning to be counted "similar" although not augmented views of each other.

Specifically, given a batch of N samples, we compute the label similarity matrix $Y \in \mathbb{R}^{N\times N}$ where each entry Y_{ij} represents the cosine similarity between the multi-hot label vectors y_i and y_j. Similarly, the feature similarity matrix $X \in \mathbb{R}^{N\times N}$ is obtained via the dot product of feature embeddings z_i and z_j, extracted from the model encoder.

The soft contrastive loss is formulated as:

[10] https://huggingface.co/datasets/wangyi111/SSL4EO-S12/tree/main/ssl4eo-s12-ml

$$L^{SoftCon} = -\sum_{i=1}^{N}\sum_{j=1}^{N}\left(Y_{ij}\cdot log\sigma(X_{ij}) + (1-Y_{ij})\cdot \log\left(1-\sigma(X_{ij})\right)\right)$$

And the overall loss we optimize is defined as:

$$L = L^{Contrast} + \lambda \cdot L^{SoftCon}$$

Where: $L^{Contrast} = -\sum_{i=1}^{N} \log \frac{\exp(z_i \cdot z_i'/\tau)}{\sum_{j=1}^{N}\exp(z_i \cdot z_j'/\tau)}$

2.2.2.3.3 Continual Pre-training

To improve efficiency and transferability, we introduced continual pretraining from strong natural image models (DINO for ResNet, DINOv2 for ViTs) to EO modalities. Input layers are randomly initialized to handle spectral differences. A *Siamese masking* strategy is applied during contrastive training, where only a subset of patches is visible to the encoder, reducing memory and compute requirements while acting as implicit augmentation.

2.2.2.4 CromSS: Semantic Segmentation by Weak Labels to Fuse Sentinel-1 & -2

CromSS (Cross-modal Sample Selection) is a weakly supervised pretraining strategy optimized for semantic segmentation. The loss to optimize the network enforces cross-modal consistency to mitigate label noise. Unlike conventional methods, CromSS utilizes noisy labels to improve the pre-training of segmentation models.

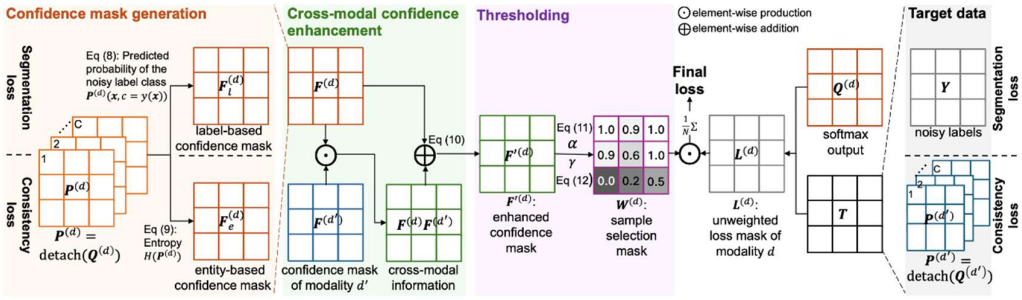


Figure 19. Two losses drive the optimization of a CromSS model: A semantic segmentation loss (top row) and a cross-modal consistency loss (bottom row, separated by a dashed line). Both are generated by three logical steps: pixel-wise confidence mask generation, cross-modal confidence enhancement, and thresholding for pixel weighting.

CromSS employs cross-modal learning to enhance feature representations and mitigate label noise through sample selection and spatial-temporal label smoothing (see Figure 19). The pretrained encoders can be transferred to various

segmentation tasks using different decoders. The individual components of the methods are described below.

2.2.2.4.1 Model Architecture

CromSS explores two fusion strategies (see Figure 20) to integrate information from Sentinel-1 SAR and Sentinel-2 multi-spectral data: *Late fusion*. Trains separate models for each modality and enforces consistency at the output level.
Middle fusion: Uses a shared decoder, imposing a stricter feature-level constraint across modalities.

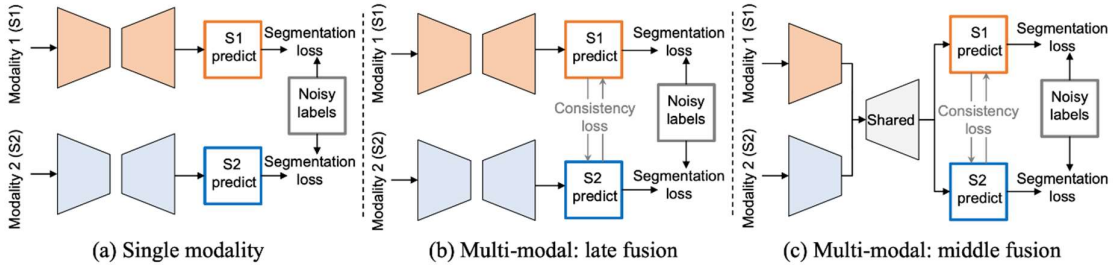


Figure 20. Comparison of single-modal training and multi-modal (middle/late) fusion strategies.

Both strategies incorporate segmentation losses for each modality and a consistency loss to align feature representations between the two. The segmentation loss combines cross-entropy (CE) loss and Dice loss, while the consistency loss is computed using Kullback–Leibler (KL) divergence to align predictions between modalities.

2.2.2.4.2 Sample Selection Strategy

To filter noisy labels, CromSS applies a three-step sample selection strategy. This process dynamically assigns weights to training samples based on their noise-level, ensuring that noisy labels are down-weighted without being fully discarded.

Confidence mask generation: Estimates label confidence using softmax probabilities for segmentation labels and entropy-based uncertainty for consistency losses.
Cross-modal confidence enhancement: Refines confidence estimates by incorporating information from both modalities. Samples with consistent predictions across modalities receive higher confidence scores.
Thresholding: Assigns selection weights based on confidence values. A dynamic soft weighting strategy ensures that low-confidence samples are not entirely discarded but contribute less to training.

This process ensures that only reliable labels contribute significantly to learning, improving robustness against noisy annotations.

2.2.2.4.3 Spatial-Temporal Label Smoothing

To further mitigate overconfidence and enhance robustness, CromSS introduces spatial-temporal label smoothing. Instead of conventional uniform label smoothing, it integrates spatial and temporal dependencies:

Spatial smoothing: Uses a Gaussian kernel to smooth labels based on neighbouring pixel information.
Temporal smoothing: Incorporates information from multiple seasonal observations by averaging labels across different timestamps before applying smoothing.

By leveraging spatial and temporal correlations, this approach reduces the impact of mislabelled samples and improves feature consistency across different time frames.

2.2.2.5 SpectralEarth: SSL on EnMAP Data

2.2.2.5.1 Pre-training Methods

We explore a broad range of SSL methods to establish a set of baselines pretrained models for hyperspectral remote sensing imagery. We utilize MoCo-v2 and DINO given their effectiveness in frozen encoder evaluation. Moreover, they are compatible with both CNN and ViT backbones. We also include MAE training for its established performance in fine-tuning scenarios. Together, these methods form a representative set of algorithms to investigate SSL in the hyperspectral domain.

2.2.2.5.2 Network Architectures

Building on the established efficacy of ResNet and ViT backbones in computer vision, we seek to adapt these architectures for hyperspectral imaging. A key challenge: standard ResNet and ViT architectures treat spectral bands independently, failing to capture the correlations between adjacent bands and the overall characteristics of the spectrum. Our architecture design addresses three critical requirements for hyperspectral imaging:

- **Spectral Feature Extraction:** We aim to efficiently extract features from the spectral domain. This involves capturing the unique spectral signatures and correlations among different bands necessary for hyperspectral data analysis.
- **Adaptability to Diverse Sensors:** The architecture must accommodate variability in the number of spectral bands. This adaptability is crucial when working with data from various sensors, each with distinct spectral properties.
- **Simplicity and Efficiency:** 3D convolution or tokenizing the image in the spectral dimension can be computationally expensive for large models, even

with moderately sized hyperspectral data cubes. We seek adaptation of classical architectures with minimal overhead.

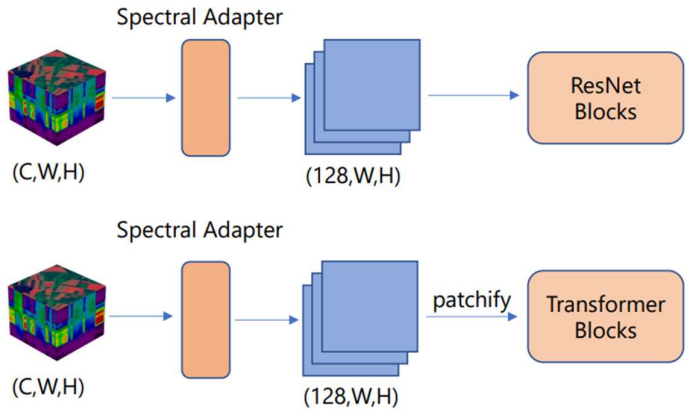


Figure 21. Illustration of SpectralEarth model architectures adapting ResNet and ViT to hyperspectral data input through a Spectral Adapter.

To meet these requirements, we implement a simple modular design, integrating 1D convolutional layers as a Spectral Adapter at the onset of the standard vision backbones, as depicted in Figure 21. These layers perform convolutions across the spectral dimension, generating new feature maps that feed into the core backbone. Specifically:

- **Spectral-ResNet:** We replace the stem layers with three 1D convolutional layers, allowing the network to process spectral information effectively before entering the residual blocks.
- **Spectral-ViT:** The 1D convolutions are placed prior to the patch embedding layer, ensuring that the transformer receives a spectrally enriched representation of the input.

This simple strategy results in an architecture that can be tailored to various hyperspectral sensors while leveraging a lightweight adapter for spectral feature extraction. The resulting models are capable of harnessing both the spatial and spectral information, while avoiding the extra cost incurred by 3D convolutions or spectral tokens in ViTs. These backbones will be pre-trained on SpectralEarth and fine-tuned on several downstream datasets with task-specific heads, as illustrated in Figure 22.

Source code and dataset of SpectralEarth will become publicly available once the corresponding paper under review has been accepted.

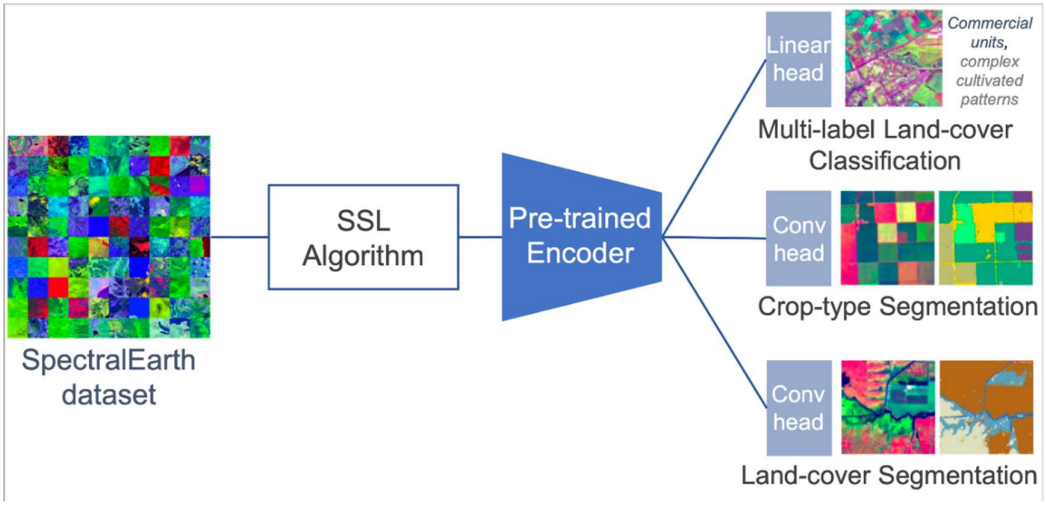


Figure 22. Schematic overview of the pre-training/fine-tuning pipeline. An encoder is pre-trained on SpectralEarth and adapted to the several downstream tasks with task-specific heads.

2.2.2.5.3 Multispectral to Hyperspectral with Pretrained Foundational Model

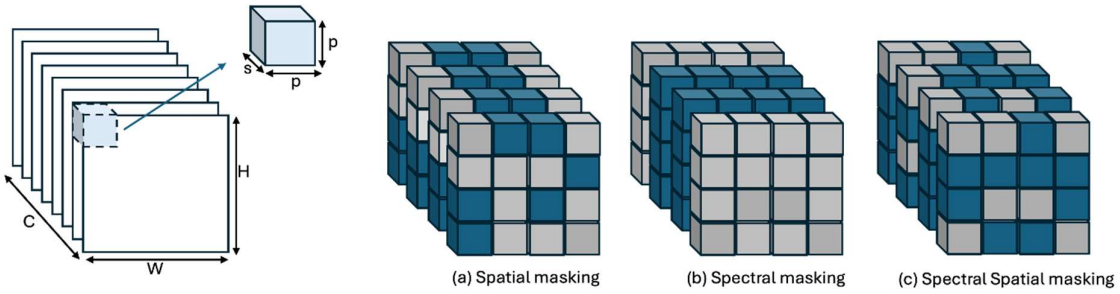


Figure 23. Visualization of masking strategies for spatial-spectral data cubes with C channels and WxH spatial pixels (left). From left to right most: (a) all channels of a given spatial pixel get masked out, (b) all spatial pixels of a (hyper)spectral channel gets masked, and (c) a (random) mix of spatial pixels get masked for each channel

In order to link hyper-spectral and multi-spectral models, we developed a self-supervised methodology that exploits the correspondence of spectral bands through their wavelength (ranges) assigned. In particular, in (Gonzalez et al., 2025) we utilize spectrally masked (Figure 23 (b)) autoencoders (MAE) with vision transformer (ViT) backbones to reconstruct EnMAP data (Figure 24). When fine-tuned, such models subsequently allow to reconstruct hyperspectral data from multi-spectral input such as provided by Sentinel-2 imagery.

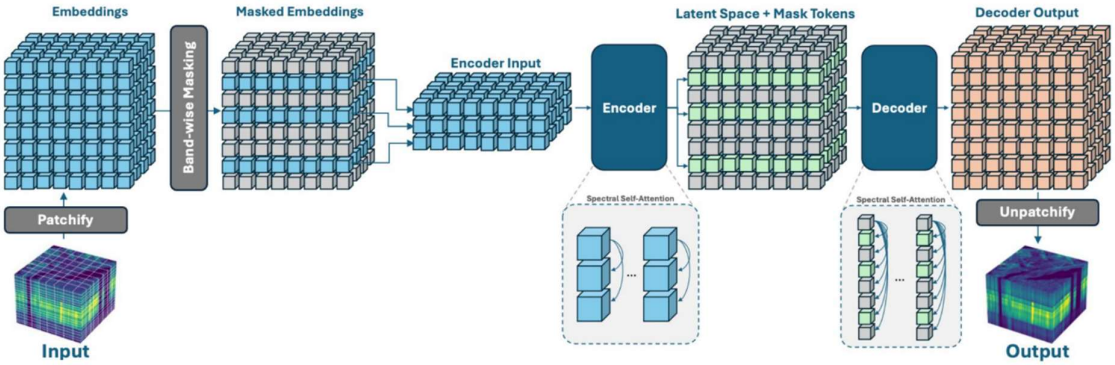


Figure 24. Masked Autoencoder (MAE) with Vision Transformer (ViT) for hyperspectral imagery as utilized in our approach. The hyperspectral data cube gets patchified into spatial-spectral data cubes that get spectrally masked for reconstruction.

2.2.2.6 MMDC: Single-date Sentinel-1 & -2 Model

MMDC model is able to extract common representations between Sentinel-1 and Sentinel-2 images exploiting the auxiliary data, such as meteorological variables and Digital Elevation Model (DEM).

While the MMDC model only exploits single timestamp data to produce latent representations, it serves as a base algorithm for other more complex models. Currently, it has been integrated into multi-modal and multi-temporal self-supervised algorithms, such as MALICE, able to extract spatio-temporal features from satellite image time series.

Moreover, the latent representations $Z = E(X)$ produced by MMDC model are not physically interpretable, therefore, the physical models might be integrated into MMDC model (Zérah et al., 2023), so the produced latent variables represent different phenological parameters, such as Leaf-Area-Index or Leaf Water Content.

The main objective of Multi-Modal Data Cube (MMDC) algorithm is to find common representations between Sentinel-1 and Sentinel-2 images and, therefore, improve the overall quality of downstream tasks, such as variables regression or landcover classification. It can be achieved by using the embedded representations instead of original images, or by making two images' sources – radar and optical – interchangeable.

To accomplish this task, we use a self-supervised encoder-decoder structure (see Figure 25), common for the embedding algorithms.

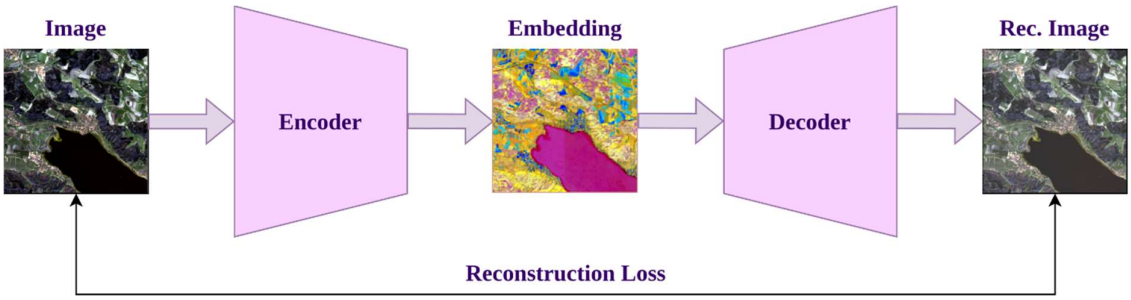


Figure 25. Encoder-decoder model. The input image is passed to the Encoder to produce the embeddings, which are then passer to the decoder to reconstruct the input image.

To train the model, the input data is passed to the encoder model to produce the embeddings, then the embeddings are passed to the decoder model to reconstruct the input data. The whole model is optimized to minimize the reconstruction loss. Once the desired model performance is achieved, we discard the decoding part and keep only the encoder to produce the image embeddings.

Our model deploys Variational AutoEncoder algorithm (Pu et al., 2016) that represents the embedded data as Gaussian distribution parametrized by its mean μ and variance σ. We will assess two different model configurations: baseline Figure 26 and experts' model Figure 27.

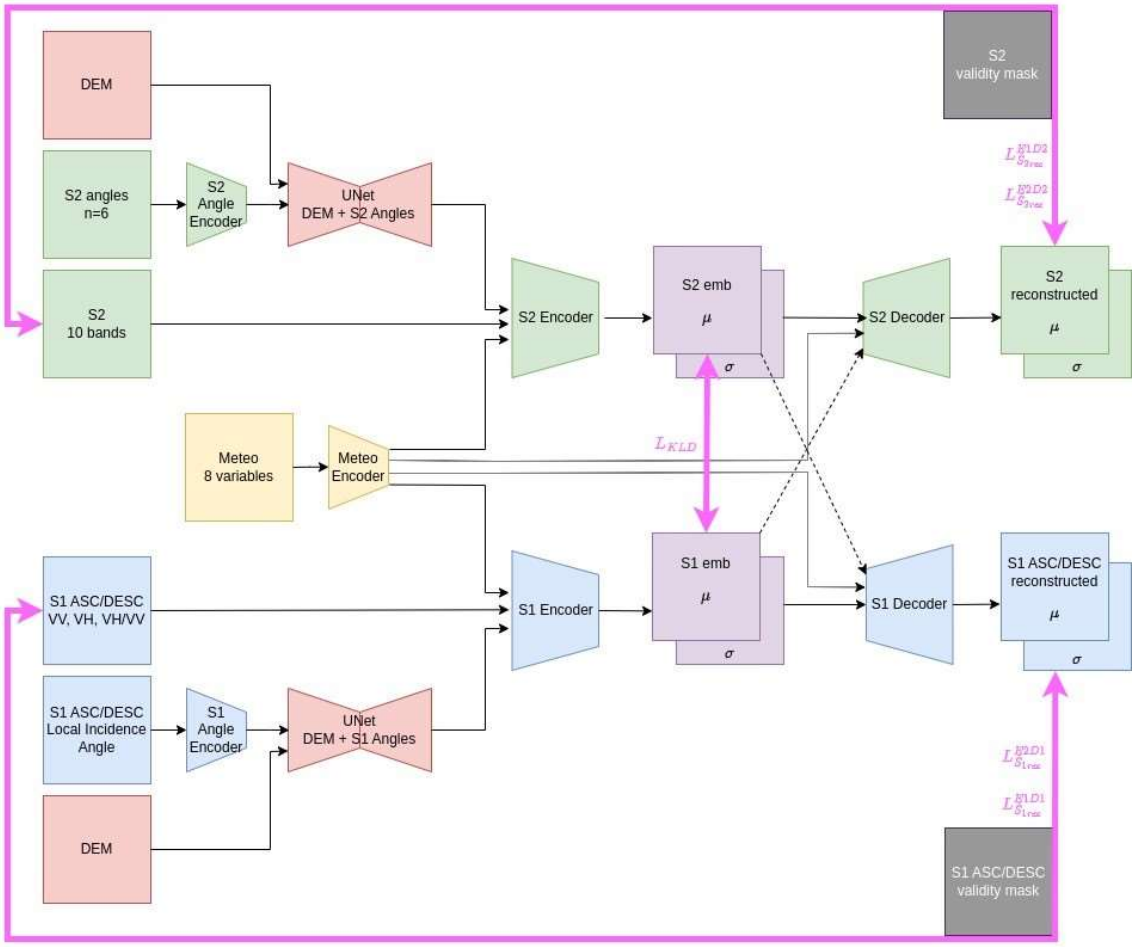


Figure 26. MMDC baseline. The model consists of two modality-specific encoders, as well as an encoder for weather data. Each modality-specific encoder first encodes the angles and DEM model, then this encoded data is concatenated with encoded weather data and image data and passed to the final encoding model to extract a modality-specific embedding. The decoding part consists of two modality-specific decoders, which use the produced embeddings to reconstruct the input data, in mono-modal and cross-modal way.

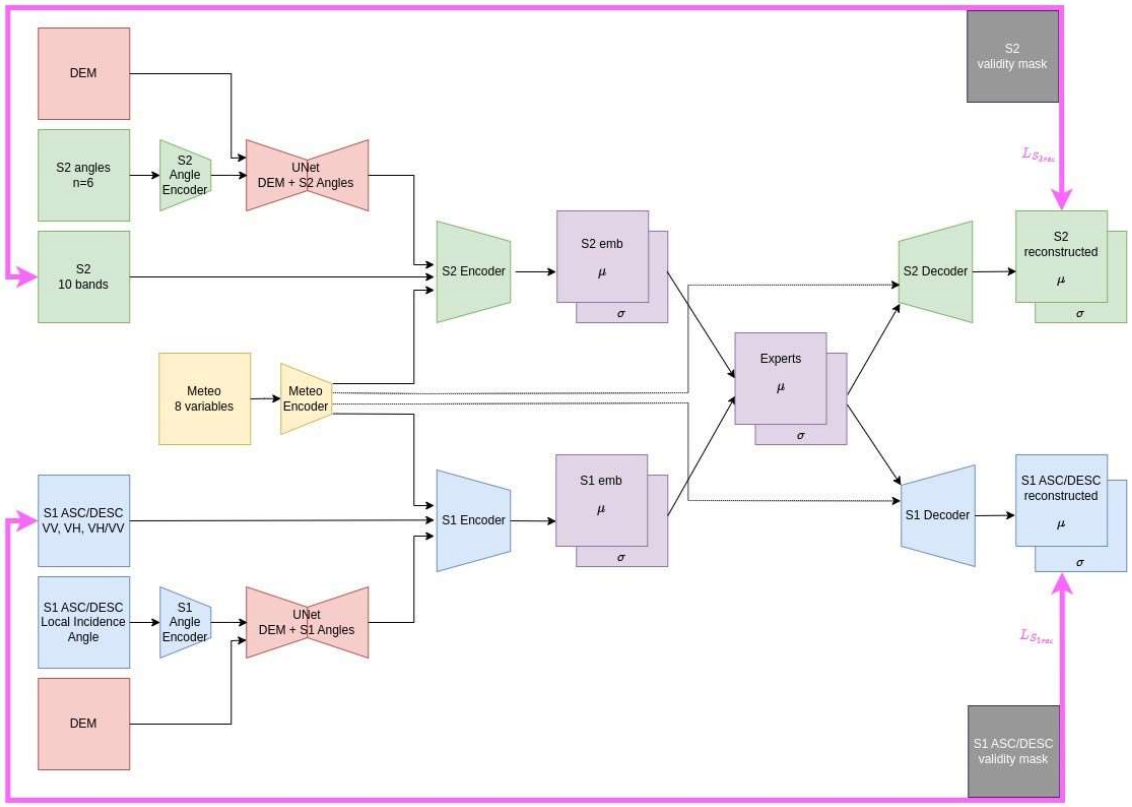


Figure 27. MMDC experts' model. The encoder is designed in the same way as in the baseline model and produces two modality-specific embeddings that are then fused using a chosen expert strategy. The decoder part, as in the baseline model, is composed of two modality-specific decoder which reconstructs the input data from the expert embeddings.

The encoder part of both baseline and experts models is the same: we use two different modality-specific encoder models E^{S1} and E^{S2} that process a pair of single-date imagesX^{S1}, X^{S2} and its corresponding auxiliary data, such as DEM D, angles A^{S1}, A^{S2} and weather data W. However, before being passed to each modality-specific encoder, the weather data is encoded by its proper encoder E^{W}. The same is done with angles and DEM data that are passed to their own modality-specific encoders E^{S1_a} and E^{S2_a}. Each encoder performs a two-step encoding process: first, it encodes the acquisition angles alone, followed by encoding the combination of the previously encoded angles and DEM data. Two encoding modules extract the probability distributions of two latent variables $Z^{S1} \sim N(\mu^{S1}, \sigma^{2\,S1})$ of same sizes for Sentinel-1 and Sentinel-2. Note, due to the algorithm design, all VAEs produce the *logvar* (logarithm of variance) in the latent space instead of σ^2, but we omit it for the simplicity in our explanations.

The encoder part can be written as follow:

$$p(Z^{S1}|X^{S1}, W, A^{S1}, D) \sim N(\mu^{S1}, \sigma^{2S1}) \sim E^{S1}\left(X^{S1}, W_{enc}, E^{S1_a}(A^{S1}, D)\right),$$

$$p(Z^{S2}|X^{S2}, W, A^{S2}, D) \sim N(\mu^{S2}, \sigma^{2S2}) \sim E^{S2}\left(X^{S2}, W_{enc}, E^{S2_a}(A^{S2}, D)\right)$$

with

$W_{enc} = E^W(W)$.

Both Sentinel-1 orbits are processes as a single input modality, with the corresponding orbit validity mask, so only one embedding for two orbits is obtained. The number of features of the embedded representation is chosen by user, however, the height and width remain the same, as for the input data. We choose to preserve spatial dimensions as most downstream tasks in remote sensing require a 2D map output.

For the **baseline**, the decoding part is the following: we reconstruct the Sentinel-1 image $X^{S1z^{S1}}$ from Sentinel-1 sampled latent variable z^{S1} by using the associated Sentinel-1 decoder D^{S1}. The image is reconstructed in form of probability distribution $p(\widehat{X^{S1}}|Z^{S1})$ with $\mu^{\widehat{X^{S1}}}$ and $\sigma^{2\,\widehat{X^{S1}}}$. We equally perform the cross-reconstruction by reconstructing Sentinel-1 image $X^{S1z^{S2}}$ from Sentinel-2 embedding by using the Sentinel-1 decoder. The similar procedure is used to reconstruct the Sentinel-2 image. The whole model is optimized by minimizing Sentinel-1 and Sentinel-2 mono-modal reconstruction losses ℓ_{rec_m}, cross-reconstruction losses $\ell_{re\ _m}$, as well as the invariance between two latent spaces (Symmetric Gaussian Kullback–Leibler divergence (SGKLD) loss) ℓ_{inv}. By using the cross-reconstruction and KLD losses, we impose to the embeddings to be similar to each other. The decoder can be formulated as follow:

$z^{S1} = \mu^{S1} + \epsilon\sigma^{2S1}$,

$z^{S2} = \mu^{S2} + \epsilon\sigma^{2S2}$,

$p(\widehat{X^{S1}}|Z^{S1}) = D^{S1}(z^{S1})$,

$p(\widehat{X^{S1}}|Z^{S2}) = D^{S1}(z^{S2})$,

$p(\widehat{X^{S2}}|Z^{S2}) = D^{S2}(z^{S2})$,

$p(\widehat{X^{S2}}|Z^{S1}) = D^{S2}(z^{S1})$,

where $\epsilon \sim N(0,1)$.

Finally, the global loss ℓ can be expressed as:

$$\ell = w_{inv}\ell_{inv} + w_{rec_m}\ell_{rec_m} + w_{rec_c}\ell_{rec_c}$$

With

$$\ell_{inv} = \text{SGKLD}\big(p(Z^{S1}|X^{S1}, W_{enc}, A^{S1}, D), p(Z^{S2}|X^{S2}, W_{enc}, A^{S2}, D)\big) \ell_{rec_m} =$$
$$\frac{1}{2}\Big(GaussianNLL\left(X^{S1}, p(\widehat{X^{S1}}|Z^{S1})\right) + GaussianNLL\left(X^{S2}, p(\widehat{X^{S2}}|Z^{S2})\right)\Big) \qquad l_{rec_c} =$$
$$\frac{1}{2}\Big(GaussianNLL\left(X^{S1}, p(\widehat{X^{S1}}|Z^{S2})\right) + GaussianNLL\left(X^{S2}, p(\widehat{X^{S2}}|Z^{S1})\right)\Big)$$

where w_{inv}, w_{rec}, w_{rec_m} are the weights applied to their corresponding loss values.

In turn, the **experts model** requires less optimization steps, as modality-specific embeddings are merged into producing an experts latent representation $p(Z|X^{S1}, X^{S2}, A^{S1}, A^{S2}, W, D)$. The expert representation can be computed by using product of experts (PoE - product of two Gaussian distributions) or mixture of experts (MoE – mixture of two Gaussian distributions).

After computing the experts, we reconstruct Sentinel-1 and Sentinel-2 images using modality specific decoders D^{S1}, D^{S2} from experts' latent representation. The whole model is optimized by using only Sentinel-1 and Sentinel-2 reconstruction losses and the invariance loss between three latent spaces. We equally minimize Kullback-Leibler divergence between a monovariate Gaussian $p(Z)$ and $N(0,1)$.

We can express this as follow:

$$p(Z|Z^{S1}, Z^{S2}) = \xi\big(p(Z^{S1}|X^{S1}, W, A^{S1}, D), p(Z^{S2}|X^{S2}, W, A^{S2}, D)\big),$$

$$z = \tau\big(p(Z|X^{S1}, X^{S2}, A^{S1}, A^{S2}, W, D)\big),$$

$$p\big(\widehat{X^{S1}}|Z\big) = D^{S1}(z),$$

$$p\big(\widehat{X^{S2}}|Z\big) = D^{S2}(z),$$

where $\xi(\ \ )$ and $\tau(\ \ )$ are experts and random sampling functions that depend on the chosen expert strategy.

The global loss can be expressed as follow:

$$\ell = w_{inv}\ell_{inv} + w_{rec}\ell_{rec}$$

with

$$l_{inv} = KLD\big(p(Z^{S2}), p(Z^{S1})\big) + KLD\big(p(Z), p(Z^{S1})\big) + KLD\big(p(Z), p(Z^{S2})\big) + KLD\big(p(Z),\ N(0,1)\big)$$

$$l_{re\ m} = \frac{1}{2}\Big(GaussianNLL\left(X^{S1}, p(\widehat{X^{S1}}|Z)\right) + GaussianNLL\left(X^{S2},\ p(\widehat{X^{S2}}|Z)\right)\Big),$$

Moreover, to improve the quality of embeddings for vegetation-related DST, we add LAI loss that is expressed as root Mean Squared Error between LAI computed from the input Sentinel-2 image and the reconstructed one.

Note, that for the DST tasks, we can use only mean values of latent space, or both mean and variance.

Experts models

In general, the experts models can be expressed as follow: We consider a dataset $\{X^{(i)}\}_{i=1}^{N}$ of N samples, each of which is a set of M modalities $X^{(i)} = \{x^{(i)}\}_{j=i}^{M}$.

In the VAE setting, we use an encoder (a NN with parameters Φ) to produce $q_\Phi(z|X^{(i)})$ as an approximation of $p(z|X^{(i)})$.

With one encoder per modality, we have M estimates of $q_\Phi(z|X^{(i)})$ that we denote $q_\Phi^j\left(z\middle|X_j^{(i)}\right)$ where $j \in [1, M]$ indicates each of the modalities of input data.

To get a single latent space from those modalities, several solutions have been proposed:

MVAE (Wu et al., 2018) (Product of Experts, PoE) with $\widetilde{q_\Phi}(z|X^{(i)}) \propto \prod_{j=1}^{M} \widetilde{q_\Phi^j}\left(z\middle|X_j^{(i)}\right)$

MMVAE (Shi et al., 2019) (Mixture of Experts, MoE) with $\widetilde{q_\Phi}(z|X^{(i)}) \propto \frac{1}{M}\sum_{j=1}^{M} \widetilde{q_\Phi^j}\left(z\middle|X_j^{(i)}\right)$

The MVAE uses a geometric mean, which enables learning a sharp posterior, resulting in a good approximation of the joint distribution. On the other hand, the MMVAE applies an arithmetic mean which allows better learning of the unimodal and pairwise conditional distributions.

As in our case, as both distributions are Gaussian, *the product of experts* can be expressed as follow:

$$\sigma_{PoE}^{-2} = \sigma_{S1}^{-2} + \sigma_{S2}^{-2},$$

$$\mu_{PoE} = \sigma_{PoE}^{2}\left(\frac{\mu_{S1}}{\sigma_{S1}^{2}} + \frac{\mu_{S2}}{\sigma_{S2}^{2}}\right)$$

The Mixture of Experts, however, can be expressed as a mixture of two Gaussian distributions with equal weights.

Data Masking

To improve the robustness of the model, we equally propose to progressively mask the input satellite data.

For the baseline model, the masking process is the following: we pre-train the model until it is converged, then, we start to randomly mask a small number of entire input images bands, progressively increasing their number over time until a

certain threshold is reached. To mask a band, we replace its original pixel values with Gaussian noise, using precomputed statistics with dataset mean and standard deviation values. Note that we do not take the type of satellite into account while masking, so the bands are randomly chosen from both Sentinel-1 and Sentinel-2 ascending and descending orbits. Bands from different satellites can be masked at the same time.

The masking of the experts model starts in the same way, although, when we reach a certain number of masked bands (usually 50%), we start randomly masking only one input modality within a patch - Sentinel-1 or Sentinel-2 - progressively increasing the number of masked bands until it reaches 100% within one satellite. We continue the training while masking an entire modality within a patch until the model converges. Note, that S1 ASC and DESC orbits are considered as one satellite modality.

2.2.2.7 MALICE and MALICE Auxiliary - Spatio-Temporal Sentinel-1 and Sentinel-2 models

MALICE Auxiliary algorithm is based on multi-modal SITS MALICE algorithm developed by Iris Dumeur during her PhD. In turn, MALICE is based on mono-modal SITS embedding algorithms: ALISE and U-Barn (Dumeur et al., 2024b) and, therefore, we are not going to detail the inherited modules in this document.

2.2.2.7.1 MALICE

Here we give a simplified description of MALICE algorithm, in particular, the novelties compared to ALISE model. The training and inference processes of MALICE differ significantly. To provide a more intuitive understanding, we begin by introducing the inference process first.

The inference process is based on the MALICE encoder, the latter can be resumed as follow. MALICE's encoder architecture is a two-branch neural network (Figure 28) where each branch processes a single modality by using spatial, spectral and temporal encoder (SSTE), which corresponds to the proposed U-BARN architecture (Dumeur et al., 2024), and a temporal projector. The temporal projector produces aligned representations thanks to a cross-attention mechanism with learnable queries. Although the SSTE weights differ for each modality, a common temporal projector is shared between both branches. Hence, we consider that the two modalities are aligned on the same latent temporal features. The modality-specific encoders are denoted by $ALISE^{S2}$ and $ALISE^{S1}$.

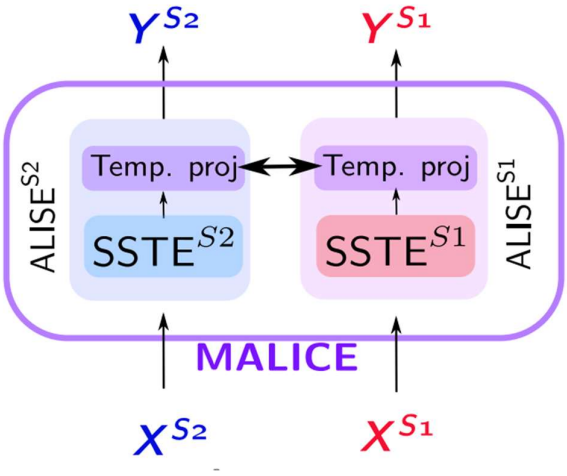


Figure 28. MALICE encoder architecture. MALICE is composed of two branches each associated to one modality. The branches have different SSTE, but they share the same temporal projector.

As it was mentioned earlier, the training strategy differ from the inference. To optimize the model, we use invariance loss ℓ_{inv}, as well as mono-modal ℓ_{rec_m} and cross-modal ℓ_{rec_c} reconstruction losses, almost in the same manner as in MMDC without experts:

$$\ell = w_{inv}\ell_{inv} + w_{rec}\left(w_{rec_m}\ell_{rec_m} + \left(1 - w_{rec_m}\right)\ell_{rec_c}\right),$$

where $w_{inv}, w_{rec}, w_{rec_m}$ are the weights applied to their corresponding loss values, and ℓ is the global loss.

However, to impose better invariance between generated modalities, we use multi-modal multi-view self-supervised learning (SSL) strategy. The main idea of this approach can be resumed as follow:

Given two input time series X^{S1} and X^{S2} corresponding to Sentinel-1 and Sentinel-2 SITS acquired over the same ROI, we divide each of them into views X_A^{S1}, X_B^{S1}, X_A^{S2}, X_B^{S2}.

The views correspond to alternative ways of observing the input data. Different views can be, for example, acquisitions from different sensors or artificially augmented versions of the input data. In our work, *views* refer to two sub-series X_A^i and X_B^i, extracted from the same time series X^i of modality i, but corresponding to different observation dates. The views are non-overlapping within the same modality and have the equal length. For instance, consider a time series X^i with acquisition dates $T = [t_1, \ldots, t_6]$. Two possible views could be X_A^i with dates $T_A = [t_1, t_3, t_5]$ and X_B^i with dates $T_B = [t_2, t_4, t_6]$. As demonstrated in ALISE, employing a specific date selection strategy does not yield significant performance improvements. Therefore, the dates for the views are randomly selected from the available dates, and the example provided is not an exhaustive representation of possible combinations.

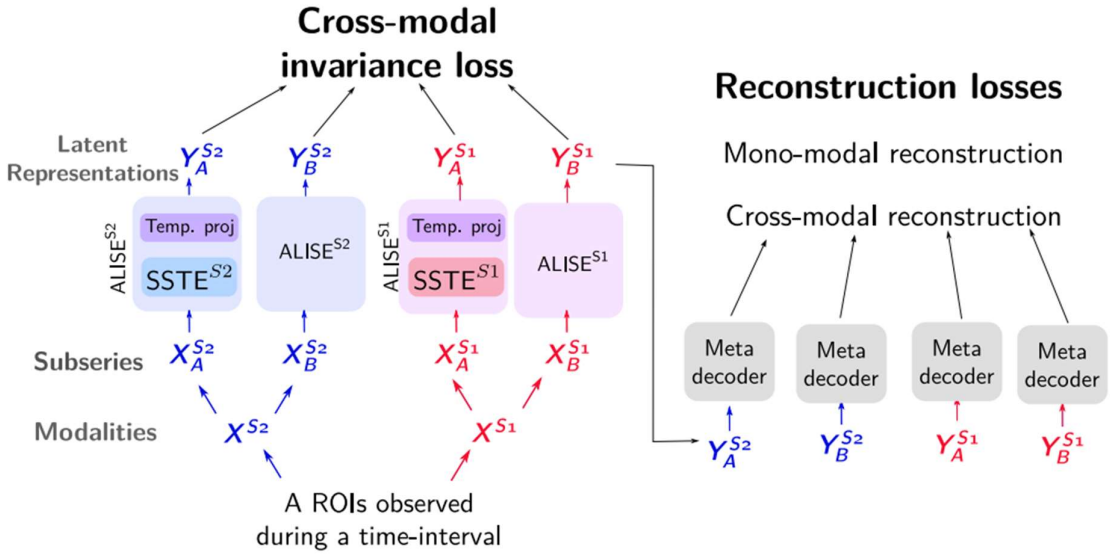


Figure 29. MALICE pre-training overview. Discriminative losses as well as reconstruction losses are considered for pre-training.

Each view is passed to a corresponding modality-specific encoder $ALISE^i$ producing 4 latent representations $Y_A^{S1}, Y_B^{S1}, Y_A^{S2}, Y_B^{S2}$, see Figure 29.

$$Y_A^{S1} = ALISE^{S1}(X_A^{S1}),$$

$$Y_B^{S1} = ALISE^{S1}(X_B^{S1}),$$

$$Y_A^{S2} = ALISE^{S2}(X_A^{S2}),$$

$$Y_B^{S2} = ALISE^{S2}(X_B^{S2}).$$

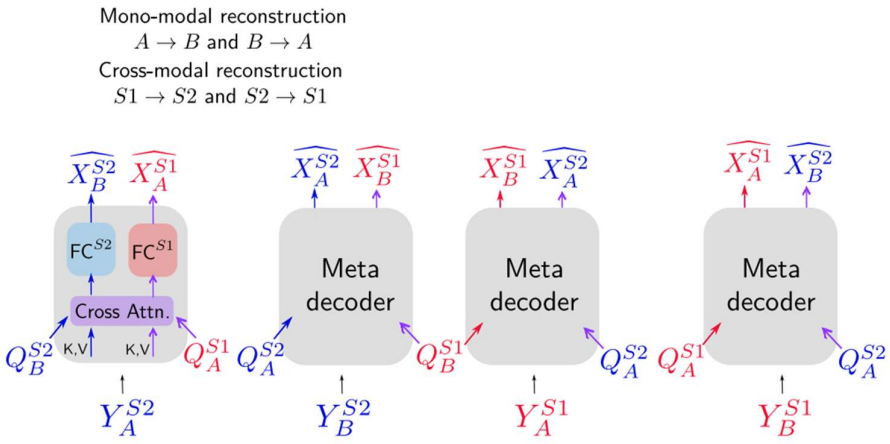


Figure 30. Description of the reconstruction SSL losses employed to pre-train MALICE. Reconstruction losses are either mono-modal or cross-modal. The metadecoder performs a cross-attention between the latent representations Y (used as keys and values) and the queries Q which contain the

information about the modality as well as the dates to be reconstructed. Superscripts refer to the modality S1 or S2 and subscripts to the sub-series A, B.

The mono-modal and cross-modal reconstructions are achieved by implementing a meta-decoder, see Figure 30. A similar architecture to that of the decoder used for ALISE pre-training, which exclusively processes pixel time series, is proposed. Notably, it relies on a cross-attention mechanism between the latent representation (involved as keys K and values V) and queries which contain the information of the date to reconstruct. The cross-attention module is common for both modalities and is followed by two modality-specific Fully-Connected (FC) layers FC^{S1} and FC^{S2}. These layers enable to output the required number of spectral bands of the modality to reconstruct. Compared to ALISE, to enable the reconstruction of different modalities, MALICE meta-decoder proposes two important modifications:

The construction of the queries involved in the cross-attention is modified. Notably, two learnable tokens $m_{S2} \in \mathbb{R}^{d_{tok}}$ and $m_{S1} \in \mathbb{R}^{d_{tok}}$ are employed, each associated to the modality to reconstruct. d_{tok} is a hyper-parameter controlling the size of the learnable tokens. In addition, instead of summing the Positional Encoding (PE) to the learnable token, m_{S2} (resp. m_{S1}) is concatenated to the positional encoding of the series to reconstruct. The positional encoding function is similar to the one employed in ALISE.

After the cross-attention layer, two modality-specific fully-connected layers (FC^{S2} and FC^{S1}) are employed (see Figure 30). These layers enable to output the required number of spectral bands of the modality to reconstruct.

In our work, to overcome the memory limitations, we propose to compute the following combinations of mono- and cross-modal reconstructions:

$$\widehat{X_A^{S1}} = FC^{S1}\left(Meta\left(Y_B^{S1}\right)\right),$$

$$\widehat{X_B^{S1}} = FC^{S1}\left(Meta\left(Y_A^{S1}\right)\right),$$

$$\widehat{X_A^{S2}} = FC^{S2}\left(Meta\left(Y_B^{S2}\right)\right),$$

$$\widehat{X_B^{S2}} = FC^{S2}\left(Meta\left(Y_A^{S2}\right)\right),$$

$$\widehat{X'^{S1}_A} = FC^{S1}\left(Meta\left(Y_A^{S2}\right)\right),$$

$$\widehat{X'^{S1}_B} = FC^{S1}\left(Meta\left(Y_B^{S2}\right)\right),$$

$$\widehat{X'^{S2}_A} = FC^{S2}\left(Meta\left(Y_A^{S1}\right)\right),$$

$$\widehat{X'^{S2}_B} = FC^{S2}\left(Meta\left(Y_B^{S1}\right)\right).$$

Finally, the losses components can be computed as follow:

$$\ell_{inv} = \frac{1}{2}\left(\rho\left(Y_A^{S1}, Y_A^{S2}\right) + \rho\left(Y_B^{S1}, Y_B^{S2}\right)\right), \widetilde{X^{S1}}$$

$$\ell_{re\ m} = \frac{1}{4}\left(\rho\left(\widehat{X_A^{S1}}, \widetilde{X_A^{S1}}\right) + \rho\left(\widehat{X_B^{S1}}, \widetilde{X_B^{S1}}\right) + \rho\left(\widehat{X_A^{S2}}, X_A^{S2}\right) + \rho\left(\widehat{X_B^{S2}}, X_B^{S2}\right)\right),$$

$$\ell_{re\ c} = \frac{1}{4}\left(\rho\left(\widehat{X'^{S1}_A}, \widetilde{X_A^{S1}}\right) + \rho\left(\widehat{X'^{S1}_B}, \widetilde{X_B^{S1}}\right) + \rho\left(\widehat{X'^{S2}_A}, X_A^{S2}\right) + \rho\left(\widehat{X'^{S2}_B}, X_B^{S2}\right)\right),$$

where $\rho(\ \ )$ is the root Mean-Squared Error loss function.

Furthermore, it must be noticed that the Sentinel-1 ground truth data does not correspond to the input data X^{S1}, but to the Sentinel-1 data filtered with a speckle filter, named $\widetilde{X^{S1}}$. Indeed, without this speckle filter (J.-S. Lee, 1986), the Sentinel-1 data is highly noisy, and using it as-is, would significantly penalize the reconstruction loss, even when the reconstructions themselves are accurate.

As it can be seen from the equation, in the original version of the model, the invariance loss is computed in a cross-modal way. For the ablation study, we equally propose to deploy a same modality invariance loss:

$$\ell_{in\ same} = \frac{1}{4}\left(\rho\left(Y_A^{S1}, Y_A^{S2}\right) + \rho\left(Y_B^{S1}, Y_B^{S2}\right) + \rho\left(Y_A^{S1}, Y_B^{S1}\right) + \rho\left(Y_A^{S2}, Y_B^{S2}\right)\right).$$

2.2.2.7.2 MALICE with Auxiliary data

As mentioned earlier, MALICE Aux is based on MALICE algorithm, the main difference lies in the integration of the auxiliary data: AGERA5 time series, acquisition angles, and DEM. As in MMDC, for each acquisition date of Sentinel-1 and Sentinel-2 SITS we add a corresponding 6-day time series of weather data , as well as the associated DEM. Unlike MMDC, the weather data loses spatial dimension and instead represents mean spatial value per day per variable. Its shape is per element of the batch. Contrary to MMDC, the acquisition angles are concatenated with the spectral/backscatter band values and do not have angle-specific encoders.

To integrate the auxiliary data to MALICE model, we modify SSTE module of the encoder, while the temporal projector and decoder module stay intact. Figure 31 shows the modified SSTE module that we name SSTE Aux. As in MALICE, SSTE Aux module is proper to each modality.

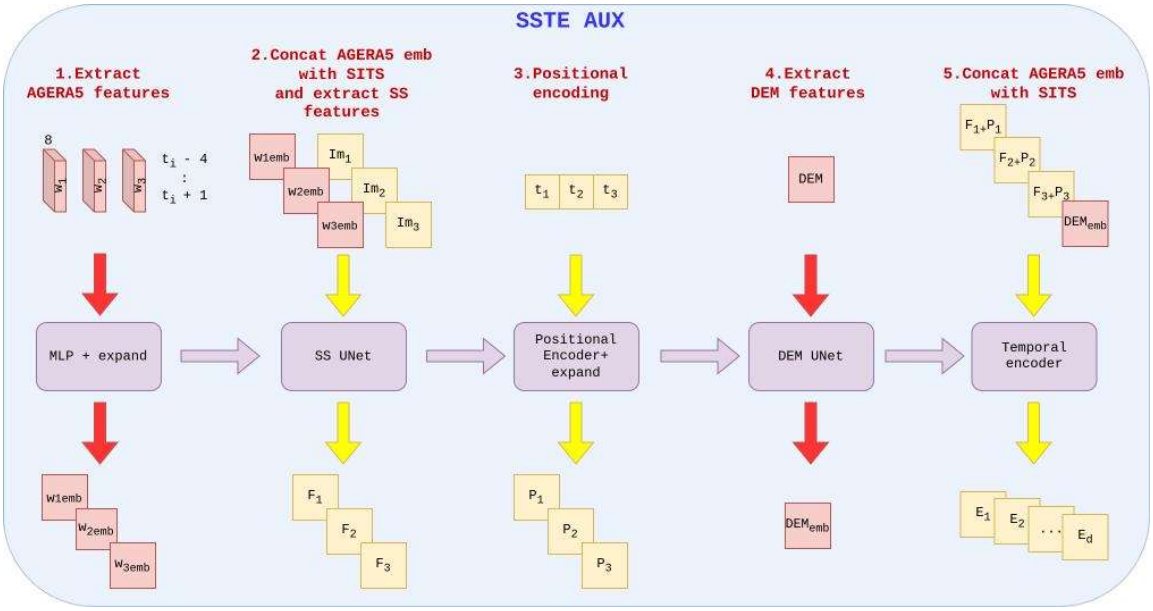


Figure 31. SSTE Aux module. Each module is modality-specific, except for the DEM encoder that is shared by the modalities. In red we show the newly added data and operations, compared to the original SSTE module. We start by 1/encoding the AGERA5 data. Then 2/we concatenate it with the SITS data that is passed to the SS module. 3/The positional encoding of the acquisition dates is done. 4/DEM features are encoded by a DEM UNet module. Finally, 5/those features are concatenated in temporal dimension with the sum of encoded SITS and PE, and passed to the temporal encoder.

The basic SSTE module consists of UNet module for Spatio-Spectral feature extraction of individual images of SITS, positional encoding of DOYs and a temporal encoder that process the sum of image features and encoded DOY.

We describe the SSTE Aux module as follow. Given the input data of shape:

> SITS: $T \times C_{sa\ i} \times H \times W$, where T is the length of a modality-specific time series, H and W are their height and width, $C_{sa\ i}$ is the number of channels of the corresponding modality (includes image bands and acquisition angles). Note that here we omit the batch dimension for the simplicity;
> DEM: $C_{DEM} \times H \times W$, where C_{DEM} is the number of channels of DEM data;
> AGERA5 series: $T \times T_w \times C_w$, where T_w is the length of each weather mini-series, here $T_w = 6$, C_w is the number of weather variables, here $C_w = 8$,

we proceed with the following steps:

> We encode AGERA5 time series with an MLP network, each 6-days mini-time series is encoded separately to preserve the temporal dimension T, resulting in the output shape $T \times C_{w_{emb}}$. We spatially extend the obtained features into $T \times C_{w_{emb}} \times H \times W$ by repeating the same feature value across the spatial dimensions H and W.
> We concatenate spatially expanded AGERA5 features with their corresponding SITS on the channel dimension and pass the new series of

shape $T \times (C_{sat_i} + C_{w_{emb}}) \times H \times W$ to the SS UNet block. We extract SS features for each individual date of T, resulting in shape $T \times C_{im_{emb}} \times H \times W$.
We perform the positional encoding of DOYs of length T.
We extract SS features of DEM data resulting in shape $C_{DEM_{emb}} \times H \times W$, where $C_{DEM_{emb}} = C_{im_{emb}}$. As DEM data is shared by both modalities, we use a single DEM Unet module for both SSTE Aux modules.
As in SSTE, we sum the positional encoding values with SITS features and we concatenate the encoded DEM data in the temporal dimension of the feature SITS, resulting in length $T + 1$. Finally, we pass this extended SS feature series to a temporal encoder to obtain an unaligned embedding if the SITS with new dimensions $Q \times F \times H \times W$, where Q and F are the numbers of queries and new features respectively.

We use global validation loss, both for MALICE and MALICE Aux, to choose the best model. In the default configuration, we set $Q = 10$ and $F = 64$.

2.2.3 Benchmarking & evaluation

For benchmarking, we perform:

- **linear probing** (freezing the pre-trained encoder E model parameters), and/or
- **fine-tuning** (continue training all model parameters)

for the downstream tasks, because

- linear classification with frozen encoder is a standard evaluation protocol in self-supervised learning thanks to its simplicity and ease of use, and
- fine-tuning is the most straightforward way to evaluate the upper limit of pre-trained models.

2.2.3.1 SSL4EO-S12

The pre-trained models are transferred to various downstream tasks for evaluation on scene classification and semantic segmentation.

- **Scene classification:** we include
  - EuroSAT (Helber et al., 2018)
    (single label land cover classification) and
  - BigEarthNet (Sumbul et al., 2019)
    (multilabel land cover classification).
- **Semantic segmentation:** we focus on the SSL4EO-EU-Forest dataset.

2.2.3.2 DeCUR

2.2.3.2.1 Pre-training datasets

We pretrain DeCUR in three multimodal scenarios: SAR-optical, RGB-DEM and RGB-depth. For SAR-optical, we utilize the SSL4EO-S12 dataset. One random season is selected to generate each augmented view. For RGB-DEM, we conduct pretraining on the training set of GeoNRW dataset (Baier et al., 2022). The dataset includes orthorectified aerial photographs (RGB), LiDAR-derived digital elevation models (DEM) and OpenStreetMap refined segmentation maps from the German state of North Rhine-Westphalia. We crop the raw 6.942 training scenes to 111.000 patches of size 250x250 pixels. For RGB-Depth, we use the SUN RGB-D dataset (Song et al., 2015) which consists of 10.335 RGB-Depth pairs of various image sizes. Following (J. Zhang et al., 2023), we preprocess the depth images to HHA format (Gupta et al., 2014).

2.2.3.2.2 Model Details

As detailed in section 2.2.2.2, the model architecture comprises of a multimodal extension of Barlow Twins where each branch holds a separate backbone and a 3-layer MLP projector of output dimension 819). DeCUR is trained on embedding representations after the projector, whose dimensions are split into *common* and modality-*unique* components. We perform a light grid search to estimate the best split ratio: For our experiments with Sentinel-1/2 we numerically determined the common dimensions to occupy 87.5%; for RGB-DEM and RGB-depth this figure reduces to 75%. We employ a ResNet-50 for all scenarios, with an additional SegFormer (E. Xie et al., 2021) or RGB-Depth.

2.2.3.2.3 Downstream Datasets

We evaluate our pre-trained models on three downstream tasks:

- **Scene Classification:** we use the EuroSAT dataset as a popular benchmark to assess the ability of our pre-trained models to learn semantic representations. We report the overall accuracy for each model.
- **Multi-label Classification:** we evaluate our models on BigEarthNet-MM (Sumbul et al., 2021b) using 100% and 10% of the labels to showcase the label efficiency of SSL pre-training. We report the mean average precision (mAP) of each model.
- **Semantic Segmentation:** We evaluate RGB-DEM pretraining on the SUN RGB-D (37 classes) and NYU-Depth-v2 (Silberman et al., 2012) (40 classes) datasets for semantic segmentation. Fully convolutional networks (FCN) (Shelhamer et al., 2016) with standard ResNet50 backbone for each modality are used as the segmentation model.

2.2.3.3 SoftCon

2.2.3.3.1 Pre-training

We pretrain SoftCon with ResNet50 (MoCo-v2) and ViT-S/ViT-B (MoCo-v3) backbones on the SSL4EO-S12-MLdataset. Multispectral and SAR images are normalized to 8-bit. When multiple seasonal images exist, two are randomly selected for the base and momentum encoders. Data augmentation follows previous work, including random crops (224×224), colour jitter, grayscale conversion, Gaussian blur, and flips. For continual pretraining, we initialize ResNet50 from DINOv2 and ViT-S/ViT-B from DINOv2 (without register) and train for 100 epochs.

2.2.3.3.2 Downstream Tasks

The pretrained models are evaluated using linear probing and fine-tuning on 11 downstream tasks, including land cover classification, segmentation, and change detection. Tasks include:

- Land Cover Classification: BigEarthNet, EuroSAT, fMoW-Sentinel
- Segmentation: DFC2020

For classification tasks, models are trained using 1% or 10% of labeled data where applicable, and mAP or top-1 accuracy is reported. Segmentation tasks use mIoU scores. A grid search determines the optimal learning rate, using either SGD or AdamW, with training durations ranging from 30 to 100 epochs. All datasets use a 224×224 input size, and for GEO-Bench tasks, only the linear head or decoder is trained while the encoder is frozen.

2.2.3.4 CromSS

2.2.3.4.1 Pre-training

CromSS is pretrained on the SSL4EO-S12@NoL dataset using a U-Nets architecture with Sentinel-2 L1C imagery. The training uses the Adam optimizer with an initial learning rate of 5×10^{-3}, adjusted via a ReduceLROnPlateau scheduler. A 1% validation split is held out from training. The sample selection ratio (α) decays from 1 to 0.5 over 80 epochs, while the confidence weighting factor (γ) ramps from 0 to 1 in parallel.

2.2.3.4.2 Downstream Datasets

The pretrained encoders are transferred to three segmentation benchmarks:

SSL4EO-S12@DW: Similar to the pretraining data, with cleaner DW labels.
DFC2020

For comparison, CromSS is evaluated against seven state-of-the-art SSL methods: DINO, MoCo, DeCUR, scaleMAE, satMAE, satMAE++, and DOFA, as well as a supervised ResNet-50 from Satlas. ViT-based baselines (e.g., DOFA, satMAE++) are fine-tuned using UperNet.

2.2.3.5 SpectralEarth

We pre-train Spectral-ResNet-50 and spectral Spectral-ViT-S employing the MoCo and DINO algorithms. We utilize the augmentation recipe of MoCo-v2 without colour-jittering and Gray-scale to avoid distorting spectral information.

2.2.3.5.1 Downstream Tasks

Multilabel classification on EnMAP-CORINE: We append a linear classification layer to the pretrained ResNet and ViT backbones to continue training the models with Binary Cross-Entropy (BCE) loss for 100 epochs.

Semantic segmentation on EnMAP-CDL/NLCD/TreeMap/BDForet/BNETD and DESIS/EO1-CDL: We follow a lightweight adaptation transfer protocol for the segmentation tasks like in BYOL (Grill et al., 2020b) and MoCo. We train for 100 epochs on all datasets. The pretrained backbones are converted into segmentation models as follows:

- *ResNet:* We remove all strides from convolutions, replace them with atrous convolutions, and append two final convolutional layers.
- *ViT:* We reshape the tokens from the last layer and add two additional convolutional layers and upscaling steps.

Soil parameter regression: We use the training set of the Hyperview Challenge dataset (Nalepa et al., 2022) for the evaluation of cross-sensor transfer learning. Hyperview is a soil parameter estimation dataset consisting of 1.732 hyperspectral patches with 150 bands. Each patch is annotated with four soil parameters including potassium, phosphorus pentoxide, magnesium, and acidity. We append two fully connected layers to the pretrained ResNet and ViT backbones. The models were optimized based on a normalized mean squared error criterion.

2.2.3.5.2 Evaluation Protocols

We consider a plurality of evaluation protocols with various levels of fine-tuning the pre-trained models:

- **Frozen Backbone:** evaluates the quality of learned representations without any fine-tuning of the backbone.
- **Fine-tuning:** full model fine-tuning on downstream datasets.
- **Fine-tuning Adapter:** While frozen encoder evaluation is an effective measure of the quality of learned representations, it may not always yield the best results.

In the context of satellite imagery, it is often beneficial to adapt the model to the specific distribution of the downstream dataset. The distribution depends on several factors: geo-location, illumination conditions, seasonal changes, etc. Additionally, the relevance of spectral bands may vary depending on the downstream task. Therefore, we consider an intermediate setting where only the initial spectral adapter block is fine-tuned.

2.2.3.5.3 Evaluation Metrics

Depending on the nature of the downstream task, we report the following metrics:

- **Classification:** The multi-label F1-score is reported for the EnMAP-CORINE dataset.
- **Segmentation:** The mean Intersection over Union (mIoU) is reported for all segmentation datasets.
- **Regression:** For the Hyperview dataset, we follow the metric defined in the challenge: let MSE1, MSE2, ... be the mean squared errors of the predictor for soil parameter 1, 2, ... Let MSI1bsae, MSI2base, ... be the corresponding mean squared errors of a base predictor returning the mean value of the soil parameter calculated over the training set. Then, the reported Normalized MSE is defined as the ratio of the sum of these two mean squared errors.

2.2.3.5.4 Multi-Spectral to Hyperspectral Super-Resolution

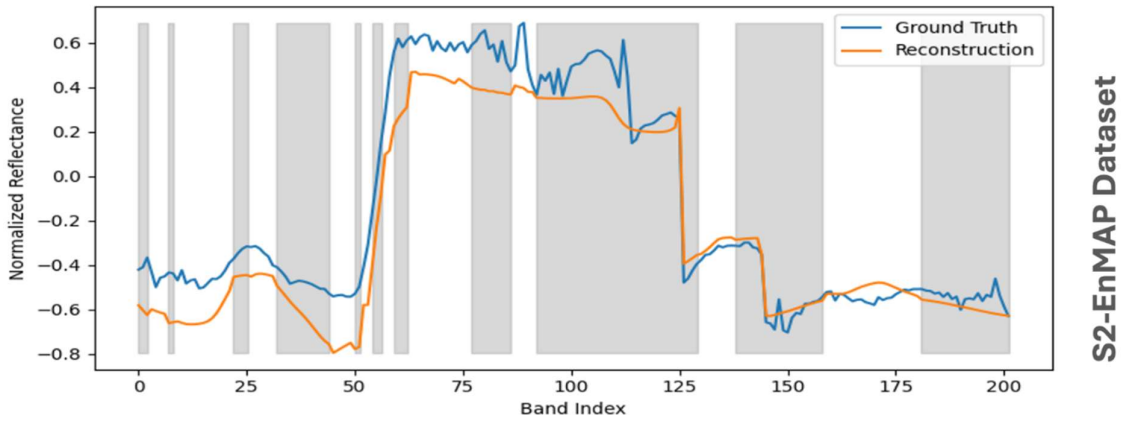


Figure 32. Sample of reconstructed spectrum of a single spatial pixel. While gray bands mark masked bands (e.g., not covered by Sentinel-2), the Band Index labels correspond to EnMAP channels.

In our specific implementation, the ViT model is small, with 4 encoder blocks and 2 decoder blocks. The embedding dimensions are 768 for the encoder and 512 for the decoder. The input images have dimensions of 128 × 128 × 240, and the model is trained using a masking ratio of 75%. Instead of employing a loss function that measures only the loss over masked pixels, a holistic loss function is used here, which computes the loss over both masked and unmasked pixels. This approach demonstrates better convergence compared to traditional methods.

The pretraining methodology involves masking strategies to capture both spectral and spatial-spectral relationships inherent in hyperspectral data. We explore two masking strategies for benchmarking: spectral and spatial-spectral masking. To avoid computational complexity, band groups are used instead of individual bands. In the spatial-spectral masking strategy, patches are similarly divided, but the masking is applied randomly to entire patches based on a specified ratio. These strategies are designed to enhance the reconstruction of hyperspectral images by emphasizing both spectral and spatial relationships.

As illustrated by Figure 32, at a masking ratio as strong as 75%, band reconstruction correctly captures qualitative trends compared to the ground truth – an initial indicator that our model has potential to reconstruct hyperspectral information from multi-spectral data such as Sentinel-2.

2.2.3.6 MMDC, pVAE, ALISE, MALICE, MALICE Aux

To evaluate the quality of the proposed embedding models, we test them on various real-life downstream tasks:

Leaf-Area-Index (LAI) regression - (single-date).
Pastis crop classification - (multi-temporal).
Tree Cover Density (TCD) regression - (multi-temporal).
Above Ground Biomass (AGB) and Forest Canopy Height (FCH) regression - (multi-temporal).

All the embedding models are frozen for the DST. Moreover, we do not produce the embeddings “on the go” during DST model training but compute them once as a data preparation step.

For each embedding algorithm, we choose several sets of hyperparameters for model training and evaluation. However, here we present only the results obtained with the best sets of hyperparameters. The embedding models with those sets of hyperparameters are considered the most optimal ones and, therefore, will be distributed in the future on openEO platform.

The following hyperparameters were chosen as the best for each model:

MMDC: MoE with masking strategy, LAI loss, embedding size of $C_{out} = 6$ features,
pVAE: default configuration,
ALISE: $Q = 10, C_{out} = 64, w_{inw} = 1, w_{rec_m} = 1, w_{rec_c} = 0$
MALICE: $Q = 10, C_{out} = 64, w_{inw} = 1, w_{rec_m} = 1, w_{re\ _c} = 0$
MALICE Aux: $Q = 10, C_{out} = 64, w_{inw} = 1, w_{re\ _m} = 1, w_{rec_c} = 0$, and same modality invariance loss.

Note that for some algorithms we use MMDC S1 embeddings computed for both orbits, while for others, only the ascending orbit is used. This choice is driven by the co-occurrence of same day acquisitions in the available datasets.

We use the rMSE (root Mean-Squared Error) and R2 score metrics for regression tasks. And mIoU (mean intersection over union) and Accuracy for classification.

2.2.3.6.1 LAI regression

LAI regression is a single-date estimation of a biophysical variable, usually achieved with high accuracy with Sentinel-2 data alone.

We evaluate this DST task with the raw Sentinel-1 ascending and Sentinel-2 data input, as well as with Sentinel-1 and Sentinel-2 MMDC embeddings. The other embeddings cannot be evaluated, as ALISE, MALICE, and MALICE Aux process multi-temporal data, and pVAE already contains LAI value among the extracted latent variables.

We use a simple 2D convolution model for all types of inputs. As it was stated in the preliminary experiences, the model struggle to predict high LAI values due to their underrepresentation in the dataset. For this reason, we apply loss weights, proportional to the GT values.

For MMDC embeddings, we use both Sentinel-1 orbits. However, due to the model design, we only use ascending orbit for the raw data.

2.2.3.6.2 SITS classification

Pastis Crop Classification task exploits SITS rather that single date images in order to detect the crop phenology.

We use Pastis-R dataset and its variations for this DST. As some of our embedding algorithms (MMDC and pVAE) perform single-date image embeddings and do not take into account the temporal aspect, we encode Pastis SITS day-wise to produce a new encoded time series of the same temporal dimension with those embedding algorithms.

We use different DST models for predictions, depending on the input data:

For the raw input data, as well as for MMDC and pVAE encoded series, we use U-TAE (U-net with Temporal Attention Encoder) model (Garnot et al., 2022; Garnot & Landrieu, 2021) (Figure 33) for spatio-temporal feature extraction followed by a classification layer. As in the original paper (Garnot et al., 2022), we perform multi-modal predictions, as well as single-modality predictions. For multi-modal predictions, we use a late fusion feature strategy: the features are extracted for each modality separately with U-TAE model and then concatenated and passed to the final classification layer.

For the multi-temporal embeddings, we use a simple MLP classifier (3 dense layers with ReLU activations), as we consider that the extracted multi-temporal embeddings already contain the essential information. Note that contrary to the single-date models, we only perform single-modality embeddings.

For all the DST tasks, we train the models for 100 epochs and choose the one with the best validation metric - mIoU (mean intersection over union). We use Cross Entropy loss for model optimization.

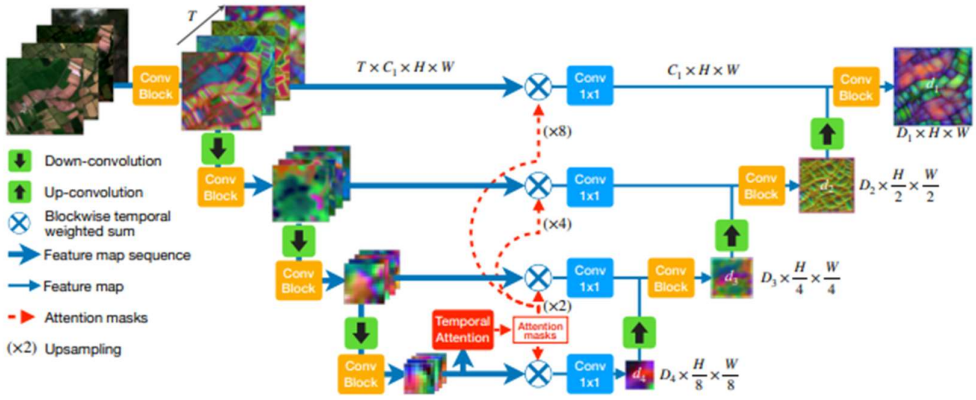


Figure 33. UTAE model (taken from (Garnot & Landrieu, 2021)).

2.2.3.6.3 TCD regression

To assess the quality of our embeddings, we compare them to the original benchmark VLCC 2018 TCD of GAF partners. For the raw Sentinel-2 data, we choose all the images available from March to October 2018 (leaf season) with cloud coverage less than 60%, plus 4 less cloudy images per month. However, contrary to the original benchmark, we equally assess Sentinel-1 data performance, exploiting all the available images for the given timestamp. For the selected dates, we compute the feature-wise median values for each month to generate a monthly synthesis. Finally, the processed data is passed to the CatBoost Regression model.

While GAF partners use the following hyperparameters to train their model: iterations=150, l2_leaf_reg=17, depth=7, we have empirically chosen the following parameters: iterations=750, l2_leaf_reg=17, depth=10.

We use the same date selection strategy for our single-date embedding algorithms (MMDC, pVAE). However, as the multi-temporal embedding algorithms (ALISE, MALICE, MALICE Aux) already produce the cloud-free temporal synthesis, there is no need for the laborious date selection process. We use the same CatBoost algorithms hyperparameters for all the tests.

2.2.3.6.4 BioMass and Tree Height Regression

Sentinel-1 dataset contains 7 years of data over the whole country of Sweden. We select tiles 33VVK and 33VWK for the test (as in the original benchmark of our partners), and we divide the resting tiles in training and validation tiles with 70/30% ratio.

For our method, we use two separate MLP models, for AGB and FCH. We apply loss weights to the high values, as they are underrepresented in our dataset.

2.3 Implementation and results

2.3.1 SSL4EO-S12

2.3.1.1 Classification

We benchmark different SSL methods through linear probing on EuroSAT and BigEarthNet, following standard practices from the literature. As detailed in Table 4, all methods outperform random initialization (rand. init.) by a substantial margin. As expected, linear probing on BigEarthNet with all labels performs worse than fully supervised training. Promisingly, the gap stays below 5%. On small datasets like BigEarthNet with 10% labels or EuroSAT, linear probing provides results comparable to supervised training within approx. ±1%.

Table 4. Linear probing results for EuroSAT, BigEarthNet (BE) with 10% and 100% labels. We report overall accuracy for EuroSAT and mean average precision for BigEarthNet.

Model	Backbone	EuroSAT	BE-10%	BE-100%
rand. init.	ResNet-50	82.0	61.0	60.0
	ViT-S	81.3	60.0	60.0
MoCo	ResNet-50	98.0	82.1	84.2
	ViT-S	97.7	82.3	83.1
DINO	ResNet-50	97.2	82.0	83.9
	ViT-S	97.7	81.7	83.4
MAE	ViT-S	94.1	77.5	78.2
data2vec	ViT-S	96.9	77.3	79.4
supervised	ResNet-50	98.0	83.4	87.4
	ViT-S	96.7	81.3	87.4

Furthermore, we benchmark fine-tuning results in Table 5. All self-supervised methods outperform supervised learning with a margin from 1% to 6%. Top SSL-models score 99.1% on EuroSAT (MoCo/DINO) and over 90% on BigEarthNet (MoCo/DINO). Comparing linear probing and fine-tuning results, one interesting phenomenon shows up: in linear probing contrastive methods (MoCo and DINO) consistently score better than image-masking methods (MAE and data2vec), which is consistent with existing literature.

Table 5. Fine-tuning results for EuroSAT and BigEarthNet.

Model	Backbone	EuroSAT	BE-10%	BE-100%
MoCo	ResNet-50	99.1	86.2	91.8
	ViT-S	98.6	86.1	89.9
DINO	ResNet-50	99.1	87.1	90.7
	ViT-S	99.0	86.9	90.5
MAE	ViT-S	98.7	84.8	88.9
data2vec	ViT-S	99.1	85.6	90.3

We implement embedding generation of Sentinel-2 data by SSL4EO-S12 pre-trained models in the `evoland` code package. A sample Python code snipped reads:

```
# configuration & Sentinel-2 image preprocessing
from evoland.data import patchify
IMAGE_DIR  = './input-patch-dir'
PATCH_SIZE = 224
patchify(
    intput_file = 'sentinel2.tif',
    output_dir  = IMAGE_DIR,
    patch_size  = PATCH_SIZE,
)

# load pre-processed data into AI model-consumable form
from evoland.data import TreeCoverDataModule
data = TreeCoverDataModule(
    root         = '.',
    img_dir_name = IMAGE_DIR,
    batch_size   = 1,
    num_workers  = 0,
    patch_size   = PATCH_SIZE,
    use_masks    = False,
)

# load SSL4EO-S12 model for embedding generation
from evoland.models import FeatureExtractorModule
model = FeatureExtractorModule(
    backbone_name = 'resnet18',
    pretrained    = 'ssl4eo',
)

# generate embeddings (with PyTorch Lightning framework)
from lightning.pytorch.trainer import Trainer
embeddings = Trainer(
    accelerator = 'cpu',
    devices     = 1,
).predict(
    model              = model,
```

```
    datamodule         = data,
    return_predictions = True,
)

# save geo-referenced embeddings to GeoJSON file
data.save_geojson(embeddings, './')
```

It will be generated a GeoJSON vector file `feature_vectors.json` that is consumable by geospatial analytics tools such as QGIS. A sample record for an embedding `embeddings[0]` in the saved GeoJSON file reads:

```
{
  "type": "Feature",
  "geometry": {
    "type": "Point",
    "coordinates": [
      26.828725053307725,
      60.497680302616295
    ]
  },
  "properties": {
    "feature_vector": [
      0.22618064284324646,
      0.003551470348611474,
      0.0658784806728363,
      0.09929963201284409,
      ...
    ]
  }
}
```

Figure 34 exemplifies the geolocation of the (512-dimensional) embeddings relative to the input image patches (224x244 pixels). We pick the centre of the corresponding input patch.

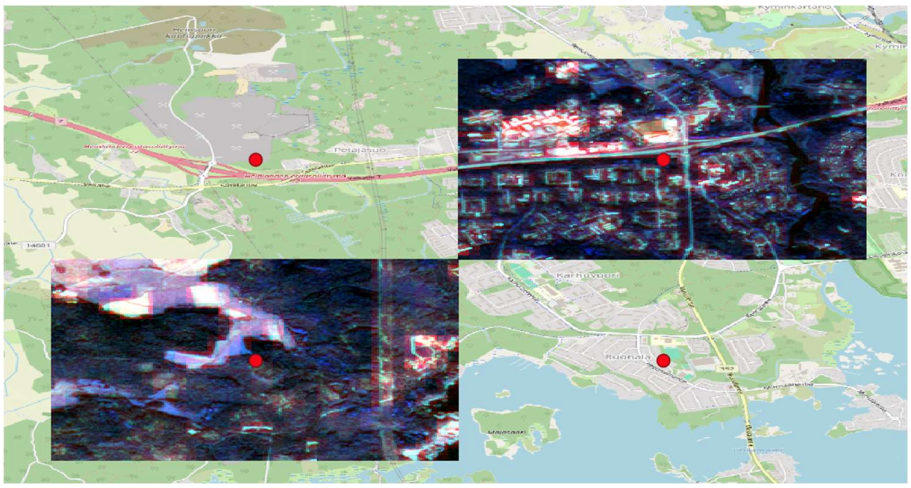


Figure 34. Visualization of four geolocated image-patch-level SSL4EO-S12 embeddings (red dots) with two out of the four Sentinel-2 input images, background map: OpenStreetMap.

2.3.1.2 Segmentation

To assess the performance of our pre-trained models on segmentation tasks, we leverage the SSL4EO-EU-Forest dataset. For a thorough evaluation, we consider three backbones: ResNet-18, ResNet-50, and ViT-S coupled with 3 segmentation models: UNet, DeepLabV3+, and a simple convolutional decoder for the vision transformer backbone.

We solely consider full model fine-tuning for optimal performance, and we compare all results against training from scratch as a baseline. Table 6 summarizes our experiments.

Table 6. Summary of benchmark results for forest semantic segmentation.

Segmentation Protocol	Encoder	Pre-training Weights	Overall Accuracy	Mean IoU
UNet	ResNet-18	Random	84.05	72.49
		MoCo	**85.19**	**74.19**
		DINO	85.14	74.13
	ResNet-50	Random	84.25	72.79
		MoCo	**85.27**	**74.33**
		DINO	84.84	73.68
DeepLabV3+	ResNet-18	Random	84.15	72.63
		MoCo	84.94	73.83
		DINO	**85.15**	**74.14**
	ResNet-50	Random	84.17	72.67
		MoCo	**85.31**	**74.39**
		DINO	84.76	73.55
UpConv	ViT-S	Random	83.47	71.63
		MoCo	85.08	74.03
		DINO	**85.26**	**74.31**

As a general trend, we observe that our pre-trained models outperform training from scratch in every setting. Moreover, MoCo pre-trained models perform on average slightly better than DINO. Interestingly, the ResNet-18 backbone, which is significantly smaller than ResNet-50 and ViT-S, performs comparably well. This indicates that the tree cover segmentation task does not benefit from very deep models.

Qualitative results are provided in Figure 35 and Figure 36. We observe that the segmentation maps produced by our pre-trained models are more accurate and closer to the ground truth than the models trained from scratch, which is aligned with the quantitative results discussed above. Moreover, Figure 36 illustrates that different backbones yield similar segmentation maps. We do not observe significant differences or characteristic patterns for each architecture. Given these observations, the simplest and most efficient backbone, i.e., ResNet-18, stands out as a natural candidate picked out of an overall set of 15 models trained.

Remarkably, our models yield accurate predictions across seasons, including fall and winter, as illustrated by the last two rows in Figure 36. This robustness arises from the seasonal contrast during pre-training and the seasonal augmentation during fine-tuning. Both make the model capture seasonal patterns.

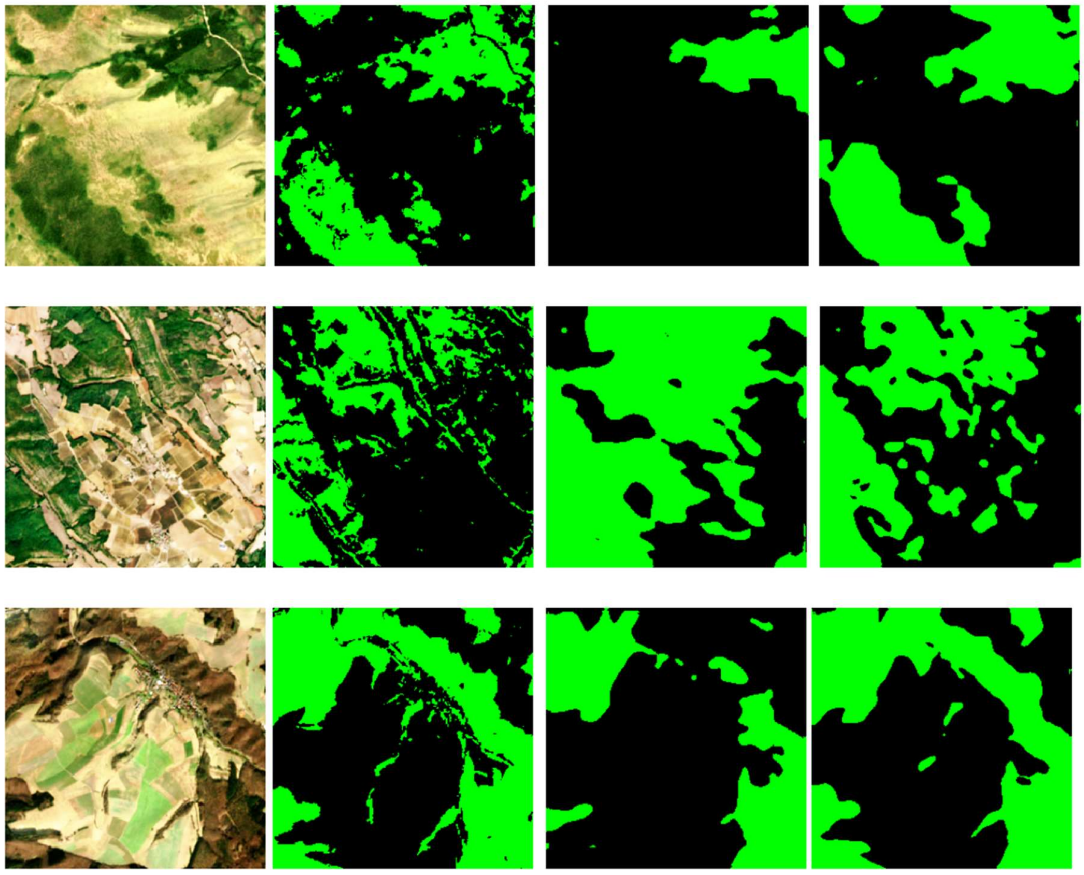
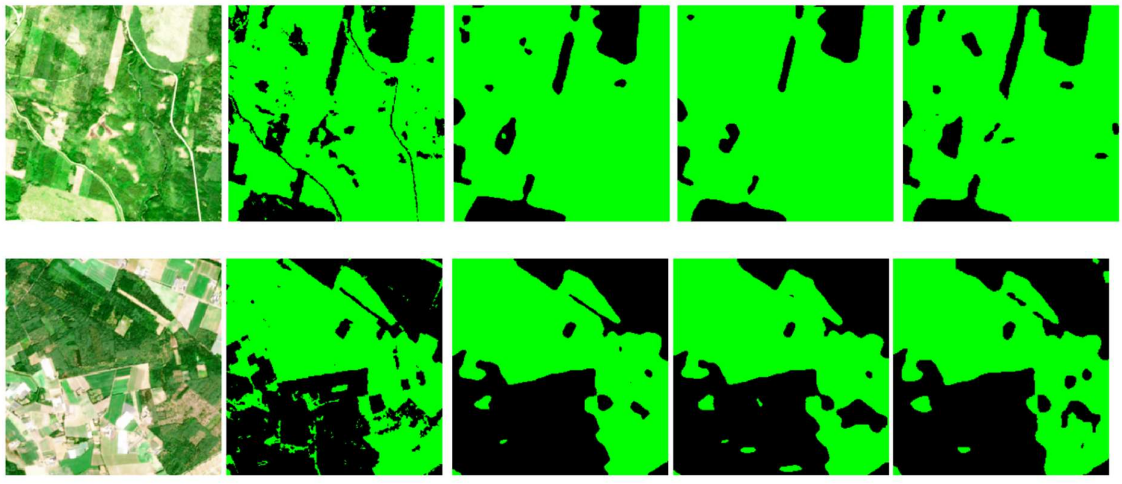


Figure 35. Comparison between SSL and Supervised Learning. From left to right: Sentinel-2, HRL forest mask (ground truth), Supervised Learning prediction, SSL & fine-tuning (UNet with ResNet-18 backbones for the encoder and decoder).



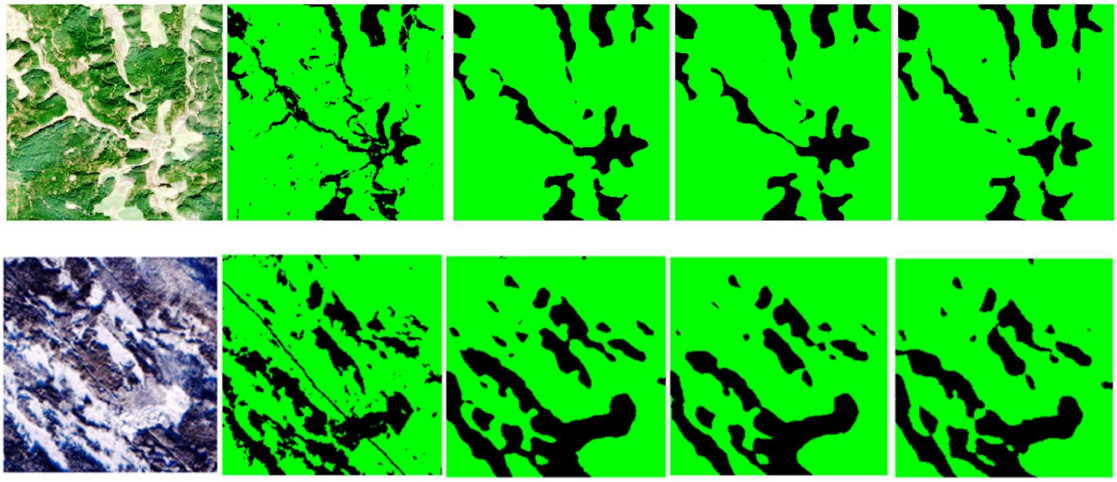


Figure 36. Comparison of pre-trained backbones with a UNet model. From left to right: Sentinel-2, HRL forest mask (ground truth), ResNet-18 prediction, ResNet-50 prediction, ViT-S prediction.

Although the segmentation models yield accurate results, with an overall accuracy of 85% on a single timestamp of Sentinel-2 data, their predictions tend to lose fine-grained details, leading to oversmoothed maps. This phenomenon can be explained by a combination of factors:

- **Inductive bias:** Convolutional neural networks and vision transformers have strong inductive biases allowing to leverage spatial information and locality in images. Looking at neighbouring pixels, either through convolution or attention/tokenization, is a very useful prior information to determine the class of a pixel. This is particularly true in computer vision, where small objects containing only a few pixels are rare. In contrast, for medium resolution remote sensing imagery, a few pixels can represent a building, a road infrastructure, dozens of trees, etc. Pixel-level classifiers, which treat every pixel independently, can preserve high frequency details - at the price of more noisy and less accurate outputs.
- **Noisy labels:** The SSL4EO-S12-Forest masks are based on the HRL-2018 product. While very accurate, the labels are still not perfect as demonstrated in Figure 37. In fact, there is an inherent uncertainty stemming from a product aggregated over the span of a year: the tree cover in regions could have changed over the course of 2018. These uncertainties in the ground truth (typical for small, isolated patches rather than large regions) may adversely impact the model performance. By focusing on fine-grained details, the network will inevitably fit the noise in forest maps losing accuracy.
- **Training objective:** We fine-tune our models using a cross-entropy loss - as commonly reported in the literature. However, cross-entropy and other popular segmentation losses (e.g., Dice loss), optimize over all pixels in the scene with uniform weighting. Consequently, small tree/non-tree areas do not significantly contribute to the overall training objective. One could over-

emphasize such regions; however, this comes at a price of degrading overall model performance.

- **Engineering design choices:** Certain design choices in popular convolutional neural networks such as ResNets can exacerbate the loss of fine-grained details. For example, the original ResNet-18 & -50 architectures have a stem layer (convolution followed by max-pooling) which downscales the feature maps' height and width by a factor of 4. While this improves the computational efficiency of the model, avoiding full-resolution feature maps contributes to the observed "blobiness" effect where predicted regions of trees exhibit rounded boundaries loosing pixel-level details and fine-grained variation.
  Design choices like what we employ can be found in popular semantic segmentation models such as UNet. It typically consists of an encoder and a decoder part. The encoder reduces the spatial dimensions of the feature maps, subsequently upscaling them by the decoder. While the skip connections greatly help in preserving details, the maps that are produced remain smoother than what a pixel-level classifier would produce.

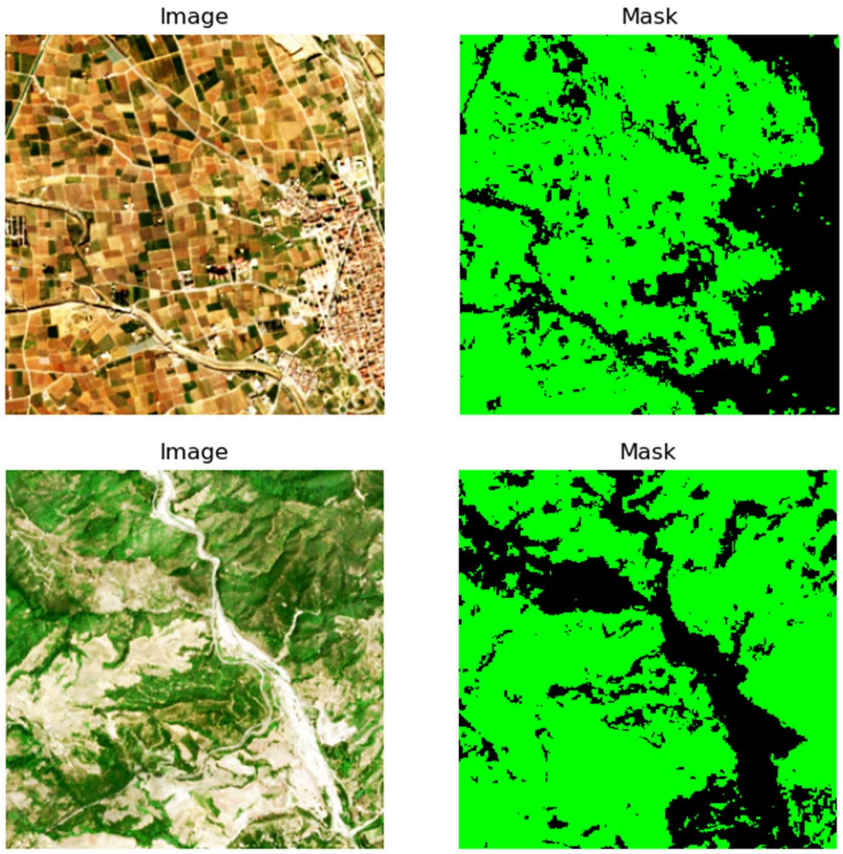


Figure 37. Two illustrative examples where the 2018 HRL forest masks (right column) do not well match the ground truth Sentinel-2 imagery for a random season that same year (left column): Forest seem to have converted into agriculture land (top) and forests might have undergone clearance by logging, etc.

When designing a segmentation model in a single timestamp setting, it requires balancing the inherent trade-offs between overall accuracy, noise level of the output, and computational efficiency. We explored these dimensions through a series of experiments.

First, we aggressively down sampled the ResNet architecture according to:

- **Convolution Kernel:** The original ResNet models use 7x7 kernels in the first convolution layer. Large kernels may contribute to the loss of details by aggregating extended regions of pixels. Therefore, we replaced the first layer with 3x3 kernels.
- **Strides and Pooling:** The first convolution layer in ResNet has a stride of 2 and is followed by max-pooling with an identical stride. As a result, the features maps entering the first residual blocks have their dimensions reduced by a factor of 4, i.e., 16 times smaller in terms of number of pixels. Motivated by the discussion above, we set the stride to 1 in the first convolution and remove the max pooling layer.

We will refer to the resulting model as “Custom-ResNet-18". A comparison between the original ResNet-18 and our modified version if provided by the following:

Table 7. ResNet-18 and Custom-ResNet-18 results.

Architecture	Overall Accuracy	Mean IoU
ResNet-18	85.14	74.13
Custom-ResNet-18	**86.08**	**75.55**

Custom ResNet-18 outperforms the original one, as well as all models in Table 7 reaching an overall accuracy of ~86%.

Figure 38 provides a qualitative assessment: The custom ResNet-18 better captures smaller areas without trees thanks to the network layers operating on full resolution feature maps. Nevertheless, the predictions still exhibit a certain degree of smoothness when compared to the 2018 HRL forest masks.

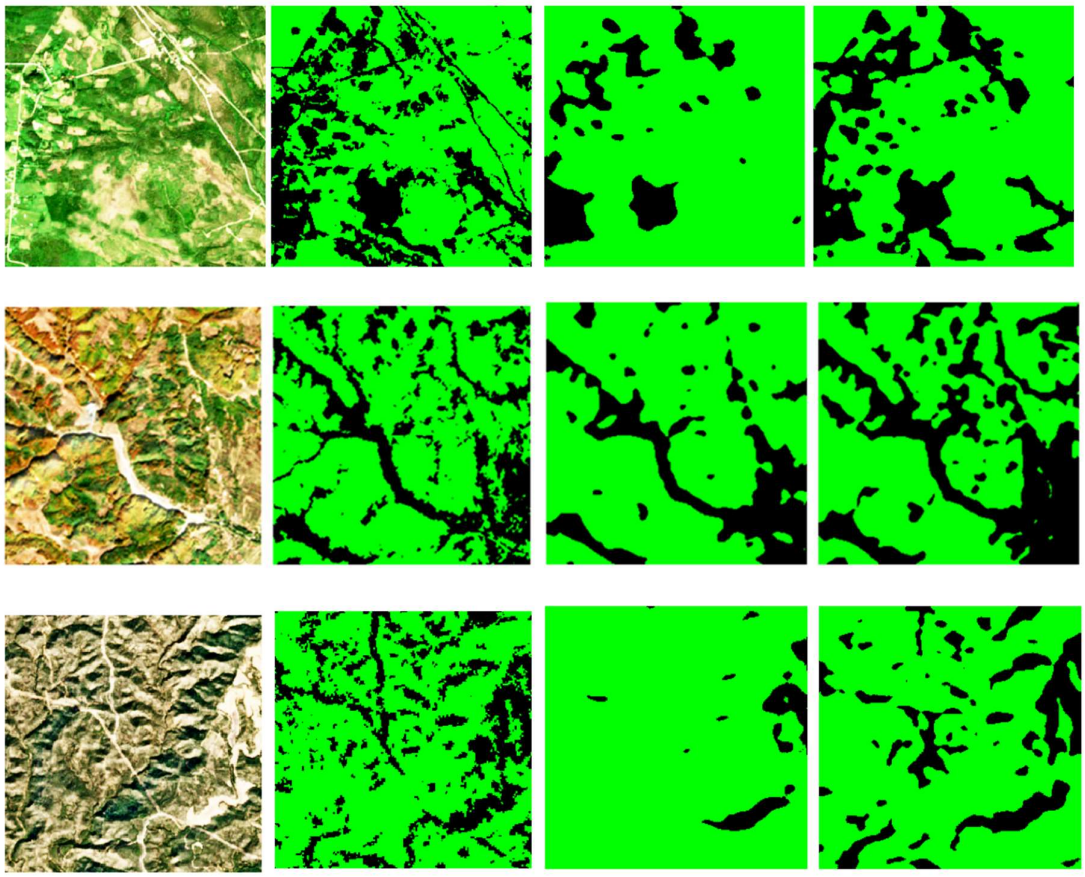


Figure 38. From left to right: Sentinel-2, HRL forest mask (ground truth), ResNet-18 prediction, Custom-ResNet-18 prediction.

Segmentation models such as UNet utilize an encoder-decoder architecture. Shallow layers in the network are low-level feature extractors (e.g., edge detectors), while deeper features represent more abstract concepts at a lower resolution. One possible way to keep full-resolution features in the entire network: adopt a design like in Fully Convolutional Networks (FCN). Specifically, all the strides in the ResNet encoder are set to one, and a final segmentation head taking as input the concatenation of feature maps from different layers. We refer to this model as FCN-ResNet-18. We investigate how deep and shallow features affect predictions and accuracy by repeatedly using a subset of the residual layers in the encoder. Quantitative and qualitative results are provided below with illustrations in Figure 39.

Table 8. Forest benchmark.

# of layers	1	2	3	All
Accuracy	76.22	81.05	82.55	84.31
mIoU	61.88	68.14	70.29	72.78

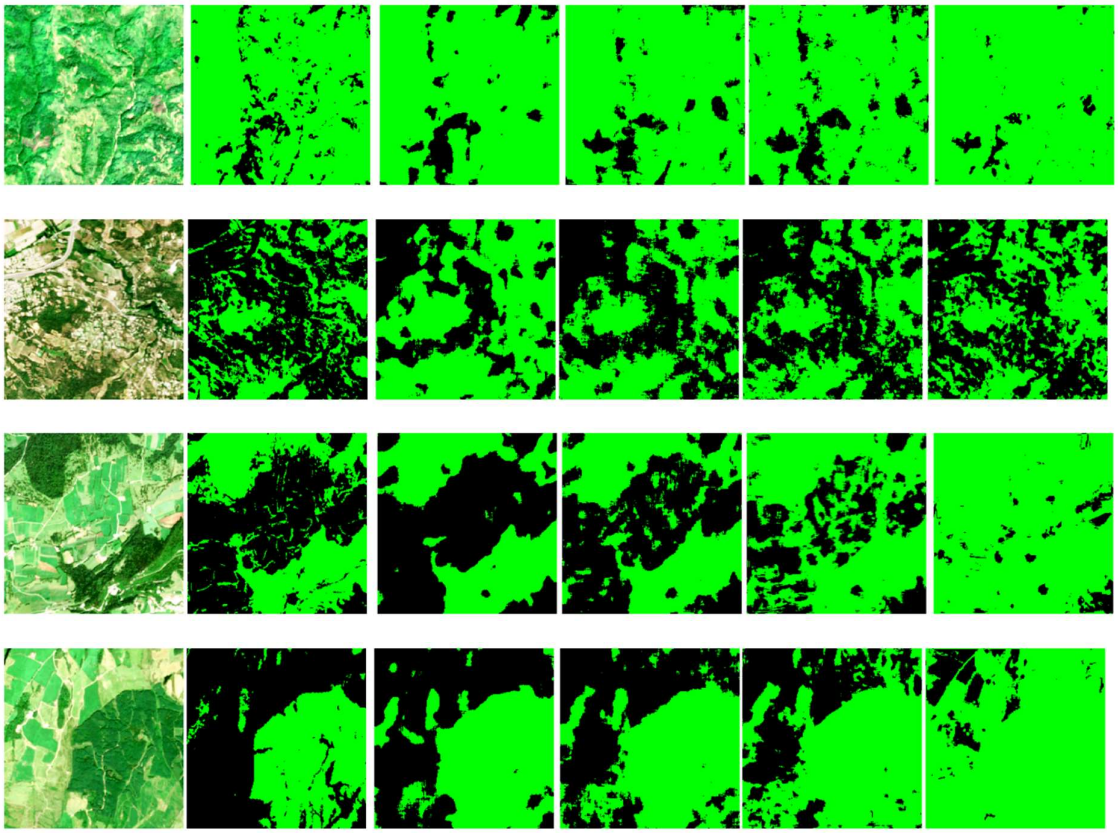


Figure 39. Illustration of forest prediction performance with FCN-ResNet-18. From left to right: Sentinel-2, HRL forest mask (ground truth), FCN-ResNet-18 prediction All layers, FCN-ResNet-18 prediction 3, 2, and 1 layer(s).

Our experiments confirm that the overall accuracy and mIoU scores for the FCN-ResNet are worse when compared to the UNet. This is consistent with existing literature. Moreover, using fewer layers systematically leads to a drop in performance. As our illustrations confirm: While shallow models can better capture fine-grained patterns (e.g., second row in Figure 39), they fail to distinguish any type of vegetation from trees. This demonstrates that such low-level features lack a semantic understanding of the scene. On the other hand, the deeper models have higher accuracy, but, in general, these models lose spatial details.

Finally, we explore how a deep, pure pixel-level model performs in forest semantic segmentation. When timeseries of Sentinel-2 data are available, the spatial context may not be critical for accurate pixel-wise classification. However, in the case of single scene settings, it is challenging to accurately classify pixels solely from their spectral signature given large variance in the data. Our experiments confirm this assumption. We pick a ResNet-18 backbone and restrict the kernel size to one for all convolutions: *ResNet-18/1*. For this model, we obtain an overall accuracy of 76.59 corresponding to a mean Intersection-over-Union value of 62.05. This result is significantly worse compared to our best UNet result, Custom-ResNet-18, cf. Table 8. Qualitative results are provided in Figure 40. When multi-spectral reflectance

values are discriminative (e.g., green leaves in an urban area such as in row 1), the predictions of ResNet-18/1 are close to ground truth. However, the model is noisier and tends to classify all vegetation as trees, like the shallow FCN-ResNet-18 models.

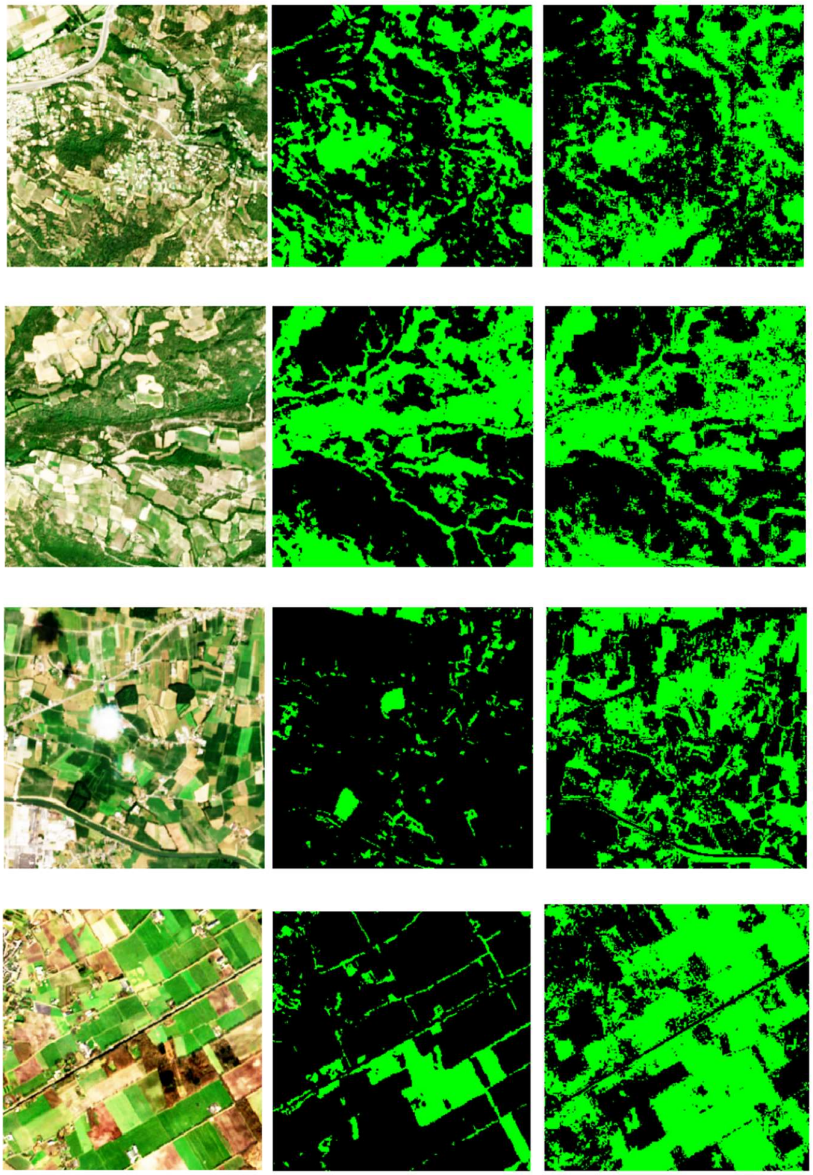


Figure 40. Visual demonstration of the single-pixel convolution model ResNet-18/1. From left to right: Sentinel-2, HRL ground truth, ResNet-18/1 prediction.

In the light for this analysis, we conclude that deeper semantic features are more accurate and robust to input style changes, while shallow features can preserve fine spatial patterns and details. Depending on the emphasis, different trade-offs can be achieved by combining the features learned through the network. As an example, we design a simple segmentation model, *ResNet-18-Segment*, that uses the features from every ResNet-18 stage, maps each of them to a constant size using 1x1 convolutions and applies a simple segmentation head on top of the

stacked features. This architecture yields 85% accuracy while providing outputs that are less oversmoothed than from what is observed with UNet architectures, see Figure 41.

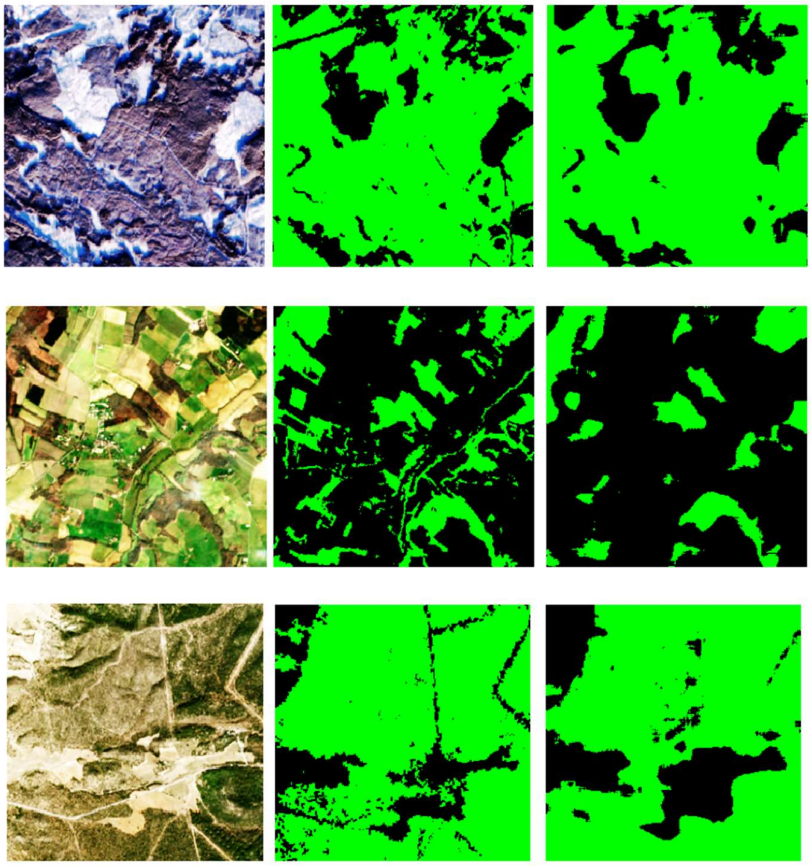


Figure 41. Visual demonstration of the single-pixel convolution model ResNet-18-Segment. From left to right: Sentinel-2, HRL ground truth, ResNet-18-Segment prediction.

2.3.1.3 Segmentation: Refined tree mapping

Building upon the initial segmentation experiments, we implemented a series of refinements aimed at improving **geometric accuracy, data quality of the SSL4EO-EU-Forests dataset, along with the sharpness of our model predictions**. These modifications have significantly improved our models both quantitatively and qualitatively.

2.3.1.3.1 Sentinel-2 Data Reprocessing

To ensure higher geometric fidelity, we reprocessed the Sentinel-2 data in accordance with the updates introduced in SSL4EO-S12 v1.1. This includes enhanced co-registration of Sentinel-2 imagery, mitigating spatial misalignment between spectral bands. By improving pixel alignment, we reduce interpolation artifacts and geometric noise, leading to cleaner inputs for segmentation models.

In addition to geometric corrections, we further filter out cloudy samples that were not detected by Sentinel-2's built-in cloud mask. To achieve this, we employ a Mobile Net V2 model, which was trained on the CloudSEN12+ dataset. We apply a conservative threshold of 20% cloud cover, ensuring that we do not discard too many valid samples. Figure 42displays examples of such images that are obfuscated with clouds to be used for training.

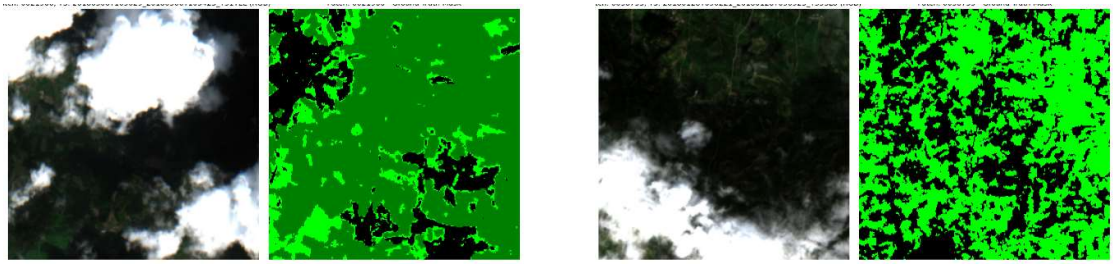


Figure 42. Samples of cloudy images detected by the deep learning cloud segmentation model

2.3.1.3.2 Weighted Cross-Entropy for Fine-Grained Segmentation

A key limitation observed in earlier results was the over-smoothing of segmentation outputs, particularly for isolated trees/small areas without trees. To mitigate this, we introduced a weighted cross-entropy loss function, which assigns higher weights to pixels located at class boundaries. This encourages the model to focus more on edge details, ultimately leading to sharper segmentation masks. The weight map is computed based on the distance transform of the segmentation masks, ensuring that boundary pixels are prioritized while retaining the overall class distribution. The definition of the loss for a HXW image is given in the equation below:

$$L_{WCE} = -\frac{1}{HW}\sum_{x=1}^{H}\sum_{y=1}^{W} w(x,y)\sum_{c=1}^{C} y_{x,y,c}\log\left(\widehat{y_{x,y,c}}\right)$$

Where $w(x,y) = w_b$ if (x,y) is a boundary pixel, else 1

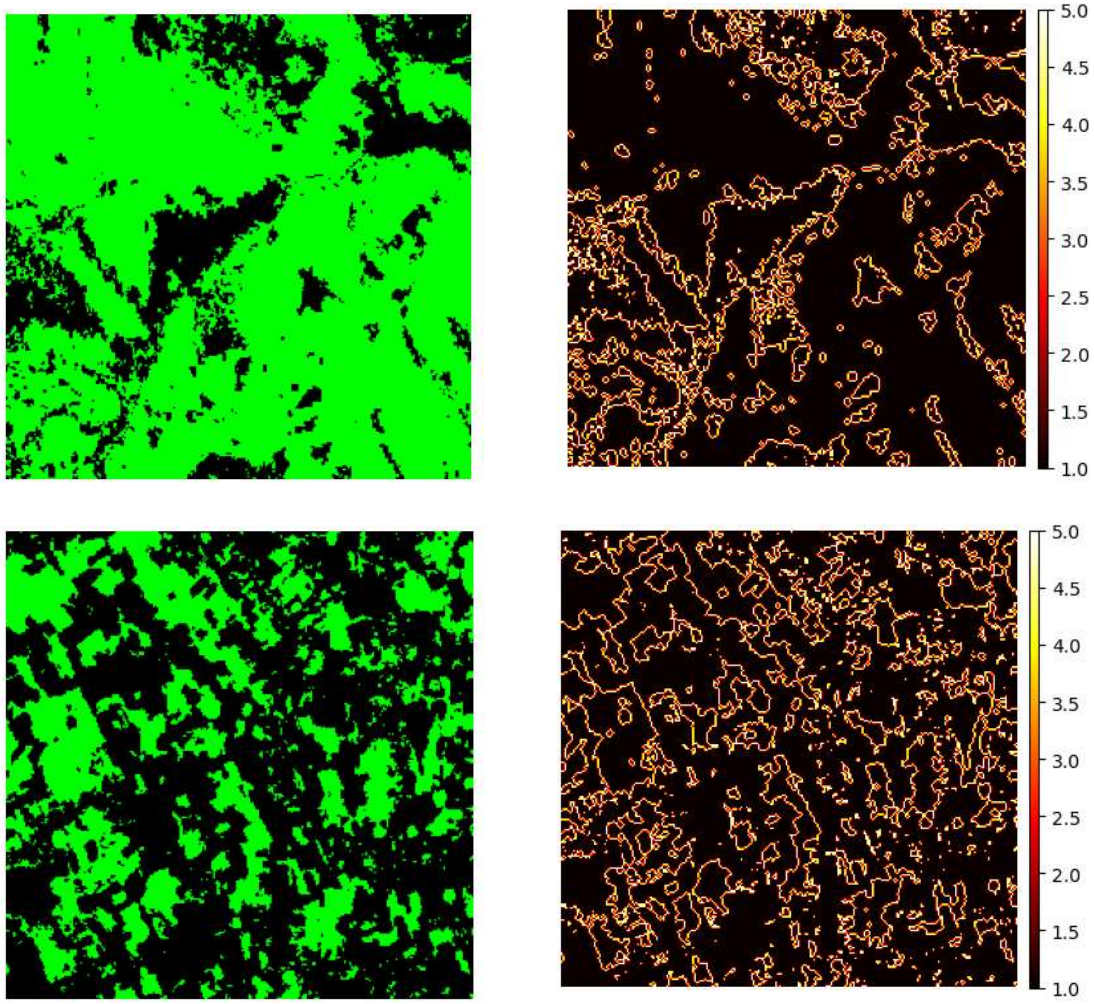


Figure 43. Samples illustrating the weighted cross entropy loss – Left: ground truth mask and Right: corresponding weight mask.

Figure 43 illustrates the areas of focus of our weighted cross entropy loss, showing how the model is trained to better preserve fine-grained details.

2.3.1.3.3 Automated Data Cleaning Pipeline

The SSL4EO-EU-Forests dataset inherently contain noisy labels due to errors in ground truth generation, land cover changes, and environmental factors such as clouds or snow. To address this, we implemented an iterative data cleaning pipeline that systematically identifies and removes low-quality samples from training.

The approach follows a semi-automatic filtering pipeline, wherein a pre-trained model is fine-tuned iteratively, and samples with poor segmentation performance are progressively excluded. A key aspect is that we start from a pre-trained model and only fine-tune for a few epochs per iteration. This prevents the model from

overfitting to noisy labels while still allowing it to refine its predictions. The pipeline follows these steps:

1. Fine-tune the model for a small number of epochs on the current dataset.

2. Evaluate per-sample segmentation quality using a predefined metric.

3. Identify the lowest-performing samples (bottom P\%) and exclude them from subsequent training iterations.

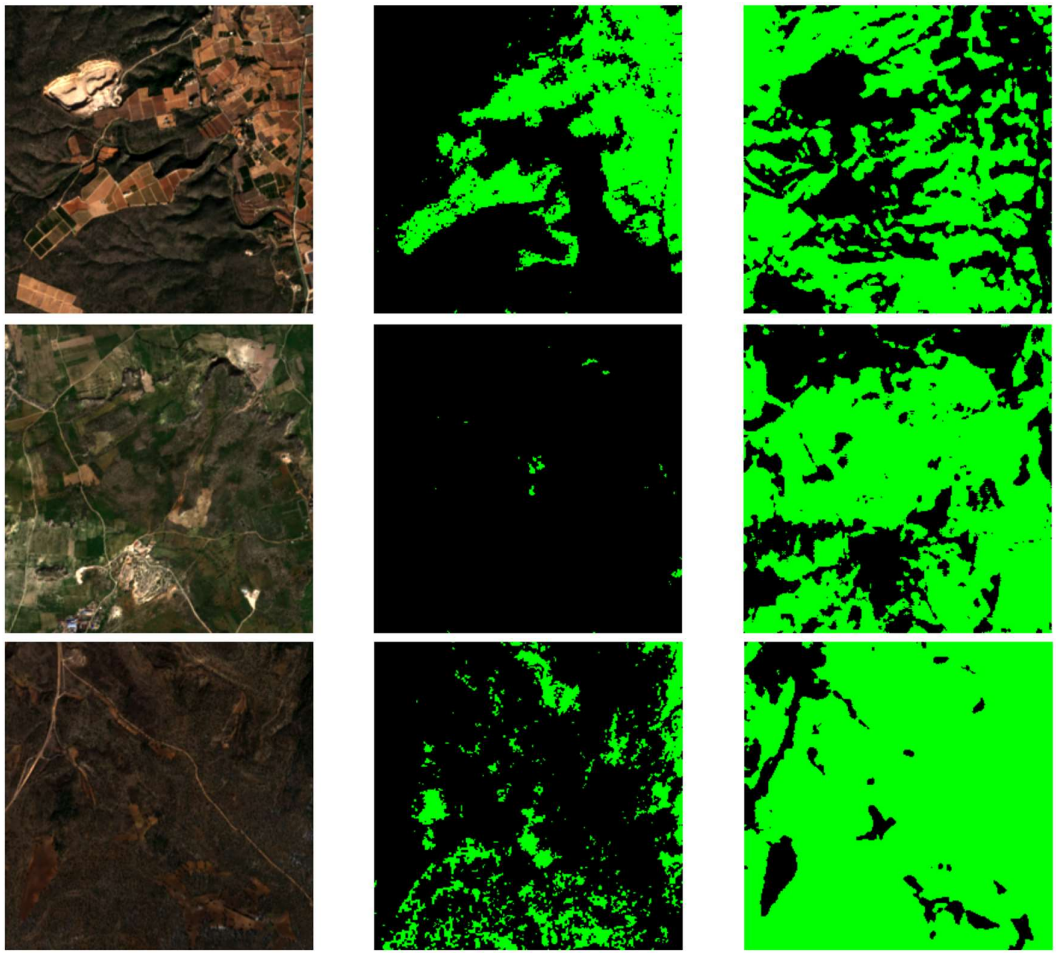
4. Repeat the process until no significant improvements are observed.

The process is formally detailed in Figure 44.

Data: Training dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$
Result: Fine-tuned model M, updated ignore list $\mathcal{I}$
Initialize ignore list $\mathcal{I} \leftarrow \emptyset$
for *iteration* $t = 1$ *to* T **do**
  `// Prepare training data excluding ignored samples`
  $\mathcal{D}_{\text{train}} \leftarrow \mathcal{D} \setminus \{(x_i, y_i) \mid i \in \mathcal{I}\}$
  `// Load pre-trained model`
  Initialize model M with pre-trained weights $W_{\text{pretrained}}$
  `// Fine-tune the model`
  Fine-tune M on $\mathcal{D}_{\text{train}}$ for E epochs
  `// Evaluate model on training data`
  for *each sample* $(x_i, y_i) \in \mathcal{D}_{train}$ **do**
    Compute prediction $\hat{y}_i = M(x_i)$
    Compute per-sample metric $m_i = \text{Metric}(y_i, \hat{y}_i)$
  end
  `// Identify low-performing samples`
  Rank samples based on m_i
  Select bottom $P\%$ samples $\mathcal{L} \leftarrow \{i \mid m_i$ is among lowest $P\%\}$
  `// Update ignore list`
  $\mathcal{I} \leftarrow \mathcal{I} \cup \mathcal{L}$
end
Algorithm 1: Iterative Fine-tuning with Noisy Sample Removal

Figure 44. Algorithm used to filter out noisy samples from the SSL4EO-EU-Forests dataset.

To ensure that the removed samples are genuinely noisy rather than challenging cases where the model fails, we visually inspect the discarded samples along with the model predictions obtained during the filtering process. The results are shown in Figure 45, where we observe that many removed samples exhibit a mismatch between the Sentinel-2 input and the ground truth masks. Interestingly, in some cases, the model still manages to correctly predict the tree cover despite discrepancies between the mask and the image.



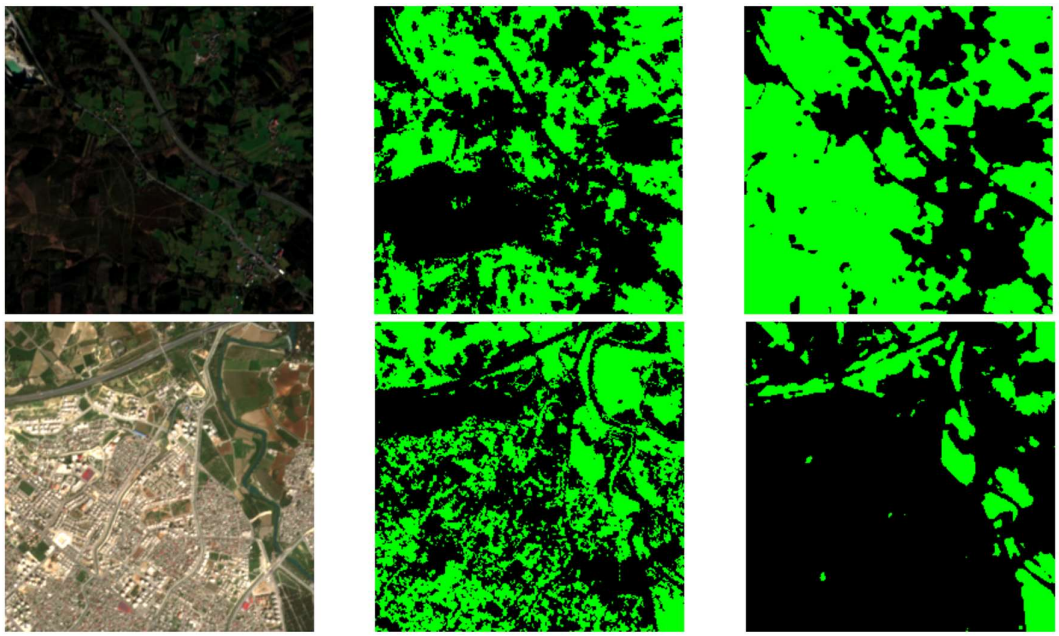


Figure 45. Noisy samples filtered out from SSL4EO-EU-Forests - Left: S2 image, Middle: target mask and Right: model prediction.

We run the algorithm with a percentage of 1% on individual timestamps from the SSL4EO-EU-Forests dataset for 100 iterations. The data distribution of the resulting filtered dataset is shown in Table 9. Comparison of forest-segmentation results from the first phase of the project (Custom-ResNet18) and the refined results from the second phase (Custom_ResNet18-WCE-Filtered).

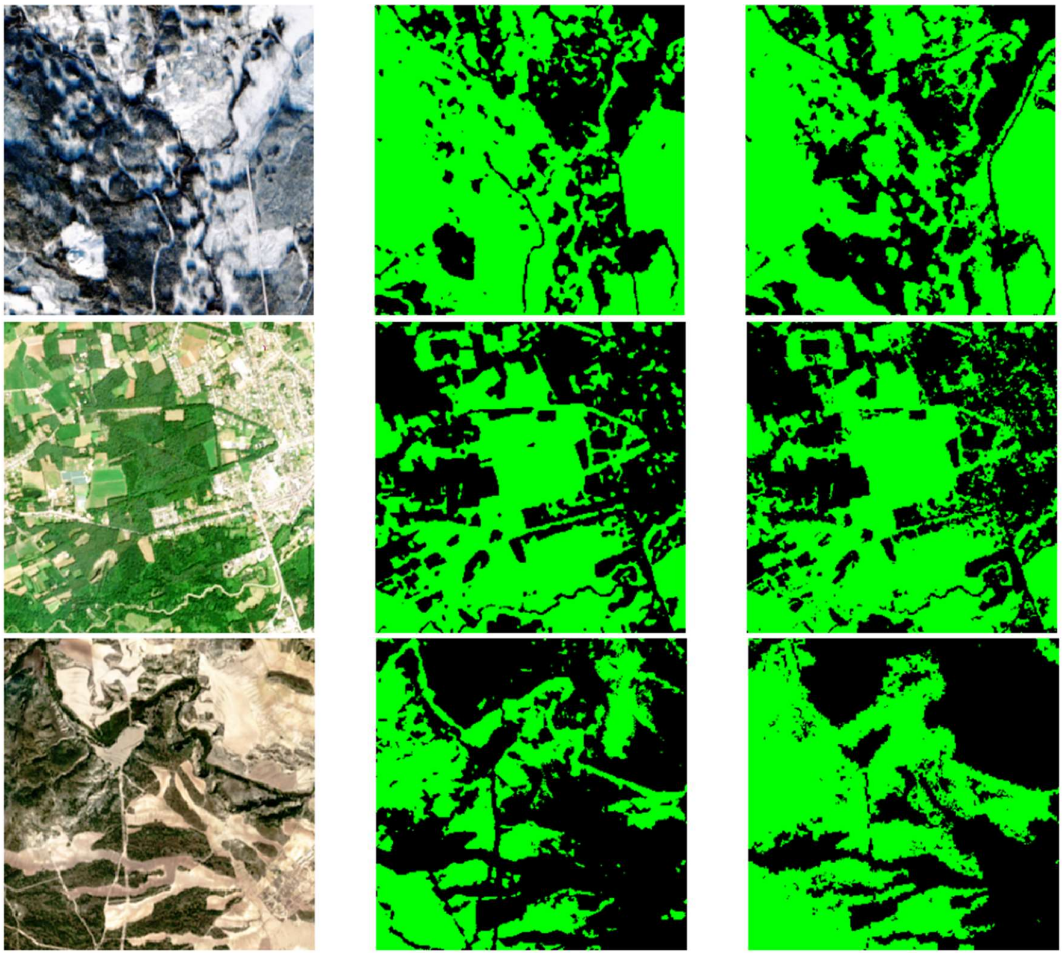
2.3.1.3.4 Updated Tree Mapping Results

We fine-tune a Custom-ResNet-18 model, pre-trained on SSL4EO-S12, using the weighted cross-entropy loss on the filtered dataset obtained after cloud removal and noisy mask filtering. The updated results are summarized in Table 9. Comparison of forest-segmentation results from the first phase of the project (Custom-ResNet18) and the refined results from the second phase (Custom_ResNet18-WCE-Filtered)., which compares the previous and new performance of Custom-ResNet-18. We observe a gain of 5 points in overall accuracy and 10 points in mean IoU (mIoU). This improvement demonstrates the effectiveness of our data filtering strategy in reducing noise-induced errors.

Table 9. Comparison of forest-segmentation results from the first phase of the project (Custom-ResNet18) and the refined results from the second phase (Custom_ResNet18-WCE-Filtered).

Model	Overall Accuracy	Mean IoU
Custom- ResNet18	86.08	75.55
Custom-ResNet18-WCE-Filtered	91.5%	85.3%

These improvements are also evident in qualitative results, as shown in Figure 46. In addition to higher accuracy, the updated segmentation masks exhibit more fine-grained spatial features, particularly along tree boundaries and fragmented vegetation areas. Compared to previous results, the new masks are less smoothed, capturing smaller tree clusters with higher precision. Moreover, our model performs well in challenging conditions, such as with snow cover (see row 1). With a single timestamp as input, it can yield more precise results than the masks it was trained on for some samples – see row 3 and 5.



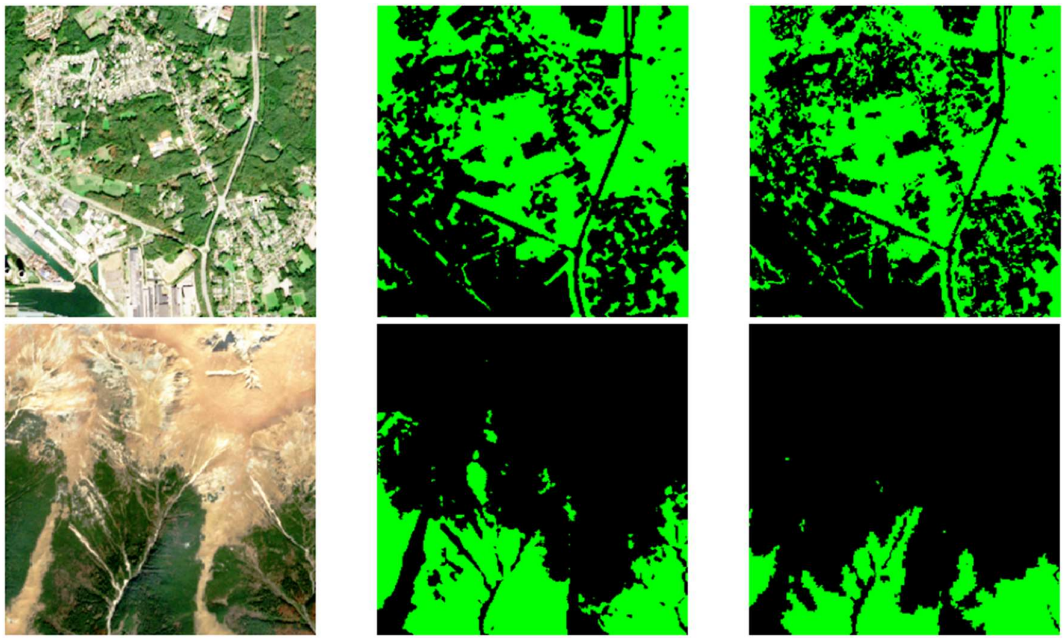


Figure 46. Samples demonstrating our model's predictions– Left: Sentinel 2 patch, Middle: model prediction and Right: HRL mask.

2.3.1.3.5 Dominant Leaf Type Results

Beyond tree cover segmentation, we extend our analysis to predicting the dominant leaf type (DLT)—broadleaved or coniferous. While DLT classification typically benefits from multi-temporal observations, we assess how well our models can perform with a single Sentinel-2 timestamp.

To this end, we fine-tune a Custom-ResNet-18 model, pre-trained on SSL4EO-S12, using the same Sentinel-2 patches, but now with a three-class segmentation mask: Non-tree, Broadleaved trees and Coniferous trees

The model achieves an overall accuracy of 87.5% and a mean IoU (mIoU) of 78.2%. These results indicate that, while single-date Sentinel-2 imagery poses limitations for DLT classification, the model still captures relevant spectral patterns distinguishing between broadleaf and coniferous forests.

Qualitative results in Figure 47 further illustrate the model's performance. While certain areas exhibit some confusion between broadleaved and coniferous classes, the segmentation masks retain fine-grained spatial structures, suggesting that spectral differences from a single observation, combined with spatial patterns, contribute to a meaningful degree of separability.

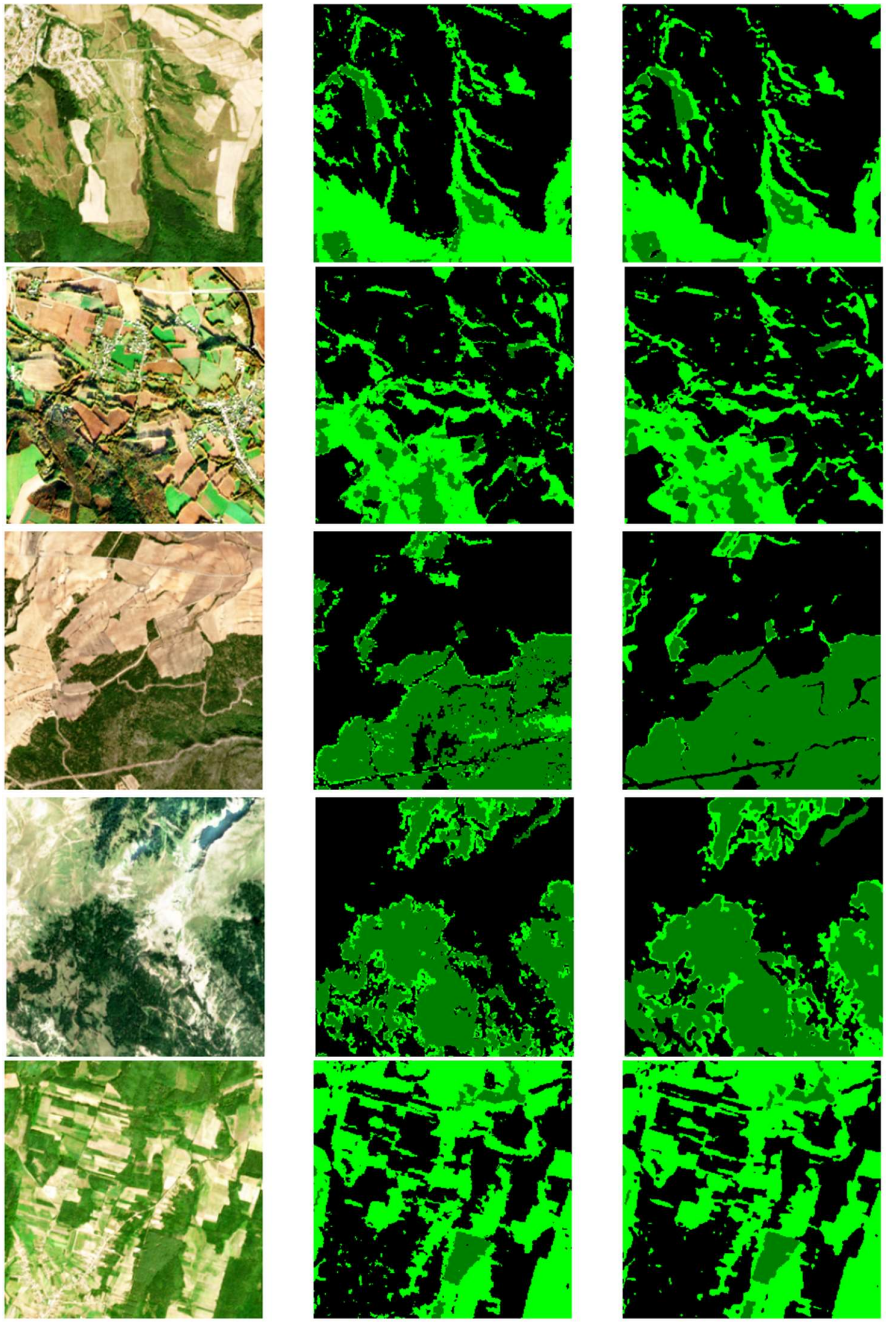


Figure 47. Qualitative results for DLT classification – Left: S2 patch, middle: HRL 2018 mask, and right: model predictions.

2.3.2 DeCUR

2.3.2.1 Classification

We pretrain SAR-optical encoders on SSL4EO-S12 and transfer them to BigEarthNet-MM, a multimodal, multi-label scene classification dataset covering 19 distinct classes. We apply late fusion for multimodal transfer learning, i.e., we concatenate the encoded features from Sentinel-1 & -2, followed by a classification layer. We report multimodal linear classification and fine-tuning results with 1% and 100% training labels in Table 10.

In most scenarios, DeCUR outperforms existing cross-modal SimCLR-like contrastive learning by 2%-4.8%, and we observe similar results in performance when fine-tuning. Compared to Barlow Twins, we improve by 0.7% and 1.4% on linear evaluation and fine-tuning with 1% labels. Corresponding figures read 2.2% and 0.2% with all labels (100%). Additionally, we report SAR-only results in Table 11. In practical scenarios cloud may render Sentinel-2 unusable, while Sentinel-1 may carry relevant signal for scene classification. DeCUR outperforms other methods in most scenarios by a large margin, while achieving comparable performance on fine-tuning with full labels. In addition, DeCUR outperforms single-modal Barlow Twins pretraining by 2.1%-2.5% with 1% labels and 0.8%-2.1% with full labels. This indicates the value of joint, multimodal pretraining.

Table 10. Benchmark of DeCUR on BigEarthNet-MM when compared to other models.

SAR-optical	1% labels		100% labels	
	Linear	**Fine-tune**	**Linear**	**Fine-tune**
Rand. Init.	58.7	58.7	70.1	70.1
Supervised	77.0	77.0	88.9	88.9
SimCLR-cross	77.4	78.7	82.8	89.6
CLIP	77.4	78.7	82.8	89.6
Barlow Twins	78.7	80.3	83.2	89.5
VICReg	74.5	79.0	81.9	89.5
DeCUR (ours)	**79.4**	**81.7**	**85.4**	**89.7**

Table 11. Benchmark of DeCUR utilizing the Sentinel-1 encoder for the downstream classification task on BigEarthNet, only.

SAR	1% labels		100% labels	
	Linear	**Fine-tune**	**Linear**	**Fine-tune**
Rand. Init.	50.0	50.0	54.2	54.2
Supervised	67.5	67.5	81.9	81.9
SimCLR-cross	68.1	70.4	71.7	83.7
CLIP	68.0	70.2	71.7	83.4
Barlow Twins	72.3	73.7	77.8	83.6
VICReg	69.3	71.9	74.1	83.6

Barlow Twins-SAR	71.2	73.3	77.5	81.6
DeCUR (ours)	**73.7**	**75.4**	**78.3**	**83.7**

2.3.2.2 Semantic Segmentation

We pretrain and evaluate RGB-DEM encoders on GeoNRW for semantic segmentation (10 classes). We use simple fully convolutional networks (FCN) as the segmentation model, which concatenates the last three-layer feature maps from both modalities, up-samples and sums them for the prediction map. Like with the classification task, linear classification is conducted by freezing the encoder, and fine-tuning is conducted by training all model parameters. Mean intersection over union (mIoU) is used as the evaluation metric.

Table 12. DeCUR evaluated for semantic segmentation on the GeoNRW dataset.

RGB-DEM	**1% labels**		**100% labels**	
	Linear	**Fine-tune**	**Linear**	**Fine-tune**
Rand. Init.	14.1	14.1	23.0	23.0
Supervised	22.1	22.1	44.0	44.0
SimCLR-cross	23.0	30.2	35.2	47.3
CLIP	22.8	28.8	35.0	46.7
Barlow Twins	31.2	33.6	43.0	48.4
VICReg	27.4	32.8	38.0	45.1
DeCUR (ours)	**34.9**	**36.9**	**43.9**	**48.7**

Table 13. DeCUR evaluated for semantic segmentation on the GeoNRW dataset, but without feeding digital elevation (DEM) information when compared to Table 12.

RGB	**1% labels**		**100% labels**	
	Linear	**Fine-tune**	**Linear**	**Fine-tune**
Rand. Init.	14.2	14.2	18.5	18.5
Supervised	17.5	17.5	38.8	38.8
SimCLR-cross	20.1	25.9	29.6	42.5
CLIP	20.0	25.7	29.4	42.3
Barlow Twins	29.4	33.4	38.0	45.9
VICReg	23.7	28.7	32.4	41.6
BarlowTwins-RGB	28.6	32.6	36.2	45.7
DeCUR (ours)	**31.4**	**34.5**	**43.9**	**46.5**

We report multimodal, linear classification and fine-tuning results with 1% and 100% training labels in Table 12. Promisingly, DeCUR outperforms other methods in all scenarios by a large margin (up to 12.1% compared to CLIP). Meanwhile, we report RGB-only results in Table 13 - in practice, DEM data is not always available. DeCUR

demonstrates a significant improvement compared to others in all scenarios (up to 11.4% compared to CLIP).

2.3.2.3 Interpretability and feature decoupling

To gain insight into how shared and modality-specific information are encoded in the embeddings Z, we calculate the cross-modal alignment of every embedding dimension. This is done by counting the diagonal elements of the cross-correlation matrix `C'=corr(Z1', Z2')`, i.e., we compute the trace of C'. The closer diagonal values reach 1, the better such embedding values have been aligned for a given dimension.

We plot the histograms of all diagonal elements of C' for a random batch of images - for both, DeCUR and cross-modal Barlow Twins. The former explicitly decouples unique dimensions (decoup.), while the latter assumes that all dimensions are common (w/o decoup.). As depicted by Figure 48, for DeCUR when compared to Barlow Twins, the alignment of dimensions is strong, and a small fraction of dimensions strongly misaligns, as expected per our design of the unique representations.

We further visualize such effects with t-SNE by clustering a batch of images where each embedding dimension becomes a separate data point. This analysis contrasts with the standard t-SNE setting where each input sample constitutes a data point. Figure 49 illustrates modality-unique dimensions are well separated, and common dimensions are perfectly overlapped.

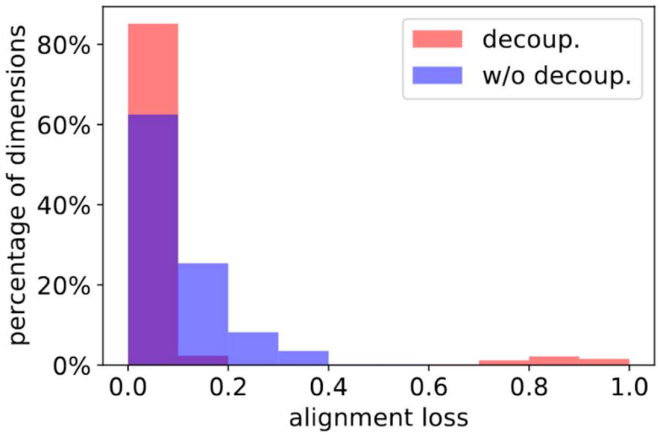


Figure 48. Analysis of embedding alignments for DeCUR (red) vs. Barlow Twins (blue).

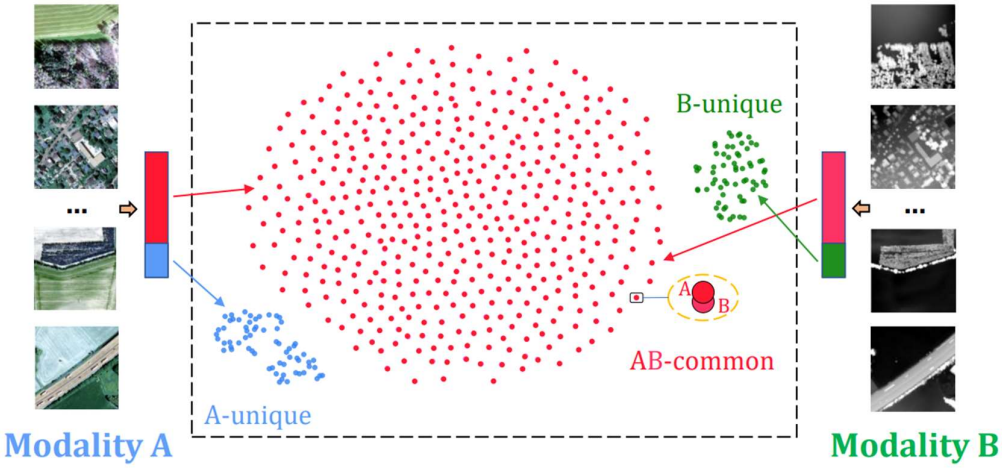
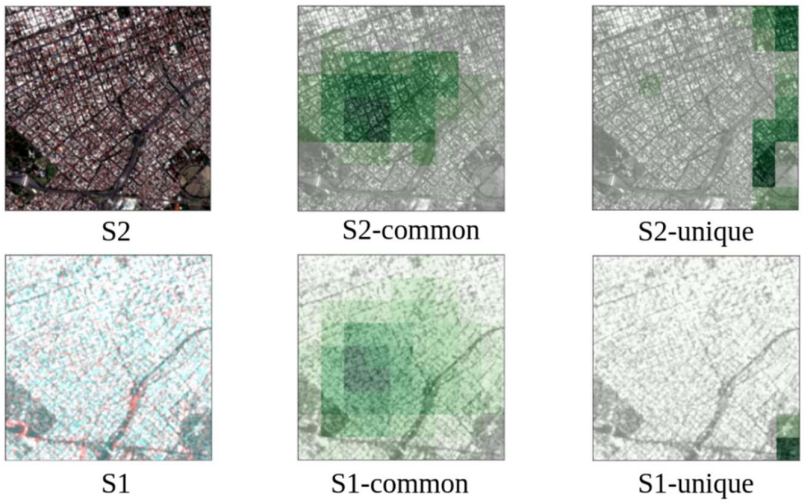


Figure 49. t-SNE analysis of unique (blue/green) vs. common (red) representations for Sentinel-1 & -2 joint embeddings generated by DeCUR.

Moreover, we employ GradCAM (Selvaraju et al., 2017) to visualize the spatial saliency of input modalities corresponding to the common and unique embedding representations. For preparation, we separately average the common and unique dimensions. Subsequently, a backpropagation pass is performed w.r.t the corresponding output target: 0 for common and 1 for unique. This generates the GradCAM saliency map after the last convolutional layer of DeCUR. Then, we up-sample the saliency maps to the size of the input for overlay. A *common* and a *unique* saliency map is generated per modality. We present an example in Figure 50 (top) for Sentinel-1 & -2 imagery (S1 and S2). While S1-common and S2-common highlight (in green) similar geospatial areas, S1-unique and S2-unique focus on separate geospatial regions.



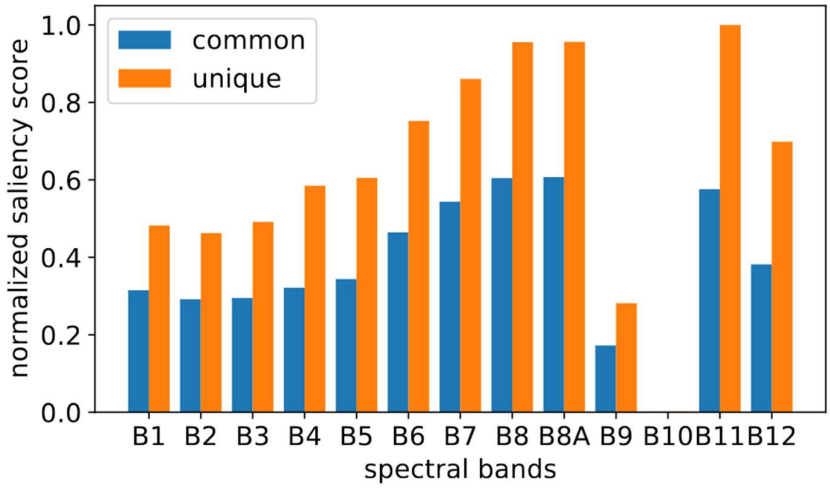


Figure 50. GradCAM analysis of common and unique embeddings with DeCUR (top). Integrated Gradient analysis of Sentinel-2 spectral bands (bottom).

To analyse spectral saliency for Sentinel-2 imagery we apply Integrated Gradients (Sundararajan et al., 2017) counting statistics over spectral bands. An importance score gets assigned to each input feature by approximating the integral of gradients of the output (0 for common and 1 for unique, as with GradCAM above) w.r.t. the inputs. We average the importance scores of each band to obtain the spectral saliency for both, common and unique representations. We normalize the scores and do statistics over the entire SSL4EO-S12 dataset. Figure 50 (bottom) presents the corresponding histograms. The band-wise importance distribution appears consistent with the domain knowledge:

- near-infrared bands (B5-B8A, including vegetation red edge) are relevant
- red (B4) is more important compared to blue (B2)
- water vapour (B9) and cirrus (B10) are strongly related to the atmosphere and thus, they render less important for land surface monitoring

2.3.3 SoftCon

2.3.3.1 Land Cover Classification

SoftCon achieves state-of-the-art (SOTA) performance across multiple land cover classification tasks, see Figure 51. On BigEarthNet-10%, SoftCon significantly outperforms existing methods. ResNet50 improves upon the best existing models by 2.7% in linear probing and 1.6% in fine-tuning, with a 6.2% and 5.2% improvement over SeCo, respectively. Remarkably, this is the first time a self-supervised ResNet50 surpasses fully supervised training from scratch, demonstrating the strength of contrastive learning with SoftCon. Additionally, SoftCon's ResNet50 outperforms several large Vision Transformer (ViT) models, further validating the effectiveness of ConvNet backbones for Earth observation (EO) tasks.

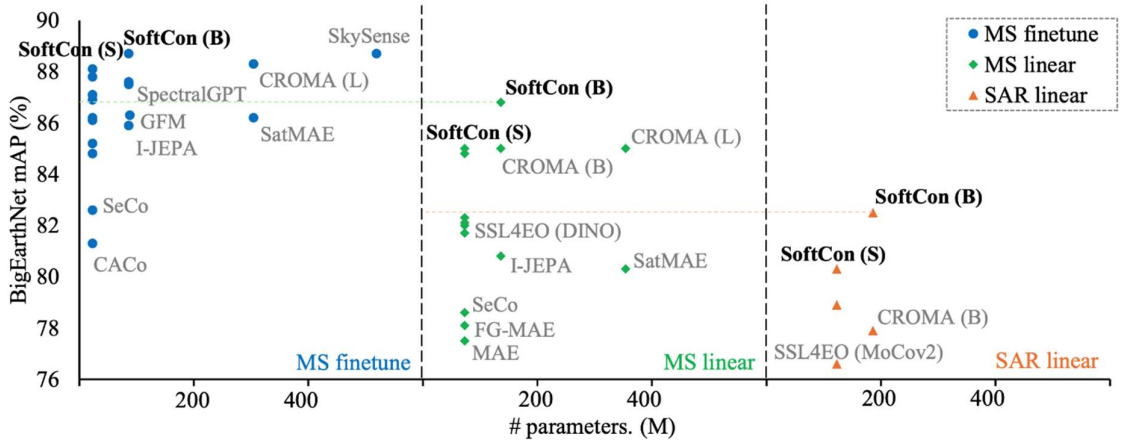


Figure 51. A detailed comparison of transfer learning performances on BigEarthNet-10%. S/B/L represents ViT-small/base/large. SoftCon (ours) achieves SOTA results with lighter backbones on both linear probing and fine-tuning, and both multispectral and SAR.

For ViT architectures, SoftCon's ViT-Small achieves performance comparable to larger models such as SpectralGPT, CROMA, and SkySense, while ViT-Base reaches 86.8% mAP on BigEarthNet-10% in linear probing. Notably, this performance is achieved using single-modality pretraining (MS vs. MS+SAR), fewer parameters (ViT-S vs. ViT-L), and significantly fewer training epochs (100 vs. 600). Similarly, SoftCon achieves fine-tuning results on par with SkySense, while requiring much less pretraining data.

Table 14. Linear probing results on BigEarthNet-SAR-10% and EuroSAT-SAR.

Method	Backbone	BE-SAR-10%	EuroSAT-SAR
MoCo-v2	RN50	76.6	82.4
SoftCon (ours)	RN50	**78.9**	**87.1**
MAE	ViT-S	69.8	79.3
FG-MAE	ViT-S	71.7	80.7
CROMA	ViT-B	77.9	87.5
SoftCon (ours)	ViT-S	80.3	87.1
SoftCon (ours)	ViT-B	**82.5**	**89.1**

For SAR-based classification (see Table 14 and Table 15), SoftCon substantially improves performance on BigEarthNet-SAR and EuroSAT-SAR, confirming its adaptability across modalities. On BigEarthNet-SAR, SoftCon's ResNet50 surpasses MoCo-v2 by 2.3%, while achieving a 4.7% improvement on EuroSAT-SAR. SoftCon's ViT-S outperforms MAE and FG-MAE by up to 11.6%, demonstrating that contrastive learning is more effective than masked image modeling for producing strong out-of-the-box representations. Furthermore, SoftCon's ViT-Base achieves 81.4% on BigEarthNet-SAR, outperforming CROMA by 3.5%.

Table 15. Linear probing/fine-tuning accuracy on fMoW-sentinel-10%.

Method	Backbone	Linear	Finetune
DINO	ViT-S/16	32.6	52.8
MAE	ViT-S/16	27.7	51.8
SatMAE	ViT-S/16	35.2	57.2
I-JEPA	ViT-S/16	32.4	53.5
CROMA	ViT-L/8	39.2	59.0
SoftCon (ours)	ViT-S/14	39.9	59.7
SoftCon (ours)	ViT-B/14	**44.0**	**60.6**

SoftCon also achieves SOTA performance on fMoW-Sentinel, a more challenging land cover classification dataset. ViT-Small already outperforms state-of-the-art models, while ViT-Base outperforms the best prior model (CROMA) by 4.8% in linear probing and 1.6% in fine-tuning. These results further validate the effectiveness of SoftCon's multi-label supervision and cross-domain pretraining strategies.

2.3.3.2 Land Cover Segmentation

SoftCon also delivers strong results in land cover segmentation, as illustrated by Table 16. On the DFC2020 segmentation benchmark, SoftCon's ResNet50 outperforms MoCo-v2 by 3%, while also surpassing several large-scale ViT models. ViT-S improves upon CROMA by 0.9%, and ViT-Base further refines performance by an additional 0.3%, highlighting the effectiveness of SoftCon's representations. Additionally, qualitative results indicate that SoftCon captures more fine-grained and accurate semantic information than SSL4EO-S12, further demonstrating its advantages for downstream EO tasks.

Table 16. Fine-tuning mIoU results on DFC2020.

Method	Backbone	mIoU
Supervised	RN50	42.9
MoCo-v2	RN50	47.3
SoftCon (ours)	RN50	**50.3**
MAE	ViT-S	48.0
SatMAE	ViT-L	44.1
CROMA	ViT-L	49.8
SoftCon (ours)	ViT-S	50.7
SoftCon (ours)	ViT-B	**51.0**

2.3.4 CromSS

2.3.4.1 Transfer Learning Performance on DFC2020

Our results on the DFC2020 dataset are summarized in Table 17. CromSS significantly improves remote sensing segmentation when pretrained on Sentinel-2 (S2) multispectral data, outperforming state-of-the-art SSL methods across most evaluation metrics. Middle fusion performs slightly better than late fusion, confirming the benefit of stronger feature alignment across modalities. However, for Sentinel-1 (S1) radar data, the improvement is less pronounced. The higher sensitivity of S1 to label noise is likely responsible for this discrepancy, suggesting that stronger cross-modal strategies are needed to improve segmentation performance on radar imagery in future research.

Table 17. Transfer learning results (%) on the DFC2020 dataset from PSPNet with ResNet-50 backbones and UPerNet with ViT-Large backbones. 3B/9B/10B/13B in brackets indicate the number of S2 bands as inputs for each model. Frozen and fine-tuned represent no updates of encoders and optimizing encoders along with decoders in the transfer learning setting. The best results are highlighted in bold. The annotations are the same in the following tables.

Inputs	Encoder		Frozen				Fine-tuned			
	Backbone	Weights (#S2 bands)	OA	AA	mIoU	mF1	OA	AA	mIoU	mF1
S1	ViT-Large	DOFA	49.15± 1.89	39.59± 1.40	27.51± 0.72	36.74± 1.13	57.41± 1.44	43.74± 1.37	31.12± 1.43	41.35± 1.80
	ResNet50	random	53.89± 0.56	40.65± 0.38	28.96± 0.16	38.74± 0.22	52.03± 0.56	42.02± 0.34	28.02± 0.23	37.45± 0.29
		MoCo	60.74± 0.36	**47.48± 0.49**	34.17± 0.27	44.40± 0.36	60.30± 0.45	45.00± 0.65	31.78± 0.42	41.78± 0.42
		DeCUR	60.22± 0.18	46.75± 0.58	34.02± 0.18	44.24± 0.35	57.79± 0.48	44.53± 0.55	31.23± 0.40	41.20± 0.39
		CromSS-midF	60.75± 0.30	45.66± 0.66	34.12± 0.28	**44.47± 0.31**	60.07± 0.50	**46.44± 0.63**	33.48± 0.43	**44.05± 0.48**
		CromSS-lateF	**61.65± 0.34**	44.80± 0.09	**34.20± 0.10**	44.14± 0.15	**60.69± 0.19**	45.75± 0.36	**33.63± 0.12**	43.86± 0.10
S2	ViT-Large	scaleMAE (3B)	63.49± 0.26	53.31± 0.29	38.51± 0.20	49.90± 0.12	68.42± 0.46	55.13± 0.25	40.87± 0.31	52.18± 0.27
		DOFA (9B)	63.49± 0.58	48.79± 0.20	37.19± 0.14	48.87± 0.44	64.44± 0.50	49.50± 0.11	37.65± 0.31	48.91± 0.59
		satMAE (10B)	67.02± 0.41	51.91± 0.04	40.13± 0.16	51.52± 0.31	66.47± 0.21	50.51± 0.18	39.02± 0.16	50.48± 0.66
		satMAE++ (10B)	68.41± 0.06	53.15± 0.77	40.96± 0.27	52.93± 0.32	67.21± 0.17	51.44± 0.04	39.90± 0.26	51.32± 0.66
	ResNet50	random (9B)	61.16± 1.32	46.76± 0.34	34.00± 0.53	44.76± 0.61	63.63± 0.53	50.28± 0.33	36.57± 0.31	47.58± 0.21
		random (13B)	56.11± 0.55	44.63± 0.22	31.26± 0.14	41.89± 0.19	58.83± 0.87	45.89± 0.36	33.63± 0.24	44.14± 0.29
		Satlas (9B)	65.91± 0.35	51.03± 0.41	38.74± 0.33	50.27± 0.46	64.76± 0.35	52.02± 0.61	37.85± 0.13	49.23± 0.21
		DINO (13B)	64.65± 0.12	48.53± 0.30	37.65± 0.13	48.77± 0.15	63.75± 0.70	49.80± 1.06	37.04± 0.55	48.95± 0.73
		MoCo (13B)	63.19± 0.70	50.76± 0.34	37.57± 0.72	49.83± 0.60	60.98± 0.64	47.07± 0.36	34.75± 0.74	46.64± 0.40
		DeCUR (13B)	63.02± 0.70	47.99± 0.57	36.82± 0.37	47.90± 0.44	64.69± 1.03	50.90± 0.97	38.02± 0.94	50.10± 1.10

		CromSS-midF (9B)	66.00± 0.15	55.81± 0.07	40.31± 0.40	52.59± 0.17	69.00± 0.26	**57.07± 0.07**	42.17± 0.13	54.02± 0.24
		CromSS-lateF (9B)	68.95± 0.67	**56.01± 0.41**	41.69± 0.78	53.61± 0.94	68.82± 0.17	55.75± 0.10	41.51± 0.28	53.34± 0.20
		CromSS-midF (13B)	**69.56± 0.32**	54.61± 0.25	**42.52± 0.11**	**54.12± 0.16**	**70.41± 0.35**	55.11± 0.30	**43.21± 0.16**	**54.91± 0.23**
		CromSS-lateF (13B)	67.96± 0.49	52.62± 0.64	41.38± 0.34	53.25± 0.49	68.49± 0.49	53.19± 0.59	41.91± 0.47	53.59± 0.39

2.3.4.2 Transfer to SSL4EO-S12@DW

CromSS demonstrates a significant advantage over other SSL methods when applied to SSL4EO-S12@DW, proving that noisy labels can be effectively leveraged for pretraining. However, an interesting observation is that models using fewer bands (9B) perform better than those using all 13 bands (13B). This is likely due to the low-resolution spectral bands introducing more ambiguity in classifying the dataset's 9 classes.

For S1 data, CromSS-pretrained encoders initially underperform compared to DeCUR, which benefits from modality decoupling during pretraining. However, after fine-tuning, CromSS models achieve comparable results, highlighting its potential for SAR-based segmentation.

Visual segmentation maps reveal that CromSS-pretrained models capture finer details than competing SSL methods, particularly when encoder weights remain fixed. However, these maps also exhibit more pepper-and-salt noise, suggesting that while CromSS improves detail sensitivity, it may also amplify minor inconsistencies or artifacts. This trade-off could be advantageous for applications where preserving fine structures is critical.

Table 18. Transfer learning results (%) on the SSL4EO-S12@DW dataset from DeepLabV3+ with ResNet-50 and UPerNet with ViT-Large.

Inputs	**Encoder**		**Frozen**				**Fine-tuned**			
	Backbone	**Weights (#S2 bands)**	OA	AA	mIoU	mF1	OA	AA	mIoU	mF1
S1	**ViT-Large**	**DOFA**	62.65± 0.20	51.53± 0.26	36.92± 0.34	49.97± 0.45	66.91± 0.24	57.75± 0.83	41.31± 0.19	54.66± 0.10
	ResNet50	**random**	56.84± 0.53	44.67± 0.55	29.85± 0.45	41.10± 0.58	65.77± 0.28	54.98± 0.41	39.80± 0.36	53.04± 0.50
		MoCo	68.47± 0.20	62.24± 0.22	45.19± 0.47	58.95± 0.48	68.69± 0.22	60.71± 0.69	**44.76± 0.37**	**58.41± 0.42**
		DeCUR	**69.38± 0.16**	**62.49± 0.38**	**45.53± 0.08**	**59.36± 0.21**	**69.10± 0.18**	**61.67± 1.62**	44.72± 0.46	58.37± 0.45
		CromSS-midF (13B)	66.25± 0.16	54.29± 0.23	39.63± 0.12	52.17± 0.26	68.04± 0.54	59.75± 1.19	43.71± 0.42	57.21± 0.58
		CromSS-lateF (13B)	66.03± 0.14	53.05± 0.12	39.63± 0.14	52.22± 0.18	69.00± 0.44	60.47± 1.01	44.00± 0.82	57.29± 0.94
S2	**ViT-Large**	**scaleMAE (3B)**	75.89± 1.23	72.60± 1.67	57.37± 1.33	68.95± 1.40	78.80± 1.50	73.57± 2.12	59.89± 2.07	72.67± 1.78
		DOFA (9B)	82.18± 0.27	72.07± 0.09	58.45± 0.40	71.21± 0.39	82.25± 0.16	73.24± 0.89	60.11± 1.13	72.53± 0.93

		satMAE (10B)	77.54± 0.10	71.34± 0.10	58.37± 0.11	71.52± 0.05	77.65± 0.11	71.84± 0.36	57.84± 0.26	70.58± 0.19
		satMAE++ (10B)	76.64± 0.18	70.44± 0.18	57.04± 0.17	69.71± 0.05	78.18± 0.10	72.83± 0.24	58.87± 0.23	71.54± 0.22
	ResNet50	**random (9B)**	66.79± 0.32	62.68± 0.47	48.22± 0.32	61.22± 0.37	74.64± 0.43	70.91± 0.75	56.51± 0.68	69.56± 0.70
		random (13B)	64.80± 0.47	60.30± 0.20	46.21± 0.33	59.11± 0.30	73.36± 0.21	68.58± 0.49	54.72± 0.25	67.76± 0.26
		Satlas (9B)	75.15± 0.10	70.26± 0.15	56.39± 0.12	69.29± 0.24	77.92± 0.67	72.59± 0.86	59.17± 1.02	71.87± 1.00
		DINO (13B)	77.64± 0.16	72.55± 0.53	59.36± 0.46	71.98± 0.50	77.81± 0.23	71.36± 0.58	58.05± 0.26	70.38± 0.39
		MoCo (13B)	77.44± 0.62	73.05± 0.40	59.52± 0.90	72.14± 0.83	77.53± 0.08	72.76± 0.61	58.82± 0.56	71.57± 0.48
		DeCUR (13B)	76.37± 0.18	71.28± 0.51	57.45± 0.15	70.15± 0.28	77.20± 0.22	72.05± 0.30	58.54± 0.39	71.32± 0.33
		CromSS-midF (9B)	78.85± 0.18	**73.77± 0.28**	**61.39± 0.09**	**73.79± 0.10**	79.19± 0.34	73.89± 0.56	61.12± 0.43	73.64± 0.37
		CromSS-lateF (9B)	**78.98± 0.21**	73.39± 0.60	60.93± 0.60	73.37± 0.52	**79.93± 0.14**	**74.03± 0.45**	**62.05± 0.24**	**74.18± 0.20**
		CromSS-midF (13B)	78.60± 0.35	73.04± 0.23	60.56± 0.28	73.01± 0.24	78.64± 0.19	73.16± 0.14	60.79± 0.12	73.20± 0.13
		CromSS-lateF (13B)	78.48± 0.16	72.24± 0.18	60.50± 0.13	72.86± 0.12	79.33± 0.24	73.46± 0.32	61.10± 0.66	73.33± 0.61

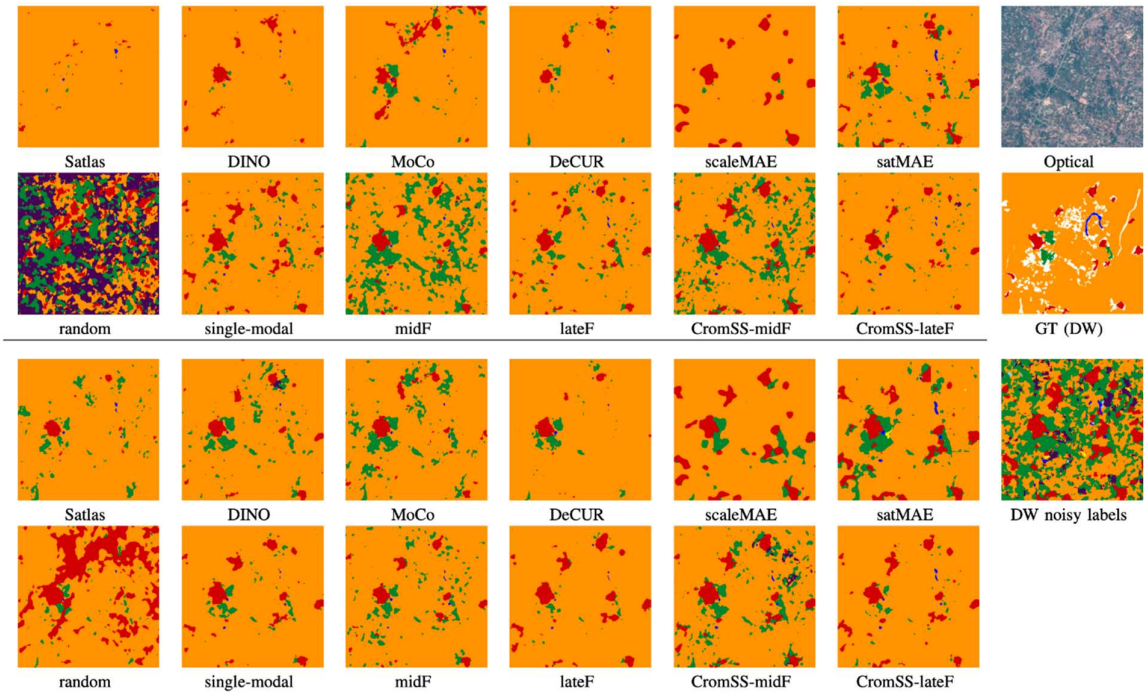


Figure 52. The segmentation maps obtained on the SSL4EO-S12@DW downstream dataset with S2 as inputs from UperNet with ViT-large pretrained by scaleMAE and satMAE, and the rest from PSPNets with ResNet-50 pretrained by different methods (upper: with frozen encoders, lower: with fine-tuned encoders).

2.3.5 SpectralEarth

Our key findings are summarized in Table 19. We include random initialization baselines on each downstream task, along with our pretrained models. Due to 202 spectral channels in SpectralEarth patches, a comparison with RGB-ImageNet weights is of little value.

Since the last version of the report, we have updated our results and conducted more experiments. We report our latest experiments below.

2.3.5.1 Comparison of SSL Algorithms

Table 19 summarizes our results for three SSL algorithms with the Spec. RN50 and Spec. ViT-S backbones.

Table 19. Performance of SpectralEarth hyperspectral model on 7 downstream tasks. Best performance per model architecture is highlighted in bold, and the second best is underlined.

Model Configuration			Classif. (F1)	Segmentation (mIoU)					
Arch.	Weights	Protocol	CORINE	CDL	NLCD	EUROCROPS	TREEMAP	BDFORET	BNETD
Spec. RN50	Random	Frozen	69.92 ± 0.51	62.65 ± 0.80	35.94 ± 0.38	54.83 ± 0.83	35.87 ± 0.23	56.42 ± 0.85	35.67 ± 0.22
		Full FT	**78.01 ± 0.62**	77.28 ± 0.13	47.95 ± 0.04	68.93 ± 1.41	46.47 ± 0.04	75.57 ± 0.53	48.59 ± 0.14
	MoCo-V2	Frozen	74.07 ± 0.14	70.97 ± 0.19	41.87 ± 0.06	61.70 ± 0.08	41.67 ± 0.05	67.01 ± 0.11	41.55 ± 0.05
		Full FT	**77.53 ± 0.25**	**48.11 ± 0.03**	**70.95 ± 0.63**	46.94 ± 0.03	**77.50 ± 0.35**	46.94 ± 0.03	
		Adapter	76.82 ± 0.39	75.84 ± 0.11	44.62 ± 0.03	66.71 ± 0.23	44.84 ± 0.04	73.24 ± 0.16	44.84 ± 0.04
	DINO	Frozen	76.61 ± 0.31	71.02 ± 0.13	41.83 ± 0.03	60.29 ± 0.29	42.03 ± 0.03	66.20 ± 0.17	42.03 ± 0.03
		Full FT	77.71 ± 0.55	77.49 ± 0.23	47.62 ± 0.07	69.52 ± 1.00	**47.08 ± 0.13**	75.45 ± 0.61	**49.21 ± 0.18**
		Adapter	76.96 ± 0.37	75.51 ± 0.15	44.38 ± 0.03	66.62 ± 0.35	44.70 ± 0.07	72.90 ± 0.15	44.70 ± 0.07
Spec. ViT-S	Random	Frozen	70.59 ± 0.62	65.96 ± 0.72	36.94 ± 0.30	55.78 ± 0.52	38.72 ± 0.25	60.61 ± 0.84	38.64 ± 0.20
		Full FT	77.63 ± 0.37	74.77 ± 0.22	45.31 ± 0.62	66.87 ± 0.62	44.43 ± 0.13	73.10 ± 0.51	44.35 ± 0.13
	MoCo-V2	Frozen	73.99 ± 0.04	71.40 ± 0.32	39.99 ± 0.08	60.21 ± 0.39	41.76 ± 0.06	67.06 ± 0.25	41.84 ± 0.12
		Full FT	78.37 ± 0.34	75.59 ± 0.14	45.92 ± 0.05	67.27 ± 0.29	45.03 ± 0.10	73.97 ± 0.41	44.58 ± 0.48
		Adapter	76.72 ± 0.35	75.28 ± 0.19	43.18 ± 0.08	66.46 ± 0.21	44.09 ± 0.07	72.57 ± 0.53	44.10 ± 0.08
	DINO	Frozen	75.11 ± 0.13	71.56 ± 0.35	40.47 ± 0.06	58.76 ± 0.27	41.89 ± 0.13	66.08 ± 0.82	41.89 ± 0.13
		Full FT	78.67 ± 0.34	76.10 ± 0.31	45.64 ± 0.15	65.56 ± 0.76	45.44 ± 0.15	72.30 ± 0.38	45.32 ± 0.06
		Adapter	77.00 ± 0.50	74.83 ± 0.07	42.86 ± 0.08	64.78 ± 0.65	43.54 ± 0.11	70.85 ± 0.13	43.62 ± 0.20
	MAE	Frozen	72.85 ± 0.20	72.01 ± 0.28	41.08 ± 0.05	62.14 ± 0.18	41.19 ± 0.15	68.54 ± 0.35	41.07 ± 0.15
		Full FT	**79.17 ± 0.55**	**77.16 ± 0.31**	**47.61 ± 0.10**	**69.87 ± 0.16**	**45.80 ± 0.15**	**74.75 ± 0.55**	**45.50 ± 0.13**
		Adapter	76.75 ± 0.35	75.21 ± 0.21	43.64 ± 0.07	67.18 ± 0.58	43.45 ± 0.14	74.06 ± 0.26	43.33 ± 0.13

2.3.5.1.1 Classification Results

Pretrained models outperform random weights in linear evaluation for MoCo-V2 and DINO on EnMAP-CORINE, with DINO yielding the best results, demonstrating the ability of joint-embedding methods to learn robust representations, even in the absence of fine-tuning.

Additionally, we observe that MAE under-performs in linear probing, which is consistent with existing literature. This can be attributed to the reconstruction pre-training objective, which focuses on low-level features, whereas joint-embedding methods learn more global and semantic representations. For full fine-tuning, CNN pretrained models with MoCo and DINO are on par with/slightly worse than training from scratch. On the contrary, the Spec. ViT-S model benefits from pre-

training in all settings compared to random initialization, with MAE yielding the best results.

Fine-tuning the spectral adapter yields a notable performance boost compared to linear evaluation. This significant performance improvement aligns with reports in the literature: fine-tuning early network layers is an efficient compromise between frozen weights vs. complete fine-tuning.

In terms of F1-score, only fine-tuning the adapter (0.3% of network parameters) is, in most settings, less than one point below training from scratch, highlighting the efficient adaptability of our pretrained models.

2.3.5.1.2 Segmentation Results

Pretrained models with a frozen encoder consistently outperform random weights across all segmentation tasks. This shows that even without fine-tuning, the models learn useful features for segmentation. MoCo and DINO pre-trained models yield similar results for both the CNN and the ViT model. However, MAE as a feature extractor consistently outperforms MoCo and DINO (for the same backbone). This can be attributed to the pixel lvl objective, more suitable for pixel level tasks (as opposed to image lvl problems).

For full fine-tuning, MoCo-V2 pretrained CNNs consistently outperform training from scratch, while DINO is on par with the random initialization baseline. Notably, the gaps are more pronounced for BDForet and Eurocrops, which are the dataset with the lowest estimated level of label noise. For the ViT model, MAE + fine-tuning performs best in all settings, yielding improvements of up to 3 points of mIoU. The second-best models are, depending on the dataset, either MoCo or DINO, with smaller gaps w.r.t. training from scratch.

Fine-tuning the spectral adapter significantly outperforms frozen encoder performance in all settings, showing the importance of tailoring the spectral feature extraction to the characteristics of the downstream task. Notably, fine-tuning the spectral adapter outperforms training from scratch on CDL, EuroCrops and BD-Foret.

Compared to CNNs, ViTs perform slightly worse than CNN on segmentation problems. This can be attributed to multiple factors: 1. the limited size of the fine-tuning dataset, even combined with the larger pre-training data, is not sufficient to compensate for the lack of inductive biases of ViTs; 2. the spatial patterns of the masks, which present high frequency features, partly due to label noise, are harder to learn with a ViT due to the fixed patch size; compared to CNNs which can preserve full resolution feature maps. This effect can be better observed in the patch size experiment below.

2.3.5.2 Scaling up pre-trained models

Recent studies have shown significant performance gains through model scaling in SSL, e.g. DINOV2. Consequently, recent geospatial foundation models are exploring ViT architectures with hundreds of millions of parameters, sometimes exceeding a billion parameters. We investigate model scaling on SpectralEarth by pretraining Spec. ViT architectures of varying sizes: Base (B), Large (L), Huge (H), and Giant (g). Our findings are presented in Table 20. We observe that MAE consistently outperforms training from scratch across all model sizes. Remarkably, fine-tuning the adapter, which corresponds to 56K trainable parameters, matches or surpasses training from scratch for CORINE, CDL, EuroCrops and BDForet. Moreover, we observe initial benefits from scaling up the models up to Spec. ViT-L/H, especially for frozen encoder performance. However, additional increases in model size do not improve performance, possibly due to the size of the pretraining dataset. While SpectralEarth is significantly larger than previous HSI datasets, it remains smaller than the datasets used in computer vision for training state-of-the-art models. This highlights the importance of scaling the data when training large vision transformers in remote sensing.

Table 20. Performance of SpectralEarth hyperspectral ViT models pre-trained with MAE on 7 downstream tasks. Best performance per model architecture is highlighted in bold, and the second best is underlined.

Model Configuration		Classif. (F1)	Segmentation (mIoU)					
Arch.	Protocol	CORINE	CDL	NLCD	EUROCROPS	TREEMAP	BDFORET	BNETD
Spec. ViT Models								
Spec. ViT-B	Random Full FT	77.17 ± 0.72	74.65 ± 0.60	45.57 ± 0.14	66.14 ± 0.43	44.51 ± 0.13	73.55 ± 0.39	44.41 ± 0.07
	MAE Frozen	75.04 ± 0.30	72.20 ± 0.16	40.82 ± 0.14	61.29 ± 0.22	39.74 ± 0.27	68.39 ± 0.20	39.74 ± 0.27
	MAE Full FT	79.49 ± 0.26	**77.44 ± 0.26**	47.59 ± 0.23	**69.34 ± 0.43**	46.10 ± 0.04	76.30 ± 0.52	**45.69 ± 0.09**
	MAE Adapter	77.63 ± 0.25	75.30 ± 0.32	43.81 ± 0.04	67.31 ± 0.43	43.65 ± 0.18	74.51 ± 0.15	43.71 ± 0.08
Spec. ViT-L	Random Full FT	76.85 ± 0.67	74.28 ± 0.34	45.49 ± 0.13	65.66 ± 0.46	44.52 ± 0.08	72.93 ± 0.59	44.56 ± 0.11
	MAE Frozen	73.98 ± 0.21	72.72 ± 0.23	42.67 ± 0.03	62.13 ± 0.18	42.62 ± 0.19	71.07 ± 0.31	42.66 ± 0.13
	MAE Full FT	79.25 ± 0.65	77.41 ± 0.26	47.84 ± 0.03	68.73 ± 0.36	**46.10 ± 0.17**	76.26 ± 0.73	44.62 ± 0.04
	MAE Adapter	77.24 ± 0.07	75.94 ± 0.13	44.15 ± 0.07	67.69 ± 0.32	44.09 ± 0.09	74.89 ± 0.26	43.82 ± 0.14
Spec. ViT-H	Random Full FT	77.18 ± 0.44	73.72 ± 0.47	45.33 ± 0.12	64.77 ± 0.48	44.45 ± 0.17	72.98 ± 0.45	44.35 ± 0.21
	MAE Frozen	73.96 ± 0.30	71.68 ± 0.21	41.87 ± 0.13	61.94 ± 0.11	41.95 ± 0.17	68.90 ± 0.12	41.97 ± 0.11
	MAE Full FT	**79.51 ± 0.30**	77.26 ± 0.18	**47.85 ± 0.14**	69.13 ± 0.49	46.05 ± 0.13	**76.57 ± 0.35**	45.68 ± 0.14
	MAE Adapter	77.23 ± 0.32	75.34 ± 0.22	43.90 ± 0.08	66.51 ± 0.14	43.75 ± 0.06	74.13 ± 0.23	43.74 ± 0.13
Spec. ViT-g	Random Full FT	76.94 ± 0.61	73.44 ± 0.38	45.03 ± 0.04	64.40 ± 0.40	44.38 ± 0.11	72.83 ± 0.50	44.42 ± 0.05
	MAE Frozen	72.50 ± 0.26	68.39 ± 0.20	39.68 ± 0.03	57.11 ± 0.13	40.21 ± 0.07	62.99 ± 0.09	40.21 ± 0.08
	MAE Full FT	78.20 ± 0.66	76.14 ± 0.23	46.61 ± 0.17	66.75 ± 0.49	45.18 ± 0.12	73.47 ± 0.58	45.02 ± 0.11
	MAE Adapter	77.26 ± 0.50	75.25 ± 0.24	43.17 ± 0.09	66.84 ± 0.29	43.69 ± 0.05	74.06 ± 0.47	43.75 ± 0.15

2.3.5.3 Cross-Sensor Transferability

To assess the transferability of our pretrained models to other sensors, we conduct evaluations on the Hyperview, DESIS-CDL and EO1-CDL datasets. Images in both datasets have been collected by sensors with spectral characteristics different from EnMAP. In Table 21, we evaluate the Spec. ViT-L architecture across various fine-tuning protocols. We observe consistent improvements compared to training from scratch when fine-tuning both datasets. Remarkably, while the frozen encoder

performance on the EO1/DESIS-CDL datasets is relatively poor, it yields the best results for Hyperview. This can be explained by the severe overfitting observed on this dataset. For EO1/DESIS-CDL, fine-tuning the spectral adapter only is sufficient to close most of the performance gaps with fine-tuning, and significantly outperforms training from scratch.

Table 21. Cross-sensor transferability of the Spec. ViT-L model pre-trained on SpectralEarth.

Model Configuration	HYPERVIEW (Norm MSE ↓)	CDL (mIoU) ↑	
	ENMAP	EO1	DESIS
MAE Frozen	**81.29 ± 2.11**	59.46 ± 1.02	62.78 ± 0.31
MAE Full FT	86.71 ± 0.35	66.34 ± 0.77	**69.05 ± 0.50**
MAE Adapter	86.45 ± 0.38	**66.58 ± 0.41**	68.45 ± 0.16
Training from Scratch	89.19 ± 0.90	65.09 ± 0.76	66.06 ± 0.57

2.3.5.4 Computational Efficiency

Model convergence: We evaluate the training efficiency of SSL pretrained models relative to training from scratch on the CORINE and CDL benchmarks. Figure 53 demonstrates that pretrained models converge more rapidly - in particular, for early stages of model training. Notably, fine-tuning matches the 100 epochs performance of training from scratch in ~10 epochs for EnMAP-CORINE, and EnMAP-CDL.

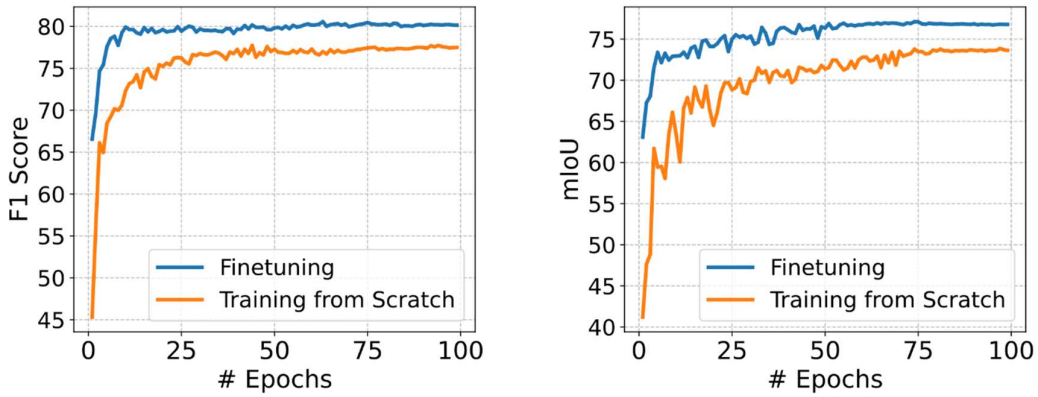
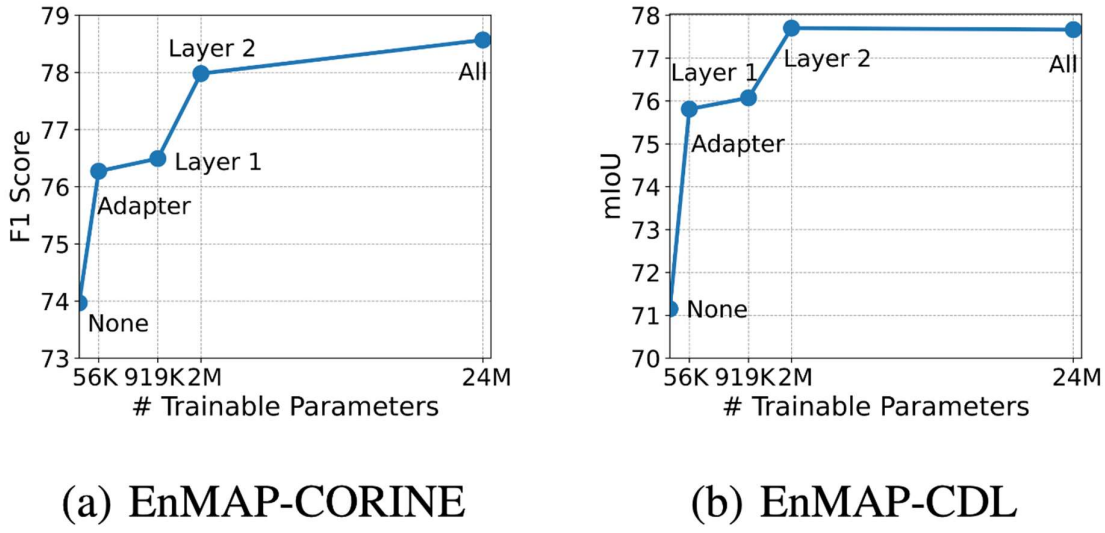


Figure 53. Evaluation of SpectralEarth pre-trained model performance on downstream tasks when compared to training the model from scratch.

Parameter-efficient fine-tuning: SSL pre-trained models are typically evaluated in linear probing to assess the quality of the learnt representations. While a frozen encoder can achieve competitive performance, fine-tuning often surpasses linear probing – trading for a higher computational cost. Our experiments with progressive unfreezing of network layers are a compromise between both

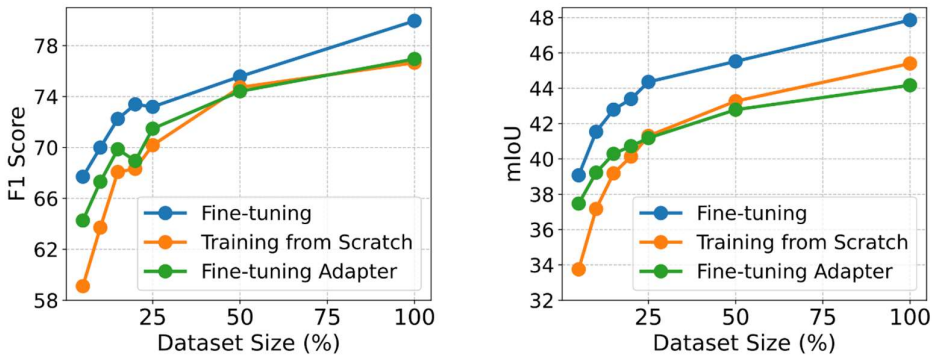
scenarios. As depicted in Figure 54, fine-tuning the first two blocks of a pretrained Spec. ResNet-50 produces results that closely resemble full fine-tuning. This implies cost- and energy-efficiency of our pretrained models for downstream tasks. It highlights the computational advantage of leveraging our pretrained models when compared to training from scratch.



(a) EnMAP-CORINE (b) EnMAP-CDL

Figure 54. Performance of Spec. ResNet-50 on downstream tasks based on the fine-tuning setup. None: linear probing, Adapter: fine-tuning of Spectral Adapter (cf. Figure 13), Layer 1: fine-tune Spectral Adapter and 1st layer of ResNet-50 backbone, Layer 2: fine-tune Spectral Adapter, 1st and 2nd layers of ResNet-50 backbone, All: full fine-tuning.

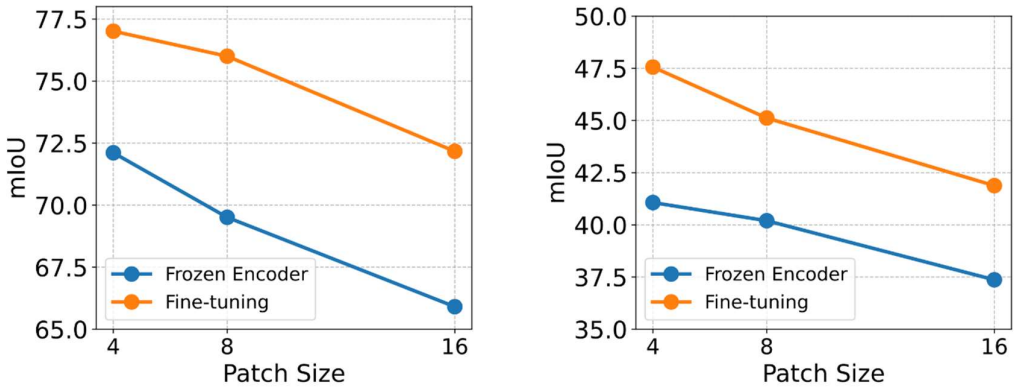
Training with limited labels: Pre-trained models are very useful when labelled data is scarce. To assess this, we evaluate the performance of the Spec. ViT-L model, pre-trained with MAE, on varying subset sizes of EnMAP-CORINE and EnMAP-NLCD datasets. The results, reported in Figure 55 demonstrate the benefits of pre-training across all data regimes. In particular, we observe large gaps of 9 points in F1 score for EnMAP-CORINE, and 5 points in mIoU for EnMAP-NLCD, when using 5% of the labels compared to training from scratch. Interestingly, in lower data regime scenarios, fine-tuning the adapter only suffices to outperform training from scratch. This demonstrates the generalizability of the features learned from pre-training.



(a) EnMAP-CORINE (b) EnMAP-NLCD

Figure 55. Impact of the size of the downstream dataset evaluated on EnMAP-CORINE and EnMAP-NLCD for the Spec. ViT-L model.

Importance of the patch size: Medium-resolution remote sensing imagery requires more fine-grained detail preservation than natural imagery, especially for segmentation tasks. This aspect is sometimes overlooked in the literature, where vision transformers with a patch size of 16x16 or 8x8 are commonly used as backbones to pretrain foundation models to reduce the computational cost. We argue that a smaller image size with a smaller patch size can lead better trade-offs, particularly for HSI imagery where pixel information is rich and should not be heavily compressed in the embedding space. To analyse the impact of the patch size, we pre-trained Spec. ViT-S using MAE with patch sizes of 4x4, 8x8 and 16x16. Results in Figure 56 show that smaller patches consistently improve performance for segmentation tasks. These results are consistent with our observations in the forest segmentation use-case, showing that applying vanilla computer vision models without extra care can lead to a loss of fine-grained details.



(a) EnMAP-CDL (b) EnMAP-NLCD

Figure 56. Impact of the size of the downstream dataset evaluated on EnMAP-CORINE and EnMAP-NLCD for the Spec. ViT-L model.

2.3.5.5 Reconstructing EnMAP from Sentinel-2

Before fine-tuning for multi-modal data fusion any hyperspectral foundation model trained by generative methods, it is paramount to investigate on the quality of reconstruction. We utilize mean square errors (MSE) evaluated overall, exclusively masked, or exclusively unmasked pixels. In addition, we compute the Structural Similarity Score (SSIM) to capture spatial-spectral consistency (by correlation analysis).

Table 22 lists these metrics for models pretrained by spectral (Spec) and spatial-spectral (Spa-Spec) masking (75% of patches).

Table 22. Quality scores of reconstructing EnMAP from EnMAP.

Masking	MSE			SSIM
Train	**Total**	**Masked**	**Unmasked**	
Spec	7.11 E-4	9.17 E-4	9.25 E-5	0.7372
Spa-Spec	**4.27 E-4**	**5.6 E-4**	**2 E-5**	**0.8571**

Table 23 summarizes our initial experiments of fine-tuning a hyperspectral MAE foundation model to reconstruct hyperspectral EnMAP data from multi-spectral Sentinel-2 imagery.

Table 23. Quality scores of reconstructing EnMAP from Sentinel-2

Dataset	Frozen/FT	Masking	Total MSE
S2-EnMAP	Frozen	Spec	9.96 E-3
S2-EnMAP	Finetuned	Spec	**3.89 E-3**

2.3.6 MMDC, pVAE, ALISE, MALICE, MALICE Aux

The results of MMDC method can be presented as two parts:

1. Qualitative and Quantitative evaluation of the obtained latent spaces
2. Quantitative evaluation of downstream task results.

2.3.6.1 Latent spaces

Whilst the quality of MMDC trained model can be only evaluated with downstream tasks, the visual analysis of the produces latent representations and overall loss trend may indicate if the model can be exploitable:

- more the mean values of S1 and S2 latent spaces look alike, the higher the chance that the downstream tasks will be performed with better accuracy as the learnt embeddings assure the interchangeability of two sources of data.
- the reconstructed S1 ascending/descending and S2 images should resemble as much as possible to the original inputs.

Figure 57 and Figure 58 present the visualization of the results of MMDC model. One can state good direct reconstruction of the input data, slightly worse cross-reconstruction. Two latent spaces have the overall resemblance, however, S1 embedding contains more “granular” textures proper to radar images. Finally, the logvariance is often high for saturated pixels corresponding to fields or buildings.

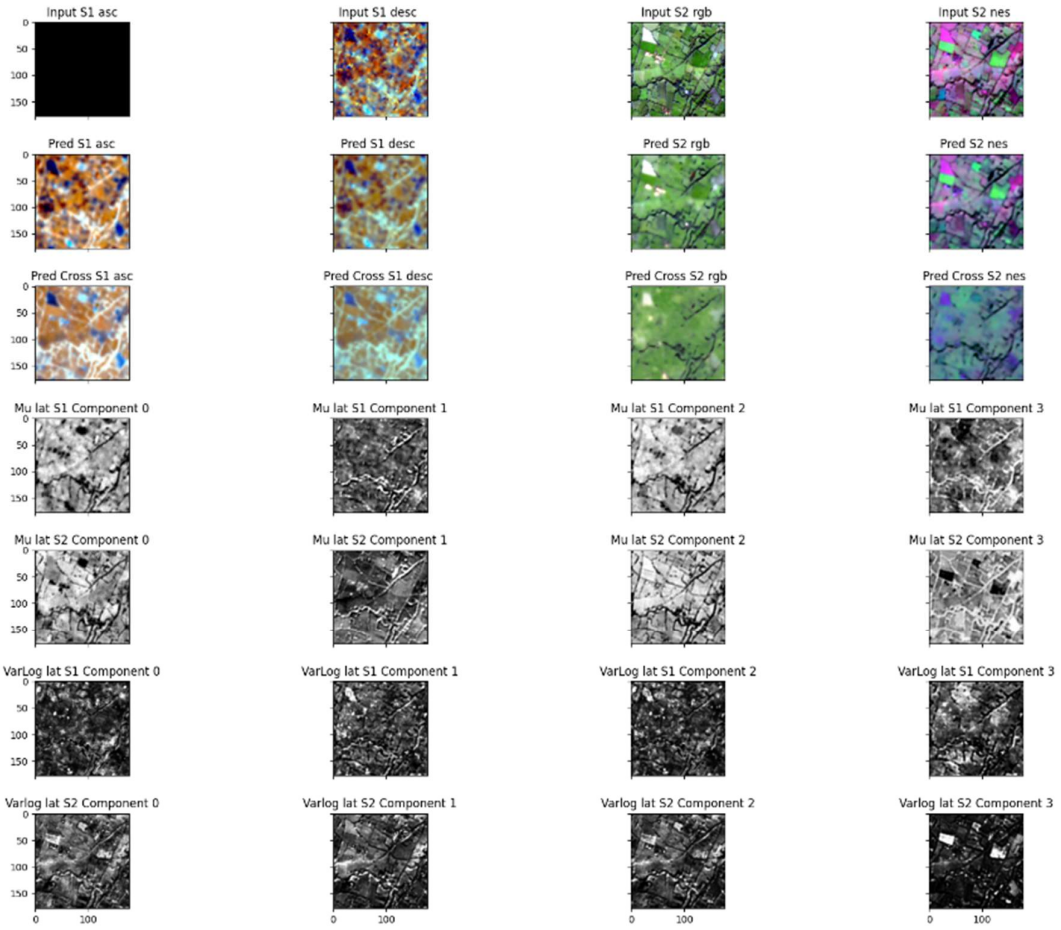


Figure 57. Latent spaces obtained with MMDC baseline. First row: the original input data. Second row: data reconstructed from its proper latent space (S1 and S2 images reconstructed from S1 and S2 embeddings respectively). Third row: cross-reconstructed data (S1 and S2 images reconstructed from S2 and S1 embeddings respectively). Fourth and Sixth rows – mean and logvar of S1 embeddings. Fifth and Seventh rows – mean and logvar of S2 embeddings.

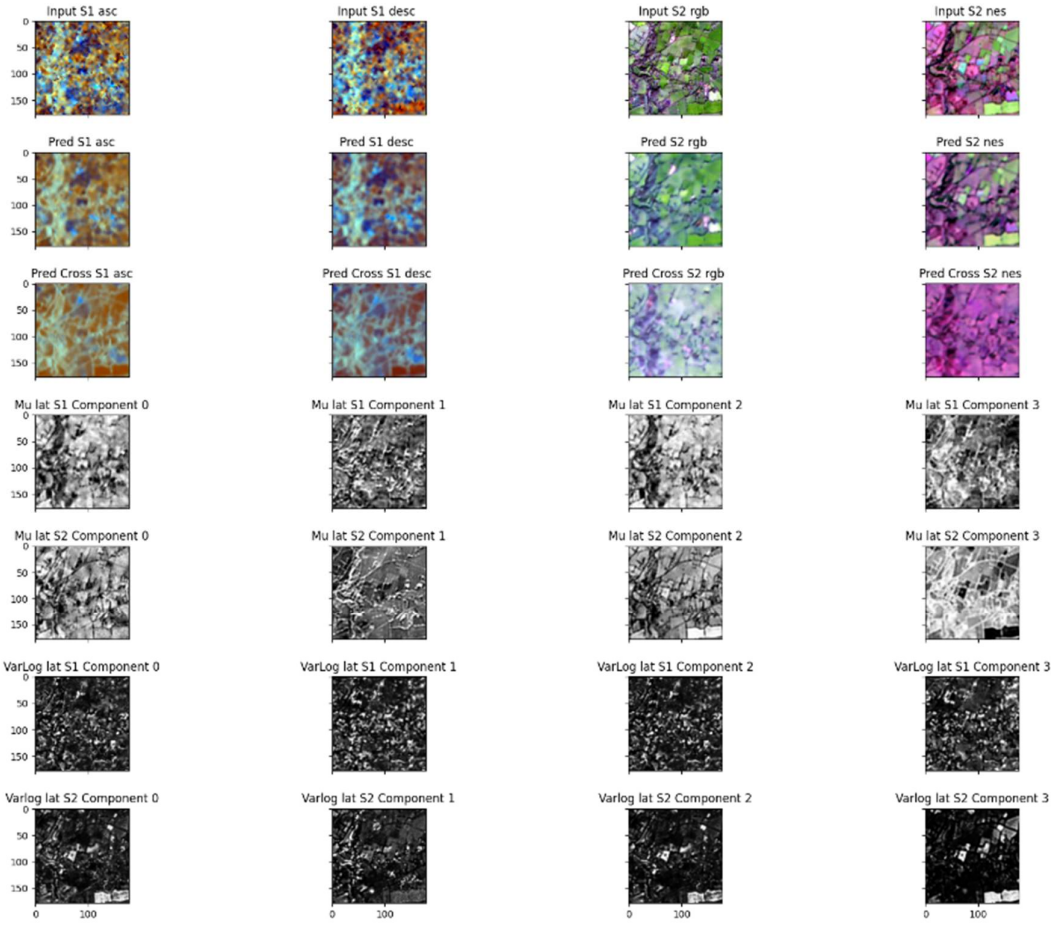


Figure 58. Latent spaces obtained with MMDC baseline. First row: the original input data. Second row: data reconstructed from its proper latent space (S1 and S2 images reconstructed from S1 and S2 embeddings respectively). Third row: cross-reconstructed data (S1 and S2 images reconstructed from S2 and S1 embeddings respectively). Fourth and Sixth rows – mean and logvar of S1 embeddings. Fifth and Seventh rows – mean and logvar of S2 embeddings.

2.3.6.2 Downstream tasks

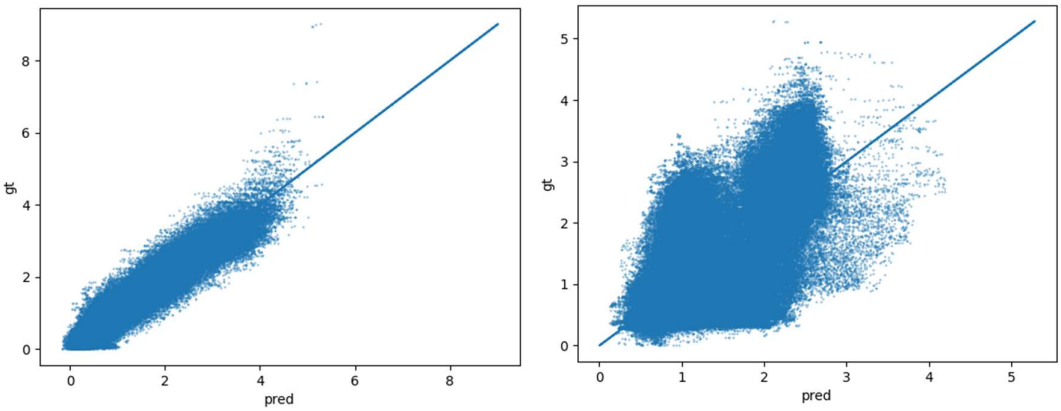
2.3.6.2.1 LAI regression

Table 24. Regression performance with raw image input and MMDC embeddings.

Model	Raw	MMDC
	rMSE/R2	
S2	0.12/0.95	0.23/0.82
S1	0.62/0	0.43/0.39

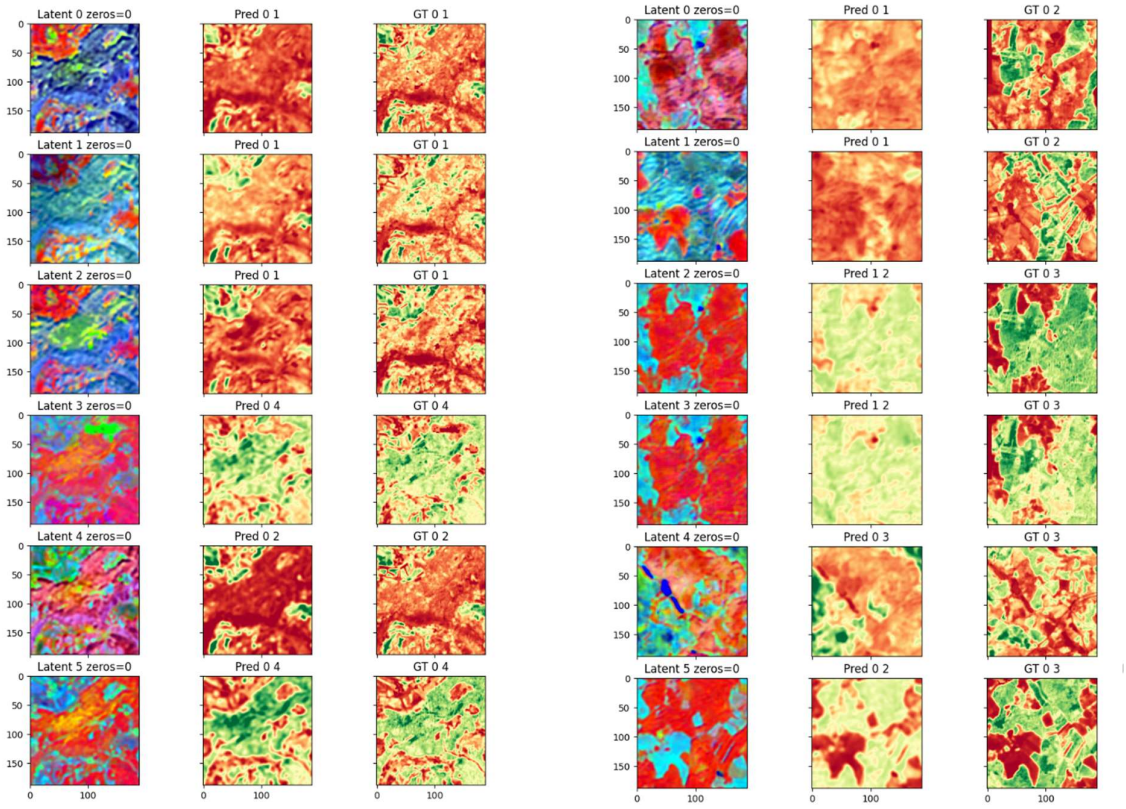
We can state from the Table 24 that the DST achieves the best performance with raw Sentinel-2 data, and the worst one with Sentinel-1 raw input giving almost zero

correlation between predicted and GT values. Sentinel-2 embeddings give slightly worse performance than the raw data. Sentinel-1 embeddings significantly outperform the raw data; however, we still consider their performance poor and insufficient for real-life applications. In effect, as it can be seen on Figure 59 and Figure 60, the model struggle to predict high LAI values with Sentinel-1 embeddings. We equally stated that weighted loss did not bring any major improvements.



(a) S2 MMDC embeddings input. *(b) S1 MMDC embeddings input.*

Figure 59. Scatterplot of predicted LAI values vs ground truth values.



(a) S2 MMDC embeddings input. *(b) S1 MMDC embeddings input.*

Figure 60. LAI regression predictions with MMDC embeddings inputs. Note that in "Pred 0 2" 0 and 2 stand for minimum and maximum normalization values for the visualization of the results.

2.3.6.2.2 SITS classification

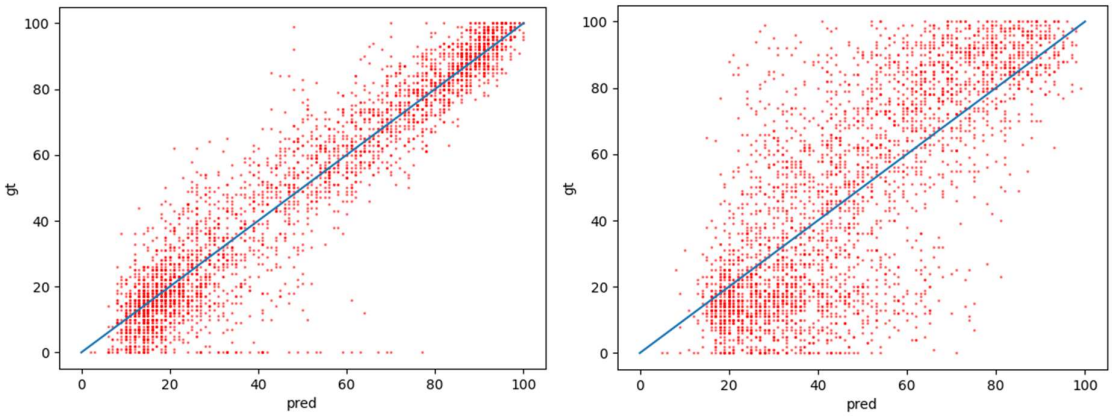
Table 25. Performance of different embedding inputs on crop classification task. For S1, A/D denote ascending and descending orbits. Experiments were conducted using 5 folds; however, if an experiment was performed only with 1-fold, the results are labelled with "(1f)".

Model	Raw	MMDC	pVAE	MALICE	MALICE Aux
	mIoU,% / Accuracy,%				
S2	63.1/83.2	56.3/83.9 (1f)	65.7/89.1	-	61.6/86.8 (1f)
S1	A54.9/? D54.7/?	A53.7/83.0	-	A 57.9/85.1	A 59.0/85.2
S1+S2	66.3/84	63.5/86.9 (1f)	-	-	-

As can be observed from the Table 25, the best performance for Sentinel-2 is achieved with pVAE embeddings, as it extracts the bio-physical variables, which helps to distinguish some crop classes. As for LAI regression task, MMDC S2 embeddings perform worse that the raw Sentinel-2 input. Other Sentinel-2 embeddings achieved the performance similar to the raw input data.

In turn, Sentinel-1 embeddings extracted with multi-temporal embedding algorithms produce better results that Sentinel-1 alone, or than a single-date MMDC S1. Moreover, MALICE Aux achieve the noticeable improvement compared to the raw Sentinel-1 data.

2.3.6.2.3 TCD regression



(a) S2 MALICE Aux embeddings input. (b) S1 ASC MALICE Aux embeddings input

Figure 61. Scatterplot of predicted TCD values vs ground truth values with MALICE Aux embeddings. We do not present scatterplots for other embeddings, as they do not visually show any notable difference.

Table 26. TCD regression performance with different embedding inputs.

Model	GAF raw	Our raw	MMDC	pVAE	Alise	Malice	Malice Aux
	rMSE, %						
S2	18.0	10.7	11.8	10.9	10.7	10.3	10.6
S1	-	22.9	18.3	-	-	-	-
S1 ASC	-	25.7	20.3	-	-	21.8	20.6

As it can be seen from the Table 26 and Figure 61, all Sentinel-2 embeddings perform relatively similarly, except for MMDC with a slightly worse performance.

For Sentinel-1 embeddings, we notice that MMDC embeddings with both orbits perform better, than other embeddings with only ascending orbit. This behaviour may be related to the poor quality of S1 ASC data within our study area (missing data and artefacts); however, we did not conduct any further evaluations.

2.3.6.2.4 BioMass and Tree Height Regression

We compare the obtained results to JOANNEUM research (JR) results. For the comparison, we only dispose of the results of their single year method, where they only use data from 2021 year, from 10 June to 15 September. We denote :

"JR, rMSE" - their best method based on rMSE value, uses the following input data: S2 optical filtered data + DEM + S2 extracted textures;
"JR, R2" - their best method based on R2 value, uses the following input data: S2 optical filtered data + DEM + S2 extracted textures + calibrated SAR values.

Table 27. Tree Height Regression performance with MALICE S1 ASC embeddings.

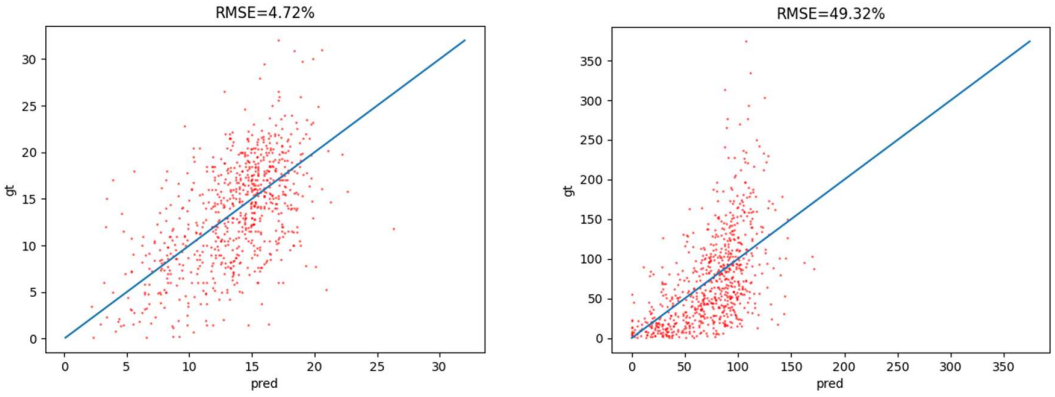
	RMSE, m	MAE, m	R2
Our S1 ASC	4.7	3.7	0.31
JR,~rMSE	3.8	3.1	0.65
JR,~R2	3.9	3.0	0.65

Table 28. BioMass Regression performance with MALICE S1 ASC embeddings.

	RMSE, t/ha	MAE, t/ha	R2
Our S1 ASC	49.1	36.2	0.32
JR,~rMSE	45.1	31.2	0.44
JR,~R2	57.5	30.9	0.48

Table 27,

Table 28 and Figure 62 present the performance of S1 ASC MALICE embeddings. Our model struggle to predict high values, despite weight application. Therefore, we observe significantly lower R2 values for our methods. We equally observe good rMSE value for AGB prediction, compared to concurrent methods, and worse FCH regression performance of our S1 embeddings.



a) Forest canopy height *b) Above ground biomass*

Figure 62. Scatterplot of predicted values vs ground truth values with MALICE S1 embeddings.

2.4 Summary and conclusions

We introduced, developed, and benchmarked a set of novel weakly-supervised deep learning methods for Sentinel-2, Sentinel-1 & -2 data fusion, and a first-of-its-

kind large-scale hyperspectral foundation model. These methodologies we ship with corresponding datasets curated: SSL4EO-S12 as spatio-temporally aligned, globally distributed Sentinel-1/2 patches with an extension SSL4EO-S12-EU-Forest with sampling distribution aligned to European forests.

We explored the SSL4EO-S12-EU-Forest dataset to evaluate the specific downstream task of semantic forest segmentation evaluating advantages and disadvantages of the SSL4EO-S12 models – including a tailored Custom-ResNet-18 with improved accuracy.

To advance the state-of-the-art in deep learning methodologies for Sentinel-1, Sentinel-2, and EnMAP data, we developed the following methodologies:

SoftCon leverages existing landcover annotations to improve contrastive self-supervised pretraining of foundation models for Sentinel-2 imagery.
DeCUR introduces a novel method for Sentinel-1/2 data fusion towards explainable multi-modal SSL based on engineered, modality-specific feature extraction (explainable embedding vector).
CromSS explores weak supervision models utilizing noisy semantic segmentation maps by multi-modal cross-validation of confidence given Sentinel-1/2 imagery of a scene.*SpectralEarth* provides a hyperspectral foundation model and large-scale dataset based on DLR's EnMAP satellite. Further, we propose an energy-efficient fine-tuning strategy. We also started exploring the fusion of hyperspectral and multi-spectral data utilizing the EnMAP and Sentinel-2 data archives. The MMDC approach is a different take on the problem, since the embeddings are pixel-wise and therefore keep the full spatial resolution of the input data. Their usefulness needs to be assessed using the 3 downstream tasks proposed. The tree cover density (TCD) is related to the evolution of the CLMS HRL portfolio in the forest domain (i.e., HRL VLCC). In phase 2 of the project, the MMDC approach will be further analysed as candidate method for the EvoLand C1 forest prototype.

Equipped with such methods to fuse SAR, multi- and hyperspectral satellite imagery of single and multiple timestamps, we pave the way to intelligent data analysis of the Copernicus Programme, including future missions such as CHIME for global coverage of Earth observation by hyperspectral data.

3 Improved resolution

3.1 Single Image Super-Resolution of Sentinel-2 images

3.1.1 State of the art

3.1.1.1 Introduction

Single Image Super-Resolution (SISR), is the process of reconstructing a higher resolution image from a single low-resolution observation. Task 2.5.1 of EvoLand aims at designing and training a Convolutional Neural Network (CNN) for the Single Image Super-Resolution of 10 Sentinel-2 bands (B2, B3, B4, B5, B6, B7, B8, B8A, B11, B12) to 5-meter spatial resolution. This improvement in spatial resolution will support prototypes C5 Gross Primary Production, C6 Small Landscape Features, C7 Water bodies mapping, C8 Imperviousness monitoring, C9 Urban dynamics mapping, C10 Global land surface characteristics LSC and C11 On-demand LC mapping in achieving the target performances.

In the general computer vision field, resolution often refers to the image size. Processed images are acquired under a variety of (most of the time unknown) viewing angles, scene depth, optics, motion blur and noise. Improving the resolution therefore means representing the same image content with more pixels and improved level of details. This improved level of details intuitively relates to the how much small objects can be resolved and separated in the image, but in the absence of homogeneous and known observation conditions, this notion may be difficult to quantify. This is however the situation of most of training dataset fostered from the internet.

In remote sensing optical imaging however, observation conditions are particularly stable and known. Optics and detectors properties are carefully engineered and monitored during the lifetime of the instrument, motion is known with great accuracy and depth can be considered constant, because the flight altitude (several hundreds of kilometres) is very large with respect to the focal length of the instrument (a few meters to a few tenths of meters). Variation of blur caused by variation of depth can therefore be neglected. The instrument can therefore be modelled by a Low-Pass spatial filter representing the acquisition conditions (motion, optics, diffraction, detector shape ...) followed by a 2-dimensional discrete sampling process. In this model, a single value, called Modulation Transfer Function (MTF) can be used to characterize the acquired level of details: MTF is the attenuation of this spatial Low-Pass filter at half of the spatial sampling frequency yielded by the detector size. Lower MTF values mean well sampled images according to the Nyquist-Shanon theorem, but blurry, while higher MTF values mean crisper images with aliasing artifacts caused by under-sampling of the signal. Of course, MTF varies with wavelength, and a single optic is usually used for

all spectral bands, which means that for a given instrument MTF varies across spectral bands.

In ground segments, dedicated processing is already implemented in order to restore damped spatial frequencies, in a process called image restoration. This involves accurately measuring the Point Spread Function (PSF) of the instrument (spatial shape of the Low Pass filter modelling the system) and inverting it in a trade-off between amount of sharpening and amount of noise and aliasing artifacts.

Therefore, depending on the target of super-resolution factor, deep-learning based SISR can be seen as a super-restoration process (Levin et al., 2011), that tries to leverage the non-linear modelling ability of Neural Network in order to estimate the best possible restoration in a data driven approach. A data driven approach is beneficial because in user-level products (from level 1C onward), images have already been resampled and processed by many algorithms, which modifies (often in an unknown way) the precise knowledge of image quality obtained closer to the sensor.

In this work, we are targeting a super-resolution factor of 2 for the 10-meters band, and we are therefore in the image restoration domain, in contrast with other work targeting higher scale factor (Figure 63). On the other hand, we aim at exploiting the strong correlation between spectral bands (Inglada et al., 2022).

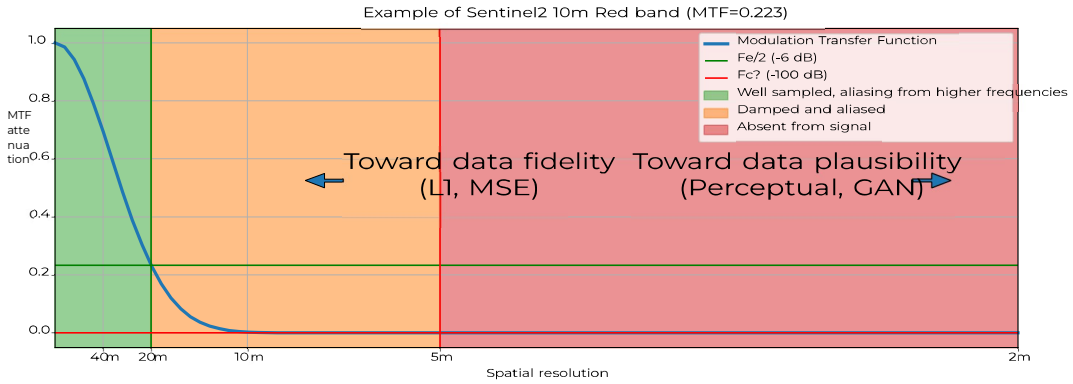


Figure 63. Image Restoration versus Image Super-Resolution.

3.1.1.2 Review of existing methods

3.1.1.2.1 Taxonomy of methods

SISR in the general computer vision domain is largely addressed in the literature of the last 30 years. In this section, we will focus on the major trends that lead to performances breakthroughs. The reader can reference to the following recent review papers for a more exhaustive view of proposed methods (Anwar et al., 2020; H. Chen et al., 2022; H. Li et al., 2021; H. Liu et al., 2021).

Seminal work: Super-Resolution Convolutional Neural Network (SRCNN)

SRCNN (Dong et al., 2015) is recognized as the first successful CNN-based SISR Network (Figure 64). Though advertised to be a Deep Learning architecture, it is in fact composed of only 3 layers, the middle one being a linear feedforward layer, while the input and output layers are standard convolutional layers. The input Low Resolution (LR) image is upsampled beforehand to the target High Resolution (HR) by means of bicubic interpolation. The network is trained using the MSE (L_2) loss, which is known to favour high Peak Signal to Noise Ratio (PSNR). The authors use images from ImageNet (Deng et al., 2009), and achieves +3dB of PSNR for an upsampling factor of 2 with respect to the baseline bicubic upsampling. (+2dB for an up-sampling factor of 2). The authors also make two useful observations for our work: first, adding more layers to the network does not necessarily improve the performances, and second, using RGB colour inputs yields better results than processing those channels independently with a single channel network. For the latter observation, the authors hypothesize that the high correlation between the RGB channels actually help the network.

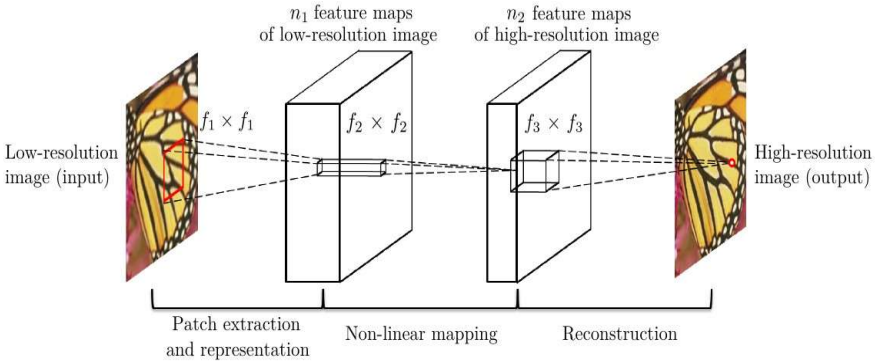


Figure 64. SRCNN network architecture (figure from [Dong et al., 2015]).

Evolution of CNN-based architectures

After SRCNN, many works aim at increasing the depth of the CNN, such as the Very Deep Super-Resolution network (VDSR) (J. Kim et al., 2016), which uses up to 20 convolution layers. Deeper networks are more difficult to train because of the well-known vanishing and exploding gradient issues, and the author achieve convergence by using gradient clipping and more importantly by learning to predict HR residuals between the HR target image and LR bicubic-upsampled image instead of directly predicting the HR target image. VDSR achieves a boost of around +1dB with respect to SRCNN.

In (Lim et al., 2017), the authors proposed the Enhanced Deep Super-Resolution network, based on a residual (ResNet) architecture. This introduces skip connections that are a natural extension of residuals prediction: they allow better gradient flows during training and further solve the difficulties of training deep networks. The authors also abandon the early up-sampling scheme in favour of the late up-sampling one: the residual blocks are applied at LR, which is more time and

memory efficient because less pixels must be processed by those layers. The final HR prediction is obtained by unfolding the feature domain into the spatial domain in the final layers, an operation known as PixelShuffle and presented Figure 65. The PixelShuffle is equivalent to a subpixel convolution scheme, while being more computationally effective (Shi et al., 2016). Finally, the authors train the network using a L_1 loss function instead of the MSE. The author reports a performance boost of around +1dB with respect to VDSR.

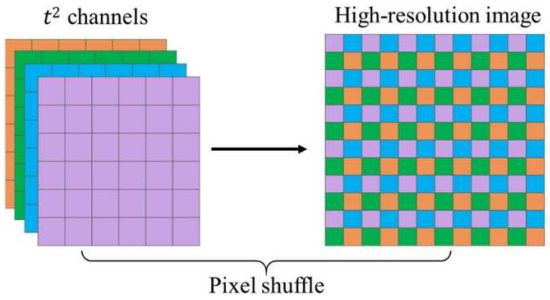


Figure 65. Understanding the PixelShuffle operation (figure from [Qin et al., 2020]).

At this point, the number of parameters yielded by the deeper architecture started to be a concern and in (He et al., 2016b), the authors propose Deeply-Recursive Convolutional Neural Network (DRCNN), to achieve desired depth while reducing the number of parameters. In DRCNN, some blocks are recurred several times, with skip connections at each recurring step. This idea is further combined with residual blocks in (Tai et al., 2017).

In Cascading Residual Network (CARN) [Ahn et al., 2018], the authors elaborate on skip connections and recurrent network by multiplying skip connections between residual blocks. As shown in Figure 66, they define two levels of block: the base residual block uses grouped convolution where the group size allows to trade accuracy for efficiency, while the outer cascading blocks combine several, possibly shared, residual blocks. The global workflow of CARN is shown in Figure 67. The author seeks for an optimal trade-off between complexity and accuracy, which is confirmed in the benchmark proposed in (Anwar et al., 2020), where CARN is the method that achieves the best trade-off between number of additions and multiplications versus achieved PSNR.

It is noteworthy that most of those models have multiple output resolution variants, where the same backbone architecture is connected to scale-specific heads, allowing to jointly train for several up-scaling factors. Authors report that such multi-resolution training scheme allow to boost individual scale performances. In our work however, we are looking for a fixed upsampling rate (e.g., the target up-scaling factor is fixed and known), and therefore multi-resolution training is of little interest.

In summary, the major improvements to CNN-based SISR networks are:

- Using the late upsampling scheme instead of the early upsampling one,

- Using residual blocks with multiple skip connections between blocks
- Using deeper network and leverage shared weights or recurrent layers to keep number of parameters under control.

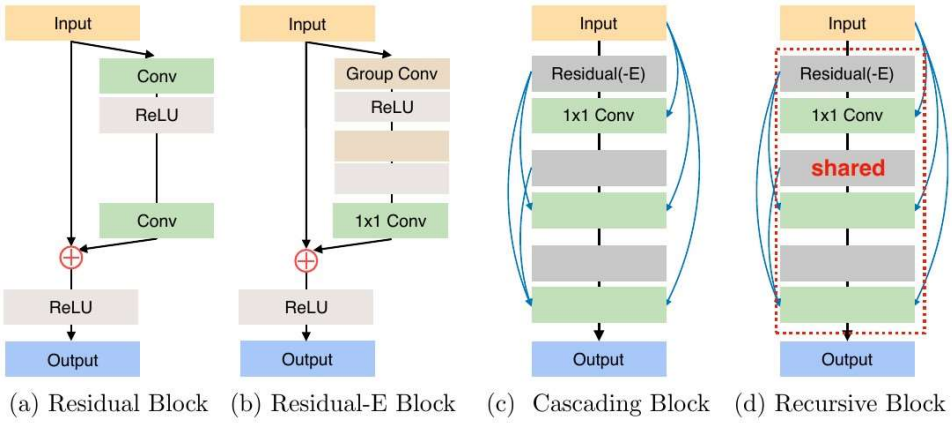


Figure 66. Residual and cascading blocks of CARN architecture (figure from [Ahn et al., 2018]).

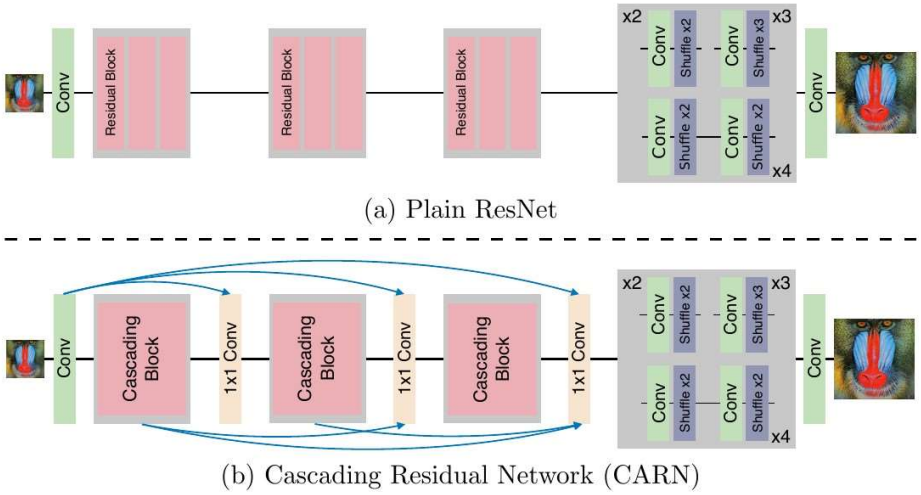


Figure 67. Overview of CARN architecture (figure from [Ahn et al., 2018]).

While a performance plateau has been reached by CNN-based methods for relatively small (<4) up-sampling factor, those models fail to achieve good performances for larger up-scaling factor. Indeed, when up-sampling factor increases, moving to the right of the graph presented in Figure 63 the problem becomes more ill-posed, as many HR predictions can correspond to the LR input. In this situation, PSNR-oriented loss functions, such as L_1 and L_2 norm, are too local to push the network toward realistic details.

One solution to this problem is the use of a perceptual loss (Johnson et al., 2016), which is essentially a L_2 or L_1 loss applied to the output of intermediate layers from a pretrained model such as VGG16. Intermediate layers output corresponds to

latent features that describe all the pixels in the perceptive field of the network and thus are less local than pixel-wise methods.

Another solution is to use adversarial training, which became popular with the pix2pix image translation network (Isola et al., 2017). The idea behind adversarial training is to jointly train the network performing some tasks (called the Generator) and a second network, called the Discriminator, which tries to discriminate between examples yielded by the Generator and examples sampled from the ground truth. When the Discriminator can no longer achieve proper discrimination, the distribution of the output of the Generator and the distribution of the ground truth data closely matches. Thus, during training, gradient flowing from the Discriminator to the Generator weights pushes the latter toward the distribution of the ground truth data.

The first Super-Resolution oriented work is SR-GAN (Ledig et al., 2017). In SR-GAN, the generator follows a classical residual block, late up-sampling network, with the notable addition of Batch Normalization, which had been discarded in [Lim et al., 2017], while the discriminator is a standard CNN, with a linear feedforward layer followed by a sigmoid. Regarding loss function, a content loss (either MSE or Perceptual Loss based on VGG) is combined with the adversarial loss. It should be noted that the network is pretrained with a single MSE loss, and that adversarial training can therefore be considered as a fine-tuning.

Enhanced SR-GAN (ESRGAN) (X. Wang, Yu, Wu, Gu, Liu, Dong, & Qiao Yu and Change Loy, 2018) elaborates on the SR-GAN work by making three key modifications:

- The basic residual block is replaced by the Residual-in-Residual Dense Block (RDDB).
- Formulation of adversarial training is made relativistic: this amount to estimate if a prediction is comparatively more plausible than the associated ground truth data.
- Improving the VGG based perceptual loss.

Transformers, diffusion models

Following trends in Natural Language Processing, some literature elaborates on the Vision Transformer (ViT) (Z. Liu et al., 2021; Ranftl et al., 2021) to develop new Super-Resolution architectures (J. Liang et al., 2021). The probabilistic generative approach of diffusion models has also been applied to SISR (H. Li et al., 2021; Rombach et al., 2022; Saharia et al., 2023).

3.1.1.2.2 SISR in remote sensing

In the past 10 years, SISR progress in the general computer vision domain has raised an interest in the remote sensing community. SISR has been applied to Very High Resolution (VHR) optical images (Chouteau et al., 2022; Tao & Muller, 2021) as a restoration process allowing a scaling factor of 2, trained with high precision

simulation of data from aerial imagery. In (D. Pouliot et al., 2018), the author uses Sentinel-2 as a target to train SRCNN in super-resolving Landsat images. Though limited in term of spatial coverage of the training set, this study is similar to the work presented in this document.

Sentinel-2 makes no exception, and a number of works are dedicated to training SISR models to improve its spatial resolution beyond 10 meters. In (N. L. Nguyen et al., 2023a), the author demonstrates that due to its instrument properties, with high aliasing and spatial shift between bands in the focal plane, Sentinel-2 is especially well suited for SISR.

In (Galar, Sesma, & Ayala Christian and Aranda, 2019; Galar et al., 2020), the authors trained EDSR on 12 pairs of Sentinel2 and Planetscope images over California, U.S., in order to super-resolve the 4 10-meter bands of Sentinel-2 to either 5-meters or 2.5 meters.

In (Romero et al., 2020) leverage ESRGAN and Worldview acquisitions to superresolve the 10-meter resolution bands from Sentinel-2 to 2 meters. It should be noted that only three pairs of Sentinel-2 and Worldview images have been used for training, and that the dataset used for this paper is not publicly available.

In (Pineda et al., 2020), 41 pairs of Sentinel-2 and PeruSat-1 images in the northern coast of Peru have been used to train SR-GAN in order to improve the spatial resolution of the Red, Green and Blue bands with a scaling factor of 4.

In (Kapilaratne et al., 2022), ESRGAN is trained to super-resolve Sentinel-2 against Worldview3 images. If the number of image pairs is significant (several hundreds), they only cover Japan. Moreover, only Red, Green and Blue are processed, and they are converted to unsigned 8-bits radiometric quantization, which is a massive loss of valuable information for downstream applications.

3.1.1.3 Strengths and weakness of existing methods

Existing works on SISR for general computer vision have come up with a large panel of efficient architecture and training strategies that the optical remote sensing community should benefit from.

However, current applications of such methods to Sentinel-2 aim for a super-resolution scale factor, which imposes the following limitations:

They require VHR commercial imagery as reference data, which is by its very nature scarce and lead to closed-source, limited dataset for training (even if some initiatives such as the WorldStrat dataset (Cornebise et al., 2022) may change the game in the future).

In turn, the VHR images used in the training set suffer from insufficient similarities with the Sentinel-2 sensor, such as spectral bands limited to Red, Green, Blue and Near Infra-Red channels, absence of constant viewing angles yielding radiometric and geometric discrepancies, and lower radiometric quality.

They require very deep networks and perceptual and adversarial training so as to address the large-scale factor, which make the problem more ill-posed and pushes generated images toward plausibility instead of fidelity.

In the work described in this document, we intend to:

- Aim for super-restoration with reasonable scale factors in the super-restoration domain: a factor of x2 for the 10 meters bands and x4 for the 20 meters band,
- Seek for radiometric consistency with LR observations as the most important criterion,
- Train on a large dataset with representative data, using a publicly available dataset,
- Process the 10 useful bands of Sentinel-2 and not only RGB or RGB-NIR, and
- Make a cost-effective model that can be deployed at scale for EvoLand prototypes and beyond.

3.1.2 Method description

3.1.2.1 Input data and study sites

3.1.2.1.1 Training dataset

In this work, we will make use of the publicly available Sen2Venμs dataset (Michel et al., 2022). This dataset has been fostered by identifying cloud-free same-day acquisitions of Sentinel-2 and Venμs (Dick A. & Moreau, 2022) images. Venμs is not a global coverage mission, as it only covers a hundred of sites worldwide, but it has the same spectral bands as Sentinel-2 (to the notable exception of B11 and B12), all sampled at 5 meter resolution, as shown in Figure 68 and Table 29. Moreover, Venμs acquire images under constant (though maybe large) viewing angles for a given site, and undergoes the L2A same processing as the one applied to Sentinel-2 in CNES ground-segment (Lonjou et al., 2016), distributed by Theia[11]mm.

Table 29. Corresponding Sentinel-2 and VENμS bands used in the SEN2VENμS dataset. Note that B11 is used as high resolution band for both B8 and B8A Sentinel-2 bands.

Sentinel-2	10m bands	20m bands
	B2 B3 B4 B8	B5 B6 B7 B8A
VENμS	5m bands	5m bands
	B3 B4 B7 B11	B8 B9 B10 B11

The dataset covers 29 locations shown in Figure 69 on earth with a total of 132,955 patches of 256 × 256 pixels at 5 m resolution covering Sentinel-2 spectral bands B2, B3, B4, B5, B6, B7, B8, B8A, as well as 128x128 pixels at 10 meter resolution patches covering B2, B3, B4, B8 and 64x64 20 meter resolution patches covering B5, B6, B7, B8A. Patches have been processed for linear radiometric harmonization and spatial

[11] www.theia.cnes.fr

registration, using Sentinel-2 as the reference. The dataset has an open license and can be freely downloaded from Zenodo (Michel et al., 2022). Figure 70 display one sample patch per location with colour composition covering all spectral bands.

In EvoLand, we aim at super-resolving all useful bands of Sentinel-2, including the SWIR bands that are missing from the Sen2Venμs dataset. During the project, we therefore sampled the B11 and B12 bands for all patches of Sen2Venμs, generating an extended dataset called sen2venμs_extended.

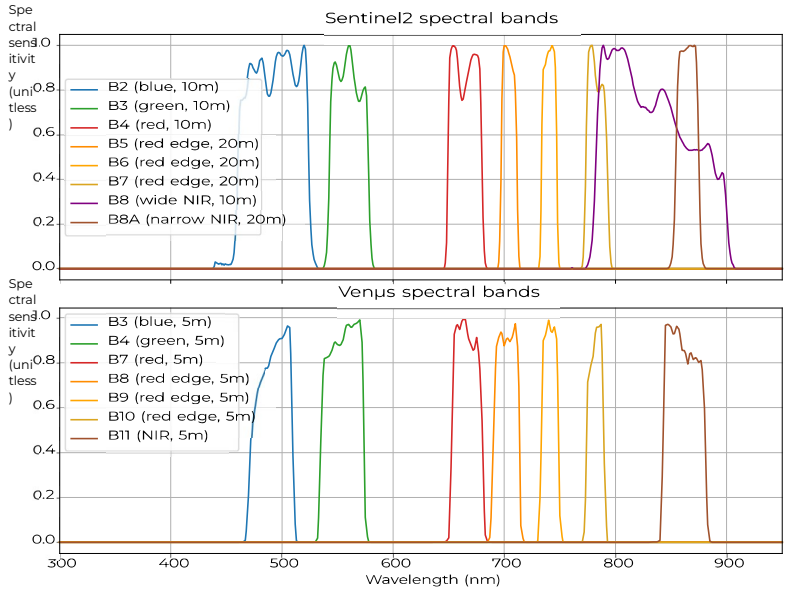


Figure 68. Spectral Sensitivity Response of Sentinel-2 and Venμs sensors.

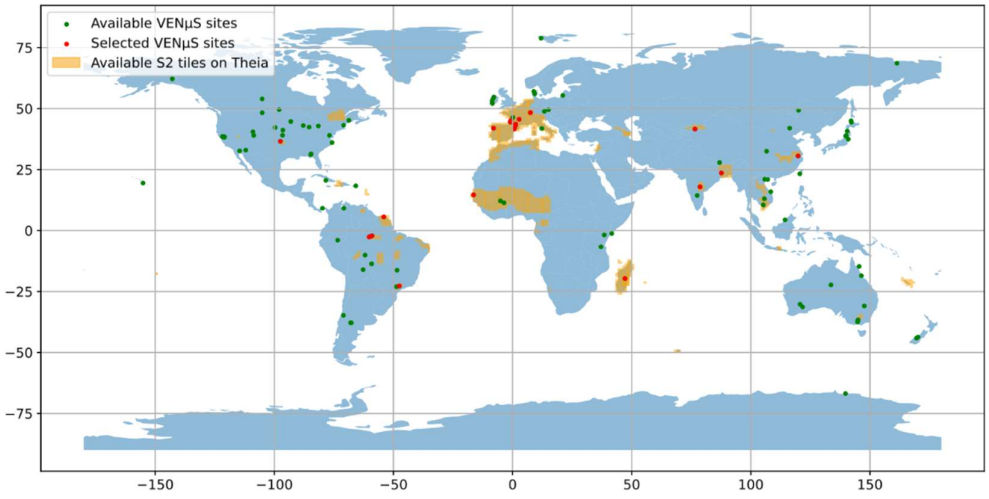


Figure 69. Location of the 29 sites composing the Sen2Venμs dataset.

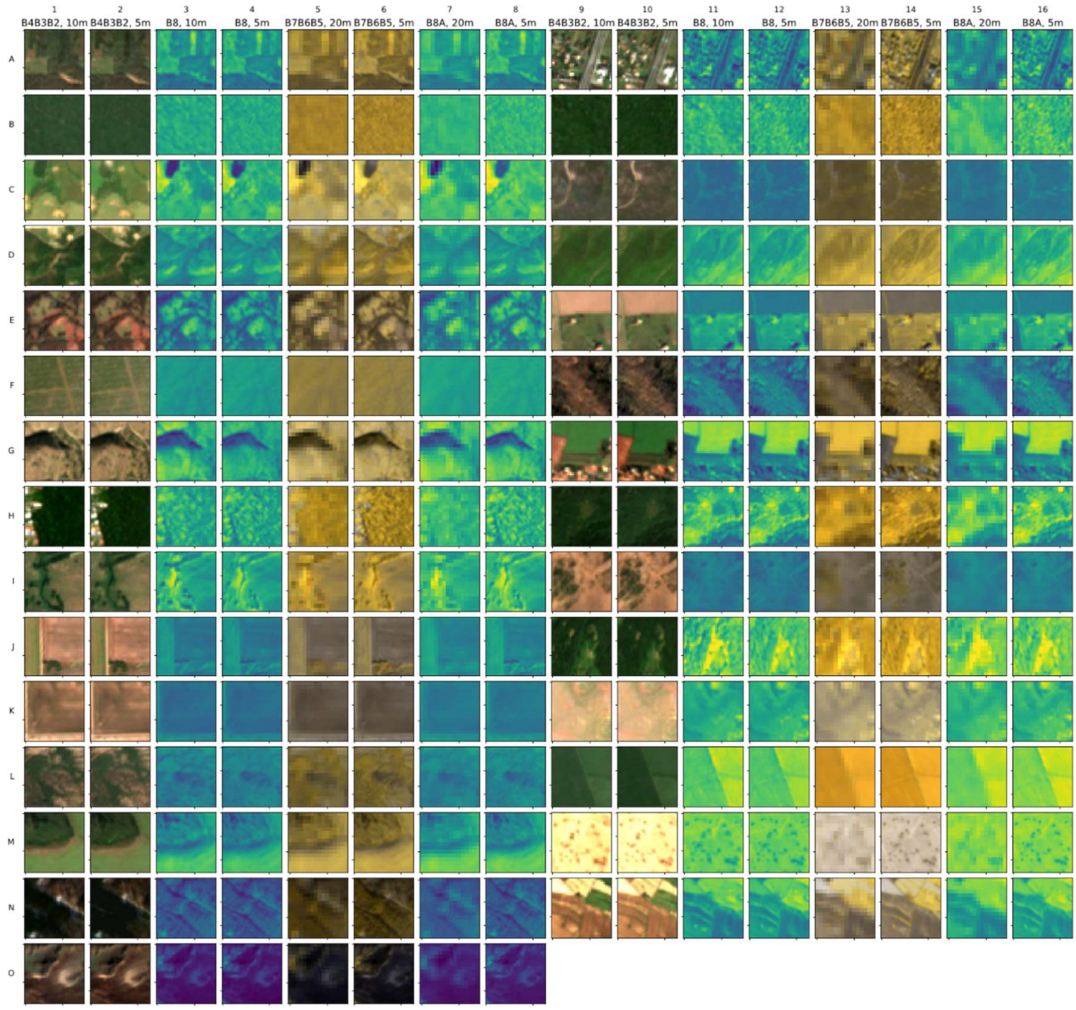


Figure 70. Examples of patches. From left to right: columns 1-8 and 9-16 show rendering of two different patches. Columns 1 and 9: B4, B3, B7 (RGB natural) at 10m, columns 2 and 10: B4, B3, B7 (RGB natural) at 5m, columns 3 and 11: colour-mapped B8 at 10m (Wide Near Infra-Red), columns 4 and 12 : colour-mapped B8 at 5m (Wide Near Infra-Red), columns 5 and 13 : B7, B6, B5 colour composition (Red Edge 3 to 1) at 20m, columns 6 and 14: B7, B6, B5 colour composition (Red Edge 3 to 1) at 5m, columns 7 and 15: colour-mapped B8A at 20m (Narrow Near Infra-Red), columns 8 and 16: colour-mapped B8A at 5m (Narrow Near Infra-Red). 29 patches are displayed, one random patch for each site. Only 64 × 64 pixels crops of the patches are displayed to improve readability. High resolution and low resolution patches radiometry was scaled to 8 bits with the same scaling factors.

3.1.2.2 Test dataset

To evaluate the performances of the trained SISR models, some sites of the Sen2Venμs_extended dataset have been kept apart for testing. The selection of testing and training/validation sites is given in Table 30.

Table 30. Country and province of each of the selected VENμS sites. The usage column indicates if the site is used for model training and validation, or for testing (model performances evaluation).

Site	Number of patches	Number of pairs	VENμS Zenith Angle	Usage
FR-LQ1	4.888	18	1.795402	Testing
NARYN	3.814	25	5.010906	
FGMANAUS	129	4	7.232127	
MAD-AMBO	1.443	19	14.788115	Testing
ARM	15.859	39	15.160683	
BAMBENW2	9.018	34	17.766533	
ES-IC3XG	8.823	35	18.807686	Testing
ANJI	2.314	16	19.310494	Testing
ATTO	2.258	9	22.048651	Testing
ESGISB-3	6.057	19	23.683871	
ESGISB-1	2.892	13	24.561609	
FR-BIL	7.105	30	24.802892	
K34-AMAZ	1.385	21	24.982675	
ESGISB-2	3.067	13	26.209776	
ALSACE	2.654	17	26.877071	
LERIDA-1	2.281	6	28.524780	
ESTUAMAR	912	13	28.871947	
SUDOUE-5	2.176	20	29.170244	
KUDALIAR	7.269	20	29.180855	
SUDOUE-6	2.435	14	29.192055	
SUDOUE-4	935	7	29.516127	
SUDOUE-3	5.363	14	29.998115	
SO1	12.018	36	30.255978	
SUDOUE-2	9.700	27	31.295256	
ES-LTERA	1.701	19	31.971764	Testing
FR-LAM	7.299	22	32.054056	
SO2	738	22	32.218481	Testing
BENGA	5.858	29	32.587334	
JAM2018	2.564	18	33.718953	

3.1.2.3 Description of the candidate method

This section describes in detail the model and the training strategy used in EvoLand.

3.1.2.3.1 Model

Backbone SISR model

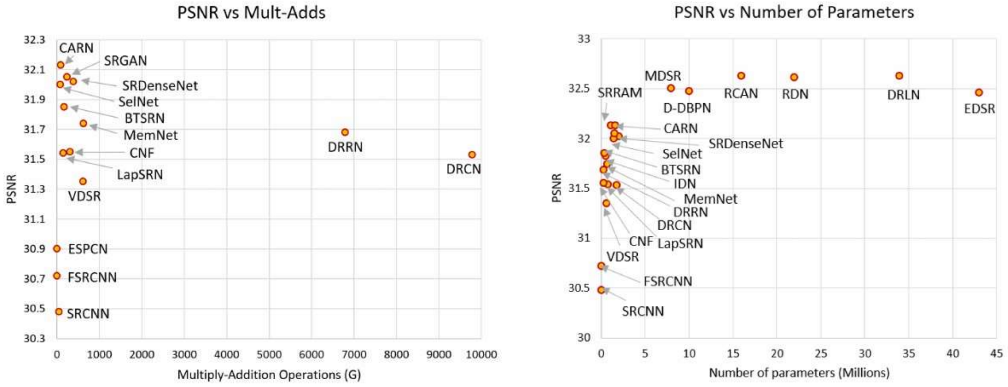


Figure 71. Achieved PSNR with respect to model complexity (figures from [Anwar et al., 2020]) (a) PSNR vs. number of mult-add operations (b) PSNR vs. number of parameters.

In section 2, the key factor of SISR network architecture performances have been identified: late up-sampling scheme combined with a chain of residual blocks, which can be recurred or have their weight shared to reduce the size of the model. Many architectures have been proposed with this design. Figure 71 shows that CARN (Ahn et al., 2018) model stands out as being the most computational cost-effective model among the simpler model (with fewer parameters). CARN will therefore be our backbone model of choice in this work. Overview of the CARN network is presented in Figure 65 and Figure 66. Since we are only aiming to a single up-scaling factor however, our CARN network will only have a single head.

The reader can refer to the CARN paper for the complete mathematical formulation of the model. The equation below gives the half size of the receptive field of the network given r the radius of convolution kernels, s the up-sampling factor, Nc the number of cascading blocks and Nr the number of residual blocks per cascading blocks. The receptive field correspond to the size of the border margin in predicted data that are computed from incomplete input data, e.g. for which some of the convolution layers are padding inputs instead of using real input values. This margin should be left aside when computing metrics and losses, as well as when stitching together patch for blocked inference.

$$f(r, s, N_c, N_r) = r \times (1 + s \times (2 + 2 \times N_c \times N_r))$$

Joint 10m - 20m SISR model

Sentinel-2 bands B5-8A and B11-12 have a spatial resolution of 20 meters, while bands B2-4 and B8 have a spatial resolution of 10 meters. To input the full set of bands into the network, the 20-meter bands are first up-sampled to 10-meters, following the idea in (Shao, Cai, et al., 2019) and (Lac et al., 2023). The corresponding workflow is described in Figure 72.

In the remaining of this document, the joint 10m-20m with parameters θ will be noted M_θ.

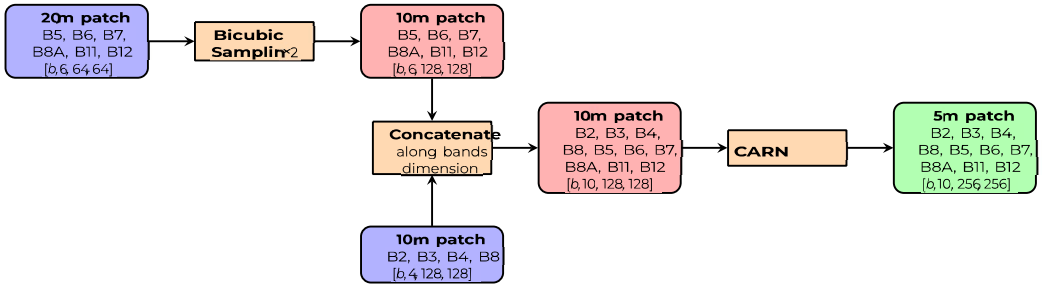


Figure 72. Workflow of the joint 10m-20m SISR model, using the CARN backbone. Tensor shapes are indicated, where b stands for the batch dimension.

3.1.2.3.2 Training strategy

Following (Ledig et al., 2017) and (X. Wang, Yu, Wu, Gu, Liu, Dong, & Qiao Yu and Change Loy, 2018), we adopt a two-stages training strategy. The model is first pretrained with a PSNR oriented loss function. In the second stage the pretrained model serves as a starting point for adversarial fine-tuning. In this section, both stages will be described in detail.

Common steps

In all stages, both input Sentinel-2 patches and reference Venμs patches are standardized band-wise with μ_i and σ_i the band-wise mean and standard deviation estimated on the training set and n the batch dimension and w and h the spatial dimension (note that for Venμs reference patch the band count only goes up to 8).

$$P_{norm}(n,i,w,h) = \frac{P(n,i,w,h) - \mu_i}{\sigma_i}, i = 1 \dots 10$$

All losses computation and gradient back-propagation therefore happen in standardized space, which avoid network to favor toward bands with higher standard deviation. In order to obtain the final predictions from the network, patches need to be de-standardized.

$$P^*(n,i,w,h) = \sigma_i \times (P^*_{norm}(n,i,w,h) + \mu_i), i = 1 \dots 10$$

Moreover, in order to ensure the best possible radiometric consistency between input Sentinel-2 patches and target Venμs patches, the latter are rescaled band-wise to the minimum and maximum value of the corresponding band in the Sentinel-2 patch, before applying standardization. The process where m^{input}_i (resp. M_i^{input}) is the minimum (resp. maximum) value of band i over the input Sentinel-2

patch, and m^{target}_i (resp. M_i^{target}) is the minimum (resp. maximum) value of band i over the target Venμs patch.

$$P^{target}_{minmax}(n,i,w,h) = m_i^{input} + \frac{M_i^{input} - m_i^{input}}{M_i^{target} - m_i^{target}} \times (P^{target}(n,i,w,h) - m_i^{target})$$

Pre-training stage

(a) Smooth L_1 loss on simulated data

It must be stressed that most of the work described in the State-of-the-Art employ simulated dataset, where the low-resolution image is actually a downsampled version of the high-resolution target, yielding data with a very high local consistency. On the other hand, the sen2venμs is composed of image data from 2 different sensors, and despite all the filtering and processing aiming at improving its consistency described in (Michel et al., 2022), the dataset still has discrepancies that may impair local pixel-wise losses. For this reason, the pre-training step will be carried by simulating the Low-Resolution Sentinel-2 patches from the target Venμs patches. The target data are first spatially convolved with a Gaussian Radial Basis Kernel. The width of the gaussian is given by $\sigma^{spat}(s,m)$, where s is the downsampling factor and m relates to the value of the MTF function at Nyquist rate. Output of the spatial convolution is then downsampled by factor s, and a gaussian noise with user-defined per-band standard deviation σ_i^{spect} is added. The 10-meter Sentinel-2 bands (i = 1...4) simulation require a downsampling factor of s = 2, while the 20-meter Sentinel-2 bands that are available in the target Venμs patch require a downsampling factor of s = 4. Finally, the SWIR bands, which are not available in the target Venμs patch, are kept from the input Sentinel-2 patch (i = 9,10).

$$P^{input}_{sim}(n,i\ldots) = \begin{cases} \left(P^{target}(n,i\ldots) \circledast \phi_{\sigma^{spat}(2,m)}\right)\downarrow_2 + \epsilon_i \sim \mathcal{N}(0,\sigma_i^{spect}) & i = 1\ldots4 \\ \left(P^{target}(n,i\ldots) \circledast \phi_{\sigma^{spat}(4,m)}\right)\downarrow_4 + \epsilon_i \sim \mathcal{N}(0,\sigma_i^{spect}) & i = 5\ldots8 \\ P^{input}(n,i\ldots) & i = 9,10 \end{cases}$$

$$\phi_\sigma(x,y,x_c,y_c) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-x_c)^2-(y-y_c)^2}{2\sigma^2}}$$

$$\sigma^{spat}(s,m) = \frac{s}{\pi}\sqrt{-2*ln(m)}$$

This simulation process can actually be seen as some kind of data augmentation, and during the pre-training, we will take full advantage of this by choosing randomly drawing values for m and σ_i from a uniform distribution within user-defined range. This allows to regularize the learned model and ensure better generalization to unseen data.

Using those notations, the full resolution loss term L_{FR} for a minibatch for size N, where L^{smooth}_1 is the Smooth L_1 Loss and introduced in (Girshick, 2015b). Smooth L_1

loss behaves like L_1 loss far from 0 and like L_2 loss in the vicinity of 0. It therefore retains the good properties of L_2 while being less sensitive to outliers. The full resolution loss term corresponds to the smooth L_1 loss between the output of the model M_θ applied to data simulated from the reference $P_{sim}^{input}(n,i...)$) and the

$$\mathcal{L}_{FR}(\theta) = \frac{1}{N} \sum_{\substack{i=1...8 \\ n=1...N}} \mathcal{L}_1^{smooth}(\mathcal{M}_\theta(P_{sim}^{input}(n,i\ldots)), P^{target}(n,i\ldots))$$

$$\mathcal{L}_1^{smooth}(x,y) = \begin{cases} 0.5(y-x)^2 & if|y-x|<1 \\ |y-x|-0.5 & otherwise \end{cases}$$

reference $P^{target}(n,i...)$, restricted to the 8 bands that are available in reference Venμs patch. Also note that loss computation is spatially restricted to pixels that are further away from patch borders than the receptive field, in order to avoid computing loss values on pixel that did not see the full receptive field of the network.

(b) Smooth L_1 loss on downscaled data for B11 and B12 training

The full resolution loss term cannot be computed for the 2 last bands (B11 and B12), which are not available in the reference Venμs patch. In order to supervise the output of model $M(\theta)$ for those bands, we will make use of the Wald trick (Wald et al., 1997a), and downscale all the input data by a factor of 4 to make up a low resolution input, while down-scaling the 10-meter input bands (i = 1...4) by a factor of 2 and retaining the 20-meter input bands (i = 5...10), in order to make low resolution target, as described in the equation below.

$$P_{wald}^{input}(n,i\ldots) = \left(P^{input}(n,i\ldots) \circledast \phi_{\sigma^{spat}(4,m)}\right) \downarrow_4 + \epsilon_i \sim \mathcal{N}(0,\sigma_i^{spect})$$

$$P_{wald}^{target}(n,i\ldots) = \begin{cases} \left(P^{input}(n,i\ldots) \circledast \phi_{\sigma^{spat}(2,m)}\right) \downarrow_2 + \epsilon_i \sim \mathcal{N}(0,\sigma_i^{spect}) & i=1...4 \\ P^{input}(n,i\ldots) & i=5...10 \end{cases}$$

Those simulated patches now allow to supervise the reconstruction of B10 and B11, albeit at lower resolution. This is achieved by defining L_{Wald}, which is defined as the smooth L_1 loss between model $M(\theta)$ applied to low resolution input $P_{wald}^{input}(n,i...)$ and low-resolution target $P^{target}_{wald}(n,i...)$, restricted to the last 2 bands (B11 and B12). Note that this with sen2venμs_extended patch sizes, model output for low resolution patch will 64x64. Depending on the depth of the network, this might be smaller than the margin imposed by receptive field. Since we only compute loss values on pixel that see the full receptive field, patch size for reduced resolution training set have been artificially augmented by mirroring.

$$\mathcal{L}_{Wald}(\theta) = \frac{1}{2} \sum_{\substack{i=11,12 \\ n=1...N}} \mathcal{L}_1^{smooth}(\mathcal{M}_\theta(P_{wald}^{input}(n,i\ldots)), P_{wald}^{target}(n,i\ldots))$$

(c) Total minibatch loss and optimization

The total minibatch loss for pre-training is the sum of the full resolution loss term $L_{FR}(\theta)$ and reduced resolution loss term $L_{Wald}(\theta)$. It allows to jointly optimize the reconstruction of the 10 spectral bands within a single model and optimization process.

$$\mathcal{L}_{pretrain}(\theta) = \mathcal{L}_{FR}(\theta) + \mathcal{L}_{Wald}(\theta)$$

The model parameters θ are obtained through the minimization of $L_{pretrain}(\theta)$ by Stochastic Gradient Descent (SGD) using the ADAM method (Kingma & Ba, 2014a). The learning rate is scheduled with the Cosine Annealing with Warm Restarts method (Loshchilov & Hutter, 2016b) which slowly decreases the learning rate, and periodically reset it to the highest value. The period of these restarts also increases during learning. The rationale for using such a learning rate scheduler is that we want to avoid very narrow local minima, because narrow local minima will lead to poor generalization if the unseen data slightly differ from the training data. Periodically resetting the learning rate allow to get out of such minima and promote wider minima.

Adversarial fine-tuning stage

Once the model has been pre-trained on simulated data, the fine-tuning step aims at pushing the pre-trained model towards the manifold of real data, by means of adversarial training. We adopt the generic framework of (X. Wang, Yu, Wu, Gu, Liu, Dong, & Qiao Yu and Change Loy, 2018), with some adaptations as described in the following sections.

(a) Discriminator

The discriminator D is a standard UNet (Ronneberger et al., 2015), with 4 levels and without skip connections. It has 64 features at the highest resolution, and the last layer outputs a single value as the discriminator decision on the input patch. Additionally, we perform high pass filtering of the input patches in order to focus the adversarial training on high spatial frequencies.

(b) Adversarial formulation

Similar to (X. Wang, Yu, Wu, Gu, Liu, Dong, & Qiao Yu and Change Loy, 2018), we use the Relativistic Average Discriminator, where D is the discriminator network and σ the sigmoid function. Given real (target) data x_r and fake (super-resolved) data x_f, we train discriminator network D so that $D_{Ra}(x_r,x_f) \to 1$ and $D_{Ra}(x_f,x_r) \to 0$. Intuitively, this means that real data should have higher discriminator response than the average of fake data over the minibatch, and that fake data should have lower response than the average of real data over the minibatch.

$$\mathcal{D}_{Ra}(x, y) = \sigma(D(x) - \mathbb{E}[D(y)])$$

From there, we can define the loss function for the discriminator, which minimize Binary Cross Entropy (BCE) between target label 0 for fake data and (1 − *l*) for real data and output from D_{Ra}. *l* denotes a label smoothing value preventing the network to become overconfident in recognizing the real (Venμs target) data.

$$\mathcal{L}_D^{Ra} = -(1-l) \cdot \mathbb{E}_{x_r}[log(\mathcal{D}_{Ra}(x_r, x_f))] - \mathbb{E}_{x_f}[log(1 - \mathcal{D}_{Ra}(x_f, x_r))]$$

Likewise, we define generator loss, which essentially plays the opposite game and tries to trick the discriminator network into labelling real data with 0 and fake data with 1. Note that we do not employ label smoothing this time.

$$\mathcal{L}_G^{Ra} = \mathbb{E}_{x_f}[log(\mathcal{D}_{Ra}(x_f, x_r))] - \mathbb{E}_{x_r}[log(1 - \mathcal{D}_{Ra}(x_r, x_f))]$$

(c) Adversarial training

During adversarial training, we alternate between optimization steps of the Discriminator and optimization steps of the Generator. The generator loss combines the loss from pre-training with the relativistic average Generator, as stated in the equation below. This ensure that we remain in the vicinity of the optimum found during pre-training in the hyper-parameters space, while trying to push the model output further into the manifold of real data.

$$\mathcal{L}_{finetune}(\theta) = \mathcal{L}_{pretrain}(\theta) + \lambda \cdot \mathcal{L}_G^{Ra}(\theta)$$

3.1.2.4 Benchmarking & evaluation

This section describes the set of metrics that will be used to assess the performance of the super-resolution model and compare models with each other.

Each metric is measured on each band separately. During testing time, best model inference is performed on both simulated and real input data. This allows to assess perform in real conditions as well as with perfect data that do no not suffer from discrepancies such as local mis-registration.

Standard image quality metrics

The first set of quality metrics that will be used is a selection of standard Image Quality (IQ) metrics (Vivone et al., 2014). This includes Root Mean Square Error (RMSE) and Peak Signal-To-Noise Ratio (PSNR), which is derived from RMSE, as well as the Structural Similarity Image Metric (SSIM).

Radiometric consistency-oriented metrics

One very important property of the trained models for end-users is their ability to preserve the radiometric properties of the input Sentinel-2 image. In order to do

that, 2 specific metrics have been designed: the LR fidelity loss, and the low gradient Strata RMSE.

(a) LR Fidelity

For band *i* and corresponding up-sampling factor *s* = 2 or *s* = 4 and superresolved patches $P^{\star}$, the Low-Resolution Fidelity term.

Low Resolution Fidelity is the root mean-squared error between the input Sentinel2 image and the super-resolved output, down-scaled back to the input resolution after filtering by MTF based spatial kernel.

$$\mathcal{L}_{LR}(i) = \sqrt{\sum_{n,w,h} \left(P^{input}(n,i,w,h) - \left(P^{\star}(n,i,w,h) \circledast \phi_{\sigma^{spat}(s,m)}\right) \downarrow_s \right)^2}$$

(b) Low Gradient Magnitude Strata RMSE

A second metric that is used to determine radiometric consistency is the RMSE on the low gradient magnitude strata. First, the distribution of gradient magnitude is estimated on the training set. Then, the 25\quantile of gradient magnitude is computed from this distribution. Low Gradient Magnitude Strata RMSE is defined as the RMSE of pixels whose Gradient Magnitude lies below this 25\threshold. Intuitively, pixels in the low Gradient Magnitude Strata are in areas where there is no textures or edges. They should therefore be radiometrically stable through SISR. Figure 73 gives an example of Low Gradient Stratas over a Venμs image.



Figure 73. Example of 25% Low Gradient Magnitude Strata (blue) and 25% High Gradient

Magnitude strat (red)

High spatial frequencies oriented metrics

Metrics in this section have been designed to assess how much High Spatial frequencies (e.g. fine details) have been recovered by the super-resolution algorithm.

(a) HR Fidelity

The High-Resolution Fidelity metric is essentially the RMSE between high-pass filtered super-resolved image and the high-pass filtered reference (Venμs) image.

$$\mathcal{HPF}(P) = P - P \circledast \phi_{\sigma^{spat}(s,m)}$$

$$\mathcal{L}_{HR}(i) = \sqrt{\sum_{n,w,h} \left(\mathcal{HPF}(P^{target}(n,i,w,h)) - \mathcal{HPF}(P^{\star}(n,i,w,h))\right)^2}$$

(b) High Gradient Magnitude Strata RMSE
Similar to the Low Gradient Strata RMSE metric, the High Gradient Strata RMSE is the RMSE of pixels belonging to the highest 25\strata. Such pixels should exhibit strong textures or sharp edges and should benefit the most from SISR.

No Reference IQ metrics

Because the Sen2VEnμs dataset is composed of source and target images from different sensors, and despite all the efforts to harmonize those images, discrepancies are still present, which impair Full Reference IQ metrics evaluation. Therefore, no reference IQ metrics will also be used to assess the intrinsic quality of super-resolved image.

(a) Variation of Total Variation

Total Variation (TV) is the sum of gradient magnitude over image. Intuitively, high image noise should yield higher TV values that low image noise, and sharp edges should yield higher TV than blurry images.

$$\mathcal{TV}(P,i) = \sum_{n,w,h} \sqrt{|P(n,i,w+1,h) - P(n,i,w,h)|^2 + |P(n,i,w,h+1) - P(n,i,w,h)|^2}$$

Values of TV are unbounded and dependent on the radiometric range of the image. Therefore, in this work, we will compute the percentage of variation of TV with respect to the input bicubic image as well as to the reference Venμs image.

(b) Variation of High Frequencies Power in Fourier Domain

In Figure 74, the benefits of SISR with respect to plain bicubic up-sampling can be observed in the spatial frequency domain: it can be observed that frequencies past the input Nyquist rate (around 1/4th in each image quarter) have much more power in the SISR predicted image than in the bicubic up-sampled one. In order to get a quantitative measure of this gain, we define the HF Power Variation metric as:

$$\mathcal{HFPV}(P*, P^{ref})(i) = \frac{1}{n} \sum_{\substack{n \\ f_x > F_e/2 \\ f_y > F_e/2}} \frac{|\mathcal{FFT}(P*(n,i,\ldots)|^2 - |\mathcal{FFT}(P^{ref}(n,i,\ldots)|^2}{|\mathcal{FFT}(P^{ref}(n,i,\ldots)|^2}$$

HFPV will be computed against bicubic up-sampled input Sentinel-2 image as well as against Venμs reference image.

(c) BRISQUE and BRISQUE Variations

The Blind Referenceless Image QUality Evaluator (BRISQUE) has been proposed in [Mittal et al., 2012] is a holistic measure of IQ based on local statistics that are mapped to quality score using Support Vector Machine Regressor (SVR) trained on a quality annotated images database. BRISQUE yields a score between 0 and 100, 0 being the best quality and 0 being the poorest quality. In this work, we compute both the BRISQUE score of the super-resolved image as well as the variation of the BRISQUE score with respect to bicubic up-sampled images as well as to reference Venμs image. We use Pytorch implementation of (Kastryulin et al., 2022).

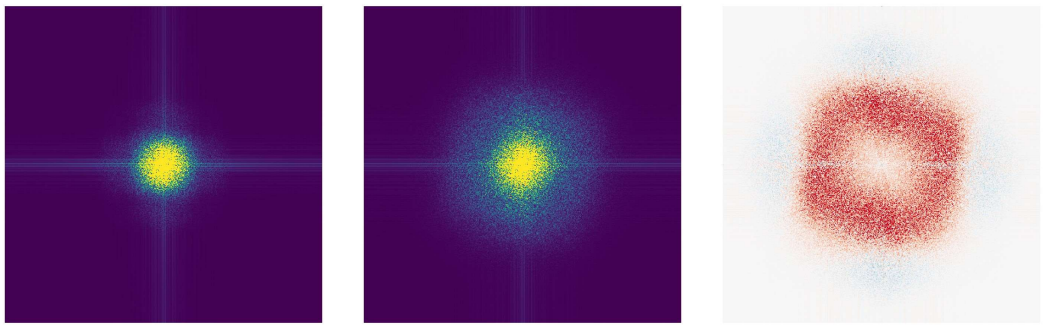


Figure 74. From left to right : FFT of bicubic up-sampled Sentinel-2 image, FFT of superresolved image, difference between both FFT (red means high FFT on SISR), for band B7.

3.1.3 Implementation and results

3.1.3.1 Models training

All models were implemented using Pytorch

(Paszke et al., 2017; Paszke, Gross, Massa, Lerer, Bradbury, Chanan Gregoryand Killeen, et al., 2019)

. This work was performed using HPC resources from GENCI-IDRIS (Grant 2023-AD010114835) and all experiments have been conducted on NVIDIA A100 GPU nodes of the Jean-Zay cluster in France. Batch size is set to 64 for all experiments.

3.1.3.1.1 Configurations

Models configuration investigated in the study are presented in Table 31 with their associated number of parameters and half size of receptive field. CARN models with different number of blocks from 1 to 4 are investigated. We also tested one configuration of ERSGAN Generator network (X. Wang, Yu, Wu, Gu, Liu, Dong, & Qiao Yu and Change Loy, 2018). Model carn_3x3x64g4sw will be our reference model to assess various loss functions and compare with other parameters.

3.1.3.1.2 Pre-training

Unless stated otherwise, all models were pre-trained with the loss described in section 2c. Additionally, model carn_3x3x64g4sw has been pre-trained with classic Smooth L_1 loss, as well as with a combination of HR and LR loss as described in sections 2a and 3a. Models are trained for 20 epochs. Learning rate is set to 2e-4, with an initial restart period of 750 which doubles every time. Loss values are computed for pixels that have full input data, e.g. excluding a margin that correspond to the size of the receptive field in Table 31. Metrics are computed on the testing set (see Table 31) with a fixed validation margin of 64 pixels, which is larger than the largest receptive field. This ensure that all models will be compared on the same set of pixels.

Table 31. Models investigated in the study. Naming of configuration for carn is nb_cascading_blocks x nb_eres_blocks_per_cascading_blocks x nb_features. g1 indicates that grouping is set to 1 in convolution layers, and sw indicates if weights are shared between ERes blocks of a given cascading block.

Model.	Nb. of parameters	Receptive field
carn_1x3x64g4sw	1.243.530	29
carn_2x3x64g4sw	1.867.914	17
carn_3x3x64g4sw	2.492.298	41
carn_4x3x64g4sw	3.116.682	53
carn_3x3x64g4	4.659.594	41
carn_3x3x64g1sw	5.146.506	41
esrgan_3x64x32	2.280.662	36

3.1.3.1.3 Fine-tuning

Fine-tuning stage tarts from the best checkpoint according to validation loss (e.g. loss computed on the validation set) identified during pre-training. Models are fine-tuned for 10 additional epochs. Learning rate for Discriminator optimization is set to 1e-5, while learning rate for Generator is set to 2e-5. Both are schedule with Cosine Annealing with Warm Restarts with a restart period of 160 000, which is higher than the number of fine-tuning epochs. This ensures a very slow decrease of the learning rate throughout the fine-tuning stage. Label smoothing is set to 0.1. Discriminator uses all bands except B11 and B12 which are missing from the full resolution target Venμs patches. MTF value for HPF prior to discriminator is set to 0.2.

One difficulty with adversarial training is to know how to select good checkpoints. Indeed, as the Discriminator and Generator are both optimized against each other during training, the relativistic adversarial loss cannot be used to compare different steps of fine-tuning : a weak discriminator may lead to a very low value even for predictions with poor quality, whereas a sharp discriminator may beat the Generator even for very high-quality predictions. For that reason, in order to automate the model selection and avoid relying on manual qualitative assessment of checkpoints, we use an external metric to assess the quality model at current step. BRISQUE shows very interesting properties to serve this purpose (see analysis of pre-training in section 3.2.3) and will thus be used to select the best model.

3.1.3.2 Performances assessment

While PSNR has often been used as the metric to go for comparing models, in the case of sen2venμs dataset, where data come from different sensors, it does not yield meaningful results. Figure 75 shows PSNR obtained over the testing set for the carn_3x3x64g4sw model trained with different loss functions, while Figure 76 shows sample predictions from those models. One can observe that according to PSNR, the SISR predictions do not bring significant improvements for 10 meters bands. The l1_real and hr_lr_l1_real yield slightly higher PSNR values than bicubic up-sampling while l1_sim yield slightly lower PSNR values. However, looking at predicted patches on Figure 76, the latter clearly outperforms both bicubic-up-sampling and the two other models visually.

This means that in this work, we cannot rely on traditional FR metrics with respect to the target Venμs image : other discrepancies are of higher order of magnitude than the improvement in spatial details that we intend to measure. We will therefore rely on metrics presented in section 2 to assess radiometric consistency with input Sentinel-2 images, metrics presented in section 3 to assess improvement of high-resolution details, and metrics presented in section 4 to assess the overall image quality of predictions. Though all those metrics have been computed, we will only present the most meaningful ones in this report.

In the remaining of this section, we will focus on the influence of training loss, the influence of CARN parameters, and the influence of different model depth. This analysis is carried on pre-trained model. Performances gain related to fine-tuning will be discussed in the last sub-section.

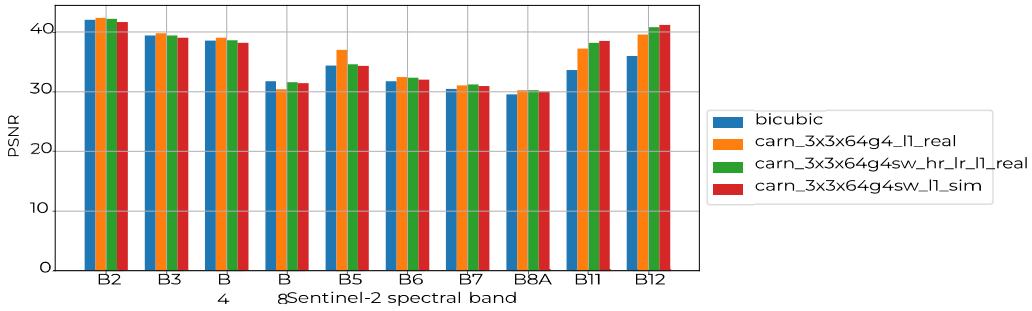


Figure 75. PSNR achieved by the carn_3x3x64g4sw trained with different training losses, during inference on real input data over the testing set.

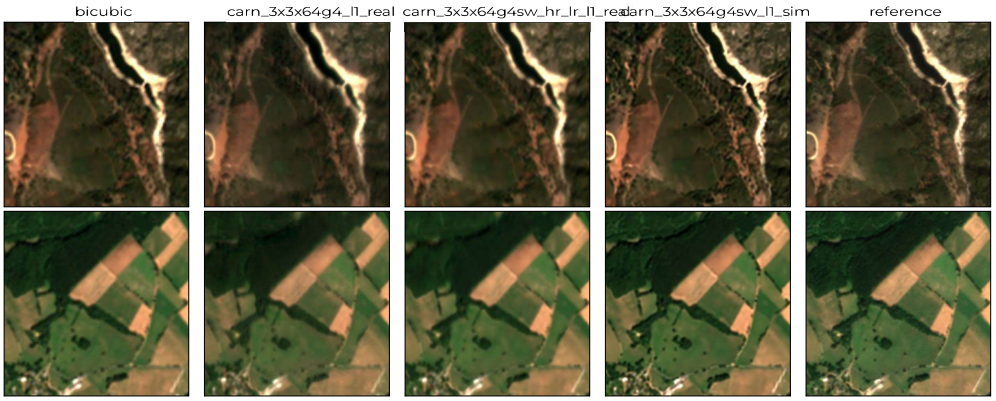


Figure 76. RGB Rendering of predictions from the carn_3x3x64g4sw trained with different training losses, on real patches from the testing set.

3.1.3.2.1 Influence of training loss

In this section, we discuss the choice of loss for pre-training. We compare the L^{smooth}_1 loss on simulated data, as described in section 2a, with plain L^{smooth}_1 on real data, which would be the default choice for SISR methods. We also compare with the HR-LR L^{smooth}_1 loss on real data, following the principles presented in sections 3a and 2a. The rationale for the latter is to train the network to optimize simultaneously the radiometric consistency and the HR details improvement.

Radiometric consistency with input Sentinel-2 images

Figure 77 shows the LR Fidelity metric for both the bicubic up-sampling and the carn_3x3x64g4sw pre-trained with the 3 losses. It can be observed that training the model with plain L^{smooth}_1 on real data leads to significantly worst radiometric consistency with the input data. This is because Venμs has indeed radiometric inconsistencies with Sentinel-2, because of different Spectral Sensitivity Response and differences in viewing angles. The model trained against real data unfortunately learns this mapping and predict reflectance values that are more consistent with Venμs than with the input Sentinel-2 image. On the contrary, the HR-LR L^{smooth}_1 loss includes the LR Fidelity term that enforces better consistency with Sentinel-2. Still the best consistency is obtained with the L^{smooth}_1 loss on

simulated data. Indeed, simulated data are perfectly coherent: there is no radiometric distortion between the input and the target to be learned by the model. These differences in radiometric distortion can be observed in Figure 76, Figure 78 and Figure 79, that shows the prediction from the different models for different colour composites.

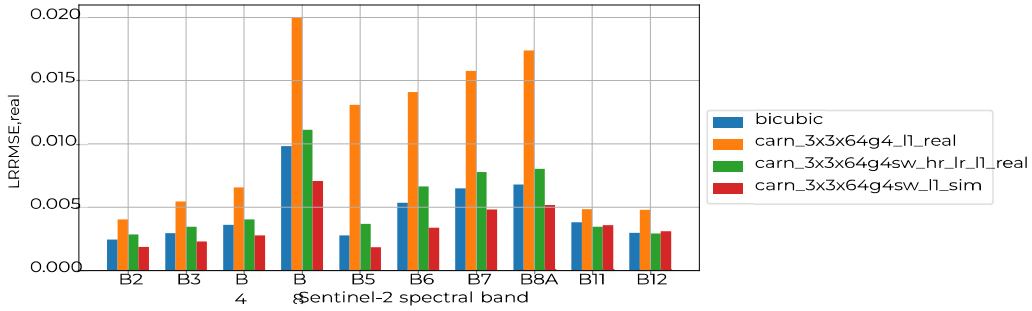


Figure 77. LR RMSE on real data for the carn_3x3x64g4sw trained with different losses functions.

Improvement of HR details

Figure 80 shows the HR Fidelity metric computed from inference on real data, while Figure 81 shows the same HR Fidelity metric computed on simulated data. Both graphs show a clear improvement of HR details for the 20-meter bands. However, for the 10-meter band, HR Fidelity on real data barely show any improvement with respect to the bicubic up-sampling, which is inconsistent with the visual inspections on Figure 76, Figure 78 and Figure 79. HR fidelity on simulated data is more consistent with qualitative observations and shows that training $L^{smooth}{}_1$ on real data does not outperform bicubic up-sampling for the 10-meter bands. The HR-LR $L^{smooth}{}_1$ seems to perform equally with the $L^{smooth}{}_1$ on simulated data, which is not backed by visual inspection, for which the latter is clearly better performing. This shows the limitations of the HR Fidelity metric and justify the use of NR QI metrics in the following sections.

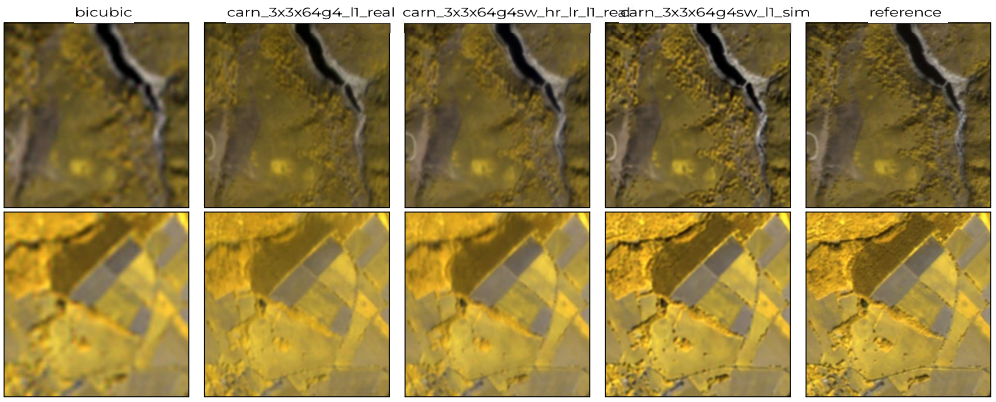


Figure 78. Red Edges bands (B7,B6,B5) rendering of predictions from the carn_3x3x64g4sw trained with different training losses, on real patches from the testing set.

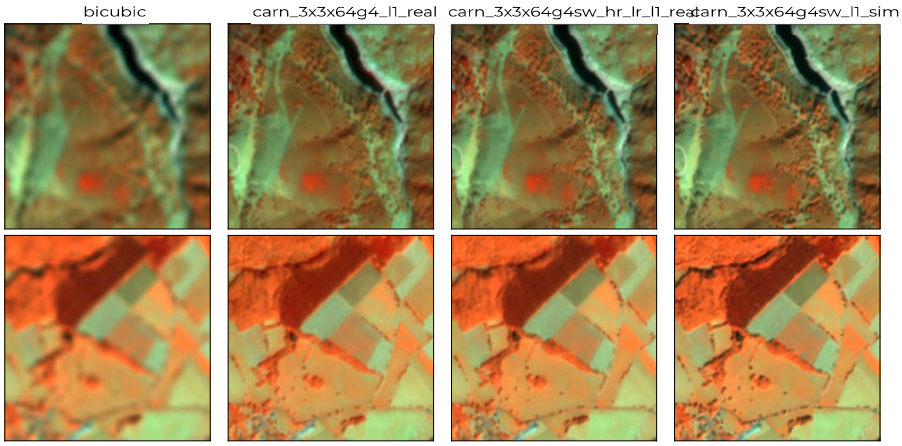


Figure 79. NIR and SWIR bands (B8,B9,B10) rendering of predictions from the carn_3x3x64g4sw trained with different training losses, on real patches from the testing set.

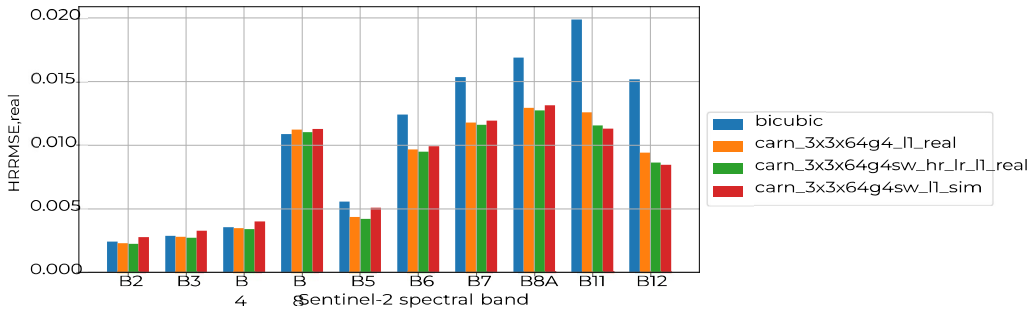


Figure 80. HR RMSE on real images for the carn_3x3x64g4sw trained with different losses functions.

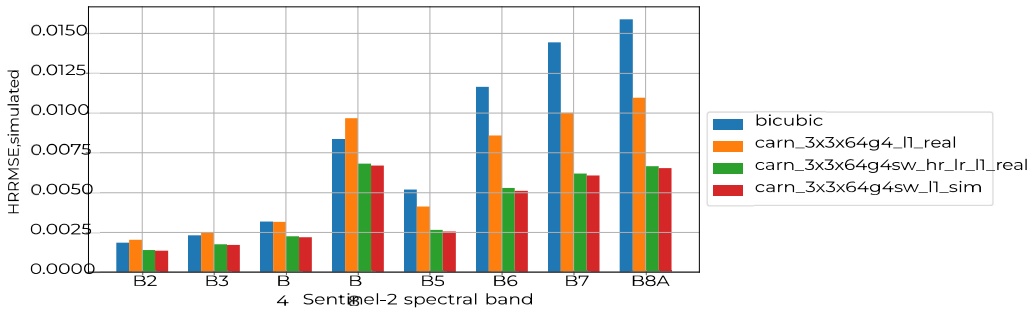


Figure 81. HR RMSE on simulated images for the carn_3x3x64g4sw trained with different losses functions.

Conclusion on loss functions for pre-training

Among the 3 options that were tested, the L^{smooth}_1 on real data fails to achieve radiometric consistency with the input Sentinel-2 images, and to perform better than bicubic up-sampling on 10-meter bands. Though the HR-LR L^{smooth}_1 performs better in both criteria, The L^{smooth}_1 on simulated data is a best option both qualitatively and quantitatively. It will therefore be used for the pre-training step of all remaining experiments.

3.1.3.2.2 Influence of shared weights and grouping

In this section, we investigate 2 parameters of the CARN model: shared weights within a given cascading block (sw) and features groups in convolution layers of the residual blocks. Starting with the base model carn_3x3x64g4sw, which has shared weights and a grouping of 4, we disable weight sharing on one hand and reduce grouping to 1 on the other. It must be stressed than though both yields a significantly higher number of parameters as shown in Table 31, they should have limited impact on inference time, since the overall structure of the model does not change. To compare the 3 settings, we will make use of HR RMSE in Figure 82 and BRISQUE score in Figure 83, as well as visual inspection on Figure 84. HR RMSE does not show any significant difference. BRISQUE score seems to indicate that disabling weight sharing may improve image quality slightly, but this can hardly be noted on Figure 84. At least for the pre-training set-up those two options seem to have no impact on performances. We will therefore stick to our default (grouping of 4 and shared weights).

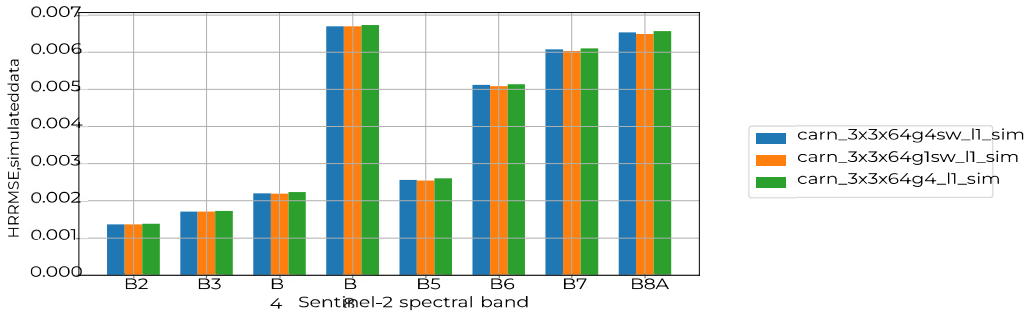


Figure 82. HR RMSE metric on simulated data for different CARN parameters.

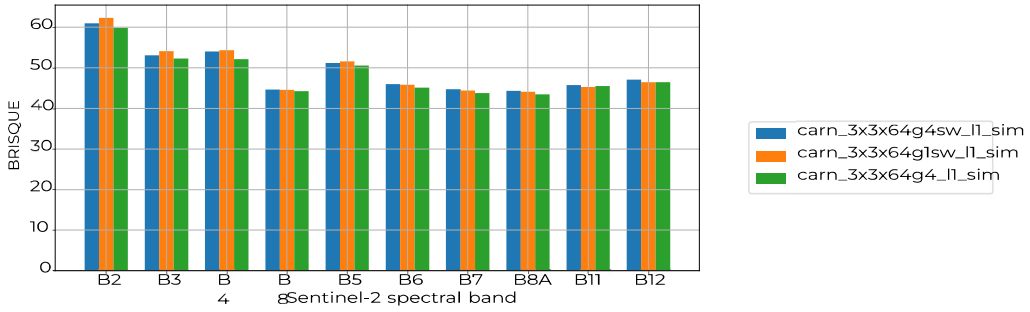


Figure 83. BRISQUE score on simulated data for different CARN parameters.

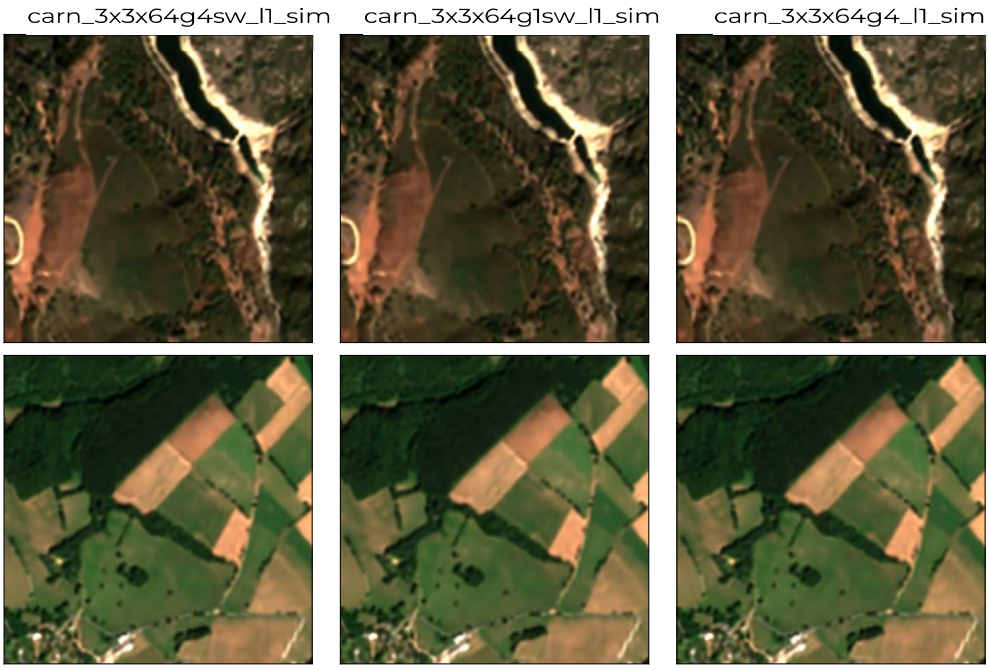


Figure 84. RGB Rendering of prediction starting from carn_3x3x64g4sw, and disabling shared weights and using a grouping of 1, on real patches from the testing set.

3.1.3.3 Model depth

Since our goal is to provide a model that can be applied at large scale on Sentinel-2 data to streamline SISR with the EvoLand prototypes, we are looking for the best trade-off between model complexity (and thus, inference time) and performances. Model depth is probably the most important parameter in that perspective. In this section we will vary the number of cascading blocks of CARN from 1 to 4, all other parameters being equal. We will also compare with an instance of ESRGAN generator network, trained in the same exact pre-training conditions. To compare the different models, we will use the HR RMSE metric on simulated data, as well as the BRISQUE score and the HF power variation with respect to bicubic up-sampling described in section 4b. The two latter NR QI metrics are computed on super-resolved patches from real data and are therefore more akin to measure the quality of the model in operational conditions.

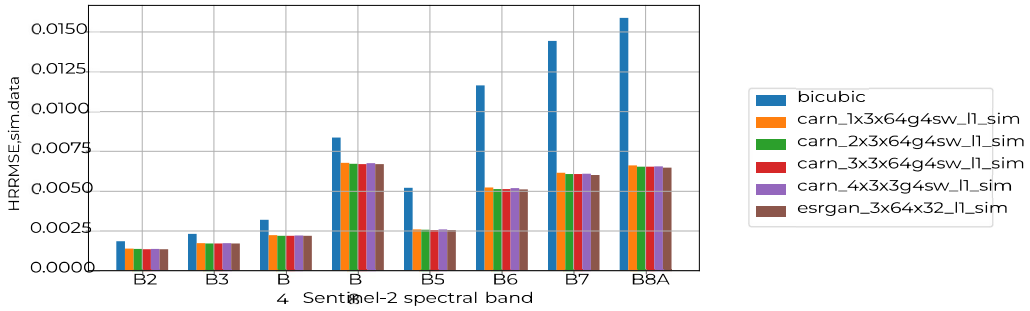


Figure 85. HR RMSE metric on simulated data, for varying model depths.

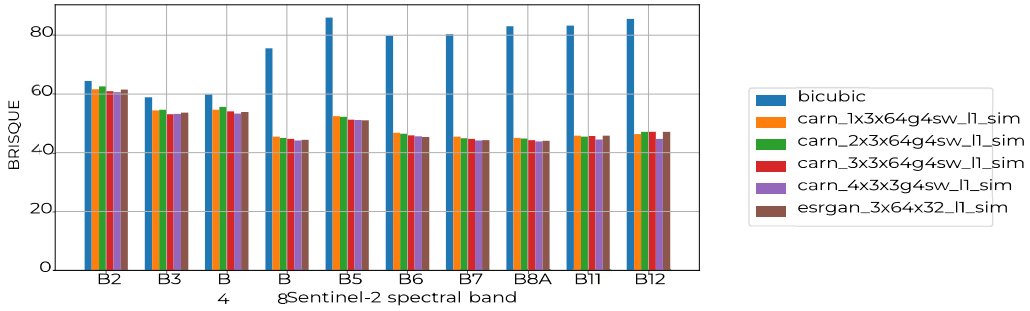


Figure 86. BRISQUE score on real data, for varying model depths

In Figure 85, HR RMSE metric on simulated data does not allow to distinguish between the different models. They all seem to perform equally on the 10-meter bands, with very subtle variations for the 20-meter bands. It should be noted however that all models outperform bicubic up-sampling by a vast margin. BRISQUE score displayed in Figure 86 enables a better separation of models, with deeper models performing slightly better than shallower ones. It can be observed that ESRGAN performs slightly worse than CARN with 3 or 4 cascading blocks on the 10-meter bands but seems a slightly better option on B5 and B6. In Figure 87, it can be observed that the CARN model with 3 cascading blocks seems to bring the largest increase of HF power with respect to bicubic up-sampling for the 10-meter bands, while the CARN model with 4 cascading blocks brings more HF power increase for the 20-meter bands. Both out-perform ESRGAN generator in almost all spectral bands. It should be noted that larger increase of HF power may be caused by either sharper spatial details or higher level of noise. A good measure of the level of noise can be obtained by the TV, as presented in section 4a. In Figure 88, we compare the variation of TV with respect to Venμs reference. We can observe that in all cases, the absolute variation does not exceed 0.25\predicted images is consistent with the level of noise in Venμs image, for all model depths. Last, visual inspection on Figure 89 does not allow to qualitatively distinguish between model depths.

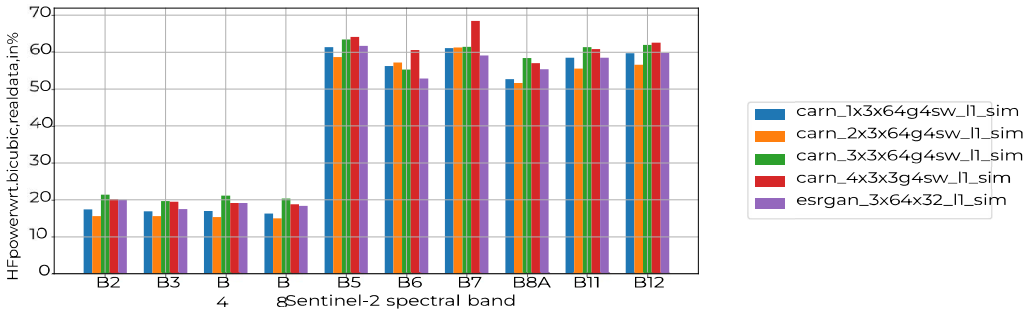


Figure 87. HF power variation with respect to bicubic on real data, for varying model depths.

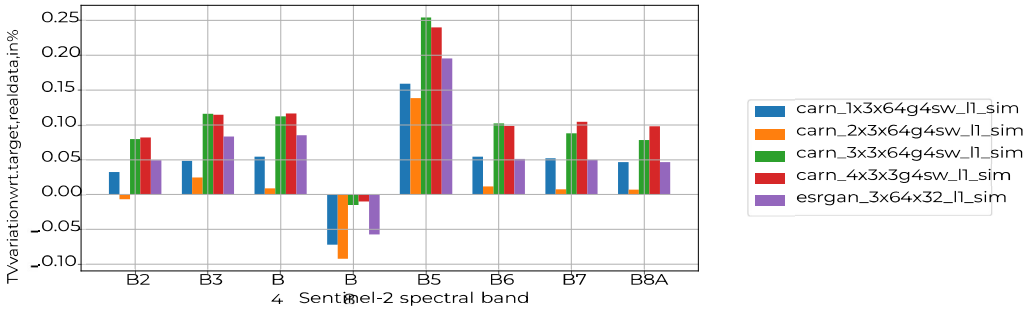


Figure 88. TV variation with respect to target on real data, for varying model depths.

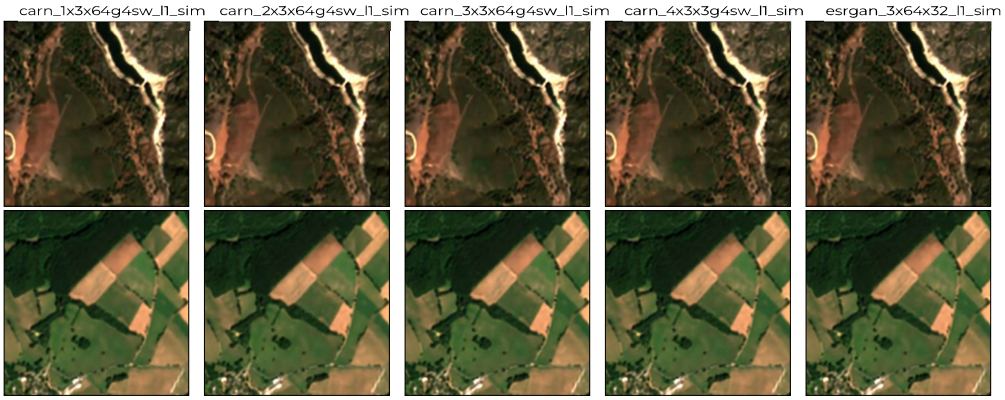


Figure 89. RGB Rendering of predictions of models with different depths, on real patches from the testing set.

CARN model with 4 cascading blocks brings more HF power increase for the 20-meter bands. Both out-perform ESRGAN generator in almost all spectral bands. It should be noted that larger increase of HF power may be caused by either sharper spatial details or higher level of noise. A good measure of the level of noise can be obtained by the TV, as presented in section 4a. In Figure 88, we compare the variation of TV with respect to Venμs reference. We can observe that in all cases, the absolute variation does not exceed 0.25\predicted images is consistent with the level of noise in Venμs image, for all model depths. Last, visual inspection on Figure 89does not allow to qualitatively distinguish between model depths.

In pre-training conditions, there are no significant differences among the deeper CARN configuration with 3 and 4 cascading blocks. There are also no differences between CARN with 3 cascading blocks and ESRGAN with 3 blocks, two models that have approximately the same number of parameters (see Table 31). This backs up our choice of carn_3x3x64g4sw as our reference model in this work.

3.1.3.4 Fine-tuning

In this section, the performance of the fine-tuning stage is investigated. Fine-tuning has been performed with the reference architecture carn_3x3x64g4sw, as well as with same architecture with varying number of cascading blocks (1 to 4). Fine-tuning has also been performed for esrgan_3x64x32 and carn_3x3x64g4, which has same inference cost but more parameters to train than the reference architecture. Last, fine-tuning has also been performed for reference architecture using the discriminator without input HPF, for the sake of comparison. After 10 epochs, the model with the best average BRISQUE score over all bands is selected.

Figure 90 shows the evolution of the BRISQUE score for all fine-tuned models with respect to the pre-trained models. For all models, fine-tuning allows an improvement of 10 to 15 points of BRISQUE score depending on the spectral band (except for band B11 and B12, which are not processed by the discriminator). Interestingly, 20-meter resolution bands B5 to B8A since to benefit more from fine-tuning than the 10-meter resolution bands. For those band, esrgan_3x64x32 as well as carn_3x3x64g4 slightly outperform the other models, while for

10-meter bands, there is no model standing out. The effect on BRISQUE score of disabling the HPF in discriminator is unclear. Figure 93, Figure 94 and Figure 95 present predicted patches from the training set, in a colour composition with Red Edge bands, for all models, pre-trained and fine-tuned. Fine-tuned images look indeed slightly sharper that pre-trained one, but this improvement is barely noticeable.

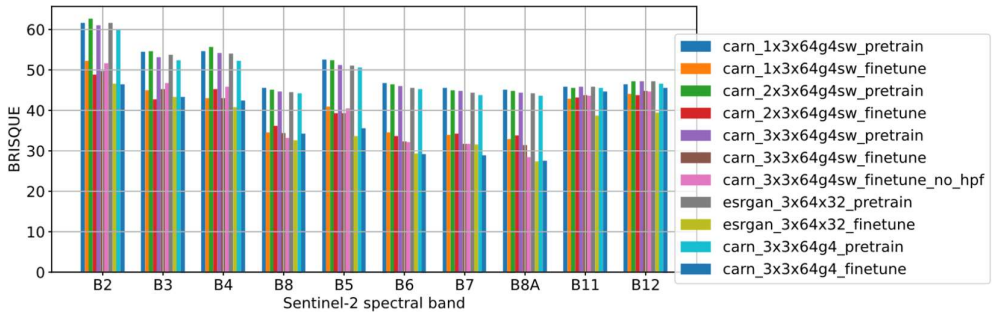


Figure 90. BRISQUE score on real data, for pretrained vs. fine-tuned models.

Figure 91 shows the HF power variation with respect to bicubic on real data, for both finetuned models and pre-trained models. Conclusions are similar to the analysis of the BRISQUE score, but interestingly esrgan_3x64x32 and carn_3x3x64g4 show a larger increase of power between fine-tuned and pre-trained version. If we look at Figure 95, we can observe strong patterns on forest

areas that are probably hallucinated and may be the cause for a higher increase of HF power. This is also slightly noticeable in Figure 94 for the carn_3x3x64g4sw trained with HPF.

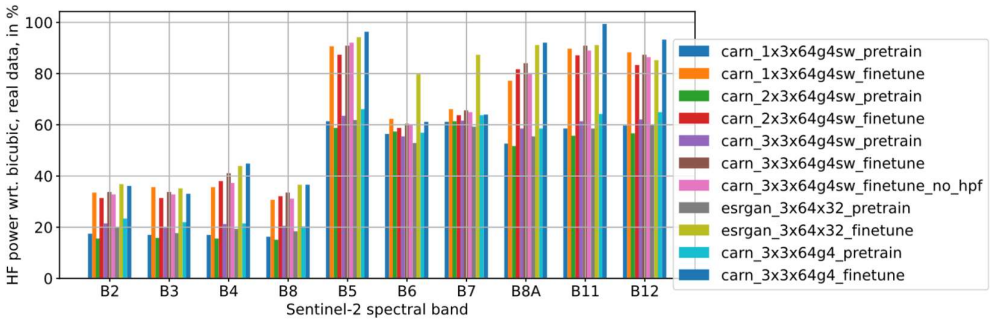


Figure 91. HF power variation with respect to bicubic on real data, for pretrained vs. finetuned models.

Figure 92 shows the variation of Total Variation on real data with respect to target Venμs patch. It can be observed that pre-trained models all have a TV value close to the target, and that fine-tuning seem to increase TV by up to 0.6\TV. This may suggest that finetuning may be introducing a bit of noise and artifacts, especially for the carn_3x3x64g4 and esrgan_3x64x32 which have the higher increase. Band B5 is the spectral band with the most pronounced effect. Note however that this effect is probably very limited, since the maximum increase value is relatively low.

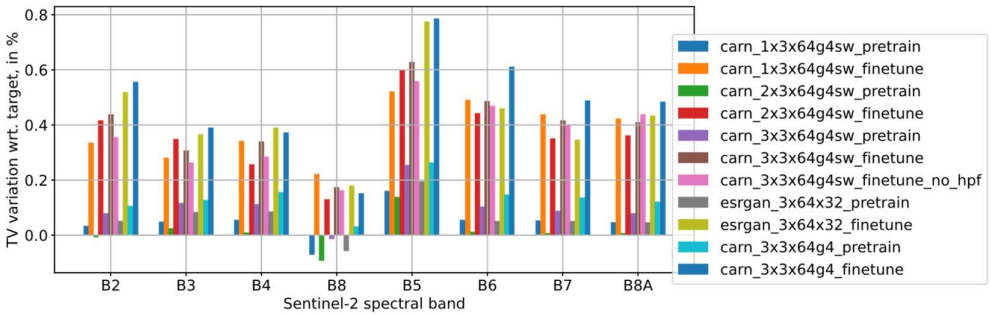


Figure 92. Total Variation with respect to target on real data, for pretrained vs. fine-tuned models.

3.1.3.5 Selected model and inference time

From those experiments, we select carn_3x3x64g4sw_finetuned as our best model, as it offers the best trade-off between performances and complexity. However, if processing time is considered critical for downstream applications, carn_1x3x64g4sw_finetuned would be a perfectly acceptable option, already offering solid resolution gains.

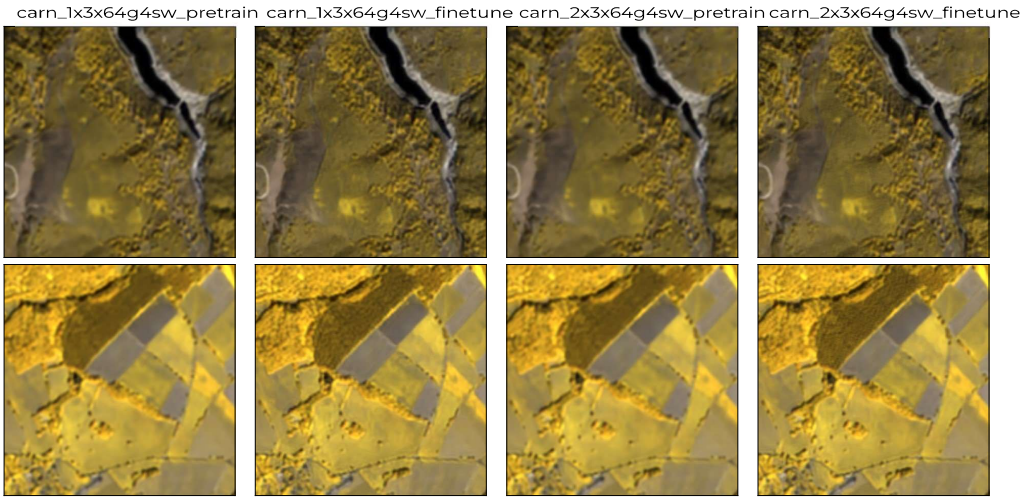


Figure 93. Red Edges bands (B7,B6,B5) rendering of predictions of fine-tuned vs. pre-trained models, form carn_1x3x64g4sw and carn_2x3x64g4sw, on real patches from the testing set.

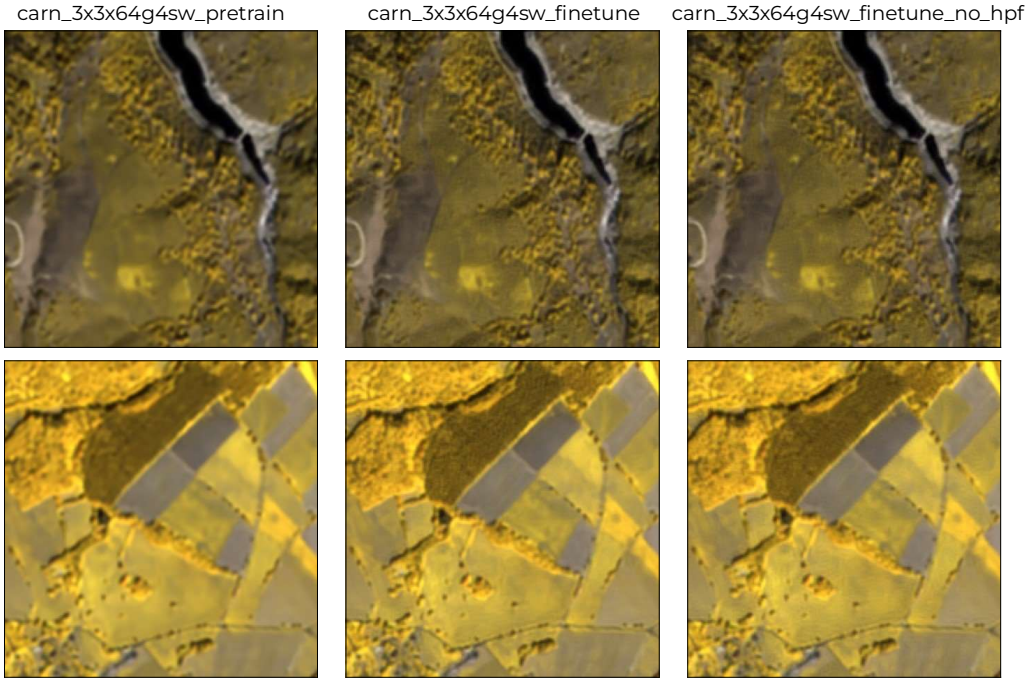


Figure 94. Red Edges bands (B7,B6,B5) rendering of predictions of fine-tuned vs. pre-trained models, form carn_3x3x64g4sw with and without HPF in discriminator, on real patches from the testing set.

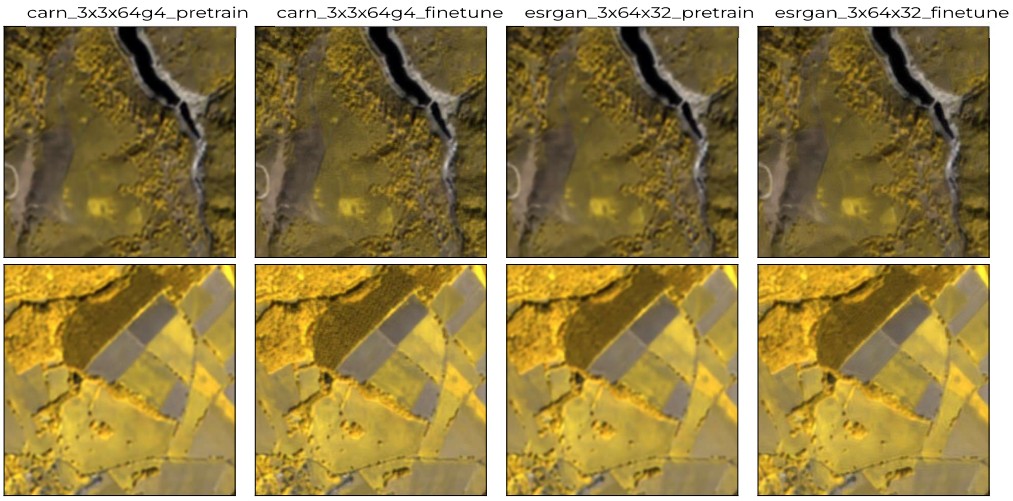


Figure 95. Red Edges bands (B7, B6, B5) rendering of predictions of fine-tuned vs. pre-trained models, form carn_4x3x64g4sw and esrgan_3x64x32, on real patches from the testing set.

With the reference model carn_3x3x64g4sw, Table 32 shows the order of magnitude of full inference time (including I/O) for different processing configurations.

Table 32. Order of magnitude of inference time of the carn_3x3x64g4sw model on full Sentinel-2 products for different processing configurations.

	CPU (1 core)	CPU (8 cores)	GPU (A100)
L1C	6 hours	1 hour	6 minutes
L2A	5 hours	50 minutes	5 minutes

3.2 Fusion of Satellite Image Time Series including Thermal Infrared

3.2.1 Introduction

Global coverage of the Earth surface with Satellite Image Time Series (SITS) is a critical component of the continuous monitoring of our planet constants, such as Essential Climate Variables (Bojinski et al., 2014) or Essential Biodiverisity Variables (Jetz et al., 2019). The Landsat series alone provides 50 years of such constant monitoring (Wulder et al., 2022), and for more than 10 years now Sentinel-2 has complimented the Landsat archive with higher revisit, higher spatial resolution images. SITS are used to monitor vegetation (Misra et al., 2020), derive Land Cover and Land Use maps (Phiri et al., 2020), control the European Common Agricultural Policy (Vajsová et al., 2020) or monitor Water Bodies (Bhangale et al., 2020).

Spatial and temporal resolutions of SITS are important parameters that define the reachable scale of analysis both in space and time. The revisit time is even more critical for passive optical sensors, where cloud occurrences obliterate a vast

proportion of the observations (Wilson & Jetz, 2016). On the other end, there is an increasing number of orbiting sensors that could complement each other to increase either the revisit time or the spatial resolution, using an adequate methodology. Ideally such methodology should be able to transform all observations from all sensors into an omniscient prediction: all wavelengths, any time, at best spatial resolution.

In this work, we consider the case of two or more contemporary sensors with global coverage and regular revisit. Both sensors have complementary revisit cycles and spectral bands, with different spatial resolution. There are a few combinations of sensors already flying, the most popular being Landsat 8/9 with Sentinel-2, because of their complementary revisit cycles (J. Li & Chen, 2020). Other already available combinations include PlanetScope and Sentinel-2 (Latte & Lejeune, 2020; Sadeh et al., 2021). In the future, new interesting mission will appear, such as the combination of Trishna (Lagouarde et al., 2018) and LSTM (Bernard et al., 2023) with Sentinel-2 or Landsat, as well as new up-coming missions such as Sentinel-2 NG or Landsat-Next. In 2020, the French space agency studied a High-Resolution complementary mission for Sentinel-2 called Sentinel-HR (Michel et al., 2022). There is therefore a growing need for methods that can jointly leverage those continuous sources of earth monitoring.

3.2.1.1 Existing work

Several sub-domains of the methodological literature in remote sensing partially address this challenge. They are summarized in the following figure. This includes Single and Multi-Image Super-Resolution, Band-Sharpening, Spatio-Temporal Fusion and Temporal Modeling. Those works usually focuses on machine learning methodological development within the constraints of existing datasets or domain definition.

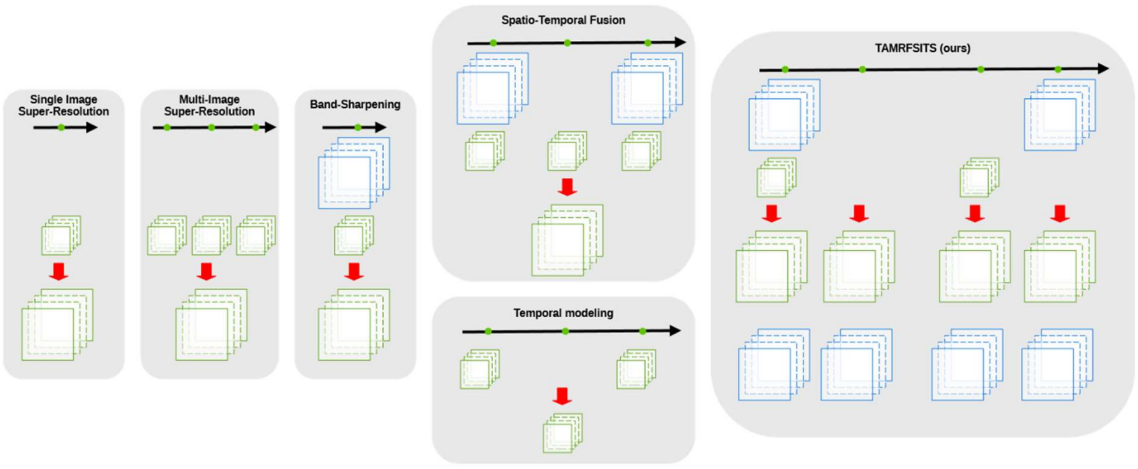


Figure 96. Overview of capabilities of methods from the literature, as compared to ours. The LR SITS is represented in green, while the HR SITS is represented in blue.

3.2.1.1.1 Single and Multi-Image Super-Resolution

Single or Multi-Images Super-Resolution (SISR or MISR) (Anwar et al., 2019; A. Liu et al., 2023) where images from higher resolution sensors are used to train a model that can then be used to improve the spatial resolution from one or several images acquired by a lower resolution sensor. SISR and MISR are inherently ill-posed, since there can be many High-Resolution explanation to the same Low-Resolution image(s). Therefore, successful SISR usually relies on adversarial training (Ledig et al., 2017; X. Wang et al., 2019) in to capture the target data manifold structure during training. Both SISR (Galar et al., 2020; Nguyen et al., 2023b; D. A. Pouliot et al., 2018; Salgueiro Romero et al., 2020) and MISR (Bordone Molini et al., 2019; Ibrahim et al., 2025; Märtens et al., 2019; Okabayashi et al., 2024) have been applied to remote-sensing imagery. Models are usually trained by using either simulated or cross-sensor dataset, each having their own strengths and weaknesses, which were investigated in a previous work (Michel et al., 2025). Though vastly investigated in the literature, SISR and MISR are not well suited for leveraging the synergies between two global coverage sensors of different spatial resolutions. Because the model itself only uses LR images, SISR does not integrate the prior HR information given by HR observations at other dates, which always exists in the case of global coverage missions. Its lack of temporal processing prevents it to overcome missing data in the LR input images. Finally, it requires HR reference data to have very similar spectral bands, and same-date acquisitions in case of cross-sensor dataset.

3.2.1.1.2 Band-Sharpening

Band-sharpening (Firozjaei et al., 2022), also known as pan-sharpening (Ciotola et al., 2025; Vivone et al., 2014), or reference-based Super-Resolution (Lutio et al., 2019; Su et al., 2025) depending on the scientific community, aims at using an auxiliary near-simultaneous High Resolution image (often from the same satellite) in order to improve the spatial resolution of an image acquired by a lower resolution sensor. Spatial sharpening can be seen as a dual method of SISR: instead of using a near-simultaneous HR image with similar channels as a reference to supervise the learning of the model, it uses a near-simultaneous HR image with different channels as an input to the model. While adding HR inputs certainly helps the model to improve the spatial resolution, there is no longer a HR reference to train the model against. To overcome this problem, researchers in various remote sensing fields have been using the scale invariance hypothesis, also known as the Wald protocol (Wald et al., 1997b), and train the model at degraded LR spatial resolutions, using the original LR images as reference (this process is well illustrated in Figure 97 (F. Palsson et al., 2018)). This practice is commonplace in thermal sharpening, a branch of spatial sharpening addressing the Thermal Infra-Red or Land Surface Temperature (LST) images (F. Gao et al., 2012; Granero-Belinchon et al., 2019), but is also used in other band-sharpening applications when sensor is equipped with different spatial resolution across bands (F. Palsson et al., 2018; Salgueiro et al., 2021). However, several studies show that the scale invariance hypothesis is weak and can lead to wrong estimation of local textures (Merlin et al., 2010; B. M. Nguyen et al., 2022). In thermal sharpening, most methods only learn

the mapping from degraded HR image and LR image and then use the learnt model at full HR resolution. This imposes to learn a new model for each pair of HR and LR mapping, as well as to resort to a mechanism called residual compensation to ensure that the LR dynamics are correct.

Band-sharpening seems more equipped that SISR and MISR to solve the problem at stake: it supports LR bands without HR reference, and in some cases might produce gap-less outputs, if several HR acquisitions close in time are used [Shao et al., 2019]. Yet it relies on the weak scale invariance hypothesis, is essentially mono temporal. It cannot extrapolate unseen date or learn the temporal dynamics.

3.2.1.1.3 Spatio-Temporal Fusion

Spatio-Temporal Fusion (STF) (Belgiu & Stein, 2019; Xiao et al., 2023) is a lineage of methods derived from the initial STARFM algorithm (F. Gao et al., 2006). STARFM is a model-based method for the fusion of Landsat and MODIS surface reflectances. To process the LR target date, it requires LR and HR simultaneous images both before and after this target date. Model-based improvement have followed (Belgiu & Stein, 2019; X. Zhu et al., 2018) most of which extend STARFM or try to overcome its limitations by using more complex models, without changing or extending this paired configuration. More recently, the model-based method has been traded for a parametric Deep-Learning (DL) model, including convolutional layers (X. Liu et al., 2019; Z. Tan et al., 2018, 2019, 2022), adversarial training (Y. Xie et al., 2024; Y. Zhang et al., 2024), transformers (W. Li et al., 2022; Y. Wu & Huang, 2024; G. Yang et al., 2022) and vision transformers (G. Chen et al., 2022). It must be stressed than though the DL models proposed in the literature became more complex, very few of them break free from the limitation of requiring one or two pairs of simultaneous LR and HR acquisitions, to the notable exception of (Goyena et al., 2023) and (X. Zhang et al., 2024). They also do not extend the model capability to full SITS with more than two dates. Additionally, several of these methods actually require the scale invariance hypothesis for training, and some of them also require retraining for each target date. It should also be stressed that there is often no temporal modelling between the pairs, since acquisition dates are not taken into account.

3.2.1.1.4 Temporal Modeling

In addition to these domains focusing on spatial resolution, there has been some effort to interpolate or forecast SITS (Moskolaï et al., 2021). One strong driver of such research is to mitigate the occurrence of clouds in passive optical data in order to generate gapless SITS (S. Li et al., 2021), and most methods handle SITS from a single sensor, with some notable exceptions (Irigireddy & Bandaru, 2025; Roy et al., 2008). In (H. Liu et al., 2024), a transfomer is used in the temporal dimension in order to reconstruct cloudy pixels in the Harmonized Landsat Sentinel2 (HLS) dataset (Claverie et al., 2018).

3.2.1.2 Contributions

We hypothesize that those different families of methods only address a sub-part of a more general problem: given the observed SITS from two or more sensors over the same area, predict any acquisition time at best observed spatial resolution all bands from all-sensors, without any further assumptions. Failing to address the big picture not only restricts the capabilities of those methods but also requires additional weak assumptions and limitations to solve the sub-part of the problem at stake. Capabilities, limitations and weak assumptions are summarized in the following Table 33.

In EvoLand, we have proposed a mathematical formulation of the general problem, a deep-learning architecture and a training procedure that allows to solve it by means of Self Supervised Learning (SSL). Our proposed model, called Time Attention Multi-Resolution Fusion of Satellite Image Time Series (TAMRFSITS) has full capabilities and none of the limitations and weak assumptions of existing families of method. Though the proposed model is not limited to a specific pair of sensors and can even handle more than two sensors, we demonstrate its properties through the fusion of Sentinel-2 and Landsat-8 time-series.

Table 33. Capabilities and limitations of the different categories of methods found in solving the general problem, (x) means that some methods in the literature partially achieve the capability or solve the limitation or weak assumption.

Category of methods	Temporal	SR	Sharpening	STF	Ours
Capability					
Improves spatial resolution		x	x	x	x
Supports bands without HR reference			x	x	x
Produces gapless outputs	x		(x)	(x)	x
Extrapolates unseen dates	x				x
Learns temporal dynamic	x				x
Limitation / assumption					
Similar spectral bands in sensors	x	x		x	
Same-day acquisitions		x	(x)	(x)	
Scale invariance hypothesis			x		
Limited to pairs or triplet			x	x	
Requires no-data masks as inputs	(x)			x	
Re-training for each target date			(x)	(x)	

3.2.2 Proposed method

3.2.2.1 Problem Formulation

For the sake of simplicity, the problem formulation will be written for two SITS, one with high spatial resolution R called S_{HR} and one with Low spatial resolution r called S_{LR}. In addition, we will assume that r is an integer multiple of R:

$$r = kR,\ k \in N\ ,$$

where k is the resolution factor between both SITS.

The spatial sampling grid of the LR SITS can be defined as:

$$R_{LR} = R(x_0, y_0, h, w, r) = x_0 + ir, i \in [0, w] \times y_0 + jr, j \in [0, h],$$

where x_0 and y_0 denotes the coordinates of the upper-left corner of the grid, h ∈ N and w ∈ N denotes the integer height and width of the image, and r is the already defined spatial resolution. Conversely, the spatial sampling grid of the HR SITS can be written $R_{HR} = R(x_0, y_0, H, W, R)$, with $H = kh$ and $W = kw$.

The acquisition times of the LR SITS can be defined as:

$$T_{LR} = t_n, n \in [0, N_{LR}],$$

where $N_{LR} \in$ N is the number of acquisition times in LR SITS. Acquisition time for HR SITS is likewise noted T_{HR}.

Finally, C_{LR} denotes the set of spectral bands found in LR SITS, C_{HR} the set of spectral bands found in HR SITS, without any further assumption on possible matching bands between C_{LR} and C_{HR}.

With those notations, both the HR and LR SITS can be seen as a collection of measurements sampled from a universal observation function $F(x, y, r, t, c)$ of spatial location (x, y), spatial resolution r, acquisition time t and spectral band c:

$$S_{LR} = \{F(x, y, r, t, c), x, y, t, c \in R_{LR} \times T_{LR} \times C_{LR}\}$$

$$S_{HR} = \{F(x, y, r, t, c), x, y, t, c \in R_{HR} \times T_{HR} \times C_{HR}\}$$

For the sake of computation as well as spatial and temporal inductive biases, collections of measurements SLR and SHR are often organized into tensors of shape $[w, h, N_{LR}, \#C_{LR}]$ (resp. $[W, H, N_{HR}, \#C_{HR}]$).

Let T_{query} denote another set of acquisition time, without any assumption whether T_{query} has non-null intersection with TLR or THR.

In our work, we aim at building a parametric function Φ as well as a training procedure to derive its optimal parameters $\Theta^\star$ so that Φ can infer measurements for all bands in $C_{HR} \cup C_{LR}$ at every query acquisition times T_{query} and sampled at high spatial resolution sampling grid defined by R_{HR}.

$$\Phi(S_{LR}, S_{HR}, T_{query}, \Theta^\star) = \{F(x, y, r, t, c), x, y, t, c \in R_{HR} \times T_{query} \times C_{LR} \cup C_{HR}\}$$

Note that the output of Φ can be further separated into SITS with C_{HR} and C_{LR} channels:

$$S^\star_{LR} = \{F(x, y, r, t, c), x, y, t, c \in R_{HR} \times T_{query} \times C_{LR}\}$$

$$S^\star_{HR} = \{F(x, y, r, t, c), x, y, t, c \in R_{HR} \times T_{query} \times C_{HR}\}$$

3.2.2.2 Architecture

Function Φ is implemented through a deep-learning architecture called Temporal Attention Multi-Resolution Fusion of Satellite Image Time Series (TAMRFSITS), leveraging residual Convolutional Neural Network (CNN) in the spatial dimensions as well as Transformer in the temporal dimension. The overall architecture follows a classical encoder - decoder scheme, as presented in the figure below. Both S_{LR} and S_{HR} are fed into the TAMRFSITS encoder which outputs a latent representation S_{latent} comprising D features, with shape $[W, H, \#T_{LR} + \#T_{HR}, D]$. It is important to note that unlike many works in auto-encoders, the time dimension is not projected to a fixed size and retains the initial number of observations $\#T_\{LR\} + \#T_\{HR\}$. To obtain predictions for T_{query} acquisition times, both T_{query} and latent representation S_{latent} are fed into the TAMRFSITS decoder, which outputs $S^{\star}_{LR}$ and $S^{\star}_{HR}$. In the following sections, the architecture of the encoder and the decoder will be detailed. Intuitively, TAMRFSITS can be seen as the combination of a Single Image Super-Resolution network in the spatial dimension and a Transformer in the temporal dimension.

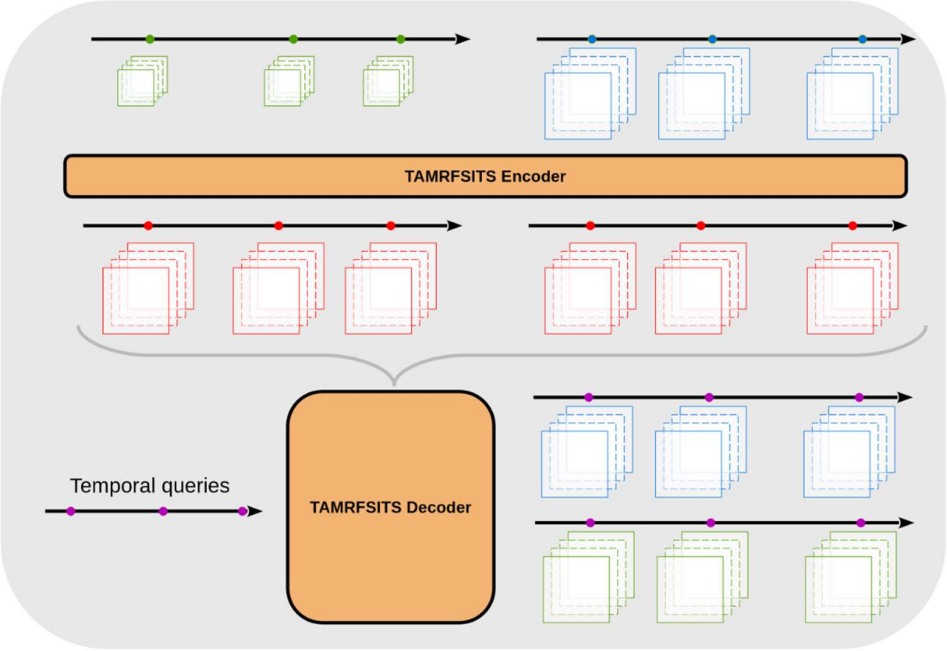


Figure 97. Overview of the proposed method.S_{LR} (in green) and (in blue) are fed into the TAMRFSITS encoder, which outputs latent representation S_{latent} (in red). The latent representation is fed into the TAMRFSITS decoder, along with T_{query}, which outputs $S^{\star}_{LR}$ (in green) and $S^{\star}_{HR}$ (in red).

3.2.2.2.1 Encoder

The workflow of the TAMRFSITS encoder is detailed in the following figure and is composed of two sequential stages. First, each acquisition date of S_{LR} and S_{HR} is processed separately by a date-wise spatial encoder. Note that a separate date-wise spatial encoder is used for the S_{LR} and S_{HR} dates. The date-wise spatial encoder outputs F features and extends the spatial dimensions to W and H for the S_{LR} SITS,

yielding S_{LR}^{latent} of shape $[W, H, \#T_{LR}, F]$ and S_{HR}^{latent} of shape $[W, H, \#T_{HR}, F]$. Note that all dimensions match except for the temporal dimension. During the second stage, the feature dimension is enriched with temporal positional encoding as well as sensor encoding, forming a feature space of size D. Then each pixel-wise sequence of $\#T_\{LR\} + \#T_\{HR\}$ tokens of D feature is processed independently by a transformer encoder architecture. Thus, the TAMRFSITS encoder follows the spatial then temporal encoding paradigm also found in UBARN (Dumeur et al., 2024b) or SITS-Former (Yuan et al., 2022).

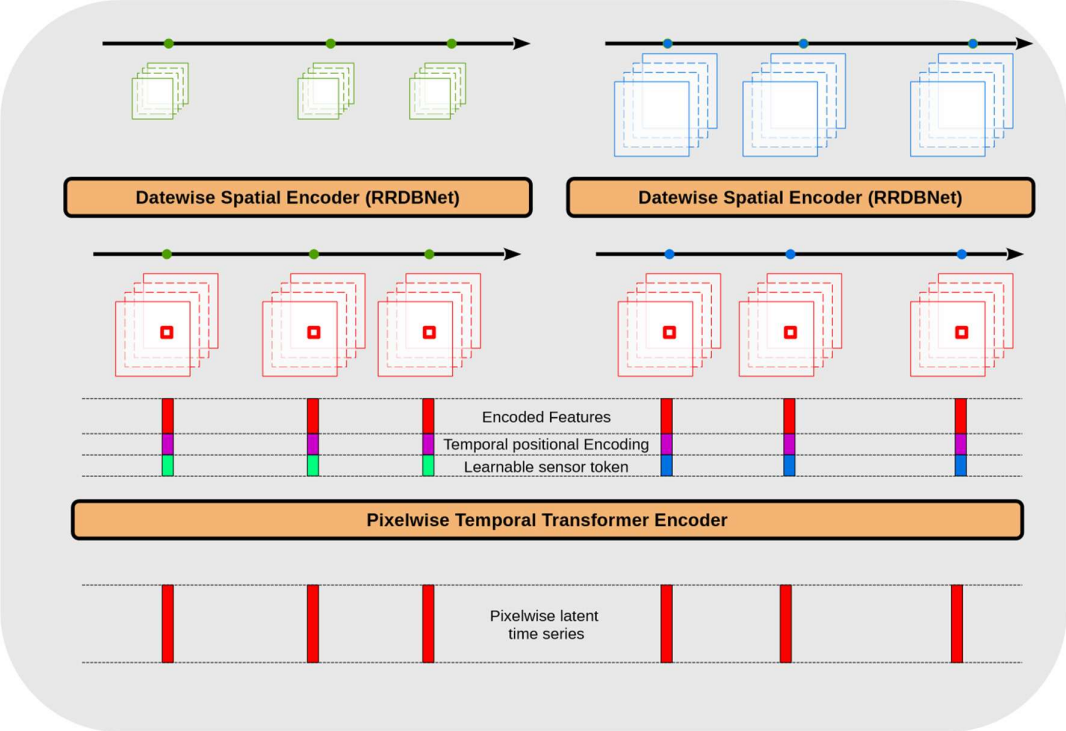


Figure 98. Details of TAMRFSITS encoder. The input SITS S_{LR} (in green) and (in blue) are first passed through separate date-wise spatial encoders based on the RRDB architecture, including spatial up-sampling of S_{LR}. The resulting series S_{LR}^{latent} and S_{HR}^{latent} have matching dimensions, except for the temporal dimension. After being informed by temporal positional encoding and a learnable sensor token, each pixel is converted to a sequence of tokens imputed to a pixel-wise transformer encoder, forming the latent SITS S_{latent}.

Spatial encoder

The date-wise spatial encoder follows the SRResNet architecture introduced in (Ledig et al., 2017) and notably used in ESRGAN (X. Wang et al., 2019). It consists in chaining several Residual in Residual Dense Block (RRDB), followed by an up-sampling operation which can be achieved either by means of Pixel-Shuffle (PS) or spatial interpolation. In the TAMRFSITS date-wise spatial encoder, the up-sampling stage is optional and not applied for SHR. The convolution layers after the up-sampling operation are also deferred to the date-wise spatial decoder.

Temporal and modality encoder

For a pixel corresponding to a spatial location, the F features extracted by the datewise spatial encoder for each date of SHR and SLR form a sequence of measurements. However, in order to allow the downstream temporal encoder to reason with sensors and acquisition date, this sequence of measurements needs to be enriched with context. The traditional positional encoding used by transformers in Natural Language Processing [Vaswani, 2017] has been extended to temporal positional encoding in many works on SITS, including (Dumeur et al., 2024b; Guo et al., 2024; Yuan et al., 2022), as alternating sine and cosine functions with varying frequencies:

$$\sin\left(t \,/\, \lambda\text{^}\{2i \,/\, L\}\right) if\ l\ is\ even,\ l \in [0, L]\ ,$$

$$\cos\left(t \,/\, \lambda\text{^}\{2i \,/\, L\}\right) if\ l\ is\ even,\ l \in [0, L]\ ,$$

where L is the number of temporal positional features and λ is a normalization factor. The L features of the temporal positional encoding are concatenated in the feature dimension with the F features produced by the date-wise spatial encoder. To inform the downstream transformer model, two dedicated learnable sensor token of size M, one for S_{HR} and one for S_{LR} is also concatenated to the feature dimension. The resulting feature dimension is therefore D = F +L+M. A pixel (e.g. spatial location) is therefore described by a sequence of $\#T_\{LR\} + \#T_\{HR\}$ tokens of D features. This pixel-wise sequence of token is then passed to a transformer encoder module as described by [Vaswani, 2017]: the sequence is passed through several blocks of multi-head self-attention followed by a feed-forward network. The resulting pixel-wise sequences are rearranged into spatial dimensions in order to form the S_{latent} SITS.

3.2.2.2.2 Decoder

The TAMRFSITS decoder operates on the latent SITS S_{latent} from the encoder as well as on the acquisition time queries T_{query}. As shown in following figure, it follows a scheme that is symmetrical to the encoder, and comprises a pixel-wise temporal decoder, followed by a datewise spatial decoder.

Temporal decoder

The first step of the temporal decoder is to transform the acquisition time queries Tquery into a sequence of #Tquery tokens of D features. A temporal positional encoding of size L is computed and concatenated to a learnable placeholder of size F + M. The query sequence and the pixel-wise sequence of Slatent are then fed to a modified transformer decoder instance: the first self attention block is removed, the Slatent sequence serves as keys and values for the multi-head cross-attention block, while the acquisition time query sequence serves as queries. The resulting output pixel-wise sequence can be seen as Slatent interpolated at the acquisition times in Tquery.

Spatial decoder

Following this pixel-wise temporal decoding, the resulting sequences are rearranged into spatial dimensions and inputted to the date-wise spatial decoder. This decoder is fairly simple and consist of two consecutive convolutional layers, with activation in between, that follows the up-sampling in the ESRGAN architecture (X. Wang, et al., 2018a). The spatial decoder outputs $\#\mathrm{CLR} + \#\mathrm{CHR}$ channels, which can then be split into two predicted SITS $\mathrm{SLR}\star$ and S^*_{HR}.

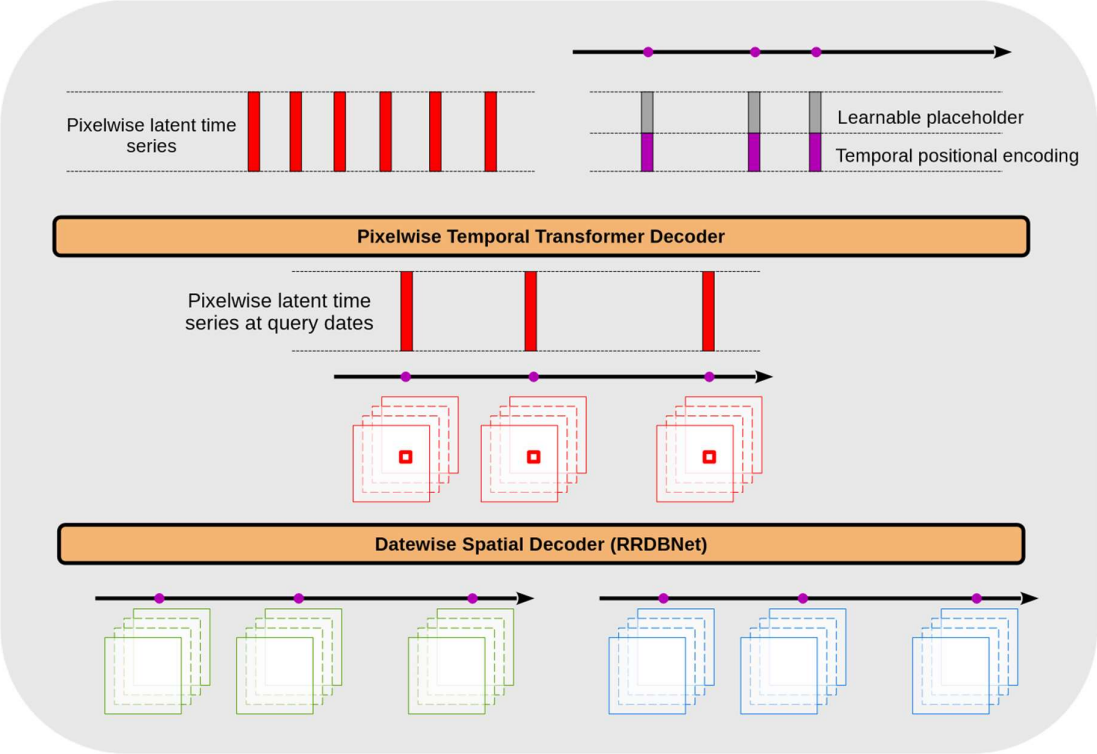


Figure 99. Details of TAMRFSITS decoder. The query acquisition times T_{query} are transformed into a sequence of tokens by means of temporal positional encoding and a learnable placeholder. The query sequence and the latent SITS S_{latent} are processed by a pixel-wise temporal decoder, starting with a cross-attention block, which interpolates S_{latent} at query acquisition times. The resulting SITS is then passed through a simple spatial decoder to predict the resulting SITS S^*_{LR} and S^*_{HR}.

3.2.2.3 Training strategy

The proposed problem formulation can be functionally implemented by the proposed architecture, but finding the optimal parameters $\Theta\star$ requires a training strategy that pushes toward the desired properties of S^*_{HR} and S^*_{LR} .

The first of these properties is that the predicted acquisition date should closely match the actual observations that could have been made on that date. This can be achieved through a SSL strategy, and more specifically a Masked Auto-Encoder (MAE) strategy (He et al., 2022; H. Liu et al., 2023; Reed et al., 2022), described in 2.3.1. The reconstruction loss term of the MAE needs to be adapted to the difference in spatial resolution of the input SITS as described in the following section.

However, this strategy by itself will not favor the migration of high frequency details from higher resolution bands to the prediction of low-resolution bands passed their initial resolution. In other words, even if all predictions will have the HR sampling grid RHR, bands from SLR that were initially of lower resolution will remain at this resolution and thus appear as blurred in the prediction. Therefore, we introduce a dedicated loss term that aims at favoring High Resolution details in all predictions, which is detailed in next section.

3.2.2.3.1 Masking Strategy

As opposed to (H. Liu et al., 2024) which adopts a purely random masking strategy, the masking strategy for TAMRFSITS is randomly drawn from a pool of strategies described in

Table 34. This allows to favor configurations that are unlikely to happen regularly with random masking, and which are of interest for downstream applications, such as completely missing the HR or LR SITS, or long-term forecasting. For each input SITS, a different strategy is drawn, with random parameters if this strategy has parameters. This maximizes input variability and help the network to better generalize.

The network is trained to reconstruct both masked and unmasked dates in S_{HR} and S_{LR}, and thus during training $T_{query} = T_{HR} \cup T_{LR}$. Reconstructing unmasked dates during training is necessary to ensure that the network will behave correctly on those dates during inference.

It should be reminded that inferring unmasked date is of interest with TAMRFSITS, since the model will provide a) higher resolution version of unmasked S_{LR} dates and b) cloud-free predictions of all unmasked inputs. In the following sections, the SITS with the masked dates will be called S_{LR}^{masked} and S_{HR}^{masked} respectively.

The total loss used to optimize TAMRFSITS is given by:

$$L = \alpha_{masked} w_{LR}^{masked} \left(L_{LR}^{masked} + \kappa_{LR}^{masked}\right) + \alpha_{clear} w_{LR}^{clear} \left(L_{LR}^{clear} + \kappa_{LR}^{clear}\right)$$
$$+\alpha_{masked} w_{HR}^{masked} \left(L_{HR}^{masked} + \kappa_{HR}^{masked}\right) + \alpha_{clear} w_{HR}^{clear} \left(L_{HR}^{clear} + \kappa_{HR}^{clear}\right)$$
$$+\alpha_{spat} w_{HR}^{clear} L_{spat},$$

where L_{HR}^{clear}, L_{HR}^{masked}, L_{LR}^{clear} and L_{LR}^{masked} are reconstruction terms, κ_{HR}^{clear}, κ_{HR}^{masked}, κ_{LR}^{clear} and κ_{LR}^{masked} are mask contrastive terms, L_{spat} is spatial reconstruction loss term, α_{masked}, α_{clear} and α_{spat} are constants weights designed to balance the different loss terms, and w_{LR}^{masked} (resp. w_{HR}^{masked}, w_{LR}^{clear}, w_{HR}^{clear}) is the ratio of masked LR dates wrt. the total number of LR and HR dates in the SITS.

The next sections will detail each of those terms.

Table 34. Pool of strategies for MAE training.

Strategy	Description	Parameter

Random	Randomly discard dates from and	Discard rate
Gaps	Generate periods with where no or are available	Length of gaps
No HR	Discard entirely	N/A
No LR	Discard entirely	N/A
Forecast	Discard all dates in and after a tipping date	Tipping date
Backcast	Discard all dates in and before a tipping date	Tipping date

3.2.2.3.2 Reconstruction term

The reconstruction loss terms make use of the Huber loss, introduced in (Girshick, 2015a) given by:

$$L_1^{smo} \quad (x) = 0.5x^2 \; if \;\; | \, x \, | < 1, \;\; | \, x \, | - 0.5 \; otherwise$$

The L_1^{smo} loss behaves like the L1 loss when differences between predicted and target values are large, which limits the impact of outliers, and like the L2 when differences are small, which is better for optimization. In the experiments, it has proven to be beneficial over the standard MSE (L2), especially in early stages of training.

Two loss terms are defined for each of S_{HR} and S_{LR}. Loss terms for S_{HR} are given by:

$$L_{HR}^{clear} = \frac{1}{\sum_{R_{HR} \times T_{HR} \times C_{HR}} V_{HR}} \sum_{R_{HR} \times T_{HR} \times C_{HR}} L_1^{smooth}(S_{HR}^{\star} - S_{HR}) V_{HR},$$

$$L_{HR}^{masked} = \frac{1}{\sum_{R_{HR} \times T_{HR}^{masked} \times C_{HR}} V_{HR}} \sum_{R_{HR} \times T_{HR}^{masked} \times C_{HR}} L_1^{smooth}(S_{HR}^{\star} - S_{HR}^{masked}) V_{HR}^{masked},$$

where integration is performed over all (x,y,t,c) ∈ R_{HR} × T_{HR} × C_{HR} (resp. R_{HR} × T_{HR}^{masked} × C_{HR}) and V_{HR} (resp. V_{HR}^{masked}) is a binary validity mask that is derived from the metadata of the products and accounts for clouds, shadows, out-of-swath areas …. Note that this mask is only use for the self-supervision and not as input to the network. Using such masks is key to learning quality predictions by avoiding setting cloudy areas as target for the $L^{smooth}{}_1$ loss.

Because $S_{LR}^{\star}$ is of spatial resolution R while S_{LR} and S_{LR}^{masked} are of coarser spatial resolution r, one cannot directly apply the same formulation for the S_{LR} loss terms:

$$L_{LR}^{clear} = \frac{1}{\sum_{R_{LR} \times T_{LR} \times C_{LR}} V_{LR}} \sum_{R_{LR} \times T_{LR} \times C_{LR}} L_1^{smooth}((S_{LR}^{\star} * \omega_{\sigma}) \downarrow_k - S_{LR}) V_{LR},$$

$$L_{LR}^{masked} = \frac{1}{\sum_{R_{LR} \times T_{LR}^{masked} \times C_{LR}} V_{LR}} \sum_{R_{LR} \times T_{LR}^{masked} \times C_{LR}} L_1^{smooth}((S_{LR}^{\star} * \omega_{\sigma}) \downarrow_k - S_{LR}^{masked}) V_{LR}^{masked},$$

where $\omega\sigma$ is a gaussian spatial kernel designed for smoothing the images in $S_{LR}^{\star}$ and $\downarrow k$ is a downsampling operator of factor k. Combining those operators allow to bring high resolution prediction $S_{LR}^{\star}$ back to its original low-resolution r.

3.2.2.3.3 Mask contrastive term

Though not explicitly enforced, the use of validity masks V_{HR}^{masked} and V_{LR}^{masked} to mask the reconstruction loss term should theoretically help the temporal transformer to learn that input invalid pixels such as cloud, cloud shadow or missing data are uninformative for the temporal reconstruction and should not be use. However, those masks from Level 2 products have commission and omission errors which might prevent to learn to ignore those pixels, especially since the transformer is also asked to reconstruct clear dates, for which the temporal incentive of having the exact same date as input is very strong. To avoid cloud or other masked invalid pixel leakage into the prediction, we introduce a mask contrastive term. This contrastive term enforces that the prediction of a masked reference pixel should be closer to the closest same pixel in time than to the invalid masked reference pixel, which can be formulated as a triplet margin loss (Balntas et al., 2016):

$$\kappa_{LR} = \max\{L_2(S_{LR}^{\star}, S_{LR}^{valid}) - L_2(S_{LR}^{\star}, S_{LR}^{invalid}) + m \arg i\, n, 0\}$$

where L2 is the Mean Squared Error, $S_{LR}^{invalid}$ represent invalid pixels of S_{LR}, S_{LR}^{valid} represents the closest valid pixel in time, and margin is a user defined threshold. κ_{HR} can be defined likewise.

In practical, a threshold on the time distance to the closest valid pixel is used, so that valid pixel that are too further away in time are not used in the computation of the contrastive loss term.

3.2.2.3.4 Spatial term

As explained in the previous section, the $L^{clear}{}_{LR}$ and $L^{masked}{}_{LR}$ reconstruction loss terms operate at initial LR resolution. As such, they cannot drive $S_{LR}^{\star}$ toward high spatial resolution details, even if $S_{LR}^{\star}$ has the spatial sampling of R_{HR}. The usual solution to this problem is to apply the Wald protocol (Wald et al., 1997b) , which consists in downscaling all the data of the target factor resolution k, train the model to perform r/k → r resolution improvement, and then apply the trained model to perform r → kr resolution improvement by assuming scale invariance. Scale invariance of resolution improvement does probably not hold, and in the TAMRFSITS global formulation of the problem the SHR SITS can be used in order to favor high spatial frequencies in all predicted bands.

Both C_{LR} and C_{HR} bands are predicted for each prediction time in T_{query}, and during training $T_{query} = T_{HR} \cup T_{LR}$. Therefore, for each acquisition in T_{HR} (either masked or clear in the MAE strategy), predicted bands in $S_{LR}^{\star}$ can be compared to reference HR images in S_{HR}. However, no assumptions have been made regarding corresponding

bands in C_{LR} and C_{HR}. Therefore, S_{HR} cannot be directly used to supervise the spatial reconstruction of $S_{LR}^{\star}$.

To overcome this issue, a linear regression with ridge regularization is performed separately for each date in T_{HR} to form an estimate of the C_{LR} bands in $S_{LR}^{\star}$ from the C_{HR} bands in S_{HR}. This operation is denoted $\Lambda(S_{LR}^{\star}, S_{HR})$. Performing this operation separately for each date allow to compensate for a) the relatively weak estimate provided by linear regression, especially for bands in C_{LR} that do not have a close match in C_{HR} b) bands C_{LR} that depend on external factors that are not measured by bands in C_{HR}. The latter case is encountered with the thermal band of Landsat-8, whose primary driver is meteorological.

Despite the date-wise estimation of the linear regression, $\Lambda(S_{LR}^{\star}, S_{HR})$ is still a weak estimate of SLR. Some local trends might not be correctly accounted for and using $\Lambda(S_{LR}^{\star}, S_{HR})$ for the direct supervision of $S_{LR}^{\star}$ might lead to radiometric distortion of the latter. Since radiometric accuracy is enforced by the reconstruction loss term, only the high spatial frequency content of $\Lambda(S_{LR}^{\star}, S_{HR})$ should supervise the high spatial frequency content of $S_{LR}^{\star}$. This is achieved using the Learned Perceptual Image Patch Similarity (LPIPS) (R. Zhang et al., 2018). Indeed, in a previous work (Michel et al., 2025), we demonstrated that LPIPS is especially well suited for the measurement of difference in spatial frequencies content and has interesting properties such as being relatively blind to radiometric distortion and spatial misalignment. The complete spatial

reconstruction loss term is given by:

$$L_{spat} = LPIPS\left(S_{LR}^{\star}, \Lambda(S_{LR}^{\star}, S_{HR})\right)$$

3.2.3 Experiments

3.2.3.1 Dataset

The LS2S2 dataset comprises joint Sentinel-2 and Landsat-8 and Landsat-9 SITS over year 2022. It is composed of 64 Areas Of Interest (AOIs) for the training set and 41 AOIs for the testing set, each covering an area of 9.9x9.9 km2 (990x990 squared pixels for Sentinel-2 and 330x330 squared pixels for Landsat) with the geographical coverage given by Figure 100. Note that the geographical coverage corresponds to EvoLand training and testing MGRS tiles. All dates with more than 25% of clear pixels over the AOI are included in the dataset. This yields a total of 2 984 Sentinel-2 images and 1 659 Landsat-8 and 9 images in the training set. Additional statistics about the number of images are presented in Table 35.

Table 35. Number of 9.9x9.9 km2 images in LS2S2 dataset for each sensor and each split.

	Sentinel2				Landsat			
	Total	**Average**	**Min**	**Max**	**Total**	**Average**	**Min**	**Max**
train	2 984	44	11	94	1 659	26	1	62
test	1 609	39	6	98	762	19	1	56

For each sensor, Top-of-Canopy surface reflectance from level 2 products are used. The spectral bands included in the dataset are presented in Table 36. It can be observed that the Landsat sensor does not have Red Edge bands or wide Infra-Red, and conversely Sentinel-2 sensor does not retrieve Land Surface Temperature (LST). In addition to the spectral bands, corresponding quality masks have been used to derive a validity mask for each date of each sensor. This dataset has been gathered through the very convenient OpenEO API (Schramm et al., 2021). Figure 101 gives an example of SITS extracted from the LS2S2 dataset.

Table 36. Spectral bands included in the LS2S2 dataset for each sensor. The table layout allows to highlight which bands are similar between sensors and which bands are only available on one sensor. Despite the variable spatial resolution across bands for both Landsat and Sentinel-2, for each sensor all bands are up-sampled to the maximum resolution (10 meter for Sentinel-2 and 30-meter for Landsat).

Sentinel-2		Landsat		Description
Band	**Resolution**	**Band**	**Resolution**	
		B1	30.	Deep blue
B02	10.	B2	30.	Blue
B03	10.	B3	30.	Red
B04	10.	B4	30.	Green
B05, B06, B07	20.			Red Edge
B08	20.	B5	30.	Near Infra-Red
B8a	10.			Wide Hear Infra-Red
B11, B12	20.	B6, B7		Short Wavelength Infra-Red
		B10	90.	Land Surface Temperature

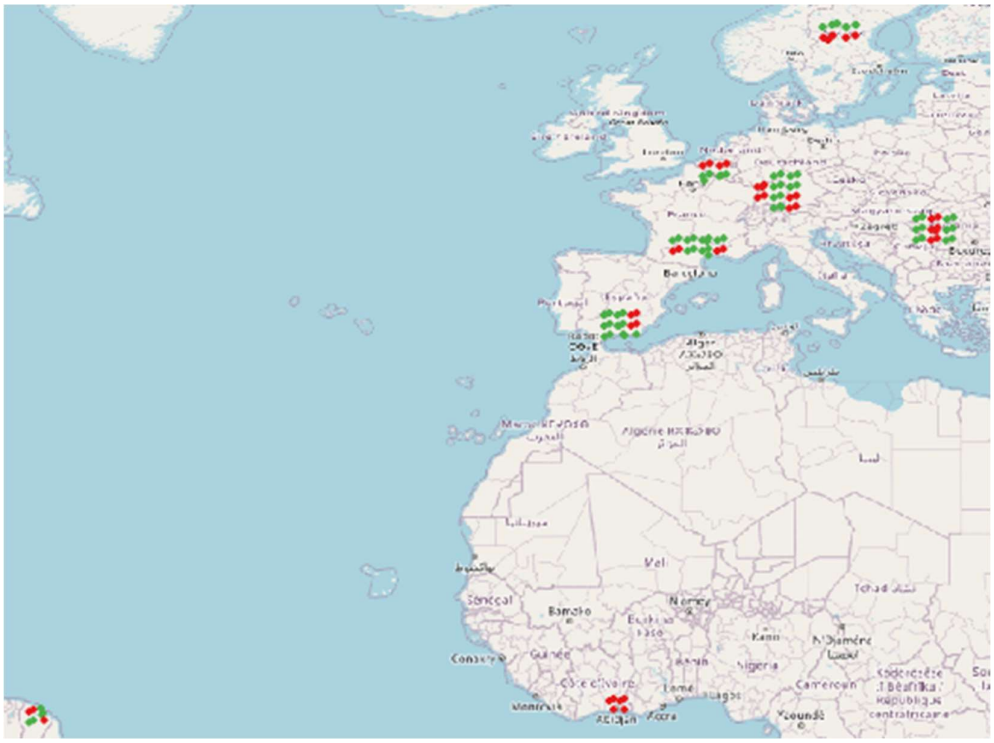


Figure 100. Geographical coverage of the LS2S2 dataset, with training AOIs in green and testing AOIs in red.

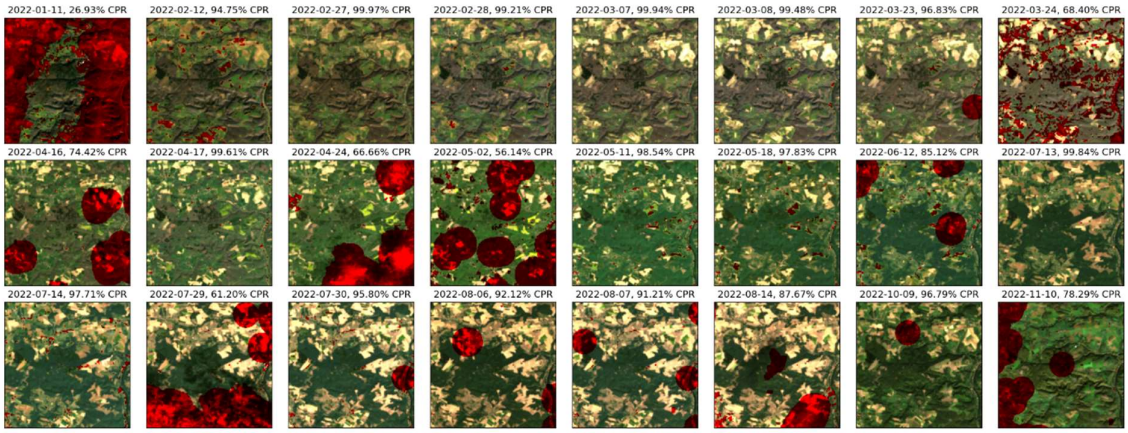


Figure 101. Landsat 8 and 9 time series.

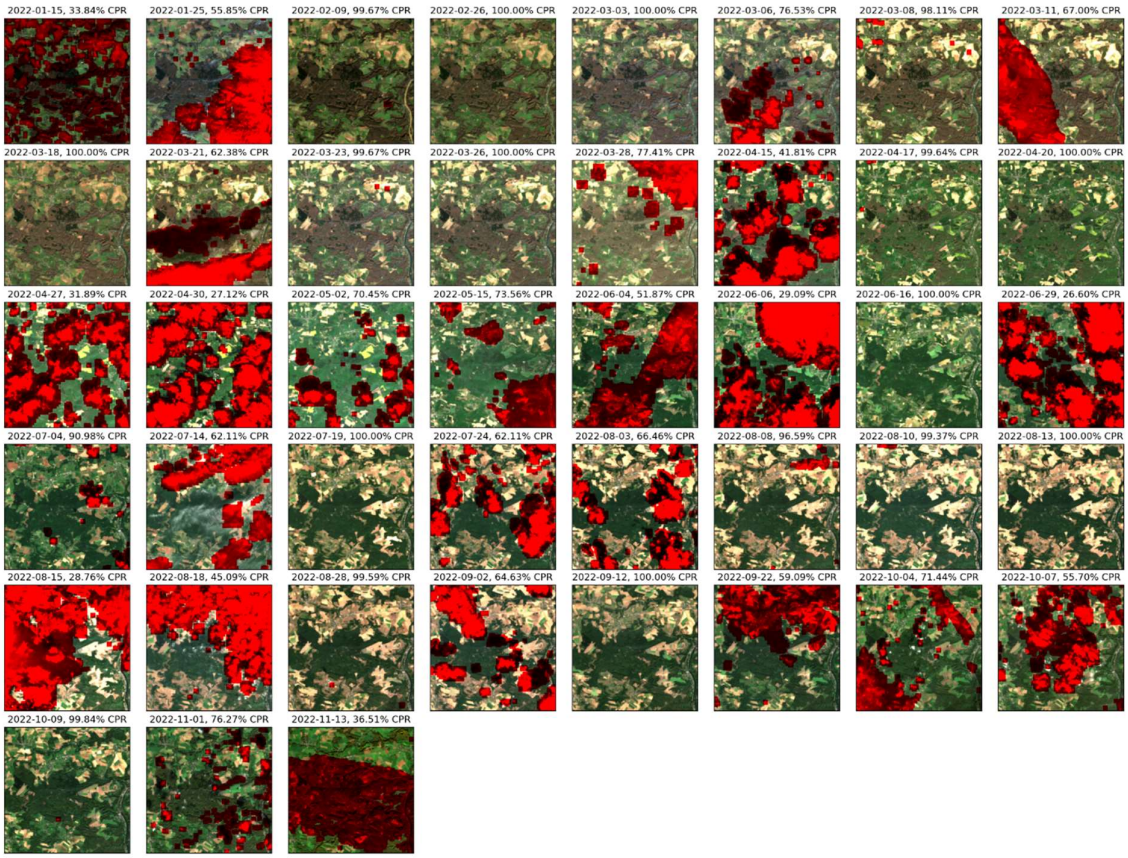


Figure 102. Example of SITS from LS2S2 dataset training slice, for UTM tile 31UFR. Pixels masked according to the quality mask included in the product are highlighted in red.

3.2.3.2 Experimental setup

3.2.3.2.1 Training hyper-parameters

The LS2S2 training split is sub-divided into 2052 training patches and 108 validation patches, each of 165x165 pixels for Sentinel-2 SITS and 55x55 pixel for Landsat SITS. Each patch contains the full time-series. TAMRFSITS is trained with the Adam optimizer (Kingma & Ba, 2014b), with one pair of Sentinel-2 and Landsat SITS passed at each step, which can be seen as using a batch size of one. In order to avoid memory errors, if the total number of images for a given step is higher than 50, exceeding images of both sensors are randomly dropped. TAMRFISTS is trained for 500 epochs, for a total of 1 026 000 steps. The best checkpoint according to the loss measured on the validation patches at the end of each epoch is used as the final model. The initial learning rate is set to 1e-4. After a linear warm-up of one epoch, the cosine annealing with warm restarts (Loshchilov & Hutter, 2016a) is used

to modulate the learning rate, with an initial period of one epoch and a multiplicative factor of 2. All codes use Pytorch (Paszke, Gross, Massa, Lerer, Bradbury, Chanan, et al., 2019) and runs on NDVIA GPU A100 and H100.

3.2.3.2.2 Metrics

Following our previous findings (Michel et al., 2025), the following metrics will be used to assess the model performances. We choose to retain only three qualified metrics according to this work, each with a distinctive purpose:

- RMSE measures the faithfulness to the radiometry of the target image at initial resolution,
- BRISQUE is a No-Reference Image Quality metrics that grades the global Image Quality,
- Frequency Restoration (FR) compares the spatial frequency content of the prediction and the target image at initial resolution.

All metrics are measured separately for each band. They are also computed on masked dates, e.g. not available as model inputs, and clear dates, e.g. used as model input, separately.

RMSE

RMSE is computed with respect to the initial resolution of the target data. Since $S_{HR}^{\star}$ has the same resolution R as S_{HR}, $RMSE_{HR}$ has a simple formulation:

$$RMSE_{HR} = L_2(S_{HR}^{\star} - S_{HR}).$$

On the other hand, $S_{LR}^{\star}$ as spatial resolution R while S_{LR} has spatial resolution r = kR.

In order to be able to compute RMSELR, SLR⋆ is first smoothed by a gaussian kernel and downsampled, in a similar way to loss term:

$$RMSE_{LR} = L_2((S_{LR}^{\star} * \omega_{\sigma}) \downarrow_k - S_{LR}).$$

Since RMSELR is computed at lower resolution r it does not measure the spatial resolution improvement of SLR⋆ . The following two metrics are designed for this purpose.

BRISQUE Score

BRISQUE (Mittal et al., 2012) is a No Reference Image Quality (IQ) which builds local features mapped to IQ score by a Support Vector Regressor. It is trained on human annotated scores, yielding a score between 0 and 100, 0 being the best IQ. According to (Michel et al., 2025), BRISQUE provides a solid criterion to measure the sharpness of the image as well as its overall quality, even if it tends to favour a bit of noise. Using a No Reference IQ metric also allows to provide an insight on the performances that is not biased by the quality of the reference data, such as unmasked clouds, which are present in the dataset described in 3.2.3.1.

Frequency Restoration

Frequency Domain Analysis is proposed in (Michel et al., 2025) in order to measure the improvement of the spatial frequency content of super-resolved images. It consists in analysing the Frequency Attenuation Profile (F_{AP}) for bandwidth [$_{fm,fM}$].

$$U_{f_m,f_M} = \{(u,v): f_m \leq \sqrt{u^2+v^2} < f_M\}$$

Let denote the set of discrete spatial frequencies ($_{u,v}$) that lies within a ring defined by $_{fm}$ and $_{fM}$ in Fourier plane, F_{AP} is given by:

$$\mathcal{F}_{\mathcal{AP}}[P](f_m,f_M) = \frac{1}{\#U_{f_m,f_M}} \sum_{(u,v)\in U_{f_m,f_M}} | \mathcal{F}[P](u,v) |$$

where $\#U_{fm,fM}$ is the number of elements in $U_{fm,fM}$, and $F[P](u,v)$ is the DFT of image $_P$.

$F_{AP}[_P](_{fm,fM})$ is successively computed over a set of non-overlapping bandwidth intervals as given by the DFT quantization. For the sake of simplicity, the resulting set of values is denoted $F_{AP}[_P](_{fn})$, $_n \in [0,_N]$ in the following, with $_{fn}$ the central frequency of each frequency intervals. It is then averaged across batches and dataset. Finally, the normalized logarithmic $_{FAP}$ is computed, which gives spatial frequency attenuation in dB:

$$FAPlog[P](fn) = 10 \cdot \log10\ FAP[P](fn) - \log10\ FAP[P](f0).$$

Since we do not have High Resolution reference images for $_{SLR}$ and conversely Low-Resolution input images for $_{SHR}$, we adopt the No-Reference Actual Frequency Restoration metric (Michel et al., 2025), which will be abbreviated Frequency Restoration (FR) in the remaining of the document:

$$FRLR = XFAP^{log}(S^{\star}_{LR}) - F^{log}_{AP}((S_{LR})\uparrow_k),$$

$$FRHR = XFAP^{log}(S^{\star}_{HR}) - F^{log}_{AP}(S_{HR}),$$

where ($_{SLR}$) $\uparrow_k$ is the bicubic up-sampling of factor $_k$ of $_{SLR}$. FR is expressed in decibels.

3.2.3.3 Performances assessment

3.2.3.3.1 Random masking

In this first experiment half of the LR dates and half of the HR dates are masked for each SITS of the testing set. Masked dates are removed from the inputs of the network but are used as additional dates for $_{Tquery}$. This experiment allows to measure how well TAMRFSITS performs in reconstructing missing dates, as well as reconstructing input dates, in a setup where both input SITS are evenly spread across the studied time frame. Figure7 shows sample prediction using the trained model on AOI 31TFJ of the testing set, with the 50% random masking strategy. One can observe that the model always output high resolution 10-meter predictions for all bands of both the Sentinel-2 and Landsat sensor, in any of the following case :

observed Sentinel-2 or Landsat image, observed but cloudy Sentinel-2 and Landsat image, and unobserved date.

Table 37 and Table 38 gives an overview of the metrics for the whole testing for both Landsat and Sentinel-2 SISTS. As expected, RMSE is lower for observed input dates with respect to unobserved input date for both sensors. RMSE follows the same trends for Landsat and Sentinel-2, with RMSE for observed dates below 0.01 for visible bands and around 0.01 for NIR and SWIR bands. For Landsat SITS, the FR is around 5 dB for observed and unobserved dates, but the BRISQUE score is lower for the unobserved dates, which suggests that integrating the Landsat observed date into the prediction causes a slight diminution of the image quality, while preserving the 10-meter resolution details. On the contrary, BRISQUE score is slightly lower for observed dates in the Sentinel-2 SITS. For Sentinel-2 dates, FR is null for 10-meter input bands and is around 4dB for all 20 meter bands. There are no significant differences in terms of BRISQUE score, suggesting equal Image Quality across predicted bands for Sentinel-2. Finally, LST of Landsat is only evaluated on observed date. The median FR gain is around 10dB, suggesting a strong restoration of 10-meter details. However, if median RMSE is below 3 K, the 95% quantile is higher than 20 K, which suggest that even for observed Landsat date, the low-resolution dynamic of LST is not always captured by the TAMRFSITS model.

A better understanding of the trends in RMSE can be observed in Figure 104, which shows a scatter plot of the average RMSE of all observed and unobserved dates for both Sentinel2 10-meter bands, Landsat 30-meter bands and Landsat LST (90 meter), sorted by AOIs from the test set. First, it can be observed that some AOIs, corresponding to the equatorial area, have very few dates. For both Sentinel-2 10 meters and Landsat 30 meters, there is a consistent distribution of the RMSE across AOIs, with observed dates consistently better than unobserved dates, and some outliers that can probably be related to either cloud mask faults leading to measuring performances against cloudy pixels, dates with snow which is not always predicted correctly by TAMRFSITS. For Landsat LST on the other end, some AOIs behave very differently from the other, with very high RMSE values for all dates, which explains the high 95% RMSE quantile.

Table 37. Landsat prediction metrics for TAMRFISTS when applying a random masking strategy of 50% of the dates, yielding 365 clear Landsat dates and 389 masked Landsat dates. RMSE is expressed in surface reflectances except for LST, where it is expressed in Kelvins. The 50% column shows median value across all dates, while the 95% shows the 95% quantile across all dates.

	RMSE (lower is better)				BRISQUE (lower is better)				FR (higher is better)			
	Clear		Masked		Clear		Masked		Clear		Masked	
	50%	95%	50%	95%	50%	95%	50%	95%	50%	95%	50%	95%
B1	0.004	0.009	0.007	0.014	47.8	70.6	40.6	75.5	5.5	4.1	5.4	4.1
B2	0.004	0.009	0.007	0.015	41.4	61.7	33.6	65.5	5.8	4.4	5.7	4.5
B3	0.004	0.011	0.008	0.020	34.8	48.9	21.8	38.4	5.9	4.6	5.9	4.6
B4	0.005	0.013	0.009	0.024	34.3	46.9	18.7	37.7	5.9	4.5	5.9	4.5
B5	0.009	0.029	0.024	0.050	34.4	49.7	12.5	30.2	5.6	3.9	5.6	3.7
B6	0.009	0.026	0.020	0.040	39.0	52.3	12.4	26.7	5.7	3.7	5.6	3.6

B7	0.006	0.020	0.015	0.034	37.3	50.5	13.2	27.8	5.7	3.9	5.6	3.7
LST	2.801	20.049			50.3	97.1			11.1	-1.9		

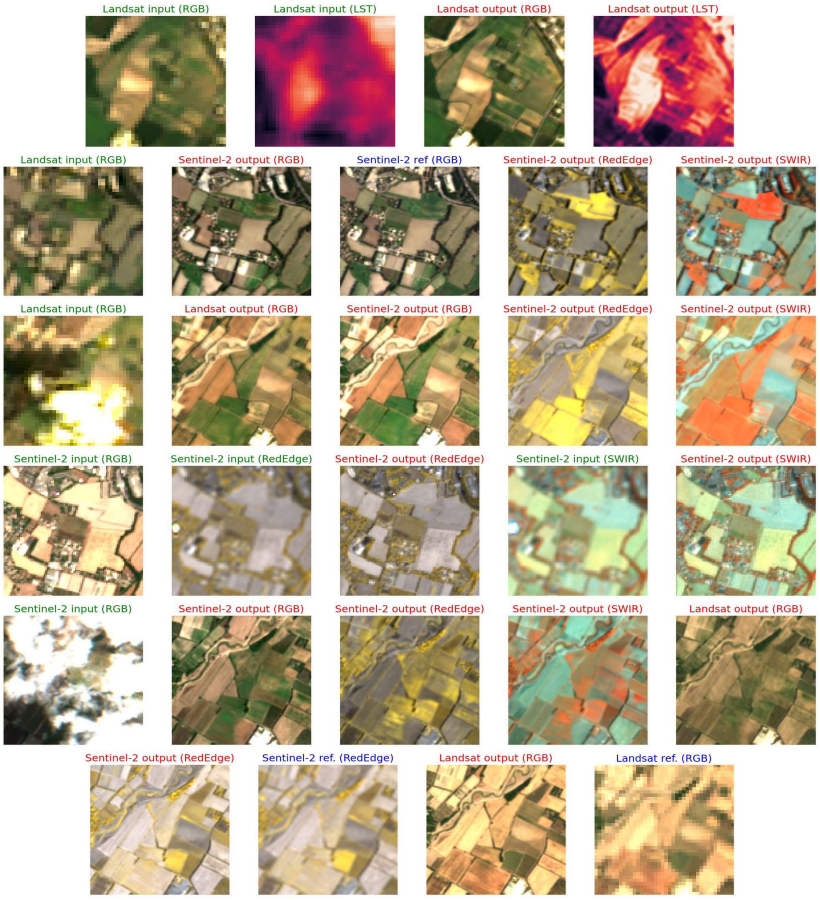


Figure 103. Illustration of the TAMRFSITS model capabilities over Area Of Interest 31TFJ for year 2022 (from the test set). Landsat SITS has 51 dates, of which 26 dates are removed from the model input and kept for validation. Sentinel-2 has 81 dates, of which 41 are removed from the model input and kept for validation. In this figure, images that are part of the model inputs are highlighted in green, model predictions are highlighted in red, and reference images are highlighted in blue. From top to bottom: rows 1 and 2 shows prediction for an observed Landsat date, row 3 shows prediction for a cloudy Landsat date, row 4 shows prediction for an observed Sentinel-2 date, row 5 shows prediction for a cloudy Sentinel-2 date, and row 6 shows predictions for an unobserved date.

Table 38. Sentinel2 prediction metrics for TAMRFISTS when applying a random masking strategy of 50% of the dates, yielding 365 clear Landsat dates and 389 masked Landsat dates. RMSE is expressed in surface reflectances (unitless). The 50% column shows median value across all dates, while the 95% shows the 95% quantile across all dates.

	RMSE (lower is better)				BRISQUE (lower is better)				FR (higher is better)			
	Clear		Masked		Clear		Masked		Clear		Masked	
	50%	95%	50%	95%	50%	95%	50%	95%	50%	95%	50%	95%
B2	0.006	0.037	0.014	0.093	19.8	40.4	25.4	50.2	-0.0	-1.2	0.0	-1.2
B3	0.006	0.032	0.015	0.096	16.1	26.8	20.2	33.0	0.1	-1.0	0.1	-0.9
B4	0.007	0.037	0.017	0.100	17.0	31.3	20.1	38.7	0.1	-0.9	0.1	-1.0
B8	0.011	0.043	0.034	0.099	12.2	25.3	15.4	29.4	0.1	-1.0	0.1	-0.8
B5	0.008	0.037	0.018	0.096	14.2	23.2	16.3	27.2	4.4	3.2	4.5	3.3
B6	0.011	0.039	0.027	0.097	12.9	23.6	14.2	27.1	4.4	3.2	4.5	3.2
B7	0.011	0.043	0.031	0.095	13.3	25.2	13.9	27.3	4.3	3.0	4.4	3.1
B8a	0.012	0.044	0.032	0.096	13.7	26.7	13.9	28.6	4.3	3.0	4.4	3.1
B11	0.011	0.029	0.026	0.061	13.1	28.3	14.6	27.8	4.9	3.8	5.0	3.8
B12	0.009	0.023	0.020	0.048	14.3	26.2	16.0	28.2	4.7	3.7	4.8	3.6

3.3.2 Comparison with Deep-Harmo

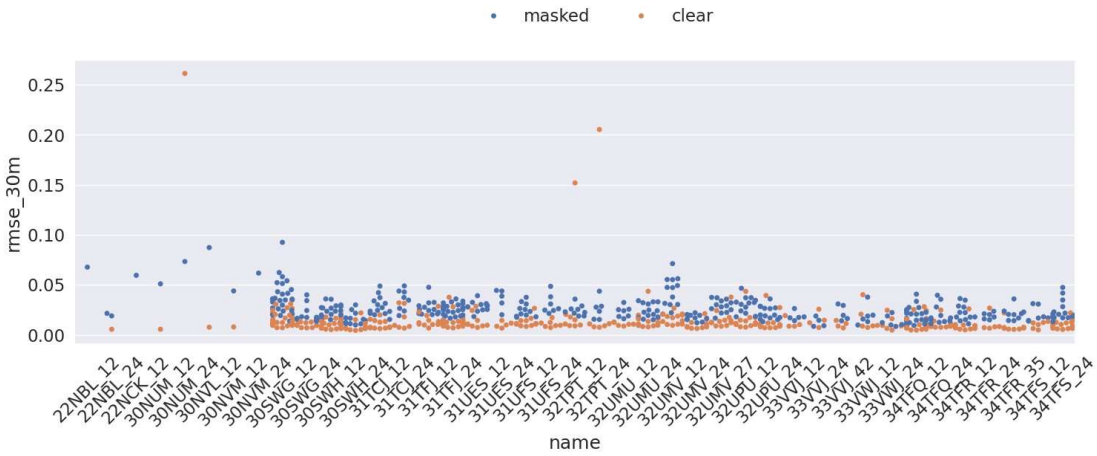
In this section, TAMRFSITS predictions are compared to those of Deep-Harmo (Sambandham et al., 2024).

In this work, an ensemble of UNet SISR models is used to predict Sentinel-2 bands B2, B3, B4, B8a, B11 and B12 from Landsat bands B2, B3, B4, B5, B6, B7, and B8 which is the Landsat panchromatic band that has 15-meter spatial resolution. It should be noted that prior to entering the UNets, all Landsat bands are up-sampled to 10 meter resolution. Deep-Harmo does not have the same capabilities as TAMRFSITS : it cannot predict unseen date, it does not predict all the Sentinel-2 bands, and it cannot predict cloudy Landsat dates. However, the estimated performances are solid, and the authors provide the pre-trained weights for the ensemble, and therefore Deep-Harmo is a good candidate to situate the performances of TAMRFSITS with respect to the literature.

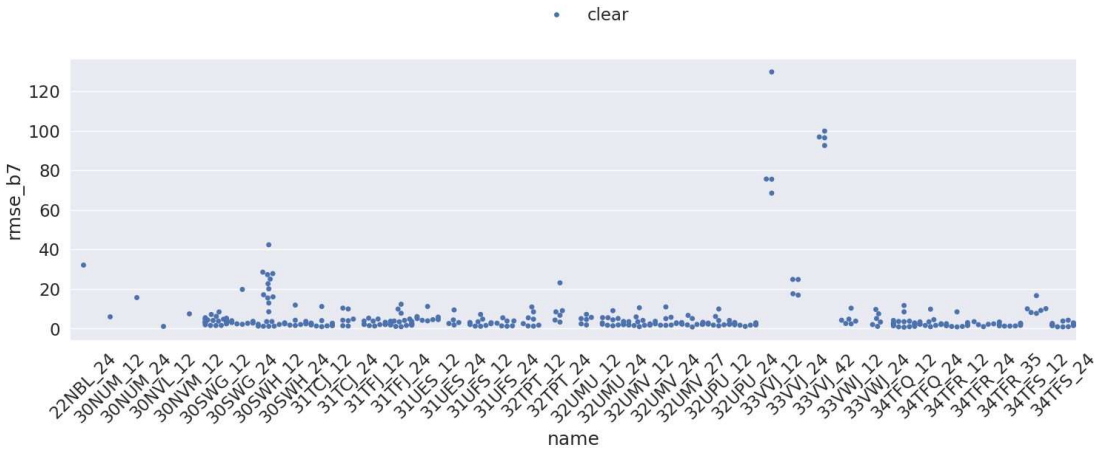
In order to do so, the following experiment is made for all AOIs of the testing set, whenever there is a same-day occurrence of Landsat and Sentinel-2, the Sentinel-2 image is removed from the model input and kept for validation. All the other dates serve as input to the TAMRFSITS model. Deep-Harmo pre-trained ensemble model is used to infer Sentinel-2 bands from Landsat for each of the same-day acquisition. Performances are then compared on the same-day Sentinel-2 images. Note that contrary to Deep-Harmo, TAMRFSITS does not use the panchromatic band from Landsat.

Table 39 shows the metrics corresponding to this experiment. Despite not using the B8, 15-meter resolution panchromatic channel from Landsat, TAMRFSITS outperforms DeepHarmo by a vast margin in all bands and all metrics. DeepHarmo has RMSE value 2 to 3 times higher than TAMRFSITS, and a BRISQUE score also 2 to 3 times higher. This means that radiometry of predicted bands is less faithful to

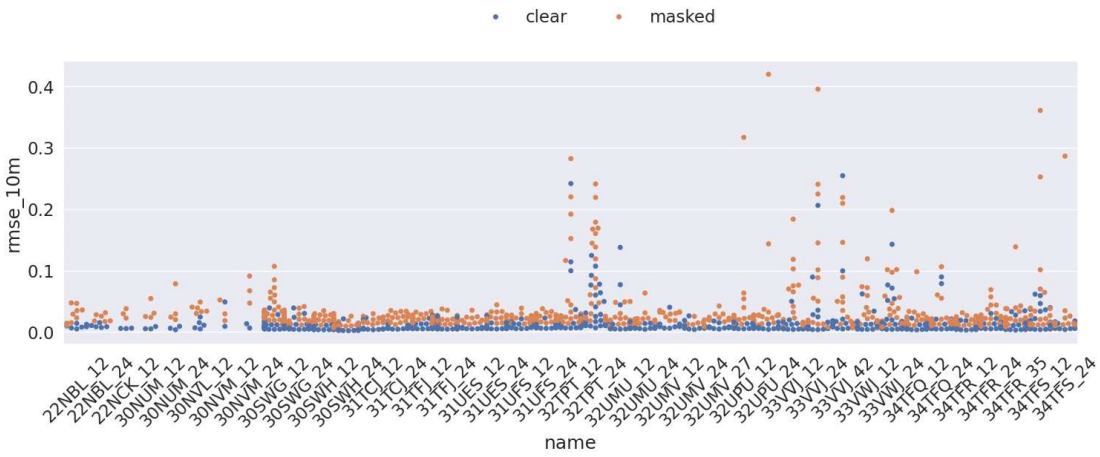
the true Sentinel-2 observation, and that the overall IQ of DeepHarmo prediction is far lower. Looking at the FR metric, it is interesting to note that for 10-meter bands, TAMRFSITS has a FR of 0, which is expected and means that the spatial resolution of the prediction matches the spatial resolution of the reference Sentinel-2 bands. On the same bands, DeepHarmo has FR values around -5 dB, which means that DeepHarmo has less high spatial frequency content than the reference Sentinel-2 bands, e.g. DeepHarmo does not yield 10-meter high-resolution details. On the 20-meter bands, TAMRFSITS provide a FR improvement of around 4 dB, while DeepHarmo shows a degradation of around -2 dB.



(a) Distribution of average RMSE on 30-meter Landsat bands across Areas of Interest for clear and masked dates.



(b) Distribution of average RMSE on 30-meter Landsat bands across Areas of Interest for clear dates only.



(c) Distribution of average RMSE on 10-meter Sentinel-2 bands across Areas of Interest for clear and masked dates.

Figure 104. Distribution of average RMSE of the output of TAMRFSITS, over 10 meter, 3022 meter and 90 meter bands.

This means that the true spatial resolution of DeepHarmo predictions is probably around 20-meter. This is confirmed by visual inspection of predictions from DeepHarmo and TAMRFSITS shown in figure 9, where DeepHarmo prediction shows radiometric distortion and blur for 10-meter bands. Note that TAMRFSITS uses the full joint Sentinel-2 and Landsat SITS for the prediction, while DeepHarmo only utilizes the Landsat image of the same date (albeit with the 15-meter panchromatic band). As such, it is not surprising that TAMRFSITS achieves better performances.

Table 39. Median values of RMSE, BRISQUE and FR compared between TAMRFSITS and DeepHarmo [Sambandham et al., 2024], over the full testing set, for 176 same-day acquisitions where Sentinel-2 images have been kept aside for validation. Best values for each metric are highlighted in bold.

	RMSE↓		*BRISQUE*↓		*FR*↑	
Band	TAMRFSITS	DeepHarmo	TAMRFSITS	DeepHarmo	TAMRFSITS	DeepHarmo
B2	**0.011**	0.036	**26.3**	52.4	**-0.0**	-5.9
B3	**0.012**	0.038	**23.2**	49.1	**0.1**	-5.6
B4	**0.013**	0.039	**25.6**	47.9	**0.1**	-5.0
B8a	**0.023**	0.047	**21.1**	64.8	**3.6**	-2.3
B11	**0.020**	0.047	**24.9**	64.1	**4.2**	-1.7
B12	**0.016**	0.040	**25.7**	63.2	**3.9**	-1.7

3.3 Summary and conclusions

In this work, we intended to provide a lightweight SISR tool for EvoLand prototypes to superresolved 10 Sentinel-2 bands to 5-meter resolution. This objective led us to favour relatively shallow networks, compared to literature in computer vision or remote sensing with higher super-resolution scaling factors. Indeed, if SISR was to

be part of the processing chain of prototypes for new Copernicus products, it would be used to process hundreds of thousands of products, and the benefit of SISR will be balanced with the additional processing and storage costs. The trained model achieves good SISR performances with remarkably low spectral distortion, which means that SISR products may be used in place of traditional Sentinel2 images by downstream applications, without modification of any parameters related to radiometry.

For relatively shallow models, we proposed a training methodology adapted to the specificity of the Sen2Venμs dataset, allowing to work around the different input resolutions, the radiometric and geometric discrepancies inherent to a multi-sensor dataset and the lack of SWIR bands in the target reference data. We also demonstrate that when trained properly, those models achieve very good performances in pre-training, and that GAN fine-tuning only marginally improves those performances, with the additional hassle of model selection during the fine-tuning stage. This divergence with the literature can have two different explanations: either pre-training is very weak in existing work and good performances could be achieved by better pre-training of the models, without the GAN part, or the GAN training benefits to deeper models, that are out of the scope of this study.

In addition, work on Sentinel-2 SISR has been extended in a scientific publication currently being reviewed for IEEE Transactions on Geo-Science and Remote Sensing (Michel et al., 2025)

Besides the single image Super Resolution, In the second phase of the project, we focussed on the fusion of multi-modal Satellite Image Time Series and application to Thermal Infra-Red. A new problem formulation has been proposed that allows to handle both temporal interpolation and super-resolution in a unified framework. This formulation led us to a new Deep-Learning architecture and training strategy, that allows to achieve good performances on the fusion task with respect to the literature, and can tackle several fusion tasks (sensor translation, cloud interpolation, super-resolution, temporal extrapolation) where models from the literature are dedicated to a single task.

4 Continuous monitoring

4.1 Change detection – Time series modelling

4.1.1 State of the art

Forest change detection is a crucial area of research, particularly in monitoring disturbances such as storm events and insect calamities (Puhm et al., 2020). Various remote sensing approaches have been utilized to monitor forest degradation, focusing on visible impacts on forest canopies detectable through remote sensing techniques (Mitchell et al., 2017). In addition, global forest change datasets have also shown potential in aiding countries with national forest monitoring and reporting requirements (Galiatsatos et al., 2020). Within CLMS, forest change detection is a key part of the HRLs on Forest and Vegetated Land Cover Characteristics (VLCC). The evolution of these layers is targeted by the C1, C2 and C3 forest prototypes in EvoLand, for which methods of time series modelling will be exploited.

4.1.1.1 Introduction

The availability of dense time series of Earth observation data has led to a growing interest in time series analysis for vegetation monitoring and change detection. Through high-quality georeferencing and atmospheric correction of the satellite images, it is possible to create consistent sequences of reflectance values measured over time. Time series analysis (TSA) can be applied to extract information out of these data signals, like identifying signal changes possibly corresponding to changes of land cover. TSA has been identified as a promising method especially for forest monitoring. The reflectance signals of undisturbed forest pixels can be captured quite well by time series models using Machine Learning (ML) methods. Deviations from the learned, undisturbed state can be isolated in an unsupervised manner based on statistical thresholds. This is a key advantage over Artificial Intelligence (AI) methods, which often require lots of training data. However, vegetation monitoring algorithms also need to deal with several signal characteristics such as seasonality, irregular sampling intervals, and signal artefacts which can make the extraction of reliable information difficult.

Dense time series of high-resolution optical data have a number of characteristics that pose a challenge to change detection applications. In addition to noise effects remaining after atmospheric correction and uncertainties in the geometric registration, these challenging characteristics include:

- Seasonality. Recurring seasonal patterns can be attributed to plant phenology and/or varying illumination conditions due to topography and

solar angle. Considering that annual variations of temperature and rainfall also affect phenology, the seasonal patterns can vary between years.

- Irregular sampling interval. Satellites with a regular nadir acquisition scheme usually have a constant revisit cycle, for example 5 days for the Sentinel-2A and -B constellation. If images from overlapping orbits are integrated though, the sampling interval becomes irregular. Data gaps due to masked clouds, cloud shadows, and snow also add to this irregularity.
- Presence of signal artefacts. Despite using state-of-the-art screening algorithms like Fmask (Z. Zhu & Woodcock, 2012), un-masked clouds, cloud shadows, and snow-covered areas remain in the pre-processed imagery. Corresponding observations have to be treated as invalid since the measured reflectance values neither represent the undisturbed land cover state nor a persistent change of it.

Accuracy and temporal resolution are crucial for the early detection of environmental disturbances, with most methods aiming for at least annual temporal resolution. Spatial resolution objectives vary, with efforts generally aiming for finer resolutions, such as 10-metres with Sentinel-2 data, to accurately characterise disturbances. The push towards automation and scalability is clear, with methods like BFAST and CCDC crafted for large-scale application (Verbesselt et al., 2010a; Z. Zhu et al., 2012a). However, the computational and storage demands for continuous monitoring pose significant challenges.

4.1.1.2 Review of existing methods

This review synthesizes the existing literature, analyses trends, compares accuracies, and identifies gaps in the current research and implementations.

Algorithms can be divided based on how they deal with the time series characteristics described in the introduction above. The different approaches used to handle these characteristics strongly affect the algorithms' suitability for monitoring changes timely and reliably. **Approaches that do not account for seasonality** form a first group of algorithms. A review of change detection studies using Landsat time series concludes that many older studies focused on mapping changes only at annual or biannual time scales based on series of cloud-free composite images which always represent the same growing season (Z. Zhu, 2017). In this context, both the Vegetation Change Tracker (VCT) (Huang et al., 2010) as well as the LandTrendr approach (Kennedy et al., 2010a) represent widely used algorithms, but they are not designed for near real-time mapping.

A **second group of algorithms explicitly accounts for seasonality** by using regression models based on trigonometric functions to capture the intra-annual variations (variations within one year, i.e. seasonality) of the spectral signatures independently for each pixel. With this approach, also frequently referred to as harmonic regression, periods of stable land cover are modelled as a deterministic, continuous function of time. Irregular sampling intervals and data gaps are therefore not a problematic issue, but the deterministic nature of the model does

not allow inter-annual variations of the seasonal pattern (variations between different years, e.g. shifts in seasonality). The seasonal model represents an average of different conditions occurring within a stable period, e.g. dry and wet years, late and early leave outbreak. A widely used algorithm belonging to this group is the Breaks for Additive Season and Trend (BFAST) algorithm (Verbesselt et al., 2010b) and its evolution BFAST Monitor (Verbesselt et al., 2012a). While the latter is tailored to near real-time mapping of new changes, the original version is intended for the analysis of historic time series. Both versions have been used in a variety of studies and can be applied to detect both abrupt and gradual changes. Concerning the robustness of BFAST to invalid observations, it has been stated that occasional signal artefacts are well handled, but temporally aggregated occurrences such as several consecutively un-masked clouds can be a source of error (DeVries, Verbesselt, et al., 2015). Also, additional pre-processing to eliminate artefacts was applied (DeVries, Decuyper, et al., 2015; DeVries et al., 2016; Hamunyela et al., 2016). The second widely used implementation of the harmonic regression approach is the Continuous Change Detection and Classification algorithm (CCDC), where the original concept (Z. Zhu et al., 2012b) is extended to include more types of land cover beside forest as well as a classification framework (Z. Zhu & Woodcock, 2014). From the beginning, CCDC was designed to work with dense Landsat time series and can handle seasonality, irregularly spaced observations, and signal artefacts to some extent. Both abrupt and gradual changes can be detected. Some further updates to the algorithm include (i) a mechanism to automatically adjust the complexity of the time series model based on the number of available clear observations as well as (ii) a different method to estimate the model parameters which reduces overfitting (E. B. Brooks et al., 2014). A third algorithm employing harmonic regression utilizes residuals from the regression together with statistical quality control charts (E. B. Brooks et al., 2014). This approach comprises signal artefact detection with Shewhart X-bar charts. After the elimination of artefacts, both abrupt and gradual changes are indicated by exponentially weighted moving average (EWMA) charts in a near real-time manner. All the described algorithms of the second group share certain basic concepts, but the individual implementations vary. They are designed to process large amounts of data in a highly automated way and therefore rely on data-driven statistical boundaries for detecting change, although the distinct nature and computation of these boundaries is quite different.

The **third group of algorithms stands out by also taking inter-annual variations of the seasonal pattern into account**. Structural time series models are set up in terms of components, such as trends and cycles, which have a direct interpretation. Their statistical treatment is based on the state space form and the Kalman filter and first described for time series analysis in econometrics (E. B. Brooks et al., 2014). Compared to the harmonic regression approach, the model is no longer deterministic. Kalman filtering denotes a versatile parameter estimation technique which yields optimal estimates in a statistical sense (Kalman, 1960). It is well established in many application fields and has been applied to numerous signal tracking problems. The combination of structural time series models with the

Kalman filter and the concept's suitability for remote sensing purposes has been investigated in (Puhm et al., 2020). The Kalman-filter has also been applied to time series of MODIS 8-day composites to detect insect induced defoliation in near real-time at a forest test area in northern Sweden (Olsson et al., 2016). This study makes use of the CUSUM control chart (PAGE, 1954) to indicate changes, but it does not combine structural time series models with the Kalman filter. Instead, the filter is used to derive a smoothed NDVI time series based on a global model trajectory. Another algorithm belonging to the third group is BEAST (K. Zhao et al., 2019). BEAST stands for Bayesian Estimator of Abrupt change, Seasonal change, and Trend. "As an ensemble algorithm, BEAST quantifies the relative usefulness of individual decomposition models, leveraging all the models via Bayesian model averaging." (Decuyper et al., 2022) introduce the AVOCADO (Anomaly Vegetation Change Detection) method, designed for the continuous monitoring of forest disturbance and regrowth across diverse forest types and conditions using satellite time series data. The AVOCADO algorithm employs kernel density estimations to create a forest reference phenology that incorporates the natural variability in forest phenology over time, rather than smoothing these variations. This approach allows for the detection of anomalies in forest condition with a spatially explicit likelihood measure, making it capable of capturing both large-scale disturbances and subtle, gradual changes in the forest landscape, such as selective logging and the effects of shifting cultivation. The advantages of the AVOCADO method lie in its flexibility and accuracy in monitoring forest changes across different ecosystems and disturbance regimes (Decuyper et al., 2022).

Cloud cover can obscure the Earth's surface, making this a significant challenge for optical data sources. The presence of frequent cloud cover significantly hinders the accuracy of disturbance detection methods, underlining a crucial gap in effective strategies to mitigate cloud impacts. This limitation is particularly pronounced in tropical and mountainous regions where cloud cover is persistent, impacting the quality of remotely sensed information and hindering accurate digital mapping (Bai et al., 2019; Lechler et al., 2020).

There is a clear trend towards integrating and fusing diverse data sources for forest monitoring. Studies have shown that the fusion of optical and radar data leads to improved classification accuracy and better results in land use mapping and monitoring. Efforts to combine optical and radar data, along with other environmental datasets, are increasing, with the aim of achieving a comprehensive monitoring approach (Frison et al., 2018; Sarzynski et al., 2020). The integration of multi-source data, such as combining Landsat with Sentinel-2 or integrating optical with radar data, presents a promising approach. However, there is substantial room for improving the synergy between multiple data sources. For example, expanding the integration to include various environmental datasets could offer a more holistic view of ecosystems and their disturbances.

Machine learning methods, such as support vector machines, neural networks, deep learning, and natural language processing, have been instrumental in processing structured and unstructured data for improved forest monitoring (Lee

et al., 2020; Li et al., 2023). The use of machine learning and AI is becoming more dominant, offering solutions to improve accuracy, reduce computational requirements and increase the level of automation in monitoring systems.

A fine balance exists between the complexity of methodologies and their computational efficiency. While advanced methods promise enhanced accuracy, they often come at the cost of significant computational resources. This trade-off raises questions about the practicality of complex methods, especially when the gains in accuracy are marginal compared to the increased resource demands.

Understanding carbon dynamics is also becoming increasingly important, linking disturbance monitoring with the quantification of carbon emissions and removals. Although directly quantifying these aspects remains challenging, it is an important direction for future research, given the global urgency of addressing climate change.

Significant progress has been made in terms of temporal resolution, automation, and scalability. However, challenges still persist in dealing with cloud cover, intra-annual variability, and computational efficiency. Future research should concentrate on integrating various data sources, applying advanced computational techniques, and tailoring methods to specific user needs and ecosystem types.

4.1.1.3 Strengths and weakness of existing methods

The literature on detecting forest disturbance presents a variety of methodologies, each with its own strengths and limitations.

Methods such as the Vegetation Change Tracker (VCT) and LandTrendr are widely used and stand out for their simplicity and user-friendly nature, making them highly accessible to a broad spectrum of users. By focusing on changes over annual or biannual timescales, these methods offer efficient means to detect significant disturbances without the computational complexities inherent in more nuanced models. However, the principal limitation of these approaches lies in their inability to perform near real-time mapping, which constrains their utility in situations that require rapid response. Moreover, their disregard for seasonal variations can lead to the mischaracterization or oversight of disturbances in ecosystems where seasonal dynamics are a significant factor (Huang et al., 2010; Kennedy et al., 2010a).

The VCT methodology, with its specific focus on forests with closed or near-closed canopy, demonstrates a tailored approach to vegetation monitoring. The utilization of biannual stacks of vegetation indices for disturbance detection aligns well with medium-term change monitoring. Nonetheless, the method's applicability is somewhat limited by the need for supervised scene selection, potentially restricting its use across diverse forest types and complicating scalability (Huang et al., 2010).

LandTrendr excels in capturing both short-duration events and long-term trends, employing a temporal segmentation algorithm that ensures seasonal consistency

in monitoring. However, the reliance on manual preprocessing steps, such as scene selection and cloud removal, may hinder its operational efficiency. Additionally, the method's performance is heavily reliant on the quality and consistency of input data (Kennedy et al., 2010a).

Methods employing harmonic regression, such as BFAST, BEAST and CCDC, adeptly handle intra-annual variations, thereby enhancing the capacity to detect changes within a single year. These methodologies show a commendable resilience to occasional signal artefacts, attributing to their robust design. Nevertheless, they are not without limitations; the presence of temporally aggregated artefacts, like consecutive unmasked clouds, can lead to inaccuracies. Moreover, the deterministic nature of harmonic regression models may not sufficiently capture inter-annual variations, potentially overlooking subtle or gradual changes (Verbesselt et al., 2010b; K. Zhao et al., 2019; Z. Zhu et al., 2012b).

BFAST stands out for its versatile application across various vegetation types, eliminating the necessity for land cover-specific adjustments. Its capability to analyse a wide array of time series data from different satellite sensors enhances its utility. However, the method's computational intensity and the requirement for continuous, consistent time series data pose challenges, particularly in regions with frequent cloud cover or data gaps (Verbesselt et al., 2010b).

BEAST introduces a unique set of advantages, particularly its departure from the need to select a singular time series model that is presumed to yield the best results. Instead, it operates within a defined parameter space and aims to identify the most appropriate model for the given time series signal. This methodological flexibility extends to the detection of structural breakpoints within both the trend and seasonal components of the data, providing a comprehensive range of estimates such as the probability of a breakpoint, its temporal location, and the magnitude and direction of the change. These capabilities position BEAST as a powerful tool for in-depth time series analysis (K. Zhao et al., 2019).

CCDC offers a continuous monitoring approach that dynamically updates with every new image, facilitating real-time change detection. This comprehensive methodology integrates image preprocessing, change detection, and classification, presenting a holistic solution for vegetation disturbance monitoring. Despite its strengths, the high computational demand and assumptions of consistent phenology for stable land covers may limit its effectiveness in variable environments (Z. Zhu & Woodcock, 2014).

Incorporating components such as trends and cycles, methods including BEAST and those based on Kalman filtering, provide a detailed perspective on temporal changes, acknowledging both intra- and inter-annual variations. Techniques like Kalman filtering and Bayesian model averaging stand out for their ability to offer statistically optimal estimates, thus enhancing the precision of change detection. However, the sophistication and complexity of these models necessitate a higher level of expertise and computational resources, which may limit their accessibility and practical application. The computational demands associated with these

techniques pose challenges for their use in large datasets or for near real-time monitoring without substantial computational support.

CODED specifically addresses the challenge of detecting low-grade forest canopy disturbances, a critical but often overlooked aspect of forest change. It provides detailed metrics on disturbance events, including magnitude, duration, and recovery. However, its focus on deforestation and degradation may omit other types of vegetation changes, and the semi-automated nature of the method, requiring manual effort for land cover classification, could limit scalability (Bullock et al., 2020).

AVOCADO introduces a new approach to forest change detection, aiming to identify anomalies in vegetation dynamics. This method enhances the spectrum of disturbance detection methodologies by focusing on the early identification of subtle changes, potentially preceding larger-scale disturbances. AVOCADO must navigate the balance between computational efficiency and the need for accurate, timely data (Decuyper et al., 2022).

After conducting a thorough analysis of the existing methodologies for detecting forest disturbances, it is clear that each approach has its own strengths and weaknesses. When evaluating the reliability and validity of these methodologies, it is evident that each has undergone rigorous peer-reviewed studies to ensure high reliability in their respective applications. Operational readiness varies, with some methods such as CCDC and BFAST exhibiting higher degrees of automation and scalability, in contrast to others that require more intensive manual preprocessing or parameter adjustment. The challenges faced in the field of forest disturbance monitoring are underscored by the common limitations across these methodologies. Significant considerations include computational demands, data requirements, and adaptability to near real-time applications. Furthermore, the accuracy and suitability of these methods may differ depending on the ecosystem and type of disturbance, emphasising the need to choose a methodology that closely aligns with the particular objectives and limitations of each monitoring approach. Among these methodologies, BFAST, VCT, LandTrendr, CCDC, CODED, Kalman filter and the more recent AVOCADO offer a range of capabilities for forest disturbance detection and monitoring.

4.1.2 Method description

This section describes one of the aspects of the methodology involved towards implementing a forest monitoring system with the ability to not only *detect* disturbances accurately and timely but also *classify* the disturbance type. The detection step is still an active area of research in the remote sensing field, as outlined in the preceding section. One goal of the first project phase was to conduct a benchmarking of several already available unsupervised change detection algorithms leveraging TSA and select the most suitable method for the production of the first prototype iteration.

4.1.2.1 Input data and study sites

The selection of study sites Figure 105 is at least partly representative for a range of different conditions encountered in Europe. There are considerable differences regarding the amount and quality of available optical input data, often linked to local properties like the average amount of cloud and snow cover or the average sun elevation over the year. For example, time series of the Spanish demonstration sites will likely exhibit a superior density (i.e. the number of usable clear-sky observations) and signal quality (the signal-to-noise ratio, SNR) compared to time series of the Swedish demonstration sites or the Alpine region. Another important factor concerning time series density is whether one or two satellite orbits cover a region. Figure 106 illustrates examples of various densities. For the benchmarking of TSA algorithms, these aspects of input data quality can be a source of unwanted bias. To reduce this bias in the benchmarking process, we reduce and homogenize the temporal resolution of the input data for all study sites by using monthly composites.

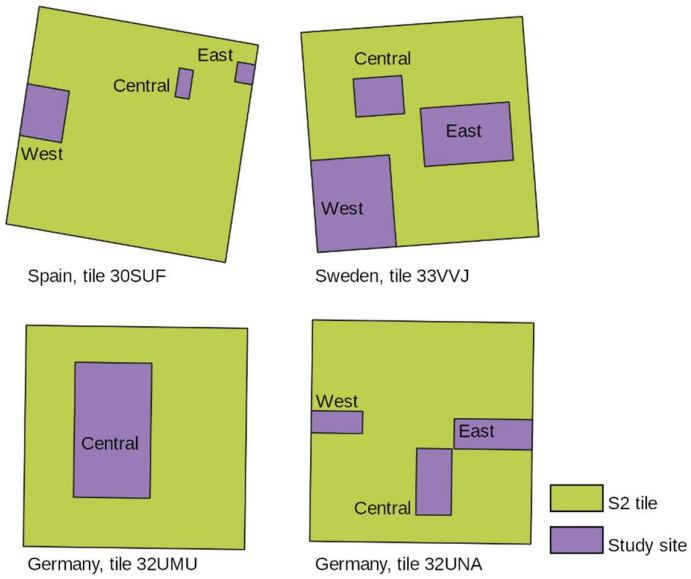


Figure 105. Benchmarking study sites and their locations within the demonstration sites.

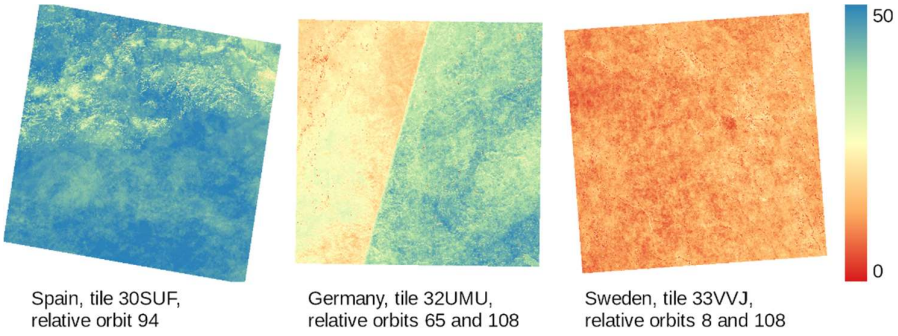


Figure 106. Number of available clear-sky observations for the year 2019.

Almost all of the change detection algorithms mentioned in Section 4.1.1.3 are univariate, i.e. change detection is based on one selected observable sensitive to the targeted disturbance types. The existing literature mentions many options like NDVI, NBR, or components of the Tasseled Cap (TC) transformation (Cohen et al., 2017; Oeser et al., 2017). We decided to use the TC Wetness (TCW) component with transformation coefficients given in (Crist, 1985). After cloud masking based on the readily available Sentinel-2 Scene Classification (SCL), monthly TCW composites have been generated by selecting the median of all un-masked observations acquired in one calendar month. Input data sets are listed in Table 40. The cloud masking approach was to consider SCL pixel values of 4 (vegetation), 5 (not vegetated), 6 (water), and 7 (unclassified) as clear, while all other SCL pixel values result in a masked input pixel (no data).

Table 40. Sentinel-2 and derived input data sets.

Dataset	Description	Spatial Resolution	Temporal resolution
Sentinel-2 data	Sentinel-2 L2A spectral bands, Scene Classification (SCL)	10 – 20 m	Single scenes
Masked Tasseled Cap Wetness (TCW)	Derived from Sentinel-2 L2A spectral bands and SCL	10 m	Single scenes
Monthly TCW Composites	Derived from masked TCW	10 m	Monthly

4.1.2.2 Description of the candidate methods

The benchmarking features five candidate methods:

- BFAST with CuSum monitoring strategy
- BFAST with MoSum monitoring strategy
- EWMA
- BEAST
- Kalman filter (KF) with CuSum monitoring strategy

Both BFAST variants and EWMA are implemented in the nrt Python package maintained by the JRC (https://github.com/ec-jrc/nrt). All these algorithms are based on the same concept. First, a deterministic harmonic time series model featuring trend and seasonal components is fitted to historic data preceding the monitoring period. The order of the harmonic model needs to be specified by the user. Second, the residuals of data acquired in the monitoring period w.r.t. the historic model are computed. It is assumed that the mean of the residual sequence remains approximately zero if no structural breaks in the signal occur. Significant deviations to this assumed behaviour (and therefore change) are signalled by either CuSum, MoSum or EWMA test statistics crossing a user-defined level of significance. The computational load of these algorithms is reasonable. The most extensive task is setting up the historic model.

The author of BEAST has released an implementation on GitHub (https://github.com/zhaokg/Rbeast). BEAST has some unique advantages

compared to the other candidates: On the one hand, the user does not have to select a single time series model deemed to give the best results. Instead, only the boundaries of a certain model parameter space have to be specified. Then the algorithm tries to find the best model within this space for the given time series signal. On the other hand, it searches for structural break points separately in both the trend and seasonal component. For each identified structural break point, the algorithm is able to give a range of useful estimates like break probability, its location in time or its magnitude and direction. There are also some disadvantages:

- BEAST is not designed for near real-time applications. There is no function for "saving" a model and continue processing once new data is available.
- The algorithm does not fully support irregularly spaced time series common in remote sensing. Although irregularly spaced data is accepted as input, the program internally aggregates or interpolates the data to a regular interval. There is no documentation about how this is done.
- For real world applications targeting large areas and a wide range of disturbance types, some required initial assumptions about the model space can be difficult to make and the consequences on the results are hard to gauge, for example
    - The minimum and maximum number of expected structural break points, separately for trend and seasonality
    - The minimum spacing allowed between adjacent structural breakpoints, separately for trend and seasonality
    - The maximum number of outliers in the time series
- Long program run time

The change detection method based on structural time series models and the Kalman filter is methodologically closer to BFAST and EWMA. The main difference is that a stochastic time series model is used instead of a deterministic harmonic model. The stochastic model is updated incrementally throughout the monitoring period and allows adaptations to phenological variation between years. These incremental model updates slightly increase the computational load in the monitoring phase compared to methods employing deterministic modelling. Changes are signalled by evaluating the sequence of Kalman filter innovations. It is assumed that the mean of the innovation sequence remains approximately zero if no structural breaks occur in the signal. Significant deviations to this assumed behaviour (and therefore change) are signalled by a CuSum test statistic crossing a user-defined level of significance.

4.1.2.3 Benchmarking & evaluation

Evaluating the performance of forest change detection algorithms must take the area proportions of disturbed and undisturbed classes into account. Even the sum of all disturbed patches only accumulates to a small percentage of the entire forested area. One answer to this problem is to implement a stratified random sampling of validation points and interpret them manually to obtain a reference classification. Because a large number of point samples is still required, this process

is often quite time consuming. We therefore chose to use a conceptually similar approach based on smaller numbers of interpreted polygons. Reference polygons for windthrow and bark beetle infestation have been extracted from a recently published dataset which partially overlaps the German demonstration site[12]. Reference polygons for wildfire damage have been acquired via the EFFIS service. We also added a few clear-cuts digitized by an in-house interpreter to the set of disturbance polygons. Additionally, also reference polygons of undisturbed forest were required. Using multi-temporal VHR imagery available in Google Earth, an interpreter identified and digitized forest stands where no disturbances occurred over the specific monitoring periods of the study sites. The number and size of the reference polygons has been tuned so that their combined area represents realistic area proportions of disturbed and undisturbed forest observed over large areas Table 41. This way an unbiased assessment of how well algorithms can isolate the often small disturbance patches and still maintain a low rate of commission errors at the same time is possible.

Table 41. demonstration cites and their respective undisturbed and disturbed areas.

Region	Reference area undisturbed [ha]	Reference area disturbed [ha]
Spain	1218.7 (95.5%)	57.0 (4.5%)
Germany	8780.8 (99.1%)	77.7 (0.9%)
Sweden	1127.6 (94.7%)	63.3 (5.3%)
Sum	11127.1 (98.3%)	198.0 (1.7%)

4.1.3 Implementation and results

All algorithms selected for benchmarking have been applied to each study site depicted in Figure 105. However, in an early phase of this process it became clear that the runtime of BEAST is a problematic issue, especially with a European roll-out at 10m spatial resolution in mind. To give an example, the overall BEAST runtime for the study site “32UNA West” with a size of ~2.8 megapixels was 3300 seconds, including file read/write operations. On the same machine, the typical runtime of methods implemented in the JRC nrt package did not exceed 25 seconds. Considering that existing reports of the accuracy of BEAST are promising, but also not exceptional, this significant runtime penalty does not seem warranted, and BEAST was excluded from further testing. The time frame and main disturbances for the study sites of each region are listed in Table 42.

Table 42. Study sites and their respective disturbance event.

Region	Historic period	Monitoring period	Main Disturbance type/event
Spain	2019-01-01 to 2020-12-31	2021-01-01 to 2021-12-31	Wildfire September 2021
Germany, 32UMU	2017-01-01 to 2019-12-31	2020-01-01 to 2020-12-31	Bark beetle season 2020

[12] https://doi.org/10.3220/DATA20231025183526-0

Germany, 32UNA	2016-01-01 to 2018-12-31	2019-01-01 to 2019-12-31	Summer storm August 2019
Sweden	2015-07-01 to 2018-07-01	2018-07-01 to 2019-06-31	Wildfire August 2018

Several runs with different sensitivity settings have been conducted for all remaining benchmarking candidates. The results listed in Table 43 represent the parameter setting which yielded the highest overall accuracy out of all tries for the specific method. In all cases a single parameter set was used for all regions. Bold numbers highlight the highest accuracy in each column, underlined numbers the second highest. User's and Producer's accuracy (abbreviated UA and PA, respectively) are given for the change class and indicate how well an algorithm isolates the disturbances. The overall accuracy often reaches very high values, because most of the forest area is correctly classified as undisturbed. A high degree of variance between regions and algorithms can be observed. Both EWMA and KF/CuSum score outperform the BFAST variants. A decision between EWMA and KF/Cusum in favour of the former is made based on the fact that EWMA scores one more second place than KF/Cusum and achieved a particularly good result in the German region representing Central European conditions.

Both BFAST variants show a higher rate of commission error (low UA) despite already low sensitivity settings. This indicates that EWMA and KF/CuSum handle the remaining signal artifacts in the monthly composites better. The very high UA of KF/CuSum in Sweden is probably the most unexpected result. Regarding the PA values, one important aspect has to be noted. The storm and wildfire events happened in August or September. For those study sites where the monitoring period ends in the same year, the resulting forest disturbances can theoretically be observed no more than 4 times using monthly composites. A more realistic assumption is 2-3 times. The achieved PA therefore gives an indication of the performance in near real-time scenarios. It can be expected that PA would increase if the monitoring period were extended.

Table 43. Accuracy assessment.

	Spain			Germany			Sweden			Overall		
Method	**UA**	**PA**	**OA**	**UA**	**PA**	**OA**	**UA**	**PA**	**OA**	**UA**	**PA**	**OA**
BFAST/CoSum	61.8	52.4	96.4	38.1	64.1	98.8	45.1	**74.4**	93.8	44.7	64.0	98.0
BFAST/MoSum	45.6	**61.5**	95.0	62.6	**72.2**	99.4	22.6	71.8	85.4	37.2	**69.0**	97.4
EWMA	**77.3**	56.1	**97.3**	**91.8**	59.4	**99.6**	51.3	51.0	94.8	71.4	55.7	98.8
KF/CoSum	74.0	46.2	96.9	67.7	60.8	99.4	**89.2**	69.2	**97.9**	**76.0**	59.3	**99.0**

All numbers in %, UA User's Accuracy (change class), PA Producer's Accuracy, OA Overall Accuracy

The following flow chart describes the proposed methodology for intra-annual detection of forest disturbances to be implemented in the first prototype iteration (Figure 107).

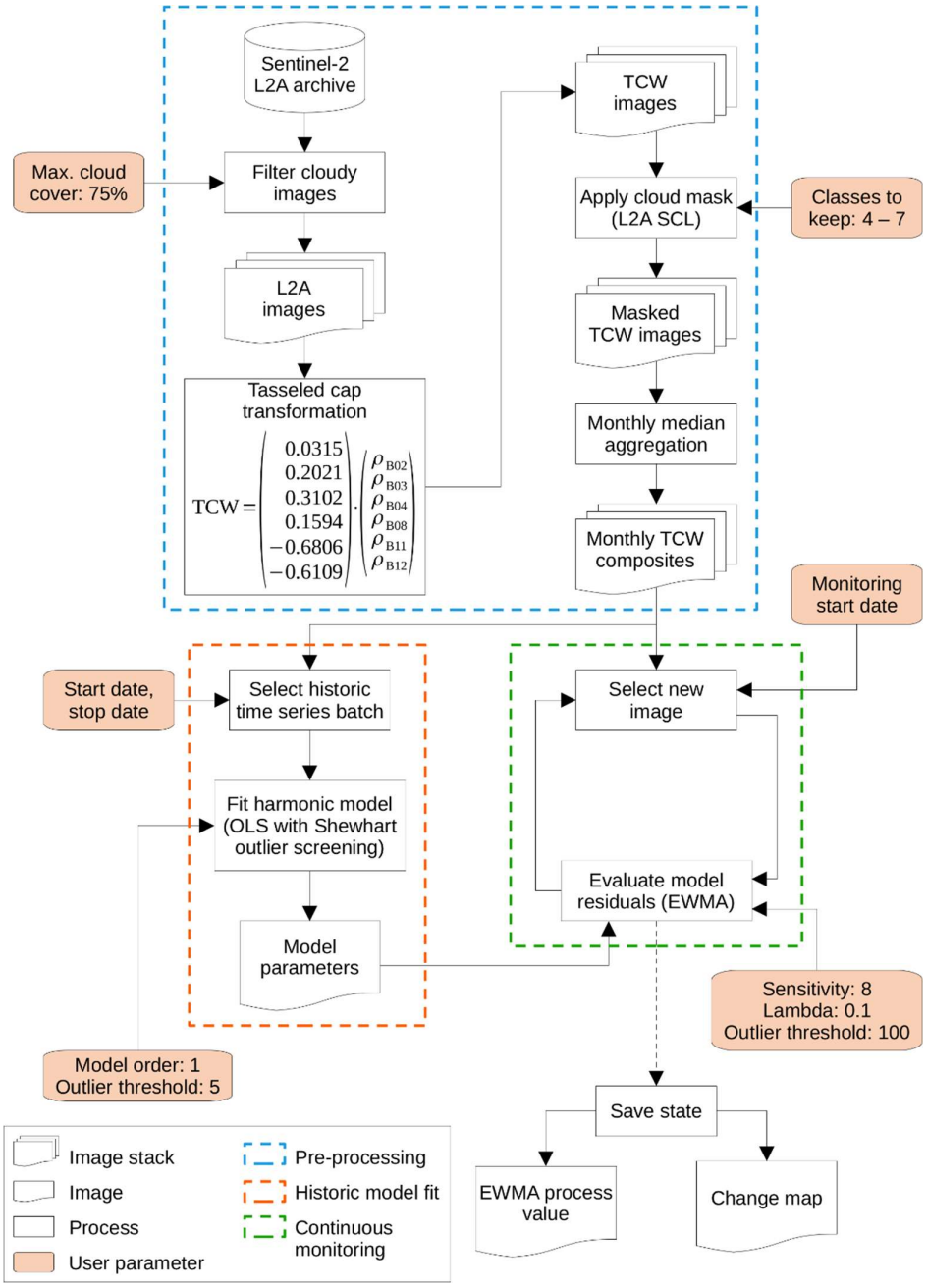


Figure 107. Proposed forest change detection methodology.

4.1.4 Summary and conclusions

The benchmarking has yielded very interesting results. Existing methods already deliver in many cases practical change detection results with reasonable computational effort. A possible combination of the individual strengths of the EWMA and KF methods presents a promising perspective towards reducing the regional variance in accuracy.

The extraction of meaningful change patches from optical time series data is an important prerequisite for the correct classification of change drivers. Although one of the benchmarking candidates proved to be too computationally expensive to be used for initial change detection, it might still play a role in the second step of change classification. It could be applied to get refined estimates of the change date or magnitude only for previously isolated patches.

4.2 Disturbance classification

4.2.1 State of the art

The classification of disturbances is an important application of remote sensing data. The capability to not just detect changes but also classify them into drivers or disturbance agents is particularly sought after in the domain of forest disturbances.

Different forest disturbances have different impacts on forest recovery and available biomass. Thus, the classification of forest disturbance agents is an important part of the sustainable management of forests. Satellite data is uniquely capable of providing a large scale, consistent and statistically validated baseline for the types of disturbances, their frequency as well as spatial and temporal patterns of occurrence.

4.2.1.1 Introduction

Forest Disturbance Agents can be split into two main types: Abiotic and biotic. Abiotic includes disturbances like windthrow and wildfire but also includes silvicultural practices like thinning and clear-cutting. Biotic disturbances are mainly due to insects. In Europe these are mainly bark beetle and gypsy moth infestations.

The separation of the disturbance agents using satellite data can be a challenge, since especially lower impact disturbances like partial windthrow, bark beetle die-off or thinning practices can have very similar spectral signatures.

An additional challenge in disturbance agent classification is the availability of ground truth data. In general, there doesn't exist a lot of openly available high quality recent ground truth data. Most available data has low temporal accuracy, only giving the year in which the disturbance happened.

4.2.1.2 Review of existing methods

In the literature there already exist algorithms for automatic forest disturbance agent classification (Paulino et al., 2024). However, most of these publications either focus on discerning a single disturbance type (M. S. Brown et al., 2025; Giambelluca et al., 2025) or are only applied to small areas of interest (M. S. Brown et al., 2025; Giambelluca et al., 2025; Sebald et al., 2021; Shimizu et al., 2019). The European Forest Disturbance Atlas (Viana-Soto & Senf, 2024) is one of the few publications of an algorithm applied to all of Europe, which discerns different disturbance classes (Wind/Bark Beetle, Fire, Harvest, Mixed). The algorithm uses a series of annual

summer composites using Landsat data as input data. The classification is done using a Random Forest classification. The model was trained on a stratified sampling which was labelled using human interpreters.

Key observations from the European Disturbance Atlas and other available literature are that few openly available reference datasets are available and labelling of samples into different disturbance agents is not possible without ancillary data like very high-resolution images or ground surveys. Another observation is that for satellite imagery like Landsat or Sentinel 2, the inclusion of spatio-temporal features like the shape, location, and timing of the disturbance increases the accuracy of classification (Huo et al., 2019; Shimizu et al., 2019; Viana-Sotoa and Senf, 2024).

4.2.1.3 Strengths and weakness of existing methods

The main strengths of the existing methods are the relative simplicity of the used methods. Using only annual summer composites of Landsat data to classify disturbance agents keeps down the amount of data which needs to be processed. Additionally random forest is a very well understood model, which is also comparatively quick to train on large amounts of data. Another strength are the features which are included in the model. The use of spatio-temporal features increases the accuracy of the models.

Weaknesses of the model stem from the used data. Landsat data is coarse compared to Sentinel 2, thus making small scale and non-stand replacing disturbances harder to detect. Additionally, the use of annual summer composites does not allow rapid attribution of disturbances, which is a key requirement of this project. Finally, the way that training data was collected through photo interpretation of Landsat scenes did not allow accurate differentiation of disturbance agents. Thus, only allowing for coarse modelled agent classes.

4.2.2 Method description

Based on the available literature we derived an appropriate method to test the potential of candidate features and statistical models to differentiate disturbance agents based on satellite imagery.

A priority was that the used ground truth data should be labelled either through ground surveys or through ancillary data like the EFFIS fire database. This gave greater confidence into the labelled disturbance agents. Additionally, the data was labelled for the entire Sentinel 2 time series, thus allowing the training and modelling of rapid disturbance agent classification.

For the training data, input features were acquired. For the input features, two different approaches were taken. On the one hand, it was tested how well spectral features (i.e. the different spectral bands of Sentinel 2) without any added spatial or temporal context would perform in the classification of disturbance agents. The other feature set that was tested were the embeddings of the Clay geospatial foundation model. For this, 64x64 pixel satellite image chips, centred on the

disturbance were used to generate the embeddings. These embeddings should contain spatial context, which could potentially aid in the classification of disturbance agents.

For both feature sets (spectral and embeddings), a hyperparameter search was carried out on three model architectures. The tested models were LightGBM, Random Forest and Support Vector Machine (SVM).

Additionally, only for the spectral features, a model gaussian mixture model was fit. This was only done on the spectral features, since this model architecture does not work well for such high-dimensional data as is provided by the embeddings.

Finally, the accuracies of all trained models were compared using a held-out test set.

4.2.2.1 Input data and study sites

The ground truth data used for training and testing was collected through ground surveys and human labelling with ancillary data.

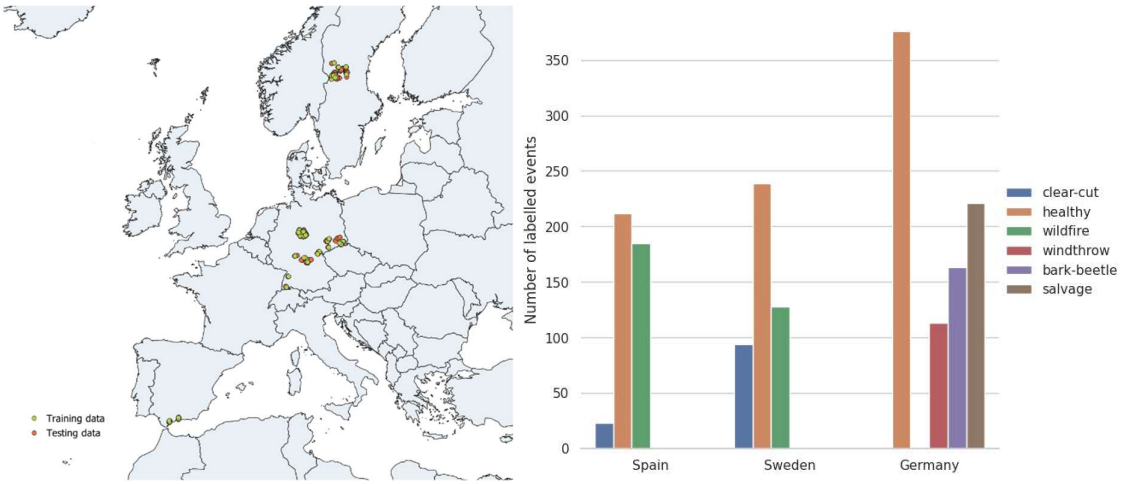


Figure 108. Spatial test and training data distribution (left) and number of unique labelled events per class (right).

In particular, most ground truth samples are from ground surveys of German forest services, these samples were delivered as disturbance polygons. The German dataset contains all the bark beetle, windthrow and salvage logging samples. Since that dataset was missing wildfire samples, it was augmented with two study areas in Spain and Sweden, which were identified through the EFFIS forest fire database.

For all of these samples, points within the provided disturbance polygons were sampled and the Sentinel 2 time-series was classified into segments based on changes in the spectral values and the recorded ground truth data. For example, a point sample in Germany could have three temporal segments: healthy forest, windthrow, salvage logging.

To validate the results of the models, a validation set was held out from the full dataset. This validation dataset includes 10% of samples in each of the classes (clear-cut, healthy, wildfire, windthrow, bark-beetle, salvage). To reduce the amount of spatial autocorrelation, the split into validation and training dataset was along the polygon ids of the original polygons which were delivered by the forest services. To also reduce possible overfitting due to the spatial context, the training and validation dataset were also split along the segments. This means that for a single unique polygon id which might have healthy, bark-beetle and salvage logging temporal segments, all healthy and bark-beetle timestamps might be part of the training dataset, but all salvage logging segments could be part of the validation dataset.

4.2.2.2 Description of the candidate methods

There are 6 candidate methods in total. They are made up of two different feature sets and 3 different models each.

The two different feature sets are geospatial foundation embeddings and pixel-based spectral features.

The selected spectral features are 10 bands of the Sentinel 2 sensor (Band numbers in parentheses): Blue (2), Green (3), Red (4), Red Edge (5-7), NIR (8), Vegetation Red Edge (8A), SWIR (11, 12). All bands were used at 10 meters resolution; bands that are at a native 20m resolution were resampled to 10 meters.

For the embeddings, the Clay geospatial foundation model was chosen due to its ability to generate embeddings on a single time step. The embeddings were generated on a 64x64 pixel image chip centred on the disturbance. The Clay model always generates 1024 embeddings per chip.

The three models that were evaluated are Random Forest, LightGBM and Support Vector Machine (SVM). These models were chosen due to their known performance in multi-class classifications and robustness to overfitting. This robustness against overfitting was especially important in the case of the embedding features, since 1024 features have the potential to uniquely encode every single sample in the dataset.

4.2.2.3 Benchmarking & evaluation

All combinations of models and features were trained through a hyperparameter search. The hyperparameter search randomly searched a predefined parameter space for the different models and picked the best performing hyperparameter set based on a 5-fold cross validation. The chosen performance metric to optimize was the macro F1 score. The macro F1 score is defined as the mean F1 score for each individual class.

All the training steps and are openly available on GitHub.

4.2.3 Implementation and results

The best performing F1 scores of each model combination are shown below. These are the values based on the inference of the validation set.

Table 44. F1 scores for each model combination.

	Embedding Macro F1	Spectral Macro F1
LightGBM	0.78	0.83
Random Forest	0.62	0.82
SVM	0.82	0.08
Gaussian Mixture	-	0.74

In general, both features perform reasonably well in separating the different disturbance classes. However, the decision tree-based models (LightGBM, Random Forest) perform much better using the spectral features than using the Embeddings as features. In the case of SVM, the embeddings lead to much better results. Overall, the best model is LightGBM using spectral features, however SVM and Embeddings achieve a very similar in overall performance. The Gaussian Mixture model also achieves quite good performance.

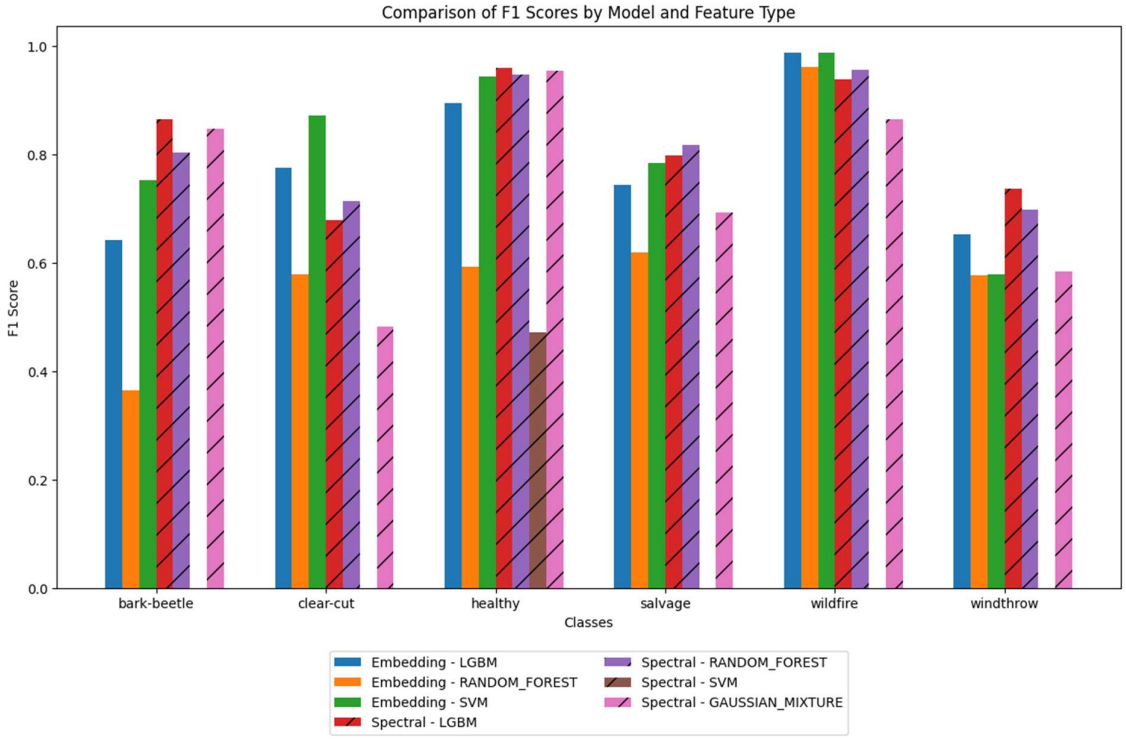


Figure 109. F1 scores for all classes and each model and feature combination

Looking at the separate F1 scores for each class and for each model we can see some differences. In general, the wildfire class is the most confidently predicted

across all models. It can also be noted that the models using spectral features are better at discerning bark-beetle and windthrow areas than the models using embeddings. Using embeddings, clear-cuts are classified with a higher confidence. For healthy and salvage samples the models using spectral features and embeddings are performing similarly. For the Gaussian Mixture model, the least confident classes are clear-cut, salvage and windthrow.

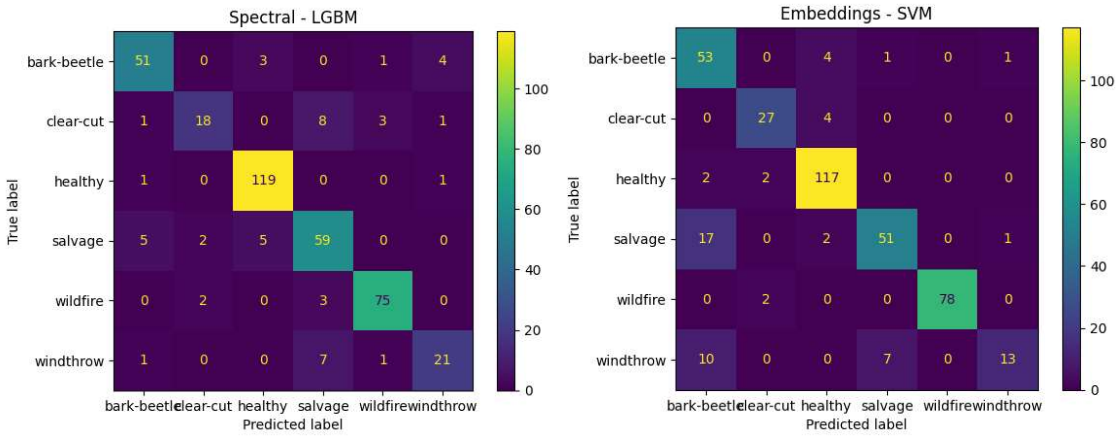


Figure 110. Confusion matrices for the best performing models for the two different input feature types.

The confusion matrices for the best model using spectral features (LGBM) and the best model using embeddings (SVM) confirm the different patterns in accuracy in the different disturbance agent classes. For the SVM model using embeddings, the main source of low confidence in the bark-beetle and windthrow classes are windthrow sample which are falsely classified as bark-beetle. Additionally for the bark-beetle class there is some confusion with the salvage logged classes. With 17 salvage samples classified as bark-beetle.

An error mode in the spectral model, which does not appear in the embeddings model is confusion of clear-cut samples with salvage logged samples. Intuitively this difference makes sense due to how the features are generated. Salvage logging and clear-cutting likely have very similar spectral signatures due to similar amount and mode of biomass removal. However, embeddings might consider some spatial information like the shape of removal, or if there are naturally disturbed areas close to the salvage logged areas. This additional information is missing from the pixel based spectral features but can increase the ability to differentiate clear-cutting from salvage logging.

Looking at a few examples of inference for the LightGBM model using Spectral features, we can see that the output can be quite noisy. It can also be noticed that the classification is affected by edge effects, with the transition from healthy forest to clear cut or salvage logged forest often identified as windthrow. Overall, for a naïve pixel-based approach the results are very promising. The noisiness of the classification could easily be lessened by applying a 3x3 median filter.

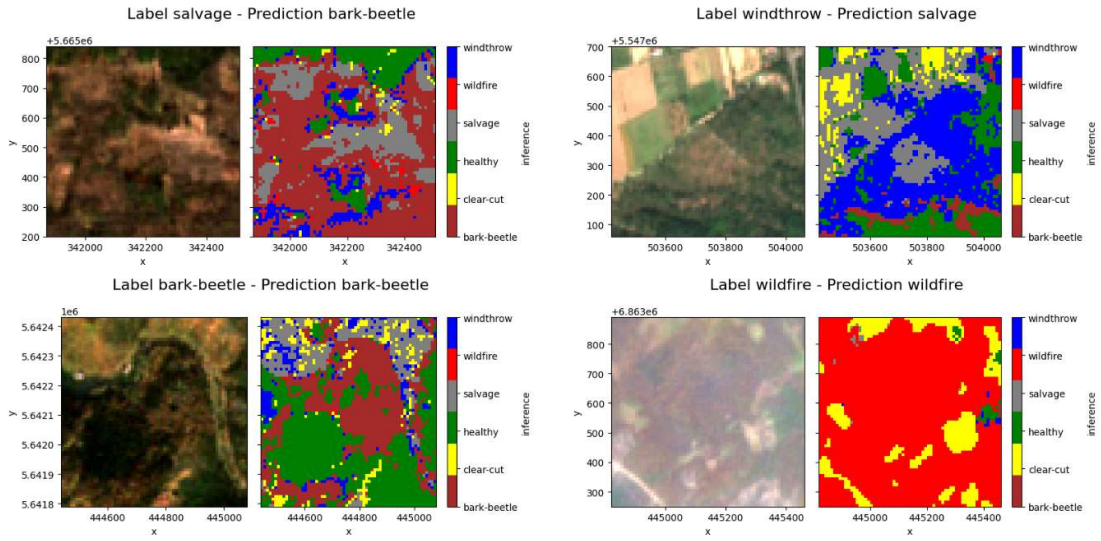
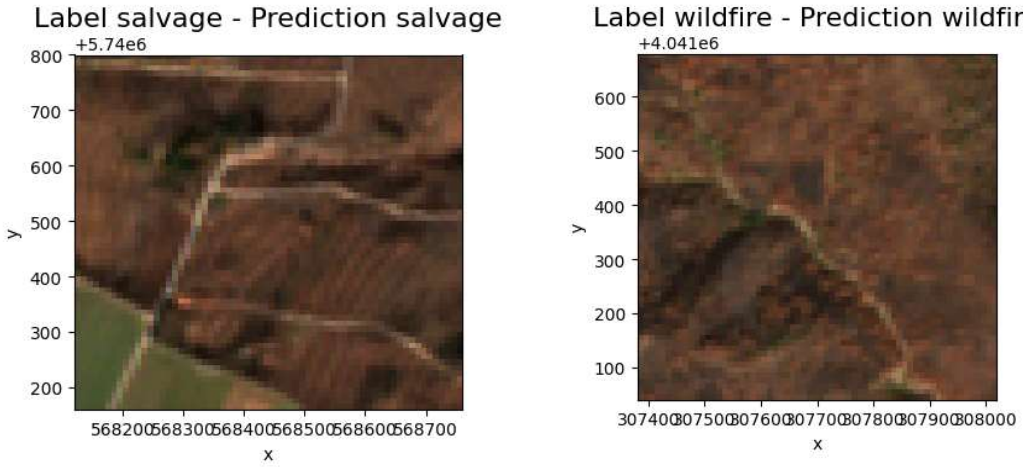


Figure 111. Examples for inference using best performing LightGBM model and spectral input.

Looking at samples of inference from the SVM model using embeddings it is possible to see that the classification mostly struggles with smaller scale, less severe disturbances. This is likely due to the embeddings being generated for the entire chip, but only the point at the centre of the image chip being used for the label and which should be predicted. In cases where this sample point only encompasses small disturbances this signal from the centre of the image could be drowned out by signal from the rest of the image chip.



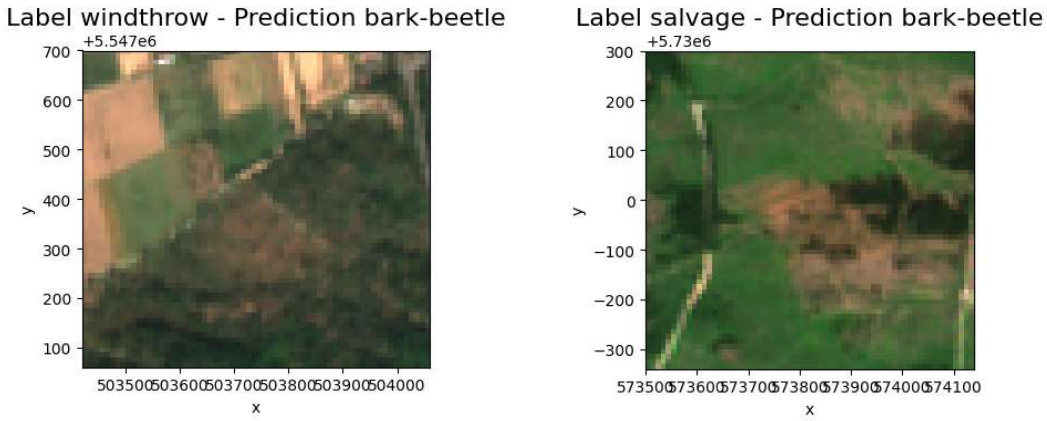


Figure 112. Examples for inference using best performing SVM model and embeddings input.

For the inference with the Gaussian Mixture Model it can be seen that as the accuracy in the windthrow class is indeed not very good. However, for other major classes, like wildfire or healthy but also bark-beetle the overall performance seems to be quite reasonable.

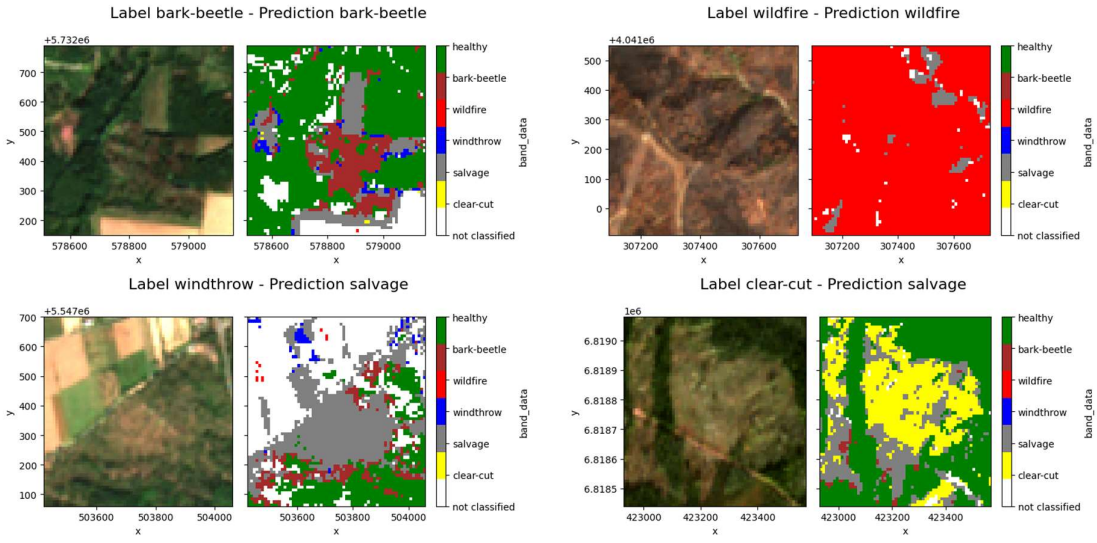


Figure 113. Examples for inference using Gaussian Mixture model and spectral input.

There are several options to try and decrease this influence. It could be tried to mask out areas which aren't forest or aren't disturbed, before generating embeddings. Another approach could be a fine-tuning of the geospatial foundation model. The model architecture can be fine-tuned to attend to the middle of the image more than the edges.

4.2.4 Summary and conclusions

The comparative analysis of spectral features and geospatial foundation model embeddings, coupled with three distinct machine learning architectures, provides critical insights into the classification of forest disturbance agents using Sentinel-2 data. The LightGBM model trained on spectral features achieved the highest macro F1 score (0.83), demonstrating strong performance across most disturbance classes, particularly for bark beetle, windthrow, and wildfire. Its pixel-based approach enables rapid, large-scale application with minimal computational overhead, aligning with the project's requirement for timely disturbance attribution. However, limitations persist, including edge effects and spectral confusion between salvage logging and clear-cutting, which arise from the lack of spatial context in pure spectral analysis.

In contrast, the SVM model using embeddings from the Clay geospatial foundation model achieved comparable performance (macro F1 = 0.82) but exhibited distinct strengths, such as improved differentiation of clear-cuts from salvage logging, likely due to implicit spatial patterns captured in the 64x64px image chips. This suggests that spatial context—such as disturbance shape, proximity to other land-cover features, or edge patterns—can resolve ambiguities that spectral data alone cannot. However, the embeddings-based approach struggled with small-scale disturbances, as the central pixel's signal was often diluted by surrounding pixels in the chip.

In general, the amount of ground truth data used in the study also is not quite sufficient to train a generalizable method for all of Europe. Lack of training and validation data in the domain of disturbance agent classification is a known issue, which was noted in other research.

In conclusion, while no single method universally outperforms others, the results underscore the feasibility of automated disturbance agent classification using Sentinel-2 data.

4.3 Change detection – AI methods

4.3.1 State of the art

4.3.1.1 Introduction

Change detection in remote sensing is key for, e.g., monitoring ecosystems, urban expansion, disaster assessment, and agricultural development. The advent of deep learning introduced a new tool to analyse remote sensing information automizing detection of changes on Earth's surface. Deep learning-based change detection leverages the power of neural networks to process and analyse vast amounts of multispectral, radar, and hyperspectral data offering an approach to identifying changes across diverse landscapes and timeframes (Figure 114).

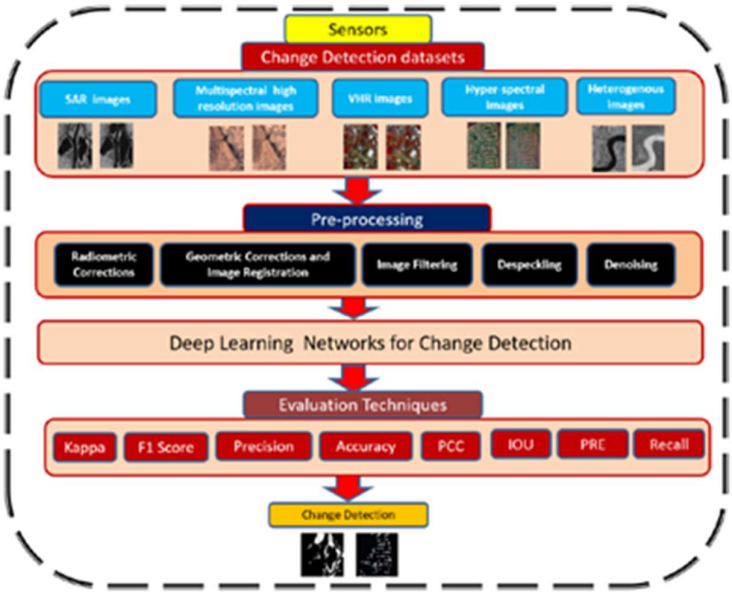


Figure 114. Typical workflow of change detection as depicted in (Shafique et al., 2022)

4.3.1.2 Review of existing methods

Existing deep learning-based change detection methods can be categorized according to different taxonomies, such as algorithm granularity, supervision level, and fusion structure. Below, we closely follow and summarize the comprehensive reviews presented in (Cheng et al., 2023a) and (Parelius, 2023). We highlight the most relevant publications, and their respective pros and cons. Readers interested in additional references and technical details related are advised to explore the cited publications.

Spatial Granularity

Based on algorithm granularity, change detection methods are either pixel-based, region-based, or hybrid methods:

- *Pixel-based* methods assign a label to each individual pixel of an image based on its spectral characteristics. The goal is to partition the image into regions of different classes. (Vinholi et al., 2022) presents an incoherent SAR change detection algorithm based on convolutional neural networks which includes a segmentation network to localize potential changes, and a classification network to further analyse the potential changes to classify them as real changes or false alarms. (X. Zhang, Tian, et al., 2022) proposes an attentive differential high-resolution change detection network, introducing a high-resolution backbone with a differential pyramid module. In contrast, (J. Liu et al., 2018) develops a symmetric convolutional coupling network with a symmetrical structure and feature transformation for heterogeneous optical and radar images.

- *Region-based* methods or object-based methods leverage image segmentation techniques to group pixels into meaningful regions based on their spatial, spectral, and contextual characteristics. These grouped regions are used as units for detecting and recognizing the changed results. (Amitrano et al., 2021) exploits multitemporal geographical object-based image analysis for SAR change detection. This approach involves classifying the data using a dictionary representation approach and combining class-specific information layers through fuzzy logic to unveil the context. (Ge et al., 2022) proposes an object-oriented change detection approach that integrates spectral-spatial saliency change information and fuzzy integral decision fusion to eliminate the detection noise. Coming from another angle, (J. Wu et al., 2021) explores how to combine graph convolutional network and a multiscale object-based technique for homogeneous or heterogeneous images.
- *Hybrid* methods involve the parallel or successive use of pixel-based and region-based techniques to detect changes from different perspectives. (Javed et al., 2020) proposes an approach for detecting recently developed urban areas in SAR imagery by utilizing the Dempster-Shafer theory to extend the results of the pixel-based method. (Jiang, Xiang, et al., 2022) introduces a dual-pathway change detection network that combines a dual-pathway feature difference network, an adaptive fusion module, and an auxiliary supervision strategy. (J.-J. Wang et al., 2024) introduces a deep adversarial network to fuse a pair of multiband images. This approach is easily complemented by a network of the same architecture to perform change detection.

In summary,

- *Pixel-based* methods shine through their ease of implementation, computational efficiency, and ability to identify small changes at high resolution. These qualities render them suitable for detecting changes in homogeneous regions. However, these methods are not tailored to detect changes in heterogeneous regions, they do not entail spatial correlations, and they are sensitive to noise.
- *Region-based* methods are robust to noise and set up to detect changes in heterogeneous regions. Region-based approaches can identify changes in shape and size, making them well-suited for detecting changes in regions with complex spectral characteristics. However, they have disadvantages such as: computationally expensive, their performance is sensitive to the choice of the segmentation algorithm picked, and they may render inefficient in detecting small changes that may require domain expertise to select appropriate segmentation parameters.
- *Hybrid* methods combine the strengths of pixel-based and region-based methods, making them effective in detecting small and large changes in complex scenes to stay robust to noise and to be suited for detecting changes in heterogeneous regions. However, they have limitations such as demanding computational requirements, they are sensitive to the choice of

a segmentation algorithm, and they are potentially inefficient in detecting changes in regions with complex spectral properties - which may require a high degree of expertise to implement.

Mode of Supervision

From another perspective, existing change detection algorithms can be categorized according to unsupervised learning, semi-supervised learning, and supervised learning:

- *Supervised* change detection methods rely on labelled datasets to train models that can identify changes between images. This approach is straightforward to implement with readily available labelled data, but it requires a large amount of accurately annotated data. Models may struggle with biased data at inference when the data distribution is different from that of the training data. Such a scenario may lead to over- or underfitting. (G. Chen et al., 2023) proposes a Stockwell scattering network for SAR change detection that combines a wavelet scattering network and a Fourier scattering network. (X. Zhang, Yue, et al., 2022) proposes the change detection network *DifUnet++* for multispectral imagery which takes a differential pyramid of two input images to incorporate a side-out fusion strategy for predicting change. (Lv et al., 2022) introduces a UNet model for heterogeneous change detection employing a multiscale convolution module backbone to cover the various sizes and shapes of objects to detect.
- *Semi-supervised* methods address the challenge of limited labels by combining a small amount of data annotations with a larger pool of unlabelled data. These methods leverage the labelled data to guide the learning process while extracting useful information from the unlabelled data at the same time. (C. Wang et al., 2022) proposes a reliable contrastive learning method for semi-supervised change detection by selecting reliable samples according to the prediction uncertainty of unlabelled images and introducing a contrastive loss for optimization. (Jiang, Li, et al., 2022) proposes a semi-supervised Siamese network based on transfer learning. It separately treats low- and high-level features.
- *Unsupervised* change detection does not require labelled data making them particularly valuable for applications where annotations are scarce or hard to come by at all. Typically, those methods employ clustering, deep autoencoders, or self-supervised learning techniques to distil representations of the data for change detection. (W. Zhang et al., 2022) proposes a sparse feature clustering network for detecting changes in SAR images which is pre-trained with a multi-objective sparse feature learning model. (Saha et al., 2021) presents a SAR change detection method based on cycle-consistent, generative adversarial network (*CycleGAN*). (Y. Chen & Bruzzone, 2022) proposes a pseudo-Siamese network that is trained to obtain pixel-wise representations to align features from shifted image pairs. (Bergamasco et al., 2022) leverages multi-resolution deep feature maps obtained from a convolutional autoencoder. Furthers, (Jia et al., 2022) builds

a bipartite adversarial autoencoder with structural self-similarity for heterogeneous remote sensing images. A structural consistency loss is introduced to transform remote sensing images into a common domain. An adversarial loss assists in translating images consistently. (Luppino et al., 2022) explores two network architectures trained with loss functions weighted by priors.

In summary,

- *Unsupervised* learning methods excel in detecting changes without labelled data accommodating diverse data sources. However, their inability to distinguish true changes from noise restricts their performance in complex environments.
- *Semi-supervised* learning methods leverage labelled and unlabelled data to achieve improved accuracy compared to unsupervised methods. On the contrary, their success relies on the quality and quantity of labelled data.
- *Supervised* learning methods can achieve high accuracy in detecting changes with enough labelled data, but their inflexibility and dependence on labelled data can pose limitations in scenarios with complex patterns and scarce labelled data.

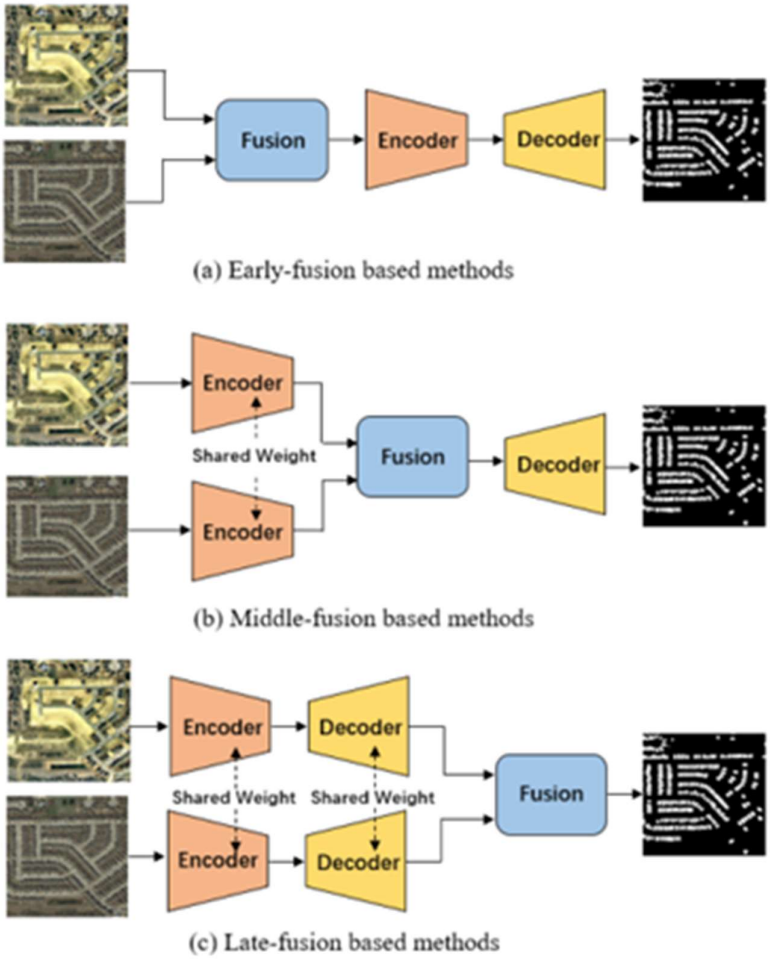


Figure 115. Change Detection fusion strategies as presented in (Cheng et al., 2023b).

Fusion Strategy

Based on where fusion enters the deep neural network (see Figure 115) change detection frameworks can be classified into early-fusion, middle-fusion, and late-fusion:

- *Early fusion* methods combine pre- and post-change data at the input layer, allowing the network to jointly learn from all available data. This approach models the interaction of different data types early on, but it may turn out challenging to implement. Preprocessing of the data input requires spatial alignment and normalization. (Caye Daudt et al., 2019) proposes a modified FC-EF UNet with added residual blocks. It combines several UNets to simultaneously perform land cover mapping and change detection. (Peng et al., 2019) constructs a UNet-style model that fuses information from various levels through up-sampling and dense skip connections. The output of such intermediate levels can be combined for predicting change.
- *Middle fusion* techniques integrate features from pre- and post-change images at intermediate layers of the artificial neural network. Common

fusion strategies allow the model to independently process each data source in the initial layers to leverage individual features for each type of data before all those get combined. Based on (Fang et al., 2022; Z. Zhou et al., 2020) uses compact information transition between the encoder and the decoder by adding up-sampling modules and skip connections between corresponding semantic levels. Similarly, (H. Zhang et al., 2023) features a UNet subnetwork for change detection. Two Siamese super-pixel sampling networks are used to extract features and perform super-pixel segmentation on the input images. Resulting features are fed to a UNet-like network, and the information about the super-pixels is used to reduce the noise and improve edge identification.

- *Late fusion* methods involve merging the outputs of separate model branches. This approach enables independent feature extraction and decision-making for pre- and post-change data, with the final change detection decision based on the aggregated outputs. Late fusion is not commonly used in supervised learning, but it is popular interest in self-supervised learning where the model separately encodes each image to subsequently detect changes from the (in)variance of encoded images in embedding space (Saha et al., 2022a).

In summary,

- *Early fusion* methods initiate the integration process at the input layer, enabling a model to capture early interactions between different data types to recognize complex change patterns. However, early fusion requires meticulous preprocessing to align and normalize data sources.
- *Middle fusion* techniques merge features from two samples in which to identify change at intermediate layers. This strategy allows for independent processing of each data source to extract unique features before fusion. The approach enables a more nuanced understanding of changes by combining the strengths of each temporal snapshot. Middle fusion is popular due to its balance between flexibility and feature integration depth.
- *Late fusion*, though less common in supervised learning contexts, proofs relevant in self-supervised setups by merging model outputs. This method supports feature extraction for each individual timestamp of an input to allow change detection decision-making for all combinations of inputs for comprehensive change detection pipelines.

4.3.1.3 Strengths and weakness of existing methods

Besides what we have alluded in section 4.3.1.2 when summarizing change detection methodologies from the angles of spatial granularity, supervision modes, and fusion strategies, below we summarize high-level pros and cons of when deep learning techniques get applied for remote sensing change detection:

Strengths

1. Deep learning-based methods excel in extracting hierarchical and complex features from remote sensing images, significantly outperforming traditional techniques in detecting subtle and complex changes.
2. Deep learning-based methods are adept at processing large-scale, high-dimensional multispectral, hyperspectral, and SAR data, enabling effective change detection across a variety geospatial scenes.

Weaknesses

1. Despite the advancements in semi-supervised and unsupervised learning, many deep learning models still heavily rely on large quantities of labelled data expensive to collect.
2. Deep learning models, especially those based on large backbones and multi-stream architectures, require significant computational resources for training and inference, posing challenges for applications in need of rapid response and decision making.
3. Models trained on biased datasets limited to, e.g., certain geographic regions may not perform well when applied to other regions or when applied to imagery with vastly different environmental conditions. Correspondingly, re-training for model fine-tuning is mandatory.

4.3.2 Method description

Given the extensive research on self-supervised learning with unsupervised pre-training for single-timestamp remote sensing imagery presented in section 2.2, we detail an unsupervised methodology for change detection (***SSL4Change***). Here, the task of detecting changes from remote sensing imagery is based on AI model-generated embeddings. The methodology does not require labels and it operates on as little as two distinct timestamps (before and after change) avoiding the need to infer change from a full time series of potentially cloud-obscured imagery.

4.3.2.1 Input data and study sites

To assess *SSL4Change* we evaluate the methodology on multi-spectral Sentinel-2 imagery (no atmospheric correction, L1C product) for detecting changes in German forests. Domain experts at GAF provided their expertise by visual inspection to identify areas where forest change such as clear cuts have been observed, cf. section 4.3.2.3 (Table 45).

Table 45. Overview on data utilized for change detection experiments with GAF.

Dataset	Description	Spatial Resolution	Temporal resolution
Sentinel-2 (L1C) tiles	GAF-specific pick with domain expert knowledge on forest change	10m	Dual timestamp

4.3.2.2 Description of the candidate method



Figure 116. General Artificial Neural Network design of the SSL4Change methodology separately convolving two remote sensing modalities (SAR, Optical) to a common embedding ($Conv5^{common}$). Our approach adapts the architecture for two Optical Sentinel-2 input at timestamps T1 and T2, i.e. "SAR input" is replaced by a 2nd "Optical input".

SSL4Change is based on (Saha et al., 2022b), adapted for multi-spectral imagery processing (cf. Figure 116) with reference implementation provided by open-source Python package `evoland` hosting[13]:

- `ChangeDetectionModule` Python (PyTorch Lightning) model class (`evoland.models`) with
- sample model weights for (German) forest change detection, `test/change_model.pth`, and
- corresponding `ChangeDetectionDataModule` PyTorch Lightning class (`evoland.data`) for data loading and model output saving.

As SSL4Change is an unsupervised methodology optimized without labels to model the notion of "**change**", it requires an algorithmic definition of the term, that is: Given two spatially aligned snapshots at times T1 and T2 of the same geospatial scene, for each pixel pair values p at location x, p(x,T1)=p1 and p(x,T2)=p2, SSL4Change assumes no change. Hence, it is key to train an SSL4Change model on pairs of scenes without any change that one would like to detect when the model predicts (inference).

Moreover, to fine-tune the model to a specific type of change such as "forest change", one may want to restrict model training by a forest landcover mask, if available. In addition, normalizing the multi-spectral imagery values p to the scene's mean and standard deviation may help to stabilize change detection results by avoiding artifacts based on bare differences of p1 and p2 due to atmospheric corrections (L2A products) or atmospheric variance (L1C products).

As experiments indicate, SSL4Change works best when T1 and T2 are close in time avoiding, e.g., seasonal changes. This setup fits EvoLand's general objective of enhanced temporal resolution for change detection to get as close as two consecutive (cloud-free) Sentinel-2 tiles covering a given geolocation x, cf. Figure 117. Sect. 4.2.2.3 reports on our findings on practical limitations with a fully unsupervised change detection prototype such as SSL4Change.

[13] https://github.com/Evoland-Land-Monitoring-Evolution/evoland-software/tree/main/src/evoland



Figure 117. Sentinel-2 inputs (before, T1 and after, T2) to SSL4change and corresponding change maps for an area in mountain range Taunus west of the German city of Wiesbaden.

The basic concept of SSL4Change is rooted in the self-supervised contrastive learning paradigm:

1. **Data preparation**: Two separate encoders E1 and E2—one for the Sentinel-2 remote sensing imagery I1 at time T1, and one for the Sentinel-2 imagery I2 at time T2—convert the input raster data cube of dimension PxBxWxH to an embedding data cube of dimension PxFxWxH. W and H represent the number of spatial pixels for image *W*idth and *H*eight, respectively. B denotes the number of Sentinel-2 *B*ands, and F is the dimension of the embedding *F*eature vector. P indicates the number of non-overlapping *P*atches with B spectral bands of size WxH that have been randomly picked from a Sentinel-2 tile. E1 and E2 are feed-forward artificial neural networks stacking a series of two-dimensional convolutions, batch-normalizations, and ReLU activation functions.
2. **Model training**: The embeddings Z1=E1(I1) and Z2=E2(I2) get utilized to optimize the parameters of the encoders E1 and E2 as follows:
    a. *pseudo-clustering:* a cross-entropy loss incentivizes the encoders E to produce a single dominant component in the embeddings Z=E(I) of dimension F. This way each Sentinel-2 patch effectively gets classified in analogy to a K-means algorithm with K=F.
    b. *temporal persistence*: bring closer the two embeddings Z1 and Z2 for a given patch p=1…P by reducing their distance (L1-norm). This loss entails that two patches of size WxH covering the same spatial area at two timestamps T1 and T2 are assumed to show no change.
    c. *spatial contrast:* Maximizing the distance of two patches that do not geospatially align avoids model collapse by which the encoders E1 and E2 always generate one single embedding Z1=Z2=Z=const. in order to optimize for temporal persistence. In practice, SSL4Change randomly shuffles the patches of I2 to I3 dissolving the spatial alignment with I1 before encoding, i.e., the distance of Z1 and Z3=E2(I3) gets maximized.

3. **Change prediction**: Once the parameters of E1 and E2 are numerically optimized by iterating over the data multiple times (epochs), change for two spatially aligned Sentinel-2 imagery I1 and I2 is obtained by threshold value C through: | E1(I1)-E2(I2) | > C generating a single-layer, binary geospatial image of dimension 1xWxH with the aid of the L1-norm | . |. As reported below, C is a critical hyper-parameter determining the accuracy of SSL4Change.

4.3.2.3 Benchmarking & Evaluation

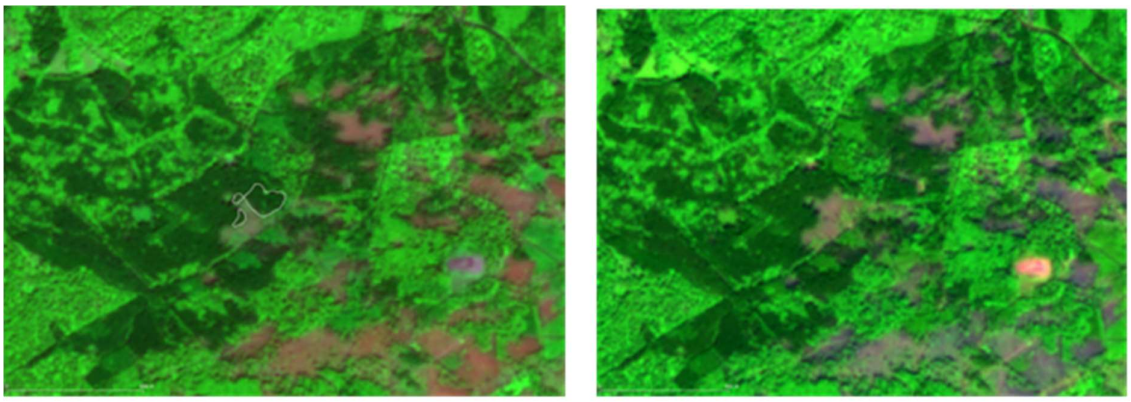


Figure 118. Sentinel-2 sample snapshots of a forest area in Germany (left: before, right: after) as demonstrator of our unsupervised change detection methodology. The forest that has been cleared from left to right is enclosed by white borders close to the image centre.

Towards a potential forest change detection prototype that does not require manual labelling as supervised input, we applied SSL4Change to predict forest changes in Germany guided by GAF experts, referred to here as *SSL4ForestChange*. Figure 118 depicts two timestamps of a sample scene separated by about a month in summer 2019. The centre of the imagery exhibits clear-cut; the remaining area does not involve any change in forestry.

A general barrier to evaluate SSL4ForestChange at scale is lack of accurate ground truth. The bare notion of what constitutes a **change** is not trivial and requires a careful definition. While the direction of a change can be formally identified with the difference of embedding vectors Z1-Z2, extraction of a label “gain in forest“ or “loss in forest” is non-trivial. It requires labelled sample data to define such terms. A data-driven, unsupervised approach ships with the opportunity of no human labour in collecting annotated data on the one side, but it buys the challenge on interpretation and control of results on the flip side.

Based on benchmarks with the Onera Satellite Change Detection dataset (Daudt et al., 2018), in urban areas, SSL4Change reports a sensitivity (relative value of true positive when compared to all positive samples) varying from as low as 36% up to about 50%. These numbers are competitive with other unsupervised methodologies. However, rural changes in natural environments may exhibit lower signal-to-noise ratios. In fact, as our experiments report below, there exist limits in

converting SSL4ForestChange into a full-fledged, industry-standard prototype ready for production.

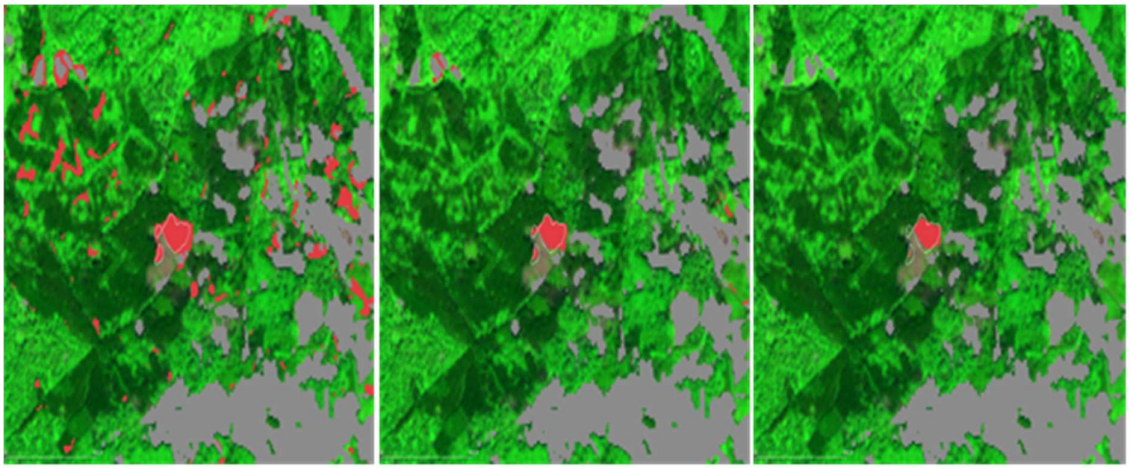


Figure 119. Detection of change (red) on demo scene based on contrastive unsupervised AI model SSL4ForestChange depending on threshold hyperparameter C. The background is filled by the Sentinel-2 data after the change. Known areas without forest are visually blocked in grey.

Figure 119 provides a visual example of how SSL4ForestChange performs when varying the threshold parameter for the two sample timestamps from Figure 118. The higher the numerical value of C is tuned—i.e., the farther corresponding pixel embeddings of the Sentinel-2 images before (I1, timestamp T1) and after (I2, timestamp T2) need to separate to qualify as change in terms of the L1-norm—the less change is detected. Domain expert guided experiments confirmed that tuning C by hand is not trivial. It is challenging to identify a trade-off between commission and omission errors: As Figure 119 demonstrates, a low value of C=2 (left) detects too much of change (red areas), and the scenario to the right with C=6 omits areas of real forest change. Although the compromise C=4 (centre) reasonably covers the forest change, fixing C implied poor performance of SSL4ForestChange on other areas.

As we alluded at the beginning of Sect. 4.2.2.2, there exists a variety of fine-tuning options to improve results. If there are available masks to indicate where no forest is present, e.g., due to urban infrastructure, masking away these pixels at inference does refine the SSL4ForestChange prototype “by definition in data”. However, training SSL4ForestChange on forest areas employing a corresponding mask, did not improve model performance for the forest change scenarios picked.

Similarly, there is an option to auto-adapt the threshold C in case the percentage of change, q, is known for one given scene: Assuming that the most prominent change is reflected by the largest L1-norm, one can order the set of all N pixel change proxy values c=| E1(I1)-E2(I2) | from largest to smallest. Starting from the largest c0>c1>…>cM>…>cN down to the smallest cN, it holds when (M-1) < q*N < M, then C can be estimated to the value of cM. Of course, the choice of C is assumed to hold valid for other scenes, too. As reported above, the assumption of a constant

C across diverse sets of scenes may be too strong for change detection in forest use cases.

4.3.3 Implementation and results

Th small-sized neural network architecture with 5 convolutional layers generating 64 feature layers each before reducing to a F=4-dimensional embedding space should enable training of SSL4Change models on most commodity hardware as of 2024. If a GPU accelerator is available, optimizing a model by stochastic gradient descent on a 2kx3k = 6 million pixels scene with B=12 bands utilizing 8 workers over 10 epochs (sufficient for model convergence) at a learning rate of 0.001 cropping WxH=128x128 pixel patches stacked in batches of P=8 requires an NVIDIA A100 GPU about 5 minutes. As demonstrated by the `evoland` code package provided with this report, it takes about 30 seconds to perform model inference employing Lightning trainer functionality:

```
evoland.models.ChangeDetectionModule.predict()
```

As highlighted above, the definition of **change** is encoded in the assumptions listed in section 4.3.2.2, Model Training, items 2a-2c. These compensate for the lack of labels. Further, we note that change detection results employing SSL4Change are sensitive to the training data provided. Additional data augmentations such as randomizing brightness and contrast of spectral bands may mitigate such sensitivities.

Given the relatively inexpensive model training,

- Semi-automated selection of the threshold hyper-parameter C for fine-tuning model inference may be automatized on scenes where there exists ground truth on the percentage of forest change, which is unlikely available on operational scale (cf. our discussion at the end of section 4.3.2.3). However, this approach requires a relatively constant C across scenes.
- Defaulting to histogram-based heuristics for determining an estimate of C is another option for “on-the-fly” regional models.
- Running a batch of models with various artificial neural network parameter initializations to pick the best based on some ground truth available is yet another strategy to utilize a limited amount of data.

4.3.4 Summary and conclusions

SSL4Change provides an unsupervised, multi-modal, contrastive change detection methodology relatively inexpensive to run on commodity hardware. We make it available as PyTorch Lightning class

```
evoland.models.ChangeDetectionModule
```

in the open source `evoland` Python package.

Initial experiments with GAF domain experts on the forest change detection use case revealed sensitivity of the SSL4ForestChange model on training data and the

choice of thresholding for the binary change maps from embedding space. The former may be mitigated by additional data augmentations, cf.:

```
evoland.data.changedetection.Random{Brightness,Contrast}
```

For the latter there exist techniques for semi-automated adjustments given a limited number of labels (cf. section 4.3.3). However, following our benchmarking in section 4.3.2.3 above, we rate SSL4ForestChange not the most promising approach for a production-ready forest change detection layer.

Benchmarking results on urban areas indicate that change detection accuracies are not expected to reach sensitivity values above 50%. In fact, one central assumption of SSL4Change is that pairs of training images over a given area should not contain any relevant change to meaningfully indicate a change at inference time. However, compared to urban areas, natural ecosystems like forests may contain complex fluctuations in spectral and spatial patterns. Those contribute to the confusion of the SSL4ForestChange model, degrading its performance.

Nonetheless the above limitations, the intermediate product of contrastive change based on the L1-norm | E1(I1) - E2(I2) | may provide a proxy for the likelihood of change in forest when combined with ground truth masks such as land cover.

Given the extensive development of (patch-level) single-timestamp, self-supervised methodologies in section 2.2.2 (SSL4EO, DeCUR, and CromSS), we focused on improving the granularity of semantic segmentation downstream tasks in phase 2 of the EvoLand project. For the remaining period of the project, we have options to explore (i) “patch-level change detection” (per image, not per pixel) and (ii) change-type identification combining SSL4EO patch-level image classification with disturbance-labelled pixel-based change detection results. However, both avenues will require a decent amount of accurate ground truth labels for satisfactory prototype performance,

4.4 Continuous mapping of Land Surface Categories

4.4.1 State of the art

4.4.1.1 Introduction

Most operational global land cover datasets, such as the Copernicus Global Land Service (CGLS) Land Cover product at 100m resolution, National Aeronautics and Space Administration (NASA) MCD12Q1 dataset at 500 m resolution, or the European Space Agency (ESA) Climate Change Initiative (CCI) land cover dataset at 300 m resolution, focus on classifying the Earth surface at a yearly interval. While they are an invaluable source of information for a suite of applications, including monitoring and modelling the Earth’s surface, their fixed legend system does not always serve the needs of all users. In addition, the yearly update interval does not

support the near-real time monitoring of fast changes (e.g. forest loss) or seasonal change (e.g. cropland rotations or changing water levels in ephemeral water bodies). Timely information on land cover and land cover change is however crucial to understand anthropogenic pressures on our environment, the state of ecosystems, *and their evolution*.

The near-real time mapping of a set of basic, independent diagnostic categories and characteristics describing the properties of the Earth's surface, called Land Surface Categories (LSC), may complement the information provided in the yearly land cover maps. For instance, continuous, near-real time information on the presence LSC has the potential to capture both land cover dynamics (e.g. forest loss) and to get information on seasonal dynamics (e.g. seasonal presence of water, snow, and ice). The dynamic nature of the LSC may support downstream applications, such as the production of land cover and land cover change products. For instance, temporal information on the presence of bare and herbaceous vegetation may support the classification of cropland and grassland on an annual basis. Other applications may extend to mapping of dynamics (e.g. the mapping of snow, water, or vegetation cycles) or the generation of user-defined classification systems (e.g., the mapping of ephemeral versus persistent water bodies based *on user-defined thresholds).*

4.4.1.2 Review of existing methods

A suite of methods has been developed to map individual LSC classes. For instance, several models have already been developed that allow to monitor intra-annual and inter-annual surface water dynamics (Pekel et al., 2016; Pickens et al., 2020). An intercomparison of a set of 14 inland surface water monitoring approaches using optical, SAR, or both data sources shows that these methods can achieve very high accuracies and yield the best performance when optical and SAR data are combined (Tottrup et al., 2022).

Several global products are available, summarizing water dynamics. The Global Surface Water Explorer (Pekel et al., 2016) classifies water occurrence, recurrence, seasonality and their change globally at 30m resolution over from the year 1984 to 2018. It is based on the combination of expert systems, visual analytics and evidential reasoning, applied on Landsat data. (Pickens et al., 2020) also mapped open inland surface waters from 1999-2018 based on Landsat and elevation data. They classified each scene into land, water, cloud, shadow, haze, and snow and ice using ensembles of classification trees. These scene-based classifications were subsequently aggregated to derive information on both intra- and interannual water dynamics and classify them in dynamic types.

Next to water monitoring, forest loss and loss and gain are commonly monitored. For instance, a range of near real-time change detection algorithms are already available, that are applicable to monitor forest change and loss (E. Brooks et al., 2017; Csillik et al., 2022; Verbesselt et al., 2012b) and near real-time forest loss alerts are being produced for the tropics (e.g. RADD and GLAD alerts (Hansen, Krylov, et al., 2016; Reiche et al., 2021).

Most methods that focus on mapping multiple classes simultaneously focus on annual updates. Yet, recently, a multi-class deep learning approach (a Fully Convolutional Neural Network) has been developed and implemented to monitor global scale land cover dynamics near real-time at 10m resolution (Dynamic World; (C. F. Brown et al., 2022)). Nine classes were covered in this product, including water, forest, grass, flooded vegetation, crops, shrub & scrub, built, bare ground, and snow & ice. Both the class probabilities and final classification are provided at global scale for scenes with a cloud probability lower than 35% since June 2015. The product was trained on polygon-based annotations and validated on 409 test tiles, resulting in an overall agreement of 73.8%. The highest agreement was found for the water, trees, built area, snow & ice classes, while the lowest agreement was found for the grass, flooded vegetation, shrub & scrub, and bare ground classes.

4.4.1.3 Strengths and weakness of existing methods

Although methods that focus on single classes are very relevant and may reach high accuracies, they are computationally less sustainable for an operational service that focusses on multiple classes. In addition, most operational products that focus on single (and multiple) classes work on longer time intervals, often yearly.

Dynamic World however provides near-real time updates of multiple classes, highlighting the benefits of near real-time land cover monitoring. Yet, the product has the drawback that it was trained using spatially coarse polygon-based annotations, resulting in a lack of fine-grained details in the maps. In addition, its nomenclature is not adapted to continuous monitoring. The nomenclature for instance includes a cropland class, which is characterized by temporal alternation between herbaceous vegetation cover and bare cover (after harvest). When classifying single scenes, this temporal information is not available. This results in confusion between grassland and cropland when the field is vegetated and confusion between bare land and cropland after harvest.

4.4.2 Method description

The general purpose of this subtask is to develop a methodology for near-real time mapping of LSC at 10m resolution.

4.4.2.1 Input data and study sites

In the first phase of the project, we focussed the development of the LSC methodology on the study sites in Europe. We trained and benchmarked the LSC models using around 1000 patches located in Europe (see also in the next section) and predicted the LSC over the pre-selected test sites within Europe.

Sentinel-2 data and the Copernicus DEM were used to train the LSC models and to predict maps over the test tiles (Table 46). The WorldCover land cover map of the year 2021 was additionally used to generate annotations (Zanaga et al., 2022).

Table 46. input datasets.

Dataset	Description	Spatial Resolution	Temporal resolution
Sentinel-2 data	Sentinel-2 L2A spectral bands	10 - 20m	Single scenes
Copernicus DEM (GLO-30)	Digital Elevation Model	30m	NA
WorldCover 2021	WorldCover land cover dataset of 2021	10m	Year 2021

4.4.2.2 Description of the candidate method

The aim of the subtask is to develop a candidate method that focuses on mapping LSC on single scenes using a legend covering 10 classes (Table 47).

Table 47. Overview of the LSC legend classes.

Label	Description	Colour
10	Woody vegetation	
30	Herbaceous vegetation	
50	Built-up areas	
60	Bare	
70	Snow/ice	
80	Water	
110	Thick clouds	
120	Thin clouds	
130	Cloud shadows	
0	No data	

The proposed methodology combines two models to include contextual information while preserving fine spatial details. First, we take advantage of the capabilities of deep learning models to extract contextual information. The deep learning model has the drawback that it has difficulties to preserve fine spatial details. To reduce these effects, we subsequently apply a machine learning (CatBoost) model that works on single pixels.

Annotations

In order to train the Deep Learning (Unet++) and Machine Learning (CatBoost) model, an extensive training dataset is required. Yet, since manually collecting annotations from single scenes is very costly and time consuming, we applied an automated workflow combining auxiliary data to collect the annotations.

Specifically, we first used the WorldCover 2021 labels to derive the most likely LSC label (Table 48). For the WorldCover Snow & Ice class, we only set the LSC label to 'snow & Ice' if the SCL also indicated the presence of snow and ice.

Table 48. Translation of the WorldCover labels into LSC labels.

WorldCover label	WorldCover class	LSC label	LSC class
10	Tree	10	Woody vegetation
30	Grassland	30	Herbaceous vegetation
40	Cropland	30	Herbaceous vegetation
60	Bare	60	Bare
70	Snow/ice	70	Snow/ice
80	Water	80	Water
95	Mangrove	10	Woody vegetation
100	Lichen	30	Herbaceous vegetation

To ensure that the annotations are able to capture temporal water fluctuations, we subsequently labelled all pixels that had a Normalized Difference Water Index (NDWI) larger than 0.3 to water. Pixels that were permanent water according to WorldCover 2021, had reflectance for the SWIR band (B12) larger than 0.2, and an FCOVER larger than 0.15 were set to herbaceous vegetation

Seasonal change in the occurrence of the bare class was annotated by setting the pixels that had a WorldCover label equal to 30 (grassland) or 40 (cropland) and an FCOVER lower than 0.1 to bare. Pixels that were bare according to WorldCover 2021 but had an FCOVER larger than 0.15 were set to herbaceous vegetation.

Seasonal change in snow and ice cover was captured by setting all pixels that simultaneously had a Normalized Difference Snow Index (NDSI) larger than 0.4, a Snow Water Index (SWI) larger than 0.21, and their reflectance for band B03 larger than 0.2 to the snow/ice class. In addition, pixels that were mapped as snow or ice in the Scene Classification Layer (SCL) were set to the snow & ice class.

Since some of the WorldCover 2021 classes may contain a mixture of LSC classes (e.g. wetland) or are prone to confusion (e.g. the shrubs class shows confusion with trees and herbaceous vegetation), these classes were set to ‘no data’. In addition, pixels that had an SCL value equal to zero (no data) or one (saturated or defective pixel) were set to ‘no data’.

Finally, we applied a UNet model, trained on the CloudSen12 dataset (Aybar et al., 2022a) to identify clouds and cloud shadows.

We prepared the annotations using single Sentinel-2 scenes of the year 2020 over around 1000 patches across Europe (Figure 120). For each location, we tried to collect four scenes with a cloud cover lower than 50% spread throughout the year (i.e. one in each quarter of the year). The patches covered an area of 1280x1280m and had a spatial resolution of 10m.

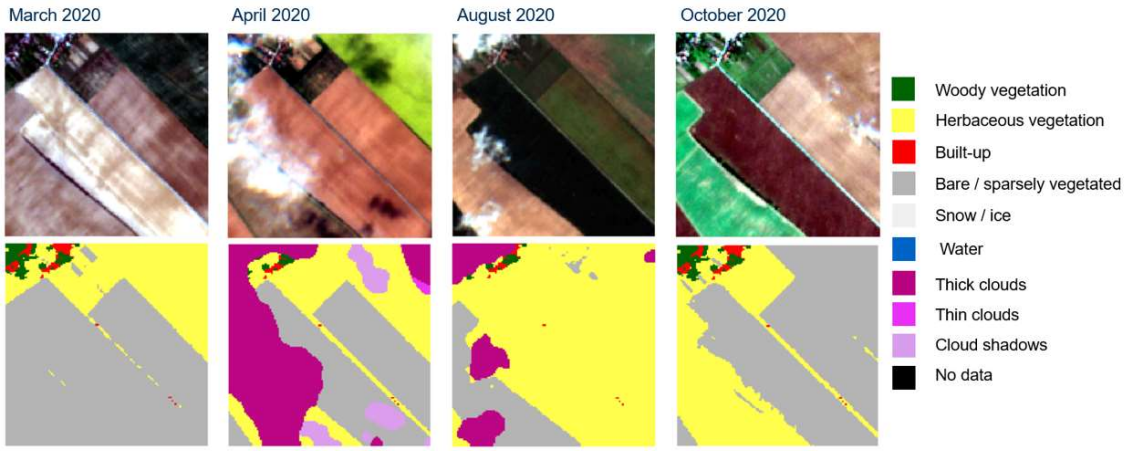


Figure 120. Example of annotations generated over a patch located in a cropland area.

Unet++ model

A Unet++ model with ResNet50 backbone was subsequently trained on the patch annotations using six spectral bands (BO2, B03, B04, B8A, B11, B12) as input. To help the model to localize, we also provided topography (DEM), the location (latitude and longitude), and circularly encoded Day of Year as input. Finally, we provided information on the 90th percentile of the NDVI over the previous year. This feature summarizes the past vegetation profile and may help the model to more robustly classify vegetated pixels during months when the spectral signal of herbaceous and woody vegetation is very similar.

Binary Cross Entropy with logits was used as loss function, where classes were weighted based on their abundance in the training dataset. Pixels annotated as no data, clouds or cloud shadows were ignored while computing the loss.

CatBoost model

Next, we trained a CatBoost model that used the probabilities of the U-Net model as input. Similar to the U-Net model, we also used localization information (DEM, location, and circularly encoded DOY), and six spectral Sentinel-2 bands (B02, B03, B04, B8A, B11, and B12). We additionally supplied the model vegetation indices (NDVI, NDSI, and NDWI) and a set of features to summarize the vegetation profile over the past year: the 10th, 25th, 50th, 75th, and 90th NDVI percentile over the previous year.

The training database for the CatBoost model was extracted from the annotated patches, where we sampled for each patch a maximum of 10 percent of the pixels per label, while not exceeding 20 observations per label and excluding pixels at the edge between two LSC classes. Observations for a label were excluded if only very few observations (less than ten) were found within the patch.

Cloud masking

Clouds and cloud shadows were detected using a U-Net model trained on the CloudSen12 dataset (Aybar et al., 2022b) and using the Sentinel-2 bands B02, B03, B04, B8A as input.

4.4.2.3 Benchmarking & evaluation

A benchmark has been performed on models that use different sets of input data (Table 49). Specifically, the effect of using features that represent the past vegetation profile in the UNet and CatBoost models has been compared.

Table 49. Overview of the benchmarked models.

Model name	Input features UNet model	Input features CatBoost model
UNet_CB	S2 bands, DEM, lat, lon, DOY	S2 bands, VI, DEM, lat, lon, DOY
UNet+_CB	S2 bands, DEM, lat, lon, DOY, **90th NDVI percentile**	S2 bands, VI, DEM, lat, lon, DOY
UNet_CB+	S2 bands, DEM, lat, lon, DOY	S2 bands, VI, DEM, lat, lon, DOY, **10th, 25th, 50th, 75th, 90th NDVI percentile**
UNet+_CB+	S2 bands, DEM, lat, lon, DOY, **90th NDVI percentile**	S2 bands, VI, DEM, lat, lon, DOY, **10th, 25th, 50th, 75th, 90th NDVI percentile**
CB	[no UNet model used]	S2 bands, VI, DEM, lat, lon, DOY
CB+	[no UNet model used]	S2 bands, VI, DEM, lat, lon, DOY, **10th, 25th, 50th, 75th, 90th NDVI percentile**

The different models were evaluated on a validation set of 10% of the annotated patches presented in the previous section. In addition, a qualitative comparison was performed on the results.

The f1-scores per LSC class show that the inclusion of features related to vegetation profile of the past year increase their accuracy. Although the increase is mostly apparent for the vegetated classes, the non-vegetated classes also show an enhancement.

Table 50. f1-score of the LSC classes and overall accuracy (OA) for each of the models.

	UNet_CB	UNet_CB+	UNet+_CB	UNet+_CB+	CB	CB+
Woody vegetation	0.79	0.85	0.78	0.86	0.76	0.83
Herbaceous vegetation	0.81	0.86	0.81	0.86	0.80	0.82
Built-up areas	0.78	0.82	0.84	0.85	0.62	0.78
Bare	0.85	0.87	0.88	0.88	0.82	0.84
Snow/ice	0.91	0.95	0.92	0.95	0.95	0.92
Water	0.95	0.96	0.95	0.93	0.93	0.95
OA	0.83	0.88	0.85	0.88	0.80	0.85

Since the model based on the UNet+_CB+ model seems to achieve the best results, this model was selected as the candidate model (Table 50).

4.4.3 Implementation and results

The candidate model has been applied to test tiles of the prototype areas within Europe. To minimize the data volume and facilitate the inspection of the dataset, we selected and produced LSC maps for single scenes for each month of the year 2020 having a minimal number of missing data and cloud cover. These maps were made available at Figure 121 and Figure 122.

The LSC maps capture the temporal fluctuations between sparsely vegetated areas, and herbaceous cover. In addition, changes in water bodies and snow cover seem to be captured well.

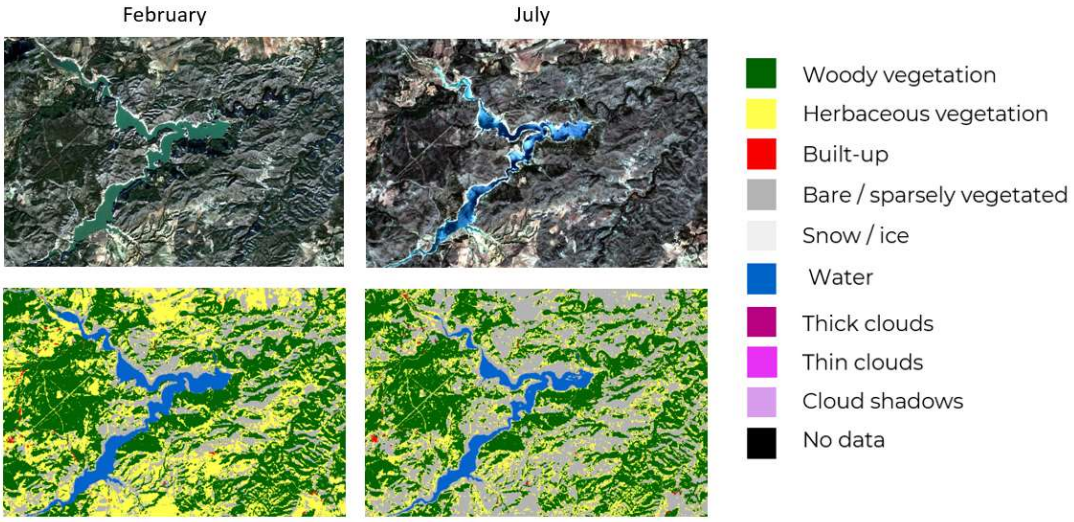


Figure 121. Example of LSC predictions over an area in Spain, with fluctuations in herbaceous cover, and minor changes in water level.

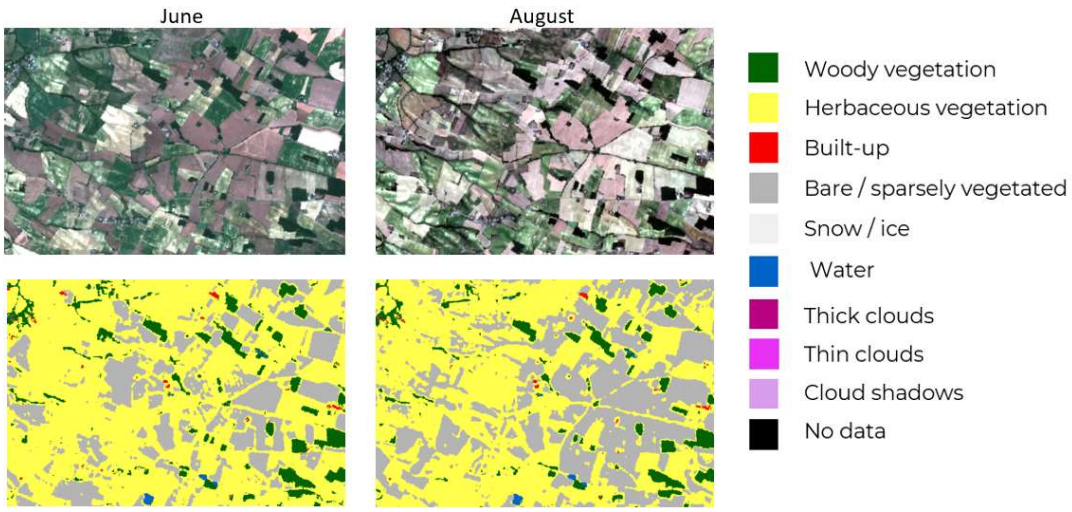


Figure 122. Example of LSC predictions over a cropland area, where harvest, crop growth, and planting/sowing activities cause changes between bare and herbaceous cover over time.

During some months, the spectral signature of herbaceous and woody vegetation is similar. This causes confusion between the vegetation classes and hence also seasonal fluctuations. The usage of the vegetation profile of the past year has the advantage that it reduces seasonal fluctuations in the estimated woody cover as the model can use information over a longer time span to provide a better estimate. Yet, some seasonal change in woody vegetation cover is still noticeable.

Using information of the past has the drawback that the model might be less sensitive to identify vegetation change (e.g. deforestation) near real-time. As change is typically occurring rarely, it is not frequently annotated, resulting in a reduced incentive for the model to learn to capture this change near real-time.

To avoid this, we plan to test the introduction of synthetic change events in the training dataset and train a model dedicated to woody vegetation changes.

In addition, we noticed that there is sometimes still some confusion between bare and built-up areas. We plan to produce a new set of annotations, where we improve the built-up annotations (using e.g. Open StreetMap data).

Finally, some commonly occurring artifacts (prediction of water over topographic shadows or prediction of snow of undetected clouds) need to be handled.

4.4.4 Summary and conclusions

A candidate model has been implemented that allows to predict NRT changes in LSC. The model uses the NDVI profile of the past year to reduce seasonal fluctuations in woody vegetation cover. The implemented model captures the water, snow, and cropland dynamics well. Improvements will focus on further reducing seasonal fluctuations in woody vegetation cover and confusion between bare and built-up areas.

5 Biomass mapping

5.1 Introduction

The following chapter presents the forest biomass, see section 5.2, and grassland/cropland GPP biomass mapping, see section 5.3. The objective is to develop novel methods for mapping AGB in forests and GPP mapping in cropland and grasslands, respectively.

5.2 Forest biomass

5.2.1 State of the art

5.2.1.1 Introduction

Forests play a key role to ensure sustainable development through biodiversity conservation and climate change mitigation. Vegetation height and above ground biomass have been ranked as high-priority biodiversity variables to be observed from space (Skidmore et al., 2021). In the Copernicus Land Monitoring Service (CLMS) portfolio, spatially explicit, timely information on Above Ground Woody Biomass as well as on forest canopy height is currently absent but would be key to support EU policies on renewable resources, climate change and biodiversity. Therefore, this candidate service aims to complement the CLMS portfolio with spatio-temporal consistent estimates of above-ground woody biomass (AGB) and forest canopy height (FCH). By applying parametric as well as non-parametric models, AGB and FCH are estimated at pixel level, which can then be aggregated according to stakeholder requirements. The candidate forest biomass mapping service should allow a European wide wall-to-wall yearly coverage with a minimum mapping unit of 0.25 ha.

5.2.1.2 Review of existing methods

The estimation methods for AGB and FCH can be broadly categorized into parametric and non-parametric methods. **Parametric methods** are based on a predetermined model that describes the relationship between the AGB and the remote sensing data. While parametric methods can provide interpretable models and be computationally efficient, they rely on the assumed model being a good fit for the data. On the other hand, **non-parametric methods** do not make strong assumptions about the form of the relationship between AGB and FCH and remote sensing data. Instead, they learn this relationship directly from the data.

5.2.1.2.1 Non-parametric methods

Forest Canopy Height was mapped at the global level by (Potapov et al., 2021) through integration of GEDI and Landsat data. They implemented a bagged regression trees ensemble method (Breiman, 1996) to model forest height. They

expect that the integration of GEDI data with time-series optical imagery enables multidecadal historic analysis and operational forward monitoring of forest height and its dynamics.

The primary science goal of the NASA's GEDI mission is the provision of accurate estimates of forest aboveground biomass density (AGBD). (Duncanson et al., 2022) describe the development of the models used to create GEDI's foodprint-level (~25m diameter) AGBD product, including a description of the procedure for final model selection by stratification. For strata in Europe, they report R^2 values between 0.64 and 0.68.

The consistency of forest height retrievals between GEDI and ICESat-2 was analysed by (X. Zhu et al., 2022). Their results show that the accuracy of forest heights extracted from GEDI data (R^2=0.93, RMSE=2.99m for power beams at night) is higher than that extracted from ICESat-2 data (R^2=0.78, RMSE=4.62m for strong beams at night)(X. Zhu et al., 2022). They demonstrate the possibility of combining different modes of space-borne LiDAR data to obtain consistent forest heights.

Amazon rainforest regrowth with GEDI and ICESAT-2 data was assessed by (Milenković et al., 2022). Their quality assessment of GEDI and ICESat-2 forest heights over four 50km long airborne LIDAR strips determined a root mean square error of 4.14m and 5.91m. They implemented a linear calibration model between satellite- and airborne-LIDAR heights for data subgroups. Their study revealed that omitting either the calibration step or the removal of secondary-forest-border pixels would result in an underestimation of the regrowth rate by more than 20%.

Quantifying aboveground biomass dynamics in Mozambique using GEDI Lidar and Landsat data was performed by (M. Liang et al., 2023). Monitoring forest degradation as compared to deforestation presents technical challenges, as degradation in general involves low-intensity AGB removal under varying temporal dynamics. In their study they combined GEDI Lidar with Landsat optical time-series data to quantify biomass losses associated with charcoal related forest degradation over 10 years. In a first step they created annual medoid composites (Flood, 2013) for selected time periods. These composites were then time-series fitted to minimize temporal noise/variations in the Landsat pixel-level spectral trajectories, by applying the LandTrendr approach (Kennedy et al., 2010b). By applying the best performing Random Forest model to the temporal stabilized annual composites, they produced annual forest AGBD maps. The Random Forest model achieved an RMSE of 7.05Mg/ha and R^2 value of 0.64 using a 10-fold cross-validation dataset.

Forest aboveground biomass mapping using multisource remotely sensed data is demonstrated by (Ehlers et al., 2022). They used multiseasonal Landsat images, LiDAR data for canopy height, RaDAR data as input. All data layers were fed to a Random Forest regression model by using USFS Forest Inventory and Analysis plots to train and test the model performance. The best model achieved an R^2 of 0.625 based on "out-of-bag" samples at 30x30m spatial resolution. Their variable importance analysis demonstrates the high information content of LiDAR data (the

top five most important variables are percentile heights of the LiDAR points). However, airborne LiDAR data is only available for large areas in exceptional cases, which limits the large area applicability of the approach.

Woody biomass and tree cover mapping including areas outside of forests was performed by (S. Liu et al., 2023). They used PlanetScope imagery with approximately 3 m pixel resolution and four multispectral bands. For training two deep learning models, a segmentation model predicting binary canopy/no-canopy and a regression model predicting canopy height, they used airborne LiDAR data across Europe. They further included a GEDI- and Sentinel-2 based canopy height map from (Lang, Jetz, Schindler, et al., 2023) as an additional model input. They used Danish NFI data to establish allometric relationships between plot-based AGB and height from aerial CHMs (S. Liu et al., 2023). The relationship between biomass and canopy height was then used to convert canopy height and cover predicted from PlanetScope imagery into biomass at 30m resolution, which was lastly aggregated to hectare level, regional level and at national level. At the national level they report a systematic bias of 7.6% compared to national inventories of 30 countries.

An evaluation of eight machine learning regression algorithms for forest above ground biomass estimation from multiple satellite data products was performed by (Y. Zhang et al., 2020). Their results showed that five tree-based ensemble algorithms (Random Forest, extremely randomized trees, categorical boosting, gradient-boosted regression tree and stochastic gradient boosting) performed better than the three non-ensemble algorithms (multivariate adaptive regression splines, support vector regression, and multilayer perceptron) for AGB estimation. Compared with the random sampling scenario, the stratified sampling scenario and separate modelling did not significantly improve the AGB estimates. For the performance metric R-squared, based on validation data which was not used for training, the three approaches Random forest (R^2=0,70), extremely randomized trees (R^2=0,70) and categorical boosting (R^2=0.71) provided very similar quality and outperformed the other approaches (Y. Zhang et al., 2020)

5.2.1.2.2 Parametric methods

Contrarily to the optical domain where measurements are linked to the greenness of the vegetation, measurements in the microwave domain have been shown to be physically linked to the biomass of vegetation, making it possible to consider a parametric approach for the estimation of biomass. Passive microwave sensors, such as SMOS or SMAP, have too coarse resolutions to meet the requirements identified in this project. Active microwave sensors, namely Synthetic Aperture Radars, provide images with a sufficient spatial resolution. Their sensitivity to forest above-ground biomass depends on the wavelength at which they operate (Le Toan et al., 2004). Shorter wavelengths, such as C-band, are limited by their inability to penetrate the canopy, leading to a saturation of the SAR measurement already at small biomass values (around 30-50 $Mg.ha^{-1}$). Longer wavelengths, such as L-band or P-band, are more sensitive to woody components and can therefore exhibit

saturation values at higher values, around 100-150 $Mg.ha^{-1}$ for L-band and 300 $Mg.ha^{-1}$ for P-band. Until the launch of the BIOMASS satellite by ESA planned in late 2025, only L-band SAR systems are available in space.

The dependence of L-band SAR measurements to forest AGB has been studied already for several decades. The most commonly used model that relates SAR measurements (backscattering coefficients) to vegetation parameters is the Water Cloud Model (WCM) (Attema & Ulaby, 1978). In recent years, most large-scale applications have used a simplified formulation of the WCM which uses 3 parameters:

$$\gamma^0 = \gamma^0_{ground}.e^{-c.AGB} + \gamma^0_{veg}.(1-e^{-c.AGB}) = a.e^{-c.AGB} + b.(1-e^{-c.AGB})$$

where γ^0 represents the backscattering coefficient of a forest (for a given polarisation), $a = \gamma^0_{ground}$ represents the backscattering coefficient of the ground below the trees, $b = \gamma^0_{veg}$ represents the backscattering coefficient of a dense forest, and c represents an attenuation coefficient which describes how quickly the backscatter rises from a to b when the AGB increases.

The three parameters a, b and c need to be calibrated because of their dependence to local conditions. Parameter a is sensitive mostly to environmental conditions, such as soil moisture, while b and c depend predominantly on the forest type.

Once calibrated, the model needs to be inverted in order to estimate AGB. While a direct inversion is possible, a more powerful approach, the Bayesian inversion, has also been used and applied at large scale in Africa (Bouvet et al., 2018).

5.2.1.2.3 Validation methods

Calibrating and validating biomass mapping methodologies is crucial to ensuring the accuracy and reliability of the derived biomass estimates. Different methods can be applied for calibration and validation. For example, forest backscatter model utilized in the CCI Biomass project does not require ground data for calibration, besides maximum possible biomass at regional level. Proper validation instead require ground truth data. Ground truth data collected through field measurements provide essential reference points for verifying the accuracy of the biomass estimates derived from remote sensing data.

Good practice of biomass validation protocol is presented by Duncanson et al. (2021) where different sources of reference data are discussed together with associated errors.

The utilize of slam national forest inventory plots in validation is discussed in (Araza et al., 2022), including dedicated R-package *PlotToMap*[14].

In addition to ground truth data, comparison with other biomass products is another valuable approach for validating biomass mapping methodologies. This

[14] https://github.com/arnanaraza/PlotToMap

comparative analysis is particularly insightful when the alternative biomass products are generated using independent methods or datasets.

Overall, the state-of-the-art calibration and validation of biomass mapping methodologies involve a combination of ground truth data collection, model calibration, and inter-comparison with other biomass products to ensure the accuracy and reliability of the final biomass estimates.

5.2.1.3 Strengths and weakness of existing methods

5.2.1.3.1 Non-parametric methods

This group of methods relates sample-based reference data on above ground biomass or forest canopy height to wall-to-wall acquired remote sensing data. The basic assumption is that the area measured at each reference sample location can spatially and temporally be matched with the same area as observed via remote sensing. Four main sources of discrepancy are reported in the CEOS Aboveground Woody Biomass Product Validation Good Practices Protocol (Duncanson et al., 2021):

- Mismatch between field plot and pixel sizes and shapes
- Spatial co-registration errors
- Lack of agreement between the forest components measured
- Temporal difference between RS and reference measurements

High quality pre-processing for both, the reference data as well as the satellite time series is therefore a prerequisite. Detailed recommendations are provided in the Aboveground Woody Biomass Product Validation Good Practices Protocol (Duncanson et al., 2021).

In general, it is observed, that AGB modelling algorithms underestimate high AGB and overestimate low AGB values. Specific efforts should therefore be made to improve the AGB estimates at low and high AGB values by calibration, see e.g. (Y. Zhang et al., 2020). Also, the study by (Milenković et al., 2022) revealed that calibration of satellite LIDAR heights with airborne LiDAR heights was necessary to avoid erroneously large regrowth intercepts and an underestimation of forest regrowth rate. A simple linear calibration model was able to correct systematic distortions.

For GEDI data, the study from (Milenković et al., 2022) shows, that erroneous forest heights were excluded when GEDI shots with a beam sensitivity of 0.95 or larger were selected. For comparing GEDI with ICESat-2 measurements (Milenković et al., 2022) prefer GEDI observations because of better forest height accuracy, and lower internal uncertainty.

Tree-based ensemble algorithms such as random forests frequently outperform non-ensemble algorithms, see e.g. (Y. Zhang et al., 2020). Further, the tree-based ensemble algorithms allow deriving confidence intervals and uncertainty estimates at pixel level. From the group of the non-parametric estimation methods we

therefore recommend using a tree-based ensemble approach for operational provision of a potential future CLMS service for mapping AGB and FCH.

5.2.1.3.2 Parametric methods

One advantage of the parametric approach tested here is that it generally requires less training samples for the calibration than non-parametric approaches such as random forest.

However, proper calibration of this model is challenging and requires taking into account all the potential perturbations that affect the relationship between the SAR backscatter and the AGB estimates. These error sources include firstly mismatches between the size of the field plots and the pixel size of the SAR images (Réjou-Méchain et al., 2014), tree canopy layover and border effects, errors in data geolocation, and topographic effects. To minimize these errors, it is usually recommended to use *in situ* reference AGB plots with a size larger than 1ha and located in flat areas, which are often difficult to find.

Another important source of error is the effect of environmental conditions on the SAR measurement, such as: variations in soil moisture, the presence of snow, the presence or absence of leaves (in deciduous forests). One way to reduce the perturbing effect of environmental conditions is to use multi-temporal datasets (several SAR acquisitions in the year) to select the most suitable dates (typically, acquired in a dry season). This is unfortunately not possible to do this with the yearly PALSAR-2 mosaics, but it will become possible when the first NISAR images will be accessible (launch planned on 30 March 2024).

5.2.2 Method description

5.2.2.1 Input data and study sites

5.2.2.1.1 Mediterranean study site

The Mediterranean study site is located in the southern area of Spain in the region Andalusia and Murcia. The forest ecosystems are diverse and complex, primarily located in the mountains. The study site which is used for testing of the non-parametric models covers the extend of the Sentinel-2 granule 30SWH with an area of 12.000 km². The main forest types from CLC+ 2018 are shown in Figure 123.

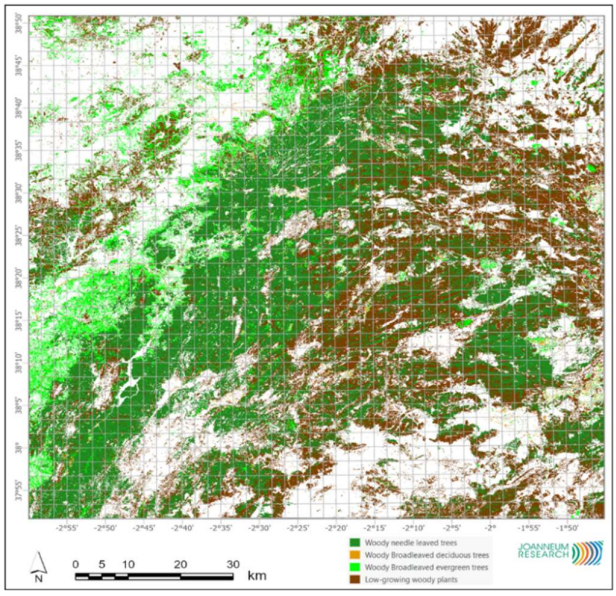


Figure 123. Mediterranean study site used for testing of non-parametric models.

The testing of parametric models in the Mediterranean site was carried out over the 16 PALSAR-2 tiles that cover most of the prototype site (9 MGRS tiles, see Figure 124)

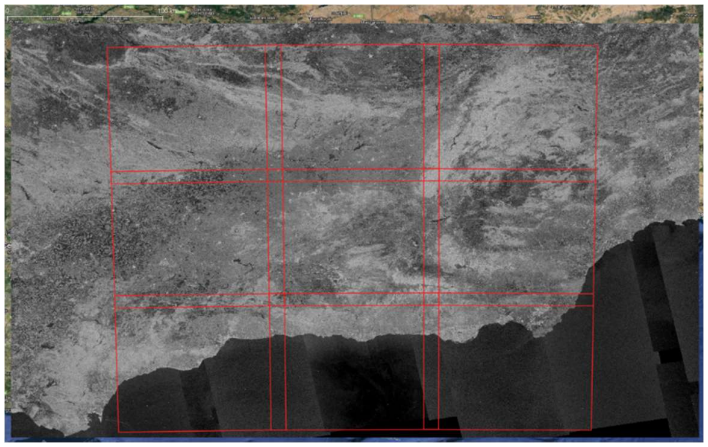


Figure 124. The filtered HV backscatter of 2020 over the 16 PALSAR-2 tiles of the Mediterranean prototype site.

Table 51. Input datasets.

Dataset	Description	Spatial Resolution	Temporal resolution
GEDI	LiDAR	25m	2022
Sentinel 2	Multispectral	10 to 20m	2022
Sentinel 1	SAR C-Band	10m	2022
COPDEM 10m	Copernicus Digital Surface Model	10m	2015
CLC+	Copernicus Land Cover	10m	2018
ALOS PALSAR	SAR L-Band	1”	2022

GEDI – LiDAR

The NASA Global Ecosystem Dynamics Investigation (GEDI) is a spaceborne lidar instrument operating onboard the International Space Station to provide high-quality measurements of forest vertical structure in temperate and tropical forests between 51.6° North and South latitude. The GEDI instrument is a geodetic-class laser altimeter/waveform lidar comprised of 3 lasers that produce 8 transects of structural information (Dubayah et al., 2020). These data are used to derive a variety of footprint and gridded products, including canopy height, canopy foliar profiles, Leaf Area Index (LAI), sub-canopy topography and biomass at a footprint resolution of 25 m (Figure 125).

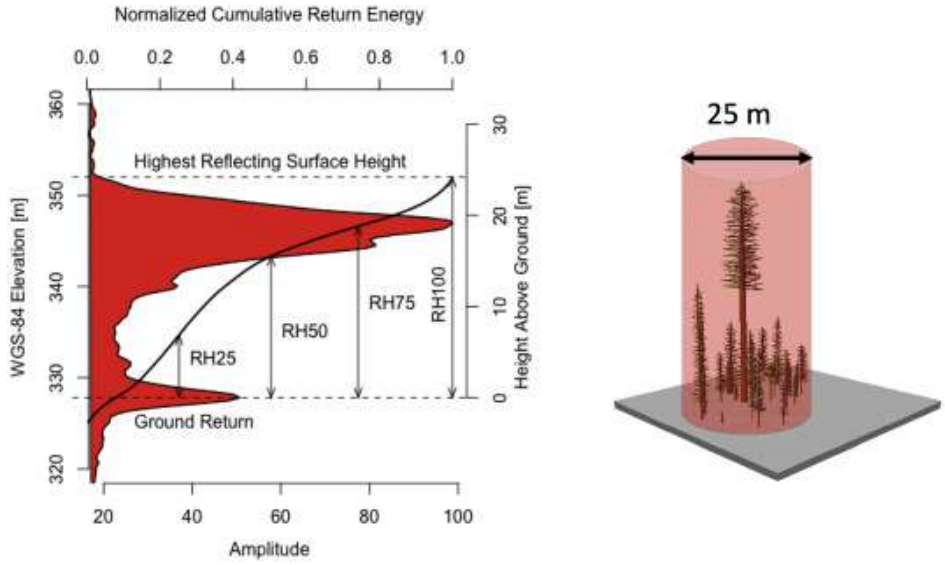


Figure 125. GEDI waveform. A near-infrared pulse of laser energy is fired towards the surface where it is reflected by leaves and branches within a nominal 25 m diameter footprint (shown at right). The returned waveform (shown on left) is processed to find ground topography, canopy height, and various relative height (RH) metrics. From these metrics a variety of other products are derived, including leaf density profile, canopy cover and aboveground biomass (Dubayah et al., 2020).

The L4A footprint data product represents aboveground biomass density for individual GEDI footprints. Footprint AGBD is derived from linear models that relate GEDI L2 waveform height metrics to field-estimated aboveground biomass (see Figure 126).

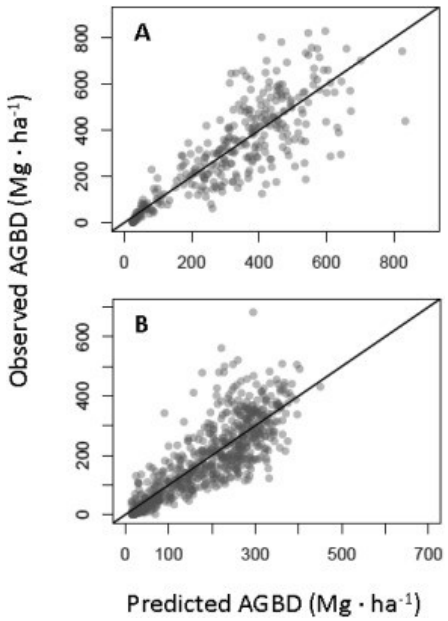


Figure 126. Footprint-level relationships between predicted and observed AGBD. The predictions are from generalized linear models developed using waveform relative height metrics. (A) Evergreen Broadleaf Trees in Australia; (B) Deciduous Broadleaf Trees in North America. The percentage RMSE is 30.7%, and 38.2%, respectively (Dubayah et al., 2020).

For estimation of AGB we used the Level 4A data which is publicly available. Detailed documentation and links for downloading is provided by NASA (2024).

Sentinel-2 optical timeseries

Sentinel-2 is a European wide-swath, high-resolution, multi-spectral imaging mission. The full mission specification of the twin satellites flying in the same orbit but phased at 180°, is designed to give a high revisit frequency of 5 days at the Equator. Sentinel-2 carries an optical instrument payload that samples 13 spectral bands: four bands at 10 m, six bands at 20 m and three bands at 60 m spatial resolution. The orbital swath width is 290 km. Full details of the mission is provided by ESA (2024). From the time-series different temporal metrics were derived as basis for detailed analysis and testing, such as e.g. vegetation indices, texture parameters, median and percentiles over defined time periods, further composites based on medoids were used which are provided by the Sentinel-2 Global Mosaic Service[15].

Sentinel-1 SAR timeseries

The Sentinel-1 mission comprises a constellation of two polar-orbiting satellites, operating day and night performing C-band synthetic aperture radar imaging, enabling them to acquire imagery regardless of the weather. We pre-processed the time-series and derived temporal features which were used for testing. Further we

[15] https://land.copernicus.eu/en/technical-library/sentinel-2-global-mosaic-product-user-manual

used Sentinel-1 monthly mosaics for the years 2020 and 2022 which were made available in the last months.

ALOS – PALSAR Mosaics

In the parametric approach, the main source of data is the global L-band SAR mosaics made available for free by JAXA[16].

It consists of global mosaics of the L-band Synthetic Aperture Radar sensor ALOS-2 PALSAR-2 at 25m resolution produced yearly over the 2015-2022 period, at HH and HV polarization. The mosaic is available for download as 1°×1° tiles.

After downloading of the 8 available years over the tiles covering the test sites, some pre-processing is needed:

- Calibration to transform the Digital Numbers to backscattering coefficient.
- Multi-temporal filtering using the approach developed in (Quegan & Jiong Jiong Yu, 2001). This filtering aims at reducing the speckle noise without degrading the spatial resolution (Figure 127). The number of looks is increased, leading to a more accurate estimation of the backscatter. All the images were filtered with 16 input images (2 polarisations and 8 years), and a local window size of 5×5 pixels.

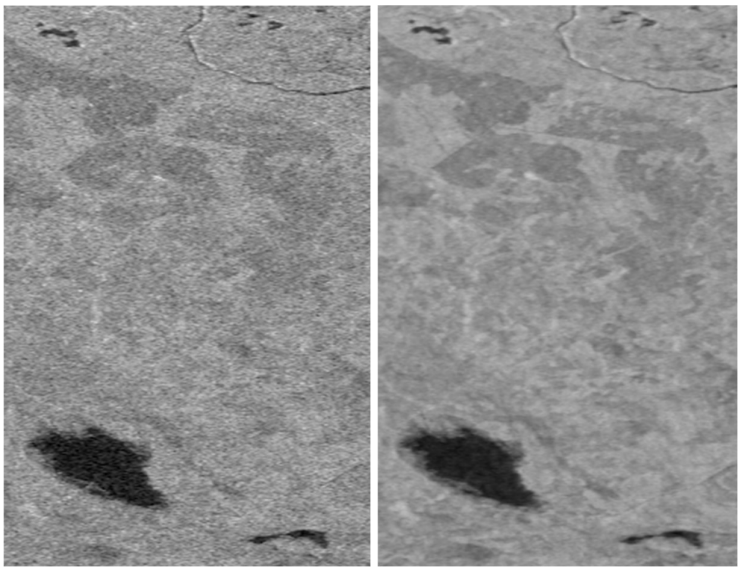


Figure 127. Effect of the multi-temporal filter: unfiltered (left) vs filtered (right) image.

5.2.2.1.2 Boreal study site

The Boreal study site is located in Central Sweden. To cover a sufficiently large sample of NFI plots for training of non-parametric models with S2 optical data, the study site was enlarged compared to the original plan. The study site which is used

[16] https://www.eorc.jaxa.jp/ALOS/en/dataset/fnf_e.htm

for testing of the non-parametric models therefore covers several S2-granules, as shown in Figure 128.

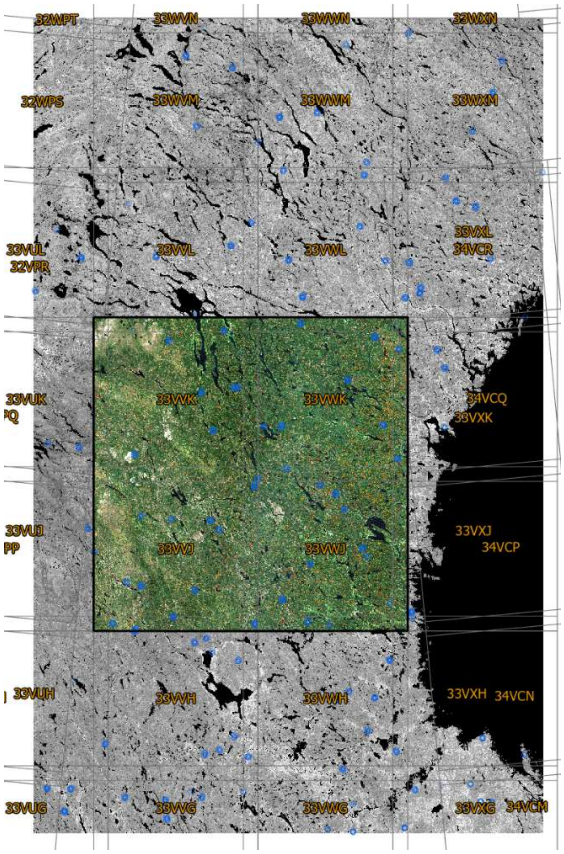


Figure 128, Extended study area used for training of the non-parametric models. Sentinel-2 NIR band overlaid with the location of the NFI plots in blue. In the centre, the location of the demo site is drawn in Sentinel-2 RGB band combination, which cover 4 Sentinel-2 granules.

Within this study site is the demonstration site, which covers four granules (drawn in RGB in above figure). The forest parameters AGB and FCH are estimated for these four granules, to cover the full demonstration site.

Within the S2-granules 33VVK and 33VWK a sub-study site is used for intensive testing of the non-parametric models, which is shown in Figure 129. This sub-study site is covered to a large extend by LIDAR data. Table 52 lists the datasets which were used for the development of the parametric as well as for the non-parametric methods.

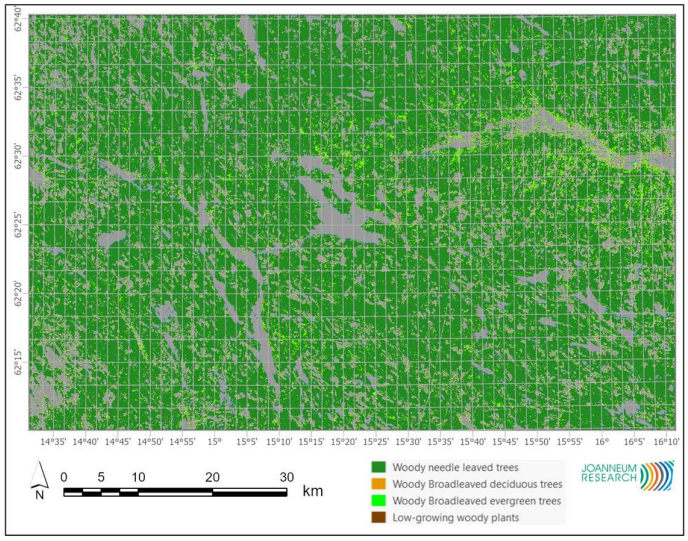


Figure 129. Sub-test site which is covered by LIDAR data to a large extend, which was used for non-parametric model testing.

Table 52. Input datasets used in the boreal study site.

Dataset	Description	Spatial Resolution	Temporal resolution
Sweden NFI	National Forest Inventory	Diameter 14m, 20m	2020, 2022
Sentinel 2	Multispectral	10 to 20m	2020, 2022
Sentinel 1	SAR C-Band, monthly mosaics	20m	2020, 2022
COPDEM 10m	Copernicus Digital Surface Model	10m	~2015
CLC+	Copernicus Land Cover	10m	2018, 2021
ALOS PALSAR	SAR L-Band	1”	2018 - 2022

ALOS – PALSAR Mosaics

The same data and pre-processing as in the Mediterranean site were used. The testing of the boreal site was carried out over the 8 PALSAR-2 tiles that cover most of the prototype site (4 MGRS tiles, Figure 130).

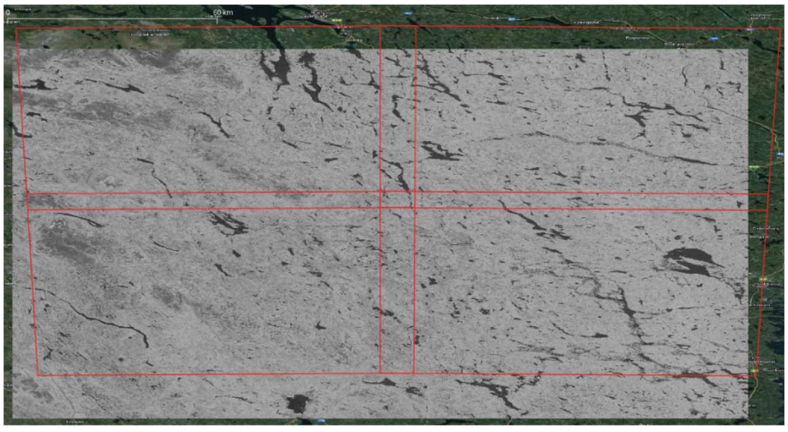


Figure 130. The filtered HV backscatter of 2020 over the 8 PALSAR-2 tiles of the boreal prototype site.

5.2.2.1.3 Tropical study site

The Tropical study site is located in the north of French Guiana and covers coastal areas as well as tropical rain forests. This site is used to test the non-parametric UNet approach. To increase the amount of training data, which is needed for this approach, the study area was expanded to cover entire French Guiana. For data processing the area is split into 1x1° tiles with 0.1° overlap Figure 131.

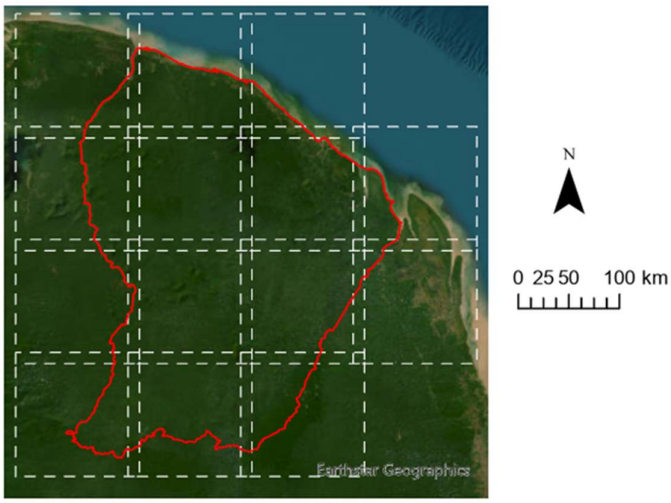


Figure 131. French Guiana border and processing extent (red). Processing is split up into overlapping tiles (grey).

Reference data from the TropiSAR campaign

The TropiSAR campaign was conducted in French Guiana in the summer 2009 in the framework of the Phase A studies pertaining to the BIOMASS mission. Tree measurements and inventories have been made over the Nouragues and Paracou sites, as well as two airborne Lidar acquisitions which were converted to AGB maps (Figure 132).

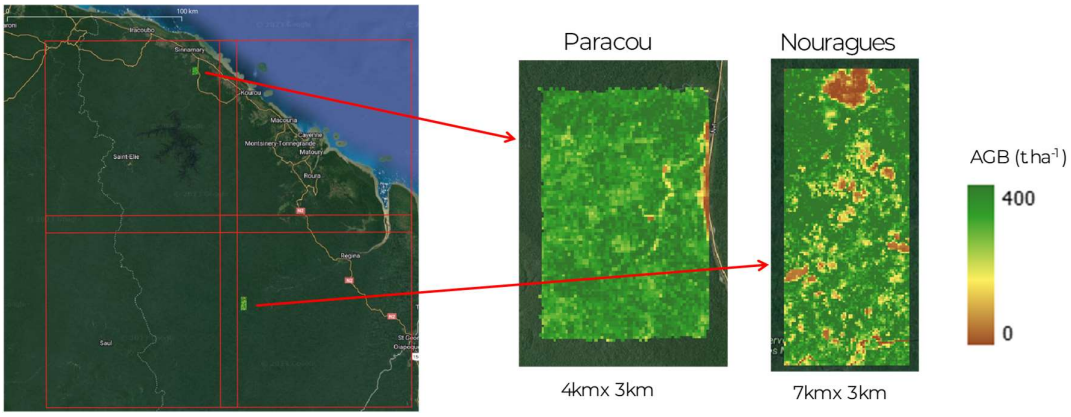


Figure 132. The two available AGB maps derived from Lidar flights during the TropiSAR campaign.

Reference data from GEDI L4A products

This dataset contains Global Ecosystem Dynamics Investigation (GEDI) predictions of the forest canopy height (L2A) and aboveground biomass density (L4A). For both, L2A and L4A, all derived products are available in six algorithm groups. For tropical forests Ngo et al. (2023) identifies the a5 algorithm as the best fit and this group is therefore used for GEDI data processing. The variable rh98 is used as an approximation for canopy height. Ngo et al., (2023) calculated that, by using rh98, the RMSE compared to LIDAR acquired canopy height is lowest for tropical forests.

To obtain high-quality and unbiased training and testing data, multiple filters were applied to the raw data. The GEDI instrument facilitates laser beams of different physical properties. Power beams are twice as powerful as coverage beams and therefore power beams are performing better in penetrating dense canopies (Lahssini et al., 2022). Therefore, only power beams are used for training and testing. Further filtering steps and the source of the chosen criteria are listed in Table 53.

Table 53. Filtering criteria for GEDI data.

GEDI Type	Filter criteria	Source
L2A and L4A	solar_elevation < 0	Dong et al. 2023
L2A and L4A	quality_flag = 1	Oliveira et al. 2023
L2A and L4A	sensitivity > 0.99	Own criteria
L2A and L4A	degrade_flag = 0	Oliveira et al. 2023
L2A	num_detectedmodes > 0	Schwartz et al. 2022
L2A	abs(elev_lowestmode - digital_elevation_model) < 100	Schwartz et al. 2022
L2A	0 <= rh98 < 7000	Own criteria
L4A	surface_flag = 1	Own criteria
L4A	0 <= agbd < 900	Own criteria

The filtered GEDI data is split into a training and a test dataset via a 1x1 km grid. In this grid, 30% of the cells are randomly assigned as testing cells and the remaining cells are training cells. All GEDI plots within the testing cells are assigned as testing points and all GEDI plots within the training cells are assigned as training points. The grid is shown in Figure 133.

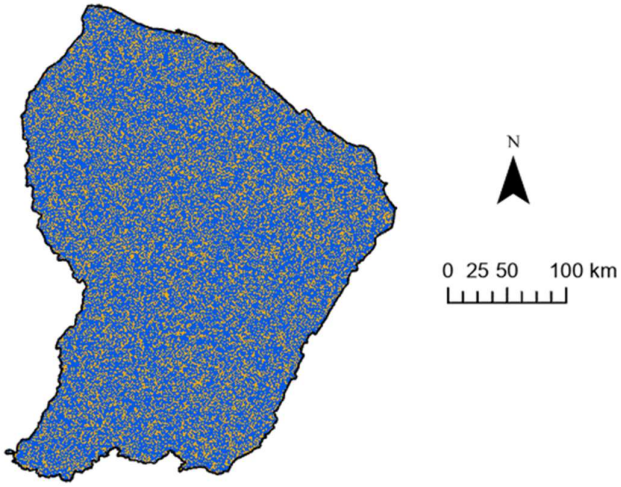


Figure 133. 1x1 km testing and training grid with French Guiana outline.

Canopy height and aboveground biomass density values roughly follow a standard distribution around 400 Mg/ha and 4000 cm respectively. To ensure that the model training is effective over a wider range, training data is moderately balanced. For testing purposes, on the other hand, a completely balanced dataset is produced, where in every interval an equal number of GEDI plots is present (Figure 134). The bin size and maximum number of plots within one bin is presented in Table 54.

Table 54. Bin size and maximum number of plots per bin for data balancing for different GEDI datasets.

GEDI Type	Balancing	Bin size	Max. number of plots
L2A – Training 2020	moderately balanced	500 cm	1500
	completely balanced	500 cm	149
L2A – Training 2023	moderately balanced	500 cm	1500
	completely balanced	500 cm	108
L2A – Testing 2020	moderately balanced	500 cm	450
	completely balanced	500 cm	44
L2A – Testing 2023	moderately balanced	500 cm	450
	completely balanced	500 cm	57
L4A – Training 2020	moderately balanced	100 Mg/ha	2500
	completely balanced	100 Mg/ha	404
L4A – Training 2023	moderately balanced	100 Mg/ha	2500
	completely balanced	100 Mg/ha	676
L4A – Testing 2020	moderately balanced	100 Mg/ha	700
	completely balanced	100 Mg/ha	186
L4A – Testing 2023	moderately balanced	100 Mg/ha	700
	completely balanced	100 Mg/ha	245

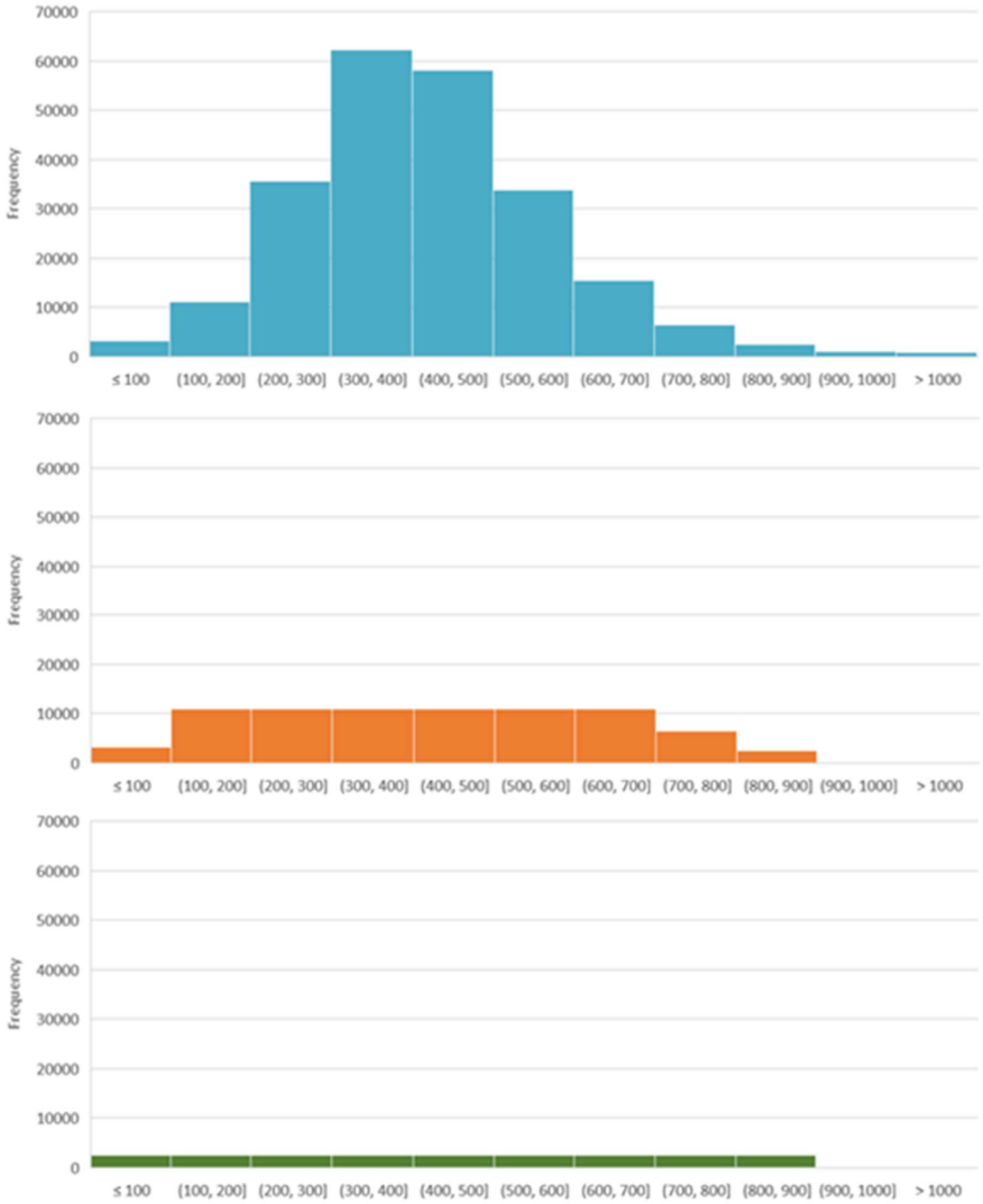


Figure 134. AGBD training data distribution for unbalanced data (blue), moderately balanced data (orange) and completely balanced data (green).

ALOS – PALSAR Mosaics

The same data and pre-processing as in the Mediterranean site is used. An exemplary site is visualized in Figure 135.

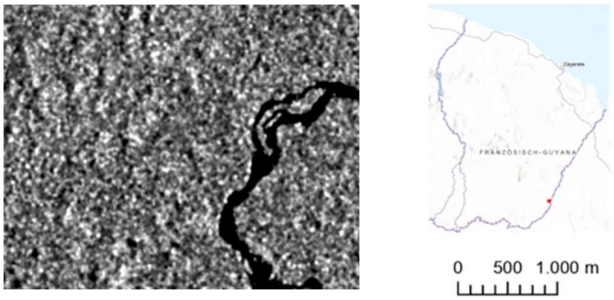


Figure 135. HV backscatter of 2020 for an example site at the border of French Guiana.

Sentinel-2 Global Mosaics

The acquisition through the Sentinel-2 Global Mosaic Service is the same as for the Mediterranean site. For every year and every 1x1° processing tile, a three-month, 10 m resolution SIAC NBAR normalized mosaic is acquired. The mosaicking period for every year is 1 August to 31 October. An exemplary site is visualized in Figure 136.

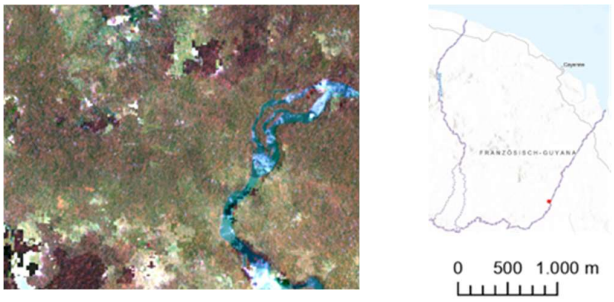


Figure 136. RGB Sentinel-2 image of 2020 for an example site at the border of French Guiana.

Planet NICFI Mosaics

The Norway's International Climate and Forest Initiative (NICFI)[17] provides monthly 5m, 4 band (RGB, NIR) mosaics over tropical areas. Yearly mosaics are calculated through the median of the 12 monthly mosaics. An exemplary site is visualized in Figure 137.

[17] https://www.planet.com/nicfi/

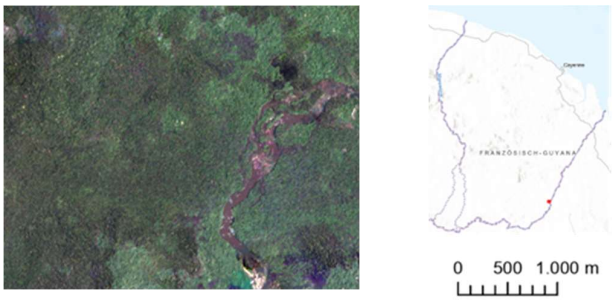


Figure 137. RGB Planet image of 2020 for an example site at the border of French Guiana.

Sentinel-1 SAR Mosaics

Sentinel-1 monthly mosaics are provided through Copernicus Data Space Ecosystem (CDSE) in the Sentinel tiling grid. The linear VV and VH gamma0 20 m monthly mosaics were recalculated to a yearly mosaic using the median. An exemplary site is visualized in Figure 138.

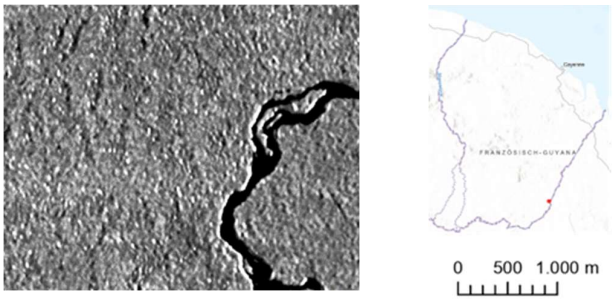


Figure 138. VV backscatter of 2020 for an example site at the border of French Guiana.

CLCplus Backbone

The CLCplus Backbone 2021 10 m raster maps 11 basic land cover classes. To be used as a basic forest mask, the classes 2 (Woody – needle leaved trees), 3 (Woody – Broadleaved deciduous trees), 4 (Woody – Broadleaved evergreen trees) and 5 (Low-growing woody plants) were reclassified to 1, all other classes to 0. The result of an exemplary site is visualized in Figure 139.

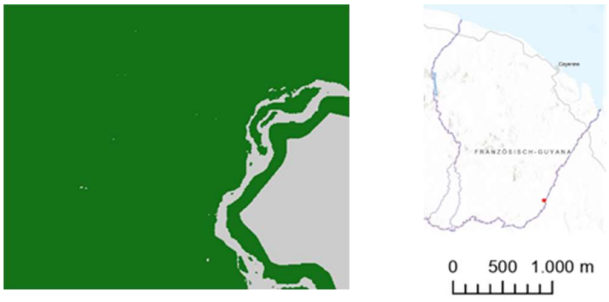


Figure 139. Forest mask based on CLCplus Backbone 2021 reclassification for an example site at the border of French Guiana. Forest areas are green, non-forest areas are grey. The large grey areas are caused by their location outside the French Guiana border.

Copernicus DEM

The 10 m Copernicus DEM is obtained from an internal JR data server. Slope and aspect were calculated from the DEM. The results of an exemplary site are visualized in Figure 140.

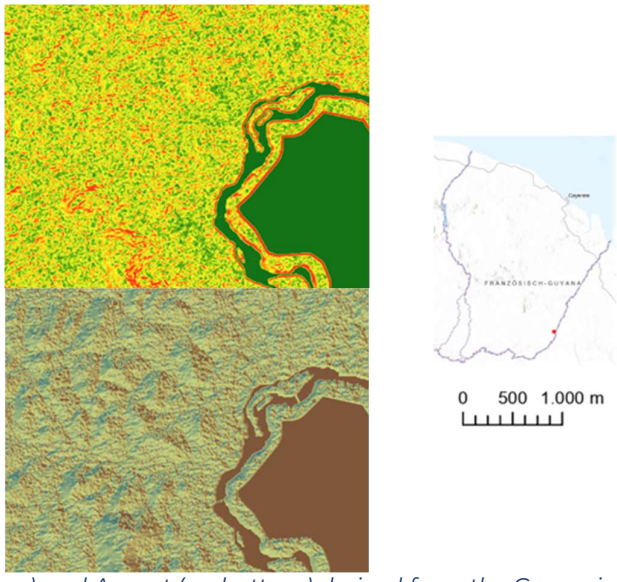


Figure 140. Slope (on top) and Aspect (on bottom) derived from the Copernicus DEM for an example site at the border of French Guiana.

5.2.2.2 Description of the candidate method

5.2.2.2.1 Non-parametric

5.2.2.2.1.1 Random Forest Regression

Based on the state of the art review, the random forest approach is one of the most promising approaches within the non-parametric methods for operational estimation of above ground biomass and forest canopy height at the continental to global level.

The regression tree machine learning algorithm has been successfully used for operational forest structure and biomass mapping at the regional to global scales (Hansen, Potapov, et al., 2016; Potapov et al., 2021). A regression tree is a machine learning algorithm that allows modelling of continuous variables. To reduce model overfitting, ensemble learning methods are usually employed, e.g. in the form of a random forest. The forest-based model creates thereby many independent decision trees, collectively called an ensemble or a forest. Each decision tree is created from a random subset of the training data and generates its own prediction. The final predictions are not based on any single tree but on the entire forest, which helps to avoid overfitting the model to the training data. Appropriate software is widely available, such as e.g. COTS tools integrated within ESRI's ArcPro software environment. Random Forest regression tools are further included in open-source

software packages such as within R[18] and also Python's scikit-learn library includes Random Forest regression tools[19].

Reference data

A basic assumption is that the area measured on the ground at each sample location can be spatially and temporally matched with the same area as viewed via remote sensing (Duncanson et al., 2022). Selection of the highest quality reference data for training and validation is therefore an important prerequisite.

A main source of reference data on above ground woody biomass and forest canopy height is from National Forest Inventories (NFI's), which is only available for a limited number of countries. This is e.g. the case for Sweden, where field measured NFI data is provided for every year, with the required high geolocation accuracy. For forests in the boreal ecosystem region, we therefore recommend using the Swedish NFI data as reference. Alternatively, ICESat-2 data could be used, which is especially relevant for global applications.

Forest ecosystems outside the boreal region we recommend using GEDI data as reference. For estimating AGB, the L4A data is publicly available. Only observations should be used as reference data which are collected in the power beam mode, during the night (to limit background noise effects of reflected solar radiation) (Potapov et al., 2021), with beam sensitivity above 0.95 (Milenković et al., 2022). Further filtering e.g. by selecting only measurements during the "leave-on" period may be applied depending on the ecosystem-region, or further quality criteria may be applied (Dubayah et al., 2022).

Model training

Optical Sentinel-2 and ALOS PALSAR satellite time series as well as parameters derived from digital elevation models serve as independent variables for the regression tree model. The sampling procedure is repeated to create n>15 independent training sets, to provide information on the robustness of the models and on the uncertainty of the estimates. We recommend calibrating the models separately for forest eco-system strata. In general, an iterative testing can be performed by comparing model out of bag errors, mean squared error. Such out of bag prediction errors are calculated using data that is a part of the training dataset that is not seen by a subset of the trees in the forest. However, in addition, for validation we recommend setting aside a portion of the reference data, which is not used at all in the training process.

Calibration

As recommended by (Milenković et al., 2022) and (Y. Zhang et al., 2020), calibration of the above ground biomass as well as the forest canopy height estimates is performed. To derive calibration models, an Ordinary Least Squares (OLS) linear

[18] https://www.randomforestsrc.org/articles/getstarted.html
[19] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

regression between the reference data and the satellite-based estimates can be applied. Only reference data, which is not used for the validation, should be used for determining the calibration coefficients. As reference data, the in-situ measured observations e.g. from a national forest inventory or LiDAR measurements such as airborne or spaceborne from GEDI / ICESat-2 is used. Open as well as commercial off the shelf (COTS) software can directly be applied for this calibration step.

5.2.2.2.1.2 UNet Regression

Beside the well-established Random Forest approach, a second approach that utilises the spatial image information and the spectral information is tested. This extra information is especially promising in biomass-rich tropical forest. Compared to an approach that is only using spectral information these image-chip based methods place higher demands on computing power. With the ongoing advances in computing infrastructure, this requirement is meanwhile well tradable. This approach is tested at the Tropical study site.

Based on the results of Dong et al. 2023, an UNet approach was chosen. The utilised architecture builds on 15x15 px image patches and features 3 encoder and 3 decoder layers plus a central bottleneck. Adam is used as the optimizer with Leaky ReLU as the activation function. Convolutional blocks are used for downsampling and feature extraction and transposed convolutions are used for upsampling. Skip connections with padding adjustments are integrated to align feature maps between encoder and decoder. The architecture is implemented in Python, using the deep learning library PyTorch[20].

Reference data
As for the Random Forest approach, pre-processed and filtered GEDI L2A and L4A plots are used for training.

Model training
Six input domains (ALOS – PALSAR Mosaics, Sentinel-2 Global Mosaics, Planet NICFI Mosaics, Sentinel-1 SAR Mosaics, CLCplus Backbone and Copernicus DEM) with 22 bands in total are used for training. Different combinations of hyperparameters and reference data preprocessing levels are compared to each other through multiple training runs.

Calibration
Calibration of the above ground biomass and the forest canopy height estimates is performed as described for the Random Forest approach.

5.2.2.2.2 Parametric

The parametric method is based on the approach described in (Bouvet et al., 2018). It aims at estimating AGB from the HH and HV backscatter of L-band SAR mosaics, through a Bayesian inversion of a model that relates HH and HV to biomass. The

[20] https://pytorch.org/

model is a simplified version of the Water Cloud Model (WCM), with 3 parameters (a, b, c), as described in section 5.2.1.3.2:

$$\gamma^0 = \gamma^0_{ground}.e^{-c.AGB} + \gamma^0_{veg}.(1-e^{-c.AGB}) = a.e^{-c.AGB} + b.(1-e^{-c.AGB})$$

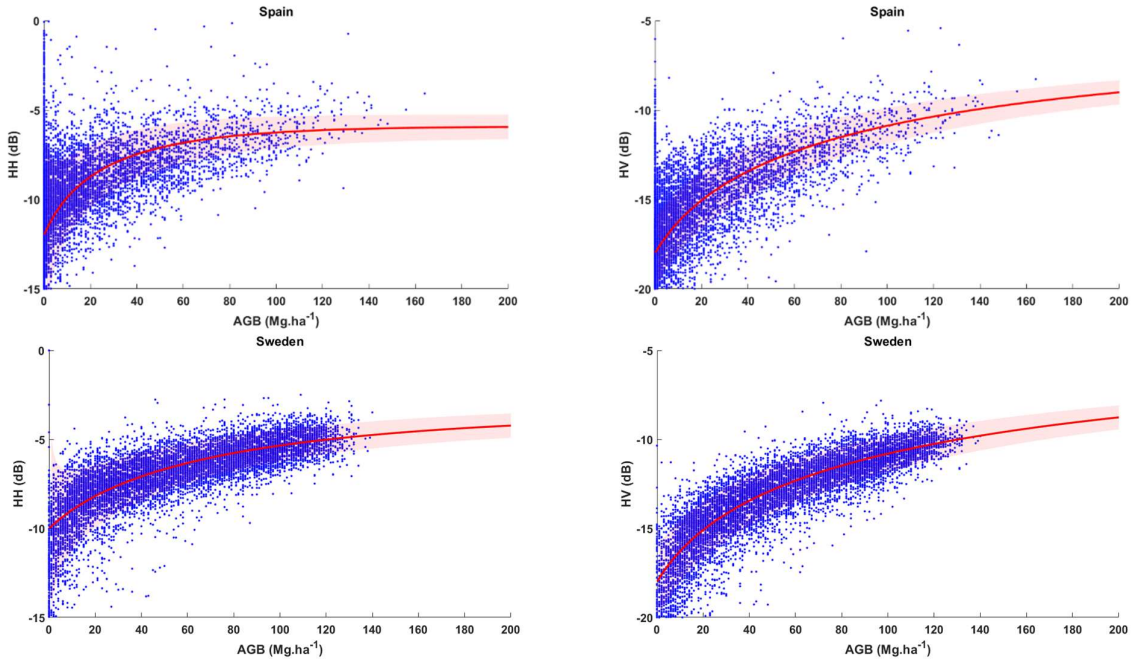
The method comprises two main steps:

Model calibration

The parameters of the model are estimated by curve-fitting approaches applied over reference AGB estimates. The AGB reference dataset used over the 3 sites is the v5.0 AGB map from the CCI Biomass project[21].

Using the same reference dataset (the CCI Biomass data) over the 3 sites allows for a better intercomparison of the results across different biomes.

Figure 141 shows the obtained regression curves of HH vs AGB and HV vs AGB for the three sites.



[21] https://catalogue.ceda.ac.uk/uuid/af60720c1e404a9e9d2c145d2b2ead4e.

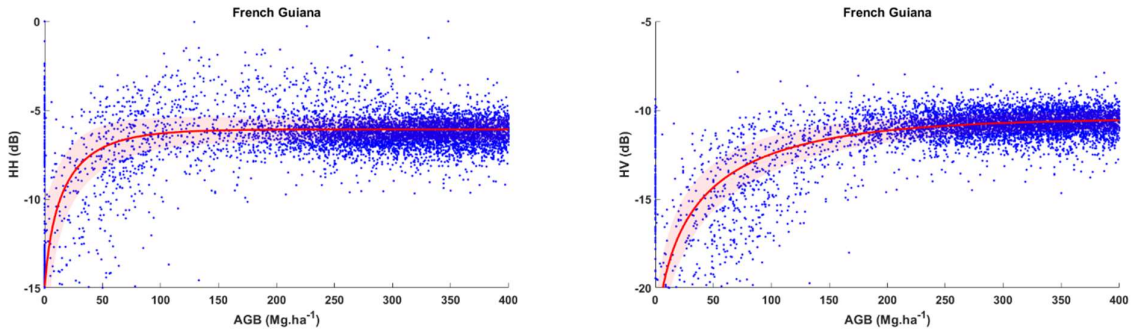


Figure 141. Calibration of the Water Cloud Model for HH (left) and HV (right) polarisations over the Sweden (top), Spain (middle) and French Guiana (bottom) sites with CCI AGB as a refence. The red curve is the regression curve obtained from the fitting. The pink area represents the standard deviation of the backscatter as modelled by MIPERS.

Furthermore, in the boreal site, we also tested to calibrate the WCM using the temporary plots of the NFI data, although the small size of each plot makes it not very suitable for the calibration of remote sensing models.

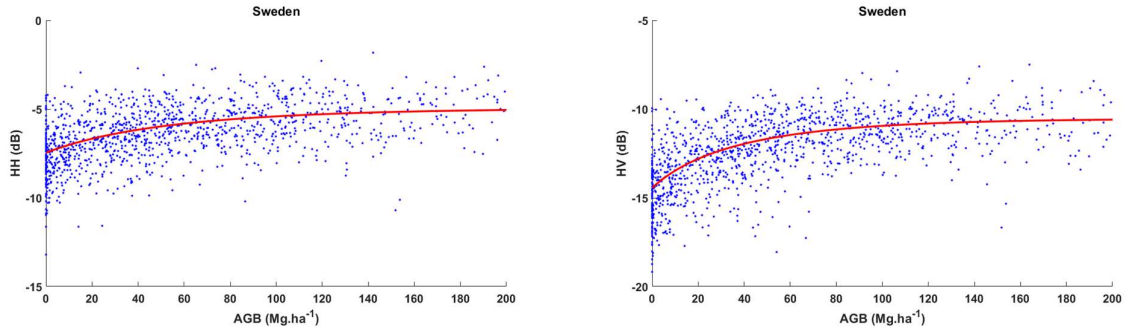


Figure 142. Calibration of the Water Cloud Model for HH (left) and HV (right) polarisations over the Sweden sites with the NFI as a reference. The red curve is the regression curve obtained from the fitting.

Bayesian inversion

The inversion maximizes the posterior probability $p(\mathrm{AGB}|\gamma^0 = \gamma^0_{obs})$ of AGB given the observation γ^o_{obs} using Bayes' theorem:

$$p(\mathrm{B}|\gamma^0_{HH\ obs}, \gamma^0_{HV\ obs}) = \frac{p(\gamma^0_{HH\ obs}, \gamma^0_{HV\ obs}|\mathrm{B})p(\mathrm{B})}{p(\gamma^0_{HH\ obs}, \gamma^0_{HV\ obs})}$$

The likelihood function $p(\gamma^0_{HH\ obs}, \gamma^0_{HV\ obs}|\mathrm{B})$ accounts for the dispersions of HH and HV backscatters caused by environmental conditions and forest structure for a given AGB value. These dispersions are considered independent and the joint conditional probability density function can thus be expressed a:

$$p(\gamma^0_{HH\ obs}, \gamma^0_{HV\ obs}|\mathrm{B}) = p(\gamma^0_{HH\ obs}|\mathrm{B}) \cdot p(\gamma^0_{HV\ obs}|\mathrm{B})$$

For a given polarization, $p(\gamma^0_{obs}|\mathrm{B})$ is modelled using a Gaussian distribution characterized by a mean equal to γ^o_{theo}, which is obtained from the regression curves obtained in the first step, and a standard deviation σ_{simu} estimated by

simulation results from the electromagnetic model MIPERS. This term σ_{simu} models the variability of γ^0_{theo} with environmental conditions or forest structure.

The term p(AGB) is the prior probability of AGB. For simplicity, this term is considered uniform in this study over the [0 AGB_{max}] range, where AGB_{max} is the assumed maximum AGB value occurring within the study area. AGB_{max} was fixed to 200Mg.ha^{-1} for the boreal and Mediterranean site, and 400Mg.ha^{-1} for the tropical site. In the future, more realistic distributions can be considered (e.g. by analysing the distribution of AGB in the NFI).

The term $p(\gamma^0)$ is a normalization factor and does not need to be computed aside of the other terms.

For a given HH and HV backscatter $\gamma^0_{HH\ obs}$ and $\gamma^0_{HV\ obs}$, the probability density function $p(\mathrm{B}|\gamma^0_{HH\ obs},\gamma^0_{HV\ obs})$ is therefore estimated as follows:

$$p(\mathrm{B}|\gamma^0_{HH\ obs},\gamma^0_{HV\ obs}) = K\frac{e^{-\frac{1}{2}\left(\frac{\gamma^0_{HH\ obs}-\gamma^0_{HH\ theo}(B)}{\sigma_{HH\ simu}(B)}\right)^2}}{\sigma_{HH\ simu}(B)}\frac{e^{-\frac{1}{2}\left(\frac{\gamma^0_{HV\ obs}-\gamma^0_{HV\ theo}(B)}{\sigma_{HV\ simu}(B)}\right)^2}}{\sigma_{HV\ simu}(B)}$$

where K is a normalization factor. This probability density function summarizes the available information about biomass. Its mode gives the maximum likelihood estimate of AGB, i.e. the most likely value of AGB for a given observation and for a uniform distribution of AGB. Here, we prefer to use the MMSE estimator which is given by:

$$AGB_{estim} = E[AGB/\gamma^0_{HH\ obs},\gamma^0_{HV\ obs}] = \int_0^{AG\ \ max} AGB.p(B/\gamma^0_{HH\ obs},\gamma^0_{HV\ obs}).dB$$

5.2.3 Implementation and results

5.2.3.1 Non-parametric

5.2.3.1.1 Random Forest Regression

In the following, selected benchmarking metrics for the boreal and Spain study site are described.

As for the boreal site, comprehensive ground measurement data from the Swedish National Forest Inventory (Swedish NFI) is available, the focus was on testing in this site. For training and benchmarking, we used the temporary plots of the Swedish NFI, which are openly available. For the year 2020, a total of n=847 and for the year 2022, a total of n=866 plots are located within the study site. For benchmarking 20% of the plots were set aside, which were not used for training, but only for statistical evaluation. For the year 2020, n= 678 plots were used for training and n=169 for evaluation. For the year 2022, n=693 plots were used for training and n=173 plots were used for evaluation. Please note, that in addition, permanent plots are used for an independent validation, which are not used in the training, nor in the testing, to allow a fully independent validation. Further benchmarking was performed in

the Spain test site, where GEDI measurements are used for training and testing. Like for the boreal site, testing data was not used in the training phase.

Benchmarking for estimating AGB:

Following figure shows the variable importance for the boreal study site based on the random forest model.

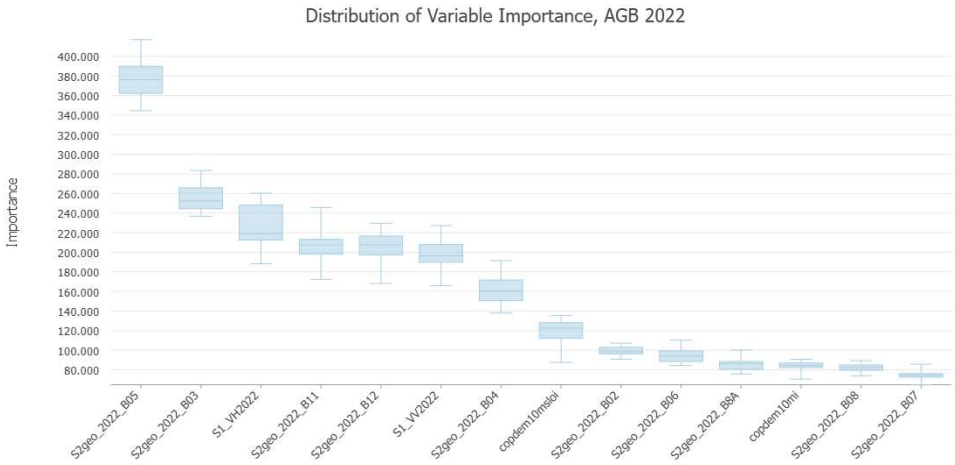


Figure 143. Boreal site, feature importance for estimation of AGB for the year 2022 based on Sentinel-2 optical bands, medoid from 1st June to 15th September 2022, the median of Sentinel1-VH and VV monthly mosaics from May to September 2022, and height and slope derived from the Copernicus-DSM.

The variable importance shows, that in addition to the optical data, also the Sentinel-1 monthly mosaics, as well as the slope derived from the 10m resolution COPERNICUS DSM contribute significantly to the information content.

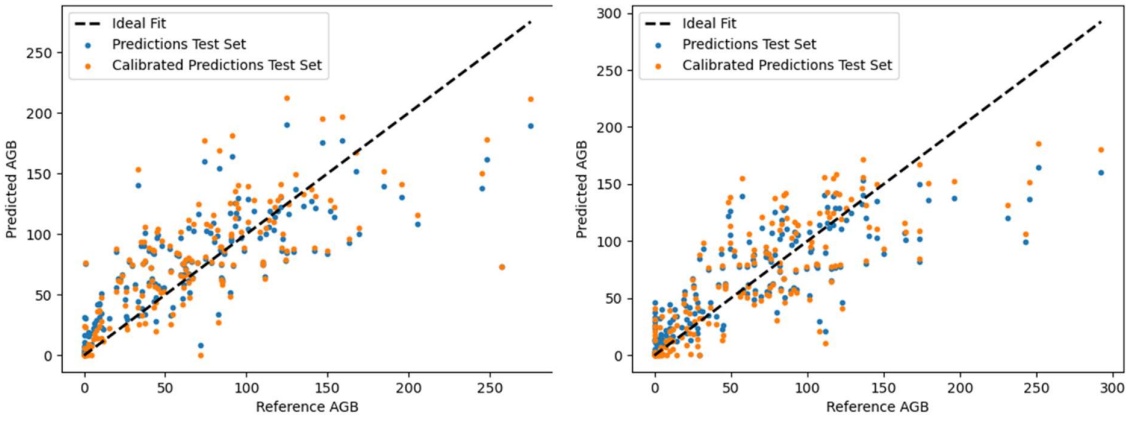


Figure 144. Boreal site, estimation of AGB for the year 2020 (left) and 2022 (right). The scattergrams show the testing plots, which were not used in the training nor in the calibration. The AGBD estimates are drawn in blue, and the estimation results after calibration are drawn in orange.

High values tend to be overestimated by the random forest approach and low values tend to be underestimated. Therefore, a linear regression is applied to calibrate the estimations, which is based on the training plots.

Various tests have been performed regarding combination of data from different sensors, see Table 55.

As in the boreal site, various tests have been performed in the Spanish test site. GEDI footprint level estimates on above ground biomass (L4A) are used for reference in the Spain site.

A high number of explanatory features have been derived from the Sentinel-2 data and filtered by applying a Gauss-kernel (sigma=0.7). The variables are listed in the variable importance figures below. QX refers to the quarterly composite, e.g. Q2 for quarter 2. PERXX refers to the percentile, e.g. PER90 for percentile 90%. MED refers to the median. CX refers to the image channel, e.g. C9 for Sentinel-2 Band 9. CXS refers to the texture feature, e.g. C4S to the standard deviation of band 4.

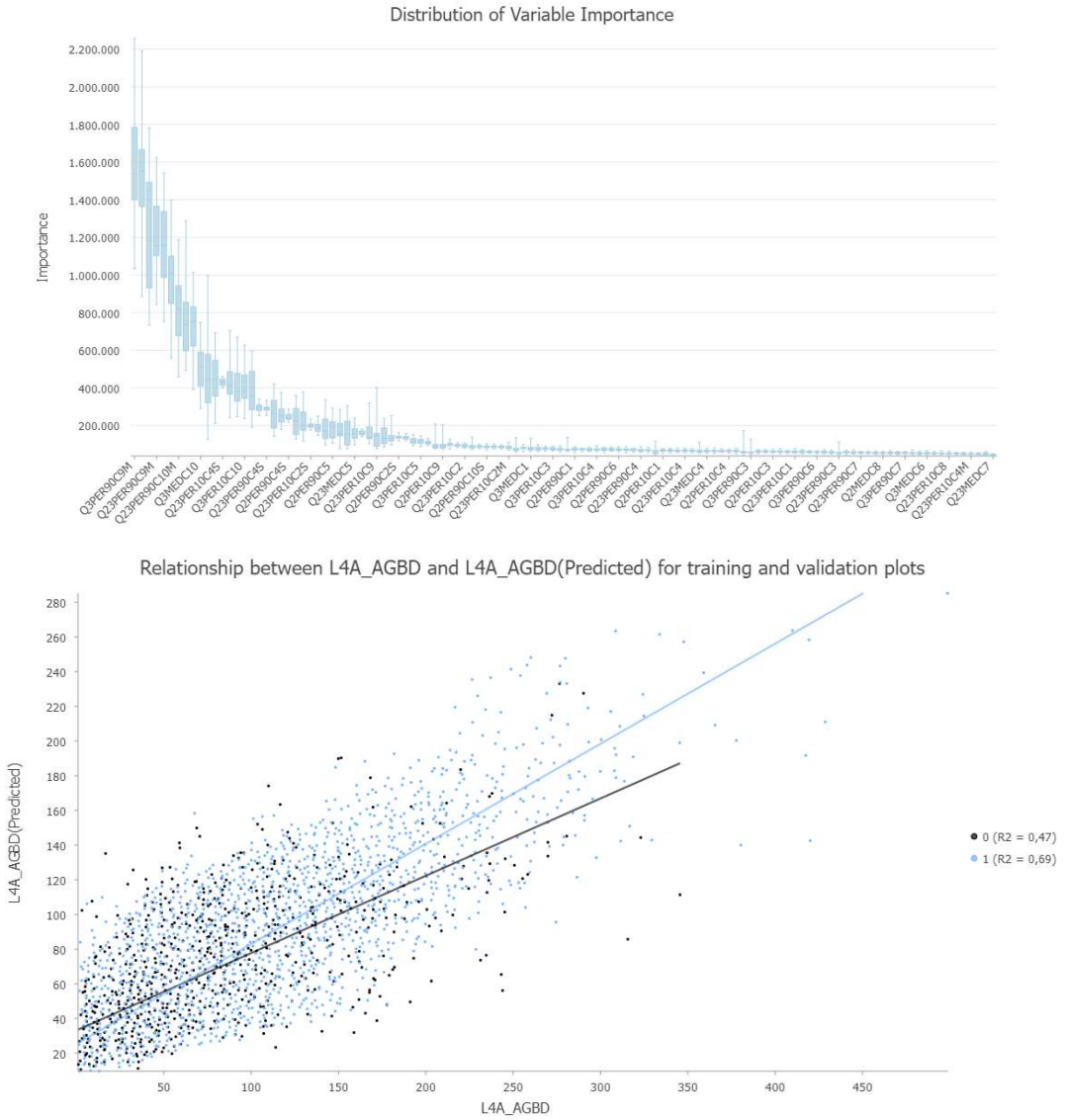


Figure 145. Selected RF model for estimating AGB in the Spain test site, based on Sentinel-2 optical features.

Table 55. Benchmarking of different sensor combinations for the random forest based estimation of AGB in the boreal test site.

Area	Sensors	r^2	MAE	RMSE
Boreal	S1-VH&VV, PALSAR-HH&HV, S2-B2/3/4/5/8/8A/11/12, CopDEM, CopDEM-Slo	0.54	29.3	41.9
Boreal	S1-VH&VV, S2-B2/3/4/5/8/8A/11/12, CopDEM, CopDEM-Slo	0.53	26.9	37.7
Boreal	S2-B2/3/4/5/8/8A/11/12, CopDEM, CopDEM-Slo	0.49	33.0	45.9
Boreal	S2-B2/3/4/5/8/8A/11/12	0.46	31.8	41.5
Boreal	S1-VH&VV, PALSAR-HH&HV, CopDEM, CopDEM-Slo	0.43	30.3	44.9
Boreal	S1-VH&VV	0.40	35.1	47.4
Spain	S2-B2/3/4/5/8/8A/11/12, CopDEM, CopDEM-Slo	0.47		

As the benchmarks demonstrate, the inclusion of the S1 monthly mosaics as well as inclusion of the Copernicus DSM leads to improved accuracies in the boreal site. Inclusion of the ALOS L-Band mosaics is not further improving the accuracy there, but leads to increased noise, and thereby to larger MAE and RMSE. The tests in the Spain study site have been performed in the first project phase at a time, when the Sentinel-1 monthly mosaics were not available. The highest correlation was found by combining the S2 optical data with the PALSAR L-Band data and parameters derived from the COPDEM.

Benchmarking for estimating Forest Canopy Height:

As for estimating AGB which is described above, the focus is on the boreal test site, because of the availability of comprehensive ground measurements from the Sweden national forest inventory. In the following, selected benchmarking metrics for the boreal as well as the Spain study sites are described.

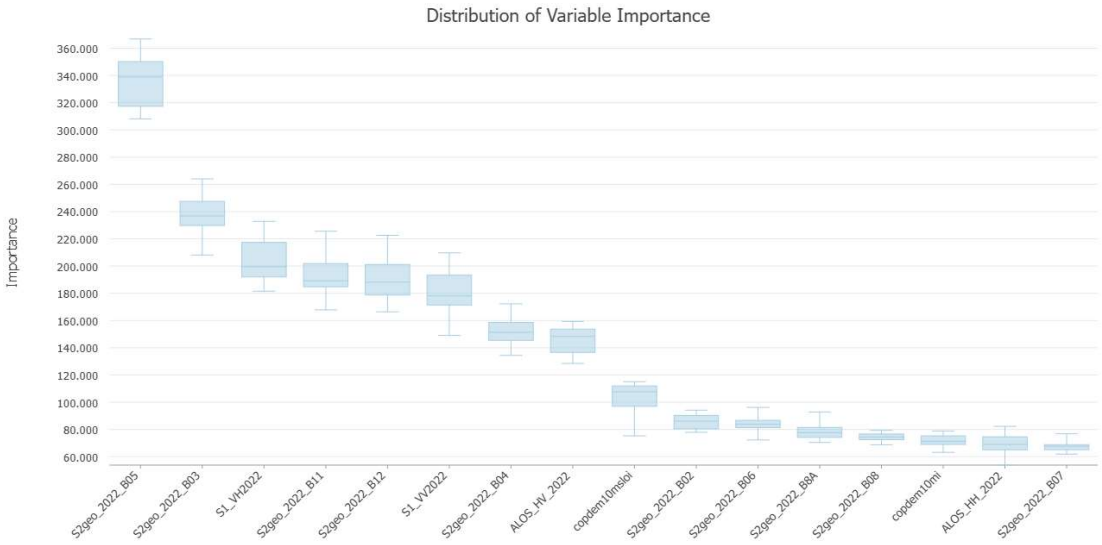


Figure 146. Boreal site, feature importance for estimation of FCH for the year 2022 based on Sentinel-2 optical bands, medoid from 1st June to 15th September 2022, the median of Sentinel1-VH and VV monthly mosaics from May to September 2022, the ALOS-PALSAR annual mosaics, and height and slope derived from the Copernicus-DSM.

The variable importance shows, that as described above for estimation of AGB, in addition to the optical data, also the Sentinel-1 monthly mosaics, as well as the slope derived from the 10m resolution COPERNICUS DSM contribute significantly to the information content.

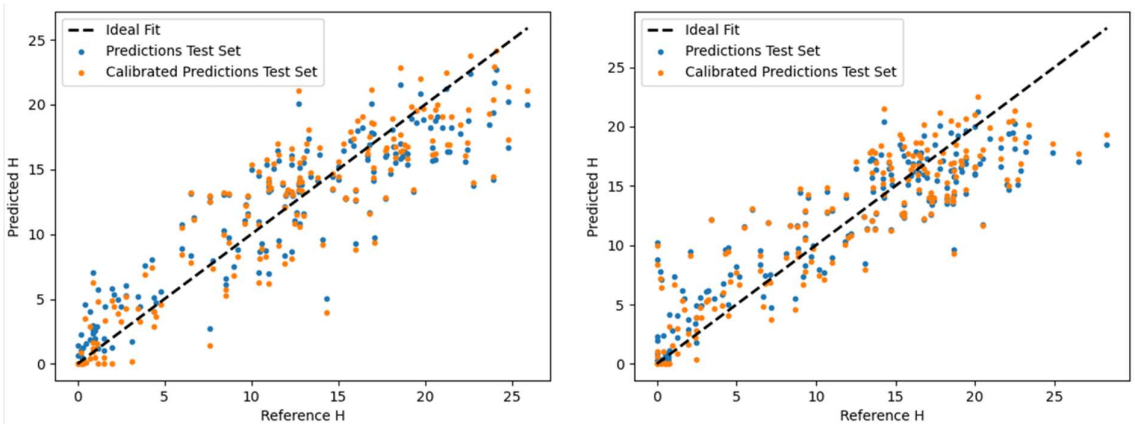


Figure 147. Boreal site, estimation of FCH for the year 2020 (left) and 2022 (right). The scattergrams show the testing plots, which were not used in the training nor in the calibration. The forest canopy height estimates are drawn in blue, and the estimation results after calibration are drawn in orange.

High values tend to be overestimated by the random forest approach and low values tend to be underestimated. Therefore, a linear regression is applied to calibrate the estimations, which is based on the training plots.

Various tests have been performed by combining data from different sensors, see Table 56.

As in the boreal site, various tests have further been performed in the Spanish test site. GEDI footprint level measurements with RH100 height (L2) are used as reference there.

A high number of explanatory features has been derived from the Sentinel-2 data and filtered by applying a Gauss-kernel (sigma=0.7). The variables are listed in the variable importance figures below. QX refers to the quarterly composite, e.g. Q2 for quarter 2. PERXX refers to the percentile, e.g. PER90 for percentile 90%. MED refers to the median. CX refers to the image channel, e.g. C9 for Sentinel-2 Band 9. CXS refers to the texture feature, e.g. C4S to the standard deviation of band 4.

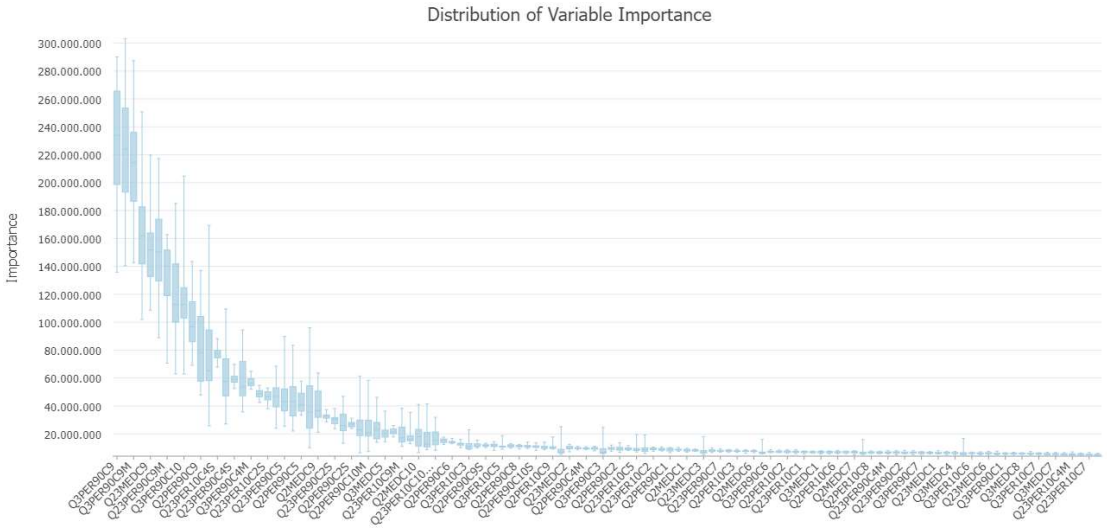




Figure 148. Selected RF model for estimating FCH in the Spain test site, based on Sentinel-2 optical features.

Table 56. Benchmarking summary table with validation results of forest canopy height.

	Sensors	r^2	MAE	RMSE
Boreal FCH	S1-VH&VV, PALSAR-HH&HV, S2-B2/3/4/5/8/8A/11/12, CopDEM, CopDEM-Slo	0.69	3.1	4.1
Boreal FCH	S1-VH&VV, S2-B2/3/4/5/8/8A/11/12, CopDEM, CopDEM-Slo	0.68	3.0	4.0
Boreal FCH	S2-B2/3/4/5/8/8A/11/12, CopDEM, CopDEM-Slo	0.65	3.0	4.1
Boreal FCH	S2-B2/3/4/5/8/8A/11/12	0.62	3.4	4.4
Boreal FCH	S1-VH&VV, PALSAR-HH&HV, CopDEM, CopDEM-Slo	0.54	3.8	4.8
Boreal FCH	S1-VH&VV, CopDEM, CopDEM-Slo	0.50	3.7	4.7
Boreal FCH	S1-VH&VV	0.47	4.0	5.0
Spain	S2-Filtered: B2/3/4/5/8/8A/11/12, CopDEM, CopDEM-Slo	0.59		

Mapping results of AGB and FCH

Based on the benchmarking the random forest models with the highest R^2 values were applied for mapping of AGB and FCH. In the following, examples of the mapping results for AGB and FCH are drawn. For the validation results see the benchmarking Table 56.

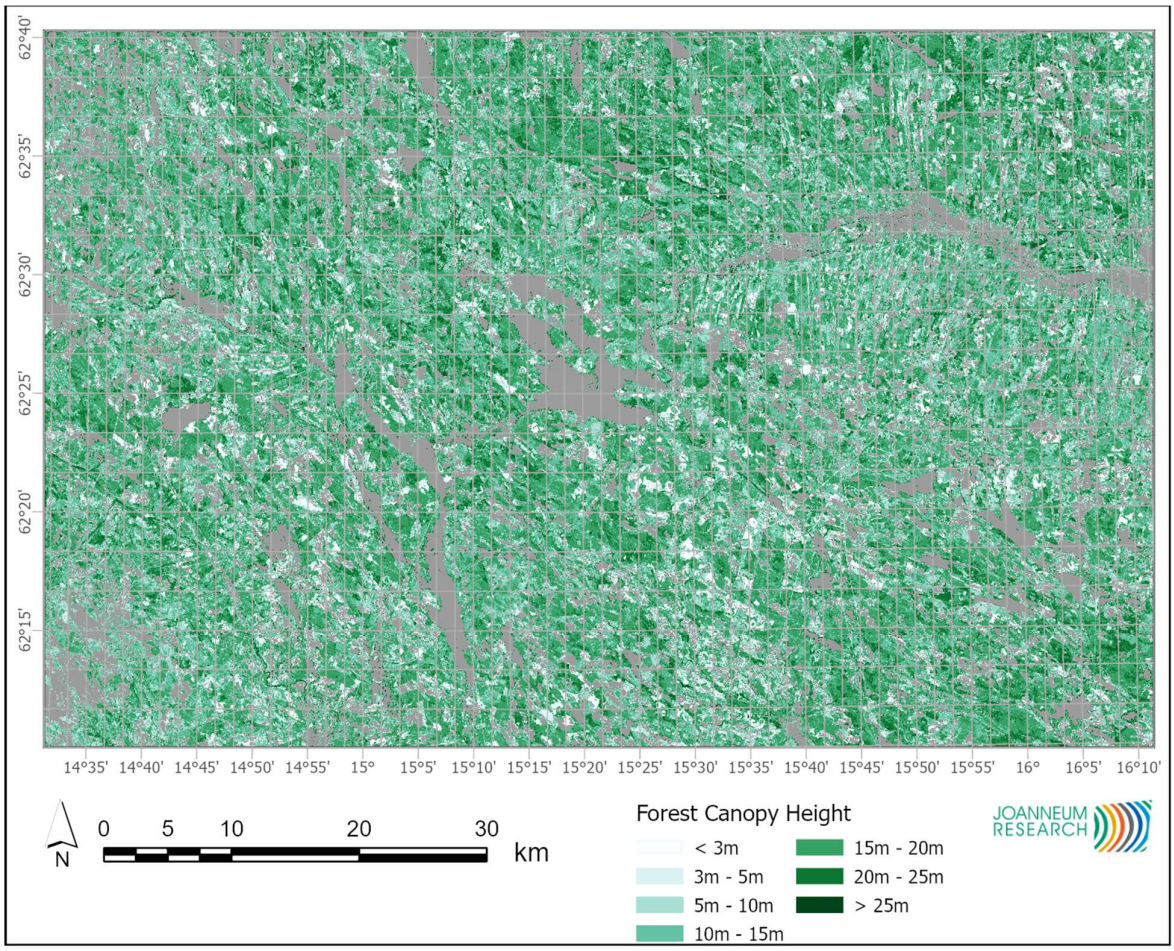
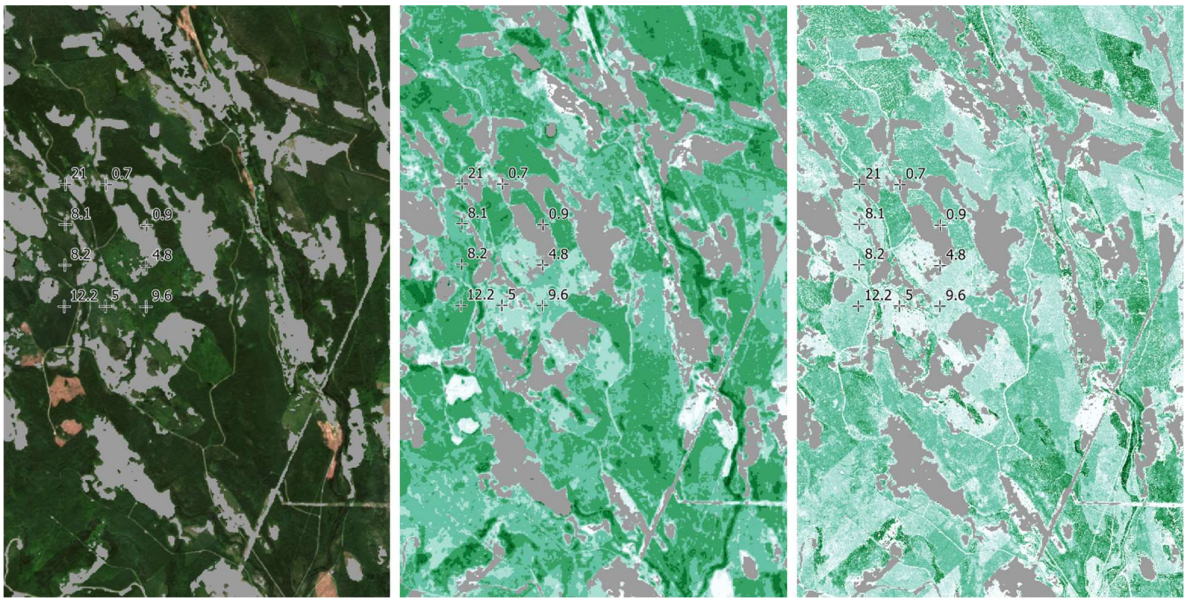


Figure 149. Result of the Forest Canopy Height estimation for the test site in Sweden.



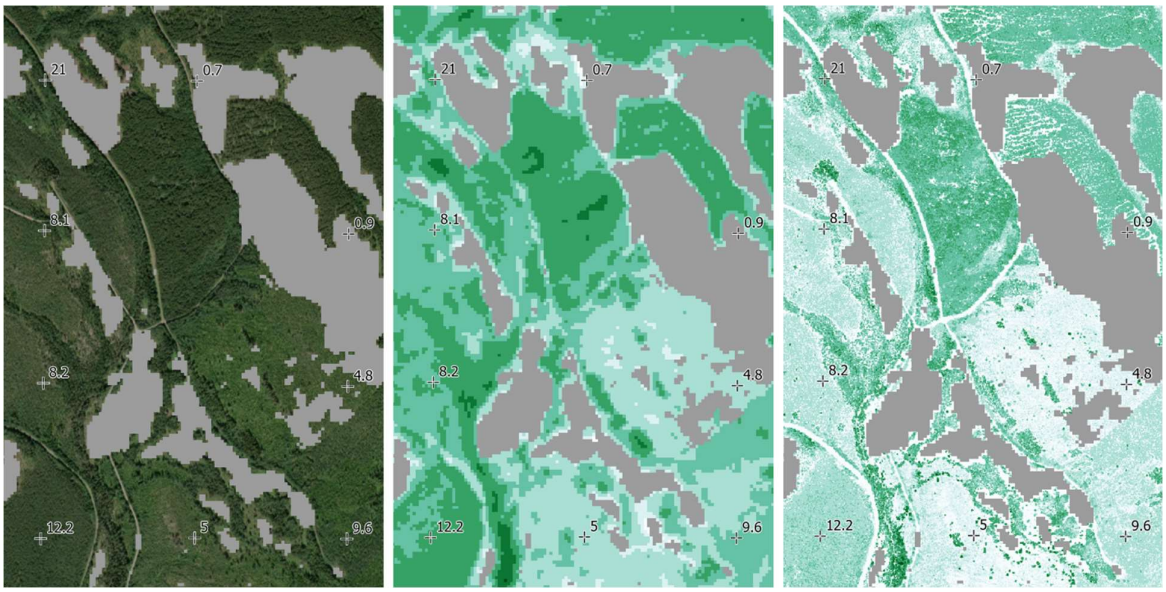


Figure 150. Detail zoom-in of the Forest Canopy Height estimation within the test site in Sweden (the crosses and labels are from NFI Sweden).

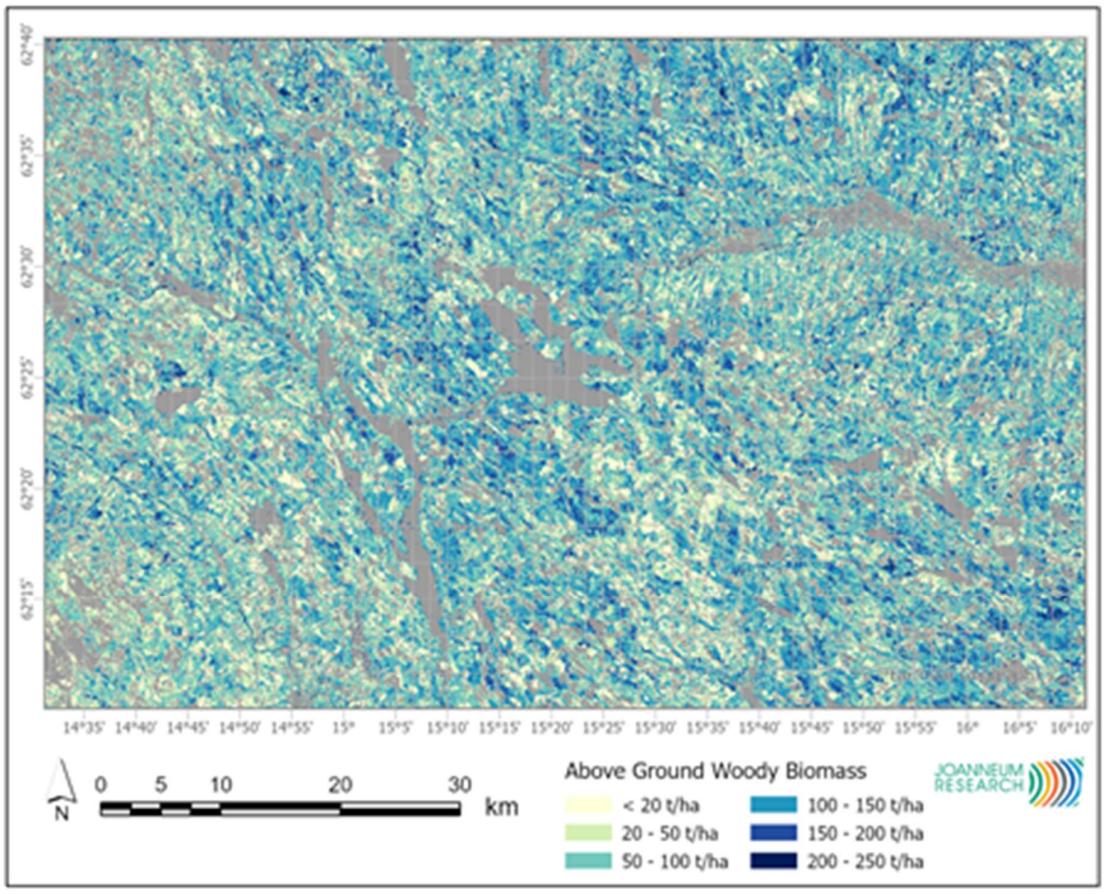


Figure 151. Result of the Above Ground Biomass estimation for the test site in Sweden.

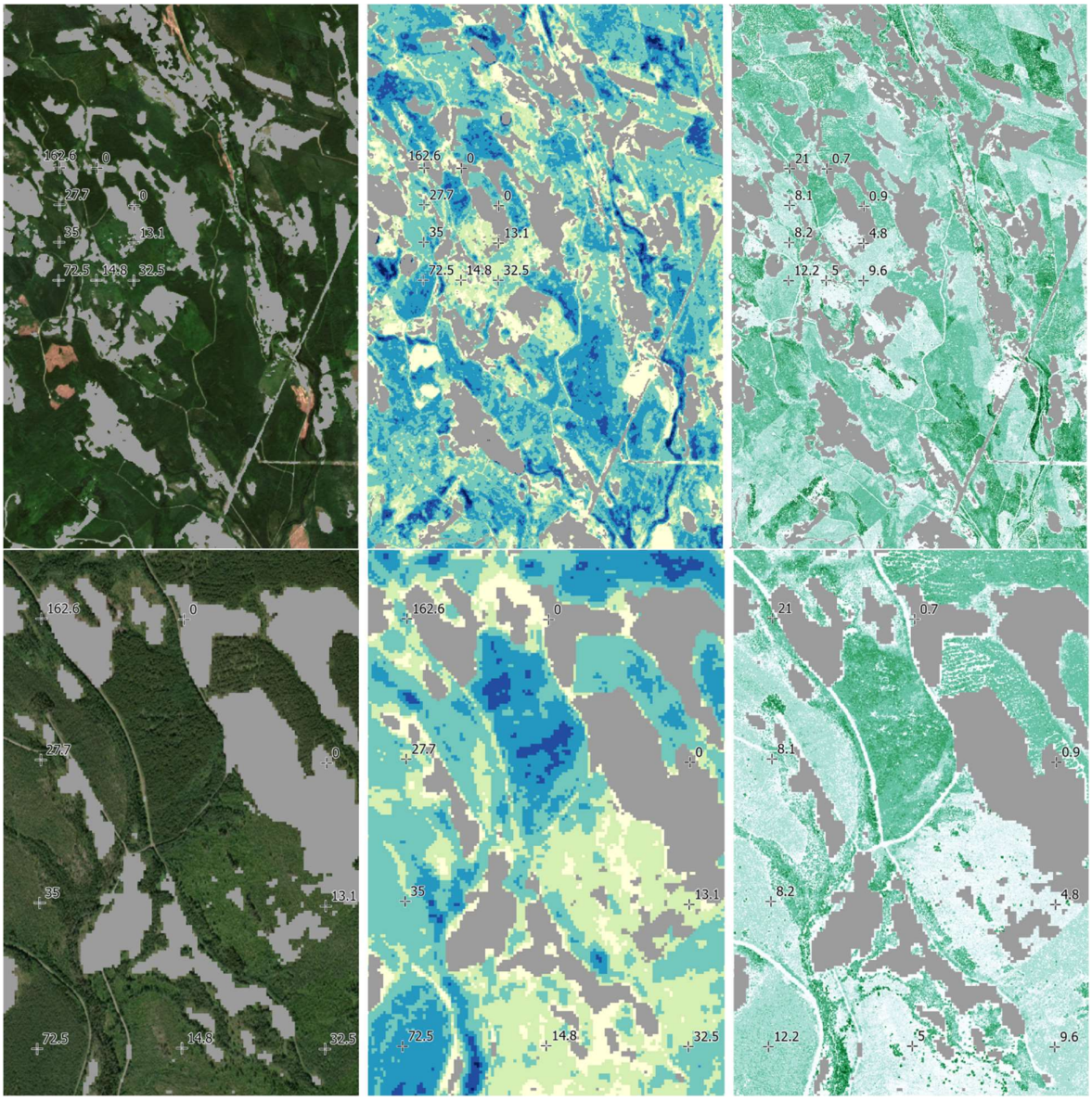


Figure 152. Detail zoom-in of the Above Ground Biomass estimation within the test site in Sweden (the crosses and labels are from NFI Sweden).

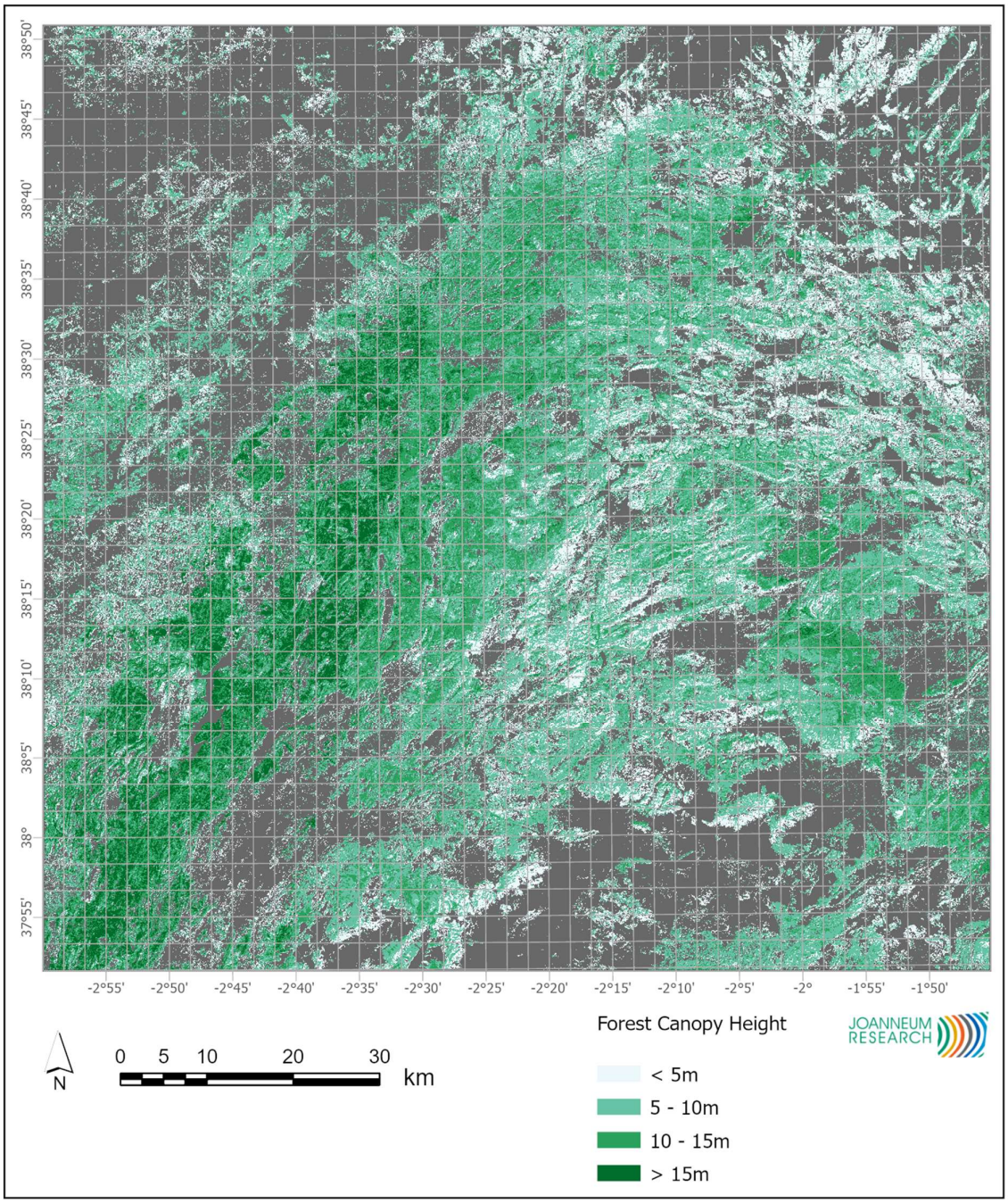


Figure 153. Result of the Forest Canopy Height estimation for the test site in Spain.

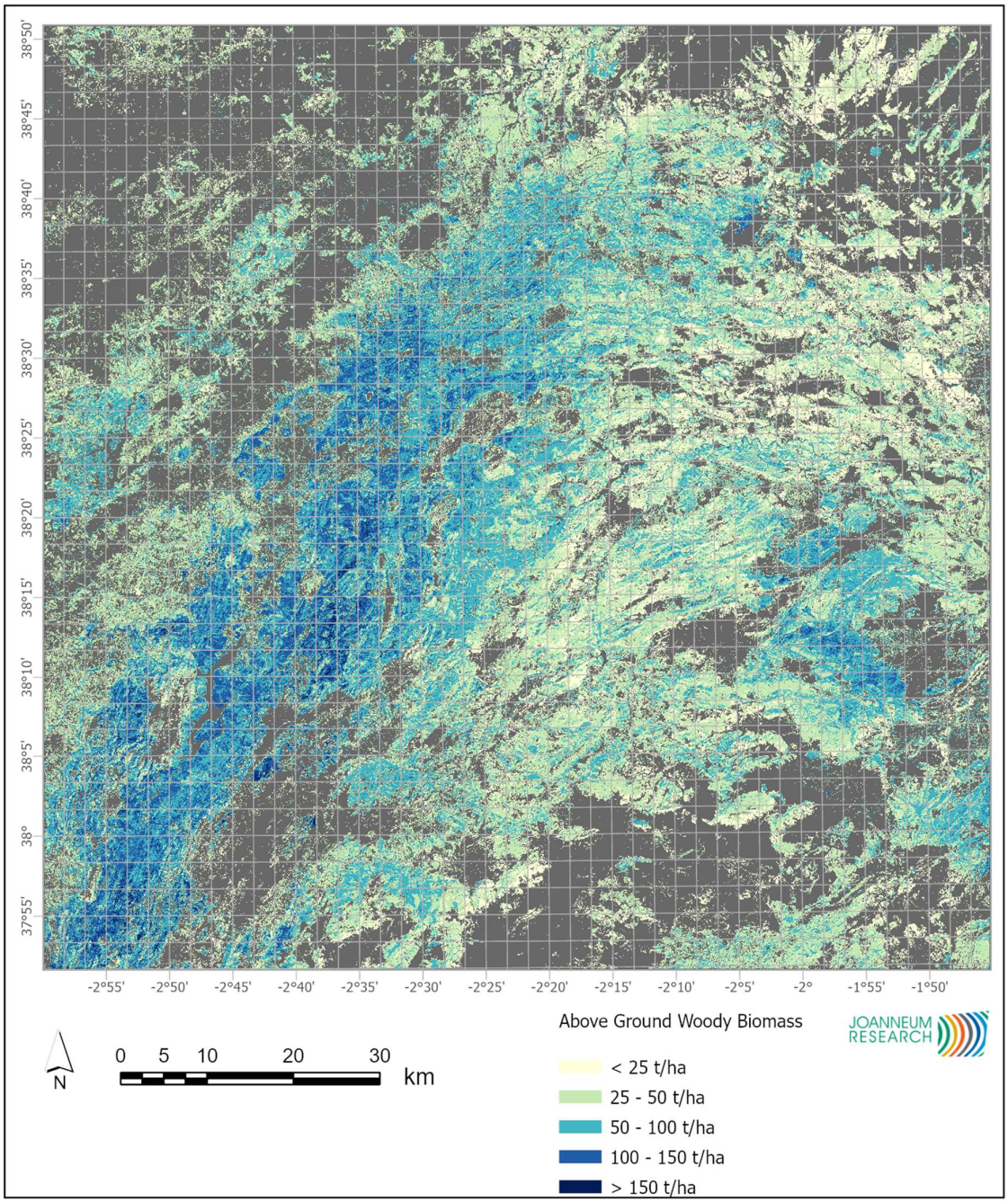


Figure 154. Result of the Above Ground Biomass estimation for the test site in Spain.

We note that despite the nominal 10-m ground sampling distance of our global map, the effective spatial resolution at which individual vegetation features can be identified is lower. Further, as a consequence of the sampling reference data used to train the model, each map pixel effectively indicates the largest canopy-top height within the footprint of the reference data, e.g. for GEDI ≈ 25 m diameter centred at the pixel.

5.2.3.1.2 UNet Regression

Benchmarking

In the following, selected benchmarking metrics for the tropical study site are described. Various variations of the UNet model and training data processing versions were combined and compared. Each result presented in Table 57 and Table 58 is the best result after multiple hyperparameter tuning runs. For both, FCH and AGBD, the best performing model measured by Pearson correlation coefficient is the following combination: Moderately balanced training data and a UNet sub-model that utilises Sentinel-2 Global Mosaics, Planet NICFI Mosaics, Sentinel-1 SAR Mosaics, CLCplus Backbone and Copernicus DEM. ALOS – PALSAR Mosaics are not part of this sub-model as it does not noticeably improve the result. These comparisons are based on the data of 2020.

Table 57. RMSE and Pearson correlation coefficient values for FCH 2020 for different combinations of UNet model type and training dataset processing.

RMSE	moderately balanced and augmented	moderately balanced	completely balanced and augmented	moderately balanced	unbalanced
without Slope and Aspect	1278.4	1268.1	1369.7	1370.6	1300.9
without PALSAR, Slope and Aspect	1281.7	1259.7	1365.8	1393.8	1300.2
without PALSAR, CLC+, Slope and Aspect	1276.0	1448.4	1516.9	1611.5	1362.0
without PALSAR and Slope	1256.2	1248.1			
without PALSAR and Aspect	1237.5	1228.4			
without PALSAR	1206.5	1200.2	1266.1		
without CLC+, Slope and Aspect	1281.9	1404.0	1482.3	1546.5	1312.3
without NICFI, Slope and Aspect	1292.6	1260.5	1368.6	1379.4	1318.6
without Sentinel 1, Slope and Aspect	1301.4	1296.1	1380.9	1389.4	1346.8
without Sentinel 2, Slope and Aspect	1355.4	1372.4	1416.5	1466.3	1394.5
without Sentinel 2 60 m bands, Slope and Aspect	1274.0	1285.7	1386.1	1371.3	1310.4

Pearson Correlation Coefficient	moderately balanced and augmented	moderately balanced	completely balanced and augmented	moderately balanced	unbalanced
without Slope and Aspect	0.777	0.779	0.738	0.737	0.766
without PALSAR, Slope and Aspect	0.775	0.782	0.743	0.730	0.766
without PALSAR, CLC+, Slope and Aspect	0.776	0.709	0.682	0.616	0.741
without PALSAR and Slope	0.784	0.786			
without PALSAR and Aspect	0.790	0.793			
without PALSAR	0.801	0.803	0.779		
without CLC+, Slope and Aspect	0.775	0.725	0.696	0.656	0.761
without NICFI, Slope and Aspect	0.771	0.782	0.739	0.737	0.759
without Sentinel 1, Slope and Aspect	0.766	0.768	0.733	0.733	0.747
without Sentinel 2, Slope and Aspect	0.744	0.737	0.718	0.695	0.727
without Sentinel 2 60 m bands, Slope and Aspect	0.777	0.772	0.734	0.739	0.763

Table 58. RMSE and Pearson correlation coefficient values for AGBD 2020 for different combinations of UNet model type and training dataset processing.

RMSE	moderately balanced and augmented	moderately balanced	completely balanced and augmented	moderately balanced	unbalanced
without Slope and Aspect	198.7	197.8	202.5	212.2	197.2
without PALSAR, Slope and Aspect	196.9	198.4	202.8	214.0	197.5
without PALSAR, CLC+, Slope and Aspect	200.2				
without PALSAR and Slope	194.3	195.0			
without PALSAR and Aspect	187.2	187.8			
without PALSAR	185.0	180.9	194.8		
without CLC+, Slope and Aspect	197.3	206.9	212.9	229.3	202.9
without NICFI, Slope and Aspect	200.1	201.7	204.9	214.0	200.3
without Sentinel 1, Slope and Aspect	199.8	201.1	204.7	213.8	200.5
without Sentinel 2, Slope and Aspect	206.3	210.1	214.3	221.1	209.3

Pearson Correlation Coefficient	moderately balanced and augmented	moderately balanced	completely balanced and augmented	moderately balanced	unbalanced
without Slope and Aspect	0.620	0.621	0.595	0.528	0.625
without PALSAR, Slope and Aspect	0.628	0.618	0.595	0.515	0.623
without PALSAR, CLC+, Slope and Aspect	0.610				
without PALSAR and Slope	0.645	0.637			
without PALSAR and Aspect	0.678	0.675			
without PALSAR	0.689	0.712	0.654		
without CLC+, Slope and Aspect	0.625	0.567	0.538	0.376	0.591
without NICFI, Slope and Aspect	0.610	0.599	0.581	0.514	0.607
without Sentinel 1, Slope and Aspect	0.611	0.601	0.579	0.516	0.605
without Sentinel 2, Slope and Aspect	0.569	0.543	0.518	0.457	0.549

Detailed performance indicators for the best performing model combinations for FCH and AGBD are shown in Table 59. It can be seen that the indicators for 2020 are showing better numbers than for 2023. The scatter plots of the testing dataset are shown in Figure 155. A saturation above 5000 cm canopy height and 500 Mg/ha is apparent.

Table 59. Performance indicators for the 2020 and 2023 FCH and AGBD UNet models.

	# train	rRMSE	RMSE	ME	r	R^2
FCH French Guiana 2020 UNet	14799	0.339	1200.2 Mg/ha	60.3 Mg/ha	0.803	0.645
FCH French Guiana 2023 UNet	16456	0.412	1453.1 Mg/ha	19.2 Mg/ha	0.692	0.479
AGBD French Guiana 2020 UNet	18311	0.401	180.9 cm	2.3 cm	0.712	0.507
AGBD French Guiana 2023 UNet	21064	0.483	217.5 cm	0.3 cm	0.544	0.296

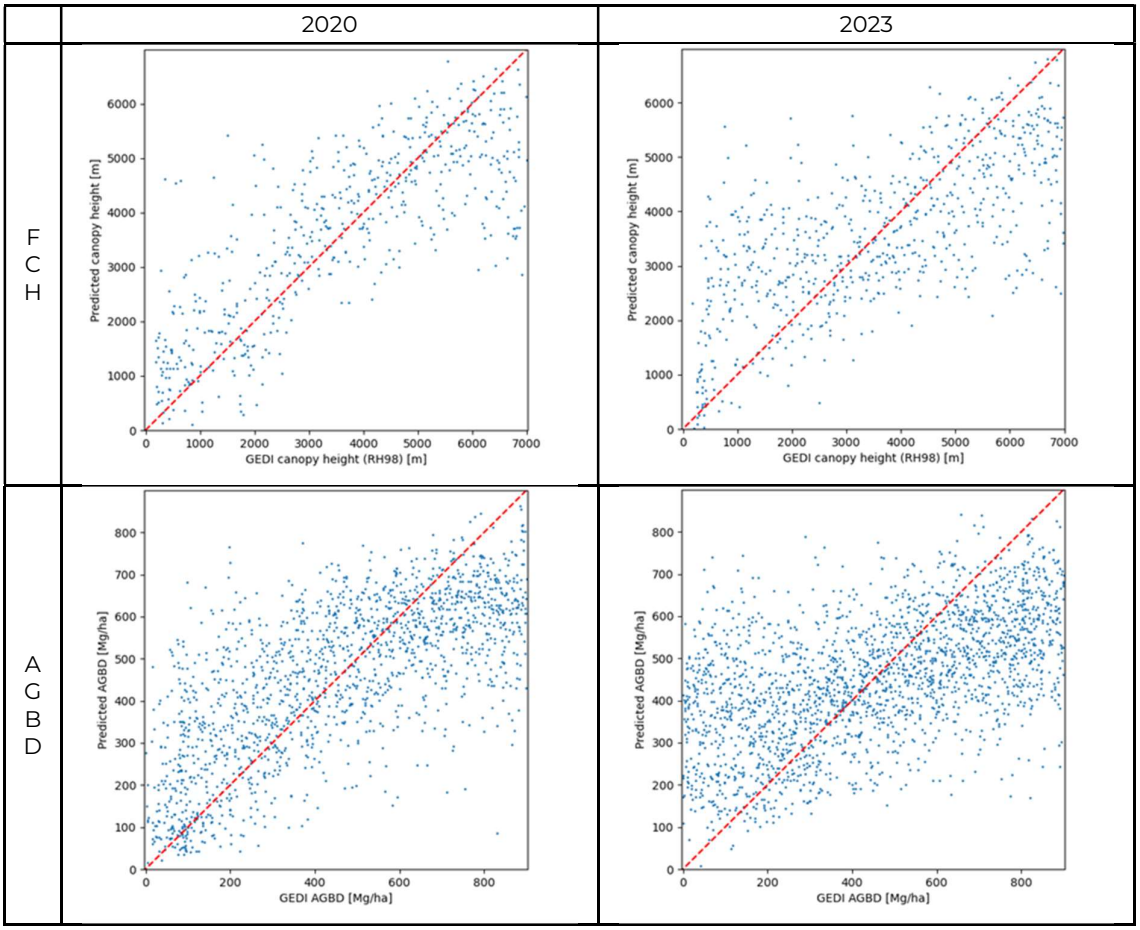


Figure 155. Testing Plots for all four trained UNet models (2020 and 2023, FCH and AGBD). Note that units for FCH are cm.

Mapping results

Mapping results of FCH and AGBD present a good visual representation (Figure 156 and Figure 157). Non-forest areas like rivers, streets and agricultural areas are correctly mapped. However, two main weaknesses of the trained model are visible. Firstly, around non-forest areas constant border effects are visible. This border areas are around 40 m wide and show medium canopy height and biomass levels (Figure 158). Secondly, even after the calibration step, a bias is present in the mapping results. 2020 FCH and AGBD values are overall higher than 2023 values. This observation is confirmed when comparing the UNet mapping results to the TropiSAR biomass data. The 2020 Unet values are higher than the TropiSAR values and the 2023 UNet values are lower than the TropiSAR values (Figure 159). We expect that the current implementation of the UNet is still too sensitive against variations in the input data quality. Further model development or more robust input data could help to mitigate this problem.

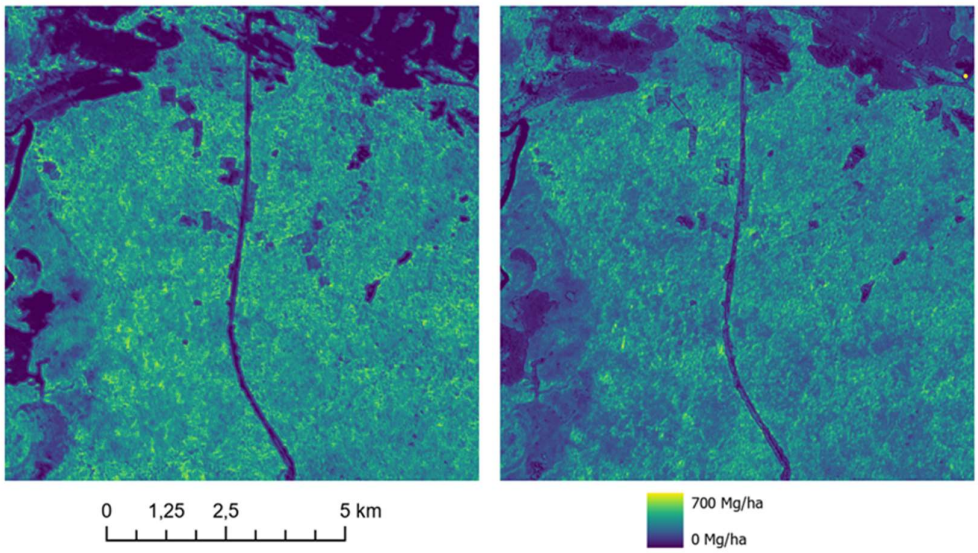


Figure 156. Example (52.91°W 5.29°N) map of AGBD for 2020 (left) and 2023 (right).

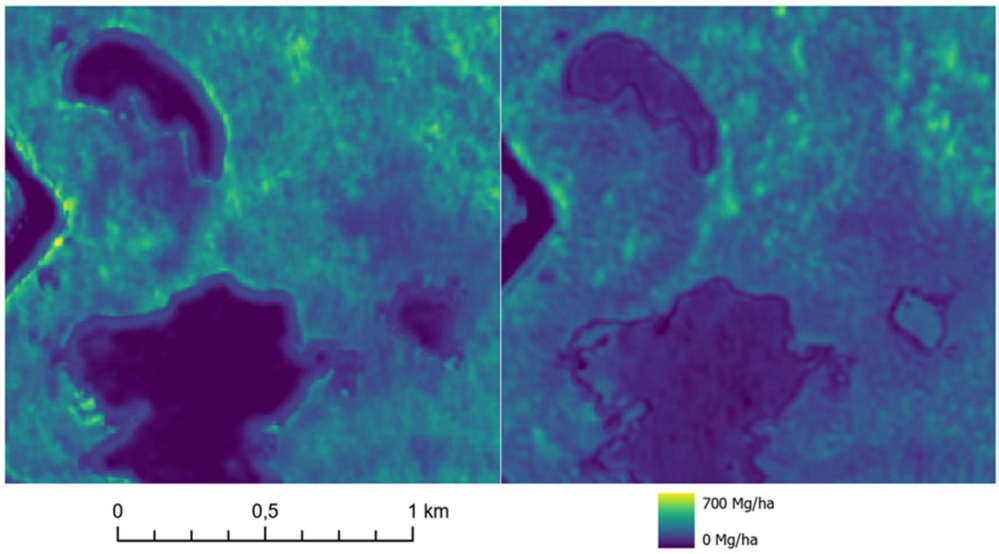


Figure 157. Detailed example (52.95°W 5.28°N) map of AGBD for 2020 (left) and 2023 (right).

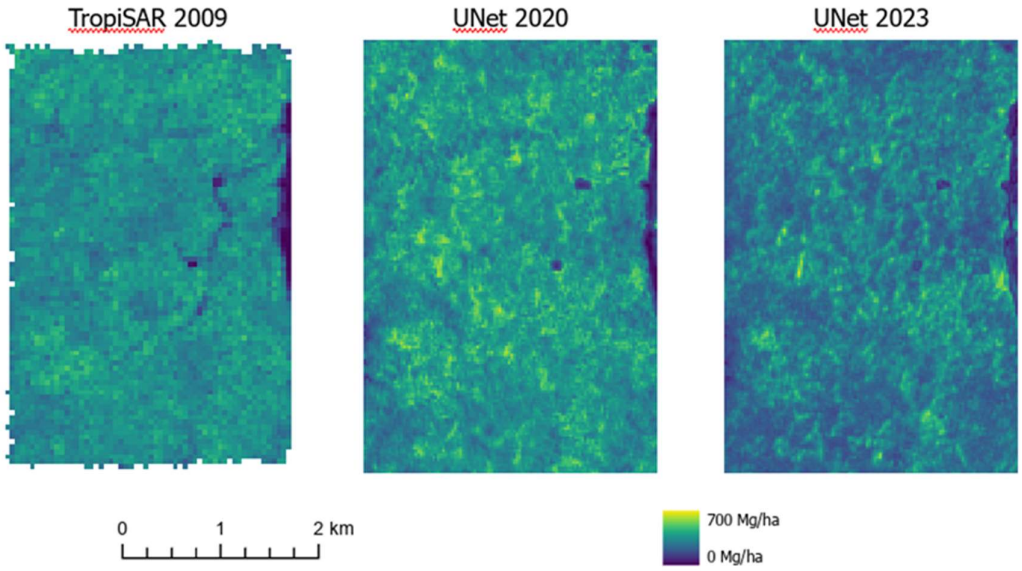


Figure 158. Comparison of TropiSAR Lidar data (2009) and UNet (2020 and 2023) maps at the Paracou site.

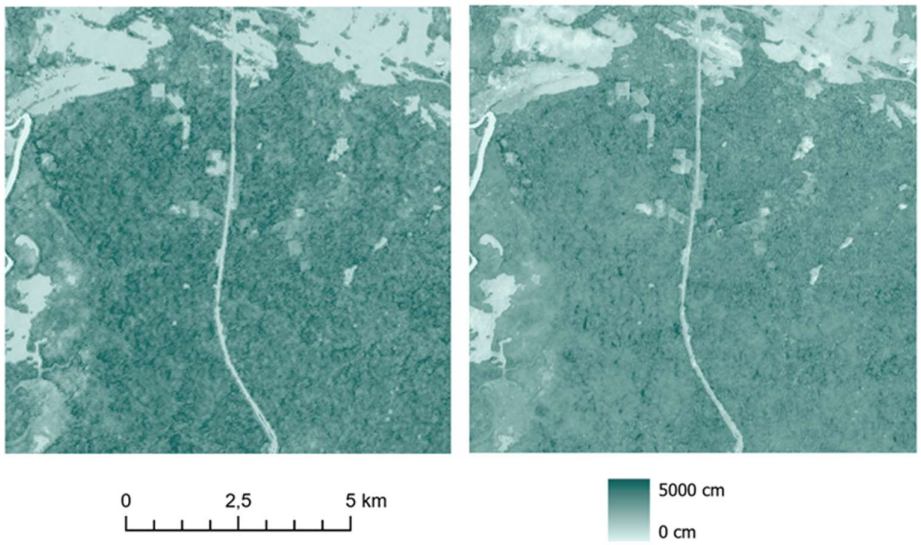


Figure 159. Example (52.91°W 5.29°N) map of FCH for 2020 (left) and 2023 (right).

5.2.3.2 Parametric

After AGB is estimated with the Bayesian inversion of the WCM, some post-processing steps are applied to mask out non-forest areas. To do so, we use the ESA WorldCover products and mask out the classes 40 (crops), 50 (urban areas) and 80 (permanent water).

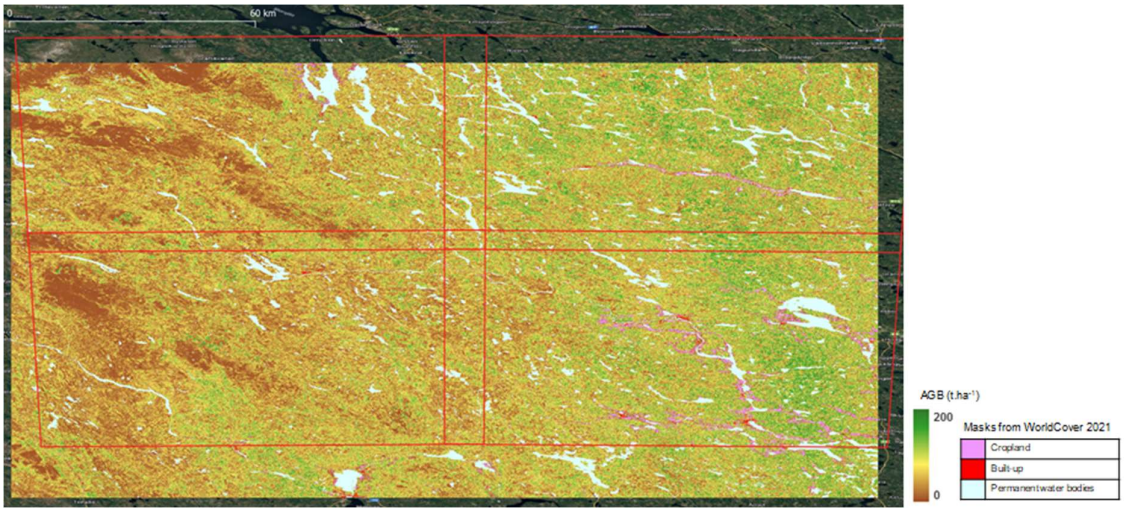


Figure 160. AGB map of the boreal site from the parametric approach for year 2020, calibration with CCI AGB 5.0.

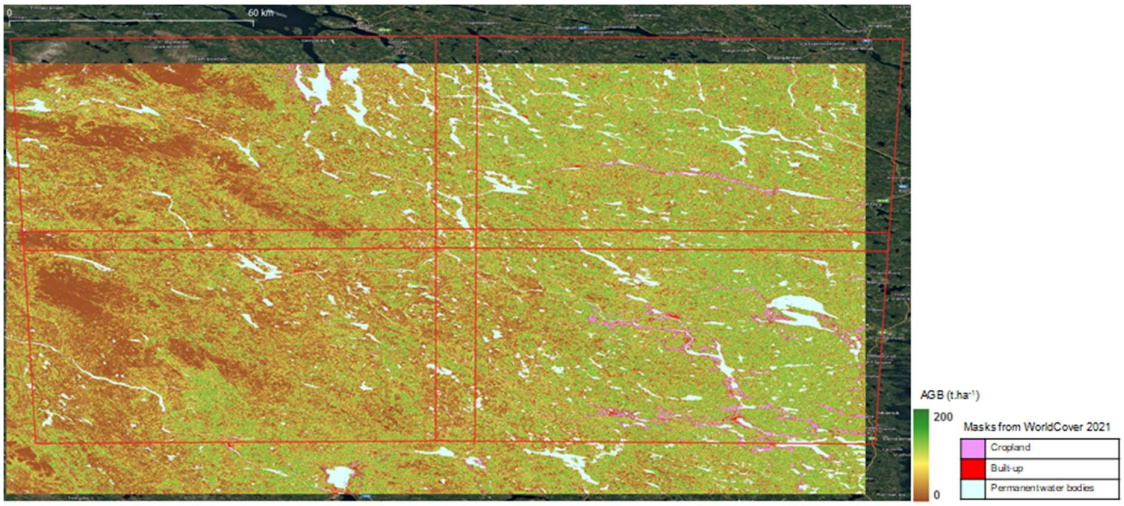


Figure 161. AGB map of the boreal site from the parametric approach for year 2020, calibration with NFI.

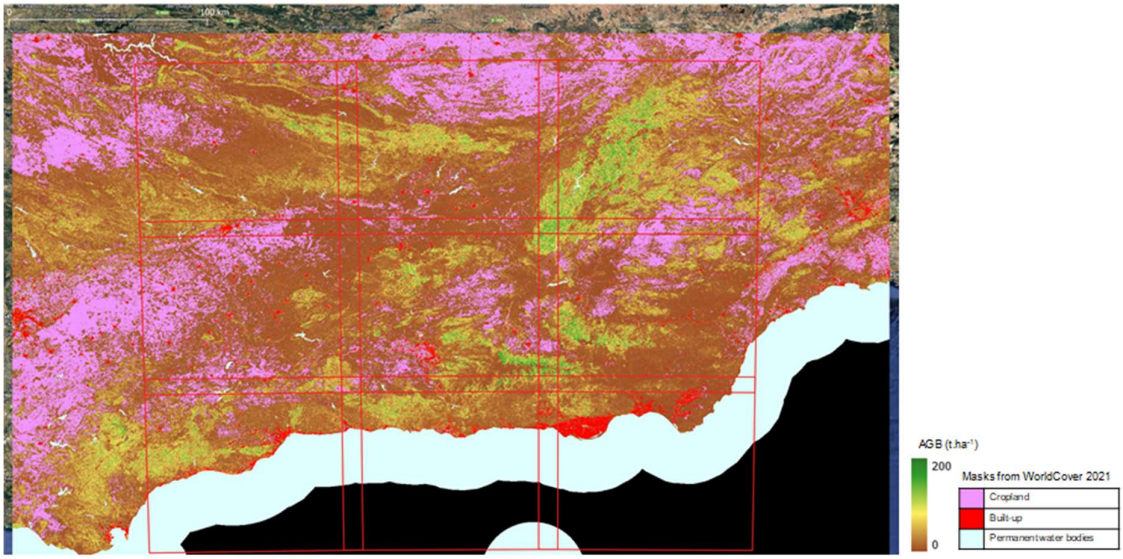


Figure 162. AGB map of the Mediterranean site from the parametric approach for year 2020, calibration with CCI AGB 5.0.

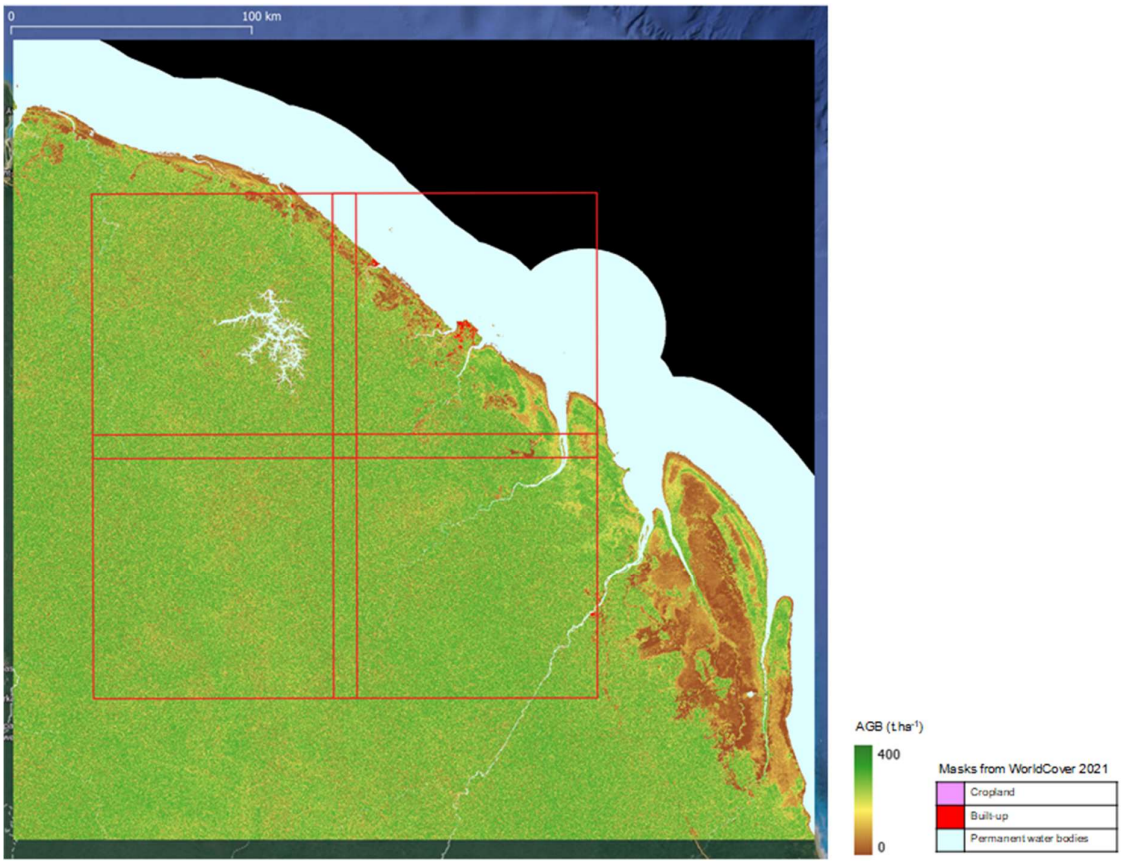


Figure 163. AGB map of the tropical site from the parametric approach for year 2020, calibration with CCI AGB 5.0.

5.2.4 Validation and comparison with other results

To allow a comparison of different approaches, we derived statistical quality measures for the boreal site, as for this site, comprehensive ground measurements from the Swedish National Forest Inventory are available for the year for which the estimations are performed. Whereas the publicly available temporary NFI plots were used for training and calibration, which is described above, only permanent plots from the Swedish NFI were used for the independent validation, which were used solely for the validation. Please note, that the temporary plots, which were used for training and calibration represent a reference unit with a circle of a diameter of 14 meters, whereas the permanent plots, which are only used for validation, represent a reference unit with a circle of a diameter of 20 meters. The below described validation results reflect the quality when ground measured NFI plot data is available for training and calibration. In case when GEDI space based measurements are used as input, the quality is in general lower. In such cases, accuracy measures can be derived from the testing data (e.g. GEDI data which was not used for training or calibration). These accuracies are given in the previous chapters, but it should be noted that these are not independent validations, as the

testing data (e.g. GEDI) can be subject to systematic inaccuracies as well as the satellite estimates.

5.2.4.1 Above Ground Woody Biomass

In the following, the results of different approaches for estimating above ground woody biomass are compared, specifically, (i) results based on random forest and calibration developed by project partner JR, and (ii) parametric estimation approach developed by UT3 and the current version 5.0 of the CCI maps are compared. Statistics within the scattergram plots allow a quantitative comparison. For a qualitative comparison, e.g. to get an impression on the effects of varying spatial resolution or noise levels, a small subset within the demonstration site is drawn in below figures.

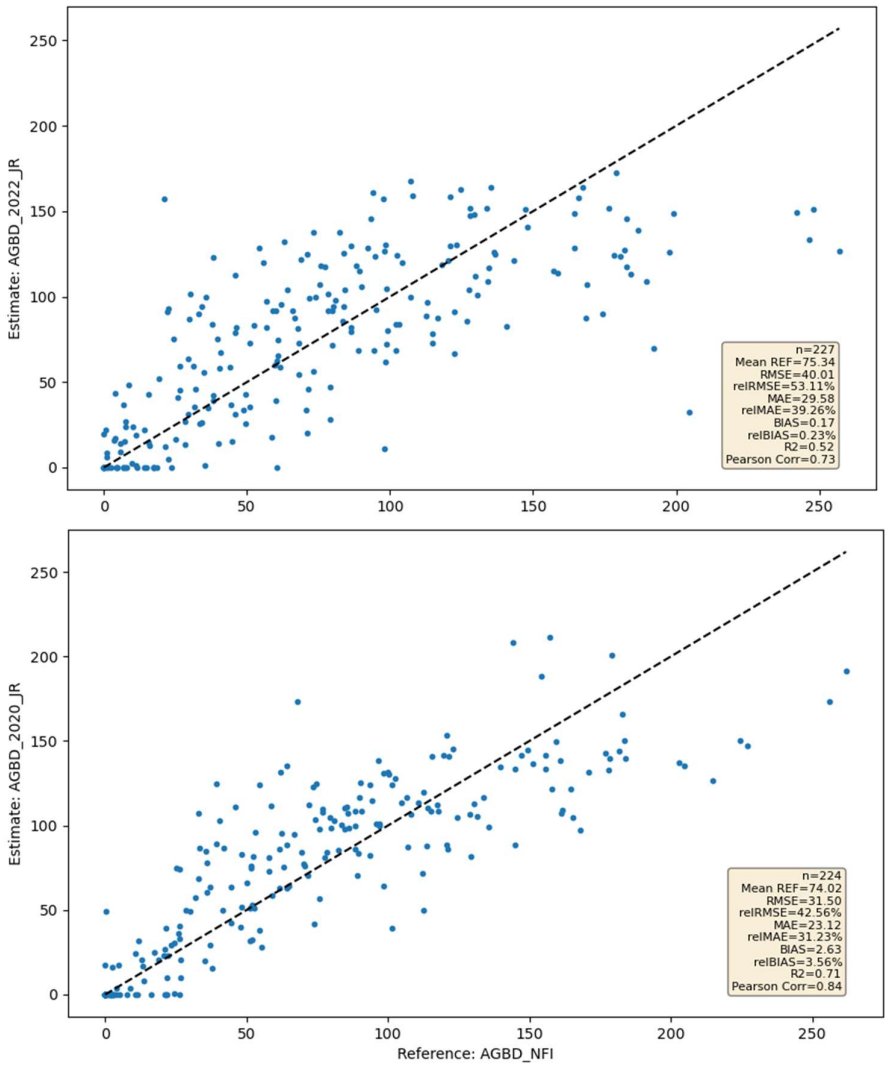


Figure 164. Validation result: AGB 2022 (above) and 2020 (below) for the Sweden demonstration site. Approach by JR with random forest estimation and calibration based on temporary NFI plots. The validation is based on the permanent NFI plots.

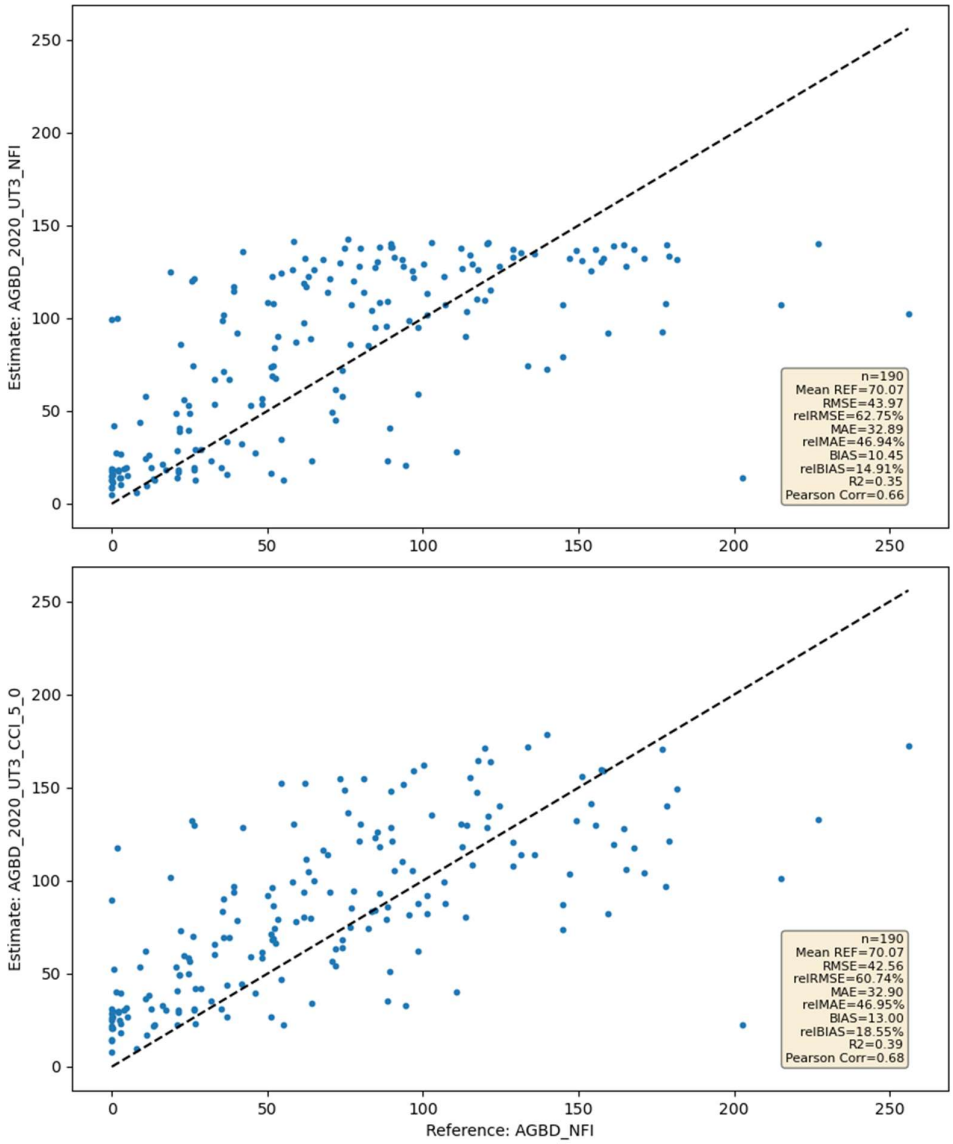


Figure 165. Validation result: AGB for the year 2020, above with calibration based on NFI temporary plots, and below with calibration based on CCI data, for the Sweden demonstration site, based on permanent NFI plots. Approach by UT3, parametric estimation with PALSAR-2.

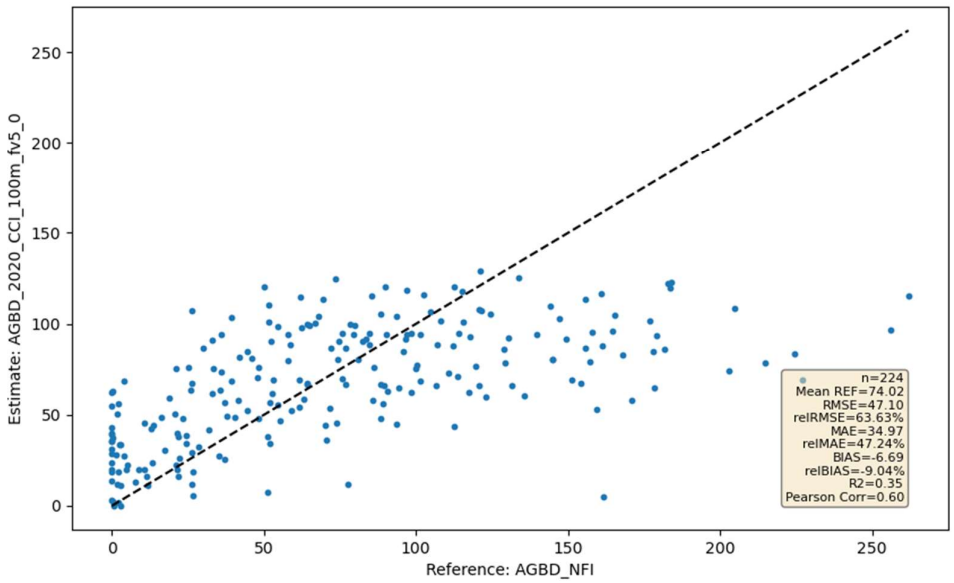


Figure 166. Validation result: AGB for the year 2020, CCI AGBD map 2020, with spatial resolution of 100m per pixel. The statistics are for information purposes only, as the observation units differ. The permanent NFI plots represent a diameter of 20m and the CCI map has a pixel resolution of 100m by 100m.

Table 60. Validation summary table for above ground woody biomass density for the boreal test site.

Approach for estimation of AGBD, boreal test site	Source	n	RMSE	BIAS	MAE	R2	r
Random Forest and linear calibration, year 2020	JR	224	31.5	2.6	23.1	0.71	0.84
Random Forest and linear calibration, year 2022	JR	227	40.0	0.2	29.6	0.52	0.73
Parametric Model, calibrated with NFI, year 2020	UT3	190	44.0	10.4	32.9	0.35	0.66
Parametric Model, calibrated with CCI5.0, year 2020	UT3	190	42.6	13.0	32.9	0.39	0.68
Parametric Models, calibrated	ESA, CCI	224	47.1	-6.7	35.0	0.35	0.60

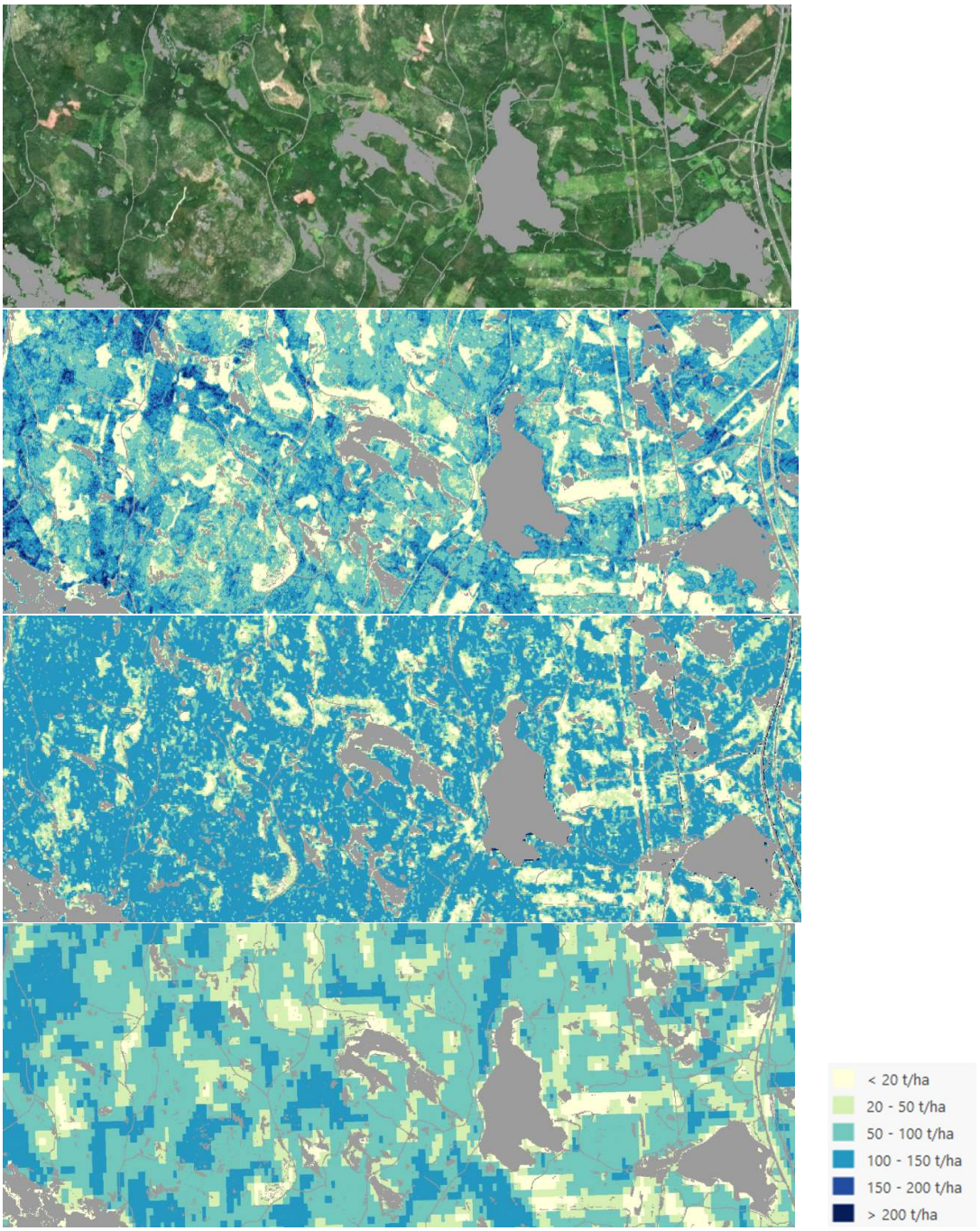


Figure 167. Qualitative comparison of AGB.

For qualitative comparison, different approaches are drawn for a subsite within the demonstration site for the year 2020. Displayed at Figure 167 above, the Sentinel-2 image mosaic in RGB is drawn, then below, the AGB estimates of the random forest and calibration approach developed by JR, then, the parametric estimation approach based on NFI calibration developed by UT3 and then the current version

5.0 of the CCI map. In gray are non-forest areas derived from the national land cover map of Sweden.

5.2.4.2 Forest Canopy Height

As above for above ground woody biomass, in the following, the forest canopy height is compared. First, scatterplots and quality metrics are provided to allow a quantitative comparison, then for qualitative comparison, the estimated forest canopy height is drawn for a subsite within the Sweden demonstration site in below figures.

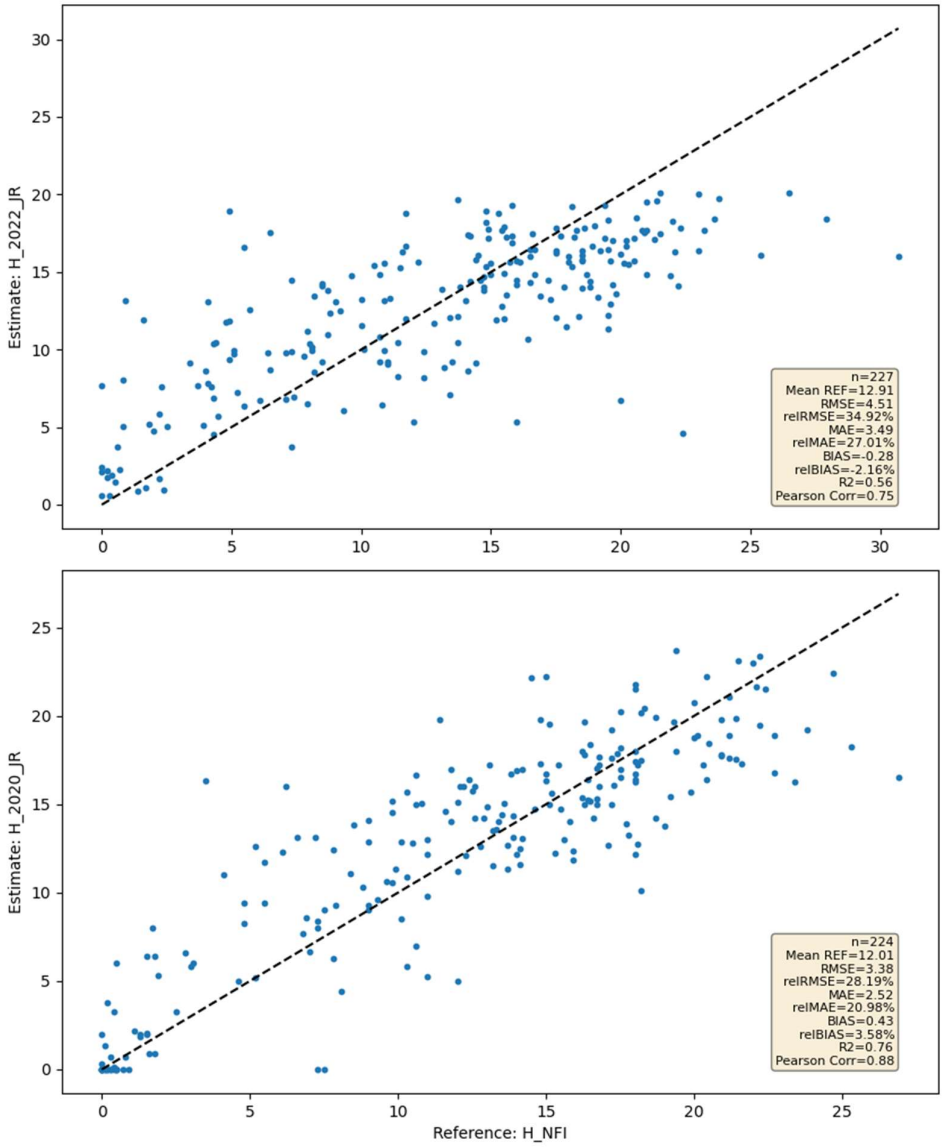


Figure 168. Validation result: FCH 2022 (above) and 2020 (below), for the Sweden site. Approach by JR with random forest estimation and calibration based on temporary NFI plots. The validation is based on the permanent NFI plots.

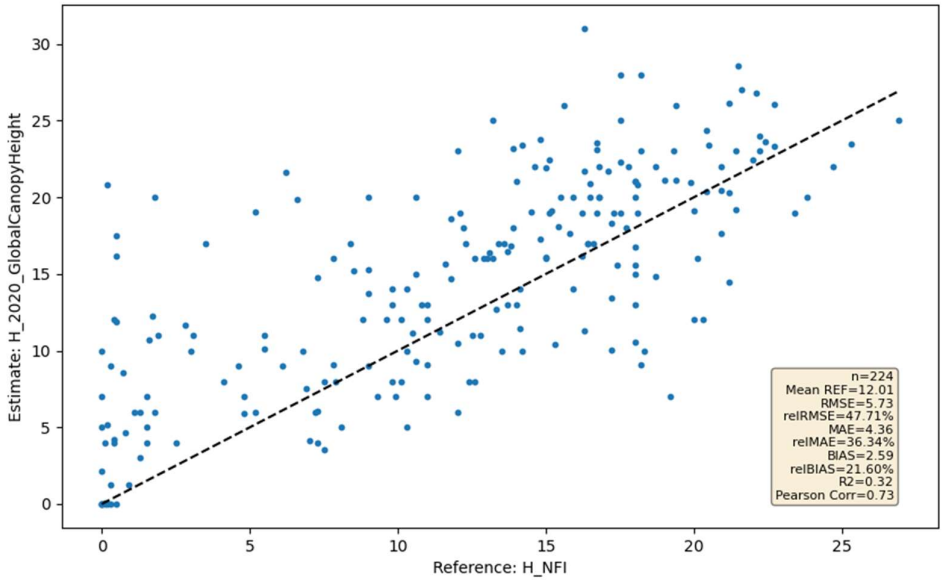


Figure 169. Validation result: Global Canopy Height 2020 dataset (Lang, Jetz, & Wegner, 2023), for the Sweden demonstration site, based on permanent NFI plots.

Table 61. Validation summary for forest canopy height for the boreal test site.

Approach for estimation of FCH, boreal test site	Source	n	RMSE	BIAS	MAE	R2	r
Random Forest and linear calibration, year 2020	JR	224	3.4	0.4	2.5	0.76	0.88
Random Forest and linear calibration, year 2022	JR	227	4.5	-0.3	3.5	0.56	0.75
Global Canopy Height, year 2022	(Lang, Jetz, & Wegner, 2023)	224	5.7	2.6	4.4	0.32	0.73

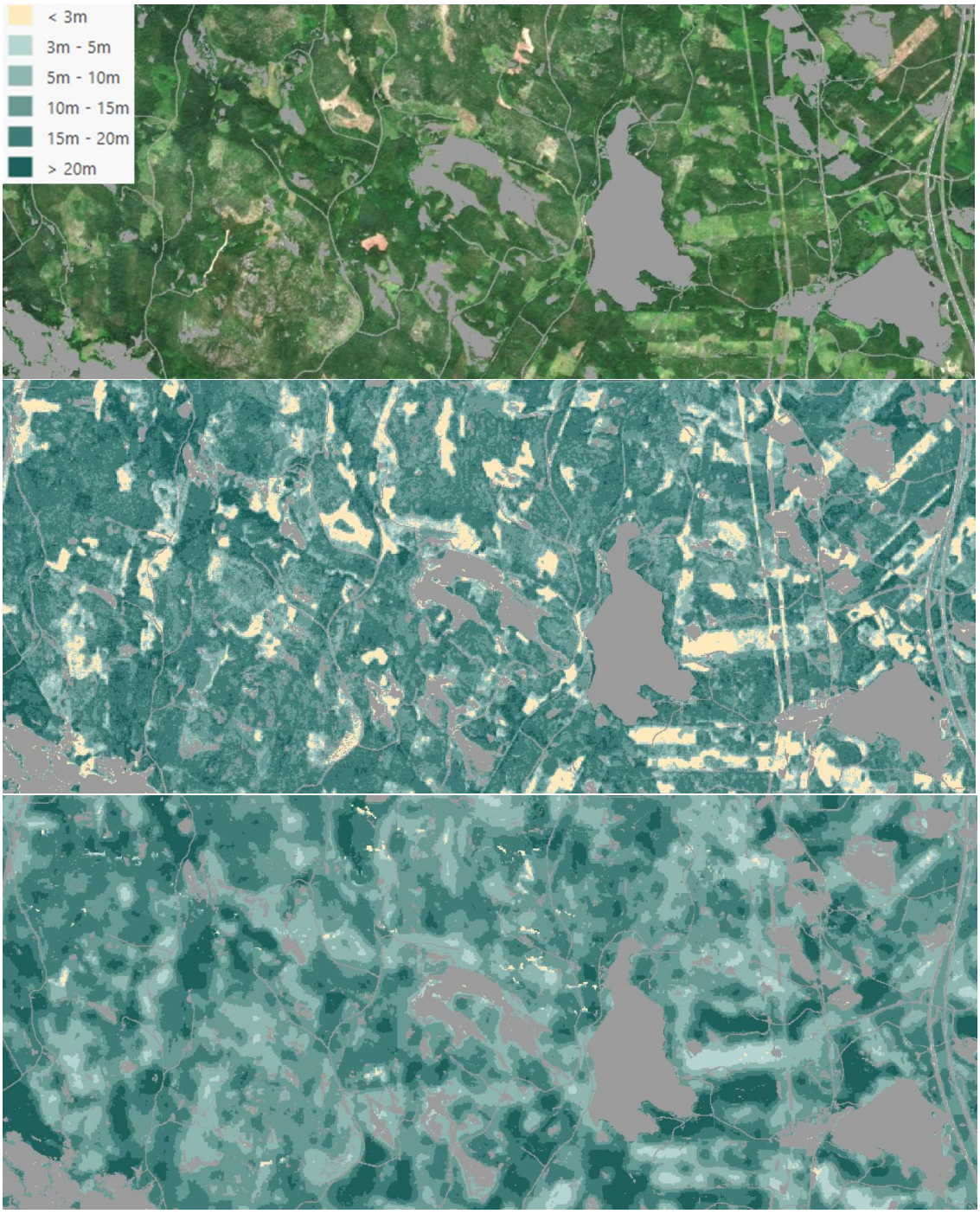


Figure 170. Qualitative comparison forest canopy height.

For qualitative comparison, different approaches are drawn for a subsite within the demonstration site for the year 2020, for the same extend as for above figure regarding AGBD. Figure 170 displays the Sentinel-2 image mosaic in RGB is drawn, then the FCH estimates of the random forest and calibration approach developed by JR, then the global forest canopy height map by (Lang, Jetz, & Wegner, 2023).

5.2.5 Summary and conclusions

The results achieved demonstrate the potential of mapping above ground woody biomass and forest canopy height based on satellite time series and sampling reference data. Compared to optical observations, data acquired in the microwave part of the spectrum contains additional information on forest structure because of the deeper penetration of microwaves into the canopy. New sensors to be launched soon, such as NISAR L-band SAR with planned launch date in 2025, or BIOMASS P-band SAR are expected to considerably contribute to enhanced mapping of woody biomass and forest canopy height. However, testing within EvoLand WP2 was not possible, because of the delayed launch dates for both satellites. However, integration of the recently available Sentinel-1 monthly mosaic data improved the quality and were therefore integrated in the mapping in the boreal and the tropical forest site.

Despite the nominal spatial resolution of 10 m, the effective spatial resolution at which individual forest patches can be identified is lower. However, the 10m resolution results are provided as expert products, as they are finally used as input for aggregating according to specific user requirements into larger assessment units with a minimum mapping unit of 0.25 ha or above. This allows generating final products according to different nomenclature specifications by including also further information into the aggregation process. For example, annual time-series maps may be aggregated according to the FAO nomenclature by applying multi-annual thresholds on canopy height with 5m, minimum mapping unit of 0.5 ha and tree crown cover from the CLMS service of 10%.

5.3 Grassland/cropland GPP biomass mapping

5.3.1 State of the art

5.3.1.1 Introduction

Information on the state and the evolution of agricultural productivity can be a relevant input for a wide variety of agricultural applications, including crop yield estimation, weather related impact on production and measuring carbon fluxes. Currently, a set of Copernicus CLMS which provides GPP estimates[22] at European level are already operational. However, these operational services use medium resolution satellites (e.g., Sentinel-3, Modis). Therefore, in the context of monitoring productivity dynamics at field level this has limitations. Furthermore, the parameters and algorithms being used are not specifically adjusted to capture the productivity dynamics for agricultural land like cropland and grassland. Within the EvoLand project, the aim is to address the current listed limitations. In the first place, the spatial detail will be improved by using the Sentinel-2 constellation for

[22] https://land.copernicus.eu/global/sites/cgls.vito.be/files/products/CGLOPS1_ATBD_DMP1km-V2_I2.11.pdf

productivity estimation. Secondly, also the thematic detail for productivity estimation under cropland and grassland will be improved, by parametrizing the GPP estimation model specifically for these two land cover types. Finally, existing limitations in current GPP CLMS products will be addressed like the incorporation drought stress and assess the C3/C4 temperature dependence for productivity. This results in the final goal to obtain a 10-daily crop/grassland specific GPP product at 10 m spatial detail across Europe, which is currently non-existing.

5.3.1.2 Review of existing methods

Different approaches exist to model vegetation productivity, ranging from inventory-based methods, empirical models, biogeochemical models, dynamic global vegetation models and remote sensing (Ito, 2011). The latter is the most of interest in this context, given the ability to provide productivity estimates at larger scale. Methods which use satellite information as input are most often based on the theory of Light Use Efficiency (LUE). This theory is based on the nearly constant relationship between incoming (and absorbed) solar radiation by the vegetation canopy and photosynthetic carbon uptake (McCallum et al., 2009). Incoming solar radiation is the main driver for this carbon uptake and downscaled by some efficiency factors that determines the actual uptake. These efficiency factors represent the actual conditions of the vegetation and its ability to absorb the incoming radiation and use it for carbon production. One of them, is a satellite based- indicator that expresses the Fraction of Absorbed Photosynthetic Active Radiation (FAPAR). Besides, that also some other environmental conditions like temperature, soil moisture content, soil type will impact the final productivity. The impact of these conditions is often determined based on models.

The current GPP CLMS are based on the Monteith approach which relies on the theory of LUE. Within this approach the GPP is determined based on a combination of (F)APAR and the actual efficiency of the plant to convert atmospheric CO2 to carbon. This efficiency has been subdivided into different parameters that can be used for calibration (see Table 62) for detailed explanation). In the current method, the effect of CO2 fertilization, temperature, and LUE on GPP production is considered. The LUE determines the potential biomass production based on the incoming solar radiation. However, as not every vegetation type is similarly efficient in using this radiation for biomass production, a biome specific parametrization was performed for the CLMS method. This potential LUE will be affected by environmental conditions that limits the actual biomass production like water stress, soil fertility, diseases, etc. The impact of these effects is accounted for in the residual parameter. However, in the CLMS approach the impact of these environmental conditions is omitted. In the equation the result is multiplied by 0.5 and 0.1 to convert Gross Dry Matter Productivity (GDMP) (kgDM/ha/day) into GPP (g C/m²/day).

$$GPP = R \,\&\, \epsilon_c \,\&\, FAPAR * \epsilon_{CO2} * \epsilon_T * \epsilon_{LUEc} * [\epsilon_{RES}] * 0.5 * 0.1$$

Table 62. Overview of different parameters used in the current Monteith approach of CLMS GPP[23].

Parameter	Definition	Value	Unit
GPP	Gross primary productivity	0-32	g C/m²/day
R	Total shortwave incoming radiation (0.2 – 3.0 µm	0-320	G J_T/ ha/day
ϵ_c	Fraction of PAR (0.4 – 0.7 µm) in total shortwave radiation	0.48	J_P/J_T
FAPAR	PAR-fraction absorbed (PA) by green vegetation	0-1	J_{AP}/J_p
ϵ_{CO2}	Normalized CO_2 fertilization effect	0-1	-
ϵ_T	Normalized temperature effect	0-1	-
ϵ_{LUEc}	LUE of dry matter (DM) at optimum	Land cover/biome specific	kgDM/GJ_{AP}
ϵ_{RES}	Residual stress (omitted)	1	-

These CLMS products has been validated using the GPP estimates from the ICOS flux towers[24]. Overall, the bias has been found small with reference data. However, the uncertainty is high, and precision compared to the reference data is low.

5.3.1.3 Strengths and weakness of existing methods

As mentioned above, the current method has some limitations especially as it does not take all environmental effect into account for GPP estimation, resulting in high uncertainty and low precision compared to reference data. Especially, the impact of drought stress is not included in the model, which results in overestimation of GPP during drought stress. Part of this stress will be reflected in the FAPAR signal, but the first impact of drought cannot be properly derived from solely FAPAR.

To account for drought stress, a drought stress factor can be added to the model equation. Based on literature following candidate drought stress indicators were identified:

- An indicator based on soil water content combined with an aridity index. As described by (Stocker et al., 2019).
- An indicator based on potential and actual evapotranspiration. As described by (Maselli et al., 2009).
- An indicator based on vapour pressure deficit

[23] https://land.copernicus.eu/global/sites/cgls.vito.be/files/products/CGLOPS1_ATBD_DMP-NPP300m-V1.1_I1.30.pdf
[24] https://land.copernicus.eu/global/sites/cgls.vito.be/files/products/CGLOPS1_QAR_DMP-NPP300m_V1.1_I1.50.pdf

5.3.2 An indicator based on vapour pressure deficit. Method description

5.3.2.1 Input data and study sites

Table 63 provides a list of datasets used for the method development. As ground truth data, ICOS towers located on crop and grassland were selected for the period between 2017-2023. This resulted in total in about 21 locations across Europe that could be used for training. This ICOS data contain GPP derived information at a high temporal scale (around 30 minutes), which can be directly used for calibrating the model parameters. Additional cleaning of the in-situ data is needed to remove less reliable measurements. The cleaned data and more information on this cleaning procedure can be found at IIASA's STAC[25] .

As satellite data, Sentinel-2 derived FAPAR is an essential input for GPP estimation to define the ability of the vegetation to absorb incoming photosynthetic active radiation. To determine, the amount of incoming radiation that could be used for photosynthesis, meteorological information from ECMWF were used. From this dataset, the daily incoming shortwave solar radiation flux could be retrieved. Furthermore, also the mean temperature provided in this dataset formed an input to the GPP estimation. All these input data was retrieved for a bounding box of 30x30 m centred around the flux sites to cover partly the sensitivity of the flux measurements according to the wind direction. Furthermore, all data was aggregated to 10-daily scale to align with the requirement to produce a 10-daily GPP product.

Table 63. List of currently used datasets for the method development.

Dataset	Description	Spatial Resolution	Temporal resolution
ICOS Fluxnet	GPP measurements from ICOS flux towers across Europe	/ (Point)	30 minutes
Sentinel-2 FAPAR	FAPAR 3-band derived from Sentinel-2 Level 2A	10 m	±5 days
AGERA5	Re-analysis of meteorological data from ECMWF	± 10 km	daily
ERA5-LAND	Re-analysis of meteorological data from ECMWF (this dataset has information on soil water content, evapotranspiration variables used in the drought stress factor assessment)	± 10 km	hourly

5.3.2.2 Description of the candidate method

The selected candidate method builds further upon the existing CLMS GPP method. This means that also the Monteith approach for GPP estimation will be used, but the high-resolution Sentinel-2 is used to define FAPAR at 10m spatial resolution. Furthermore, the different parameters in the equation will be adjusted to better reflect the GPP for crop and grassland. In the first place, the LUE (ϵ_{LUEc})

[25] https://stac.iiasa.ac.at/fluxdata_gpp.json?.language=en

parameter was separately re-calibrated for crop and grassland. In addition, the inclusion of a drought stress factor to the model is being evaluated. Three approaches (see 5.3.1.3), with for each approach several settings, are being evaluated. A semi-automatic workflow has been established that allows to add newly incoming ICOS data to the database and retrieve necessary FAPAR and meteorological data to enable model re-calibration. For the LUE re-calibration a non-linear least square fit method has been selected to find the most optimal LUE (ϵ_{LUEc}) factor for crop and grassland, separately. An input dataset containing the target GPP based on ICOS and modelled GPP by assuming ϵ_{LUEc} equal to 1, formed the input for the calibration. From this dataset 75% of the samples has been selected for training and 25% for validation. This approach was repeated for cropland and grassland data. The source code of this methodology can be consulted the EvoLand GitHub[26].

5.3.2.3 Benchmarking & evaluation

Validation of the re-calibrated ϵ_{LUEc} has been done by checking the deviation from 25% of the samples within the database. A few statistical parameters describing the accuracy of the fit were calculated, to define potential biases and issues. Furthermore, a comparison with the ϵ_{LUEc} factor from the CLMS counterpart has been done to define if the order of magnitude is in line with the expectations.

5.3.3 Implementation and results

Figure 171 highlights the accuracy of the GPP model after re-calibrating ϵ_{LUEc} specifically for crop and grassland. As can be noticed, it could be observed that overall, the outcome is quite promising with rather high R^2-values, especially for cropland. For grassland, there is somewhat more of the variation in the GPP that could not (yet) explained by the current model parameters. For cropland the ϵ_{LUEc} has been found at 3.18, while for grassland at 2.62. The generally accepted global constant for this parameter is defined at 2.54 (Haxeltine & Prentice, 1996). Hence, the one from grassland is rather close to this value. The larger value for cropland is not totally unexpected given that most of the crops are relatively efficient in turning absorbed photosynthetic active radiation into plant tissue. Similar research has also pointed out these differences according to the vegetation type (M. Wang et al., 2020). Furthermore, it should be noted that within the cropland class, large variations can exist in terms of LUE (dependent on crop type), which cannot be fully captured by this recalibrated LUE (Kukal & Irmak, 2020).

As could be noticed, still some unexplained variation remained around the 1:1 line. This means that still some environmental factors are not properly reflected in the current model setup. Based on literature and on the known limitations of the CLMS model, drought stress is not yet properly reflected into the model. Currently we are

[26] GitHub - Evoland-Land-Monitoring-Evolution/evoland-cropcarbon: Repository in which all the source code used for the Evoland GPP crop and grassland method development will be stored

still in the process of evaluating the possible drought factors (see 5.3.1.3). From the current results the indicator based on the soil water content and an aridity index gives the most promising results. For this indicator it is clear that the positive bias (indicating that the model without drought stress factor overestimates GPP) that is observed at locations where soil water content is low is reduced when this drought factor is added.

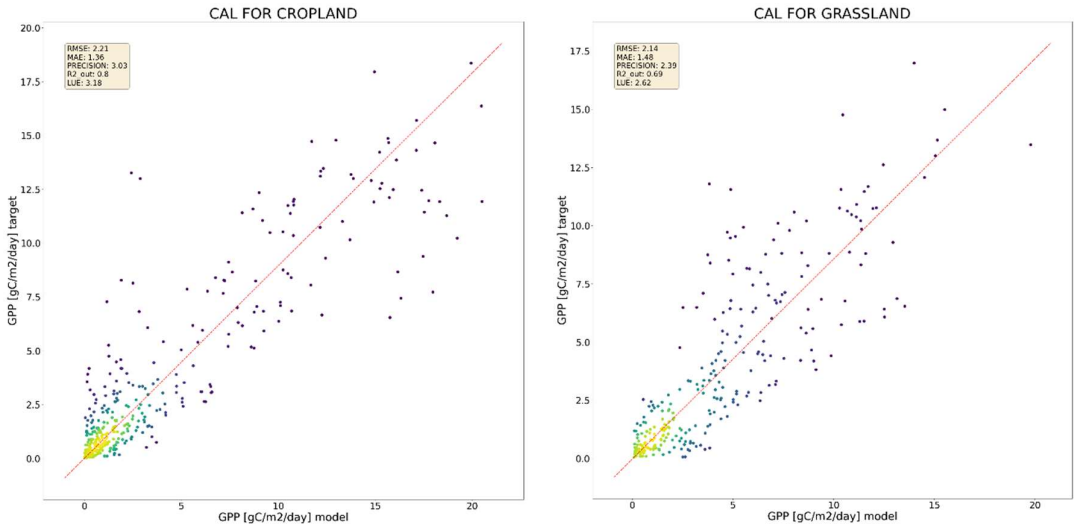


Figure 171. Left: GPP target compared to prediction for the optimized LUE for cropland. Right: GPP target compared to prediction for the optimized LUE for grassland.

Nevertheless, the current proposed procedure has still some important limitations which should be considered. In the first place, the size of the training dataset is rather limited, with only 21 flux towers spread across Europe that are located on a crop or grassland site. This means that the model parametrization will be based on the environmental settings of these specific towers and will likely not capture all possible variation across the European landscape. Furthermore, some local and isolated effects might also impact the actual GPP. For example, specific agricultural management like irrigation and fertilization will have an impact on the GPP production but also gradual impact like pest infestation. Most of these impacts could be indirectly observed on the FAPAR signal, but often there is some delay in the FAPAR response, while there is already some impact on the GPP. Besides, the difficulty to incorporate all the local effects also some limitations of using optical data for FAPAR estimation should be considered. In case of cloud persistent periods, the deviation from the actual FAPAR will grow and hence will affect the reliability of the GPP estimation. These limitations will be hardly to address, given the selected approach and used datasets.

5.3.4 Summary and conclusions

Within the EvoLand work the first developments are done to further improve the existing CLMS GPP model both in terms of thematic accuracy and spatial detail. This includes using higher resolution satellite data at 10 m spatial resolution and improving the thematic accuracy specific over cropland and grassland. As a first optimization, the LUE factor was re-calibrated for both land cover classes. This recalibration already improved the GPP estimation of crop and grassland clearly compared to using a global constant. Currently we are evaluating several drought stress indicators that can be included in the model with the aim to reduce the existing bias of the model observed during drought.

6 Conclusion and outlook

The Methods Compendium – Final Version provides a comprehensive overview of the methodology developed in Work Package 2 (WP2) of the EvoLand project, covering both development cycles. This document presents key advancements in several innovative approaches, including:

- Weakly Supervised Learning (Task 2.4) – Leveraging minimal labelled data to improve model generalization and efficiency.
- Improved Resolution (Task 2.5) – Enhancing spatial and temporal resolution to refine land monitoring applications.
- Continuous Monitoring (Task 2.6) – Developing methodologies for near-real-time land cover and vegetation change detection.
- Biomass Mapping (Task 2.7) – Advancing remote sensing techniques for accurate biomass estimation.

For each of these methodological developments, the compendium first presents a review of the state of the art, followed by a detailed description of the methods developed within EvoLand, their implementation status, and the corresponding results. These advancements are further substantiated through a series of referenced EvoLand journal articles, providing additional insights and validation.

The results from WP2 are being integrated and applied in Work Package 3 (WP3) to develop prototypes for Copernicus Land Monitoring Service applications. The final outcomes of the prototyping phase are expected to be delivered by Month 33 (M33)

References

Ahn, N., Kang, B., & Sohn, K. A. (2018). Fast, accurate, and lightweight super-resolution with cascading residual network. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11214 LNCS*. https://doi.org/10.1007/978-3-030-01249-6_16

Ait Ali Braham, N., Albrecht, C., Mairal, J., Chanussot, J., Wang, Y., & Zhu, X. (2024). *SpectralEarth: Training Hyperspectral Foundation Models at Scale*. https://doi.org/10.48550/arXiv.2408.08447

Amitrano, D., Guida, R., & Iervolino, P. (2021). Semantic Unsupervised Change Detection of Natural Land Cover With Multitemporal Object-Based Analysis on SAR Images. *IEEE Transactions on Geoscience and Remote Sensing*, *59*(7), 5494–5514. https://doi.org/10.1109/TGRS.2020.3029841

Anwar, S., Khan, S., & Barnes, N. (2019). *A Deep Journey into Super-resolution: A survey*. https://doi.org/10.48550/arXiv.1904.07523

Araza, A., de Bruin, S., Herold, M., Quegan, S., Labriere, N., Rodriguez-Veiga, P., Avitabile, V., Santoro, M., Mitchard, E. T. A., Ryan, C. M., Phillips, O. L., Willcock, S., Verbeeck, H., Carreiras, J., Hein, L., Schelhaas, M.-J., Pacheco-Pascagaza, A. M., da Conceição Bispo, P., Laurin, G. V., … Lucas, R. (2022). A comprehensive framework for assessing the accuracy and uncertainty of global above-ground biomass maps. *Remote Sensing of Environment*, *272*, 112917. https://doi.org/10.1016/j.rse.2022.112917

Attema, E. P. W., & Ulaby, F. T. (1978). Vegetation modeled as a water cloud. *Radio Science*, *13*(2), 357–364. https://doi.org/10.1029/RS013i002p00357

Aybar, C., Ysuhuaylas, L., Loja, J., Gonzales, K., Herrera, F., Bautista, L., Yali, R., Flores, A., Diaz, L., Cuenca, N., Espinoza, W., Prudencio, F., Llactayo, V., Montero, D., Sudmanns, M., Tiede, D., Mateo-García, G., & Gómez-Chova, L. (2022a). CloudSEN12, a global dataset for semantic understanding of cloud and cloud shadow in Sentinel-2. *Scientific Data*, *9*(1), 782. https://doi.org/10.1038/s41597-

Baevski, A., Hsu, W.-N., Xu, Q., Babu, A., Gu, J., & Auli, M. (2022a). *data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language*.

Bai, B., Tan, Y., Guo, D., & Xu, B. (2019). Dynamic Monitoring of Forest Land in Fuling District Based on Multi-Source Time Series Remote Sensing Images. *ISPRS International Journal of Geo-Information*, *8*(1), 36. https://doi.org/10.3390/ijgi8010036

Baier, G., Deschemps, A., Schmitt, M., & Yokoya, N. (2022). Synthesizing Optical and SAR Imagery From Land Cover Maps and Auxiliary Raster Data. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–12. https://doi.org/10.1109/TGRS.2021.3068532

Balntas, V., Riba, E., Ponsa, D., & Mikolajczyk, K. (2016). *Learning local feature descriptors with triplets and shallow convolutional neural networks*. https://doi.org/10.5244/C.30.119

Bao, H., Dong, L., Piao, S., & Wei, F. (2021). *BEiT: BERT Pre-Training of Image Transformers*.

Bardes, A., Ponce, J., & LeCun, Y. (2021). *VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning*.

Belgiu, M., & Stein, A. (2019). Spatiotemporal Image Fusion in Remote Sensing. *Remote Sensing*, *11*, 818. https://doi.org/10.3390/rs11070818

Bergamasco, L., Saha, S., Bovolo, F., & Bruzzone, L. (2022). Unsupervised Change Detection Using Convolutional-Autoencoder Multiresolution Features. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–19. https://doi.org/10.1109/TGRS.2022.3140404

Bernard, F., Manolis, I., Barat, I., Alamañac, A., Taboada, M., Mingorance, P., Ciapponi, A., Cardone, T., Nunez, I., Garcia, A., Hallibert, P., Hammar, A., Henriot, P., Codinachs, D., Patti, S., Montes, P., Skrzypek, P., Steenari, D., Weixler, S., & Holgueras, D. (2023). *The Copernicus land surface temperature monitoring (LSTM) mission: design, technology and status*. https://doi.org/10.1117/12.2679705

Bhangale, U., More, S., Shaikh, T., Patil, S., & More, N. (2020). Analysis of Surface Water Resources Using Sentinel-2 Imagery. *Procedia Computer Science*, *171*, 2645–2654. https://doi.org/10.1016/j.procs.2020.04.287

Bojinski, S., Verstraete, M., Peterson, T., Richter, C., Simmons, A., & Zemp, M. (2014). The Concept of Essential Climate Variables in Support of Climate Research, Applications, and Policy. *Bulletin of the American Meteorological Society*, *95*, 1431. https://doi.org/10.1175/BAMS-D-13-00047.1

Bordone Molini, A., Valsesia, D., Fracastoro, G., & Magli, E. (2019). DeepSUM: Deep Neural Network for Super-Resolution of Unregistered Multitemporal Images. *IEEE Transactions on Geoscience and Remote Sensing*, *PP*, 1–13. https://doi.org/10.1109/TGRS.2019.2959248

Boryan, C., Yang, Z., Mueller, R., & Craig, M. (2011). Monitoring US agriculture: the US Department of Agriculture, National Agricultural Statistics Service, Cropland Data Layer Program. *Geocarto International*, *26*(5), 341–358. https://doi.org/10.1080/10106049.2011.562309

Bouvet, A., Mermoz, S., Le Toan, T., Villard, L., Mathieu, R., Naidoo, L., & Asner, G. P. (2018). An above-ground biomass map of African savannahs and woodlands at 25 m resolution derived from ALOS PALSAR. *Remote Sensing of Environment*, *206*, 156–173. https://doi.org/10.1016/j.rse.2017.12.030

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140. https://doi.org/10.1007/BF00058655

Brooks, E. B., Wynne, R. H., Thomas, V. A., Blinn, C. E., & Coulston, J. W. (2014). On-the-Fly Massively Multitemporal Change Detection Using Statistical Quality Control Charts and Landsat Data. *IEEE Transactions on Geoscience and Remote Sensing*, *52*(6), 3316–3332. https://doi.org/10.1109/TGRS.2013.2272545

Brooks, E., Yang, Z., Thomas, V., & Wynne, R. (2017). Edyn: Dynamic Signaling of Changes to Forests Using Exponentially Weighted Moving Average Charts. *Forests*, *8*(9), 304. https://doi.org/10.3390/f8090304

Brown, C. F., Brumby, S. P., Guzder-Williams, B., Birch, T., Hyde, S. B., Mazzariello, J., Czerwinski, W., Pasquarella, V. J., Haertel, R., Ilyushchenko, S., Schwehr, K., Weisse, M., Stolle, F., Hanson, C., Guinan, O., Moore, R., & Tait, A. M. (2022). Dynamic World, Near real-time global 10 m land use land cover mapping. *Scientific Data*, *9*(1), 251. https://doi.org/10.1038/s41597-022-01307-4

Brown, M. S., Coops, N. C., Mulverhill, C., & Achim, A. (2025). Detection of non-stand replacing disturbances (NSR) using Harmonized Landsat-Sentinel-2 time series. *ISPRS Journal of Photogrammetry and Remote Sensing*, *220*, 264–276. https://doi.org/10.1016/j.isprsjprs.2024.12.014

Bullock, E. L., Woodcock, C. E., Souza, C., & Olofsson, P. (2020). Satellite-based estimates reveal widespread forest degradation in the Amazon. *Global Change Biology*, *26*(5), 2956–2969. https://doi.org/10.1111/gcb.15029

Büttner, G. (2014). *CORINE Land Cover and Land Cover Change Products* (pp. 55–74). https://doi.org/10.1007/978-94-007-7969-3_5

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., & Joulin, A. (2020). *Unsupervised Learning of Visual Features by Contrasting Cluster Assignments*.

Caye Daudt, R., Le Saux, B., Boulch, A., & Gousseau, Y. (2019). Multitask learning for large-scale semantic change detection. *Computer Vision and Image Understanding*, *187*, 102783. https://doi.org/10.1016/j.cviu.2019.07.003

Chen, G., Jiao, P., Hu, Q., Xiao, L., & Ye, Z. (2022). Swinstfm: Remote sensing spatiotemporal fusion using swin transformer. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–18.

Chen, G., Zhao, Y., Wang, Y., & Yap, K.-H. (2023). SSN: Stockwell Scattering Network for SAR Image Change Detection. *IEEE Geoscience and Remote Sensing Letters*, *20*, 1–5. https://doi.org/10.1109/LGRS.2023.3234972

Chen, H., He, X., Qing, L., Wu, Y., Ren, C., Sheriff, R. E., & Ce Zhu. (2022). Real-World Single Image Super-Resolution: a Brief Review. *Information Fusion*, *79*(nil), 124–145. https://doi.org/10.1016/j.inffus.2021.09.005

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). *A Simple Framework for Contrastive Learning of Visual Representations*.

Chen, X., Fan, H., Girshick, R., & He, K. (2020). *Improved Baselines with Momentum Contrastive Learning*.

Chen, X., & He, K. (2021a). Exploring Simple Siamese Representation Learning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15745–15753. https://doi.org/10.1109/CVPR46437.2021.01549

Chen, Y., & Bruzzone, L. (2022). A Self-Supervised Approach to Pixel-Level Change Detection in Bi-Temporal RS Images. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–11. https://doi.org/10.1109/TGRS.2022.3203897

Cheng, G., Huang, Y., Li, X., Lyu, S., Xu, Z., Zhao, Q., & Xiang, S. (2023a). *Change Detection Methods for Remote Sensing in the Last Decade: A Comprehensive Review*.

Chicco, D. (2021). *Siamese Neural Networks: An Overview* (pp. 73–94). https://doi.org/10.1007/978-1-0716-0826-5_3

Chouteau, F., Gabet, L., Fraisse, R., Bonfort, T., Harnoufi, B., Greiner, V., Le Goff, M., Ortner, M., & Paveau, V. (2022). Joint Super-Resolution and Image Restoration for PLÉIADES NEO Imagery. *The International Archives of Photogrammetry, Remote and Spatial Information Sciences*, *43*, 9–15.

Ciotola, M., Guarino, G., Vivone, G., Poggi, G., Chanussot, J., Plaza, A., & Scarpa, G. (2025). Hyperspectral pansharpening: Critical review, tools, and future perspectives. *IEEE Geoscience and Remote Sensing Magazine*, *nil*(nil), 2–29.

Claverie, M., Ju, J., Masek, J. G., Dungan, J. L., Vermote, E. F., Roger, J.-C., Skakun, S. V, & Justice, C. (2018). The harmonized landsat and Sentinel-2 surface reflectance data set. *Remote Sensing of Environment*, *219*, 145–161.

Cohen, W., Healey, S., Yang, Z., Stehman, S., Brewer, C., Brooks, E., Gorelick, N., Huang, C., Hughes, M., Kennedy, R., Loveland, T., Moisen, G., Schroeder, T., Vogelmann, J., Woodcock, C., Yang, L., & Zhu, Z. (2017). How Similar Are Forest Disturbance Maps Derived from Different Landsat Time Series Algorithms? *Forests*, *8*(4), 98. https://doi.org/10.3390/f8040098

Cornebise, J., Oršolić, I., & Kalaitzis, F. (2022). Open High-Resolution Satellite Imagery: The WorldStrat Dataset With Application to Super-Resolution. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in Neural Information Processing Systems* (Vol. 35, pp. 25979–25991). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2022/file/a6fe99561d9eb9c90b322afe664587fd-Paper-Datasets_and_Benchmarks.pdf

Crist, E. P. (1985). A TM Tasseled Cap equivalent transformation for reflectance factor data. *Remote Sensing of Environment*, *17*(3), 301–306. https://doi.org/10.1016/0034-4257(85)90102-6

Csillik, O., Reiche, J., De Sy, V., Araza, A., & Herold, M. (2022). Rapid remote monitoring reveals spatial and temporal hotspots of carbon loss in Africa's rainforests.

Communications Earth & Environment, *3*(1), 48. https://doi.org/10.1038/s43247-022-00383-z

Daudt, R. C., Le Saux, B., Boulch, A., & Gousseau, Y. (2018). Urban Change Detection for Multispectral Earth Observation Using Convolutional Neural Networks. *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, 2115–2118. https://doi.org/10.1109/IGARSS.2018.8518015

Decuyper, M., Chávez, R. O., Lohbeck, M., Lastra, J. A., Tsendbazar, N., Hackländer, J., Herold, M., & Vågen, T.-G. (2022). Continuous monitoring of forest change dynamics with satellite time series. *Remote Sensing of Environment*, *269*, 112829. https://doi.org/10.1016/j.rse.2021.112829

Deng, J., Dong, W., Socher, R., Li, L.-J., Kai Li, & Li Fei-Fei. (2009). ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. https://doi.org/10.1109/CVPR.2009.5206848

DeVries, B., Decuyper, M., Verbesselt, J., Zeileis, A., Herold, M., & Joseph, S. (2015). Tracking disturbance-regrowth dynamics in tropical forests using structural change detection and Landsat time series. *Remote Sensing of Environment*, *169*, 320–334. https://doi.org/10.1016/j.rse.2015.08.020

DeVries, B., Pratihast, A. K., Verbesselt, J., Kooistra, L., & Herold, M. (2016). Characterizing Forest Change Using Community-Based Monitoring Data and Landsat Time Series. *PLOS ONE*, *11*(3), e0147121. https://doi.org/10.1371/journal.pone.0147121

DeVries, B., Verbesselt, J., Kooistra, L., & Herold, M. (2015). Robust monitoring of small-scale forest disturbances in a tropical montane forest using Landsat time series. *Remote Sensing of Environment*, *161*, 107–121. https://doi.org/10.1016/j.rse.2015.02.012

Dick A. & Moreau, A. (2022). VENμS: Mission Characteristics, Final Evaluation of the First Phase and Data Production. *Submitted to Remote Sensing*.

Dike, H. U., Zhou, Y., Deveerasetty, K. K., & Wu, Q. (2018). Unsupervised Learning Based On Artificial Neural Network: A Review. *2018 IEEE International Conference on Cyborg and Bionic Systems (CBS)*, 322–327. https://doi.org/10.1109/CBS.2018.8612259

Doersch, C., Gupta, A., & Efros, A. A. (2015). Unsupervised Visual Representation Learning by Context Prediction. *2015 IEEE International Conference on Computer Vision (ICCV)*, 1422–1430. https://doi.org/10.1109/ICCV.2015.167

Dong, C., Loy, C. C., He, K., & Tang, X. (2015). Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine*, *38*(2), 295–307.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N.

(2020). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*.

Dubayah, R., Armston, J., Healey, S. P., Bruening, J. M., Patterson, P. L., Kellner, J. R., Duncanson, L., Saarela, S., Ståhl, G., Yang, Z., Tang, H., Blair, J. B., Fatoyinbo, L., Goetz, S., Hancock, S., Hansen, M., Hofton, M., Hurtt, G., & Luthcke, S. (2022). GEDI launches a new era of biomass inference from space. *Environmental Research Letters*, *17*(9), 095001. https://doi.org/10.1088/1748-9326/ac8694

Dubayah, R., Blair, J. B., Goetz, S., Fatoyinbo, L., Hansen, M., Healey, S., Hofton, M., Hurtt, G., Kellner, J., Luthcke, S., Armston, J., Tang, H., Duncanson, L., Hancock, S., Jantz, P., Marselis, S., Patterson, P. L., Qi, W., & Silva, C. (2020). The Global Ecosystem Dynamics Investigation: High-resolution laser ranging of the Earth's forests and topography. *Science of Remote Sensing*, *1*, 100002. https://doi.org/10.1016/j.srs.2020.100002

Dumeur, I., Valero, S., & Inglada, J. (2024a). Self-Supervised Spatio-Temporal Representation Learning of Satellite Image Time Series. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *17*, 4350–4367. https://doi.org/10.1109/JSTARS.2024.3358066

Duncanson, L., Kellner, J. R., Armston, J., Dubayah, R., Minor, D. M., Hancock, S., Healey, S. P., Patterson, P. L., Saarela, S., Marselis, S., Silva, C. E., Bruening, J., Goetz, S. J., Tang, H., Hofton, M., Blair, B., Luthcke, S., Fatoyinbo, L., Abernethy, K., … Zgraggen, C. (2022). Aboveground biomass density models for NASA's Global Ecosystem Dynamics Investigation (GEDI) lidar mission. *Remote Sensing of Environment*, *270*, 112845. https://doi.org/10.1016/j.rse.2021.112845

Eckardt, A., Horack, J., Lehmann, F., Krutz, D., Drescher, J., Whorton, M., & Soutullo, M. (2015). DESIS (DLR Earth Sensing Imaging Spectrometer for the ISS-MUSES platform). *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 1457–1459. https://doi.org/10.1109/IGARSS.2015.7326053

Ehlers, D., Wang, C., Coulston, J., Zhang, Y., Pavelsky, T., Frankenberg, E., Woodcock, C., & Song, C. (2022). Mapping Forest Aboveground Biomass Using Multisource Remotely Sensed Data. *Remote Sensing*, *14*(5), 1115. https://doi.org/10.3390/rs14051115

Fang, S., Li, K., Shao, J., & Li, Z. (2022). SNUNet-CD: A Densely Connected Siamese Network for Change Detection of VHR Images. *IEEE Geoscience and Remote Sensing Letters*, *19*, 1–5. https://doi.org/10.1109/LGRS.2021.3056416

Firozjaei, M. K., Kiavarz, M., & Alavipanah, S. K. (2022). Satellite-derived land surface temperature spatial sharpening: a comprehensive review on current status and perspectives. *European Journal of Remote Sensing*, *55*(1), 644–664.

Flood, N. (2013). Seasonal Composite Landsat TM/ETM+ Images Using the Medoid (a Multi-Dimensional Median). *Remote Sensing*, *5*(12), 6481–6500. https://doi.org/10.3390/rs5126481

Frison, P.-L., Fruneau, B., Kmiha, S., Soudani, K., Dufrêne, E., Toan, T. Le, Koleck, T., Villard, L., Mougin, E., & Rudant, J.-P. (2018). Potential of Sentinel-1 Data for Monitoring Temperate Mixed Forest Phenology. *Remote Sensing*, *10*(12), 2049. https://doi.org/10.3390/rs10122049

Fu, L., Lian, L., Wang, R., Shi, B., Wang, X., Yala, A., Darrell, T., Efros, A. A., & Goldberg, K. (2024). *Rethinking Patch Dependence for Masked Autoencoders*.

Fuchs, M. H. P., & Demir, B. (2023). *HySpecNet-11k: A Large-Scale Hyperspectral Dataset for Benchmarking Learning-Based Hyperspectral Image Compression Methods*.

Galar, M., Sesma, R., Ayala, C., Albizua, L., & Aranda, C. (2020a). Super-resolution of Sentinel-2 images using convolutional neural networks and real ground truth data. *Remote Sensing*, *12*(18). https://doi.org/10.3390/RS12182941

Galar, M., Sesma, R., & Ayala Christian and Aranda, C. (2019). SUPER-RESOLUTION FOR SENTINEL-2 IMAGES. *International Archives of the Photogrammetry, Remote & Spatial Information Sciences*.

Galiatsatos, N., Donoghue, D. N. M., Watt, P., Bholanath, P., Pickering, J., Hansen, M. C., & Mahmood, A. R. J. (2020). An Assessment of Global Forest Change Datasets for National Forest Monitoring and Reporting. *Remote Sensing*, *12*(11), 1790. https://doi.org/10.3390/rs12111790

Gao, F., Kustas, W. P., & Anderson, M. C. (2012). A data mining approach for sharpening thermal satellite imagery over land. *Remote Sensing*, *4*(11), 3287–3319.

Gao, F., Masek, J., Schwaller, M., & Hall, F. (2006). On the blending of the landsat and modis surface reflectance: Predicting daily landsat surface reflectance. *IEEE Transactions on Geoscience and Remote Sensing*, *44*(8), 2207–2218.

Gao, P., Ma, T., Li, H., Lin, Z., Dai, J., & Qiao, Y. (2022). *ConvMAE: Masked Convolution Meets Masked Autoencoders*.

Garnot, V., & Landrieu, L. (2021). *Panoptic Segmentation of Satellite Image Time Series with Convolutional Temporal Attention Networks*. https://doi.org/10.1109/ICCV48922.2021.00483

Garnot, V., Landrieu, L., & Chehata, N. (2022). Multi-modal temporal attention models for crop mapping from satellite time series. *ISPRS Journal of Photogrammetry and Remote Sensing*, *187*, 294–305. https://doi.org/10.1016/j.isprsjprs.2022.03.012

Ge, C., Ding, H., Molina, I., He, Y., & Peng, D. (2022). Object-Oriented Change Detection Method Based on Spectral–Spatial–Saliency Change Information and Fuzzy Integral Decision Fusion for HR Remote Sensing Images. *Remote Sensing*, *14*(14), 3297. https://doi.org/10.3390/rs14143297

Giambelluca, A. L., Hermosilla, T., Álvarez-Mozos, J., & González-Audícana, M. (2025). Identifying forest harvesting practices: Clear-cutting and thinning in diverse tree species using dense Landsat time series. *Forest Ecology and Management*, *578*, 122442. https://doi.org/10.1016/j.foreco.2024.122442

Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K. V., Joulin, A., & Misra, I. (2023). ImageBind One Embedding Space to Bind Them All. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15180–15190. https://doi.org/10.1109/CVPR52729.2023.01457

Girshick, R. (2015a). Fast R-CNN. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Gonzalez, R., Albrecht, C., Ait Ali Braham, N., Lambhate, D., Almeida, J., Fraccaro, P., Blumenstiel, B., Brunschwiler, T., & Bangalore, R. (2025). *Multispectral to Hyperspectral using Pretrained Foundational model*. https://doi.org/10.48550/arXiv.2502.19451

Goyena, H., Pérez-Goya, U., Montesino-SanMartin, M., Militino, A. F., Wang, Q., Atkinson, P. M., & Ugarte, M. D. (2023). Unpaired spatio-temporal fusion of image patches (ustfip) from cloud covered images. *Remote Sensing of Environment*, *295*(nil), 113709.

Granero-Belinchon, C., Michel, A., Lagouarde, J.-P., Sobrino, J. A., & Briottet, X. (2019). Multi-resolution study of thermal unmixing techniques over madrid urban area: Case study of trishna mission. *Remote Sensing*, *11*(10).

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. A., Guo, Z. D., Azar, M. G., Piot, B., Kavukcuoglu, K., Munos, R., & Valko, M. (2020a). *Bootstrap your own latent: A new approach to self-supervised Learning*.

Guo, X., Lao, J., Dang, B., Zhang, Y., Yu, L., Ru, L., Zhong, L., Huang, Z., Wu, K., Hu, D., He, H., Wang, J., Chen, J., Yang, M., Zhang, Y., & Li, Y. (2024). Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 27672–27683.

Gupta, S., Girshick, R., Arbeláez, P., & Malik, J. (2014). *Learning Rich Features from RGB-D Images for Object Detection and Segmentation* (pp. 345–360). https://doi.org/10.1007/978-3-319-10584-0_23

Hamunyela, E., Verbesselt, J., & Herold, M. (2016). Using spatial context to improve early detection of deforestation from Landsat time series. *Remote Sensing of Environment*, *172*, 126–138. https://doi.org/10.1016/j.rse.2015.11.006

Hansen, M. C., Krylov, A., Tyukavina, A., Potapov, P. V, Turubanova, S., Zutta, B., Ifo, S., Margono, B., Stolle, F., & Moore, R. (2016). Humid tropical forest disturbance alerts using Landsat data. *Environmental Research Letters*, *11*(3), 034008. https://doi.org/10.1088/1748-9326/11/3/034008

Hansen, M. C., Potapov, P. V., Goetz, S. J., Turubanova, S., Tyukavina, A., Krylov, A., Kommareddy, A., & Egorov, A. (2016). Mapping tree height distributions in Sub-Saharan Africa using Landsat 7 and 8 data. *Remote Sensing of Environment*, *185*, 221–232. https://doi.org/10.1016/j.rse.2016.02.023

Haxeltine, A., & Prentice, I. C. (1996). A General Model for the Light-Use Efficiency of Primary Production. *Functional Ecology*, *10*(5), 551. https://doi.org/10.2307/2390165

He, K., Chen, X., Xie, S., Li, Y., Dollar, P., & Girshick, R. (2022). Masked Autoencoders Are Scalable Vision Learners. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15979–15988. https://doi.org/10.1109/CVPR52688.2022.01553

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum Contrast for Unsupervised Visual Representation Learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9726–9735. https://doi.org/10.1109/CVPR42600.2020.00975

He, K., Zhang, X., Ren, S., & Sun, J. (2016a). Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. https://doi.org/10.1109/CVPR.2016.90

Helber, P., Bischke, B., Dengel, A., & Borth, D. (2018). Introducing Eurosat: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification. *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, 204–207. https://doi.org/10.1109/IGARSS.2018.8519248

Huang, C., Goward, S. N., Masek, J. G., Thomas, N., Zhu, Z., & Vogelmann, J. E. (2010). An automated approach for reconstructing recent forest disturbance history using dense Landsat time series stacks. *Remote Sensing of Environment*, *114*(1), 183–198. https://doi.org/10.1016/j.rse.2009.08.017

Ibrahim, M., Benavente, R., Ponsa, D., & Lumbreras, F. (2025). HyDA-Net: A Hybrid Dense Attention Network for Remote Sensing Multi-Image Super-Resolution. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *PP*, 1–24. https://doi.org/10.1109/JSTARS.2025.3547785

Inglada, J., Michel, J., & Hagolle, O. (2022). Assessment of the Usefulness of Spectral Bands for the Next Generation of Sentinel-2 Satellites by Reconstruction of Missing Bands. *Remote Sensing*, *14*(10). https://doi.org/10.3390/rs14102503

Irigireddy, B., & Bandaru, V. (2025). *Satflow: Generative model based framework for producing high resolution gap free remote sensing imagery*.

Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, *2017-January*. https://doi.org/10.1109/CVPR.2017.632

Ito, A. (2011). A historical meta-analysis of global terrestrial net primary productivity: are estimates converging? *Global Change Biology*, *17*(10), 3161–3175. https://doi.org/10.1111/j.1365-2486.2011.02450.x

Jain, U., Wilson, A., & Gulshan, V. (2022). *Multimodal contrastive learning for remote sensing tasks*.

Jakubik, J., Roy, S., Phillips, C. E., Fraccaro, P., Godwin, D., Zadrozny, B., Szwarcman, D., Gomes, C., Nyirjesy, G., Edwards, B., Kimura, D., Simumba, N., Chu, L., Mukkavilli, S. K., Lambhate, D., Das, K., Bangalore, R., Oliveira, D., Muszynski, M., … Ramachandran, R. (2023). *Foundation Models for Generalist Geospatial Artificial Intelligence*.

Javed, A., Jung, S., Lee, W. H., & Han, Y. (2020). Object-Based Building Change Detection by Fusing Pixel-Level Change Detection Results Generated from Morphological Building Index. *Remote Sensing*, *12*(18), 2952. https://doi.org/10.3390/rs12182952

Jenni, S., & Favaro, P. (2018). Self-Supervised Feature Learning by Learning to Spot Artifacts. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2733–2742. https://doi.org/10.1109/CVPR.2018.00289

Jetz, W., McGeoch, M., Guralnick, R., Ferrier, S., Beck, J., Costello, M., Fernández, M., Geller, G., Keil, P., Merow, C., Meyer, C., Muller-Karger, F., Regan, E., Schmeller, D., & Turak, E. (2019). Essential biodiversity variables for mapping and monitoring species populations. *Nature Ecology & Evolution*, *3*. https://doi.org/10.1038/s41559-019-0826-1

Jia, M., Zhang, C., Lv, Z., Zhao, Z., & Wang, L. (2022). Bipartite Adversarial Autoencoders With Structural Self-Similarity for Unsupervised Heterogeneous Remote Sensing Image Change Detection. *IEEE Geoscience and Remote Sensing Letters*, *19*, 1–5. https://doi.org/10.1109/LGRS.2022.3201925

Jiang, X., Li, G., Zhang, X.-P., & He, Y. (2022). A Semisupervised Siamese Network for Efficient Change Detection in Heterogeneous Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–18. https://doi.org/10.1109/TGRS.2021.3061686

Jiang, X., Xiang, S., Wang, M., & Tang, P. (2022). Dual-Pathway Change Detection Network Based on the Adaptive Fusion Module. *IEEE Geoscience and Remote Sensing Letters*, *19*, 1–5. https://doi.org/10.1109/LGRS.2021.3103991

Johnson, J., Alahi, A., & Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9906 LNCS*. https://doi.org/10.1007/978-3-319-46475-6_43

Kalman, R. E. (1960). A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, *82*(1), 35–45. https://doi.org/10.1115/1.3662552

Kapilaratne, R. G. C. J., Kakuta, S., & Kaneta, S. (2022). ENHANCED SUPER RESOLUTION FOR REMOTE SENSING IMAGERIES. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *V-3–2022*, 53–60. https://doi.org/10.5194/isprs-annals-V-3-2022-53-2022

Kastryulin, S., Zakirov, J., Prokopenko, D., & Dylov, D. V. (2022). *PyTorch Image Quality: Metrics for Image Quality Assessment*. arXiv. https://doi.org/10.48550/ARXIV.2208.14818

Kemker, R., & Kanan, C. (2017). Self-Taught Feature Learning for Hyperspectral Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, *55*(5), 2693–2705. https://doi.org/10.1109/TGRS.2017.2651639

Kennedy, R. E., Yang, Z., & Cohen, W. B. (2010a). Detecting trends in forest disturbance and recovery using yearly Landsat time series: 1. LandTrendr — Temporal segmentation algorithms. *Remote Sensing of Environment*, *114*(12), 2897–2910. https://doi.org/10.1016/j.rse.2010.07.008

Kennedy, R. E., Yang, Z., & Cohen, W. B. (2010b). Detecting trends in forest disturbance and recovery using yearly Landsat time series: 1. LandTrendr — Temporal segmentation algorithms. *Remote Sensing of Environment*, *114*(12), 2897–2910. https://doi.org/10.1016/j.rse.2010.07.008

Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2022). Transformers in Vision: A Survey. *ACM Computing Surveys*, *54*(10s), 1–41. https://doi.org/10.1145/3505244

Kim, D., Cho, D., & Kweon, I. S. (2019). Self-Supervised Video Representation Learning with Space-Time Cubic Puzzles. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 8545–8552. https://doi.org/10.1609/aaai.v33i01.33018545

Kim, J., Kwon Lee, J., & Mu Lee, K. (2016). Accurate image super-resolution using very deep convolutional networks. *Proceedings of the IEEE Conference on Computer and Pattern Recognition*, 1646–1654.

Kingma, D. P., & Ba, J. (2014a). Adam: a Method for Stochastic Optimization. *Nil*, *nil*(nil), nil. https://doi.org/10.48550/ARXIV.1412.6980

Kukal, M. S., & Irmak, S. (2020). Interrelationships between water use efficiency and light use efficiency in four row crop canopies. *Agrosystems, Geosciences & Environment*, *3*(1). https://doi.org/10.1002/agg2.20110

Lac, A., Michel, J., Poulain, V., & Sfaksi, N. (2023). *Sentinel-2 Single Image Super-Resolution with the SEN2VENμS Dataset: architecture, training strategy, performances assessment and application to Water Bodies Detection*. https://hal.science/hal-04218629

Lagouarde, J.-P., Bhattacharya, B. K., Crébassol, P., Gamet, P., Babu, S. S., Boulet, G., Briottet, X., Buddhiraju, K. M., Cherchali, S., Dadou, I., Dedieu, G., Gouhier, M., Hagolle, O., Irvine, M., Jacob, F., Kumar, A., Kumar, K. K., Laignel, B., Mallick, K., …

Ramakrishnan, R. (2018). The Indian-French Trishna Mission: Earth Observation in the Thermal Infrared with High Spatio-Temporal Resolution. *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, 4078–4081. https://doi.org/10.1109/IGARSS.2018.8518720

Lahssini, K., Baghdadi, N., Maire, G., & Fayad, I. (2022). Influence of GEDI Acquisition and Processing Parameters on Canopy Height Estimates over Tropical Forests. *Remote Sensing*, *14*, 6264. https://doi.org/10.3390/rs14246264

Lang, N., Jetz, W., Schindler, K., & Wegner, J. D. (2023). A high-resolution canopy height model of the Earth. *Nature Ecology & Evolution*, *7*(11), 1778–1789. https://doi.org/10.1038/s41559-023-02206-6

Latte, N., & Lejeune, P. (2020). PlanetScope Radiometric Normalization and Sentinel-2 Super-Resolution (2.5 m): A Straightforward Spectral-Spatial Fusion of Multi-Satellite Multi-Sensor Images Using Residual Convolutional Neural Networks. *Remote Sensing*, *12*. https://doi.org/10.3390/rs12152366

Le Toan, T., Quegan, S., Woodward, I., Lomas, M., Delbart, N., & Picard, G. (2004). Relating Radar Remote Sensing of Biomass to Modelling of Forest Carbon Budgets. *Climatic Change*, *67*(2–3), 379–402. https://doi.org/10.1007/s10584-004-3155-5

Lechler, S., Picoli, M. C. A., Soares, A. R., Sanchez, A., Chaves, M. E. D., & Verstegen, J. (2020). EXPLORING NASA'S HARMONIZED LANDSAT AND SENTINEL-2 (HLS) DATASET TO MONITOR DEFORESTATION IN THE AMAZON RAINFOREST. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XLIII-B3-2020*, 705–711. https://doi.org/10.5194/isprs-archives-XLIII-B3-2020-705-2020

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., & Shi, W. (2017). Photo-realistic single image super-resolution using a generative adversarial network. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, *2017-January*. https://doi.org/10.1109/CVPR.2017.19

Lee, B., Kim, N., Kim, E.-S., Jang, K., Kang, M., Lim, J.-H., Cho, J., & Lee, Y. (2020). An Artificial Intelligence Approach to Predict Gross Primary Productivity in the Forests of South Korea Using Satellite Remote Sensing Data. *Forests*, *11*(9), 1000. https://doi.org/10.3390/f11091000

Lee, J.-S. (1986). Speckle Suppression And Analysis For Synthetic Aperture Radar Images. *Optical Engineering*, *25*(5). https://doi.org/10.1117/12.7973877

Levin, A., Weiss, Y., Durand, F., & Freeman, W. T. (2011). Understanding blind deconvolution algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(12), 2354–2367.

Li, H., Yang, Y., Chang, M., Feng, H., Xu, Z., Li, Q., & Chen, Y. (2021). *SRDiff: Single Image Super-Resolution with Diffusion Probabilistic Models*.

Li, H., Zhang, G., Zhong, Q., Xing, L., & Du, H. (2023). Prediction of Urban Forest Aboveground Carbon Using Machine Learning Based on Landsat 8 and Sentinel-2: A Case Study of Shanghai, China. *Remote Sensing*, *15*(1), 284. https://doi.org/10.3390/rs15010284

Li, J., & Chen, B. (2020). Global revisit interval analysis of landsat-8-9 and sentinel-2a-2b data for terrestrial monitoring. *Sensors (Switzerland)*, *20*(22). https://doi.org/10.3390/s20226631

Li, S., Xu, L., Jing, Y., Yin, H., Li, X., & Guan, X. (2021). High-quality vegetation index product generation: a review of ndvi time series reconstruction techniques. *International Journal of Applied Earth Observation and Geoinformation*, *105*(nil), 102640.

Li, W., Cao, D., & Xiang, M. (2022). Enhanced multi-stream remote sensing spatiotemporal fusion network based on transformer and dilated convolution. *Remote Sensing*, *14*(18), 4544.

Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., & Timofte, R. (2021). SwinIR: Image Restoration Using Swin Transformer. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 1833–1844.

Liang, M., Duncanson, L., Silva, J. A., & Sedano, F. (2023). Quantifying aboveground biomass dynamics from charcoal degradation in Mozambique using GEDI Lidar and Landsat. *Remote Sensing of Environment*, *284*, 113367. https://doi.org/10.1016/j.rse.2022.113367

Lim, B., Son, S., Kim, H., Nah, S., & Lee, K. M. (2017). Enhanced Deep Residual Networks for Single Image Super-Resolution. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, *2017-July*. https://doi.org/10.1109/CVPRW.2017.151

Liu, A., Liu, Y., Gu, J., Qiao, Y., & Dong, C. (2023). Blind Image Super-Resolution: A Survey and Beyond . *IEEE Transactions on Pattern Analysis & Machine Intelligence* , *45*(05), 5461–5480. https://doi.org/10.1109/TPAMI.2022.3203009

Liu, C., Albrecht, C. M., Wang, Y., & Zhu, X. X. (2025). CromSS: Cross-modal pretraining with noisy labels for remote sensing image segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 1. https://doi.org/10.1109/TGRS.2025.3552642

Liu, H., Gan, J., Fan, X., Zhang, Y., Luo, C., Zhang, J., Jiang, G., Qian, Y., Zhao, C., Ma, H., & Guo, Z. (2023). Pt-tuning: Bridging the gap between time series masked reconstruction and forecasting via prompt token tuning. *Nil*, *nil*(nil).

Liu, H., Qian, Y., Zhong, X., Chen, L., & Yang, G. (2021). Research on super-resolution reconstruction of remote sensing images: a comprehensive review. *Optical Engineering*, *60*(10), 100901.

Liu, H., Zhang, H. K., Huang, B., Yan, L., Tran, K. K., Qiu, Y., Zhang, X., & Roy, D. P. (2024). Reconstruction of seamless harmonized landsat sentinel-2 (hls) time series via self-supervised learning. *Remote Sensing of Environment*, *308*(nil), 114191.

Liu, J., Gong, M., Qin, K., & Zhang, P. (2018). A Deep Convolutional Coupling Network for Change Detection Based on Heterogeneous Optical and Radar Images. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(3), 545–559. https://doi.org/10.1109/TNNLS.2016.2636227

Liu, S., Brandt, M., Nord-Larsen, T., Chave, J., Reiner, F., Lang, N., Tong, X., Ciais, P., Igel, C., Pascual, A., Guerra-Hernandez, J., Li, S., Mugabowindekwe, M., Saatchi, S., Yue, Y., Chen, Z., & Fensholt, R. (2023). The overlooked contribution of trees outside forests to tree cover and woody biomass across Europe. *Science Advances*, *9*(37). https://doi.org/10.1126/sciadv.adh4097

Liu, X., Deng, C., Chanussot, J., Hong, D., & Zhao, B. (2019). stfnet: a two-stream convolutional neural network for spatiotemporal image fusion. *IEEE Transactions on Geoscience and Remote Sensing*, *57*(9), 6552–6564.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9992–10002. https://doi.org/10.1109/ICCV48922.2021.00986

Lonjou, V., Desjardins, C., Hagolle, O., Petrucci, B., Tremas, T., Dejus, M., Makarau, A., & Auer, S. (2016). MACCS-ATCOR joint algorithm (MAJA). *Remote Sensing of Clouds and the Atmosphere XXI*, *10001*. https://doi.org/10.1117/12.2240935

Loshchilov, I., & Hutter, F. (2016a). *SGDR: Stochastic Gradient Descent With Warm Restarts*.

Loshchilov, I., & Hutter, F. (2016b). Sgdr: Stochastic Gradient Descent With Warm Restarts. *Nil*, *nil*(nil), nil. https://doi.org/10.48550/ARXIV.1608.03983

Lu, X., Zheng, X., & Yuan, Y. (2017). Remote Sensing Scene Classification by Unsupervised Representation Learning. *IEEE Transactions on Geoscience and Remote Sensing*, *55*(9), 5148–5157. https://doi.org/10.1109/TGRS.2017.2702596

Luppino, L. T., Kampffmeyer, M., Bianchi, F. M., Moser, G., Serpico, S. B., Jenssen, R., & Anfinsen, S. N. (2022). Deep Image Translation With an Affinity-Based Change Prior for Unsupervised Multimodal Change Detection. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–22. https://doi.org/10.1109/TGRS.2021.3056196

Lutio, R. d., D'aronco, S., Wegner, J. D., & Schindler, K. (2019). Guided super-resolution as pixel-to-pixel transformation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8829–8837.

Lv, Z., Huang, H., Gao, L., Benediktsson, J. A., Zhao, M., & Shi, C. (2022). Simple Multiscale UNet for Change Detection With Heterogeneous Remote Sensing

Images. *IEEE Geoscience and Remote Sensing Letters*, *19*, 1–5. https://doi.org/10.1109/LGRS.2022.3173300

Manas, O., Lacoste, A., Giro-i-Nieto, X., Vazquez, D., & Rodriguez, P. (2021). Seasonal Contrast: Unsupervised Pre-Training from Uncurated Remote Sensing Data. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9394–9403. https://doi.org/10.1109/ICCV48922.2021.00928

Märtens, M., Izzo, D., Kržič, A., & Cox, D. (2019). *Super-Resolution of PROBA-V Images Using Convolutional Neural Networks*. https://doi.org/10.48550/arXiv.1907.01821

Maselli, F., Papale, D., Puletti, N., Chirici, G., & Corona, P. (2009). Combining remote sensing and ancillary data to monitor the gross productivity of water-limited forest ecosystems. *Remote Sensing of Environment*, *113*(3). https://doi.org/10.1016/j.rse.2008.11.008

McCallum, I., Wagner, W., Schmullius, C., Shvidenko, A., Obersteiner, M., Fritz, S., & Nilsson, S. (2009). Satellite-based terrestrial production efficiency modeling. *Carbon Balance and Management*, *4*(1), 8. https://doi.org/10.1186/1750-0680-4-8

Merlin, O., Duchemin, B., Hagolle, O., Jacob, F., Coudert, B., Chehbouni, G., Dedieu, G., Garatuza, J., & Kerr, Y. (2010). Disaggregation of modis surface temperature over an agricultural area using a time series of formosat-2 images. *Remote Sensing of Environment*, *114*(11), 2500–2512.

Michel, J., Ekaterina, K., & Jordi, I. (2025). Revisiting remote sensing cross-sensor Single Image Super-Resolution: the overlooked impact of geometric and radiometric distortion. *TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING*.

Michel, J., Vinasco-Salinas, J., Inglada, J., & Hagolle, O. (2022). SEN2VENμS, a Dataset for the Training of Sentinel-2 Super-Resolution Algorithms. *Data*, *7*(7). https://doi.org/10.3390/data7070096

Milenković, M., Reiche, J., Armston, J., Neuenschwander, A., De Keersmaecker, W., Herold, M., & Verbesselt, J. (2022). Assessing Amazon rainforest regrowth with GEDI and ICESat-2 data. *Science of Remote Sensing*, *5*, 100051. https://doi.org/10.1016/j.srs.2022.100051

Misra, G., Cawkwell, F., & Wingler, A. (2020). Status of Phenological Research Using Sentinel-2 Data: A Review. *Remote Sensing*, *12*, 2760. https://doi.org/10.3390/rs12172760

Mitchell, A. L., Rosenqvist, A., & Mora, B. (2017). Current remote sensing approaches to monitoring forest degradation in support of countries measurement, reporting and verification (MRV) systems for REDD+. *Carbon Balance and Management*, *12*(1), 9. https://doi.org/10.1186/s13021-017-0078-9

Mittal, A., Moorthy, A. K., & Bovik, A. C. (2012). No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, *21*(12), 4695–4708.

Moskolaï, W. R., Abdou, W., Dipanda, A., & Kolyang. (2021). Application of deep learning architectures for satellite image time series prediction: a review. *Remote Sensing*, *13*(23), 4822.

Mou, L., Ghamisi, P., & Zhu, X. X. (2018). Unsupervised Spectral–Spatial Feature Learning via Deep Residual Conv–Deconv Network for Hyperspectral Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, *56*(1), 391–406. https://doi.org/10.1109/TGRS.2017.2748160

Nalepa, J., Le Saux, B., Longepe, N., Tulczyjew, L., Myller, M., Kawulok, M., Smykala, K., & Gumiela, M. (2022). The Hyperview Challenge: Estimating Soil Parameters from Hyperspectral Images. *2022 IEEE International Conference on Image Processing (ICIP)*, 4268–4272. https://doi.org/10.1109/ICIP46576.2022.9897443

Ngo, Y.-N., HO TONG MINH, D., Baghdadi, N., & Fayad, I. (2023). Tropical Forest Top Height by GEDI: From Sparse Coverage to Continuous Data. *Remote Sensing*, *15*, 975. https://doi.org/10.3390/rs15040975

Nguyen, B. M., Tian, G., Vo, M.-T., Michel, A., Corpetti, T., & Granero-Belinchon, C. (2022). Convolutional neural network modelling for modis land surface temperature super-resolution. *2022 30th European Signal Processing Conference (EUSIPCO)*, 1806–1810.

Nguyen, N. L., Anger, J., Raad, L., Galerne, B., & Facciolo, G. (2023a). On the role of alias and band-shift for sentinel-2 super-resolution. *ArXiv Preprint ArXiv:2302.11494*.

Noroozi, M., Pirsiavash, H., & Favaro, P. (2017). Representation Learning by Learning to Count. *2017 IEEE International Conference on Computer Vision (ICCV)*, 5899–5907. https://doi.org/10.1109/ICCV.2017.628

Oeser, J., Pflugmacher, D., Senf, C., Heurich, M., & Hostert, P. (2017). Using Intra-Annual Landsat Time Series for Attributing Forest Disturbance Agents in Central Europe. *Forests*, *8*(7), 251. https://doi.org/10.3390/f8070251

Okabayashi, A., Audebert, N., Donike, S., & Pelletier, C. (2024). *Cross-sensor super-resolution of irregularly sampled Sentinel-2 time series*. https://doi.org/10.1109/CVPRW63382.2024.00055

Olsson, P.-O., Lindström, J., & Eklundh, L. (2016). Near real-time monitoring of insect induced defoliation in subalpine birch forests with MODIS derived NDVI. *Remote Sensing of Environment*, *181*, 42–53. https://doi.org/10.1016/j.rse.2016.03.040

Oord, A. van den, Li, Y., & Vinyals, O. (2018). *Representation Learning with Contrastive Predictive Coding*.

Owens, A., Wu, J., McDermott, J. H., Freeman, W. T., & Torralba, A. (2016). *Ambient Sound Provides Supervision for Visual Learning* (pp. 801–816). https://doi.org/10.1007/978-3-319-46448-0_48

PAGE, E. S. (1954). CONTINUOUS INSPECTION SCHEMES. *Biometrika*, *41*(1–2), 100–115. https://doi.org/10.1093/biomet/41.1-2.100

Palsson, B., Sveinsson, J. R., & Ulfarsson, M. O. (2022). Blind Hyperspectral Unmixing Using Autoencoders: A Critical Comparison. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *15*, 1340–1372. https://doi.org/10.1109/JSTARS.2021.3140154

Palsson, F., Sveinsson, J. R., & Ulfarsson, M. O. (2018). Sentinel-2 image fusion using a deep residual network. *Remote Sensing*, *10*(8).

Parelius, E. J. (2023). A Review of Deep-Learning Methods for Change Detection in Multispectral Remote Sensing Images. *Remote Sensing*, *15*(8), 2092. https://doi.org/10.3390/rs15082092

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., & Lerer, A. (2017). *Automatic differentiation in PyTorch*.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 32* (pp. 8024–8035). Curran Associates, Inc.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan Gregory and Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito Zachary and Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., & Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 32* (pp. 8024–8035). Curran Associates, Inc. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

Paulino, E., Schlerf, M., Röder, A., Stoffels, J., & Udelhoven, T. (2024). Forest disturbance characterization in the era of earth observation big data: A mapping review. *International Journal of Applied Earth Observation and Geoinformation*, *128*, 103755. https://doi.org/10.1016/j.jag.2024.103755

Pekel, J.-F., Cottam, A., Gorelick, N., & Belward, A. S. (2016). High-resolution mapping of global surface water and its long-term changes. *Nature*, *540*(7633), 418–422. https://doi.org/10.1038/nature20584

Peng, D., Zhang, Y., & Guan, H. (2019). End-to-End Change Detection for High Resolution Satellite Images Using Improved UNet++. *Remote Sensing*, *11*(11), 1382. https://doi.org/10.3390/rs11111382

Phiri, D., Simwanda, M., Salekin, S., Nyirenda, V., Murayama, Y., & Ranagalage, M. (2020). Sentinel-2 Data for Land Cover/Use Mapping: A Review. *Remote Sensing*, *12*, 2291. https://doi.org/10.3390/rs12142291

Pickens, A. H., Hansen, M. C., Hancher, M., Stehman, S. V., Tyukavina, A., Potapov, P., Marroquin, B., & Sherani, Z. (2020). Mapping and sampling to characterize global inland water dynamics from 1999 to 2018 with full Landsat time-series. *Remote Sensing of Environment*, *243*, 111792. https://doi.org/10.1016/j.rse.2020.111792

Pilgrim, M. (2009). Serializing Python Objects. In *Dive Into Python 3* (pp. 205–223). Apress. https://doi.org/10.1007/978-1-4302-2416-7_13

Pineda, F., Ayma, V., & Beltran, C. (2020). a generative adversarial network approach for super-resolution of sentinel-2 satellite images. *The International Archives of Photogrammetry, Remote and Spatial Information Sciences*, *43*, 9–14.

Potapov, P., Li, X., Hernandez-Serna, A., Tyukavina, A., Hansen, M. C., Kommareddy, A., Pickens, A., Turubanova, S., Tang, H., Silva, C. E., Armston, J., Dubayah, R., Blair, J. B., & Hofton, M. (2021). Mapping global forest canopy height through integration of GEDI and Landsat data. *Remote Sensing of Environment*, *253*, 112165. https://doi.org/10.1016/j.rse.2020.112165

Pouliot, D. A., Latifovic, R., Pasher, J., & Duffe, J. (2018). Landsat Super-Resolution Enhancement Using Convolution Neural Networks and Sentinel-2 for Training. *Remote Sensing*, *10*, 394. https://doi.org/10.3390/rs10030394

Pu, Y., Gan, Z., Henao, R., Yuan, X., Li, C., Stevens, A., & Carin, L. (2016). *Variational Autoencoder for Deep Learning of Images, Labels and Captions*.

Puhm, M., Deutscher, J., Hirschmugl, M., Wimmer, A., Schmitt, U., & Schardt, M. (2020). A Near Real-Time Method for Forest Change Detection Based on a Structural Time Series Model and the Kalman Filter. *Remote Sensing*, *12*(19), 3135. https://doi.org/10.3390/rs12193135

Quegan, S., & Jiong Jiong Yu. (2001). Filtering of multichannel SAR images. *IEEE Transactions on Geoscience and Remote Sensing*, *39*(11), 2373–2379. https://doi.org/10.1109/36.964973

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). *Learning Transferable Visual Models From Natural Language Supervision*.

Ranftl, R., Bochkovskiy, A., & Koltun, V. (2021). Vision Transformers for Dense Prediction. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 12179–12188.

Rawat, W., & Wang, Z. (2017). Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. *Neural Computation*, *29*(9), 2352–2449. https://doi.org/10.1162/neco_a_00990

Reed, C. J., Gupta, R., Li, S., Brockman, S., Funk, C., Clipp, B., Keutzer, K., Candido, S., Uyttendaele, M., & Darrell, T. (2022). Scale-mae: a scale-aware masked autoencoder for multiscale geospatial representation learning. *CoRR*.

Reiche, J., Mullissa, A., Slagter, B., Gou, Y., Tsendbazar, N.-E., Odongo-Braun, C., Vollrath, A., Weisse, M. J., Stolle, F., Pickens, A., Donchyts, G., Clinton, N., Gorelick, N., & Herold, M. (2021). Forest disturbance alerts for the Congo Basin using Sentinel-1. *Environmental Research Letters*, *16*(2), 024005. https://doi.org/10.1088/1748-9326/abd0a8

Réjou-Méchain, M., Muller-Landau, H. C., Detto, M., Thomas, S. C., Le Toan, T., Saatchi, S. S., Barreto-Silva, J. S., Bourg, N. A., Bunyavejchewin, S., Butt, N., Brockelman, W. Y., Cao, M., Cárdenas, D., Chiang, J.-M., Chuyong, G. B., Clay, K., Condit, R., Dattaraja, H. S., Davies, S. J., … Chave, J. (2014). Local spatial structure of forest biomass and its consequences for remote sensing of carbon stocks. *Biogeosciences*, *11*(23), 6827–6840. https://doi.org/10.5194/bg-11-6827-2014

Ren, Z., & Lee, Y. J. (2018). Cross-Domain Self-Supervised Multi-task Feature Learning Using Synthetic Imagery. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 762–771. https://doi.org/10.1109/CVPR.2018.00086

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-Resolution Image Synthesis with Latent Diffusion Models. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2022-June*. https://doi.org/10.1109/CVPR52688.2022.01042

Romero, L. S., Marcello, J., & Vilaplana, V. (2020). Super-resolution of Sentinel-2 imagery using generative adversarial networks. *Remote Sensing*, *12*(15). https://doi.org/10.3390/RS12152424

Ronneberger, O., Fischer, P., & Brox, T. (2015). *U-Net: Convolutional Networks for Biomedical Image Segmentation*.

Roumeliotis, K. I., & Tselikas, N. D. (2023). ChatGPT and Open-AI Models: A Preliminary Review. *Future Internet*, *15*(6), 192. https://doi.org/10.3390/fi15060192

Roy, D. P., Ju, J., Lewis, P., Schaaf, C., Gao, F., Hansen, M., & Lindquist, E. (2008). Multi-temporal modis–landsat data fusion for relative radiometric normalization, gap filling, and prediction of landsat data. *Remote Sensing of Environment*, *112*(6), 3112–3130.

Sadeh, Y., Zhu, X., Dunkerley, D., Walker, J., Zhang, Y., V S, M., Chenu, K., & Rozenstein, O. (2021). Fusion of Sentinel-2 and PlanetScope time-series data into daily 3 m surface reflectance and wheat LAI monitoring. *International Journal of Applied Earth Observation and Geoinformation*, *96*, 102260. https://doi.org/10.1016/j.jag.2020.102260

Saha, S., Bovolo, F., & Bruzzone, L. (2021). Building Change Detection in VHR SAR Images via Unsupervised Deep Transcoding. *IEEE Transactions on Geoscience and Remote Sensing*, *59*(3), 1917–1929. https://doi.org/10.1109/TGRS.2020.3000296

Saha, S., Ebel, P., & Zhu, X. X. (2022a). Self-Supervised Multisensor Change Detection. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–10. https://doi.org/10.1109/TGRS.2021.3109957

Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D. J., & Norouzi, M. (2023). Image Super-Resolution via Iterative Refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(4). https://doi.org/10.1109/TPAMI.2022.3204461

Salgueiro, L., Marcello, J., & Vilaplana, V. (2021). Single-image super-resolution of sentinel-2 low resolution bands with residual dense convolutional neural networks. *Remote Sensing*, *13*(24), 5007.

Salgueiro Romero, L., Marcello, J., & Vilaplana, V. (2020). Super-Resolution of Sentinel-2 Imagery Using Generative Adversarial Networks. *Remote Sensing*, *12*, 2424. https://doi.org/10.3390/rs12152424

Sambandham, V. T., Kirchheim, K., Ortmeier, F., & Mukhopadhaya, S. (2024). Deep learning-based harmonization and super-resolution of landsat-8 and sentinel-2 images. *ISPRS Journal of Photogrammetry and Remote Sensing*, *212*, 274–288.

Sarzynski, T., Giam, X., Carrasco, L., & Lee, J. S. H. (2020). Combining Radar and Optical Imagery to Map Oil Palm Plantations in Sumatra, Indonesia, Using the Google Earth Engine. *Remote Sensing*, *12*(7), 1220. https://doi.org/10.3390/rs12071220

Scheibenreif, L., Hanna, J., Mommert, M., & Borth, D. (2022). Self-supervised Vision Transformers for Land-cover Segmentation and Classification. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1421–1430. https://doi.org/10.1109/CVPRW56347.2022.00148

Schramm, M., Pebesma, E., Milenković, M., Foresta, L., Dries, J., Jacob, A., Wagner, W., Mohr, M., Neteler, M., Kadunc, M., Miksa, T., Kempeneers, P., Verbesselt, J., Gößwein, B., Navacchi, C., Lippens, S., & Reiche, J. (2021). The openeo api-harmonising the use of earth observation cloud services using virtual data cube functionalities. *Remote Sensing*, *13*(6), 1125.

Sebald, J., Senf, C., & Seidl, R. (2021). Human or natural? Landscape context improves the attribution of forest disturbances mapped from Landsat in Central Europe. *Remote Sensing of Environment*, *262*, 112502. https://doi.org/10.1016/j.rse.2021.112502

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based

Localization. *2017 IEEE International Conference on Computer Vision (ICCV)*, 618–626. https://doi.org/10.1109/ICCV.2017.74

Shafique, A., Cao, G., Khan, Z., Asad, M., & Aslam, M. (2022). Deep Learning-Based Change Detection in Remote Sensing Images: A Review. *Remote Sensing*, *14*(4), 871. https://doi.org/10.3390/rs14040871

Shao, Z., Cai, J., Fu, P., Hu, L., & Liu, T. (2019). Deep learning-based fusion of Landsat-8 and Sentinel-2 images for a harmonized surface reflectance product. *Remote Sensing of Environment*, *235*. https://doi.org/10.1016/j.rse.2019.111425

Shao, Z., Wang, L., Wang, Z., & Deng, J. (2019). Remote Sensing Image Super-Resolution Using Sparse Representation and Coupled Sparse Autoencoder. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *12*(8), 2663–2674. https://doi.org/10.1109/JSTARS.2019.2925456

Shelhamer, E., Long, J., & Darrell, T. (2016). *Fully Convolutional Networks for Semantic Segmentation*.

Shi, W., Caballero, J., Huszar, F., Totz, J., Aitken, A. P., Bishop, R., Rueckert, D., & Wang, Z. (2016). Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2016-December*. https://doi.org/10.1109/CVPR.2016.207

Shimizu, K., Ota, T., Mizoue, N., & Yoshida, S. (2019). A comprehensive evaluation of disturbance agent classification approaches: Strengths of ensemble classification, multiple indices, spatio-temporal variables, and direct prediction. *ISPRS Journal of Photogrammetry and Remote Sensing*, *158*, 99–112. https://doi.org/10.1016/j.isprsjprs.2019.10.004

Silberman, N., Hoiem, D., Kohli, P., & Fergus, R. (2012). *Indoor Segmentation and Support Inference from RGBD Images* (pp. 746–760). https://doi.org/10.1007/978-3-642-33715-4_54

Skidmore, A. K., Coops, N. C., Neinavaz, E., Ali, A., Schaepman, M. E., Paganini, M., Kissling, W. D., Vihervaara, P., Darvishzadeh, R., Feilhauer, H., Fernandez, M., Fernández, N., Gorelick, N., Geijzendorffer, I., Heiden, U., Heurich, M., Hobern, D., Holzwarth, S., Muller-Karger, F. E., … Wingate, V. (2021). Priority list of biodiversity metrics to observe from space. *Nature Ecology & Evolution*, *5*(7), 896–906. https://doi.org/10.1038/s41559-021-01451-x

Song, S., Lichtenberg, S. P., & Xiao, J. (2015). SUN RGB-D: A RGB-D scene understanding benchmark suite. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 567–576. https://doi.org/10.1109/CVPR.2015.7298655

Stewart, A. J., Robinson, C., Corley, I. A., Ortiz, A., Ferres, J. M. L., & Banerjee, A. (2022). TorchGeo. *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*, 1–12. https://doi.org/10.1145/3557915.3560953

Stocker, B. D., Zscheischler, J., Keenan, T. F., Prentice, I. C., Seneviratne, S. I., & Peñuelas, J. (2019). Drought impacts on terrestrial primary production underestimated by satellite monitoring. *Nature Geoscience*, *12*(4). https://doi.org/10.1038/s41561-019-0318-6

Su, H., Li, Y., Xu, Y., Fu, X., & Liu, S. (2025). A review of deeplearning-based super-resolution: From methods to applications. *Pattern Recognition*, *157*(nil), 110935.

Sumbul, G., Charfuelan, M., Demir, B., & Markl, V. (2019). Bigearthnet: A Large-Scale Benchmark Archive for Remote Sensing Image Understanding. *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, 5901–5904. https://doi.org/10.1109/IGARSS.2019.8900532

Sumbul, G., de Wall, A., Kreuziger, T., Marcelino, F., Costa, H., Benevides, P., Caetano, M., Demir, B., & Markl, V. (2021a). BigEarthNet-MM: A Large-Scale, Multimodal, Multilabel Benchmark Archive for Remote Sensing Image Classification and Retrieval [Software and Data Sets]. *IEEE Geoscience and Remote Sensing Magazine*, *9*(3), 174–180. https://doi.org/10.1109/MGRS.2021.3089174

Sundararajan, M., Taly, A., & Yan, Q. (2017). *Axiomatic Attribution for Deep Networks*.

Tai, Y., Yang, J., & Liu, X. (2017). Image super-resolution via deep recursive residual network. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, *2017-January*. https://doi.org/10.1109/CVPR.2017.298

Tan, H. H., & Lim, K. H. (2019). Vanishing Gradient Mitigation with Deep Learning Neural Network Optimization. *2019 7th International Conference on Smart Computing & Communications (ICSCC)*, 1–4. https://doi.org/10.1109/ICSCC.2019.8843652

Tan, Z., Di, L., Zhang, M., Guo, L., & Gao, M. (2019). An enhanced deep convolutional model for spatiotemporal image fusion. *Remote Sensing*, *11*(24), 2898.

Tan, Z., Gao, M., Li, X., & Jiang, L. (2022). A flexible reference-insensitive spatiotemporal fusion model for remote sensing images using conditional generative adversarial network. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–13.

Tan, Z., Yue, P., Di, L., & Tang, J. (2018). Deriving high spatiotemporal remote sensing images using deep convolutional network. *Remote Sensing*, *10*(7), 1066.

Tao, Y., & Muller, J. P. (2021). Super-resolution restoration of spaceborne ultra-high-resolution images using the ucl optigan system. *Remote Sensing*, *13*(12). https://doi.org/10.3390/rs13122269

Tottrup, C., Druce, D., Meyer, R. P., Christensen, M., Riffler, M., Dulleck, B., Rastner, P., Jupova, K., Sokoup, T., Haag, A., Cordeiro, M. C. R., Martinez, J.-M., Franke, J., Schwarz, M., Vanthof, V., Liu, S., Zhou, H., Marzi, D., Rudiyanto, R., … Paganini, M. (2022). Surface Water Dynamics from Space: A Round Robin Intercomparison of Using Optical and SAR High-Resolution Satellite Observations for Regional

Surface Water Detection. *Remote Sensing*, *14*(10), 2410. https://doi.org/10.3390/rs14102410

Tunstall, L., Leandro Von Werra, & Thomas Wolf. (2022). *Natural language processing with transformers*. O'Reilly Media, Inc.

Vajsová, B., Fasbender, D., Wirnhardt, C., Lemajic, S., & Devos, W. (2020). Assessing spatial limits of Sentinel-2 data on arable crops in the context of checks by monitoring. *Remote Sensing*, *12*(14). https://doi.org/10.3390/rs12142195

Vanschoren, J. (2019). *Meta-Learning* (pp. 35–61). https://doi.org/10.1007/978-3-030-05318-5_2

Verbesselt, J., Hyndman, R., Zeileis, A., & Culvenor, D. (2010a). Phenological change detection while accounting for abrupt and gradual trends in satellite image time series. *Remote Sensing of Environment*, *114*(12), 2970–2980. https://doi.org/10.1016/j.rse.2010.08.003

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

Viana-Soto, A., & Senf, C. (2024). *The European Forest Disturbance Atlas: a forest disturbance monitoring system using the Landsat archive*. ESSD – Land/Biogeosciences and biodiversity. https://doi.org/10.5194/essd-2024-361

Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. *Proceedings of the 25th International Conference on Machine Learning - ICML '08*, 1096–1103. https://doi.org/10.1145/1390156.1390294

Vinholi, J. G., Silva, D., Machado, R., & Pettersson, M. I. (2022). CNN-Based Change Detection Algorithm for Wavelength-Resolution SAR Images. *IEEE Geoscience and Remote Sensing Letters*, *19*, 1–5. https://doi.org/10.1109/LGRS.2020.3027382

Vivone, G., Alparone, L., Chanussot, J., Dalla Mura, M., Garzelli, A., Licciardi, G. A., Restaino, R., & Wald, L. (2014). A critical comparison among pansharpening algorithms. *IEEE Transactions on Geoscience and Remote Sensing*, *53*(5), 2565–2586.

Wald, L., Ranchin, T., & Mangolini, M. (1997a). Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images. *Photogrammetric Engineering and Remote Sensing*, *63*(6), 691–699.

Wang, C., Su, W., & Gu, H. (2022). SAR Image Change Detection Based on Semisupervised Learning and Two-Step Training. *IEEE Geoscience and Remote Sensing Letters*, *19*, 1–5. https://doi.org/10.1109/LGRS.2021.3050746

Wang, J.-J., Dobigeon, N., Chabert, M., Wang, D.-C., Huang, T.-Z., & Huang, J. (2024). CD-GAN: A robust fusion-based generative adversarial network for

unsupervised remote sensing change detection with heterogeneous sensors. *Information Fusion*, *107*, 102313. https://doi.org/10.1016/j.inffus.2024.102313

Wang, L., Luc, P., Recasens, A., Alayrac, J.-B., & Oord, A. van den. (2021). *Multimodal Self-Supervised Learning of General Audio Representations*.

Wang, L., & Yoon, K.-J. (2022). Knowledge Distillation and Student-Teacher Learning for Visual Intelligence: A Review and New Outlooks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(6), 3048–3068. https://doi.org/10.1109/TPAMI.2021.3055564

Wang, M., Sun, R., Zhu, A., & Xiao, Z. (2020). Evaluation and Comparison of Light Use Efficiency and Gross Primary Productivity Using Three Different Approaches. *Remote Sensing*, *12*(6), 1003. https://doi.org/10.3390/rs12061003

Wang, Q., Ma, Y., Zhao, K., & Tian, Y. (2022). A Comprehensive Survey of Loss Functions in Machine Learning. *Annals of Data Science*, *9*(2), 187–212. https://doi.org/10.1007/s40745-020-00253-5

Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., & Change Loy, C. (2018). Esrgan: Enhanced super-resolution generative adversarial networks. *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 0.

Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., & Loy, C. C. (2019). *ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks: Munich, Germany, September 8-14, 2018, Proceedings, Part V* (pp. 63–79). https://doi.org/10.1007/978-3-030-11021-5_5

Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., & Qiao Yu and Change Loy, C. (2018). Esrgan: Enhanced super-resolution generative adversarial networks. *Proceedings of the European Conference on Computer (ECCV) Workshops*, 0.

Wang, Y., Albrecht, C. M., Braham, N. A. A., Liu, C., Xiong, Z., & Zhu, X. X. (2023). *DeCUR: decoupling common & unique representations for multimodal self-supervision*.

Wang, Y., Albrecht, C. M., Braham, N. A. A., Mou, L., & Zhu, X. X. (2022). Self-Supervised Learning in Remote Sensing: A review. *IEEE Geoscience and Remote Sensing Magazine*, *10*(4), 213–247. https://doi.org/10.1109/MGRS.2022.3198244

Wang, Y., Albrecht, C. M., Braham, N. A. A., Mou, L., & Zhu, X. X. (2022c). Self-Supervised Learning in Remote Sensing: A review. *IEEE Geoscience and Remote Sensing Magazine*, *10*(4), 213–247. https://doi.org/10.1109/MGRS.2022.3198244

Wang, Y., Albrecht, C. M., & Zhu, X. X. (2024). Multilabel-Guided Soft Contrastive Learning for Efficient Earth Observation Pretraining. *IEEE Transactions on Geoscience and Remote Sensing*, *62*, 1–16. https://doi.org/10.1109/TGRS.2024.3466896

Wang, Y., Braham, N. A. A., Xiong, Z., Liu, C., Albrecht, C. M., & Zhu, X. X. (2023). SSL4EO-S12: A large-scale multimodal, multitemporal dataset for self-supervised learning in Earth observation [Software and Data Sets]. *IEEE Geoscience and Remote Sensing Magazine*, *11*(3), 98–106. https://doi.org/10.1109/MGRS.2023.3281651

Wang, Y., Hernández, H. H., Albrecht, C. M., & Zhu, X. X. (2023). *Feature Guided Masked Autoencoder for Self-supervised Learning in Remote Sensing*.

Wei, D., Lim, J., Zisserman, A., & Freeman, W. T. (2018). Learning and Using the Arrow of Time. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8052–8060. https://doi.org/10.1109/CVPR.2018.00840

Wilson, A., & Jetz, W. (2016). Remotely Sensed High-Resolution Global Cloud Dynamics for Predicting Ecosystem and Biodiversity Distributions. *PLoS Biology*, *14*, e1002415. https://doi.org/10.1371/journal.pbio.1002415

Wu, J., Li, B., Qin, Y., Ni, W., Zhang, H., Fu, R., & Sun, Y. (2021). A multiscale graph convolutional network for change detection in homogeneous and heterogeneous remote sensing images. *International Journal of Applied Earth Observation and Geoinformation*, *105*, 102615. https://doi.org/10.1016/j.jag.2021.102615

Wu, Y., & Huang, M. (2024). A unified generative adversarial network with convolution and transformer for remote sensing image fusion. *IEEE Transactions on Geoscience and Remote Sensing*, *nil*(nil), 1.

Wulder, M., Roy, D., Radeloff, V., Loveland, T., Anderson, M., Johnson, D., Healey, S., Zhu, Z., Scambos, T., Pahlevan, N., Hansen, M., Gorelick, N., Crawford, C., Masek, J., Hermosilla, T., White, J., Belward, A., Schaaf, C., Woodcock, C., & Cook, B. (2022). Fifty years of Landsat science and impacts. *Remote Sensing of Environment*, *280*, 113195. https://doi.org/10.1016/j.rse.2022.113195

Xiao, J., Aggarwal, A. K., Duc, N. H., Arya, A., Rage, U. K., & Avtar, R. (2023). A review of remote sensing image spatiotemporal fusion: Challenges, applications and recent trends. *Remote Sensing Applications: Society and Environment*, *32*(nil), 101005.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., & Luo, P. (2021). *SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers*.

Xie, Y., Hu, J., He, K., Cao, L., Yang, K., & Chen, L. (2024). The gan spatiotemporal fusion model based on multi-scale convolution and attention mechanism for remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 1–13.

Xie, Z., Lin, Y., Yao, Z., Zhang, Z., Dai, Q., Cao, Y., & Hu, H. (2021). *Self-Supervised Learning with Swin Transformers*.

Yang, G., Qian, Y., Liu, H., Tang, B., Qi, R., Lu, Y., & Geng, J. (2022). Msfusion: Multistage for remote sensing image spatiotemporal fusion based on texture transformer

and convolutional neural network. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *15*(nil), 4653–4666.

Yang, L., Jin, S., Danielson, P., Homer, C., Gass, L., Bender, S. M., Case, A., Costello, C., Dewitz, J., Fry, J., Funk, M., Granneman, B., Liknes, G. C., Rigge, M., & Xian, G. (2018). A new generation of the United States National Land Cover Database: Requirements, research priorities, design, and implementation strategies. *ISPRS Journal of Photogrammetry and Remote Sensing*, *146*, 108–123. https://doi.org/10.1016/j.isprsjprs.2018.09.006

Yang, X., Song, Z., King, I., & Xu, Z. (2023). A Survey on Deep Semi-Supervised Learning. *IEEE Transactions on Knowledge and Data Engineering*, *35*(9), 8934–8954. https://doi.org/10.1109/TKDE.2022.3220219

Yuan, Y., Lin, L., Liu, Q., Hang, R., & Zhou, Z.-G. (2022). Sitsformer: a pre-trained spatio-spectral-temporal representation model for sentinel-2 time series classification. *International Journal of Applied Earth Observation and Geoinformation*, *106*(nil), 102651.

Zanaga, D. , Van De Kerchove, R. , Daems, D. , De Keersmaecker, W. , Brockmann, C. , Kirches, G. , Wevers, J. , Cartus, O. S. , Lesiv, M. , Herold, M., Tsendbazar, N.-E. , Xu, P. , Ramoino, F. , & Arino, O. (2022). *ESA WorldCover 10 m 2021 v200 (Version v200)*.

Zbontar, J., Jing, L., Misra, I., LeCun, Y., & Deny, S. (2021). *Barlow Twins: Self-Supervised Learning via Redundancy Reduction*.

Zérah, Y., Valero, S., & Inglada, J. (2023). Physics-Driven Probabilistic Deep Learning for the Inversion of Physical Models With Application to Phenological Parameter Retrieval From Satellite Times Series. *IEEE Transactions on Geoscience and Remote Sensing*, *61*, 1–23. https://doi.org/10.1109/TGRS.2023.3284992

Zhang, H., Lin, M., Yang, G., & Zhang, L. (2023). ESCNet: An End-to-End Superpixel-Enhanced Change Detection Network for Very-High-Resolution Remote Sensing Images. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(1), 28–42. https://doi.org/10.1109/TNNLS.2021.3089332

Zhang, J., Liu, H., Yang, K., Hu, X., Liu, R., & Stiefelhagen, R. (2023). CMX: Cross-Modal Fusion for RGB-X Semantic Segmentation With Transformers. *IEEE Transactions on Intelligent Transportation Systems*, *24*(12), 14679–14694. https://doi.org/10.1109/TITS.2023.3300537

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, W., Jiao, L., Liu, F., Yang, S., Song, W., & Liu, J. (2022). Sparse Feature Clustering Network for Unsupervised SAR Image Change Detection. *IEEE Transactions*

on Geoscience and Remote Sensing, *60*, 1–13. https://doi.org/10.1109/TGRS.2022.3167745

Zhang, X., Li, S., Tan, Z., & Li, X. (2024). Enhanced wavelet based spatiotemporal fusion networks using cross-paired remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, *211*(nil), 281–297.

Zhang, X., Tian, M., Xing, Y., Yue, Y., Li, Y., Yin, H., Xia, R., Jin, J., & Zhang, Y. (2022). ADHR-CDNet: Attentive Differential High-Resolution Change Detection Network for Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–13. https://doi.org/10.1109/TGRS.2022.3221492

Zhang, X., Yue, Y., Gao, W., Yun, S., Su, Q., Yin, H., & Zhang, Y. (2022). DifUnet++: A Satellite Images Change Detection Network Based on Unet++ and Differential Pyramid. *IEEE Geoscience and Remote Sensing Letters*, *19*, 1–5. https://doi.org/10.1109/LGRS.2021.3049370

Zhang, Y., Fan, R., Duan, P., Dong, J., & Lei, Z. (2024). Dcdgan-stf: a multiscale deformable convolution distillation gan for remote sensing image spatiotemporal fusion. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *nil*(nil), 1–15.

Zhang, Y., Ma, J., Liang, S., Li, X., & Li, M. (2020). An Evaluation of Eight Machine Learning Regression Algorithms for Forest Aboveground Biomass Estimation from Multiple Satellite Data Products. *Remote Sensing*, *12*(24), 4015. https://doi.org/10.3390/rs12244015

Zhao, K., Wulder, M. A., Hu, T., Bright, R., Wu, Q., Qin, H., Li, Y., Toman, E., Mallick, B., Zhang, X., & Brown, M. (2019). Detecting change-point, trend, and seasonality in satellite time series data to track abrupt changes and nonlinear dynamics: A Bayesian ensemble algorithm. *Remote Sensing of Environment*, *232*, 111181. https://doi.org/10.1016/j.rse.2019.04.034

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., … Wen, J.-R. (2023). *A Survey of Large Language Models*.

Zhou, C., Li, Q., Li, C., Yu, J., Liu, Y., Wang, G., Zhang, K., Ji, C., Yan, Q., He, L., Peng, H., Li, J., Wu, J., Liu, Z., Xie, P., Xiong, C., Pei, J., Yu, P. S., & Sun, L. (2023). *A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT*.

Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., & Liang, J. (2020). UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation. *IEEE Transactions on Medical Imaging*, *39*(6), 1856–1867. https://doi.org/10.1109/TMI.2019.2959609

Zhu, X., Cai, F., Tian, J., & Williams, T. K.-A. (2018). Spatiotemporal fusion of multisource remote sensing data: Literature survey, taxonomy, principles, applications, and future directions. *Remote Sensing*, *10*(4), 527.

Zhu, X., Nie, S., Wang, C., Xi, X., Lao, J., & Li, D. (2022). Consistency analysis of forest height retrievals between GEDI and ICESat-2. *Remote Sensing of Environment*, *281*, 113244. https://doi.org/10.1016/j.rse.2022.113244

Zhu, Z. (2017). Change detection using landsat time series: A review of frequencies, preprocessing, algorithms, and applications. *ISPRS Journal of Photogrammetry and Remote Sensing*, *130*, 370–384. https://doi.org/10.1016/j.isprsjprs.2017.06.013

Zhu, Z., & Woodcock, C. E. (2012). Object-based cloud and cloud shadow detection in Landsat imagery. *Remote Sensing of Environment*, *118*, 83–94. https://doi.org/10.1016/j.rse.2011.10.028

Zhu, Z., & Woodcock, C. E. (2014). Continuous change detection and classification of land cover using all available Landsat data. *Remote Sensing of Environment*, *144*, 152–171. https://doi.org/10.1016/j.rse.2014.01.011

Zhu, Z., Woodcock, C. E., & Olofsson, P. (2012a). Continuous monitoring of forest disturbance using all available Landsat imagery. *Remote Sensing of Environment*, *122*, 75–91. https://doi.org/10.1016/j.rse.2011.10.030



Evoland
LAND MONITORING EVOLUTION

EvoLand Consortium

vito Remote Sensing

International Institute for Applied Systems Analysis
IIASA www.iiasa.ac.at

CLS
COLLECTE LOCALISATION SATELLITES

CESBIO

SINERGISE

JOANNEUM RESEARCH

cnes

DLR

GAFAG
an e-GEOS (ASI / Telespazio) Company

evenflow