AGILE: GIScience Series, 7, 13, 2026. https://doi.org/10.5194/agile-giss-7-13-2026
Proceedings of the 29th AGILE Conference on Geographic Information Science, 16–19 June 2026.
Eds.: Evelyn Uuemaa, Alexander Kmoch, Holger Virro, and Simon Scheider
This contribution underwent peer review based on a full paper submission.

Generalizability of Foundation Models: A Case Study on Cocoa Mapping Across Countries Using Sparse Labels

Ruslan Mammadov [1,2], Paul Walther [3], Julius Fricke [1], and Martin Werner [2,3]

[1]osapiens Terra GmbH, Mannheim, Germany
[2]Department of Computer Science, TUM School of Computation, Information and Technology, Technical University of Munich, Munich, Germany
[3]Department of Aerospace and Geodesy, TUM School of Engineering and Design, Technical University of Munich, Munich, Germany

Correspondence: Ruslan Mammadov (ruslan.mammadov@tum.de), Paul Walther (paul.walther@tum.de)

Abstract. Remote Sensing Foundation Models (RSFMs), which are deep neural networks pre-trained on large-scale Earth observation datasets, have become increasingly popular in remote sensing in recent years. Meanwhile, regulatory changes such as the European Union's Deforestation Regulation require a mapping of agricultural activities worldwide to ensure deforestation-free supply chains. In this context, we conduct a case study on the application of Foundation Models (FMs), such as CROMA and AlphaEarth, for the mapping of cocoa production in agroforestry systems in western Africa. In our study we show, that the pre-training of FMs does not improve the performance in data-rich conditions and that the pre-training only has limited advantage in zero-shot transfer applications. Still, the FMs show a higher sensitivity towards distribution shifts when fine-tuned in new environments. Based on our experiments, we comprehend insights for the selection and application of traditional convolutional neural network-based models and FMs in sparse-label remote sensing tasks.

Submission Type. Case study, analysis.

BoK Concepts. [IP3-5] Image Segmentation, [IP3-2] Computer vision in EO, [IP6] Image processing (value) chain

Keywords. Foundation Models, Cross-Country Generalization, Sparse Labels, Cocoa Mapping

1 Introduction

Remote sensing is nowadays primarily governed by deep learning solutions, which enable large-scale and automated analysis of the environment with improved accuracy and scalability compared to traditional methods. This includes mapping land use and vegetation, as well as the detection and evaluation of environmental changes. Recently, Remote Sensing Foundation Models (RSFMs), which are comparably large deep learning models pre-trained on vast Earth observation datasets, were proposed to ease the application of advanced deep learning solutions without the need for expensive and complex task-specific training. In this context, a variety of models, e.g., CROMA (Fuller et al., 2023a), AlphaEarth (Brown et al., 2025), Prithvi (Jakubik et al., 2023), SeCo (Yao et al., 2021a), and SatMAE (Cong et al., 2022) have been proposed and demonstrate competitive results compared to specialized models in diverse applications. They, especially, claim advantages and strong predictive performance, even in sparse label contexts, but so far this advantage was not extensively demonstrated.

This feature poses these models ideal for use in mapping tasks in environments where structured data collection, e.g., by government agencies, is not conducted. Particularly in the context of tracking global supply chains, this can enable the collection of information about the sustainability of production even without detailed on-site monitoring, which is necessary to follow new regulatory guidelines such as the EU Deforestation Regulation (European Parliament and Council of the European Union, 2023) and its application (European Parliament and Council of the European Union, 2025).

One of the products influenced by this regulation is cocoa. As such, imported cocoa is required not to have caused any deforestation after 2020 (European Parliament and Council of the European Union, 2023), which necessitates a detailed mapping of cocoa production facilities using high-resolution, up-to-date cocoa maps. This is a challenging task as cocoa plantations are not systematically mapped in many countries and cocoa is grown in agroforestry, which resembles an open-canopy

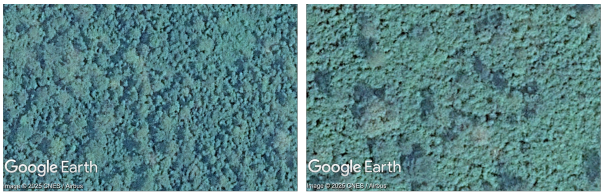

Figure 1. Google Earth imagery of forest (left) and cocoa (right), derived from CNES/Airbus data and visualized in Google Earth Pro (Google LLC, 2025a).

forest from above and is challenging to detect even in high-resolution imagery from commercial satellites (compare Figure 1) (Numbisi et al., 2019). Freely available satellite data, such as from Sentinel-2, providing only coarser resolution, make these tasks even more challenging. Therefore, the propositions for automatic cocoa detection from remote sensing imagery are mostly limited to small traditional models such as Random Forests applied to local regions in which labels were collected manually (Ashiagbor et al., 2020; Numbisi et al., 2019; Kanmegne et al., 2022; Koralewicz et al., 2025). The only known study for a deep-learning based nationwide mapping of cocoa plantations was proposed in 2023 by Kalischek et al. (2023a), who trained a CNN on 100,000 GPS-mapped cocoa samples from Ghana and Côte d'Ivoire. So far, it has not been investigated how these results generalize to other cocoa-producing countries without structured, labelled datasets.

In this study, we therefore want to examine how the mapping of cocoa can be improved in data-sparse regions by using FMs and suitable training strategies. This includes

- a compilation of a comprehensive training, validation, and test data set of cocoa agriculture in western Africa, namely the countries Côte d'Ivoire, Ghana, Nigeria, and Cameroon.
- the proposition of training strategies for different baseline CNN models and FMs for the cross-country generalization of the complex cocoa mapping task.
- the evaluation of the proposed models and the deduction of guidelines for the use of FMs for the application in mapping tasks with sparse labels.

2 Related Work

In the following, we first give a detailed overview of existing proposals for cocoa mapping from earth observation imagery before we dive into approaches to model generalizations and the resulting publication of RSFMs.

2.1 Cocoa Mapping

Cocoa mapping, as we approach it in this paper, is a specific segmentation of remote sensing imagery. In general, this image segmentation is complex as underlying image data is various in space and time and often unlabelled: Remote sensing images have a changing distribution due to atmospheric effects, climates, seasons, vegetational specifities or illumination, amongs others (Xie et al., 2023; Tong et al., 2025), and freely available data from Sentinel or Landsat has rather low resolution (European Space Agency, 2025a, b; The National Aeronautics and Space Administration, 2025). This makes labelling complex as the interpretation of the remote sensing imagery often requires an expert's knowledge. Therefore, standard baseline labelled data sets are only available to a limited extent. Further, the granularity of the tasks varies a lot, e.g., the detection of water bodies like floodings is much coarser than a mapping of different crop types in finely structured farmland.

These general issues are also apparent in cocoa mapping. While studies attempted to use Radar and Landsat data for remote sensing-based land use segmentation as early as 2001, they explain that a distinction of crop systems was not possible and only deforestation could be detected (Saatchi et al., 2001; Gerold and Lanfer, 2001). With the increasing availability of fine-grained earth observation data through the Sentinel-1 and -2 and Landsat 8 and 9 missions, more studies have been conducted for the explicit mapping of cocoa (Ashiagbor et al., 2020; Numbisi et al., 2019; Abu et al., 2021). Still, their study area is mostly limited to small regions; only four papers provide a nationwide or multinational approach for cocoa mapping (Moraiti et al., 2024; Kalischek et al., 2023a; Abu et al., 2021; Condro et al., 2020).

Main issue for most of the approaches is, that training data is only available to a limited extent as labels for cocoa plantations are mostly locally collected and not widely published. This also disqualifies complex models, as there is no reasonable amount of labelled data available to train those. Therefore, the most popular models are Random Forest approaches (Moraiti et al., 2024; Abu et al., 2021; Condro et al., 2020). Only some approaches use more complex models like convolutional neural networks (Kalischek et al., 2023a) and a U-Net (Therias et al., 2025). Another approach to improve the performance of cocoa segmentation is to input not only single images but instead time series of images (Batista et al., 2022).

2.2 Generalization in Geospatial Models

One idea to solve the limited available data for specific tasks such as the cocoa mapping is to pre-train models on different tasks, which learn general concepts to help solving the specific application with only limited amounts of labelled samples (Xiao et al., 2025). In detail, neural networks are not initialized randomly

anymore but instead inherit weights from a more general training, which does not necessarily resemble the final task. For example, models for remote sensing are pre-trained with unspecified image datasets, which allow to learn generic representations and propose a better generalization after a fine-training on the actual task, as the models have already seen more concepts than available in the limited fine-tuning dataset. Techniques used to improve such a generalization are methods for domain adaptation, adversarial learning, pseudo label generation, reconstruction, ensemble based methods (Lyu et al., 2025; Khelif et al., 2024; Gackstetter et al., 2025) and active learning (Desai and Ghose, 2022).

2.3 Foundation Models

The combined application of these pre-training methods results in the proposal of Foundation Models (FMs). These are available for the general image domain (Kirillov et al., 2023; Redmon et al., 2015) but also specified for remote sensing (Fuller et al., 2023a; Brown et al., 2025; Jakubik et al., 2023; Yao et al., 2021a; Cong et al., 2022). They are usually trained using self-supervised learning (Huo et al., 2025) to obtain generic, but meaningful representations of earth observation images, which can later be fine-tuned for specific downstream tasks without the requirements for large amounts of labelled data (Xiao et al., 2025). The self-supervised training paradigms can be classified into reconstruction-based learning, contrastive (or discriminative) learning, and self-distillation learning. *Reconstruction-based learning* masks parts of the image and tries to reconstruct those, e.g., masked autoencoders in SatMAE and Prithvi (Cong et al., 2023; Jakubik et al., 2023), *contrastive learning* tries to produce similar embeddings for similar samples, e.g., in the SeCo model (Yao et al., 2021b), and *self-distillation without labels* is used, e.g., in DINO (Caron et al., 2021; Oquab et al., 2024). Additionally, there were foundation models proposed which combine the basic training principles, e.g., CROMA (Fuller et al., 2023b), which combines reconstruction-based and contrastive learning for training, and AlphaEarth (Brown et al., 2025), which additionally utilizes metadata for learning.

To compare such foundation models, earth observation benchmarks have been introduced (Lacoste et al., 2023; Marsocci et al., 2025; Wang et al., 2025). These indicate that foundation models do not generally outperform models trained for a specific task. Additionally, none of the proposed benchmarks allows conclusions about cross-country generalizability of the foundation models. For the further study, CROMA and AlphaEarth are chosen as representative models, as they show strong performance and allow a multi-modal approach using various sensors of the earth observation imagery (Brown et al., 2025; Fuller et al., 2023b; Marsocci et al., 2025).

3 Methods

This study evaluates how different pre-training strategies affect the transfer of cocoa-mapping models across countries. An overview of the introduced methodology is given in Figure 2.

3.1 Model Architectures and Pre-Training

We consider three pre-training paradigms: (i) none, represented by a baseline CNN trained from scratch; (ii) foundational, unsupervised pre-training, represented by CROMA and AlphaEarth; and (iii) task-specific pre-training, represented by the globally trained canopy-height model GCHM. For each model family, we evaluate combinations of source-region training and fine-tuning on sparsely labeled Nigerian data.

3.1.1 Baseline CNN

The baseline model follows the encoder architecture proposed by Lang et al. (2023) and adapted by Kalischek et al. (2023a) for cocoa mapping. It begins with a pointwise block of three 1×1 convolutions with 128, 256, and 728 filters, each followed by batch normalization and ReLU, and containing a residual skip connection. This is followed by eight depthwise separable convolutional blocks, each consisting of a 3×3 depthwise separable convolution, a 1×1 pointwise convolution, ReLU, and a residual connection. All layers use 728 filters. A final 3×3 convolution produces logits for cocoa and background. In general, the CNN approach follows the design by Kalischek et al. (2023a) and hyperparameters were systematically chosen from a predefined set of potential options.

3.1.2 CROMA Foundation Model

For CROMA, we use the base optical encoder trained by Fuller et al. (2023b). The encoder is a Vision Transformer (ViT) that splits each Sentinel-2 image into 8×8 patches and projects them into a 768-dimensional space.

Patch embeddings are bilinearly upsampled to the native resolution, concatenated with the original Sentinel-2 bands, and fed to a lightweight decoder. The decoder mirrors the baseline CNN design but is reduced to two separable convolutional blocks (728 filters) followed by a 3×3 output layer with two channels.

We evaluate two regimes: (i) frozen encoder, where only the decoder is trained, and (ii) unfrozen encoder, where encoder and decoder are jointly fine-tuned. Source-region validation indicates that a learning rate of 10^{-4} for the decoder and 10^{-5} for the encoder works best; this configuration is used throughout.

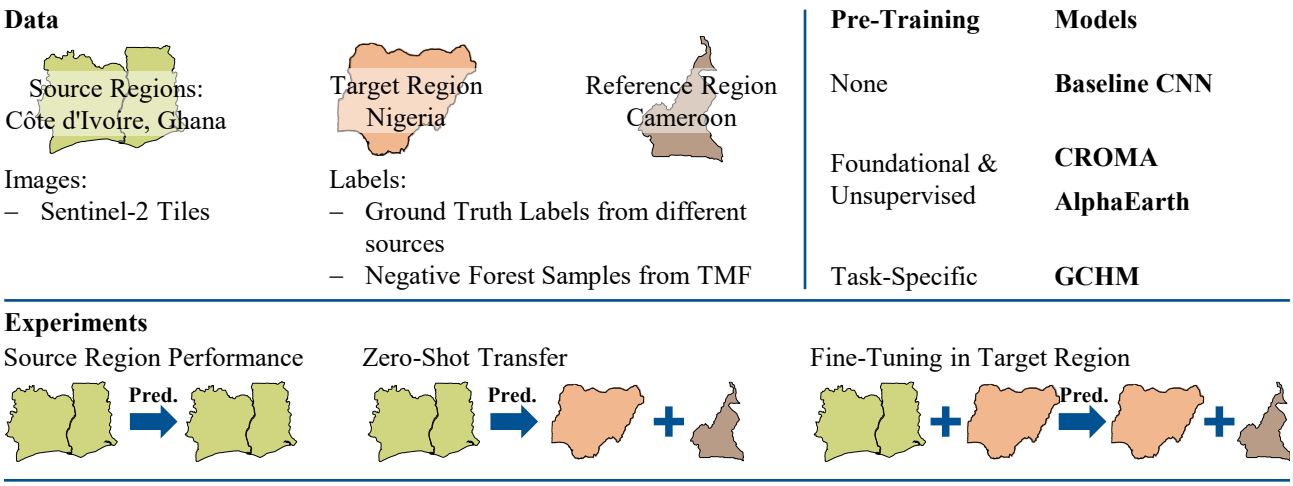

Figure 2. An overview of the applied methodologies for cross-country generalization of cocoa mapping including information on the source, target, and reference data regions, applied models with different pre-training regimes, namely, Baseline Convolutional Neural Network (CNN), CROMA, AlphaEarth and global canopy-height network (GCHM), conducted experiments, and evaluation procedures.

3.1.3 AlphaEarth Foundation Model

AlphaEarth, proposed by Brown et al. (2025), is used in an embedding-based configuration. Only annual, precomputed embeddings are available; the underlying model weights are not. The embeddings are derived from yearly multi-temporal imagery and diverse data sources (e.g. optical, radar, elevation) and map each location into a 64-dimensional feature space. As such, the model is not directly comparable to the other architectures, which operate only on Sentinel-2 imagery, but it was included in the study due to its high practical relevance as a ready-to-use, multi-modal foundation representation.

All AlphaEarth experiments operate on these fixed embeddings, available via Google Earth Engine. The decoder is identical to that of CROMA (two separable convolutional blocks plus a 3×3 output layer), and only the decoder is trained. AlphaEarth thus tests how far generic compressed representations can be exploited for cocoa mapping without updating the foundation model itself.

3.1.4 Task-specific Canopy Height Model (GCHM)

The task-specific model is the global canopy-height network GCHM trained by Lang et al. (2023) to regress canopy height from Sentinel-2 imagery using GEDI lidar data. Architecturally, GCHM mirrors the baseline CNN but with fewer filters (256 instead of 728) and an additional skip connection from the output of the initial pointwise block to the prediction head.

To keep inputs consistent, only Sentinel-2 bands are used; auxiliary inputs such as positional embeddings are removed. The regression output layer is replaced by a randomly initialized 3×3 convolution with two channels, reusing the pre-trained feature extractor.

An overview of models and their decoders, where applicable, is illustrated in Figure 3.

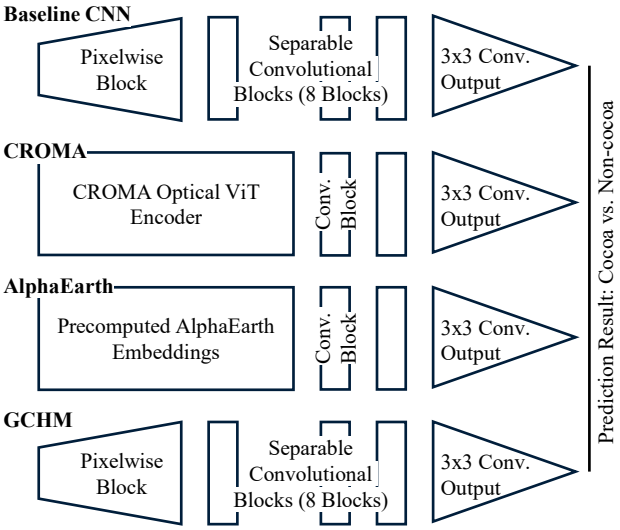

Figure 3. Overview of the model architectures and decoders used in this study. The GCHM is drawn with smaller blocks to indicate that its architecture parallels the baseline CNN but uses fewer parameters.

3.2 Dataset

The used dataset consists of two parts: the satellite imagery and corresponding labels for the source regions in Côte d'Ivoire and Ghana, the target region in Nigeria, and the reference region in Cameroon (compare Figure 4).

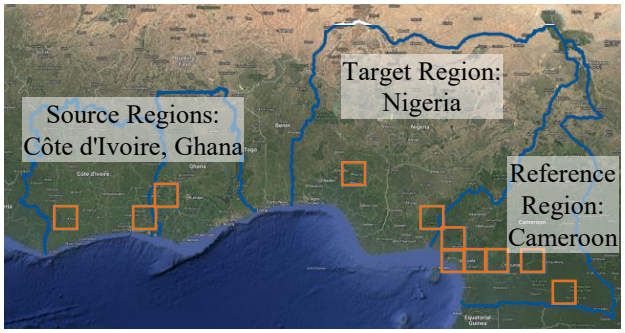

Figure 4. Study regions and the corresponding Sentinel-2 test tiles (orange). Basemap from Google Earth (Google LLC, 2025b); nation borders (blue) from Natural Earth (Natural Earth, 2025). Visualized in QGIS (QGIS Development Team, 2025).

3.2.1 Satellite Imagery and Normalization

Almost all models used in this study directly on Sentinel-2 multispectral Level-2A surface reflectance images. Imagery is accessed via the AWS Earth Search STAC API. For a given area of interest and time interval, all matching STAC items are retrieved. Duplicates with identical acquisition time are removed by keeping the product with the highest `s2:processing_baseline`. Remaining scenes are sorted by cloud cover (`eo:cloud_cover`) and selected from least to most cloudy until each pixel has at least n observations (typically $n = 5$).

For each selected scene, bands B01–B08, B8A, B09, B11, B12 and the Scene Classification Layer (SCL) are downloaded. Bands at 20 m and 60 m resolution are bilinearly upsampled to 10 m and stacked into a 12-band cube. The SCL is used to mask invalid pixels (no data, saturated/defective pixels, shadows, and clouds). The bands were selected based on the experience of other cocoa mapping approaches, such as Kalischek et al. (2023a).

Normalization is model-specific. For the baseline CNN, band-wise mean and standard deviation are computed from the source-region training tiles and used for standardization. For GCHM, we adopt the normalization statistics from the original pre-trained model. For CROMA, reflectances are clipped to $(\mu - 2\sigma, \mu + 2\sigma)$, where μ represent band-wise means and σ standard deviation, using global statistics estimated from random Sentinel-2 samples, and then rescaled to $[0, 1]$. AlphaEarth embeddings are used as provided.

3.2.2 Labels

For the source region (Côte d'Ivoire & Ghana), the labels for training and validation areas are obtained from the probability maps published by Kalischek et al. (Kalischek et al., 2023a, b). For Nigeria and Cameroon, as well as all test areas, including the test sets in Côte d'Ivoire and Ghana, higher-quality labels were extracted from various data sources. An overview of the used label sources are provided in Table 1.

3.2.3 Negative Forest Labels from TMF

In Nigeria, the most informative negative samples for cocoa mapping are dense forests with canopies similar to cocoa agroforestry. To obtain such samples at scale, we use the Tropical Moist Forest (TMF) disturbance dataset by Vancutsem et al. (2021), which provides annual maps of undisturbed forest derived from Landsat time series.

A naive strategy, labeling all undisturbed pixels as forest, was rejected after intersecting TMF with farm and cocoa datasets, which showed substantial overlap between agroforestry and TMF undisturbed-forest pixels: specifically, 2.0% of all and 16.3% of cocoa farms from eHealth Africa (2023), and 47.8% of cocoa farms from Lescuyer (2024). This outcome is consistent with the assumption that agroforestry systems experience localized disturbances (e.g. tree removal, small structures, rotational crops); therefore, agroforestry plantations can contain a mixture of disturbed and undisturbed pixels, whereas genuine natural forest remains undisturbed over larger spatial extents.

Rather than discarding TMF, we therefore derive more conservative forest samples that enforce this neighbourhood assumption. The 2024 undisturbed-forest raster is vectorized, very small polygons (below the 1st percentile of area) are discarded, and remaining polygons are subdivided until all patches are below ≤ 0.2 ha. We then retain only polygons without interior holes and with four vertices, i.e. compact rectangular patches in which both the pixels and their immediate surroundings are consistently labeled as undisturbed. This filtering removes roughly two thirds of all candidates, and a repeated intersection with the aforementioned datasets showed no remaining overlaps. The resulting set of more than 15 000 polygons in Nigeria is used as hard negative forest samples.

3.3 Training on the Source Region

Source-region training uses cocoa probability maps from Kalischek et al. (2023a) as pseudo-labels and thus distills the information. Pixels with probability $> 90\%$ are labeled as cocoa, those $< 10\%$ as non-cocoa; intermediate probabilities are treated as background and excluded. These decision boundaries are consistent to the approach by (Kalischek et al., 2023a).

All Sentinel-2 tiles intersecting Côte d'Ivoire and Ghana are identified. Three tiles with independent reference data are reserved for testing. The remaining tiles are randomly split into training and validation subsets, as shown in Figure 5, resolving overlaps by prioritizing test over validation and validation over training.

Table 1. An overview of obtained test data labels.

	Region	Sample Class	No. of Samples	Source
Source	Côte d'Ivoire	Cocoa	196	Osapiens Terra – curated
		Negative	371	Osapiens Terra – manual annotation
	Ghana	Cocoa	1772	Osapiens Terra – curated
			107	Ajagun et al. (2022)
		Negative	317	Osapiens Terra – manual annotation
Target	Nigeria	Cocoa	825	eHealth Africa (2023, 2025); The Trustees of Columbia University in the City of New York (2025) – curated
		Negative	2428	Osapiens Terra – manual annotation
			43	Manual Annotation
			24	Descals et al. (2024)
			15,513	Vancutsem et al. (2021) – curated
Ref.	Cameroon	Cocoa	223	Lescuyer (2024)
		Negative	123	Osapiens Terra – manual annotation

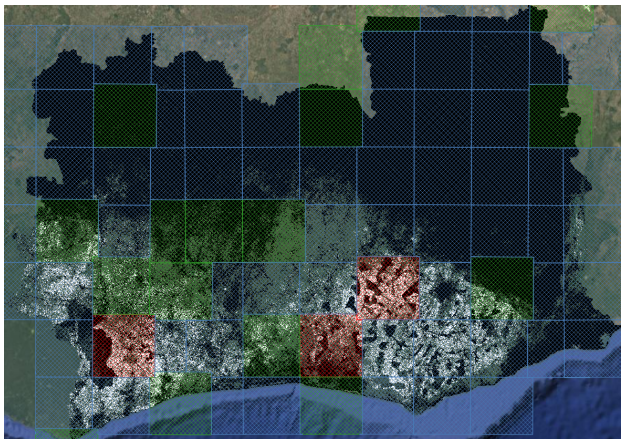

Figure 5. Train (blue), validation (green), and test (red) split for the source region, overlaid with pseudo-labels derived from Kalischek et al. (2023a) cocoa probability maps (white: >90% probability; black: <10% probability). Background imagery is from Google Earth (Google LLC (2025b)). Visualized in QGIS (QGIS Development Team (2025)).

The training protocol similar to the protocol used by Kalischek et al. (2023a) was used. For each tile, Sentinel-2 imagery is collected for two six-month periods per year (January–June and July–December) from 2019–2021, matching the temporal coverage of the probability map, with at least five usable scenes per pixel and period. Each scene and its label mask are divided into 32×32 patches. Patches with more than 90% background are discarded. For each remaining patch, a random scene is selected among those with at least 90% cloud-free area. This yields roughly 3.6 M training and 1.1 M validation patches.

Each epoch, 100,000 training patches are drawn at random; a fixed validation subset of 100,000 patches is reused across epochs. We minimize a dice loss, ignoring background pixels. Optimization uses Adam with weight decay 10^{-4}. The default learning rate is 10^{-4}; for unfrozen CROMA encoders we use 10^{-5}. A milestone scheduler reduces the learning rate by a factor of 10 at epochs 60 and 90. The checkpoint with the best validation F1 score is retained. AlphaEarth follows the same protocol using one annual embedding per year instead of multiple scenes.

3.4 Fine-Tuning on the Target Region

Fine-tuning uses Nigerian annotations after removing all samples intersecting the held-out test tiles. Each remaining sample is buffered by 320 m so that a 32×32 patch can be drawn around it with some flexibility in patch location. Samples are assigned to years based on collection date from the dataset metadata. Sentinel-2 imagery is then downloaded for the given year, ensuring at least five scenes per pixel.

During fine-tuning, 80% of samples are used for training and 20% for validation. For each training sample and epoch, a random 32×32 box containing the labeled area is drawn. Pixels within the box but outside the labeled footprint are treated as background. A random Sentinel-2 scene is selected from the corresponding cluster and year, and a cloud mask is derived from the SCL band. Image–mask pairs are rejected if fewer than 50% of pixels in the patch are cloud-free or if the labeled region is fully cloud-covered.

Fine-tuning uses the same optimizer and weight decay as source-region training but employs the sample-averaged cross-entropy loss, described in 3.5. A 10-epoch linear warm-up increases the learning rate from 0 to its nominal value. AlphaEarth uses the same sampling logic but draws from annual embeddings rather than individual scenes.

3.5 Evaluation, Loss, and Area-Normalized Metrics

Evaluation is performed on Sentinel-2 imagery, except for AlphaEarth model where the precomputed embeddings are used. For each test tile, we download enough scenes

Table 2. Overview of pre-training and transfer configurations evaluated in this study.

Pre-training Type	Trained on Source Region	Fine-tuned on Target Region	Representative Models
None	Yes / No	Yes / No	Baseline CNN
Foundation (unsupervised)	Yes / No	Yes / No	CROMA / AlphaEarth
Task-Specific (canopy height)	Yes / No	Yes / No	GCHM

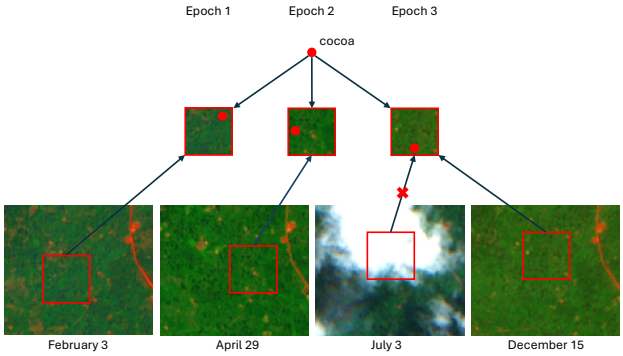

Figure 6. Illustration of the fine-tuning data sampling procedure.

to obtain at least five observations per pixel. The area of interest is split into overlapping 32×32 patches with an 8-pixel border, so that the central 16×16 regions overlap by 16 pixels in both directions. For each patch location, we select the scene with the largest cloud-free area; ties are broken in favor of the globally least cloudy scene. Predictions are computed for all patches, but only the central 16×16 pixels are retained and then stitched into a full-resolution map. Cloud-covered pixels in the selected scene are set to no data and excluded from metrics.

Because annotation sizes range from single points to large polygons, standard pixelwise losses and metrics would be dominated by large objects. We therefore use a sample-normalized formulation. For the loss, cross-entropy is first computed per pixel. For each sample s, the mean over valid pixels yields a sample loss $\mathcal{L}_s$; the batch loss is the average of $\mathcal{L}_s$ over the batch, so each sample contributes equally.

For evaluation, per-sample confusion counts (i.e., TP, FP, TN, and FN) are computed and divided by the number of valid pixels in each sample, yielding normalized counts that sum to 1 per sample. Global counts are obtained by summing these normalized values across samples, after which precision, recall, and F1-score are computed using the standard formulas. In this formulation, each sample contributes equally to the global metrics, while each pixel's contribution is inversely proportional to the number of valid pixels in its sample.

3.6 Data and Software Availability

The data is available from the corresponding author upon reasonable request. Data obtained by external authors and entities, specifically test data for Côte d'Ivoire and Ghana, and some negative classes for Nigeria and Cameroon, can only be shared with the explicit consent of the parties that provided the data.

The code is publicly available on GitHub at https://github.com/osapiens-Terra-GmbH/2026-AGILE-CocoaMapping.git.

4 Experimental Setup

The experimental space is defined by three factors: (i) pre-training type, (ii) training on the source region, and (iii) fine-tuning on the target region, as shown in Table 2.

For each model we evaluate three regimes. First, models are trained on the source region (Côte d'Ivoire and Ghana) using pseudo-labels and then evaluated on the source, target (Nigeria), and reference (Cameroon) regions in a zero-shot setting. Second, models are trained or fine-tuned only on sparse Nigerian labels, representing a country-specific, label-scarce baseline. Third, models are first trained on the source region and subsequently fine-tuned on Nigeria, combining dense pseudo-labels with limited in-country supervision.

5 Results

This section presents the empirical results for all model configurations. Unless stated otherwise, all metrics are area-normalized F1, precision, and recall as defined in Section 3.5.

5.1 Training on Source Region

The first set of experiments evaluates all models on held-out test tiles in Côte d'Ivoire and Ghana, providing a baseline under data-rich conditions.

In Côte d'Ivoire, the highest F1-scores were obtained by the baseline CNN and AlphaEarth (both 86.7%), followed by GCHM (85.3%) and CROMA (84.5%). The frozen-encoder CROMA variant performed worst (80.5%; Table 3).

On the Ghana test set the ranking is similar, but all models appear close to saturation: CROMA slightly outperforms the baseline CNN and AlphaEarth (96.4% vs. 96.0% F1), with GCHM and frozen CROMA trailing only marginally (Table 4).

Table 3. Performance on the Côte d'Ivoire test set.

Model	F1 (%)	Rec. (%)	Prec. (%)
Baseline CNN	**86.7**	86.2	87.3
CROMA	84.5	83.4	85.6
CROMA (frozen enc.)	80.5	78.3	82.8
AlphaEarth	**86.7**	88.1	85.2
GCHM	85.3	84.9	85.7

Table 4. Performance on the Ghana test set.

Model	F1 (%)	Rec. (%)	Prec. (%)
Baseline CNN	**96.0**	95.3	99.9
CROMA	**96.4**	96.7	99.4
CROMA (frozen enc.)	95.3	96.5	98.8
AlphaEarth	**96.0**	95.3	99.8
GCHM	95.8	96.1	99.8

Validation curves based on 100,000 samples (Figure 7) show a consistent ordering: the baseline CNN achieves the highest validation F1 (96.3%), followed by AlphaEarth (95.7%), GCHM (94.6%), CROMA (93.8%), and frozen CROMA (90%). The baseline CNN starts with the lowest F1 but surpasses all pre-trained models after about 60 epochs.

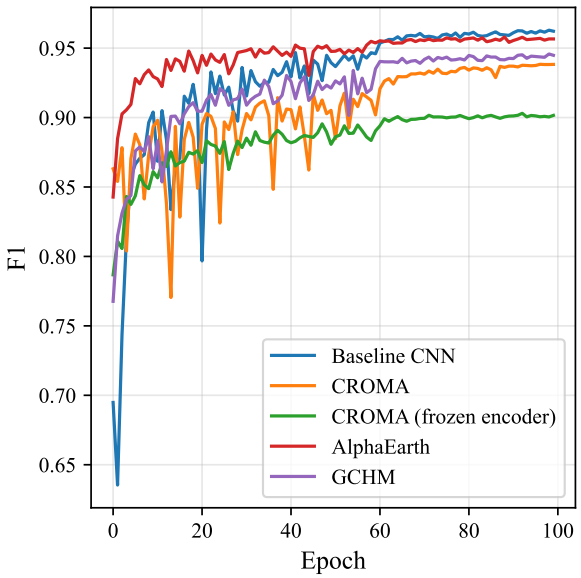

Figure 7. F1 validation scores for the main models on the source region per training epoch. The baseline CNN starts with the lowest F1 but surpasses all other models after roughly the 60th epoch.

To disentangle architecture and pre-training, GCHM was also trained from random initialization. In Côte d'Ivoire, the randomly initialized model slightly exceeds the pre-trained version (85.9% vs. 85.3% F1); in Ghana the scores are nearly identical (95.7% vs. 95.8%; Table 5). On validation, the randomly initialized GCHM reaches a higher best F1 (95.5% vs. 94.6%) and overtakes the pre-trained variant around epoch 60 (Figure 8).

Table 5. Comparison of pre-trained and randomly initialized GCHM models.

Model Variant	F1 (%)	Dataset
GCHM (pre-trained)	85.3	Côte d'Ivoire
GCHM (random init.)	**85.9**	Côte d'Ivoire
GCHM (pre-trained)	**95.8**	Ghana
GCHM (random init.)	95.7	Ghana
GCHM (pre-trained)	94.6	Best Val.
GCHM (random init.)	**95.5**	Best Val.

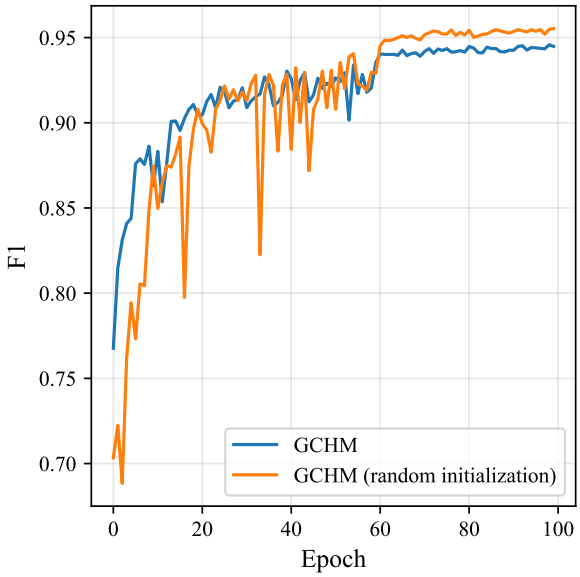

Figure 8. F1 validation scores on the source-region validation set for the GCHM model with pre-trained and randomly initialized weights per training epoch. The pre-trained model performs better in early epochs, but the randomly initialized model surpasses it at around epoch 60.

The results indicate that for models that operate on the same input (Sentinel-2), pre-training does not improve performance in this data-rich setting; the best results are obtained by the CNN trained from scratch, with GCHM close behind.

5.2 Zero-Shot Transfer to Nigeria and Cameroon

The source-trained models were next evaluated in a zero-shot setting on Nigeria and Cameroon (Tables 6 and 7). In Nigeria, the best F1 reached only 61.0% (baseline CNN), with recall at 43.9%. In Cameroon, all models remained below 50% F1 and 35% recall. This shows a substantial loss of performance when applying models trained in Côte d'Ivoire and Ghana directly to new countries, although they are still in western Africa.

Table 6. Zero-shot performance on the Nigeria test set.

Model	F1 (%)	Rec. (%)	Prec. (%)
Baseline CNN	**61.0**	43.9	99.9
CROMA	56.3	39.3	99.4
CROMA (frozen enc.)	37.7	23.3	98.8
AlphaEarth	5.8	3.0	99.8
GCHM	51.5	34.7	99.8

Table 7. Zero-shot performance on the Cameroon test set.

Model	F1 (%)	Rec. (%)	Prec. (%)
Baseline CNN	39.1	24.3	99.1
CROMA	**50.0**	33.3	99.6
CROMA (frozen enc.)	37.7	23.4	96.0
AlphaEarth	7.0	3.6	100.0
GCHM	36.7	22.5	99.5

Pre-training does not improve zero-shot transfer. In Nigeria, the baseline CNN achieves the highest F1 and recall, outperforming CROMA, GCHM, and especially AlphaEarth, whose F1 drops to 5.8%. In Cameroon, CROMA performs best, but the baseline CNN remains competitive.

A common pattern across all models is extremely high precision (typically $\geq 99\%$) combined with strongly reduced recall. This reflects the heavy class imbalance: *models learn conservative decision boundaries in the source region and carry this behaviour over to new countries, preferring to avoid false positives at the cost of missing many true cocoa samples. This is expected, as cocoa is generally not the dominant land cover; as such, unknown patterns are more likely to be predicted as non-cocoa.*

5.3 Fine-Tuning on Nigeria with Sparse Labels

To assess adaptation under sparse supervision, all source-trained models were fine-tuned on Nigerian annotations. First we compare the results against the models trained on Nigeria from scratch (Section 5.3.1), then test if the source-training has any affect by comparing the models against models not trained on source (Section 5.3.2), and finally compare all pre-training strategies to obtain the optimal strategy for cocoa segmentation (Section 5.3.3).

5.3.1 Fine-Tuning with Source-Region Training

As a baseline, a CNN and GCHM were also trained from scratch directly on Nigeria using the same fine-tuning protocol. Results are reported in Table 8.

Fine-tuning substantially improves performance for all pre-trained models. The strongest results are obtained by AlphaEarth and CROMA, followed by frozen CROMA, GCHM, and the source-trained CNN. In contrast, models

Table 8. Performance on the Nigeria test set after fine-tuning with sparse labels.

Model	F1 (%)	Rec. (%)	Prec. (%)
Random Initalization			
Baseline CNN	46.8	30.1	96.8
GCHM	72.7	57.7	98.2
Trained on source			
Baseline CNN	88.4	79.7	99.1
CROMA	**95.1**	92.4	98.0
CROMA (frozen enc.)	90.6	84.0	98.2
AlphaEarth	**95.7**	95.1	96.3
GCHM	89.1	80.8	99.4

trained only on Nigeria perform considerably worse: the randomly initialized Nigeria-only CNN reaches 46.8% F1, and the GCHM 72.7% F1, despite having identical architectures to their pre-trained counterparts.

Training and validation curves (Figure 9) show that models without pre-training quickly reach high training F1 but plateau at 60–75% validation F1, indicating overfitting and poor generalization under sparse labels. Pre-trained models attain much higher validation F1 while having similar training scores. AlphaEarth, operating on annual multi-temporal embeddings, converges particularly quickly and stably.

Overall, pre-training is highly beneficial in the label-scarce Nigerian setting: it enables much better fine-tuned performance and clearly improves sample efficiency.

5.3.2 Effect of Source-Region Training for Pre-Trained Models

The fine-tuning results in Nigeria (Section 5.3) show large gains for all pre-trained models compared to models trained from scratch. These gains could in principle come from two different mechanisms: (i) models may have learned cocoa-specific, source-region representations that transfer across countries, or (ii) they may simply have benefited from generic feature enrichment that makes fine-tuning under sparse labels easier, without strong cross-country generalization by itself.

To separate these effects, all pre-trained models were fine-tuned again on Nigeria, this time *without* an intermediate source-region training stage. The results are summarized in Table 9.

Across all three pre-trained families, the impact of including source-region training before fine-tuning is small and inconsistent in direction. These results indicate that, for models already endowed with strong pre-trained representations, additional supervised training on the source region does *not* systematically improve final performance in Nigeria.

The dominant factor behind their success after fine-tuning is therefore the quality of their generic features, not cross-

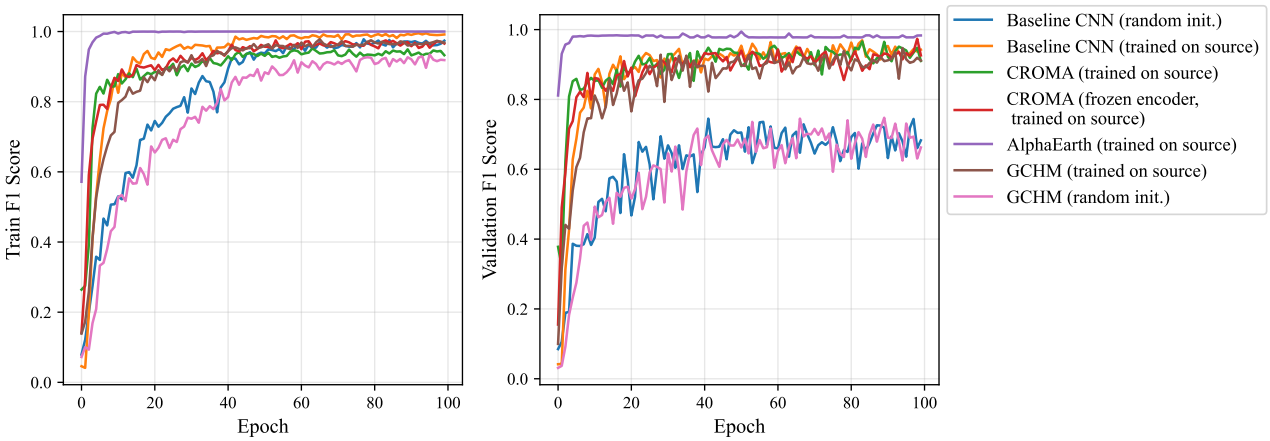

Figure 9. Training (left) and validation (right) F1 scores during fine-tuning per training epoch. All models quickly reach high training F1, but models trained from scratch (baseline CNN and GCHM with random initialization) plateau at 60–75% validation F1, indicating poor generalization under sparse supervision.

Table 9. Comparison of fine-tuning results with and without source-region training.

Model	Source Train.	F1 (%)	Rec. (%)	Prec. (%)
CROMA	Yes	95.1	92.4	98.0
CROMA	No	95.0	92.1	98.0
CROMA (frozen enc.)	Yes	90.6	84.0	98.2
CROMA (frozen enc.)	No	90.6	84.3	98.0
AlphaEarth	Yes	95.7	95.1	96.3
AlphaEarth	No	92.5	88.1	97.5
GCHM	Yes	89.1	80.8	99.4
GCHM	No	91.3	84.8	98.8

Table 10. Comparative analysis across all models.

Model	Source Train.	F1 (%)	Rec. (%)	Prec. (%)
Foundational pre-training w and w/o source training				
AlphaEarth	Yes	**95.7**	95.1	96.3
AlphaEarth	No	92.5	88.1	97.5
CROMA	Yes	**95.1**	92.4	98.0
CROMA	No	**95.0**	92.1	98.0
CROMA (frozen enc.)	Yes	90.6	84.0	98.2
CROMA (frozen enc.)	No	90.6	84.3	98.0
Task-specific pre-training w and w/o source training				
GCHM	Yes	89.1	80.8	99.4
GCHM	No	91.3	84.8	98.8
Source-region training only				
GCHM	Yes	89.7	81.6	99.7
Baseline CNN	Yes	88.4	79.7	99.1
Only fine-tuning (random initialization)				
GCHM	No	72.7	57.7	98.2
Baseline CNN	No	46.8	30.1	96.8
CROMA	No	0–18.3	0–100	0–10.1

country transfer of source-region cocoa signatures. This contrasts with models without pre-training: as shown earlier, the baseline CNN benefit strongly from source-region supervision, which effectively acts as a task-specific pre-training stage.

5.3.3 Comparison of Pre-Training Strategies

The different pre-training regimes can now be compared directly using the Nigeria fine-tuning results, including models with and without source-region training and models trained from scratch. Table 10 summarizes all configurations.

For models without any prior pre-training, source-region training has a large impact. The CNN and GCHM trained from random initialization on Côte d'Ivoire and Ghana reach 88.4% and 89.7% F1 after fine-tuning in Nigeria, compared to 46.8% and 72.7% F1 respectively when trained only on Nigerian labels. In this regime, source-region supervision acts as a classical task-specific pre-training step and clearly improves fine-tuning results.

For GCHM, the summary in Table 10 shows that global task-specific pre-training (91.3% F1) and source-region pre-training (89.7% F1 for random init. + source training) lead to very similar performance once the model is fine-tuned on Nigeria. The differences are small and mostly trade-offs between recall and precision, indicating that both strategies provide comparable benefits.

Among all model families, the foundational models CROMA and AlphaEarth reach the highest fine-tuned F1-scores (around 95–96%), with AlphaEarth slightly ahead

when source-region training is included. However, their advantage over the best task-specific and source-only pre-trained models is modest, on the order of a few percentage points. *The frozen-encoder CROMA and GCHM occupy an intermediate range, reinforcing the picture that* any *reasonable form of pre-training, foundational, task-specific, or source-region, confers substantial benefits over training from scratch under sparse labels.*

5.4 Ablation and Other Experiments

Ablation experiments with CROMA quantify the effect of key design choices in the fine-tuning pipeline. CROMA was selected as it showed strong results in the previous experiments and, unlike AlphaEarth, where only annual embeddings are available, enables using various scenes for the same year.

Removing TMF-derived forest samples from training and validation caused precision to drop from 98.0% to 51.5% and F1 from 95.0% to 66.7%, showing that these hard forest negatives are crucial for learning a robust decision boundary (compare Table 11).

Table 11. Effect of TMF-derived forest samples on CROMA fine-tuning performance in Nigeria.

Model Variant	F1 (%)	Rec. (%)	Prec. (%)
CROMA (without TMF)	66.7	96.2	51.5
CROMA (with TMF)	**95.0**	92.1	98.0

Choosing a random Sentinel-2 scene per epoch rather than always the least cloudy scene improved F1 from 90.6% to 95.0%, indicating that temporal resampling acts as effective augmentation (compare Table 12).

Table 12. Effect of scene selection strategy during fine-tuning.

Model Variant	F1 (%)	Rec. (%)	Prec. (%)
CROMA (best scene)	90.6	85.4	96.5
CROMA (random scene)	**95.0**	92.1	98.0

Randomizing the bounding box around each annotation across epochs improved F1 from 89.1% to 95.0%, highlighting the benefit of varying spatial context (compare Table 13).

Table 13. Effect of bounding-box sampling strategy during fine-tuning.

Model Variant	F1 (%)	Rec. (%)	Prec. (%)
CROMA (same b. box)	89.1	82.4	97.0
CROMA (different b. boxes)	**95.0**	92.1	98.0

Using multiple scenes at evaluation time yields a small additional gain: averaging predictions from five scenes increases F1 from 95.0% to 96.1% (compare Table 14).

Table 14. Effect of multi-scene aggregation during evaluation.

Evaluation Strategy	F1 (%)	Rec. (%)	Prec. (%)
CROMA (1 scene)	95.0	92.1	98.0
CROMA (5 scenes)	**96.1**	93.5	99.0

These ablations underline that, beyond pre-training itself, hard negative samples, temporal and spatial data augmentation, and scene selection all play a critical role in achieving robust performance in the target country.

6 Conclusion

This study examined how different pre-training strategies affect cross-country cocoa mapping. Three main conclusions emerge.

First, in data-rich conditions pre-training did not improve performance. On Côte d'Ivoire and Ghana, the baseline CNN trained from scratch matched or exceeded foundational and task-specific pre-trained models, indicating that medium-resolution EO tasks can be solved effectively without pre-training when abundant labels are available.

Second, zero-shot transfer across countries remains difficult for all model families. When models trained in Côte d'Ivoire and Ghana were applied directly to Nigeria and Cameroon, performance degraded sharply, and pre-training on large global datasets did not yield better cross-country generalization. Globally precomputed embeddings, such as those from AlphaEarth, were particularly sensitive to distribution shifts.

Third, pre-training became crucial once labels in the target country were sparse, but gains seem to arise from generic feature enrichment rather than cross-country robustness. Fine-tuning on limited Nigerian annotations yielded large gains for all pre-trained models, while architectures trained only on Nigeria performed substantially worse. However, pre-trained models reached almost the same fine-tuned performance with or without additional supervised training on the source region, and chaining multiple pre-training stages did not produce consistent improvements. Overall, the main benefit of pre-training in this setting lies in generic feature enrichment rather than inherent geographic robustness.

Overall, geospatial models do not seem to generalize across regions. However, they show strong adaptation ability even with limited samples once there's any type of pre-training.

7 Limitations and Future Work

While this study provides empirical evidence on the role of pre-training in cross-country generalization for cocoa

mapping, several important limitations define the scope of the study and suggest directions for future research.

First, the study focuses on a single use case and geographic triad. The experiments are restricted to cocoa mapping across one source-target-reference configuration (Côte d'Ivoire and Ghana, Nigeria, and Cameroon). Robust statements about generalization require replication across multiple settings. The conclusions drawn here therefore represent one step towards a broader evidence base.

Second, this work evaluates only one dimension of generalization: cross-country transfer. The analysis considers spatial transfer across national and ecological boundaries, but other forms of generalization remain unexplored, such as temporal generalization, cross-sensor transfer, robustness under atypical phenological or climatic conditions, etc.

Third, the study is constrained by the available reference data and study's scope. Although sufficient for comparative evaluation, the labeled datasets offer limited thematic diversity, particularly for non-cocoa agroforestry systems such as cashew, coffee, citrus, and mixed-tree orchards. For forest specifically, negative samples were obtained through the TMF-derived filtering procedure described in Section 3.2.3. For models to learn meaningful decision boundaries against all relevant negative classes, they should be also exposed to diverse non-cocoa agroforestry systems, even if only a small number of samples is available for each class. Future work aiming at nation-wide cocoa mapping should therefore prioritise the collection of diverse, high-quality negative samples and combine them with the transfer-learning strategies demonstrated here to produce sample-efficient maps with clear separation between cocoa, forest, and other tree-based land uses.

Declaration of Generative AI in writing

The authors declare that they have used Generative AI tools in the preparation of this manuscript. Specifically, the AI tools were utilized for language editing and improving grammar and sentence structure but not for generating scientific content, research data, or substantive conclusions. All intellectual and creative work, including the analysis and interpretation of data, is original and has been conducted by the authors without AI assistance.

Acknowledgements

This paper's results were partly obtained in a Master's Thesis by Ruslan Mammadov, unpublished, but submitted at the Technical University of Munich. The work is partly funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 507196470.

In addition, we are grateful to the researchers and organizations who kindly shared datasets or granted permission for their use. In particular, we would like to thank Ebunoluwa Ajagun and Guillaume Lescuyer; the openness and willingness to share data exemplify the collaborative spirit that advances geospatial research.

References

Abu, I.-O., Szantoi, Z., Brink, A., Robuchon, M., and Thiel, M.: Detecting cocoa plantations in Côte d'Ivoire and Ghana and their implications on protected areas, Ecological Indicators, 129, 107 863, https://doi.org/10.1016/j.ecolind.2021.107863, 2021.

Ajagun, E. O., Ashiagbor, G., Asante, W. A., Gyampoh, B. A., Obirikorang, K. A., and Acheampong, E.: Cocoa eats the food: expansion of cocoa into food croplands in the Juabeso District, Ghana, Food Security, 14, 451–470, https://doi.org/10.1007/s12571-021-01227-y, 2022.

Ashiagbor, G., Forkuo, E. K., Asante, W. A., Acheampong, E., Quaye-Ballard, J. A., Boamah, P., Mohammed, Y., and Foli, E.: Pixel-based and object-oriented approaches in segregating cocoa from forest in the Juabeso-Bia landscape of Ghana, Remote Sensing Applications: Society and Environment, 19, 100 349, https://doi.org/10.1016/j.rsase.2020.100349, 2020.

Batista, J. E., Rodrigues, N. M., Cabral, A. I. R., Vasconcelos, M. J. P., Venturieri, A., Silva, L. G. T., and Silva, S.: Optical time series for the separation of land cover types with similar spectral signatures: cocoa agroforest and forest, International Journal of Remote Sensing, 43, 3298–3319, https://doi.org/10.1080/01431161.2022.2089540, 2022.

Brown, C. F., Kazmierski, M. R., Pasquarella, V. J., Rucklidge, W. J., Samsikova, M., Zhang, C., Shelhamer, E., Lahera, E., Wiles, O., Ilyushchenko, S., Gorelick, N., Zhang, L. L., Alj, S., Schechter, E., Askay, S., Guinan, O., Moore, R., Boukouvalas, A., and Kohli, P.: AlphaEarth Foundations: An embedding field model for accurate and efficient global mapping from sparse label data, https://arxiv.org/abs/2507.22291, 2025.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A.: Emerging Properties in Self-Supervised Vision Transformers, https://arxiv.org/abs/2104.14294, 2021.

Condro, A. A., Setiawan, Y., Prasetyo, L. B., Pramulya, R., and Siahaan, L.: Retrieving the National Main Commodity Maps in Indonesia Based on High-Resolution Remotely Sensed Data Using Cloud Computing Platform, Land, 9, https://doi.org/10.3390/land9100377, 2020.

Cong, Y., Khanna, S., Meng, C., Liu, P., Rozi, E., He, Y., Burke, M., Lobell, D., and Ermon, S.: SatMAE: Pre-training Transformers for Temporal and Multi-Spectral Satellite Imagery, in: Advances in Neural Information Processing Systems, edited by Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., vol. 35, pp. 197–211, Curran Associates, Inc, https://proceedings.neurips.cc/paper_files/paper/2022/file/01c561df365429f33fcd7a7faa44c985-Paper-Conference.pdf, 2022.

Cong, Y., Khanna, S., Meng, C., Liu, P., Rozi, E., He, Y., Burke, M., Lobell, D. B., and Ermon, S.: SatMAE:

Pre-training Transformers for Temporal and Multi-Spectral Satellite Imagery, https://arxiv.org/abs/2207.08051, 2023.

Desai, S. and Ghose, D.: Active Learning for Improved Semi-Supervised Semantic Segmentation in Satellite Images, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 553–563, 2022.

Descals, A., Gaveau, D. L. A., Wich, S., Szantoi, Z., and Meijaard, E.: Global mapping of oil palm planting year from 1990 to 2021, Earth System Science Data, 16, 5111–5129, https://doi.org/10.5194/essd-16-5111-2024, 2024.

eHealth Africa: Nigerian Farmlands Dataset, https://data.grid3.org/datasets/3c2d9dfe05c246878e6635bf34cc03c0_0/, dataset curated by eHealth Africa and hosted on the GRID3 Data Hub (https://data.grid3.org/). Dataset license (custom) can be found on the publishing website, 2023.

eHealth Africa: About eHealth Africa, https://ehealthafrica.org/, non-profit organization founded in 2009 to build data-driven solutions for healthcare delivery across Africa. Accessed 2025-11-10., 2025.

European Parliament and Council of the European Union: Regulation (EU) 2023/1115 of the European Parliament and of the Council of 31 May 2023 on the making available on the Union market and the export from the Union of certain commodities and products associated with deforestation and forest degradation and repealing Regulation (EU) No 995/2010, Official Journal of the European Union, http://data.europa.eu/eli/reg/2023/1115/oj, accessed on 14.10.2025, 2023.

European Parliament and Council of the European Union: Regulation (EU) 2025/2650 of the European Parliament and of the Council of 19 December 2025 amending Regulation (EU) 2023/1115 as regards certain obligations of operators and traders, Official Journal of the European Union, http://data.europa.eu/eli/reg/2025/2650/oj, accessed on 02.01.2026, 2025.

European Space Agency: Sentinel-1, https://www.esa.int/Applications/Observing_the_Earth/Copernicus/Sentinel-1, accessed: 2025-10-31, 2025a.

European Space Agency: Sentinel-2, https://www.esa.int/Applications/Observing_the_Earth/Copernicus/Sentinel-2, accessed: 2025-10-26, 2025b.

Fuller, A., Millard, K., and Green, J.: CROMA: Remote Sensing Representations with Contrastive Radar-Optical Masked Autoencoders, in: Advances in Neural Information Processing Systems, edited by Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., vol. 36, pp. 5506–5538, Curran Associates, Inc, https://proceedings.neurips.cc/paper_files/paper/2023/file/11822e84689e631615199db3b75cd0e4-Paper-Conference.pdf, 2023a.

Fuller, A., Millard, K., and Green, J. R.: CROMA: Remote Sensing Representations with Contrastive Radar-Optical Masked Autoencoders, https://arxiv.org/abs/2311.00566, 2023b.

Gackstetter, D., Yu, K., and Körner, M.: Self-attention and frequency-augmentation for unsupervised domain adaptation in satellite image-based time series classification, ISPRS Journal of Photogrammetry and Remote Sensing, 224, 113–132, https://doi.org/10.1016/j.isprsjprs.2025.03.024, 2025.

Gerold, G. and Lanfer, N.: Agrarkolonisation und Bodennutzungsprobleme im Oriente Ecuadors: Agricultural Settlement and Land Utilization Problems in the Oriente of Ecuador, Erdkunde, 55, 362–378, https://doi.org/10.3112/erdkunde.2001.04.04, 2001.

Google LLC: Google Earth Pro, https://www.google.com/earth/about/versions/, computer software, 2025a.

Google LLC: Google Maps Satellite Imagery, https://www.google.com/maps, map tiles retrieved via Google Maps API (mt1.google.com), 2025b.

Huo, C., Chen, K., Zhang, S., Wang, Z., Yan, H., Shen, J., Hong, Y., Qi, G., Fang, H., and Wang, Z.: When Remote Sensing Meets Foundation Model: A Survey and Beyond, Remote Sensing, 17, https://doi.org/10.3390/rs17020179, 2025.

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., Ankur, K., Ramasubramanian, M., Gurung, I., Khallaghi, S., Hanxi, Li, Cecil, M., Ahmadi, M., Kordi, F., Alemohammad, H., Maskey, M., Ganti, R., Weldemariam, K., and Ramachandran, R.: Foundation Models for Generalist Geospatial Artificial Intelligence, https://arxiv.org/abs/2310.18660, 2023.

Kalischek, N., Lang, N., Renier, C., Daudt, R., Addoah, T., Thompson, W., Blaser-Hart, W., Garrett, R., and Wegner, J.: Cocoa plantations are associated with deforestation in Côte d'Ivoire and Ghana, Nature Food, 4, 384–393, https://doi.org/10.1038/s43016-023-00751-8, 2023a.

Kalischek, N., Lang, N., Renier, C., Daudt, R., Addoah, T., Thompson, W., Blaser-Hart, W., Garrett, R., and Wegner, J.: Global Cocoa Probability Map (ETH Zurich Research Collection), https://www.research-collection.ethz.ch/entities/researchdata/fa059526-e934-4fff-80b9-7d212827b76a, released under a Creative Commons Attribution (CC BY 4.0) license. Accessed 2026-11-10., 2023b.

Kanmegne, D., Latifi, H., Ullmann, T., Baumhauer, R., Thiel, M., and Bayala, J.: Modelling the spatial distribution of the classification error of remote sensing data in cocoa agroforestry systems, Agroforestry Systems, 97, 1–11, https://doi.org/10.1007/s10457-022-00791-2, 2022.

Khelif, M. K., Boulila, W., Koubaa, A., and Farah, I. R.: Domain Adaptation for Satellite Images: Recent Advancements, Challenges, and Future Perspectives, Procedia Computer Science, 246, 413–422, https://doi.org/10.1016/j.procs.2024.09.420, 28th International Conference on Knowledge Based and Intelligent information and Engineering Systems (KES 2024), 2024.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., and Girshick, R.: Segment Anything, pp. 4015–4026, 2023.

Koralewicz, A., Vlcek, J., Oliveras, I., Hirons, M., Akinyugha, A., Olowoyo, O., Ajayi-Ebenezer, M., and Owen, O.: Mapping the extent and exploring the drivers of cocoa agroforestry in Nigeria, insights into trends for climate

change adaptation, Agroforestry Systems, 99, 1–20, https://doi.org/10.1007/s10457-024-01126-z, 2025.

Lacoste, A., Lehmann, N., Rodriguez, P., Sherwin, E. D., Kerner, H., Lütjens, B., Irvin, J. A., Dao, D., Alemohammad, H., Drouin, A., Gunturkun, M., Huang, G., Vazquez, D., Newman, D., Bengio, Y., Ermon, S., and Zhu, X. X.: GEO-Bench: Toward Foundation Models for Earth Monitoring, https://arxiv.org/abs/2306.03831, 2023.

Lang, N., Jetz, W., Schindler, K., and Wegner, J. D.: A high-resolution canopy height model of the Earth, Nature Ecology & Evolution, pp. 1–12, 2023.

Lescuyer, G.: Inventaire des arbres dans 223 cacaoyères agroforestières au Cameroun, https://doi.org/10.18167/DVN1/MGDIJU, 2024.

Lyu, S., Zhao, Q., Zhou, Z., Li, M., Zhou, Y., Yao, D., Cheng, G., Zhou, H., and Shi, Z.: Deep Learning Based Domain Adaptation Methods in Remote Sensing: A Comprehensive Survey, https://arxiv.org/abs/2510.15615, 2025.

Marsocci, V., Jia, Y., Bellier, G. L., Kerekes, D., Zeng, L., Hafner, S., Gerard, S., Brune, E., Yadav, R., Shibli, A., Fang, H., Ban, Y., Vergauwen, M., Audebert, N., and Nascetti, A.: PANGAEA: A Global and Inclusive Benchmark for Geospatial Foundation Models, https://arxiv.org/abs/2412.04204, 2025.

Moraiti, N., Mullissa, A., Rahn, E., Sassen, M., and Reiche, J.: Critical Assessment of Cocoa Classification with Limited Reference Data: A Study in Côte d'Ivoire and Ghana Using Sentinel-2 and Random Forest Model, Remote Sensing, 16, https://doi.org/10.3390/rs16030598, 2024.

Natural Earth: Boundary Lines (1:10m scale), https://www.naturalearthdata.com/downloads/10m-cultural-vectors/10m-admin-0-boundary-lines/, made with Natural Earth. Free vector and raster map data @ naturalearthdata.com. Accessed 2025-11-10, 2025.

Numbisi, F. N., Van Coillie, F., and Wulf, R.: Delineation of Cocoa Agroforests Using Multiseason Sentinel-1 SAR Images: A Low Grey Level Range Reduces Uncertainties in GLCM Texture-Based Mapping, ISPRS International Journal of Geo-Information, 8, 179, https://doi.org/10.3390/ijgi8040179, 2019.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., and Bojanowski, P.: DINOv2: Learning Robust Visual Features without Supervision, https://arxiv.org/abs/2304.07193, 2024.

QGIS Development Team: QGIS Geographic Information System, QGIS Association, https://www.qgis.org, 2025.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A.: You Only Look Once: Unified, Real-Time Object Detection, https://arxiv.org/pdf/1506.02640, 2015.

Saatchi, S., Agosti, D., Alger, K., Delabie, J., and Musinsky, J.: Examining fragmentation and loss of primary forest in the southern Bahian Atlantic forest of Brazil with radar imagery, Conservation Biology, 15, 867–875, https://doi.org/10.1046/j.1523-1739.2001.015004867.x, 2001.

The National Aeronautics and Space Administration: Satellites, https://landsat.gsfc.nasa.gov/satellites, accessed: 2025-10-31, 2025.

The Trustees of Columbia University in the City of New York: Geo-Referenced Infrastructure and Demographic Data for Development (GRID3) Data Portal, https://data.grid3.org/, gRID3 Partnership. Accessed 2025-11-10, 2025.

Therias, A., Rafiee, A., Lhermitte, S., van der Lugt, P., and Lindenbergh, R.: Integrating radar and multi-spectral data to detect cocoa crops: a deep learning approach, Remote Sensing Applications: Society and Environment, 39, 101 652, https://doi.org/10.1016/j.rsase.2025.101652, 2025.

Tong, X.-Y., Dong, R., and Zhu, X. X.: Global high categorical resolution land cover mapping via weak supervision, ISPRS Journal of Photogrammetry and Remote Sensing, 220, 535–549, https://doi.org/10.1016/j.isprsjprs.2024.12.017, 2025.

Vancutsem, C., Achard, F., Pekel, J.-F., Vieilledent, G., Carboni, S., Simonetti, D., Gallego, J., Aragão, L. E. O. C., and Nasi, R.: Long-term (1990–2019) monitoring of forest cover changes in the humid tropics, Science Advances, 7, eabe1603, https://doi.org/10.1126/sciadv.abe1603, 2021.

Wang, Y., Xiong, Z., Liu, C., Stewart, A. J., Dujardin, T., Bountos, N. I., Zavras, A., Gerken, F., Papoutsis, I., Leal-Taixé, L., and Zhu, X. X.: Towards a Unified Copernicus Foundation Model for Earth Vision, https://arxiv.org/abs/2503.11849, 2025.

Xiao, A., Xuan, W., Wang, J., Huang, J., Tao, D., Lu, S., and Yokoya, N.: Foundation Models for Remote Sensing and Earth Observation: A Survey, https://arxiv.org/abs/2410.16602, 2025.

Xie, Y., Wang, Z., Mai, G., Li, Y., Jia, X., Gao, S., and Wang, S.: Geo-Foundation Models: Reality, Gaps and Opportunities, in: Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems, SIGSPATIAL '23, Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/3589132.3625616, 2023.

Yao, T., Zhang, Y., Qiu, Z., Pan, Y., and Mei, T.: SeCo: Exploring Sequence Supervision for Unsupervised Representation Learning, Proceedings of the AAAI Conference on Artificial Intelligence, 35, 10 656–10 664, https://doi.org/10.1609/aaai.v35i12.17274, 2021a.

Yao, T., Zhang, Y., Qiu, Z., Pan, Y., and Mei, T.: SeCo: Exploring Sequence Supervision for Unsupervised Representation Learning, https://arxiv.org/abs/2008.00975, 2021b.