EMBED
2SCALE

Ref. Ares(2025)6167348 - 29/07/2025

Earth Observation & Weather Data Federation with AI Embeddings

D4.5
SSL4EO-S12-Downstream Benchmark Dataset

Revision: v 1.2

D4.5 – SSL4EO-S12-downstream benchmark dataset

Work package	WP 4
Task	T4.1, T4.2, T2.2, T2.3
Due date	5/31/2025
Submission date	5/31//2025
Deliverable lead	FZJ
Version	1.2
Authors	Conrad M Albrecht (DLR)
Reviewers	Damien Roberts (UZH), Tim Reichelt (UOXF)
Abstract	In accordance with deliverable D2.3, we assembled an Earth observation benchmark dataset to develop, test, and rank neural compression methodologies. The set of downstream tasks curated include the real-world use cases that WP4 is dedicated to advance, namely: maritime awareness (SatCen), agriculture monitoring (FZJ), cloud and atmospheric physics (UOXF), and biomass estimation (UZH). Beyond we provide labelled scenes for urban heat island identification and land cover classification.
Keywords	Sentinel-1, Sentinel-2, satellite data, neural compression benchmarks, agriculture, forestry, cloud physics, ship detection, land cover, urban heat islands

Document Revision History

Version	Date	Description of change	List of contributor(s)
1.0	05/27/2025	Initial draft	Conrad M Albrecht (DLR)
1.1	07/07/2025	Incorporated reviewer feedback	Conrad M Albrecht (DLR)
1.2	07/28/2025	Fixed formatting and style	Gabriele Cavallaro (FZJ)

Funded by the European Union

Grant Agreement No: 101131841 | **Topic:** HORIZON-EUSPA-2022-SPACE-02-55
Call: HORIZON-EUSPA-2022-SPACE | **Type of action:** HORIZON-RIA

DISCLAIMER

Funded by the European Union

Project funded by

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

UK Research and Innovation

Embed2Scale (Earth Observation & Weather Data Federation With AI Embeddings) project is funded by the EU's Horizon Europe program under Grant Agreement number 101131841. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them. This work has also received funding from the Swiss State Secretariat for Education, Research and Innovation (SERI) and the UK Research and Innovation (UKRI).

COPYRIGHT NOTICE

Project funded by the European Commission in the Horizon Europe Programme		
Nature of the deliverable:	to specify R	
Dissemination Level		
PU	Public, fully open, e.g. web (Deliverables flagged as public will be automatically published in CORDIS project's page)	
SEN	Sensitive, limited under the conditions of the Grant Agreement	X
Classified R-UE/ EU-R	*EU RESTRICTED under the Commission Decision* *No2015/ 444*	
Classified C-UE/ EU-C	*EU CONFIDENTIAL under the Commission Decision* *No2015/ 444*	
Classified S-UE/ EU-S	*EU SECRET under the Commission Decision* *No2015/ 444*	

* *R: Document, report (excluding the periodic and final reports)*
DEM: Demonstrator, pilot, prototype, plan designs
DEC: Websites, patents filing, press & media actions, videos, etc.
DATA: Data sets, microdata, etc.
DMP: Data management plan
ETHICS: Deliverables related to ethics issues.
SECURITY: Deliverables related to security issues
OTHER: Software, technical diagram, algorithms, models, etc.

TABLE OF CONTENTS

Table of Contents

1 EXECUTIVE SUMMARY 5
2 INTRODUCTION: RATIONALE & OVERVIEW OF DOWNSTREAM TASKS 6
3 DATA: CHARACTERISTICS OF SSL4EO-S12-DOWNSTREAM 8
4 RESULTS: SSL4EO-S12-DOWNSTREAM SAMPLES 11
5 PUBLICATION: ACCESS TO SSL4EO-S12-DOWNSTREAM 13
6 REFERENCES 14

ABBREVIATIONS

FM	Foundation Model
GeoFM	Geospatial FM
ML	Machine Learning
LCZ	Local Climate Zone
USDA	United States Department of Agriculture
USGS	United States Geological Survey
NOAA	National Oceanic and Atmospheric Administration
NASA	National Aeronautics and Space Administration
ESA	European Space Agency
EEA	European Environmental Agency
SAR	Synthetic Aperture Radar
CC	Creative Commons
AIS	Automatic Identification System
LST	Land Surface Temperature
RGB	Red-Green-Blue (image)

1 EXECUTIVE SUMMARY

Embed2Scale aims to advance the state-of-the-art in neural data compression for Earth observation. To quantify the quality of compression, standardized and well-designed benchmarks are key. A diverse set of input data paired with annotations of real-world use cases constitutes the backbone for a mature set of downstream tasks. For those downstream tasks, Geospatial Foundation Models (GeoFMs) promise to generate embeddings serving a wide range of applications. Assembly of SSL4EO-S12-downstream combined with our efforts of an AI-compressor environment (D2.3) become the foundations of our multi-modal neural compression method developments for embedding fusion (T2.4).

2 INTRODUCTION: RATIONALE & OVERVIEW OF DOWNSTREAM TASKS

With automated vision pipelines, the concept of compression for machines shifts focus from human-perceptual fidelity to task-driven utility. Recent approaches broadly follow three categories:

1. end-to-end pipelines combine neural compressors (encoder-decoder) with convolutional task networks, jointly optimized via task-specific losses, achieving low bitrates without compromising downstream performance [1]
2. task-agnostic methods optimize image reconstruction with self-supervised consistency constraints, reducing data size while preserving downstream utility for tasks invariant to certain augmentations [2]
3. some approaches bypass reconstruction entirely, directly utilizing or optimizing compressed latent representations. References [3] train task networks directly on frozen pre-trained neural compressor latent, showing comparable performance to reconstruction-based methods. The work of [4] incorporates task-specific losses to create compressed latent optimized explicitly for particular tasks.

More specifically, we designed downstream tasks of relevance to real-world use cases guided by our consortium partners, cf. D4.2 (and references therein). In summary, these read:

- **Maritime Awareness** (*ships*): Ship detection is crucial for maritime security and safety, utilizing Synthetic Aperture Radar (SAR), optical imagery, and Automatic Identification System (AIS) data. We aim for efficiency by fusing geospatial and AIS data, refining ship monitoring with high-resolution imagery. Stakeholders like SatCen operational teams are involved in shaping the methodology and user scenarios.

- **Global Above Ground Biomass** (*biomass*): Global vegetation mapping is essential for understanding carbon cycles, human impact, and ecosystem services, requiring frequent and accurate measurements of canopy height (CH) and aboveground biomass (AGB). Traditional manual surveys are limited in scale, leading to advancements in remote sensing methods, such as satellite-based LiDAR and Sentinel imagery, which enable global vegetation analysis at high resolution. However, processing vast amounts of satellite data requires immense computational resources, prompting the need for efficient data transmission pipelines like Embed2Transfer and Embed2Infer. UZH is the key Embed2Scale consortium partner to utilize related SSL4EO-S12-downstream data.

- **Air Pollution & Cloud Physics** (*clouds*): This use case explores neural compression applications in climate science through two main strands: Strand 1 focuses on embedding models to support unsupervised cloud classification and improve storm-resolving climate models. Strand 2 aims to create a benchmark for climate data compression tailored to the needs of climate scientists (using climate datasets from ERA5, CMIP6, the Copernicus Atmosphere Monitoring Service (CAMS). The NextGEMS project (EU grant agreement No 101003470), and ESA), addressing gaps in existing neural compressor development. Collaborations of UOXF with stakeholders such as ECMWF and University of Helsinki (who are also members of ESiWACE3, EU grant agreement No 101093054) ensure relevance and community support for these advancements.

- **Agriculture Monitoring** (*crops*): The European Green Deal highlights the need for digital and green transitions to ensure climate-neutral and sustainable agriculture, as extreme weather increasingly affects crops. Sentinel-2 satellite data, combined with climate data and crop-type maps, enables large-scale monitoring of crop phenology and stress detection. AI-based tools like Embed2Transfer and Embed2Infer improve data accessibility and facilitate early identification of crop stress linked to adverse climatic conditions. This system benefits farmers, policymakers, agribusinesses, and environmental agencies by enhancing crop forecasting, risk management, and sustainability efforts.

Beyond such, we designed downstream tasks for the following use cases of relevance:

- **Urban Heat Island Detection** (*heatisland*): Climate impact requires our living spaces to adapt. In particular, large urban areas are prone to develop zones of high heat stress in summer. In fact, the classification of local climate zones (LCZ) aims at understanding the impact of natural vegetation (trees, lawns, etc.) and man-made structures like (low-, mid-, and high-rise) buildings onto the development of urban heat islands [5]. Studies show that densely built LCZs tend to exhibit higher land surface temperatures, while green spaces and water bodies contribute to cooling effects [6].

- **Land Cover Mapping** (*landcover*): The Horizon Europe project EvoLand [7] develops use cases to enable the next generation of Copernicus Land Monitoring Services [8]. Human impact and natural disasters constantly reshape the surface of our planet. Reliable and efficient mapping of change in land use and land cover benefits monitoring and managing our global resources. Accordingly, stakeholders such as the EC get enabled to develop regulations for governance.

- **Data Corruption** (*nodata*): The downloading and curation of overhead imagery may result in invalid pixels due to image tile boundaries, malfunctioned sensor elements, etc. In order to test a neural compressor's ability for such missing information, we generated an auxiliary downstream task to estimate the quality of an SSL4EO-S12 data cube stacking Sentinel-1 (GRD product) and Sentinel-2 (L1C and L2A products).

Funded by the European Union

3 DATA: CHARACTERISTICS OF SSL4EO-S12-DOWNSTREAM

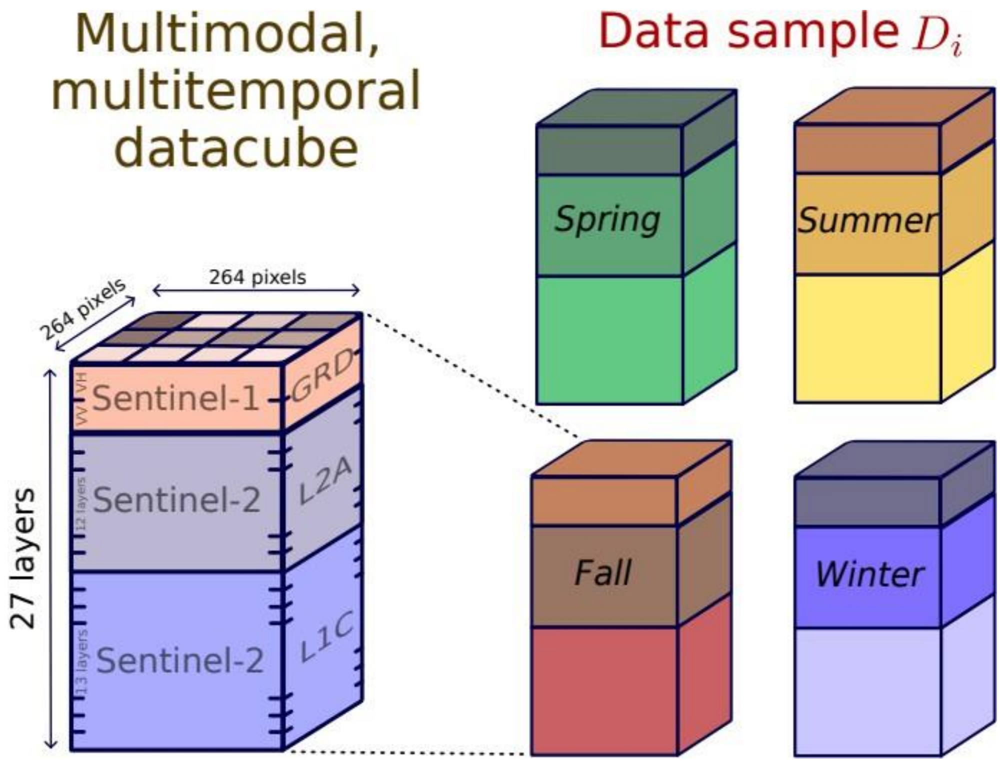


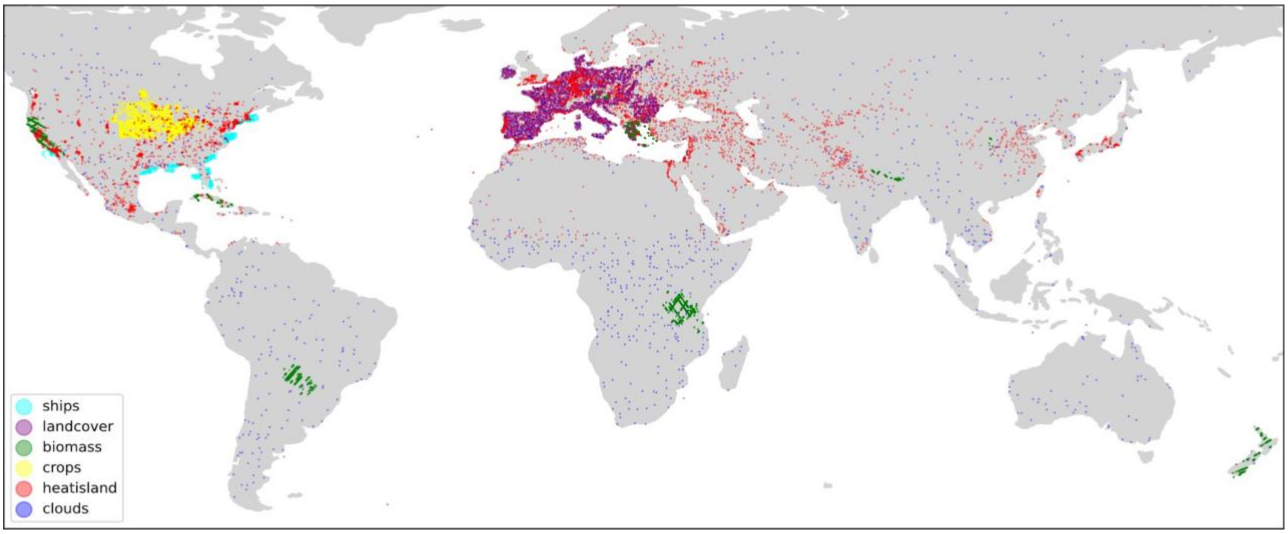
Figure 3.1: Structure of a multi-modal, multi-temporal SSL4EO-S12 data cube combining four seasons (Winter: December-February, Spring: March-May, Summer: June-August, Fall: September-November) of Sentinel-1 (GRD product, 2 channels) and Sentinel-2 (L1C and L2A products with 12+13=25 channels). The data ship in 264x264 pixels image patches. SSL4EO-S12-downstream combines data cubes of this format with labels from various real-world applications.

We provide several pre-processed, heterogeneous downstream tasks designed for continuous extension in the future. For the initial release we provide regression labels easily turned into classification tasks through thresholding.

As depicted in Fig. 3.1, SSL4EO-S12-downstream combines SSL4EO-S12 data cubes [9], incorporating 13 spectral channels from Sentinel-2 Level-1C Top-of-the-Atmosphere and 12 channels from Level-2A Surface Reflectance. In addition, these data are paired with two channels of Sentinel-1 VV and VH polarization. Four timestamps—one per season—are randomly picked to retrieve these spatial–spectral cubes over a given geolocation. Google Earth Engine [10] was used to download all relevant satellite data, including the downstream task labels, except for the ships and clouds use cases. All data were stored in cloud-ready ZIP-Store Zarr format [11]. Each downstream task contains from thousands to tens of thousands of samples globally distributed (Fig. 3.2). The associated satellite data have been pre-processed and filtered in the proper UTM projections without spatial overlap. Fig. 4.1 summarizes the downstream tasks in samples.

The **ships** downstream task concerns the location (regression) and detection of maritime vessels. Based on Automatic Identification System (AIS) ship-tracking data, we matched corresponding Sentinel-1 imagery and applied an additional fine-tuned location correction to account for temporal discrepancies between satellite data acquisitions and AIS signal transmissions. Since it is almost impossible for Sentinel-1 and Sentinel-2 to record the same area at the same time, we populated the Sentinel-2 bands of the corresponding SSL4EO-S12 data cubes with totally cloudy images available for the geolocation where the Sentinel-1 imagery was recorded.

The **landcover** task leverages aggregated land-use data from the European Environment Agency (EEA), which includes various land-cover classes such as forests, urban areas, water areas, and agricultural land. These labels represent the dominant land cover within each patch at a spatial resolution of 100 meters within Europe. Based on this data, we provide two downstream tasks: the fraction of forests and the fraction of agricultural land for the year 2018.



Dataset	Spatial Coverage	Temporal Coverage	Years Labelled	Number of Samples	Number of Downstream Tasks
ships	US Coast	2020 - 2024	2020 - 2024	3999	2
landcover	Europe	2018	2018	4691	2
biomass	global	2019	2019	2415	2
crops	US Corn Belt	2022	2022	3355	1
heatisland	Northern Hemisphere	2022	2021 - 2024	1659	2
clouds	global	2018 - 2020	2018 - 2020	1140	1
random	global	2018 - 2020	N/A	1140	1
nodata	**global**	**2018 - 2024**	**2018 - 2024**	**13260**	**1**

Figure 3.2: Global geospatial distribution and data characteristics of downstream tasks.

The **biomass** task employs above-ground biomass estimates derived from LIDAR measurements of the Global Ecosystem Dynamics Investigation (GEDI) instrument. GEDI provides structural information on vegetation height and density to estimate above-ground biomass in megagrams per hectare (Mg/ha). We exploit the GEDI Level 4A biomass estimates for as labels. Specifically, we aggregate all GEDI biomass pointwise estimates falling in the SSL4EO-S12 data cube of size 2640m x 2640m and compute the corresponding mean and standard deviation, which we use as image-level regression targets for SSL4EO-S12-downstream.

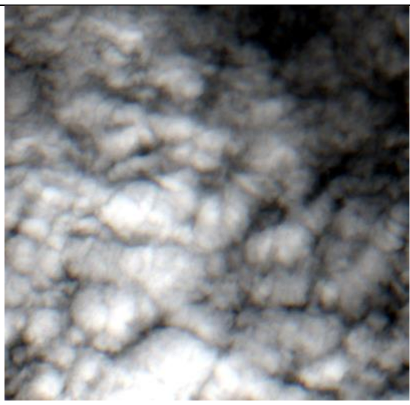
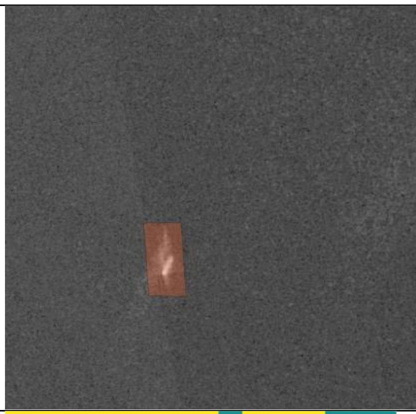
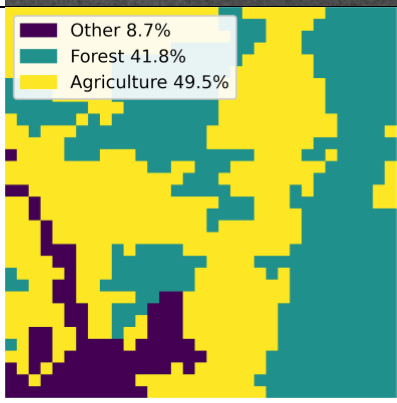
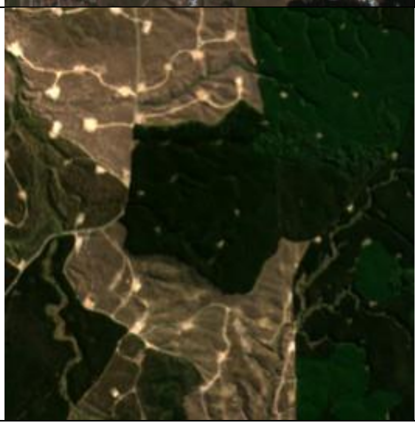
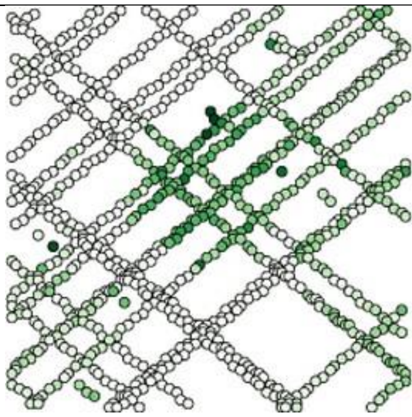
The **crops** task covers cropland in the US Corn Belt and is provided by the US Department of Agriculture (USDA). Soybean and corn we picked as classes for this downstream task, with the fraction of corn and soybeans the downstream task target. The Crop Data Layer (CDL) is published annually at 30 meters resolution. The labels are available as post-processed data, e.g., the label year 2023 corresponds to crops cultivated during the 2022 growing season.

For the **heatisland** use case, Landsat-8 Land Surface Temperature (LST) provides surface temperature data employed as labels for urban areas, band 10 (TIRS) in particular. Climate impact demands to restructure urban areas avoiding the development of urban heat islands. The corresponding downstream tasks address estimation of the mean surface temperature and its standard deviation per data cube stack in summer. We considered only cities with populations exceeding 20,000 inhabitants in the Northern Hemisphere with latitudes between 8 and 70 degrees North.

The **clouds** use case provides cloud cover fractions based on the CloudSen12+ dataset which is a mix of human annotated cloud labels on Sentinel-2 data and labels generated by a deep learning model that is trained on the human annotated dataset. Although the SAR data is not affected by clouds, Sentinel-1 is included alongside Sentinel-2 to stay consistent with the SSL4EO-S12 data structure across all downstream tasks. The artificial task random replaces the proper cloud labels with randomly picked numbers.

The **nodata** targets count invalid SSL4EO-S12 pixels that may occur due to boundary effects of the satellite image recording in swath mode. This downstream task serves as an auxiliary task to verify the stability of compression models regarding (random) input data artifacts.

4 RESULTS: SSL4EO-S12-DOWNSTREAM SAMPLES

Task	Sentinel-2 RGB image	Label	Description of downstream task
S H I P S			*Regression* task: location of ship centroid relative in image *Classification* task: vessel detection
L A N D C O V E R			*Regression* tasks: percentage of agriculture and forest pixels in scene, respectively
B I O M A S S			*Regression* tasks: biomass density (Mg/ha) for mean and standard deviation estimated on pixel-level labels derived from GEDI

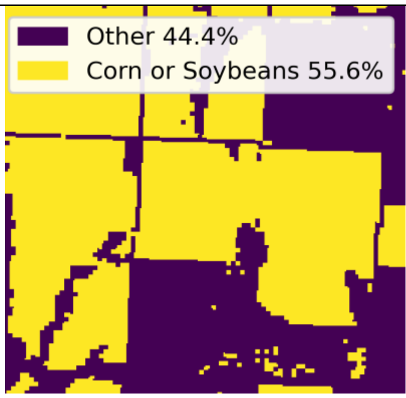
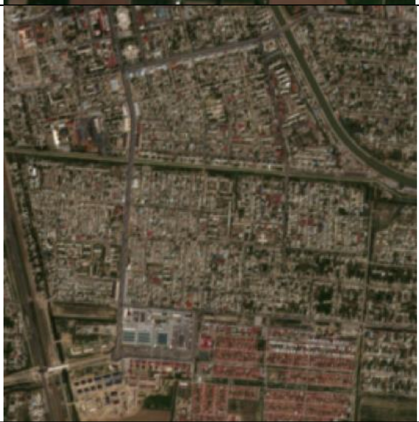
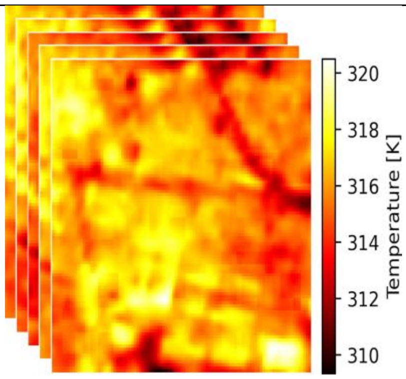
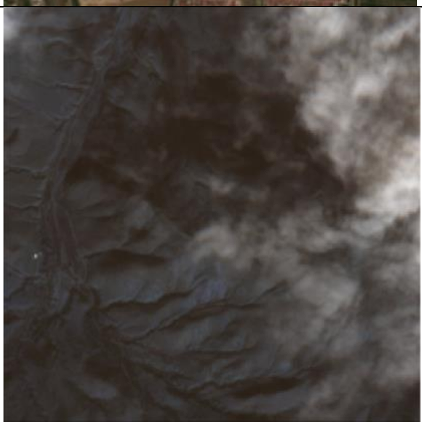
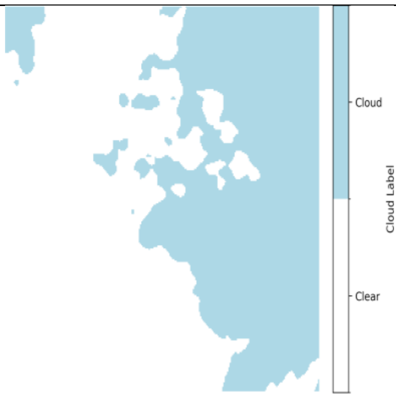
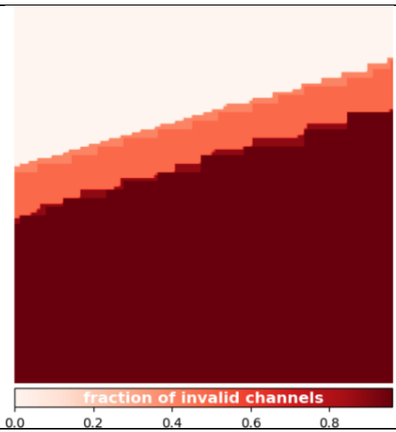
CROPS			*Regression* task: combined fraction of Soybean and Corn in scene
HEATISLAND			*Regression* tasks: summer surface temperature mean and standard deviation in Kelvin as observed by Landsat-8 thermal band
CLOUDS			*Regression task*: Average cloud cover fraction across four seasons of a year. note: the **random** regression task simply replaces the proper cloud labels by random numbers
NODATA			*Regression task:* Fraction of no-data pixels in SSL4EO-S12 data cube

Figure 4.1: Global geospatial distribution and data characteristics of downstream tasks

5 PUBLICATION: ACCESS TO SSL4EO-S12-DOWNSTREAM

We maintain public access to SSL4EO-S12-downstream through Embed2Scale's HuggingFace presence, https://huggingface.co/embed2scale. A European backup will be provided by FZJ through https://datapub.fz-juelich.de/ssl4eo-s12. Sentinel-1 and Sentinel-2 data get distributed under their respective Copernicus data license. Figure 5.1 summarizes the data label licenses:

Dataset	Origin of Data	License
Sentinel-1[1] & -2[2] [3]	ESA / Copernicus	CC BY-SA 3.0 IGO
Landsat-8[4]	USGS / NASA	Public Domain
CDL[5]	USDA NASS Cropland Data Layers	Public Domain
CORINE[6]	EEA / Copernicus Land Monitoring Service	Open and free access
CloudSen12+[7]	CloudSEN12 project	CC0 1.0
GEDI[8]	NASA	Public Domain
AIS[9]	NOAA US Coast Guard NAIS	Public Domain

Figure 5.1: Overview of data sources for the SSL4EO-S12-downstream data cubes and their respective licenses under which the data provider (Origin of Data) published them.

Given the above licenses we publish

- SSL4EO-S12-downstream: https://huggingface.co/datasets/embed2scale/SSL4EO-S12-downstream
- SSL4EO-S12-v1.1: https://huggingface.co/datasets/embed2scale/SSL4EO-S12-v1.1

under Creative Commons license CC-BY-4.0: https://creativecommons.org/licenses/by/4.0

[1] https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S1_GRD

[2] https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S2_HARMONIZED

[3] https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S2_SR_HARMONIZED

[4] https://developers.google.com/earth-engine/datasets/catalog/LANDSAT_LC08_C02_T1_L2

[5] https://developers.google.com/earth-engine/datasets/catalog/USDA_NASS_CDL

[6] https://land.copernicus.eu/pan-european/corine-land-cover/clc2018

[7] https://cloudsen12.github.io

[8] https://developers.google.com/earth-engine/datasets/catalog/LARSE_GEDI_GEDI04_A_002

[9] https://coast.noaa.gov/digitalcoast/tools/ais.html

6 REFERENCES

[1] http://doi.org/10.1109/TCSVT.2022.3195322, https://doi.org/10.1109/ICASSP39728.2021.9414465, https://doi.org/10.1109/OJCAS.2021.3126061, https://arxiv.org/abs/2011.06409, https://arxiv.org/abs/2103.04178, https://arxiv.org/abs/2111.02249

[2] https://arxiv.org/abs/2106.10800

[3] https://doi.org/10.1109/TCSVT.2023.3240391, https://arxiv.org/abs/1803.06131

[4] https://arxiv.org/abs/2007.11797

[5] https://doi.org/10.3390/rs15040884

[6] https://doi.org/10.1007/s10661-025-13705-3

[7] https://www.evo-land.eu

[8] https://land.copernicus.eu

[9] https://doi.org/10.1109/MGRS.2023.3281651, https://arxiv.org/abs/2503.00168

[10] https://earthengine.google.com

[11] https://zarr.readthedocs.io/en/stable/user-guide/storage.html

Funded by the European Union