Hy-TeC: a hybrid vision transformer model for high-resolution and large-scale mapping of canopy height

Ibrahim Fayad, Philippe Ciais, Martin Schwartz, Jean-Pierre Wigneron, Nicolas Baghdadi, Aurélien de Truchis, Alexandre d'Aspremont, Frederic Frappart, Sassan Saatchi, Ewan Sean, et al.

▶ **To cite this version:**

Ibrahim Fayad, Philippe Ciais, Martin Schwartz, Jean-Pierre Wigneron, Nicolas Baghdadi, et al.. Hy-TeC: a hybrid vision transformer model for high-resolution and large-scale mapping of canopy height. Remote Sensing of Environment, 2024, 302, pp.113945. ⟨10.1016/j.rse.2023.113945⟩. ⟨hal-04372506⟩

HAL Id: hal-04372506

https://hal.science/hal-04372506v1

Submitted on 26 Mar 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons CC BY 4.0 - Attribution - International License

Version of Record: https://www.sciencedirect.com/science/article/pii/S0034425723004972
Manuscript_648dc56f8c569fbe1d4368146cf15dfb

Vision Transformers, a new approach for high-resolution and large-scale mapping of canopy heights

Ibrahim Fayad[a,b*], Philippe Ciais[a], Martin Schwartz[a], Jean-Pierre Wigneron[c], Nicolas Baghdadi[d], Aurélien de Truchis[b], Alexandre d'Aspremont[e,b], Frederic Frappart[c], Sassan Saatchi[f], Ewan Sean[b], Agnes Pellissier-Tanon[a], Hassan Bazzi[g]

(a) Laboratoire des Sciences du Climat et de l'Environnement, LSCE/IPSL, CEA-CNRS-SCE32UVSQ, Université Paris Saclay, 91191 Gif-sur-Yvette, France.
(b) Kayrros SAS, Paris, 75009, France
(c) ISPA, UMR 1391, INRA Nouvelle-Aquitaine, Bordeaux Villenave d'Ornon, France
(d) INRAE, UMR TETIS, Université de Montpellier, AgroParisTech, CEDEX 5, 34093 Montpellier, France
(e) Department of Computer Science, École Normale Supérieure, Paris 75230, France
(f) Jet Propulsion Laboratory (JPL), California Institute of Technology, Pasadena, CA, 91125, USA
(g) Université Paris-Saclay, AgroParisTech, INRAE, UMR 518 MIA Paris-Saclay, 91120 Palaiseau, France

Abstract

Accurate and timely monitoring of forest canopy heights is critical for assessing forest dynamics, biodiversity, carbon sequestration as well as forest degradation and deforestation. Recent advances in deep learning techniques, coupled with the vast amount of spaceborne remote sensing data offer an unprecedented opportunity to map canopy height at high spatial and temporal resolutions. Current techniques for wall-to-wall canopy height mapping correlate remotely sensed information from optical and radar sensors in the 2D space to the vertical structure of trees using lidar's 3D measurement abilities serving as height proxies. While studies making use of deep learning algorithms have shown promising performances for the accurate mapping of canopy heights, they have limitations due to the type of architectures and loss functions employed. Moreover, mapping canopy heights over tropical forests remains poorly studied, and the accurate height estimation of tall canopies is a challenge due to signal saturation from optical and radar sensors, persistent cloud covers, and sometimes limited penetration capabilities of lidar instruments. In this study, we map heights at 10 m resolution across the diverse landscape of Ghana with a new vision transformer (ViT) model, dubbed Hy-TeC, optimized concurrently with a classification (discrete) and a regression (continuous) loss function. This model achieves significantly higher accuracy than previously employed convolutional-based approaches (ConvNets) optimized with only a continuous loss function. Hy-TeC results show that our proposed discrete/continuous loss formulation significantly increases the sensitivity for very tall trees (i.e., > 35 m). Overall, Hy-TeC has significantly reduced bias (0.8 m) and higher accuracy (RMSE = 6.6 m) over tropical forests for which other approaches show poorer performance and oftentimes a saturation effect. The height maps generated by Hy-TeC also have better ground sampling distance and better sensitivity to sparse vegetation. Over these areas, Hy-TeC showed an RMSE of 3.1 m in comparison to a reference dataset while the baseline ConvNet model has an RMSE of 4.3 m. Hy-TeC, which was used to generate a height map of Ghana using free and open access remotely sensed data with Sentinel-2 and Sentinel-1 images as predictors and GEDI height measurements as calibration data, has the potential to be used globally.

Keywords: Canopy heights; GEDI; Sentinel-1; Sentinel-2; Vision Transformers, Deep learning, Knowledge distillation

* Corresponding author.
E-mail address: ifayad@lsce.ipsl.fr
Preprint submitted to Remote Sensing of Environment

© 2023 published by Elsevier. This manuscript is made available under the Elsevier user license
https://www.elsevier.com/open-access/userlicense/1.0/

1. Introduction

Precise measurements of canopy height are essential to estimate important forest biophysical parameters, including above-ground biomass, carbon stock density, and habitat quality (Alexander et al., 2018; de Thoisy et al., 2016; Goetz et al., 2010). Canopy height also plays a critical role in ecosystem processes, such as carbon storage, forest productivity, and biodiversity (Lefsky et al., 2002; Zhang et al., 2016). Additionally, multi-temporal monitoring of canopy heights is crucial for assessing forest disturbance regimes, dynamics, and associated deforestation and degradation (Pourshamsi et al., 2018), making it essential for guiding forest management and conservation efforts, carbon emissions reduction, and planning afforestation programs (Chen et al., 2016; Hese et al., 2005; Miles and Kapos, 2008). As a consequence, precise estimation of forest canopy heights on a large scale has emerged as a pressing concern in current global carbon balance research (Giri et al., 2011).

Advancements in remote sensing (RS) technology have enabled the collection of an unprecedented amount of diversified data on forest structure and composition, both spatially and temporally, and has allowed researchers to estimate forest characteristics (e.g. canopy heights, carbon content, tree cover) at unprecedented resolutions ranging from 30 m products (Li et al., 2020; Potapov et al., 2021), to 10 m (Lang et al., 2023a; Schwartz et al., 2022), down to the level of individual trees (Li et al., 2023; Mugabowindekwe et al., 2023; Reiner et al., 2023). Despite the differences in the spatial dimension of the different approaches, they all rely on the same types of remotely sensed information, namely lidar, radar, and multispectral (optical) observations.

Lidar, an instrument mounted on aerial vehicles (aerial laser scanning, ALS), or operating from the Earth orbit (spaceborne), emits laser pulses at an angle close to the nadir direction and can accurately measure the time difference between transmitted emissions and their echoed returns. These returns, collected as 3D point clouds or as waveforms, have been proven to be the most accurate remote sensing way to estimate forest canopy heights (Alexander et al., 2018; Lefsky et al., 2005). Nonetheless, lidar's ground coverage remains limited to small areas due to operational costs of ALS campaigns, or technological limitations of spaceborne systems, that currently can only sparsely sample the Earth's surface (Dubayah et al., 2020; Neuenschwander et al., 2022). Optical (e.g., Sentinel-2, Landsat-8, PlanetScope) and radar (e,g., Sentinel-1, ALOS-2 PALSAR-2) systems on the other hand have repeated global coverage, with sampling resolutions ranging from high (10-30 m) to very high (3 m or less), with some missions providing free and open access to images. However, optical and radar systems do not have the ability to directly measure forest vertical structures. Indeed, optical systems only measure reflected sunlight from the Earth's surface. Over vegetation, this reflectance has shown correlation to forest vegetation structures (Woodcock et al., 1994), the green leaf area (Turner et al., 1999), and even above-ground biomass (Barbier et al., 2010; Couteron et al., 2005; Ploton et al., 2013).

However, it's important to acknowledge that the usefulness of optical sensors is not without limitations. One significant challenge is the saturation effect observed in the reflectance data. Initially, as vegetation density increases, there is also an increase in the reflectance signal, related to changes in the structural properties of the vegetation. At higher levels of vegetation density, the reflectance signal saturates due to increased scattering and absorption of light within the dense vegetation. As such, the saturation effect in optical data is mainly caused by the reduced ability of the sensor to distinguish fine variations in vegetation properties as the structural complexity of the canopy hinders the penetration of light and leads to multiple scattering events (Lu et al., 2012; Ploton et al., 2013). The saturation of the reflectance signal, however, may differ among the spectral bands. Some bands might saturate at lower vegetation density, while others might still show a linear response at higher density levels.

Side-looking radar, an active sensor that transmits electromagnetic waves at a given wavelength and measures the return power by the surface (backscatter), can penetrate through vegetation, with the depth of penetration increasing with longer wavelengths (Le Toan et al., 2011). As such, radar backscatter, which is a function of the geometric and dielectric properties of the volume of the scatterers in forests (e.g. leaves, branches, trunks) has been related to AGB (Baghdadi et al., 2015; Mitchard et al., 2012), and vegetation optical depth (Frappart et al., 2020; Wigneron et al., 2021).

Nonetheless, similar to optical imagery, radar remote sensing also experiences a saturation effect resulting from interactions with multiple scattering surfaces within the vegetation canopy. Initially, as vegetation density and structural complexity increase, for instance, in tall tropical forests, the backscatter signal exhibits a proportional rise. However, at higher levels of vegetation density, the presence of an increased volume of scatterers and multiple scattering phenomena within the dense vegetation causes the radar signals to lose their sensitivity to fine

variations in vegetation properties. Consequently, signal saturation occurs due to the overlapping of signals from numerous scattering elements, even at L-band frequencies (Imhoff, 1995; Sandberg et al., 2011).

To reduce the effect of passive optical and radar systems' saturation, which can cause a saturation effect for downstream tasks (e.g., canopy height estimation of tall trees), the sparse coverage of lidar systems, and to ensure continuous canopy height mapping, current approaches rely on non-linear mathematical models, calibrated using machine learning algorithms. These algorithms learn a function that relates information acquired from the spatially continuous data (i.e. radar and optical data) to forest vertical structures (i.e. heights proxied by lidar measurements; Lang et al., 2023a, 2019; Li et al., 2020; Morin et al., 2022; Ngo et al., 2023; Potapov et al., 2021; Schwartz et al., 2022). Machine learning algorithms can be classified into two major broad groups: conventional machine learning algorithms (ML), such as random forests (RF), or XGBoost, and deep learning algorithms (DL), which encompass the neural network-based family of algorithms. DL algorithms, in their broader definition, correspond to a series of multiple interconnected layers (e.g., convolutional layers) that form a deep neural network. As data flows through this network, each layer progressively learns to extract and transform essential features from the input data. These learned features become increasingly abstract and meaningful as they move deeper into the network. At the end of this process, the final layers integrate these abstract representations to make predictions, which in the context of this work, represent canopy height estimates.

While ML algorithms are easier to implement and are computationally less intensive than DL algorithms, they require structured data to be trained on. In essence, in the context of canopy height estimation, a feature engineering stage (after image filtering and pre-processing) to compute relevant metrics (e.g., vegetation indices, textural information) from radar and optical data is usually preferred over raw data (Li et al., 2020; Morin et al., 2022; Ngo et al., 2023; Potapov et al., 2021). In addition, while ML algorithms can handle a very large feature space, they operate on a pixel-by-pixel basis, and as such are unable to learn from the spatial context of the observation at the pixel scale, unless they are explicitly provided with textural indices. However, textural information is limited to small local windows. For these reasons, and coupled with signal saturation from optical and radar imagery, ML algorithms are severely impacted in their ability to provide accurate height estimates of tall trees (Potapov et al., 2021). In contrast, DL algorithms, while still requiring a pre-processing stage to harmonize acquired image values (i.e., reflectance from optical data or radar backscatter), have the ability to learn complex and hierarchical patterns from the data without relying on hand-crafted features. Through this hierarchical representation learning, DL models can effectively capture intricate relationships between the input features and the canopy height, even when they are not immediately apparent. DL models also offer immense flexibility in network architecture and loss function optimization. While computationally more expensive and harder to implement, DL algorithms offer the advantage of creating customized models that are specifically tailored and optimized to learn the task at hand, improving their capability to accurately capture and analyze complex patterns and relationships from the pre-processed remote sensing data (Fayad et al., 2021b).

Current state-of-the-art deep learning approaches for large-scale mapping of canopy heights, all rely on the convolutional operation, which applies a set of learnable filters or kernels to the input image to produce a set of output feature maps. These convolutional layers extract specific features by sliding over the input image and performing element-wise multiplication and summation operations. By applying a series of convolutional layers, the DL network, also known as a ConvNet, can learn a hierarchy of features, from simple edges and textures to complex shapes and objects. For example, Lang et al., 2023a used an ensemble model of simple ConvNets to produce a global canopy height map at 10 m pixel spacing using Sentinel-2 images as predictors and GEDI's RH_{98} as height proxies. Schwartz et al., 2022 leveraged both Sentinel-1 and Sentinel-2 images for the estimation of canopy heights (GEDI's RH_{95} as proxy) in the Landes forest in France using a ConvNet encoder-decoder architecture and showed higher accuracy than Lang et al., 2023a in comparison to the same independent test sets. Recently, Liu et al., 2023 used two encoder-decoder ConvNets (one to classify between tree/no tree and the other for regression) to create a 3 m map of canopy heights across Europe from PlanetScope images.

While such approaches have shown great potential for the large-scale mapping of canopy heights, they all rely on ConvNet-based architectures, which are limited in their ability to capture long-range dependencies between pixels in an image (Yuan et al., 2021). Since convolutions operate in local neighborhoods, they cannot process information from the entire image in a single step, resulting in a lack of global context and a reduction in accuracy, especially when the input images contain complex relationships between pixels. Additionally, traditional ConvNets apply a single filter to the entire image, which results in a lack of expressiveness. As such, one possible limitation

of the ConvNets approaches for canopy height mapping is their use of down-sampling strategies to increase the receptive field, which results in a loss of resolution and granularity of features in the deeper stages of the model that can be hard to recover in the decoder (Ranftl et al., 2021). To overcome this limitation, several techniques were introduced, such as the use of skip connections (Ronneberger et al., 2015), dilated convolutions (Yu and Koltun, 2016) that are meant to increase the receptive field of ConvNets, or more recently, the deployment of attention mechanisms at various stages of the network to enable selective focus on important features and better capture long-range dependencies (Chen et al., 2022, 2021; Oktay et al., 2018). While these techniques have proven to significantly improve prediction quality, ConvNet-based models are still hindered by the limitations of the convolutional operation itself within the encoder. As such, a ConvNet-based canopy height model might produce outputs with smooth details even though the input and output resolutions are the same.

The main objectives of this study are therefore three-fold. First, to mitigate some of the limitations of ConvNet-based models we will employ and evaluate a more recent family of deep learning models based on the transformer architecture. Transformer-based models have become the standard in natural language processing tasks (Yuan et al., 2021). More recently, vision transformers (ViTs) have adapted the transformer architecture for computer vision tasks, demonstrating qualities such as the ability to capture long-range dependencies and a global receptive field (Vaswani et al., 2021). Moreover, a ViT-based encoder could be an interesting alternative to a ConvNet one as it eliminates the need for explicit down-sampling operations. Instead, ViTs by design, maintain a consistent dimensional representation throughout the processing stages. However, while they have been successful with large datasets, they tend to perform worse than similarly sized ConvNets when trained on smaller amounts of data. This may be due to the lack of desirable properties in the ViT architecture that are inherent to ConvNets, like an inductive bias to focus on nearby image elements, making them less effective on smaller datasets (Touvron et al., 2021). As such, our second objective is to propose an approach for training vision transformers through a novel teacher/student learning paradigm that alleviates the need for additional data sources. This is in contrast to the recent work of Tolan et al., 2023 which proposed the use of a Self-Supervised Learning (SSL) approach, referred to as DINO (Caron et al., 2021; Zhou et al., 2022) to extract encoded features from the input images, and later use a transformer based decoder to transform the encoded features into canopy height maps over the state of California in the USA, and the state of São Paulo in Brazil.

Our final objective is related to the underestimation of the height of tall trees. This phenomenon is mainly due to two factors. First, as stated earlier, signal saturation from either optical or radar images cannot be easily mitigated in conventional ML approaches. However, DL-based approaches, such in the case of ConvNet or transformer-based models, have better spatial awareness. As such, DL models are able to extract information related to forest structures and use them in addition to spectral information to infer canopy heights. Thus, DL approaches should have better capabilities in the separation between tree vertical structures even in the presence of optical and radar saturation on some bands.

Another factor that leads to underestimation is related to the employed loss function. As we are dealing with a regression task (i.e. canopy height estimation) current approaches rely on losses such as the L1 (Mean Absolute Error, MAE) or L2 (Mean Squared Error, MSE), or a variation of those such as the Huber loss (Huber (1964)). However, these losses might not be suitable when dealing with a complex regression problem with a multi-modal or skewed distribution of targets, as is the case with height in heterogeneous land covers. Such losses tend to minimize the average loss across all modes, resulting in a fit that cannot represent extreme cases, such as tall trees. To address this issue, a more complex loss function will be proposed and is based on a discrete/continuous formulation, where the target values are classified into bins, and the model is tasked to simultaneously estimate the target value and predict the corresponding class. Similar approaches have demonstrated better accuracy for regression problems (Kundu et al. (2018); Massa et al. (2016); Mousavian et al. (2016)).

To this end, several ConvNet-based models, a transformer-based model as well as different loss functions were assessed on their ability to estimate the heights across different land covers in an end-to-end matter. The models were evaluated over the Republic of Ghana as it represents an interesting case study due to its very diverse land cover, ranging from savannas to dense tropical forests. Hence, this study should provide valuable insights into the abilities, as well as limitations, of our proposed approach on the estimation of heights in different scenarios, ranging from sparsely vegetated areas in savannas to the estimation of heights of tall trees within tropical forests.

2. Study area and datasets

2.1. Study area

The study area is Ghana, a country in West Africa with a total surface area of 238,535 km^2 (Fig. 1). It is bordered by Côte d'Ivoire to the west, Burkina Faso to the north, Togo to the east, and the Gulf of Guinea to the south. Ghana's terrain is characterized by low-lying coastal plains that rise to a central plateau. The terrain slopes gently from north to south, and the highest point is Mount Afadjato, which has an elevation of 885 meters.

Ghana has a tropical climate, with a rainy season from April to July and a dry season from November to March. The average temperature throughout the year is around 27°C, and the country experiences a relatively high level of rainfall, with an average of around 1500mm/year. The Koppen-Geiger classification (Peel et al., 2007) categorizes Ghana's climate as Aw, which is a tropical savanna climate.

Ghana's landscape is diverse, including forests, savannas, and wetlands. Forested areas, mostly located in the southern and central regions of the country, cover around 50,000 Km^2 of the total surface area. These forests are home to various tree species, such as the Ghanaian mahogany, the African teak, and the iroko. The forests in Ghana are classified into tropical rainforests and mangroves. Tropical rainforests are the most abundant type, covering around 40% of the country's forested area. Mangroves, located along the coast, also cover around 30% of the country's forested area. Finally, the remaining 30% of trees are found in savannas, mostly located in the northern and north-east regions of Ghana (Fig. 1) and characterized by grasslands and scattered trees.

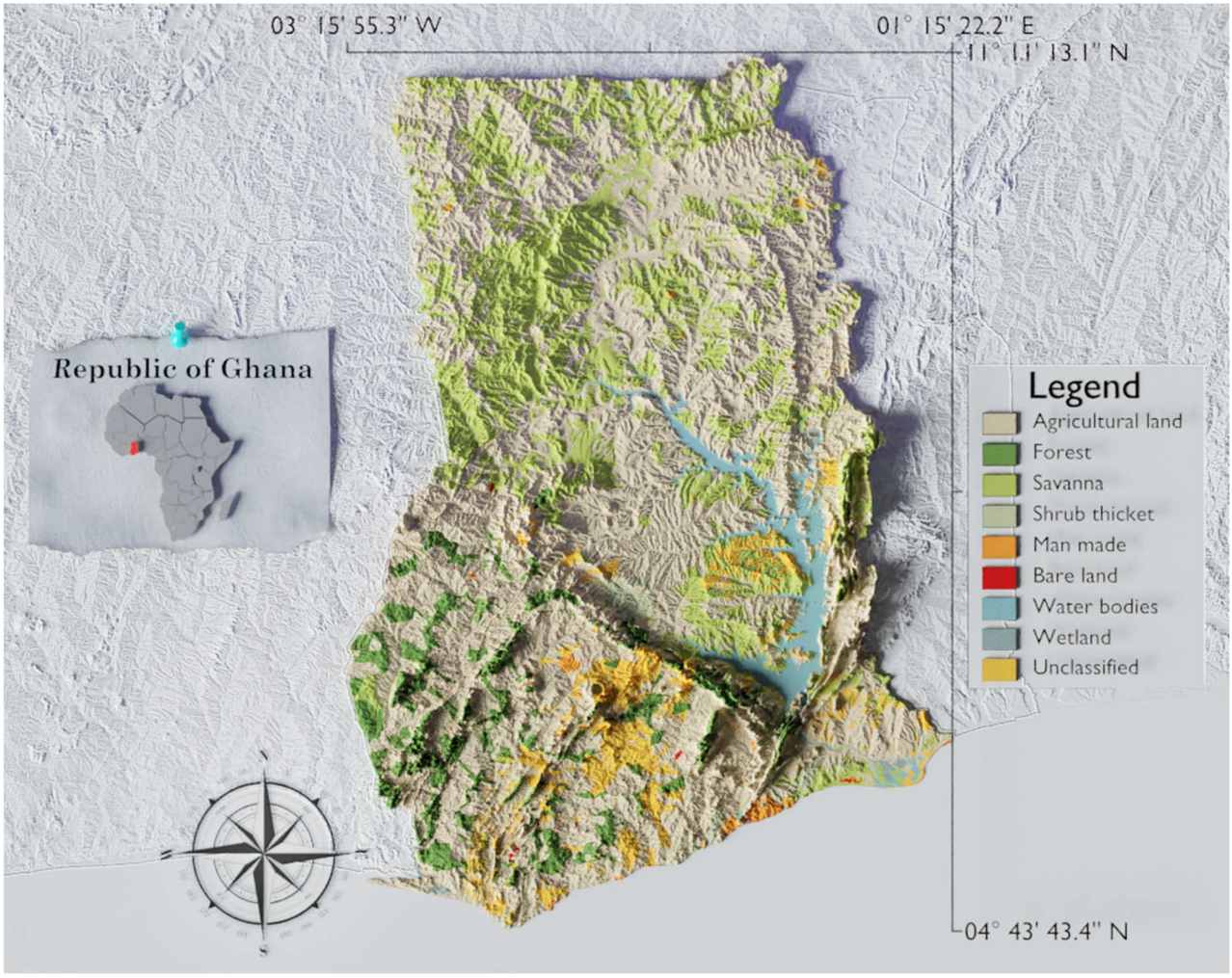


Figure 1. Location of the study area, the Republic of Ghana, and the different land cover classes (Source of the land cover classes: https://geog.sdsu.edu/Research/Projects/IPC/research/ids.html).

2.2. Datasets

Here we present the different used datasets. The datasets were divided between datasets used as model inputs, datasets used to calibrate the models or validation datasets. As model inputs, two datasets, namely Sentinel-1 and Sentinel-2 were used for model training and inference. For model calibration/validation, GEDI data were used and were split between 65% for model calibration, 10% used for hyperparameter tuning and performing model selection during the training phase, and the remaining 25% were set aside for model testing (i.e., validation). Finally, two high-resolution CHMs, one derived from an ALS dataset, and the other from Pleiades imagery were used as an additional independent validation dataset. The processing of the datasets is described below.

2.2.1. Input: Sentinel-2 Data

We used the Sentinel-2 (S2, https://scihub.copernicus.eu/) surface reflectance data (Level-2A) provided by Google Earth Engine (GEE, https://earthengine.google.com). The Level-2A product (S2-L2A) comprises bands with a spatial resolution of 10m (Blue, Green, Red, and NIR bands) and 20m (Red Edge 1, Red Edge 2, Red Edge 3, Red Edge 4, SWIR1, and SWIR2 bands). For our study, all bands were upscaled with cubic interpolation to obtain bands with a 10m ground sampling distance (GSD). Next, S2 cloudy pixels were filtered out using the S2 Cloud Probability dataset provided by SentinelHub in GEE. After cloud removal, we created a mosaic using all cloud-filtered S2 images acquired between January 2019 and March 2020 by calculating the median of all time-series images over a selected pixel. Mosaicking aids in reducing the effect of atmospheric conditions between time-series acquisitions that can lead to large differences between images. The choice to use 15 months of acquisitions is due to the persistent cloud cover in the southern part of Ghana; as creating a composite from fewer acquisitions resulted in some areas with artifacts or missing pixels.

2.2.2. Input: Sentinel-1 Data

SAR images were obtained from the Sentinel-1A (S1A) and Sentinel-1B (S1B) satellite constellation operating at the C-band (frequency of 5.405 GHz corresponding to a wavelength ~6 cm). S1 images, also provided by GEE, are calibrated using the Sentinel SNAP toolbox (https://step.esa.int/main/toolboxes/snap), developed by the European Space Agency (ESA). The calibration aims to convert the image pixel values into backscattering coefficients (σ°) in linear units, while the geometric correction was used to ortho-rectify the SAR images using the 30 m Shuttle Radar Topography Mission (SRTM) Digital Elevation Model (DEM). S1 SAR images were acquired in the Interferometric Wide swath mode (IW) in both VV and VH polarizations. These images were generated from the high-resolution Level-1 ground range detected (GRD) product, which was acquired at a spatial resolution of 20m × 22m (range × azimuth), and then resampled to produce the final product at 10m × 10m GSD. The radiometric accuracy of S1 SAR backscattering coefficients is approximately 0.70 dB (3σ) for the VV polarization and 1.0 dB (3σ) for the VH polarization (Schwerdt et al., 2017). Finally, the backscattering coefficients σ_θ^0 acquired at different incidence angles θ were normalized to a common reference incidence angle set to $\sigma_{ref} = 40°$ using the square cosine correction equation (Baghdadi et al., 2001; Topouzelis et al., 2016):

$$\sigma_{ref}^0 = \frac{\sigma_\theta^0 \, cos^2(\theta_{ref})}{cos^2(\theta)} \qquad 1$$

Like all active microwave systems, C-band S1 is sensitive to different physical phenomena that occur over a given landscape. Over bare and sparsely vegetated fields, soil moisture (SM) has a dominant influence on the C-band backscatter values (Baghdadi et al., 2017, 2008). To reduce the impact of SM, only images that were acquired in the absence of any rainfall events and up to four days prior (Bazzi et al., 2021) were retained. We define a rainfall event as rainfall (from GPM: Global Precipitation Measurement, V6) of more than 40 mm over a period of four days. Next, a composite using all images was created by calculating the median of all available acquisitions over a given pixel.

2.2.3. Calibration and validation: Gedi Data

In this study, we used GEDI's dataset Level 1B (L1B, Dubayah et al., 2021a), Level 2A (L2A, Dubayah, Ralph et al., 2021), and Level 2B (L2B, Dubayah et al., 2021b) from April 2019 to December 2021. These datasets are published by the Land Processes Distributed Active Archive Center (LP DAAC). The L1B data product contains, among others, the geolocated waveforms. The L2A data product provides metrics relating to the vertical structure of vegetation from the processed waveforms using six possible settings groups (henceforth referred to as algorithms) of differing thresholds to satisfy a variety of acquisition scenarios (Dubayah et al., 2020). Moreover, in the L2A data product, a variable named 'selected_algorithm' corresponds to the algorithm with the lowest (in terms of elevation) non-noise ground return (henceforth referred to as the "lowest mode"). Therefore, to extract canopy height recorded by GEDI and help filter GEDI data, we extracted several variables from the L1B, L2A, and L2B data products. From the L1B we extracted the waveforms, and from the L2A, and L2B data products we extracted the following variables derived from the processing algorithm identified by the selected_algorithm variable:

- The latitude, longitude, and elevation of the lowest mode.
- The latitude, longitude, and elevation of the instrument.
- The number of detected modes (num_detectedmodes).
- The acquisition date and time of the shot.
- The beam sensitivity, that is, the probability of detecting the ground below a given canopy cover, 90% of the time with a 5% chance of a false positive (Hancock et al., 2019).
- The viewing angle (VA) at acquisition time by using the latitude, longitude, and elevation of both the footprint and the instrument (Fayad et al., 2022).
- The signal-to-noise ratio (SNR) (Fayad et al., 2022).
- The relative height at 98% (RH_{98}) of returned energy as a proxy of canopy height (Blair et al., 1999).
- The elevation of the lowest mode (ELM) and the elevation of the highest return (EHR) which represent the canopy top.
- Canopy cover

Over Ghana, 27,096,322 shots were acquired during the period between April 2019 and November 2021. However, not all of these shots are viable due to unfavorable atmospheric conditions (e.g., clouds) that affected them, or other instrument-related errors. Therefore, we removed all waveforms that met any of the following criteria:

- Shots without any detected modes (num_detectedmodes = 0). These shots were noisy signals with a signal-to-noise ratio (SNR) of 0 on a linear scale.
- Shots with an SNR < 12db or VA > 5°.
- Shots with beam sensitivity lower than 95%.
- Shots with an absolute difference between the elevation of the lowest mode (ELM) and the corresponding SRTM DEM of more than 75 m ($|ELM - SRTM| > 75m|$).
- Incomplete waveforms, as in returns that weren't fully recorded by GEDI. These waveforms are detected by their search_end variable (end location of the usable part of the waveform) that is near the total number of bins in a waveform "rx_sample_count" ($rx_sample_count - search_end \leq 1\ bin$).
- Footprints with high geolocation errors. For these footprints, we relied on the difference between the footprint's derived canopy cover and the NDVI calculated from S2 (upsampled to 30 m to approximate the footprint's size). We removed all footprints that reported a difference between its canopy cover and its corresponding NDVI greater than 1.5 standard deviations (1.5σ) from the mean of all differences, following Liang et al., 2023.

After Filtering, 43% of the shots were retained over Ghana. Fig A1 shows the distribution of retained RH_{98} over the entire study area.

2.2.4. Validation: ALS Data

An airborne laser scanning dataset provided by GlobALS (Stereńczak et al., 2020), acquired over an area of ~10 km^2 of dense tropical forests in southwestern Ghana in 2012 was used for independent validation (Fig. 2, left). The ALS dataset was acquired using Optech Gemini which is a wide-area, high-altitude airborne discrete lidar mapping

system with a point density of 10/m^2. From the point cloud dataset, two rasters were generated: a digital elevation model (DEM) representing ground elevations, and a digital surface model (DSM) representing the height of the top of the canopy. Both the DSM and DEM were generated using PDal (https://pdal.io/).

For the DSM, we only considered the first returns from the laser pulses to build it at 1 m resolution, while for the DEM, we used the Statistical Multiresolution Filter (SMRF, Pingel et al., 2013) along with points classified as ground. The SMRF filter adopts a linearly increasing window and simple slope thresholding in conjunction with an image inpainting technique to filter point clouds. Given the characteristics of the location, we have set the window radius to 17 m, a slope tolerance of 0.05, an elevation threshold of 0.35 m, and a scaling factor of 0.9. The threshold parameter controls the sensitivity of the filter to distinguish ground points, while the scaling parameter adjusts the calculated elevation values for the final DEM. Finally, the canopy height model (CHM) was generated by subtracting the DSM from DEM at 1m resolution.

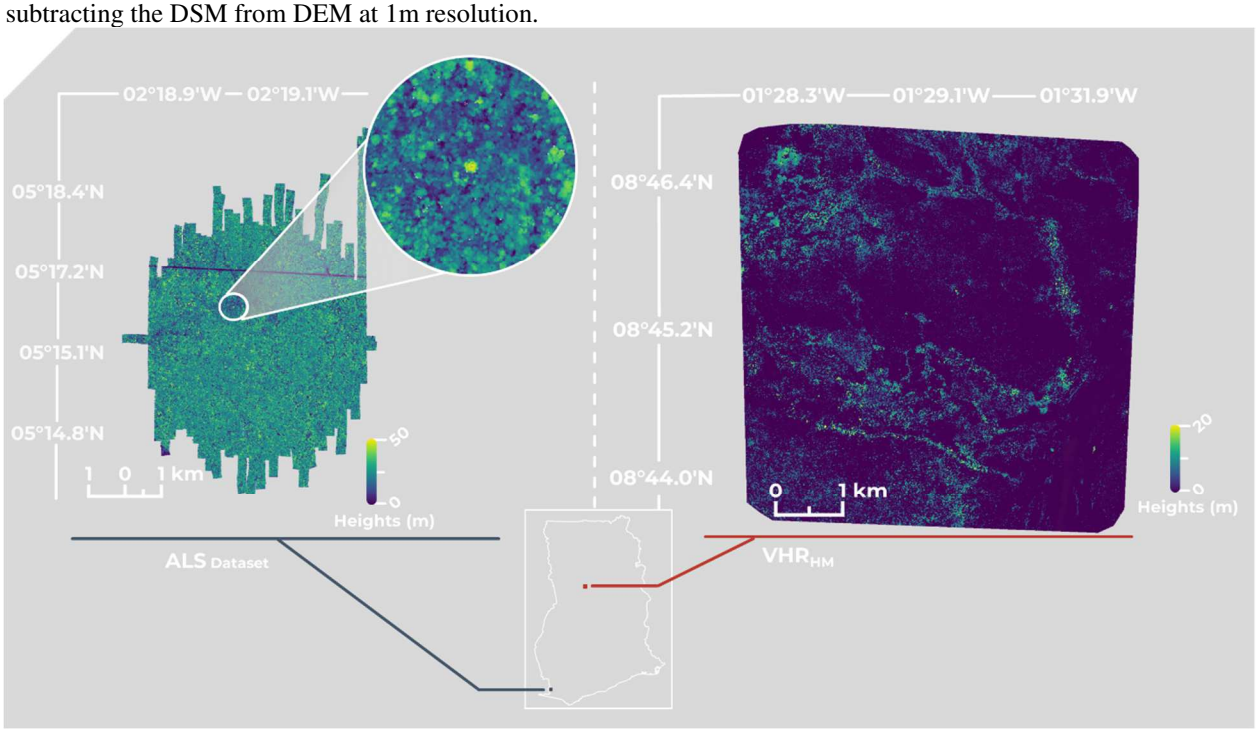


Figure 2. The ALS dataset (1 m resolution) and the Stereo 3D reconstruction (VHR_{HM}) of canopy heights from Pleiades imagery (0.7 m resolution).

2.2.5. Validation: Stereo 3D reconstruction from Pleiades imagery

In addition to the ALS dataset which covers a dense tropical forest region, we have chosen another location, also for validation purposes, that is characterized by short and sparse trees. This location was covered by a stereoscopic Pleiades image acquired in 2016. Pleiades is a constellation of optical satellites owned and operated by Airbus. Pleiades acquire very high resolution (VHR) images in panchromatic, visible, and infrared bands at 0.7 m (panchromatic) and 2.8 m (multispectral) resolutions and a geolocation accuracy better than 5 m. To derive a height map from Pleiades VHR acquisitions, we adapted the stereo 3D height reconstruction technique from the study of De Franchis et al., 2014, and the stereoscopic Pleiades acquisition. The height map was obtained by first creating a digital surface model (DSM) at 0.7 m resolution from point-cloud 3D reconstruction based on the two Pleiades images. Then a cloth simulation algorithm was applied to select the ground points from the 3D point Cloud (Zhang et al., 2016), which were then interpolated by a Laplace operator (Ousguine et al., 2016) to obtain a digital elevation model (DEM). Finally, a canopy height model (CHM) was obtained by subtracting the DEM from the DSM. Fig. 2 (right) shows the 34 Km^2 canopy height map (hereafter referred to as VHR_{HM}) derived using the pre-mentioned approach.

3. Methodology

In this study, we propose a novel vision transformer-based model and evaluate its ability to generate wall-to-wall canopy height maps. The vision transformer model was also evaluated against current state-of-the-art ConvNet-based models serving as a baseline. For the baseline models, we started with an architecture that builds on the latest architectures for canopy height mapping and we introduced several optimizations, notably at the level of the loss functions to reduce the effect of underestimation of tall trees. These modifications are first introduced in the ConvNet baseline models and later used within the vision transformer architecture. The different models were trained using the high-resolution (10 m) optical and/or radar imagery serving as independent input data and GEDI height estimates serving as a proxy of ground truth height estimates (i.e., dependent data). This section details model architectures as well as the loss functions used to optimize each model.

3.1. ConvNet based models

Serving as baseline are ConvNet-based models inspired by the encoder-decoder architecture from the study of Chen et al., 2022, which reflects current advances on the UNet architecture. The UNet and its variants are a type of ConvNets that are commonly used for image segmentation, dense prediction, and other tasks involving image data. First introduced by Ronneberger et al., 2015 for the segmentation of biomedical images, it has since been successfully deployed for various applications across multiple disciplines and has shown great potential for the estimation of canopy heights (Li et al., 2023; Schwartz et al., 2022). The primary function of a UNet is to learn the mapping between the input and output images by learning the features within an input image at different scales. It consists of a compressive path (encoder) to extract features from the input image, an expansive path (decoder), and finally an output head to reconstruct the output image from the decoded features.

Our proposed implementation of the UNet model is presented in Fig. 3. This architecture is composed of two encoder paths (separate encoders for S1 and S2) with an input of size W×H and N bands for the encoder using the S1 images, and W×H and M bands for the encoder using the S2 images. The architecture comprises a single decoder path that produces feature maps of size W×H and C bands (same size as the input and 64 bands). For the output head, given that the task is a regression problem, the implemented output head will receive from the decoder the feature maps of size W×H and C bands and produces an output of size W×H and 1 band. The detailed implementation of the different parts of the model can be found in Appendix B.1 through B.3.

The first baseline model, composed of two encoder paths and producing a single output will hereafter be referred to as 2MOU (**M**ulti**M**odal, single **O**utput **U**Net).

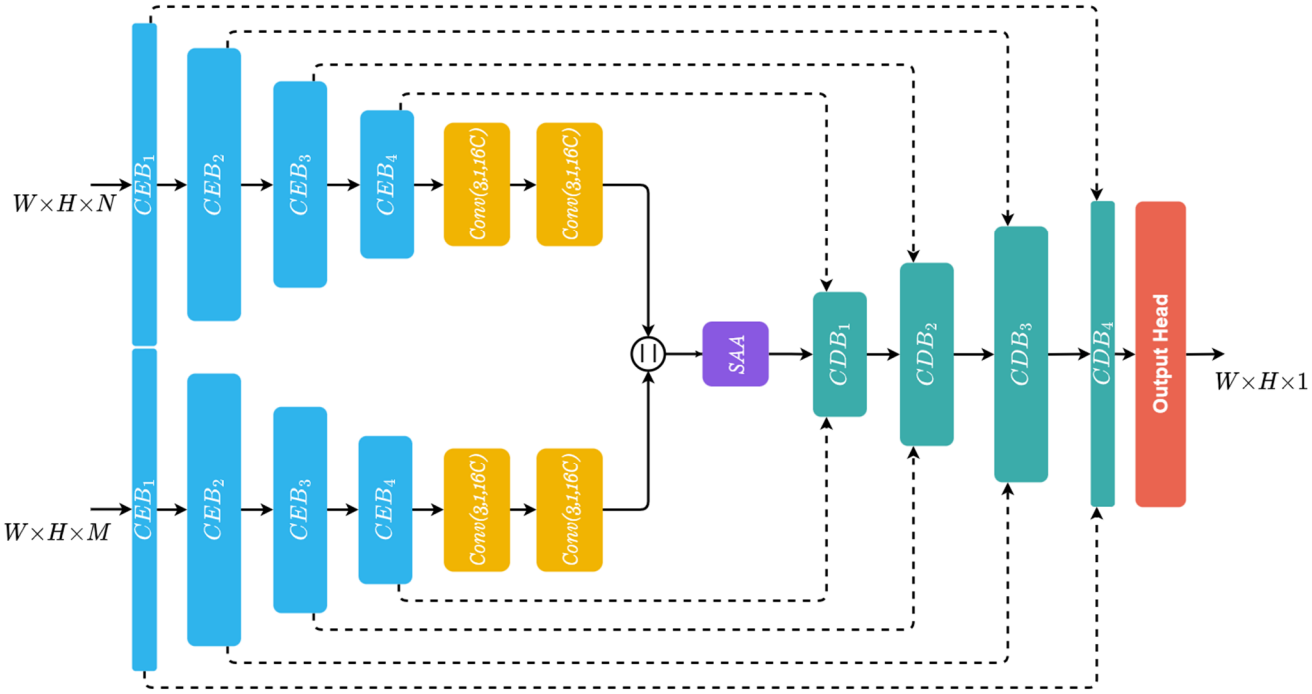


Figure 3. The architecture of the UNet-based model. CEB_i represents an encoder block, CDB_i, represents a decoder block. Skip connections are represented by dashed lines. The *Output Head* represents the output block.

3.1.1. Model optimization

In terms of model optimization, as we are dealing with a regression problem, we adopt the Huber loss (Huber, 1964) for the 2MOU model as follows:

$$\ell_{2MOU}(Y, Y') = \frac{1}{N}\sum_{i=1}^{N}\begin{cases} \frac{1}{2}(Y_i - Y_i')^2, if\ |Y_i - Y_i'| < delta \\ delta * \left(|Y_i - Y_i'| - \frac{1}{2} \cdot delta\right), otherwise \end{cases} \quad (2)$$

Where Y is the vector of target values to estimate and Y' the vector of truth values, N is the number of samples, and delta is a hyperparameter that determines the threshold for the function to switch from quadratic loss to linear. In this study, delta was set to 3 m. The choice to use the Huber loss is that it is more robust to outliers than the Mean Squared Error (MSE) loss, and preserves some of the regularity in the loss surface. Unlike the MSE loss, which penalizes large errors quadratically, Huber loss linearly penalizes large errors beyond a certain threshold, making it less sensitive to outliers. Similarly, the Mean Absolute Error (MAE) loss also linearly penalizes errors but lacks the smooth transition in the penalization for large errors, as provided by the Huber loss. As such, Huber loss is a better compromise between robustness and regularity in comparison to the MSE and MAE loss functions.

3.1.2. Alternative output head

In many machine learning regression tasks, the most common regression losses used are the L1 (MAE) or L2 (MSE), or a variation of such losses such as the Huber loss. Nonetheless, as stated earlier, when dealing with non-normal distributions of the targets, such as the case of low number of GEDI acquisitions over tall trees in comparison to shorter ones, a model trained with such a family of losses may under- or over- estimate extreme values. Over Ghana, Fig. A1 shows that the majority of acquired GEDI measurements are less than 15 m (~95% of footprints). Therefore, we suspect that by regressing the targets directly, along with the issue of signal saturation, the model will fail to generalize GEDI measurements over tall forests. To mitigate this issue, we propose an

alternative output head than presented in the previous section. The modified output head will perform two tasks. The first task is to attempt to classify the heights into *K* overlapping bins (classification sub-module, K is user-defined), while the second task will output continuous height estimates (regression sub-module). This discrete/continuous formulation has shown better accuracies for regression problems (Kundu et al., 2018; Massa et al., 2016; Mousavian et al., 2016). Moreover, the classification sub-module could also be encouraged to pay more attention to underrepresented classes by weighing the contribution of each class to the total classification loss. As such, the modified head will receive input from the decoder (Fig. 3) of shape W×H with C bands and produce two outputs, the first one, used for regression, has a shape of W×H and 1 band and the second output used for classification has a shape of W×H and K bands. This second implementation of the U-Net model will hereafter be referred to as 2MDU (**M**ulti**M**odal, **D**ual output **U**Net). Detailed implementation of the modified output head can be found in Appendix B.4.

In terms of loss functions, as we are trying to optimize the model for both the classification and regression tasks, the total loss (ℓ_{CR}) of the 2MDU model is defined as:

$$\ell_{CR}(P,T,Y,Y') = \ell_{CE}(T,P) + \alpha \cdot \ell_{R}(Y,Y') \quad 3$$

Where *P* and *T* are respectively the predicted and true probabilities for each class and for each target and truth value, ℓ_R is the regression loss (i.e., Huber, eq. 2), and α is a scaling parameter set to 5 to balance both losses. ℓ_{CE} is equal to a weighted cross-entropy loss since we are dealing with an imbalanced dataset. Cross-entropy loss is a commonly used loss function for multi-class machine learning models. It measures the difference between the predicted probability distribution and the true distribution represented by the confidence labels. In a multi-class scenario, the cross-entropy loss calculates the average negative log likelihood of the correct class. The cross-entropy loss (ℓ_{CE}) is defined as follows:

$$\ell_{CE}(P,T) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{k} w_j \cdot t_{ij} \cdot log(p_{ij}) \quad 4$$

Where t_{ij} and p_{ij} are respectively the true and predicted probabilities of class *j* for target value *i*, w_j is the weight assigned to class *j*. w_j is equal to the inverse ratio of GEDI acquisitions belonging to a certain class to the total number of GEDI acquisitions in a single batch.

For the regression loss (ℓ_{CR}), we tested two regression losses, the first one is the Huber loss (eq. 2). However, the residual (difference between predicted and true values) is not bounded in a Huber loss, with a magnitude that is linearly proportional to the difference between model estimates and the GEDI targets for large differences, and quadratically proportional to the difference for small differences. As model optimization progresses, estimated values that are close to high-quality GEDI measurements will have little contribution during backpropagation, while estimates that are far from GEDI measurements, due for example to GEDI's geographical location errors, or instrument errors (e.g., limited laser penetration in densely vegetated forests) will produce a gradient with high magnitudes, and as a result, the model will adapt to these outliers and deteriorate its performance for the inliers. To reduce the effect of these outliers, we replaced the Huber loss with an adaptive and robust regression loss developed by Barron, 2017 (Eq. D1). The 2MDU model optimized with the adaptive loss function will hereafter be referred to as α-2MDU.

3.2. Vision Transformer based model

Our novel proposed approach for canopy height mapping is based on a hybrid transformer-ConvNet architecture that is specifically tailored for the continuous prediction of canopy height. As we are dealing with a dense prediction problem, the proposed model is also based on an encoder-decoder architecture. However, in contrast to the previous approach, the ConvNet encoder is replaced with a transformer-based encoder with 12 transformer layers ($TB\ \#i, i \in \{1,2,\dots,12\}$, Fig. 4). Next, an adapted ConvNet decoder that hooks to the output of some

transformer layers ($DBi, i \in \{1,2,3,4\}$, Fig. 4) is tasked with reconstructing the encoded images and transforming them into the canopy height estimates. As we stated in the introduction, transformer-based models require an order of magnitude more data than similarly sized ConvNets (Touvron et al., 2021). Moreover, transformer-based models, unlike ConvNets which are designed to exploit the spatial structure of images, rely solely on the attention mechanism to extract features from images. This means that such models need to learn how to attend to relevant parts of the image pixels and how to combine information across different regions of the image. This requires a stronger inductive bias to guide the learning process. As such, transformers usually require long pre-training times, before being fine-tuned to the required task (He et al., 2021). To counter the issue of data limitation and long pre-training times, we get inspiration from the study of Touvron et al., 2021, where we couple the proposed transformer architecture with a pre-trained UNet-based model, namely α-2MDU and optimize the model in a teacher/student training paradigm. The transformer-based model will hereafter be referred to as Hy-TeC (**Hy**brid-**T**ransformer **e**ncoder with a **C**onvolutional decoder). The proposed architecture of Hy-TeC is presented in Fig. 4 while the detailed implementation of the encoder, decoder, and the training strategy can be found in appendix C.

Finally, due to the proposed configuration, the model will need to optimize four loss functions. At the highest resolution (i.e., 10 m), where the model outputs through DB_4 (Fig. 4) a 10 m height estimate and a height class label, the loss function is identical to eq. 3. For the KD loss, Hy-TeC outputs through the remaining decoder blocks, height estimates at resolutions of $\frac{W}{8} \times \frac{H}{8}$ (i.e., 80 m), $\frac{W}{4} \times \frac{H}{4}$ (i.e., 40 m), and $\frac{W}{2} \times \frac{H}{2}$ (i.e., 20 m). The blocks outputting height estimates at resolutions of 40 (DB_2, Fig. 4) and 80 m (DB_1, Fig. 4) will try to emulate the output of the teacher model at the same resolutions (by upscaling the output of the teacher model from 10 m to the required resolution), and thus the associated loss functions for these decoder blocks are identical to eq. 2. The choice not to use the loss function in eq. 3 for the lower resolution outputs, is simply to reduce model complexity. As such, the full loss function that the Hy-TeC model tried to minimize is defined as:

$$\ell_{HT}(H_{db4}, H'_{db4}, HC_{db4}, HC'_{db4}, H_{db2}, H'_{db2}, H_{db1}, H'_{db1}) = \beta_1 \cdot \left(\ell_{CE}(HC_{db4}, HC'_{db4}) + \alpha \cdot \ell_{AR}(H_{db4}, H'_{db4})\right) + \beta_2 \cdot \ell_R(H_{db2}, H'_{db2}) + \beta_3 \cdot \ell_R(H_{db1}, H'_{db1}) \quad (5)$$

Where H_{db4} and H'_{db4} are respectively the height estimates from DB_4 and GEDI. HC_{db4} and HC'_{db4} are respectively the height classes from DB_4 and GEDI. $H_{dbi}, i \in \{1,2\}$ are the height estimates from the different decoder blocks $DBi, i \in \{1,2\}$, while $H'_{dbi}, i \in \{1,2\}$ are the height estimates from the teacher models. $\beta_i, i \in \{1,2,3\}$ are scaling hyper-parameters that we set to 1 for β_1 and 0.75 for the rest.

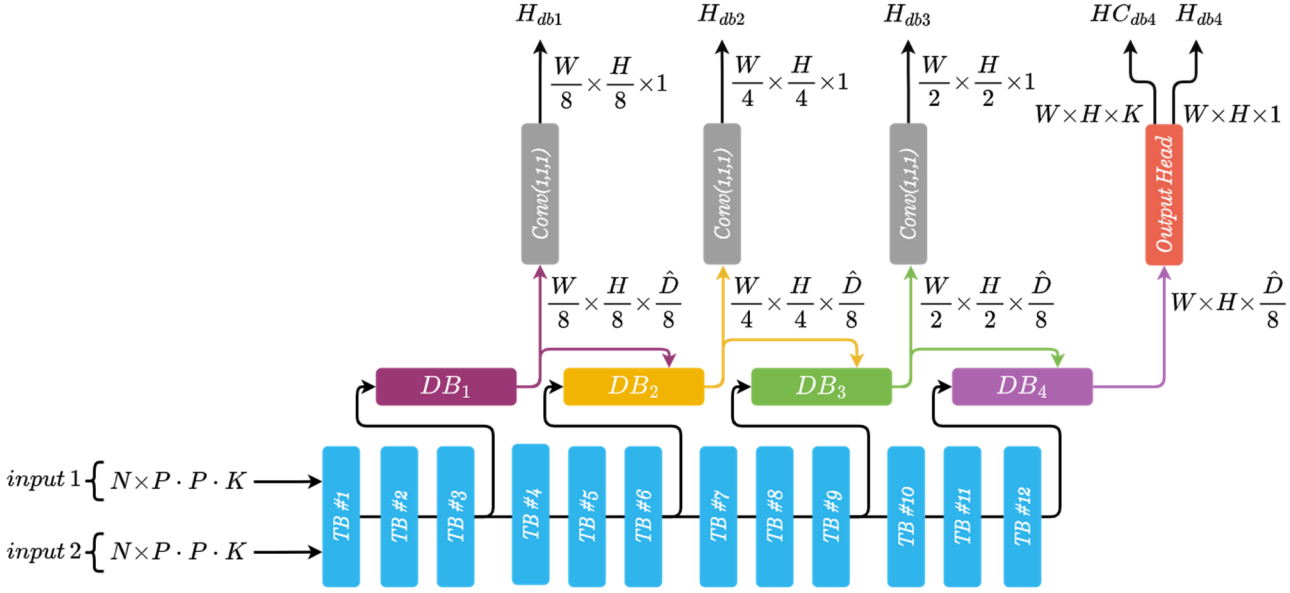


Figure 4. The implementation of the Hy-TeC model.

4. Experimental settings

The four proposed models (2MOU, 2MDU, α-2MDU, and Hy-TeC) were evaluated on their capability to estimate the canopy heights over Ghana. In this section, we will first start by introducing the data requirements for the models, followed by the training strategy, model evaluation metrics, and finally the results. Model inputs and the associated loss function of each model are presented in the following table (Table 1).

Model designation	Section	Inputs	Architecture	Loss function
2MOU	3.1	S1 and S2 data	ConvNet	Huber loss
2MDU	3.1.1	S1 and S2 data	ConvNet	Cross entropy + Huber loss
α-2MDU	3.1.1	S1 and S2 data	ConvNet	Cross entropy + Adaptive loss
Hy-TeC	3.2	S2 data	Hybrid (vision transformer encoder and ConvNet decoder)	Cross entropy + Adaptive loss

Table 1. Summary of the four evaluated models, their input data, architecture, and the loss function used for optimizing each model.

4.1. Sampling strategy

Due to the unbalanced distribution of GEDI measurements over Ghana (small surface area of tall forests), two sampling steps were defined to derive the datasets for the input image patches (i.e., S1, and S2) and the targets (i.e., GEDI measurements) that the different models used in either the training or validation stages. Before extracting the image patches that the models used as inputs, we first defined a regular grid with a cell size of 768×768 pixels (i.e., 7680 m×7680 m) that covers the entirety of Ghana. Next, for each grid cell, we recorded the distribution of available GEDI height measurements at 5 m vertical intervals. This resulted in 2559 non-overlapping cells containing at least 600 GEDI acquisitions (grid cells with GEDI acquisitions of less than 600 were not considered for training, tuning, or validation). Then, to guarantee that during training (respectively tuning or validation) the models are trained (respectively tuned or validated) across all height ranges, we further split the 2559 cells by the ratio of each height interval within each grid cell and selected the grid cells based on sufficient RH_{98} height ratios for each height range (Table E1).

Next, the non-overlapping grid cells from each set were further split to 65% for training, 10% for hyperparameter tuning and model selection, and the remaining 25% were used as a set-aside test set (i.e., validation). We also duplicated some grid cells from each set (i.e., training, tuning, or validation) n times (Table E1) to better balance the amount of available RH_{98} height ranges within the training (respectively tuning or validation) datasets. Finally, during a training step (respectively tuning or validation), a 256×256 pixels randomly defined bounding box within the larger 768×768 pixels grid cell was used to extract the necessary S1 or S2 image patches used as model inputs, and the corresponding RH_{98} as the targets or the dependent data. Even though during a training step (respectively tuning or validation), the same training (respectively tuning or validation) 768×768 pixels grid-cells could be selected multiple times, overfitting is controlled by image augmentation techniques (random crop and flip of the 256×256 pixels image patches).

4.2. Model inputs and training

For the UNet-based baseline models (i.e., 2MOU, 2MDU, and α-2MDU), the selected 256×256×2 (W×H×N) S1 images and the 256×256×10 (W×H×M) S2 images were fed to the models in batches of 12 without any further modification. For the transformer-based model (i.e., Hy-TeC), only S2 bands were considered as inputs to the encoder. While our initial testing showed that S1 data were informative to the UNet-based models, this was not the case for Hy-TeC. In fact, Hy-TeC trained with S2 only or S1 and S2 provided the same validation metrics. As

such, since SAR backscatter could be disturbed by several parameters such as forest structure, soil and vegetation moisture, and understory rebound, S1 data were not considered for Hy-TeC to reduce artifacts in the output canopy height maps. Therefore, for Hy-TeC, input 1 is composed of four S2 bands (R, G, B, and NIR), while input 2 is composed of six S2 bands (red edge 1, 2, 3, and 4 and SWIR 1 and 2). The 256×256 images with B bands for input 1 and input 2 (B = 4 for input 1 and 6 for input 2) were first divided into 256×16×16×B (N×P×P×B) non-overlapping patches and each patch flattened to a 1D vector of size 16×16×B. We then mapped each flattened patch vector to a lower-dimensional space by applying a linear embedding of size 1536, representing the tokens. Finally, we added the positional encoding to the projected tokens and concatenated the embeddings from input 1 and input 2 to form a 2D matrix of size 512×1536 (256×1536 from input 1 and 256×1536 from input 2), which are then fed in batches of 12 to the model.

In terms of optimization, a Stochastic Gradient Descent (SGD) optimizer was used for the UNet-based models with an initial learning rate of 1e-2 and a total batch size of 12 for a maximum of 250 epochs, with a cosine learning rate decay scheduler. Hy-TeC on the other hand was trained with the AdamW (Loshchilov and Hutter, 2019) optimizer with a linear warmup scheduler that gradually increases the learning rate from 1e-6 to 1e-4 over 20 epochs, after which a cosine learning rate decay scheduler is used. Hy-TeC is trained with a batch size of 12, and an image patch size of 16×16 for a maximum of 250 epochs including the warmup phase.

For each model, we kept track of the average RMSE ($\overline{RMSE}$) from the tuning dataset (i.e., 10% of the S1/S2 image patches and corresponding GEDI data) calculated by averaging the RMSE of the estimated heights at 10 m vertical intervals. The training was stopped when $\overline{RMSE}$ stopped decreasing for at least 10 epochs, after which the last model was selected for inference.

4.3. Model performance evaluation

The models were assessed on four criteria: (1) Their generalization capabilities on the set-aside GEDI data (i.e., validation dataset), (2) Their accuracy in comparison to the ALS and VHR-HM maps, (3) Model output resolution, and (4) map qualities.

Comparisons to the set-aside GEDI RH_{98} estimates as well as the ALS dataset were conducted using standard summary statistical metrics such as the correlation coefficient (r), the root mean squared error (*RMSE*), the root mean squared percentage error (*RMSPE*), the bias, the squared difference between standard deviations (SDSD, Kobayashi and Salam, 2000) and lack of correlation weighted by the standard deviations (LCS, Kobayashi and Salam, 2000). The comparison to both the RH_{98}, as well as the ALS and VHR_{HM} measurements was conducted at the resolution of 10 m, which is the target resolution of our models. This will not pose any issues since the ALS and VHR_{HM} are at a much higher resolution and were upscaled to 10 m. However, GEDI measurements are acquired over 25 m wide footprints. As such, they contain height information from ~2×2 pixels from our proposed 10 m height maps. However, photons from GEDI pulses are mostly packed within the center of the footprint, and contribute the most to the measured height profile, with weaker contribution alongside the borders of the footprint. Moreover, over homogenous areas, the measured maximum height at any given ~2×2 pixels within the footprint should be similar. Therefore, in this study, we assumed that a pixel size of 10×10 m should be representative of the measured maximum height by GEDI. *r*, RMSE, RMSPE, SDSD, and LCS are defined in Appendix F.

Next, we evaluated the proposed models based on their output image sharpness index (SI_o, Crete et al., 2007) in comparison to the image sharpness index from the S2 image over the same location (SI_{s2}). The ratio of $\frac{SI_o}{SI_{S2}}$, which we refer to as the Ground Sampling Index (GSI) should provide an indicator of the real ground sampling distance of the different products. While all the models were designed to produce height maps with a pixel size of 10 m, the real resolution might be lower due to several reasons, such as the inherent qualities of the proposed models' building blocks, or the models' attempt to remove noise and artifacts. In our formulation, GSI is calculated over 256×256 images, SI_o is the sharpness calculated from a given model's output while SI_{S2} is the average sharpness index value from the 10 m S2 bands over the same scene. GSI thus ranges from 1 to +∞, with GSI = 1 representing a scene with a GSD identical to S2. Since GSI doesn't scale linearly with the resolution, we calculated several GSI values for several smoothed S2 images at different resolutions ranging from 12.5 m down to 40 m. GSI of images at full resolution (i.e., 10m) as well as down-sampled S2 images can be found in Table 2.

Resolution (m)	10	12.5	15	17.5	20	22.5	25	27.5	30	32.5	35	37.5	40
GSI	1	1.03	1.08	1.14	1.21	1.29	1.37	1.46	1.56	1.68	1.77	1.88	2.00

Table 2. GSI at different resolutions is computed by averaging the SI over 5000 256×256 S2 image patches across different land cover types. S2 image patches at resolutions lower than 10 m were obtained by Lanczos interpolation.

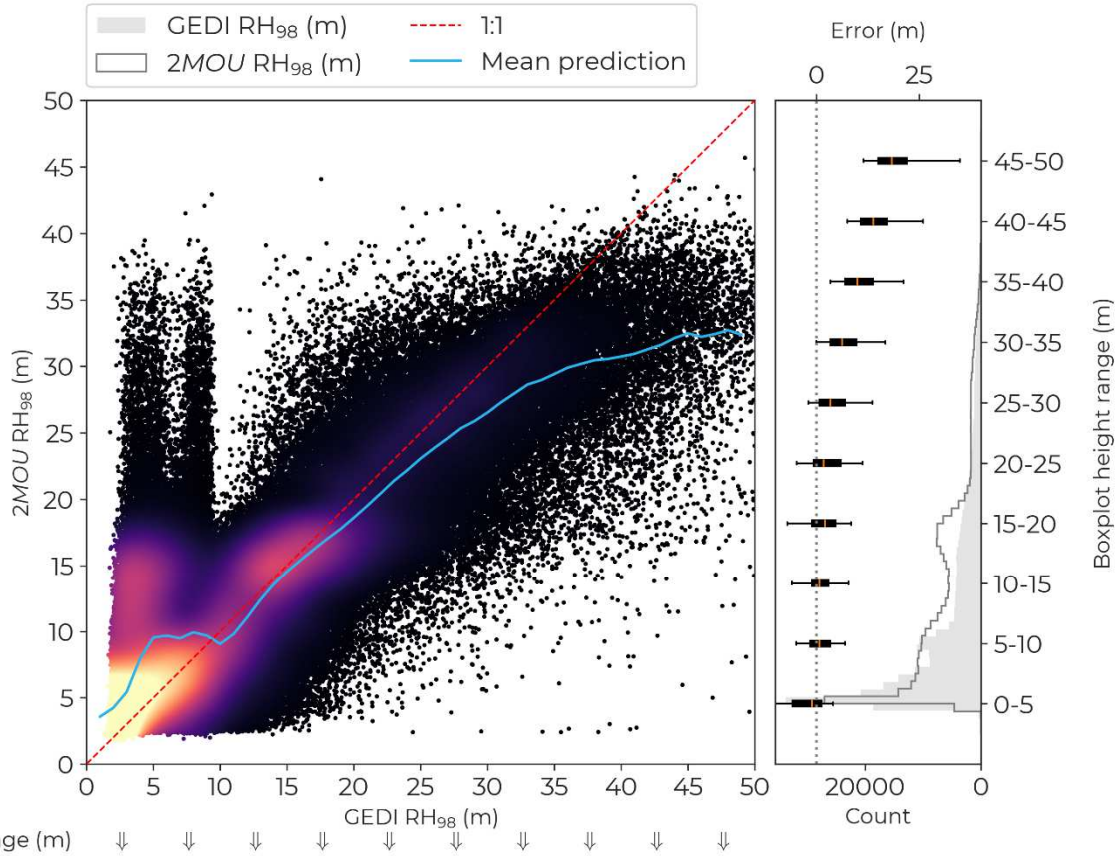
Finally, the resulting height maps from our proposed models were compared to two global height products, the product of Lang et al., 2023a (using S2 as inputs; hereafter referred to as Lang), at 10 m resolution, and that of Potapov et al., 2021 (using Landsat as inputs; hereafter referred to as Potapov) at 30 m resolution. The comparison was based on their accuracy in comparison to the ALS and VHR_{HM} datasets, their GSI, as well as the quality of the output maps. The upscaling of the ALS (respectively VHR_{HM}) from 1 m (respectively 0.7 m) to another resolution is as follows: for a given pixel from an evaluated map, we selected all the ALS pixels (respectively VHR_{HM}) within that pixel, calculated the cumulative density function (CDF) of the CHMs at 0.1 m intervals, and selected the position at 96% of the CDF for comparison with the map pixel. We chose the 96th percentile instead of the 98th as it correlated best when we compared the upscaled ALS pixels (respectively VHR_{HM}) to GEDI's $RH98$.

4.4. Results

4.4.1. Model performance in comparison to GEDI

4.4.1.1. 2MOU performance

The performance of the first tested Multi-Model UNet model 2MOU (UNet with a single continuous output) (Fig. 5), which is optimized using a Huber loss, shows that the model exhibits clear signs of saturation starting at ~30m (Fig. 5, box plots). Indeed, the bias (difference between RH_{98} and 2MOU RH_{98}) increased from 1.6 m for $RH_{98} \in [20m, 30m]$ to 14.7 m for $RH_{98} \in [45m, 50m]$. The results also show that the model, while able to distinguish between tall and short forests (inferred from the behavior of the fit), failed to capture the variability of the canopy heights. In fact, for tall canopies (RH_{98} > 20 m), the model appears to produce an increasing number of estimates centered around the mean height (~30 m). As such, to reduce the overall errors on the tall trees ($RH_{98} > 35\ m$) the model over-estimated the shorter ones ($RH_{98} \in [15m\ , 25m]$). This is observed by the RMSPE that decreases from 24.1% for $RH_{98} \in [15m\ , 20m]$ to 19.8% for $RH_{98} \in [30m, 35m]$ before increasing again for RH_{98} > 35m. For areas with bare-soil/low vegetation and short trees (RH_{98} < 15m) the model appears to estimate GEDI's RH_{98} for the majority of acquisitions with good accuracy. Nonetheless, a significant number of estimates show large discrepancies ($|2MOU_{RH98}$ - $RH_{98}|$ > 15m).

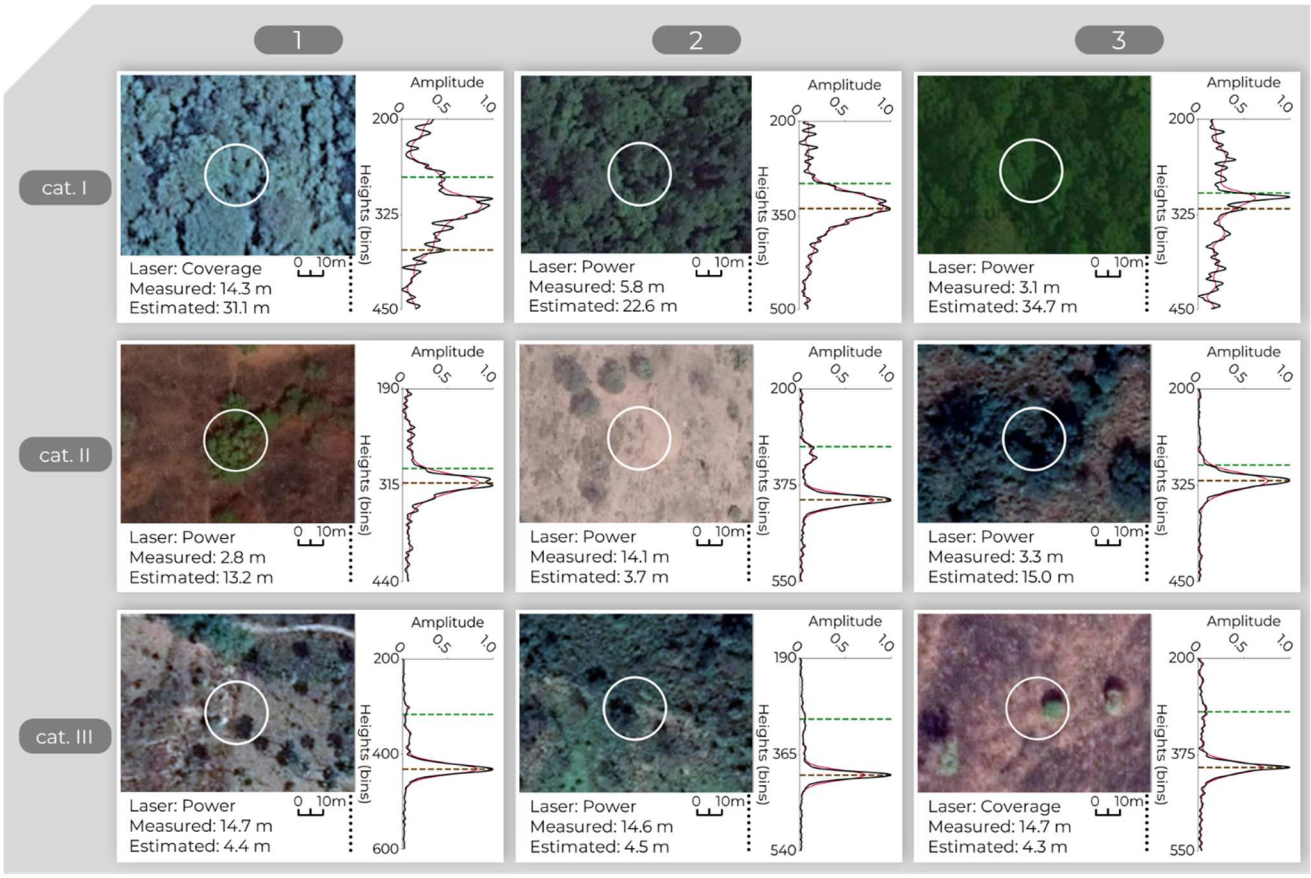


RH_{98} range (m)	⇓	⇓	⇓	⇓	⇓	⇓	⇓	⇓	⇓	⇓
RMSE (m)	5.1	5.9	3.8	4.2	5.0	5.6	6.4	8.7	12.1	15.7
RMSPE (%)	143.6	90.4	30.4	24.1	22.2	20.4	19.8	23.2	28.7	33.3
Bias (m)	-3.0	-2.5	0.8	0.9	1.6	2.6	4.2	7.0	10.8	14.7
LCS*	3.6	13.4	6.0	8.1	9.6	10.5	10.3	11.7	11.9	12.3
SDSD*	13.7	14.1	6.3	6.5	10.1	10.8	9.4	10.0	12.4	11.9
Count	105.5K	70.1K	36.2K	28.7K	16.2K	10.9K	7.8K	4.6K	2.3K	1.0K

Figure 5. Performance of the 2MOU model in comparison to GEDI's set aside RH_{98}. The summary-statistics table is calculated by ranges of RH_{98} (bounded by the x-axis tick positions).

For these estimates (i.e., RH_{98} < 15m), the error sources could mostly be classified into three categories: Category I (cat. I, Fig. 6) represents cases of instrument errors where GEDI failed to produce high-quality waveforms but yet were selected to pass our filters. The poor quality of such waveforms is either due to insufficient penetration (e.g., cat. I.1, Fig. 6), or due to unfavorable acquisition conditions, and as a consequence severely underestimated the canopy height (measured value in Fig. 6). For example, cat. I.2 and I.3 (Fig. 6) show two instances where GEDI estimated the height of tall trees as being lower than 6m, while the model accurately produced high values. Category II errors (cat. II, Fig. 6), represent the case where GEDI acquired high-quality waveforms, but with large errors on the geolocation of the footprint. For example, in Fig. 6 cat. II.1 and cat.II.3 show two waveforms with a single peak (i.e., bare soil/low vegetation) while the footprint's (wrong) location shows trees. In contrast, Fig. 6 cat.II.2 represents a waveform with two distinct peaks (indicating a tree), while the footprint location is over bare soil. Category II errors, concern mostly footprints acquired over Savannas, where the tree cover is sparse. Finally, Category III errors (cat. III, Fig. 6) represent the case where GEDI accurately

580 measured the vertical structures within the footprint with low errors on geolocation (through visual inspection),
581 however, given the small size of the canopies (characterized by a low vegetation peak in the waveform) and the
582 low resolution of S2 and S1 (10 m), the model was unable to accurately predict the height of these trees, but rather
583 it predicted the average height within the 10 m pixel.
584

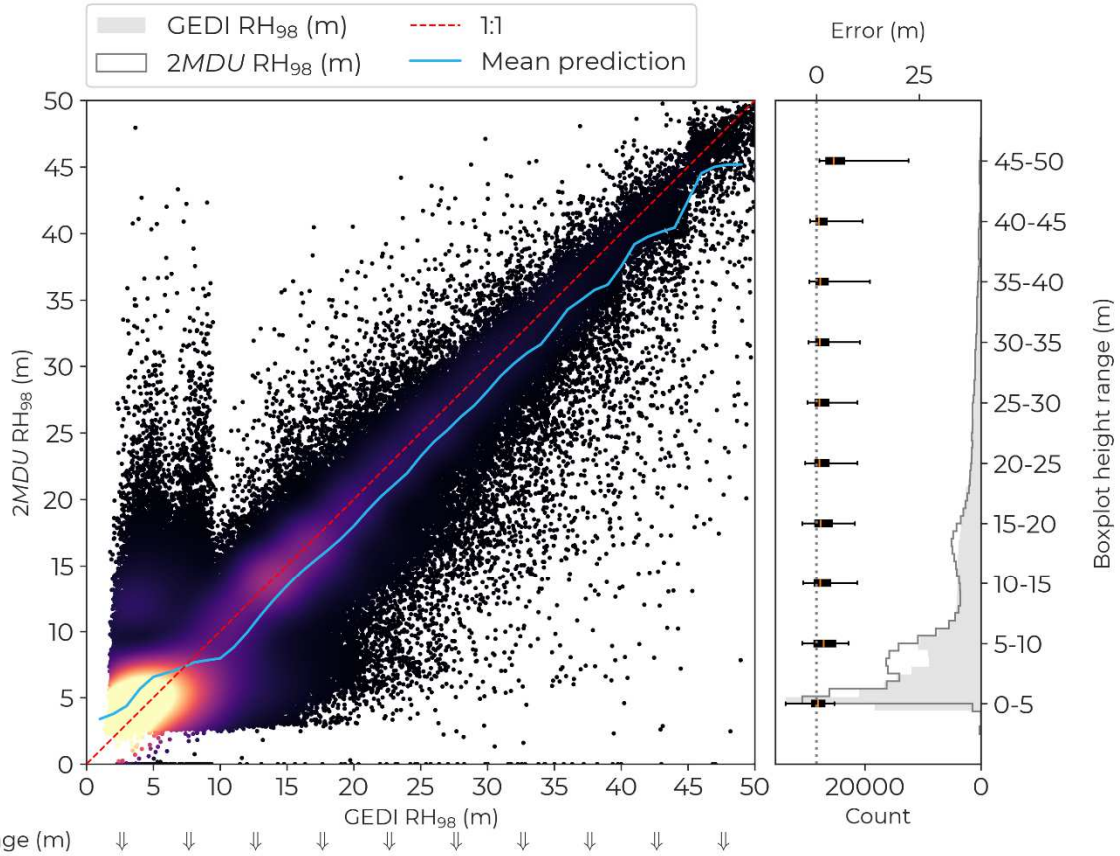


585

586 Figure 6. Example of several GEDI waveforms (black line) across different landscapes where the 2MOU model produced height estimates with
587 large differences in comparison to GEDI's RH_{98}. Footprint location is shown in white; detected ground and canopy top returns derived from the
588 smoothed waveform (red) are respectively shown in brown and green.

589 *4.4.1.2. 2MDU performance*

590 The 2MDU model employs the same encoders and decoder as the 2MOU model, however, it has a modified
591 output head that constrains the estimated outputs to ten overlapping height ranges before regressing the height
592 estimate. This translates to a high increase of accuracy across all height ranges in comparison to the 2MDU model
593 (Fig. 7). Indeed, performance results of the 2MDU show that the model no longer saturates over tall trees (RH_{98} >
594 30m, Fig. 7 boxplots), and overall has low RMSE that ranges between 3.0 m (RH_{98} < 5m) to 5.7 m (RH_{98} > 45 m).
595 Moreover, results also show that both the LCS, and SDSD are mostly similar across all height ranges, indicating
596 that the model was able to capture both the magnitude and pattern of the fluctuations to the same degree for each
597 height range. In terms of bias, the model appears to overestimate RH_{98} values smaller than 10 m, and
598 underestimate those higher than 10 m. However, for RH_{98} > 10 m, the bias is mostly constant and ranges between
599 1.7 and 2.5 m (Fig. 7).



RH_{98} range (m)	⇓	⇓	⇓	⇓	⇓	⇓	⇓	⇓	⇓	⇓
RMSE (m)	3.0	3.6	4.1	4.2	4.5	4.5	4.6	5.1	5.7	5.7
RMSPE (%)	88.6	53.0	32.9	24.6	20.1	16.4	14.3	13.5	13.4	12.0
Bias (m)	-1.5	-0.1	1.9	1.7	1.9	1.8	1.9	2.1	2.5	2.2
LCS*	2.6	7.8	5.9	7.1	7.0	6.9	6.6	7.0	8.5	7.9
SDSD*	4.0	4.6	5.9	5.7	5.8	4.9	4.2	4.0	5.5	4.1
Count	105.5K	69.8K	34.2K	28.1K	16.2K	10.7K	7.5K	4.7K	2.6K	1.2K

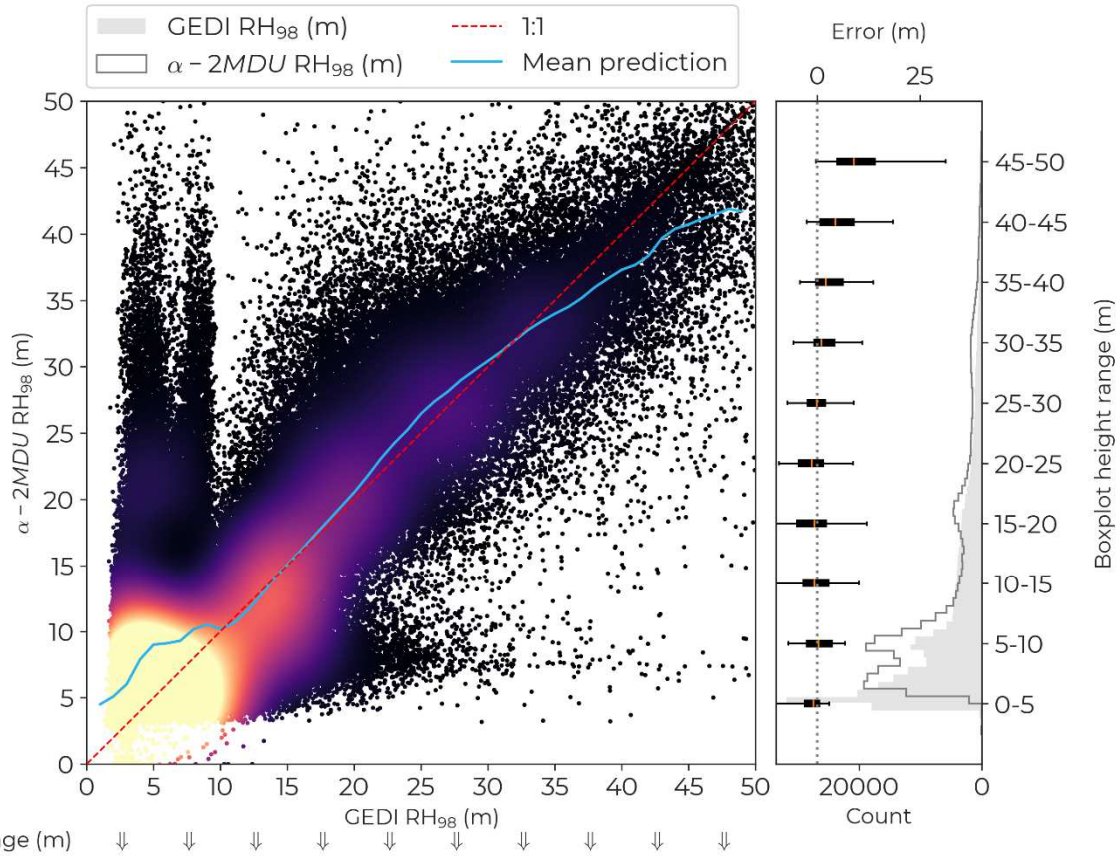
Figure 7. Performance of the 2MDU model in comparison to GEDI's set aside RH_{98}. The summary-statistics table is calculated by ranges of RH_{98} (bounded by the x-axis tick positions).

Nonetheless, as this model learned to better fit the GEDI acquisitions, it also learned to fit well the outliers. This is especially true for the category I errors (Fig. 6, cat.I) where the RMSE decreased from 5.1 and 5.9 m for respectively $RH_{98} \in [0\,m, 5\,m]$ and $RH_{98} \in [5\,m, 10\,m]$ (Fig. 5) to respectively 3.0 and 3.6 m (Fig. 7). As a consequence, the mapping capabilities of the 2MDU model degraded as it started predicting the height of some forest patches as being bare soil/low vegetation (Fig. G1).

4.4.1.3. α-2MDU performance

The α-2MDU model shares the same architecture as the 2MDU model, however, the main difference between the two models lies in the cost function associated with the regression part. The 2MDU uses the Huber loss, while the α-2MDU employs the adaptive loss which reduces the sensitivity of the outliers as training progresses (Fig. H1). As a consequence, the accuracy of the α-2MDU model in comparison to the set-aside GEDI dataset is lower than the 2MDU model across all height ranges (higher RMSE and RMSPE in general). However, the difference in

accuracy (ΔRMSPE) between the two models ranges from 1.2 to 10.8 % for $RH_{98} \geq 10$ m. But the advantage of α-2MDU is that the bias is reduced for $RH_{98} \in [10\ m, 40\ m]$ in comparison to the 2MDU model. For heights above 40 m, the model appears to underestimate taller trees with a bias of respectively 3.4 (RMSE = 7.0 m) and 5.7m (RMSE = 9.1 m) for $RH_{98} \in [40\ m, 45\ m]$ and $RH_{98} \in [45\ m, 50\ m]$.



RH_{98} range (m)	⇓	⇓	⇓	⇓	⇓	⇓	⇓	⇓	⇓	⇓
RMSE (m)	5.5	6.1	4.5	6.1	6.5	5.9	5.0	5.7	7.0	9.1
RMSPE (%)	159.1	92.5	35.9	35.4	29.5	21.6	15.5	15.2	16.6	19.2
Bias (m)	-3.3	-2.5	0.2	-0.3	-1.1	-1.1	0.1	1.8	3.4	5.7
LCS*	4.7	13.1	8.6	12.3	12.6	11.8	9.5	10.0	11.1	14.3
SDSD*	14.1	17.8	11.0	22.5	23.5	13.9	8.4	10.0	14.0	19.7
Count	102.7K	72.7K	35.5K	27.4K	15.9K	10.5K	8.1K	5.0K	2.7K	1.3K

Figure 8. Performance of the α-2MDU model in comparison to GEDI's set aside RH_{98}. The summary-statistics table is calculated by ranges of RH_{98} (bounded by the x-axis tick positions).

Finally, as the α-2MDU model is less sensitive to outliers, especially Cat. I errors due to GEDI's lower quality data (Fig. 6), the α-2MDU model no longer shows artifacts in comparison to the 2MDU model over dense tropical forests (Fig. 9).

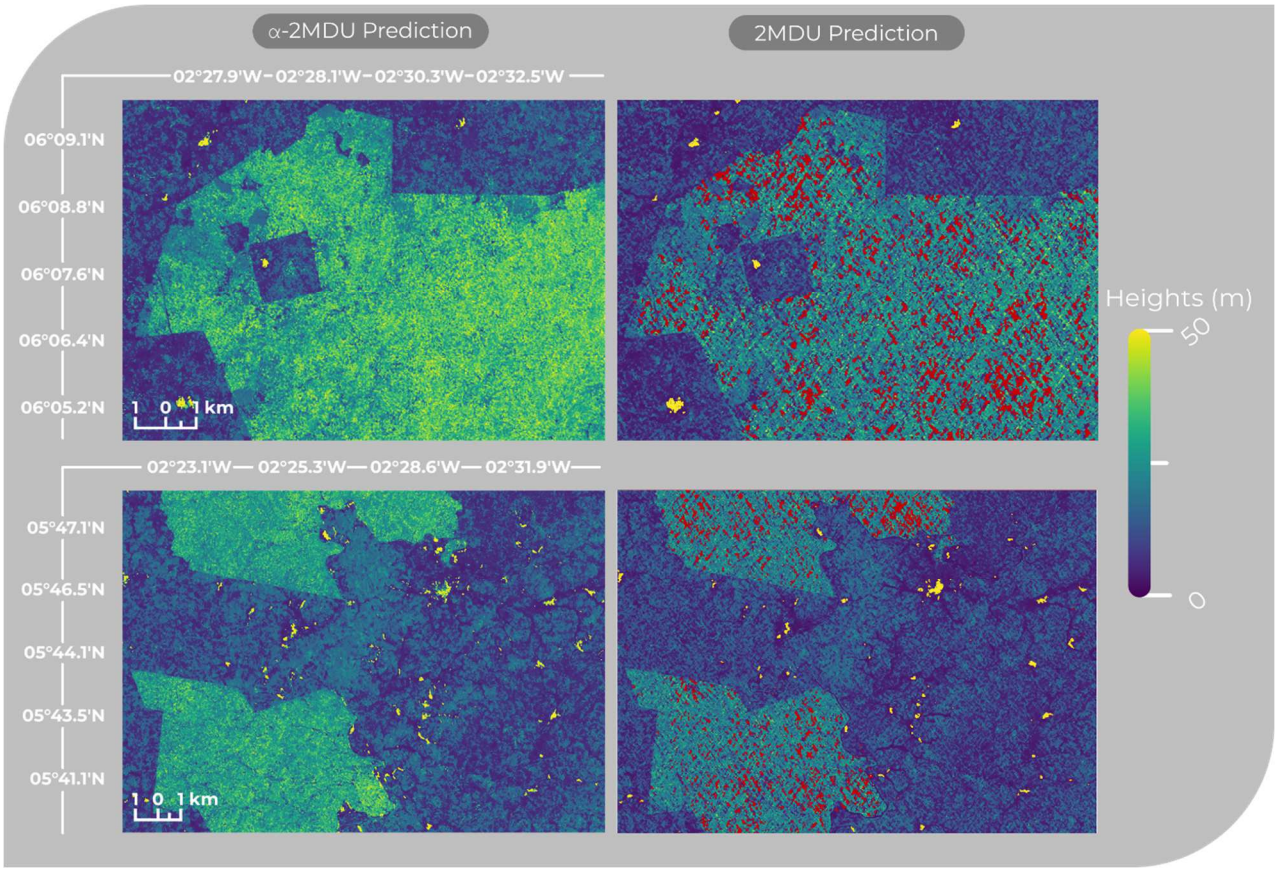
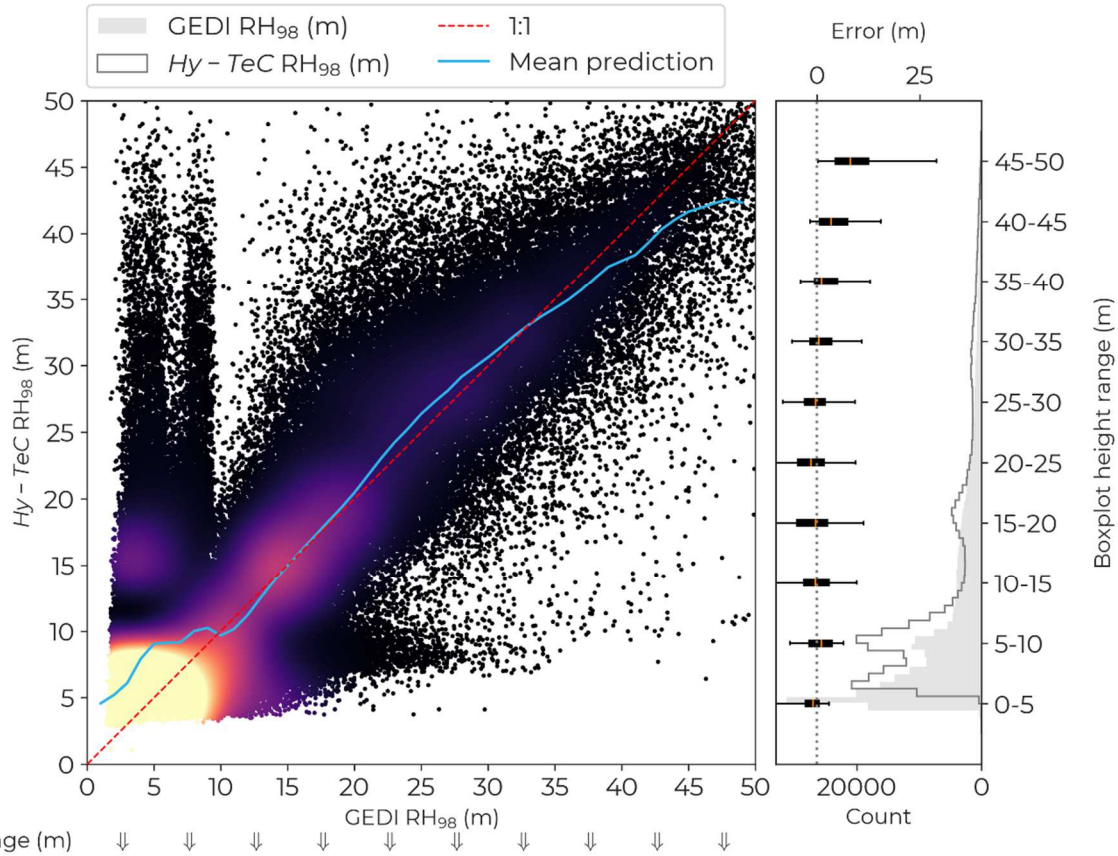


Figure 9. Comparison between the predictions of the α-2MDU and 2MDU models over two areas with tall trees. The red polygons represent forested areas where the 2MDU model estimated the heights of tall trees as being less than 2m.

4.4.1.4. Hy-TeC Performance

The Hy-TeC model which utilizes a transformer encoder and a convolutional decoder and the same output head and loss functions for its main output (10 m) as the α-2MDU model, shows improved accuracy than α-2MDU (2MDU was not considered due to its degradation over forests), especially over tall trees in comparison to the set-aside GEDI dataset (Fig. 10). For RH_{98} < 30 m, both Hy-TeC and α-2MDU seem to be on par, with almost the same performance across all metrics except for the LCS and SDSD where Hy-TeC shows slightly lower LCS and higher SDSD. This indicates a marginally lower correlation and higher uncertainty than α-2MDU (Fig. 10 vs. 8). For $RH_{98} \in [30\,m,\ 35\,m]$, α-2MDU has 0.4 m lower RMSE than Hy-TeC. However, α-2MDU has higher bias than Hy-TeC and lower LCS and SDSD, indicating some slight underestimation, lower correlation but less uncertainty. For RH_{98} > 35 m, Hy-TeC shows better performance that α-2MDU, and the difference in performance between the two models increases with increasing tree height. Indeed, the ΔRMSE increases from 0.3 m to 1.2 m for tree heights increasing from 35 m to 50 m, with Hy-TeC showing lower RMSE. Moreover, for this height range (between 35 and 50 m), Hy-TeC has lower bias and lower SDSD, but slightly higher LCS than α-2MDU.



RH_{98} range (m)	⇓	⇓	⇓	⇓	⇓	⇓	⇓	⇓	⇓	⇓
RMSE (m)	5.4	6.0	4.3	5.9	6.6	6.2	5.4	5.4	6.3	7.9
RMSPE (%)	157.9	90.4	33.9	34.2	29.7	22.7	16.9	14.4	14.9	16.6
Bias (m)	-3.4	-2.4	0.6	-0.2	-1.0	-1.1	-0.1	1.2	2.7	4.9
LCS*	4.5	13.1	7.8	11.9	12.9	12.7	10.6	9.8	9.8	12.6
SDSD*	13.2	16.7	9.7	20.7	24.6	16.8	11.4	10.1	11.2	11.7
Count	101.0K	70.2K	33.8K	26.7K	15.7K	10.4K	7.8K	4.9K	2.8K	1.2K

Figure 10. Performance of the Hy-TeC model in comparison to GEDI's set aside RH_{98}. The summary-statistics table is calculated by ranges of RH_{98} (bounded by the x-axis tick positions).

4.4.2. Comparison to ALS and VHR-HM

The analysis made in the previous sections showed that two models presented better performances overall than the others: α-2MDU and Hy-TeC. Moreover, while 2MDU has better performance metrics than both α-2MDU and Hy-TeC, generated maps by 2MDU showed major artifacts, especially over dense forests, as such it was deemed unusable. Therefore, only α-2MDU and Hy-TeC were selected in this evaluation section.

The performance of our two best models, namely the α-2MDU and Hy-TeC, were evaluated against the independent ALS campaign (Stereńczak et al., 2020) we upscaled to a resolution of 10 m to coincide with the theoretical resolution of our output maps (see section 4.3). For comparison purposes, we also evaluated three other datasets to the ALS dataset. The first dataset represents the GEDI acquisitions over the ALS area where we scaled the ALS resolution to 25 m to coincide with the 25 m wide footprints. The second and third datasets are the 10 m

global height map from Lang et al., 2023a, and the 30 m global height map from Potapov et al., 2021, respectively (ALS upscaled to 30 m for Potapov).

The results presented in Fig. 11.a show that compared to the ALS dataset, the GEDI acquisitions, which were used as calibration data for the four models, underestimated the ALS measurements by 3 m (bias), with an RMSE of 8.9 m and a correlation coefficient (r) of 0.18. Separating the GEDI acquisitions into Coverage (Cov) and Full Power (FP) acquisitions, the results show that acquisitions from the FP lasers are generally more accurate, with an RMSE of 7.9 m (r=0.30), while the RMSE for the Cov acquisitions is 9.6 m and r = 0.17. Although both the FP and Cov acquisitions underestimated the ALS measurements, the FP showed a lower bias (2.2 m) compared to the Cov (bias= 5 m).

The results of the modeled heights from the four products show that Potapov's 30 m map is the least accurate one with a RMSE of 10.0 m, a correlation coefficient (r) of 0.11, and a bias of 6.0 m (Fig. 11.b). The 10 m height map of Lang shows better accuracy in comparison to Potapov, with a lower RMSE (8.6 m), higher correlation coefficient (r=0.18), and lower bias (4.9 m) (Fig. 11.c). However, the results also show that both height products are unable to produce height estimates above 40 m over this region.

Finally, the results show that our two proposed best models, α-2MDU and Hy-TeC (respectively Fig. 11.d and Fig. 11.e), are the most accurate ones, with Hy-TeC showing significantly better performance with an RMSE of 6.6 m (r=0.39) and almost no bias (underestimation of 0.8 m) while α-2MDU shows an RMSE of 8.3 m (r=0.22, bias= 3.3 m). Moreover, unlike Potapov's and Lang's products, α-2MDU and Hy-TeC models do not appear to saturate for tall trees as both were able to produce canopy height estimates higher than 45 m.

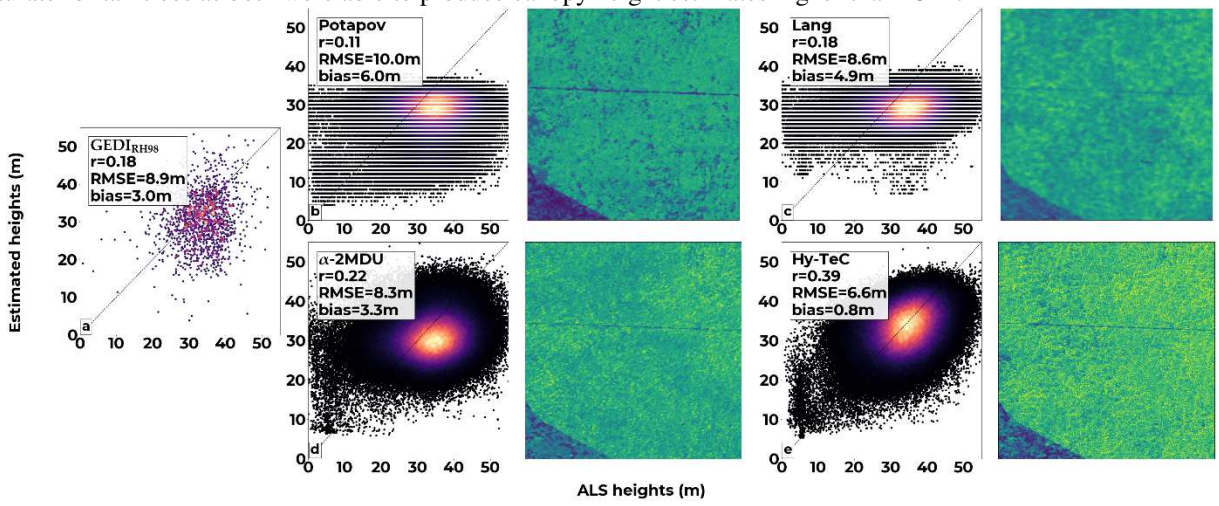


Figure 11. Comparison of estimated heights from (a) GEDI, (b) Potapov, (c) Lang, (d) α-2MDU, and (e) Hy-TeC *vs* the reference ALS data set.

In the evaluation *vs* the reference VHR_{HM} data sets over sparsely vegetated areas, GEDI's RH_{98} measurements show better accuracy in comparison to the reference dataset than over densely vegetated areas with a correlation coefficient of 0.48, an RMSE of 3.9 m, and a bias ($VHR_{HM} - RH_{98}$) of -1.3 m (Fig. 12.a). Regarding model performance (Fig. 12.b, Fig. 12.c, Fig. 12.d, and Fig. 12.e), Hy-TeC shows the highest accuracy on the estimation of trees in a sparsely vegetated context with a correlation to the reference dataset of 0.56, a RMSE of 3.1 m and bias of -1.6 m (Fig. 12.e). α-2MDU's performance (Fig. 12.d), while marginally better than Hy-TeC in terms of correlation to the reference dataset, the accuracy on the estimation of the heights with α-2MDU was significantly lower than Hy-TeC, with a RMSE of 4.3 m and a bias of -3.4 m. Regarding the global products, the Potapov product at 30 m (Fig. 12.b) showed almost no bias to VHR_{HM}, however, their product has low correlation (r=0.29) and high RMSE (6.8 m). Finally, for Lang product (Fig. 12.c), while the height estimates showed a correlation to the reference dataset (r=0.49) that is comparable to what is obtained with GEDI (r=0.49), their product has a very high bias (-8.1 m, overestimation), which in turn decreased significantly the accuracy of their estimates (RMSE=8.6 m).

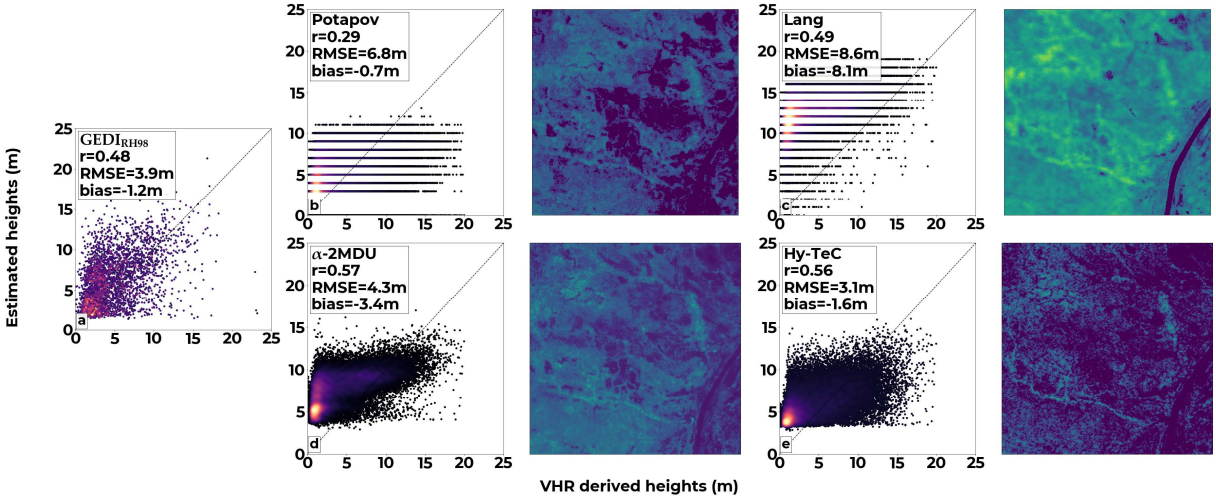


Figure 12. Comparison of estimated heights from (a) GEDI, (b) Potapov, (c) Lang, (d) α-2MDU, and (e) Hy-TeC *vs* the reference VHR_{HM} data set.

4.4.3. Height maps quality assessment

In this section, we present a comprehensive evaluation of four different canopy height map products, including our two best-performing models (α-2MDU and Hy-TeC) and two global products (Potapov and Lang). The evaluation is divided into two parts: first, we assess the ground sampling index (GSI) of each product to evaluate their real output resolution, and second, we conduct a qualitative assessment of the maps to assess their overall quality.

Regarding the ground sampling index (GSI, section 4.3) of the different products, results presented in Fig. I1 show that Hy-TeC, with an output pixel size of 10 m, stands out with the lowest GSI. In fact, around 80% of its patches have a GSI lower than 1.08, comparable to S2 patches with a resolution between 10 and 15 m. α-2MDU follows with a slightly higher overall GSI (58% of patches have a GSI lower than 1.08), but still performs significantly better than Lang (average GSI of 1.21 comparable to S2 patches with a resolution of 20 m) and Potapov models (80% of patches have comparable GSI to S2 patches with a resolution of 30 m, which is the reported resolution).

The higher resolution of our proposed model (i.e., Hy-TeC) and that of the baseline (i.e., α-2MDU) can be clearly seen in the resulting height maps. For example, over densely forested areas (Fig. 13), both Hy-TeC and α-2MDU clearly show higher ground sampling distances and this is apparent from the 3D renders, with denser predictions over the forest patches (Fig. 13, 3D Render). Moreover, in comparison to the distribution of heights from the GEDI acquisitions (Fig. 13, Height distributions), both, Hy-TeC and α-2MDU show a higher capacity to predict both low (bare soil/low vegetation) and high values (> 50 m). In contrast, Potapov's model appears to saturate above 35 m, while the Lang model shows height estimates ranging between 10 m (no low values) and 45 m.

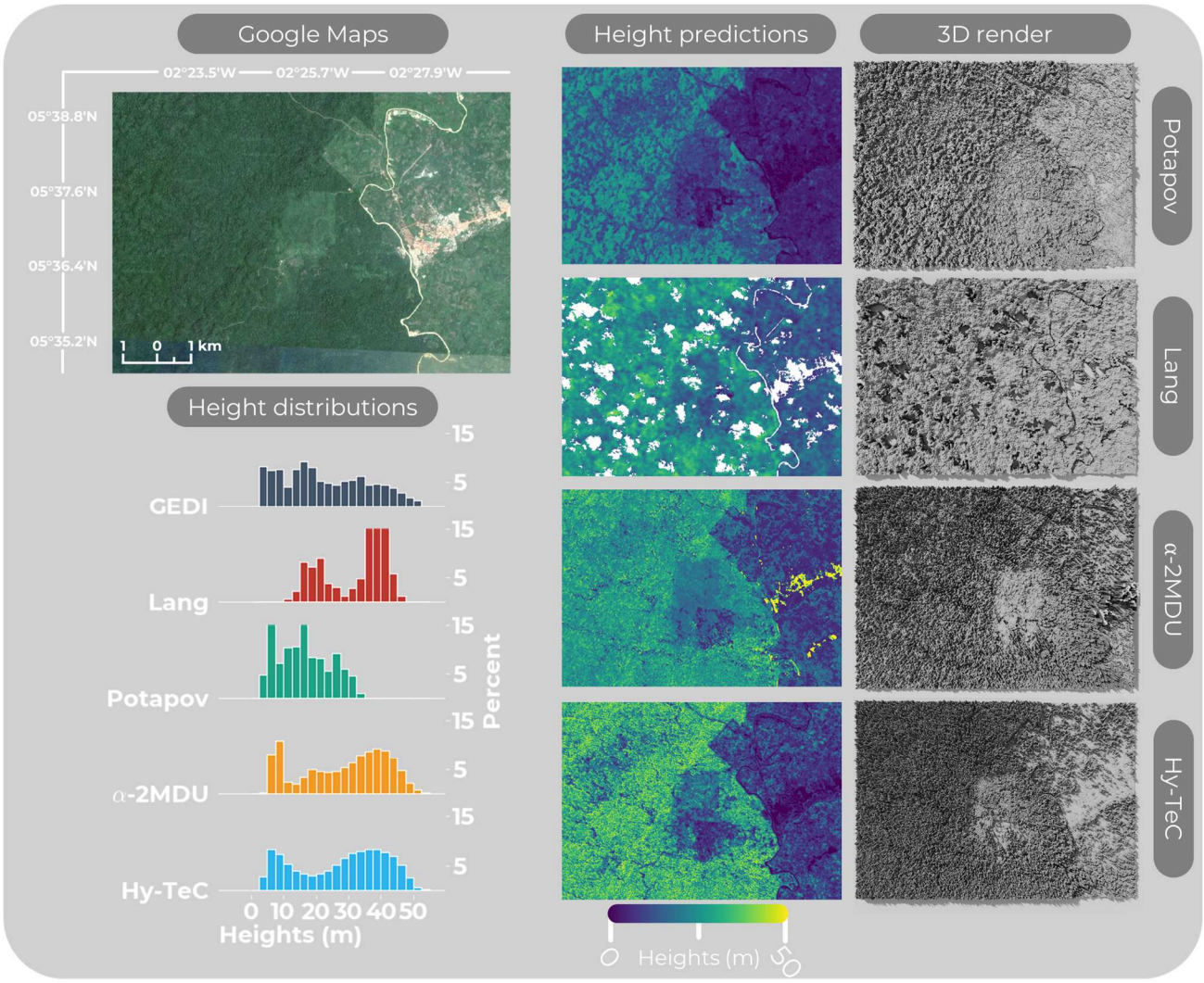
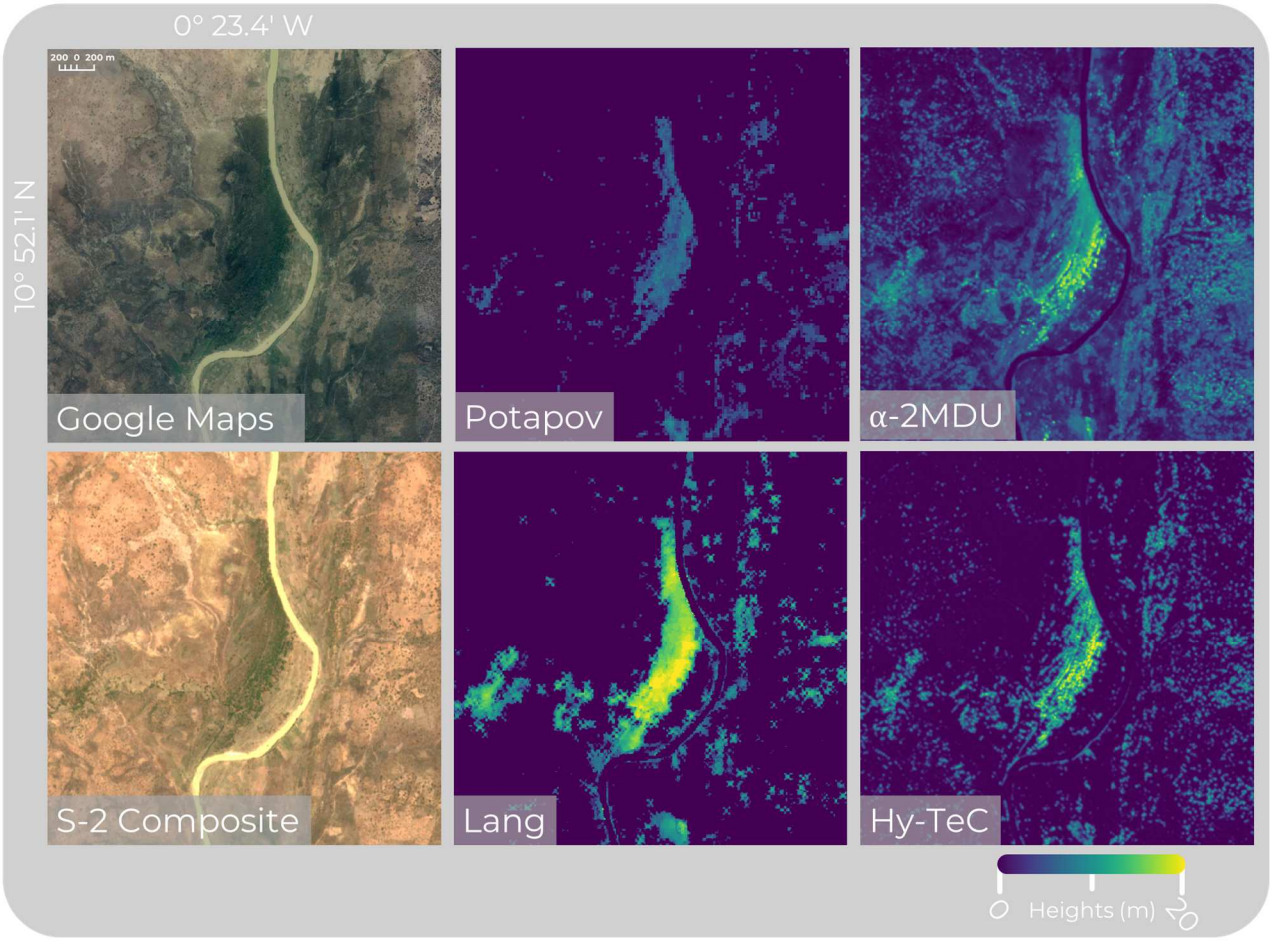


Figure 13. Canopy height predictions (Height Predictions) from GEDI estimates and the four compared models (Potapov, Lang, α-2MDU, and Hy-TeC) over an area with predominant forest areas. 3D Render is a 3D visualization of the canopy height estimates. The Height distributions represent the distribution of the canopy heights from each of the four products as well as from the available GEDI acquisitions.

Nonetheless, despite the overall agreement between the α-2MDU and Hy-Tec maps, there appear to be some differences between the two models. First α-2MDU shows slightly lower estimates for tall trees and higher estimates for short heights (< 10 m) than Hy-TeC. The higher ground sampling distance is also apparent, both in the 3D renders (Fig. 13), as well as the fine trails within the forest patch (Fig. 13, Height predictions) that are better captured by Hy-TeC than α-2MDU.

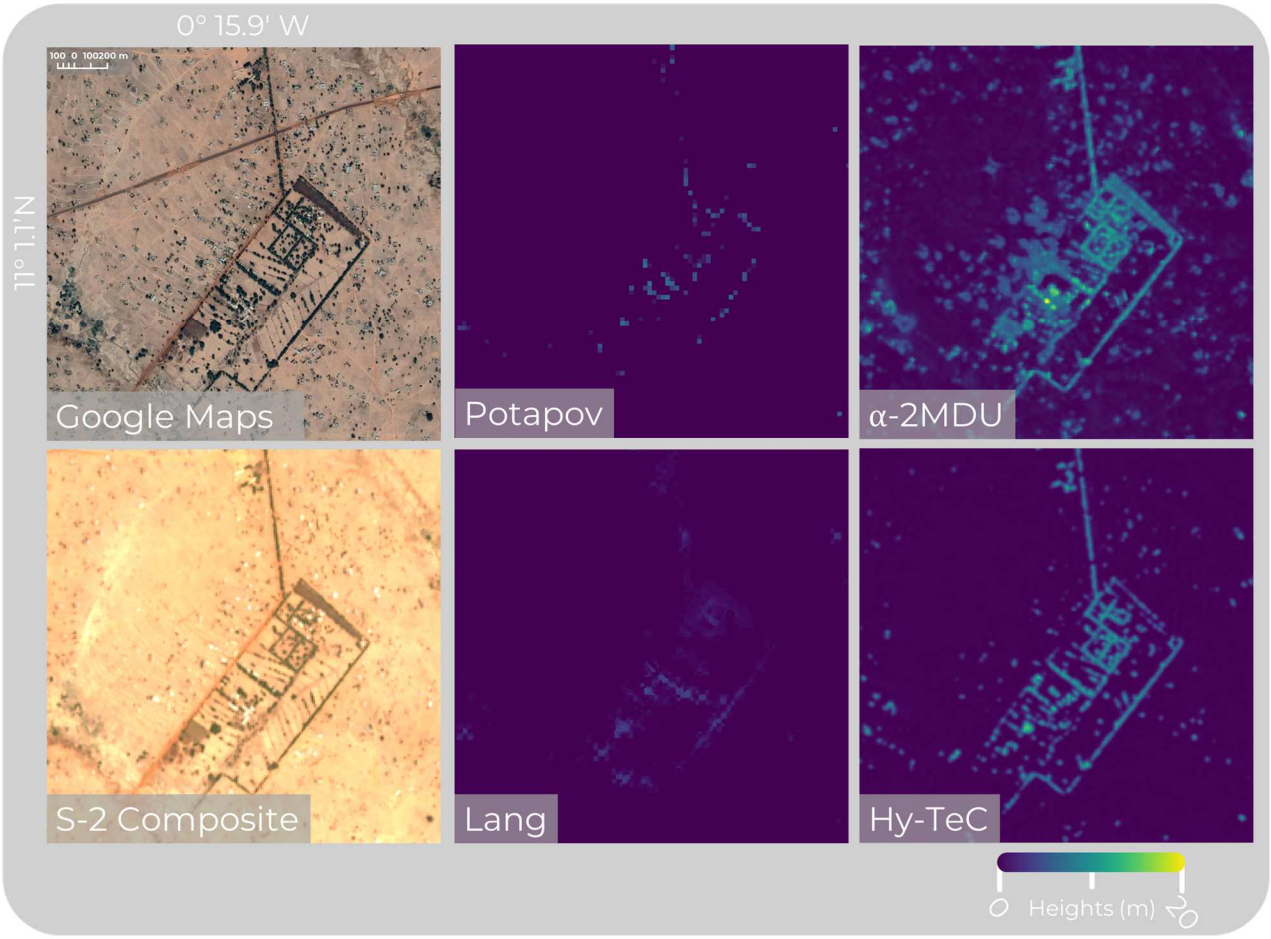
Over lightly forested areas, Fig. 14 shows the improved height estimation capabilities of Hy-TeC in comparison to the other three products. Overall, Potapov's model shows a lower estimation of canopy heights than the three other models. Nonetheless, both Potapov and Lang show estimates which are far smoother than our proposed models. Moreover, both Potapov and Lang also seem to miss most of the short and scattered trees. Finally, while α-2MDU does show some similarities to Hy-TeC, it seems to overestimate the vegetation cover in some areas.



729

730 Figure 14. Illustration of the canopy height predictions from the four compared models over an area with sparse tree cover.

731 The differences between our proposed models and existing products are more apparent in more open areas.
732 Indeed, while α-2MDU and Hy-TeC appear to estimate the heights of most trees in Fig. 15, the models of Potapov
733 and Lang only estimated the heights of the largest ones. Moreover, in this case, α-2MDU also seems to
734 overestimate the tree cover in some areas.
735
736



737

738 Figure 15. Illustration of the canopy height predictions from the four compared models over planted trees.

739 Over urbanized areas, Hy-Tec seems to perform the best (Fig. 16). Indeed, Hy-TeC accurately removed all
740 built-areas from its height predictions, and was still capable of predicting tree heights from isolated trees and forest
741 patches. The model of Potapov also seems to perform well in urban areas; however, it produces heights at coarser
742 resolutions. Finally, for α-2MDU, while the predictions appear to resemble those of Hy-TeC over forested areas
743 within the urban scene, it shows two distinct differences in comparison to Hy-TeC. First, α-2MDU appears to
744 provide in general higher height estimates for forested areas, and over built-up, it predicts very high values with no
745 particular patterns. This also seems to be the case for the Lang model (Fig. 16), as the white spots indicate that
746 urban areas have been masked in Lang's product in post-process.

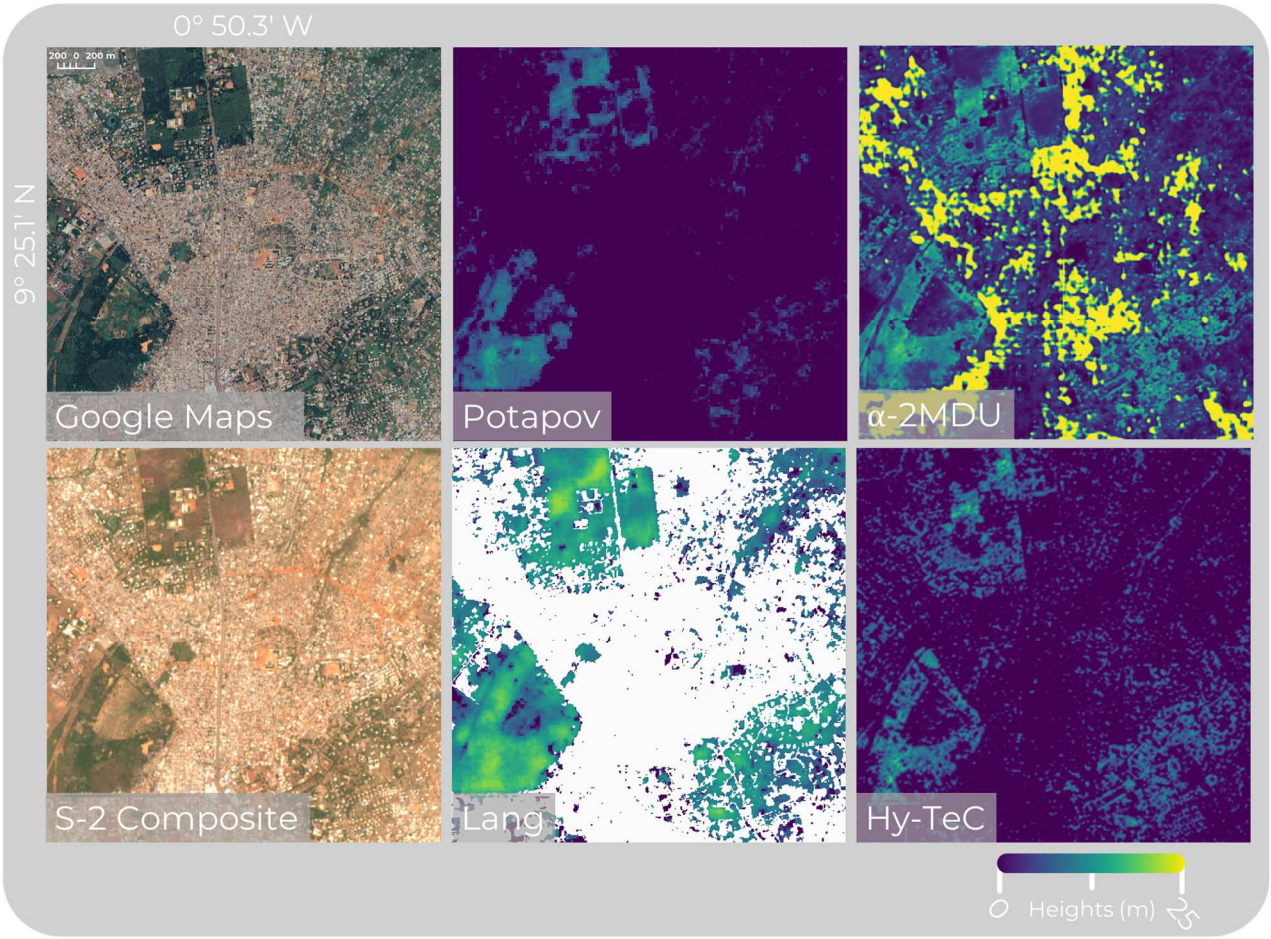


Figure 16. Illustration of the canopy height predictions from the four compared models over urban areas.

Finally, while the four models (Potapov, Lang, α-2MDU, and Hy-TeC) do agree to some extent at the national scale for very large-scale patterns of tree heights (Fig. 17), they show some major regional differences that are consistent with the results presented earlier across the different landscapes. The first major difference is the range of heights of the tall trees that the different models are able to reach (Fig 17, south of Ghana). Here, the Potapov product seems to reach a maximum of ~35 m, while the other three products can estimate a significantly higher height (more than 45 m). However, similarly to the results in Fig. 13, over tall trees, the maps from both α-2MDU and Hy-TeC have on average higher and lower bounds than those provided by Lang except for sites designated A and B in Fig. 17. For these two sites, the Lang product appears to show on average higher tree estimates than the three other products. By comparing the estimates from sites A and B to GEDI RH_{98} measurements (only RH_{98} estimates > 15 m were considered), we observe that Lang's product is overestimating the heights by 2.3 m in comparison to RH_{98}, while the comparison between GEDI's RH_{98} and the estimates from respectively Hy-TeC and α-2MDU showed a bias of -0.05 m (slight underestimation) and 0.1 m (slight overestimation). For areas with lower tree heights (e.g., Savannas, sparsely vegetated areas), the Lang product shows signs of overestimation across the country (i.e., lighter colors indicate higher heights). For Potapov and α-2MDU, at the national scale, the results are mostly consistent between the two products. Regarding Hy-TeC, it shows the highest contrast at the national scale in comparison to the three other products. Moreover, Hy-TeC shows in general lower height estimates for areas with sparse and short vegetation. This is a direct result of the significantly higher output resolution of the Hy-TeC model in comparison to previous approaches, allowing it to estimate the heights at the tree level in some instances. In contrast, the other approaches, and due to their coarser output resolution, appear to overestimate the tree cover (Fig. J1, darker is higher).

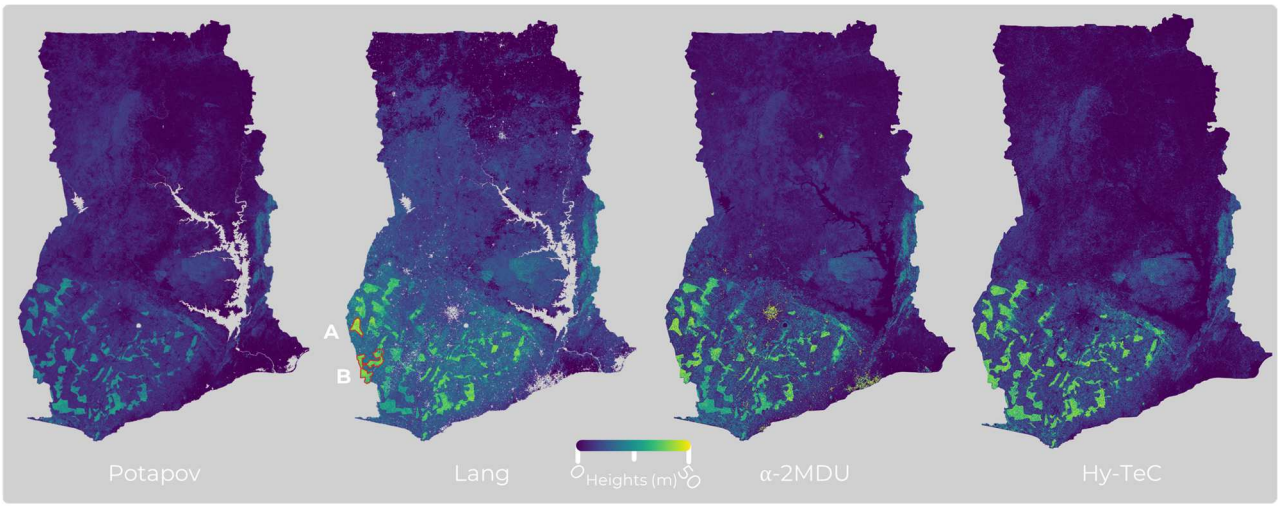


Figure 17. National Scale comparison between the four canopy height models.

5. Discussion

In this study, we presented Hy-TeC, a novel hybrid model that employs a transformer encoder and a ConvNet decoder, and evaluated its performance against a baseline model called α-2MDU based on the U-Net architecture. We found that Hy-TeC which uses as input S2 acquired imagery and calibrated with GEDI acquisitions provides significant improvements over current state-of-the-art approaches both quantitatively and qualitatively. Specifically, evaluated against the set-aside GEDI dataset, ALS dataset, and the VHR-HM dataset, our model has shown significantly reduced saturation on the height estimates of tall trees, and due to its higher output resolution, allows better detection and height estimation of trees in densely and sparsely vegetated areas. The findings regarding Hy-TeC's performance, however, are not site dependent, and are reproducible elsewhere. Indeed, over Gabon (refer to Appendix K), the generated canopy height map showed that similarly to Ghana, the vertical sampling range of Hy-TeC is larger than previous products, namely Lang's and Potapov's, indicating that Hy-TeC is able to better capture the structure of short and tall vegetation. Moreover, Hy-TeC, on average, also produces higher values for the tall vegetation in comparison to previous products. This finding is corroborated by the comparison made between the available height products and LVIS acquisitions over four Sentinel-2 tiles in Gabon. Quantitavley, the results presented in Table K1 showed that Hy-TeC outperforms the previous products in all error metrics, with a difference in accuracy that increases with increasing canopy height.

Finally, while α-2MDU has shown competitive performances in some cases in comparison to Hy-TeC, it showed some artifacts over urban areas, a tendency to overestimate the tree cover in sparsely vegetated areas, and overall smoother maps than Hy-TeC. However, it is important to acknowledge that each of the two models has its uncertainties and limitations that are influenced by various factors, including the pre-processing of the S1 and S2 images, the quality of GEDI data, model architecture, and the loss functions. In this section, we will discuss the impact of these factors on model uncertainty and explore possible approaches to further reduce it, ultimately improving the accuracy of the predictions.

5.1. S1 and S2 image pre-processing

Our main inputs for the deep learning models were S1 and S2 image mosaics acquired over 15 months. Mosaicking presents advantages as well as drawbacks. In terms of advantages, especially for optical imagery, a mosaic for a given period allows the mitigation of several problems including cloud contamination, surface directional reflectance, and view and illumination geometry (Flood, 2013). This proved beneficial in our case,

especially over the south of Ghana, a cloudy region where our models were able to produce continuous and artifact-free canopy height estimates over the dense tropical forests.

In the literature, several techniques were proposed to create a mosaic from a times series, the maximum NDVI composite (MNC; Holben, 1986; Huete et al., 2002; Roy et al., 2010), medoid composites (MC), a multi-dimensional median technique (Flood, 2013; Kennedy et al., 2018; Liang et al., 2023), or simply the median (Liang et al., 2023; Schwartz et al., 2022). Here we decided to use the median instead of MNC or MC for several reasons. First, one of the major benefits of the MCN technique is its capacity to reduce cloud contamination and cloud shadows. However, with the advancement of cloud and cloud shadow detection techniques (Goodwin et al., 2013), and the fact that MCN will produce mosaics biased towards high NDVI values, potentially reducing the contrast between forests and grasslands, the MCN was not adopted. Second, the MC technique, in contrast to the median, will select a single date for each pixel as the representative of the time series, as such, will better preserve the relationship between the bands, whereas the median method will calculate a median value for each band independently. On the other hand, the MC is more computationally expensive than the median, and while band correlation might be higher with MC, we don't believe this should have a significant impact on the height estimates as our models do not rely on calculated vegetation indices, but rather on raw reflectance and backscatter values used as input to our models. Regarding the S1 data, while mosaicking could help reduce speckle noise, in the general case, it is not necessary as SAR instruments have an all-weather operability. However, due to terrain changes that might have occurred during the 15 months, we also created a median mosaic for S1 data to stay consistent with S2 data.

In terms of drawbacks, a mosaic over a time period, especially over long periods can potentially mask large events within a landscape, which might not be desirable depending on the target application. For example, for forest monitoring applications, events such as fires or logging, especially those taking place at the end of the considered period for the mosaic, will not be detected until later. For agricultural applications, where planted crops are usually rotated each season, a mosaic of more than a season will mix information from different types of planted species, which is not desirable.

Finally, compositing, especially when spanning multiple leaf phenological cycles, can introduce various effects, particularly in forests with different phenological patterns, such as deciduous and coniferous forests. On one hand, creating a composite that encompasses multiple phenological cycles may improve the differentiation between these forest types, leading to more accurate canopy height estimates. For instance, a median composite that covers both leaf-off and leaf-on seasons may yield higher reflectance values for coniferous trees compared to deciduous ones. In models that solely rely on reflectance values, like Random Forest (RF) models without incorporating additional textural information, this could potentially lead to underestimation of canopy height over deciduous forests.

However, deep learning models, especially those with feature-rich architectures such as Convolutional Neural Networks (ConvNets) and Vision Transformers (ViTs), differ in their approach. These models leverage not only raw reflectance values but also image structural features for making predictions. As a result, they can mitigate the bias arising from differential reflectance and provide more robust canopy height estimates.

In the case of mixed forests, where the model might struggle to distinguish between mixed, coniferous, or deciduous forests, there's the potential for over- or under- estimation of canopy heights, especially when mixed forests are underrepresented in the dataset. As such, striking a balance between image quality and temporal fidelity is crucial when determining the composite time span to achieve the most accurate results.

5.2. GEDI as calibration data

Currently, GEDI is one of only two operating spaceborne lidar instruments with quasi-global coverage and will remain an essential source of information on forest vertical structures in the near future. However, data acquired by GEDI could potentially be contaminated by large sources of errors, impacting the accuracy of the produced height maps. First, GEDI is a full waveform instrument, and with its pulse width of 15 ns (~ 4.5 m; Dubayah et al., 2020), GEDI is less sensitive to vertical structures that are shorter than ~2.25 m (minimum RH_{100} value over bare soil). Therefore, by relying on the RH_{98} metric as calibration data, our proposed models are overestimating heights over bare soils, water bodies, and agricultural areas. Indeed, the lowest estimated height over Ghana was around 2.15 m over water surfaces and around 3 m over bare soils/low vegetated areas. These uncertainties could be reduced with different approaches. Di Tommaso et al., 2021 proposed an approach to classify the GEDI shots as being acquired

over tall or short crops, which could further be developed to include water bodies and bare soils. A pre-processing step, before model calibration, would then be to assign a lower height value (e.g., a lower RH metric) for the GEDI shots belonging to the low-height classes (e.g., short vegetation, water bodies, bare soil). Another approach would be to replace GEDI acquisitions for short vegetations (e.g., less than 25 m) with ICESat-2 acquisitions. ICESat-2 with its photon counting technology has shown superior accuracy in comparison to GEDI short vegetation (< 5 m), while GEDI was more accurate for tall vegetation across all height ranges (Liu et al., 2021). However, the assessment in the study of Liu et al., 2021 was conducted over geolocation-corrected GEDI and ICESat-2 measurements. Without geolocation-correction, ICESat-2 which shows good agreement with airborne lidar estimates for medium heights (Queinnec et al., 2021), should also be preferred to GEDI acquisitions for this height range (i.e., less than 25 m) due to its lower geolocation errors (<5 m; Neuenschwander et al., 2022) in comparison to GEDI (geolocation error of 10 m; Dubayah, Ralph et al., 2021). Over tall trees (e.g. > 25 m), GEDI has overall better accuracy (Liu et al., 2021). However, recent studies have shown that GEDI's accuracy is affected by several factors, including the laser used, the processing algorithm used to extract the metrics, as well as acquisition time (Fayad et al., 2021a; Lahssini et al., 2022; Liu et al., 2021). Regarding these two factors, and given the limited number of available acquisitions over the tropical forests of Ghana, all available viable GEDI acquisitions were used in this study, however, over larger areas, a better approach would be to use only high-powered shots and shots acquired during night time to reduce solar noise.

Another potential challenge for machine learning-based models is GEDI measurements acquired over urban areas, including buildings. GEDI's returns do not distinguish between man-made structures and natural canopies, and recent studies (Ma et al., 2023) suggest that GEDI is capable of deriving building heights from its returns. If GEDI shots acquired over urban areas are not accounted for, machine learning-based models may inadvertently associate features from S1/S2 data with GEDI heights over these areas, leading to the generation of non-zero, seemingly random height values within cities. In this study, this phenomenon was not observed, indicating that there weren't enough GEDI footprints acquired over buildings to skew the performance. Nonetheless, α-2MDU and Hy-TeC showed two distinctive outputs over buildings. Hy-TeC, which relies solely on S2 data, correctly linked building heights to low values, given that the reflectance signature of buildings resembles that of bare soils more than forests. Conversely, α-2MDU, which incorporates both S1 and S2 data, mistakenly associated building heights with very high values due to the increased backscatter from S1 over urban structures, an undesired outcome. To mitigate such behavior and ensure consistent model output over buildings in the context of canopy height mapping, an approach could involve setting GEDI's derived relative heights acquired over urban areas to very low values, or mask urban areas in post-processing using an auxiliary independent dataset.

Finally, while both of our models (namely α-2MDU and Hy-TeC) have shown better agreement to ALS than the current state-of-the-art, the fact remains that neither GEDI nor models calibrated with GEDI data are capable of retrieving the forest's height dynamics ranges with good accuracy. This finding is also in line with several studies that evaluated the capacity of GEDI to retrieve the canopy heights in dense forests with very heterogeneous compositions (different ages, species, etc.) (Lahssini et al., 2022; Morin et al., 2022).

5.3. Model Architectures

Regarding model architecture, our proposed approach differs from current state-of-the-art methods (Lang et al., 2023a; Morin et al., 2022; Potapov et al., 2021) in that our models are trained on larger images and they are tasked with creating a continuous height map of the same area, regardless of the density and localization of the GEDI acquisitions within. In contrast, previous approaches used as model input either a large feature space derived from vegetation phenology at a given GEDI's location (e.g. Per-pixel spectral algorithms such as regression tree models, Li et al., 2020; Potapov et al., 2021), or small windows that cover GEDI's location and neighboring pixels (Lang et al., 2023a). Processing larger images as input during training provides several benefits that can lead to improved performance. First, by leveraging the spatial information within the image, the models can better understand the overall structure of the image and generate more accurate reconstructions (Long et al., 2015). Additionally, processing larger images allows the models to learn more complex and abstract features, which can further improve the reconstruction quality. Second, training on larger images can help prevent overfitting and improve generalization performance by exposing the models to more variations and noise in the training data (Pathak et al., 2016).

Regarding the performance of α-2MDU and Hy-TeC, the results in this study showed that both models were performant in estimating canopy heights, over the tall tropical forests in central and the south of Ghana, as well as over Savannas. Nonetheless, Hy-TeC, overall, outperformed α-2MDU. Hy-TeC has better ground sampling distance, less noise, and lower RMSE when compared to GEDI, the ALS dataset, and the VHR_{HM} dataset. The difference between the two models is attributed to two main factors. First, Hy-TeC uses a transformer encoder while α-2MDU uses a ConvNet-based encoder. This difference is significant, as to obtain long-range dependencies between pixels in ConvNet-based networks, it is required to downscale the image or apply a large receptive field, leading to some loss of the finer details in the image and reducing the model's ability to capture complex structures. In contrast, transformer models such as ViTs can capture global information without downsampling or losing details, making them better suited for tasks requiring high resolution and fine-grained features (Dosovitskiy et al., 2020). The second key difference between the two models lies in the use of S1 data. Indeed, the study of Schwartz et al., 2022 showed that a model using both S1 and S2 data is significantly more accurate than a model using either dataset. Nonetheless, speckle noise as well as terrain geometry can introduce artifacts in the estimated height maps from S1 data. The effect of these can be observed in the height estimates of α-2MDU, which in several instances, overestimated the tree cover in sparse vegetated regions. Hy-TeC, on the other hand, does not have these limitations as it only relies on S2 data as its main input. Hy-TeC, however, could have potentially two main limitations. First, an untrained Hy-TeC model requires height estimates provided from a pre-trained model at lower resolutions to effectively learn an inductive bias in a timely manner. Second, Hy-TeC, as all transformer-based models, require significantly more GPU memory than their ConvNets counterparts, which might be limiting.

5.4. Loss Functions

One of the key contributions of this study is the use of two loss functions to optimize our models. The first loss, a cross entropy loss, is obtained by discretizing the output space into n-overlapping bins, while the second loss, a regression loss, is used directly on the regression estimates. This approach allowed the models to significantly reduce the underestimation effect that is usually observed when estimating the height of very tall trees. Our two models that employ such loss showed in general higher tree height estimates over areas with very tall trees in comparison to the estimates obtained in the studies of Potapov et al., 2021 and Lang et al., 2023a. Classifying the height range before estimating the height value has two main advantages. First, by constraining the range of heights that can be estimated for a given pixel, the model is allowed more latitude to learn the differences between shorter and taller canopies. Moreover, classification loss can also be weighted, which acts as a counterbalance for underrepresented height classes. By assigning a higher weight to the minority classes, the classification loss will place more emphasis on correctly predicting these classes, effectively compensating for their underrepresentation in the training data. Indeed, in our training data, only a small fraction of the available GEDI footprints were acquired over tall trees (less than 1%), yet model accuracy on the height estimation of these GEDI footprints was similar to those with significantly higher counts.

Finally, in this study, we replaced the de facto standard of regression losses, that is the L1/L2 losses, with an adaptive loss that constrains the value of the residuals to an automatically set bound. This loss which reduces the effects of outliers was used in this study as a simple alternative to a more elaborate filtering scheme of GEDI acquisitions and resulted in better quality maps in comparison to the L1/L2 losses. Nonetheless, while the adaptive loss function's ability to reduce the effect of outliers is undeniable, we are uncertain of its effects on the inliers that represent hard cases, such as the case of very tall trees present within a forest of significantly shorter ones. As such, more research needs to be conducted to quantify the advantages of using such loss in comparison to the unbounded losses such as L1/L2 coupled with a better-filtered GEDI training dataset.

6. Conclusions

The proposed transformer-based model in this study, Hy-TeC provides a novel approach for the estimation of canopy heights at large scale. Hy-TeC, which is trained using input S2 data and calibrated with GEDI has shown desirable qualities to currently available height products and even in comparison to our best ConvNet model that served as baseline (i.e., α-2MDU). Indeed, Hy-TeC showed significantly improved performances in terms of summary statistics over tall dense forests as well as over short and sparse vegetation. Hy-TeC is also capable of

producing canopy height estimates at unprecedented ground sample distances, allowing it to better capture tree structures within dense forests and sparsely vegetated areas, sometimes at the level of trees if they are large enough to be detected by S2. The model also showed reduced artifacts. These qualities render Hy-TeC more beneficial for downstream tasks such as biomass estimation, forest disturbances mapping, and forest monitoring.

The proposed loss function formulated as an adaptive discrete and continuous loss has also shown quantitatively significantly reduced bias over the tall trees and at the same time able to mitigate some of the effects from GEDI's penetration limitation and geolocation errors. As such, these both propositions, combined with a teacher/student paradigm for the transformer pre-training, allow for very efficient canopy height models that can be deployed to monitor forests (e.g., illegal logging, forest degradation, and afforestation) at very large scales, from nation-wide to continental.

Appendix A. Distribution of GEDI's RH_{98} over Ghana

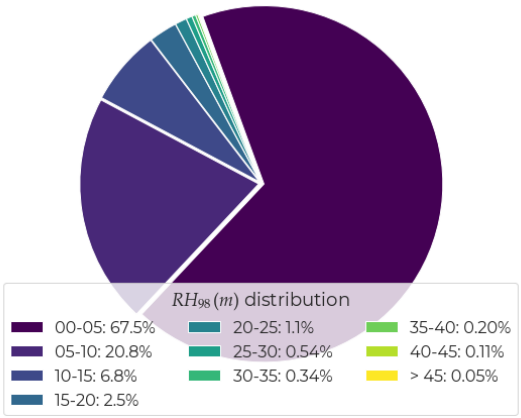


Figure A1. Distribution of RH_{98} (m) acquisitions over Ghana.

Appendix B. Implementation of the UNet-based family of models:

B.1. Encoder path

The encoder path used in the UNet-based models (i.e., 2MOU, 2MDU, and α-2MDU) consists of four convolutional based Encoder Blocks (CEB_i, Fig. B1) that extract increasingly abstract and high-level features from the input image while reducing its spatial resolution at each stage. Each CEB_i comprises two convolutional operations, followed by a Batch Normalization (Batch Norm) layer and a Leaky ReLU activation function. Leaky ReLU was chosen over the standard ReLU activation function in each block, as it overcomes some of the limitations of ReLU. Unlike ReLU, which sets all negative input values to zero, Leaky ReLU introduces a small slope for negative values (0.01), enabling the network to learn even for negative inputs. This property of Leaky ReLU ensures that no neuron becomes entirely inactive. Finally, to reduce the spatial dimension, a strided 2×2 convolutional layer was used instead of a pooling layer (e.g., max-pooling) to preserve information in the feature maps. The use of strides rather than max-pooling was preferred as max-pooling retains only the maximum values in a given area, which could result in information loss. Each encoder block with input of size: width (W), height (H), and channels (C) results in a feature map of size $\frac{\mathrm{H}}{2} \times \frac{W}{2} \times 2C$:

$$CEB_i: \mathbb{R}^{W \times H \times C} \rightarrow \mathbb{R}^{\frac{W}{2} \times \frac{H}{2} \times 2 \cdot C} \qquad A1$$

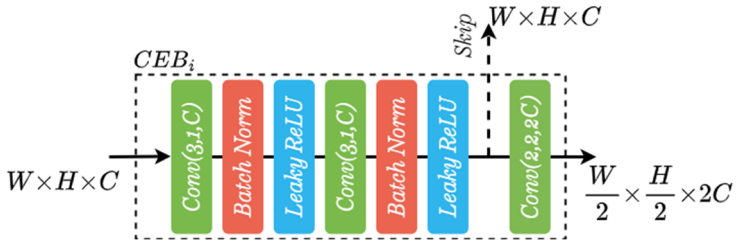


Figure B1. A U-Net encoder block (CEB_i). *Conv(K, S, F)* represents a convolution operation with a kernel of size K, stride S, and F filters (filters and channels are mentioned interchangeably). *Batch Norm* is a batch normalization layer. *Leaky ReLU* is a leaky rectified linear unit.

Optical and radar images were jointly used to produce wall-to-wall canopy height maps. Given that the two data sources represent distinct physical measurements over a scene, our U-Net-based implementation used two separate encoder paths, one for each data source. By having separate encoders, the network can learn to extract relevant features and representations independently for each modality, rather than trying to extract common features for all modalities (Kuga et al., 2017). This approach can lead to a more focused representation of the data, as each encoder can specialize in processing its own modality (Stahlschmidt et al., 2022). The outputs of both encoder paths are then concatenated channel-wise and fed into a self-aware attention (SAA) module to effectively learn the non-local interactions among encoder features. The specific implementation of the SAA module can be found in the study of Chen et al., 2022.

B.2. Decoder path

Following the SAA module, which represents the bottleneck of the model, is the decoding path, which uses the encoded features to reconstruct the output image. It does so by up-sampling the feature maps and refining them with convolutional layers. The decoding path used in this study consists of four convolutional-based Decoder Blocks (CDB_i, Fig. B2) that include up-sampling layers, concatenation layers, and convolutional layers. The up-sampling layers increase the spatial resolution of the image, while the concatenation layers connect the decoding path to the encoding path through skip connections, allowing the network to consider the features extracted in the encoding path as well (Ronneberger et al., 2015).

Each of the four decoder blocks starts with a strided 2x2 transpose convolution operation (ConvT) which effectively doubles the spatial dimension of the input feature map. ConvT allows the network to learn a mapping from low-resolution to high-resolution feature maps, whereas bilinear interpolation is a fixed and non-learnable operation. This means that the network can use its knowledge of the data to generate more meaningful high-resolution feature maps, rather than simply stretching the low-resolution maps as in bilinear interpolation. The decoder block scales the input feature maps of size W×H with C channels to a size of $2W \times 2H \times \frac{C}{2}$:

$$CDB_i: \mathbb{R}^{W \times H \times C} \rightarrow \mathbb{R}^{2 \cdot W \times \cdot H \times \frac{C}{2}} \qquad A2$$

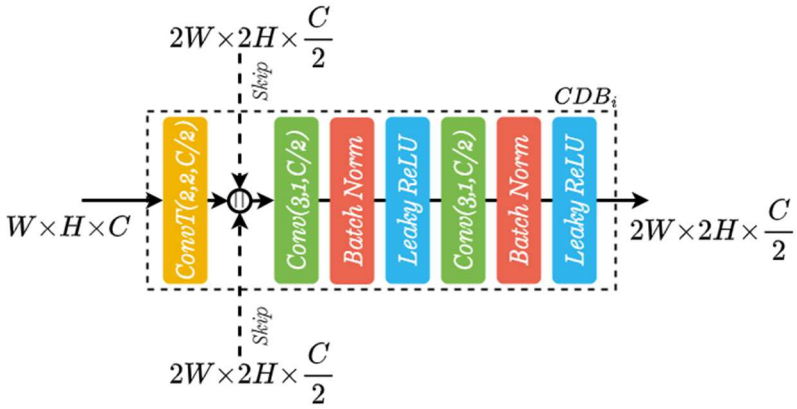


Figure B2. Implemented decoder block (CDB_i). *ConvT(K, S, F)* represents a transpose convolution operation with a kernel of size K, stride S, and F filters. The '||' operation represents a channel-wise concatenation.

B.3. Output head

The output head consists of a single convolution operation with a kernel size of 1 and a stride of 1. This operation reduces the number of channels from C channels to a single channel (i.e., the height estimates), and is followed by a Softplus activation layer (SoftPlus is a smooth approximation to the ReLU function) to ensure that the output heights are always positive (Fig. B3.a).

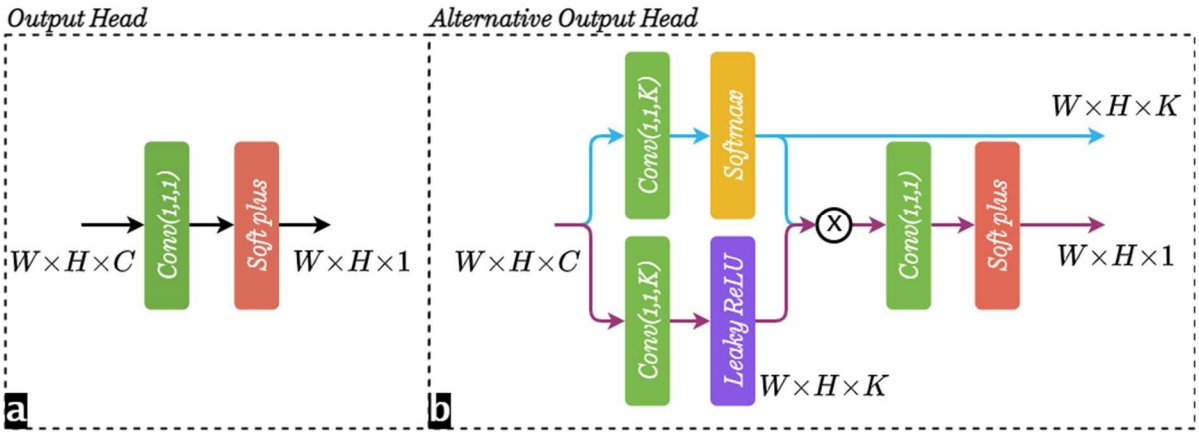


Figure B3. Proposed output blocks for the estimation of canopy heights. *ConvT(K, S, F)* represents a convolution operation with a kernel of size K, stride S, and F filters. In (b) the blue path represents the classification sub-module and the purple path represents the regression sub-module.

B.4. Alternative output head

The modified output head (Fig. B3.b) will connect the output of CDB_4 (Fig. 3) of size $W \times H \times C$ to two separate convolutional layers with size $W \times H \times K$ (K is the number of the predefined overlapping height classes). To obtain height class probabilities a *softmax* activation function is used on one of the convolutional layer's outputs (Fig. B3.b, blue path). Finally, the outputs of the two convolutional layers of size $W \times H \times K$ are multiplied element wise and go through to another convolutional layer with an output of size $W \times H \times 1$ followed by a *Softplus* activation function, making the regression task (Fig. B3.b, violet path). As such, the classification path will act as a constraint on the regression sub-module by bounding its output.

Appendix C. Implementation of Hy-TeC:

C.1. Image patching

Since ViTs use the original transformer architecture conceived for the processing of sequential data (e.g., sequence of words), fixed-sized input images with shape (width W, height H, and C channels) are first decomposed into a batch of N non-overlapping patches, hereafter referred to as tokens, of shape P×P×C. Each token is then projected by a linear layer L to a vector of size $D = P \cdot P \cdot K$. Next, to conserve positional information, a learnable positional encoding (Jiang et al., 2022) is applied to each projected token (hereafter referred to as embedded patches). Moreover, thanks to the global receptive field of ViTs which allows each pixel to attend to every other pixel within an image, we will separate the inputs to the encoder by their native resolution. In essence, bands with an original resolution of 10 m will be concatenated channel-wise in input 1, while bands such as the NIR and SWIR bands from S2 that have a native resolution of 20 m will be concatenated channel-wise in input 2. Therefore, the overall input to the model is the concatenated embedded patches from input 1 and input 2. The pre-processing of an image to be used by the ViT is shown in Fig. C1.

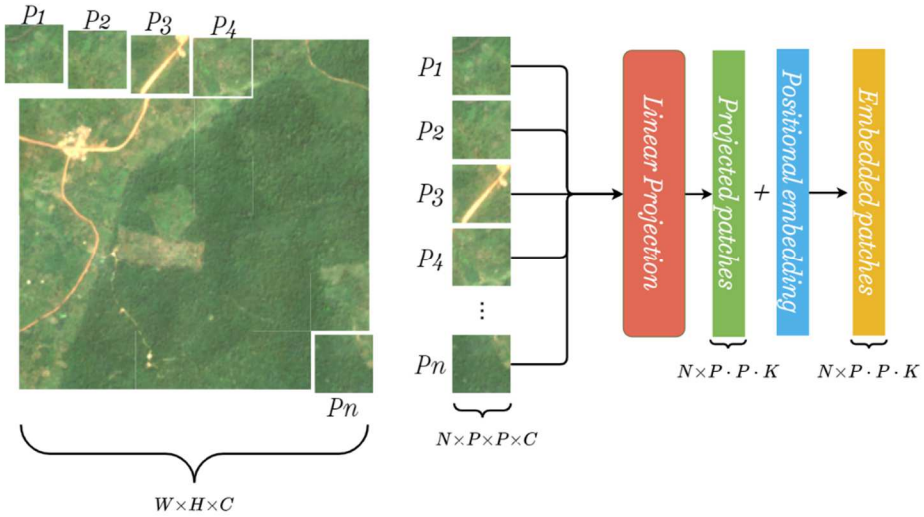


Figure C1. Visualizing the patching, projection, and positional embedding steps in the image pre-processing pipeline for Hy-TeC.

C.2. Encoder

The encoder is composed of multiple standard Transformer layers (Dosovitskiy et al., 2020). Each layer consists of a normalization layer (Norm), followed by a Multi-Head Self-Attention module (MHSA), another normalization layer, and finally a multi-layer perceptron (MLP). The input tokens fed into the encoder are transformed using T transformer blocks into a new representation t^i, where i is the output of the i-th transformer layer. In this study we used the ViT-Base proposed in (Dosovitskiy et al., 2020) which features 12 transformer blocks.

C.3. Decoder

The decoder layers are inspired by the decoder layers presented previously for the ConvNet-based models but adapted to the output of the transformer layers. Also, as the encoders are fed by 2N embedded tokens of size D from input 1 (native 10 m resolution) and input 2 (native 20 m resolution), only the N transformed tokens from input 1 are used in the decoder layers. In essence, a reprojection block (RB, Fig. C2) followed by a decoding block (DB, Fig. C3) is responsible for reshaping the transformed input 1 tokens into feature maps of different resolutions. The first operation in the reprojection block is a spatial concatenation (SP) of the tokens of size $N \times D$ to an image-like feature map of size $P \times P \times L$:

$$SP = \mathbb{R}^{N \times D} \rightarrow \mathbb{R}^{P \times P \times L} \qquad B1$$

After the *SP* operation, the feature maps from four different transformer blocks are reprojected to four different resolutions with $\hat{L}$ channels:

$$RB\ \#\ i = \begin{cases} \mathbb{R}^{P\times P\times L} \rightarrow \mathbb{R}^{\frac{P}{2}\times\frac{P}{2}\times \hat{L}}, if\ i = 1 \\ \mathbb{R}^{P\times P\times L} \rightarrow \mathbb{R}^{P\times P\times \hat{L}}, if\ i = 2 \\ \mathbb{R}^{P\times P\times L} \rightarrow \mathbb{R}^{2P\times 2P\times \hat{L}}, if\ i = 3 \\ \mathbb{R}^{P\times P\times L} \rightarrow \mathbb{R}^{4P\times 4P\times \hat{L}}, if\ i = 4 \end{cases} \quad B2$$

Next, the four reprojected blocks are input to four decoder blocks (DB), consisting of two convolutions operations (Conv K=3, S=1) and a transpose convolution operation (ConvT K=4, S=4) to effectively quadruple the spatial dimension and reduce the number of channels by eight.

$$DB\ \#\ i = \mathbb{R}^{\hat{H}\times\hat{W}\times\hat{D}} \rightarrow \mathbb{R}^{4\cdot\hat{H}\times 4\cdot\hat{W}\times\frac{\hat{D}}{8}}, i \in \{1, \ldots, 4\} \quad B3$$

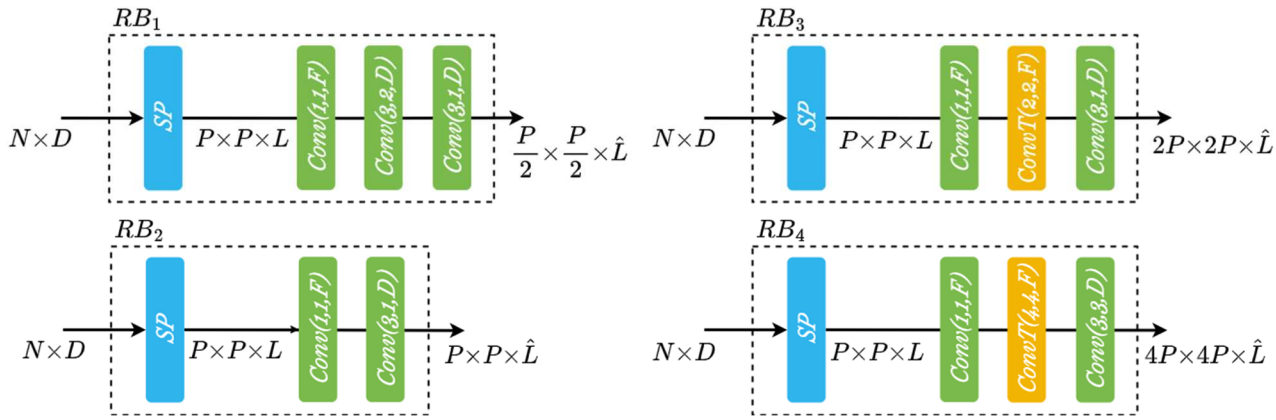


Figure C2. The implementation of the four reprojection blocks (RB_i).

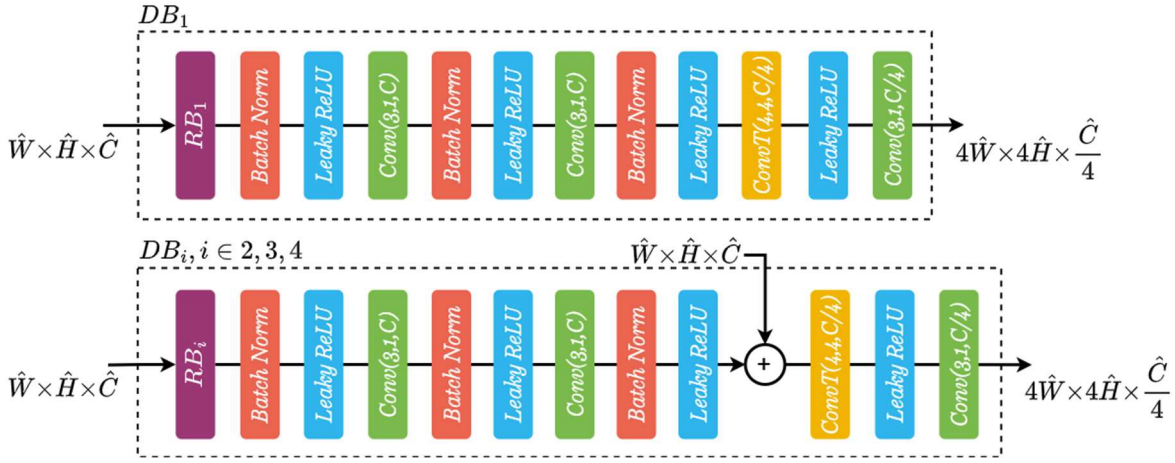


Figure C3. The implementation of the four decoder blocks (DB_i).

Finally, to produce the height estimates, we connect the same output head in Fig B3.b to the output of *DB # 4*. Effectively making our Hy-TeC model a canopy height classifier and regressor.

C.4. Knowledge distillation

Knowledge distillation (KD), introduced by Hinton et al., 2015, is a technique of transferring knowledge from a pre-trained teacher model to a smaller student model without incurring significant drops in performance. The success of KD can be attributed to the fact that deep learning models are highly complex, non-linear functions that can be difficult to optimize. By emulating the outputs of a pre-trained teacher model, the student model can inherit the knowledge and generalization abilities of the teacher model, thus simplifying the learning task. In general, KD in a classification problem is achieved through the softened cross entropy loss, which regards the teacher predictions as 'soft labels' containing knowledge of the correlation of the class labels (Hinton et al., 2015). ViT-based models can benefit from KD as demonstrated in a recent work with transformer-based models (Touvron et al., 2021). Indeed, Touvron et al., 2021) were able to train a ViT model on a reduced dataset with performances comparable to a ViT trained on large amounts of data. Nonetheless, KD works well for classification problems since the student models have access to the softened labels of the teacher model. In regression problems where the output space is continuous, the teacher model predicts values with the same characteristics as the ground truth data and an additional unknown error distribution. Moreover, since the output space is unbounded, a teacher model can provide false guidance to the student model. As such, KD is harder to achieve.

Here we propose a novel KD approach for the optimization of our proposed transformer-based model to forgo the need for additional training images or the required pre-training stage. First, acting as a teacher model is the best performing UNet-based model, namely α-2MDU (c.f. section 3.1.2). Next, since we cannot compute soft labels from the teacher model, which is a requirement for KD, we instead use the output of the teacher model to influence the transformer-based model at the lower resolutions, while leaving it to freely form its representation at the highest resolution (i.e., 10 m). In terms of implementation, the output of the teacher model at the highest resolution with size $H \times W$ will be down-sampled bilinearly to fit the output resolution of the decoder blocks of the transformer-based model at lower resolutions.

Appendix D. Barron, 2017 Adaptive Loss

The adaptive loss is a two-parameter loss function that generalizes many one-parameter loss functions (L1 or MAE, L2 or MSE, Cauchy/Lorentzian, Geman-McClure, Welsch/Leclerc, generalized Charbonnier, etc.), and is defined as follows:

$$\ell_{AR}(Y, Y') = \frac{|\alpha - 2|}{\alpha}\left(\left(\frac{\left(\frac{(Y - Y')}{c}\right)^2}{|\alpha - 2|} + 1\right)^{\frac{\alpha}{2}} - 1\right) \quad \text{D}_1$$

Where ℓ_{AR} is the robust-adaptive loss function, α is a shape parameter that controls the robustness of the loss and c is a scaling parameter. Both parameters are learned during training.

Appendix E. Selected grid cells used for the training/validation of the proposed models

Set	RH_{98} range	Ratio of RH_{98} within a cell	Cell Count	Number of duplications
1	RH_{98} <= 5m	> 50%	1152	0
2	5 m < $RH_{98} \leq$ 10m	> 25%	327	1
3	10 m < $RH_{98} \leq$ 15m	> 10%	469	0
4	15 m < $RH_{98} \leq$ 20m	> 10%	149	3

5	20 m < RH_{98} ≤ 25m	> 5%	133	4
6	25 m < RH_{98} ≤ 30m	> 2.5%	139	4
7	30 m < RH_{98} ≤ 35m	> 2.5%	52	8
8	35 m < RH_{98} ≤ 40m	> 2.5%	47	8
9	RH_{98} ≥ 40	> 2.5%	91	4

Table E1. The number of 768×768 grid cells based on the ratio of RH_{98}/cell and the RH_{98} height ranges.

Appendix F. Summary statistical metric definitions

$$\boldsymbol{r} = \frac{\sum_{i=1}^{n}(\boldsymbol{y_i} - \hat{\boldsymbol{y}})(\hat{\boldsymbol{y}} - \bar{\hat{\boldsymbol{y}}})}{\sqrt{\sum_{i=1}^{n}(\boldsymbol{y_i} - \bar{\boldsymbol{y}})^2}\sqrt{\sum_{i=1}^{n}(\hat{\boldsymbol{y}_i} - \bar{\hat{\boldsymbol{y}}})}} \qquad E1$$

$$RMSE = \sqrt{\frac{1}{n}\cdot\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad E2$$

$$RMSPE = 100\cdot\sqrt{\frac{1}{n}\cdot\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right)^2} \qquad E3$$

$$SDSD = (SD_s - SD_m)^2 \qquad E4$$

$$LCS = 2SD_sSD_m(1 - r) \qquad E5$$

Where y_i is the target value (i.e. RH_{98}), $\hat{y}_i$ the estimated value, $\bar{y}_i$ is the mean of all observed values, $\bar{\hat{y}}$ is the mean of all estimated values, n is the sample size, and SD_s and SD_m are respectively the Standard deviation of the measured heights by GEDI and the standard deviation of the estimated heights by the models. They are defined as follows:

$$S_{\square\square} = \sqrt{\frac{\sum_{\square=1}^{\square}(\square_{\square} - \bar{\square})}{\square}} \qquad E6$$

$$S_{\square\square} = \sqrt{\frac{\sum_{\square=1}^{\square}(\widehat{\square_{\square}} - \bar{\hat{\square}})}{\square}} \qquad E7$$

Appendix G. The degradation effect observed from 2MDU's generated height maps

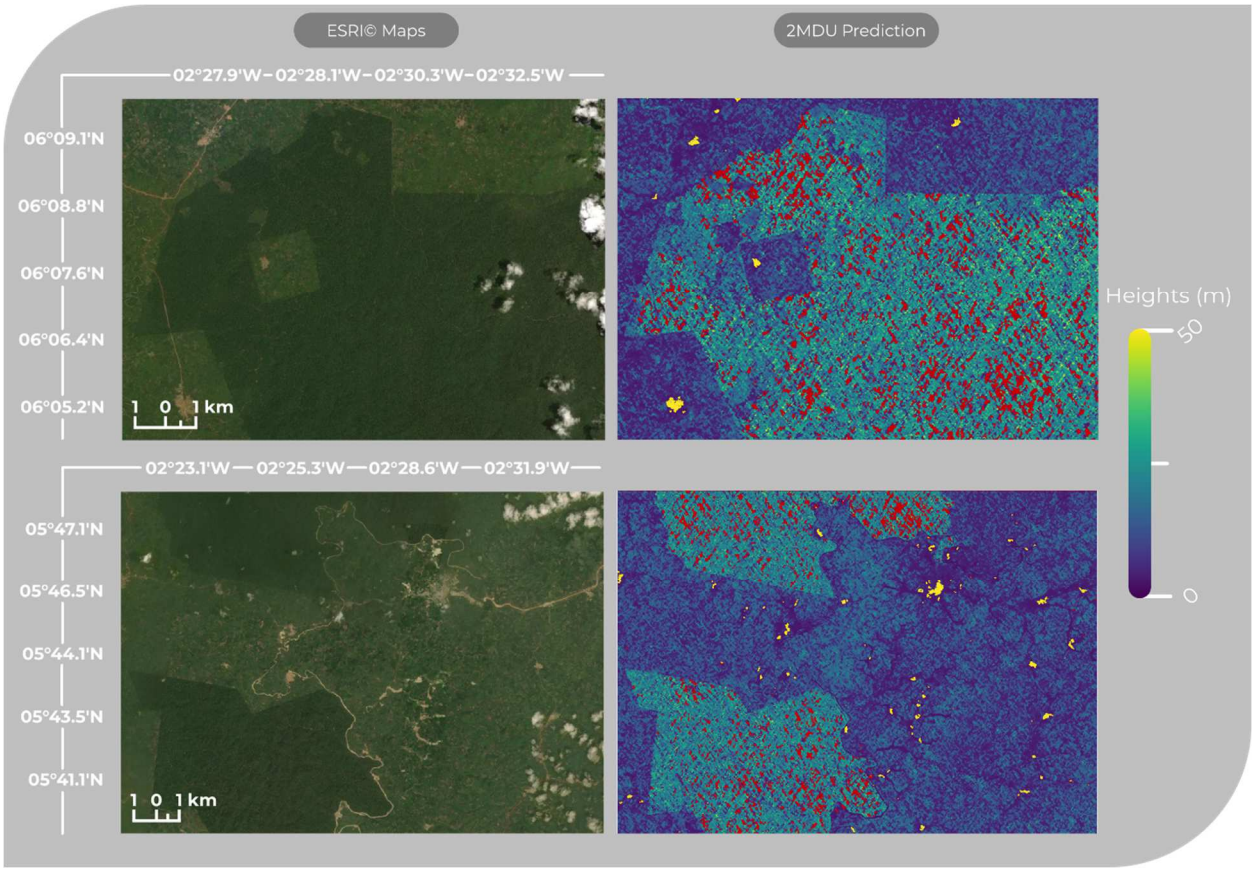


Figure G1. Illustration of the degraded performance (red polygons) of the 2MDU model over two forest patches. The red polygons represent forested areas where the 2MDU model estimated the heights of tall trees as being less than 2m.

Appendix H. Comparison between the residual and gradients of the adaptive loss and Huber loss

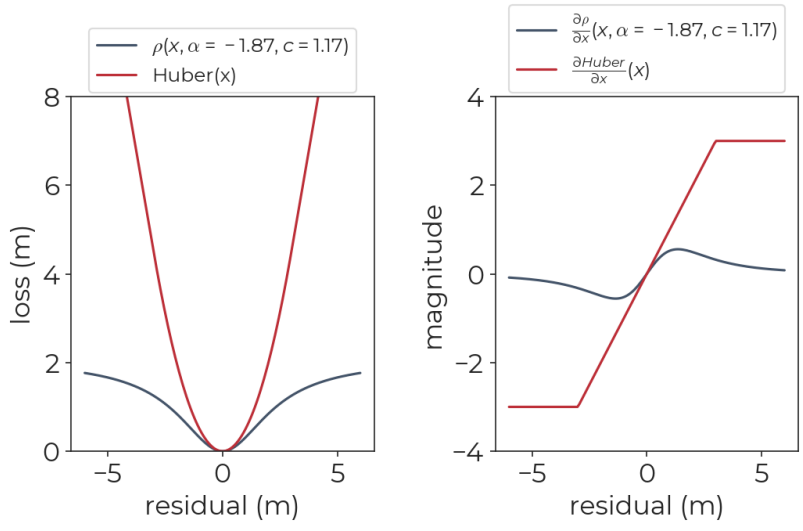


Figure H1. The adaptive (ρ) and Huber loss functions (left) and their gradients (right).

Appendix I. Distribution of GSI from the four compared height products

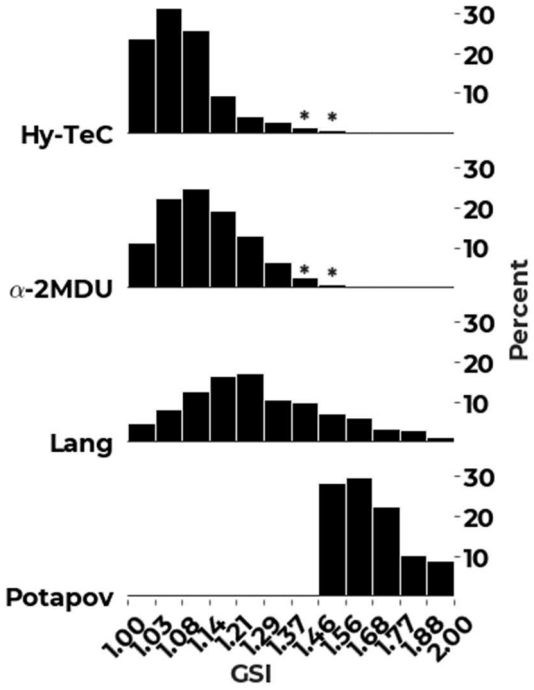


Figure I1. GSI distribution of the 5000 256×256 height patches from the four compared height models over Ghana. '*' represents less than 1% of patches.

Appendix J. Differences in output resolution of the four height products over sparse vegetation

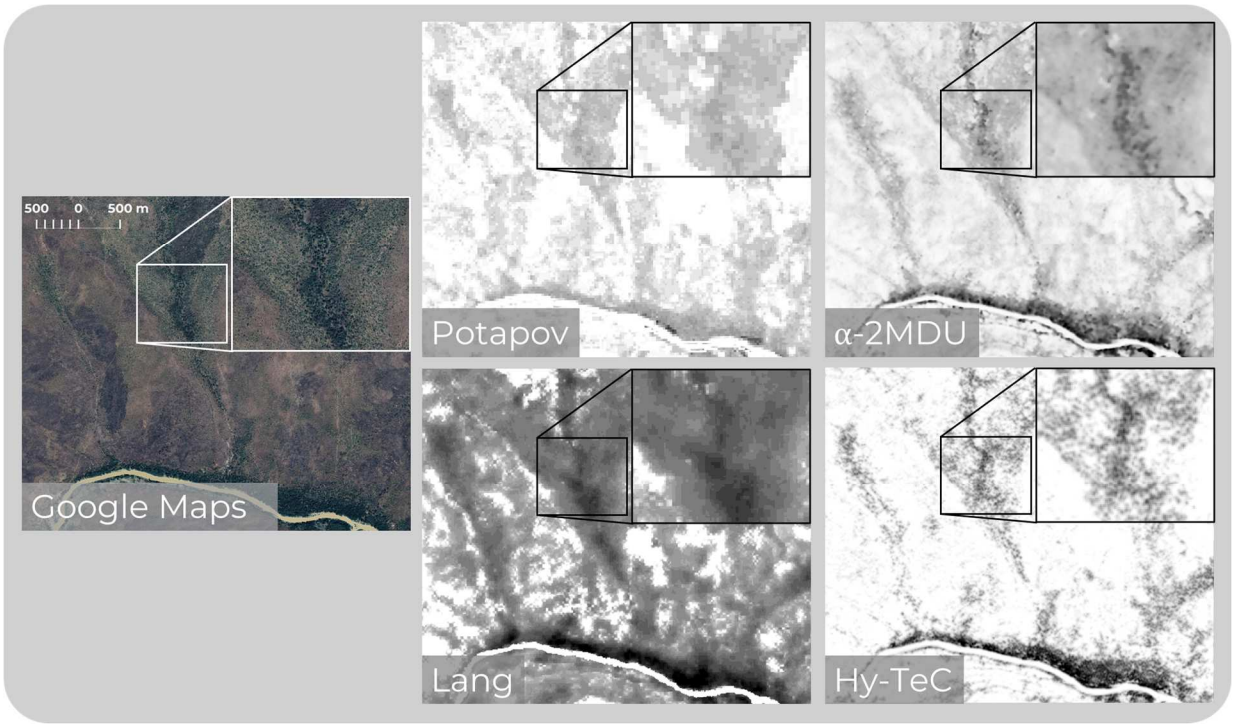


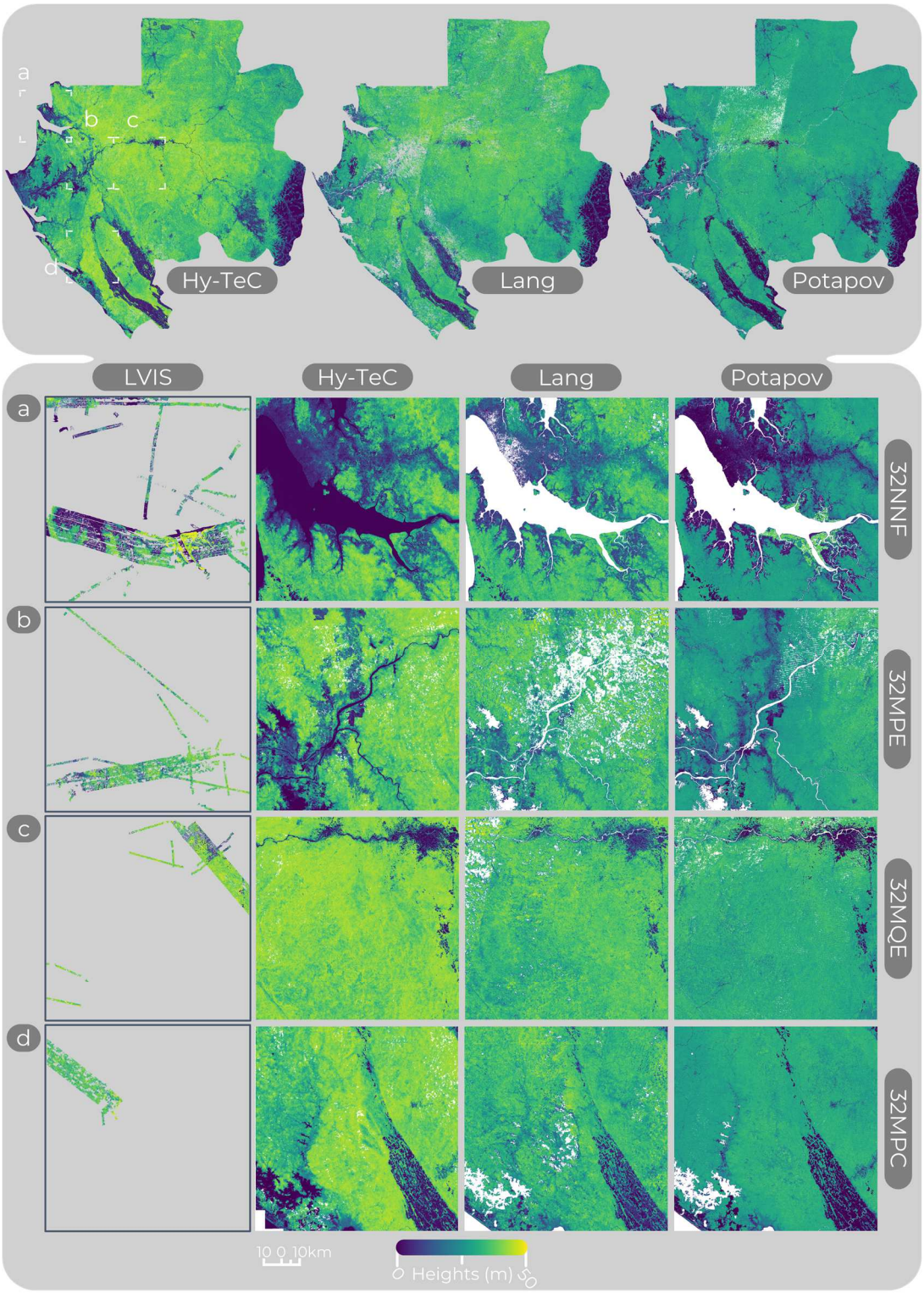
Figure J1. Illustration of the tree height predictions from the four compared models over a sparsely vegetated area. Darker colors indicate higher heights.

Appendix K. Hy-TeC's performance over Gabon

To better assess the performance of Hy-TeC, we have generated a canopy height map over Gabon, with a surface area of 267,670 km^2. The map was generated using Hy-TeC's model weights trained over Ghana, and then fine-tuned for a couple of epochs over Gabon to account for possible differences in the spectral values and differences in the forest structures between the two sites.

The generated map, along with those from the studies of Potapov et al., 2021 and that of Lang et al., 2023a were compared to four independent LVIS acquisition sites in Gabon in 2016. To stay consistent with comparison results provided in the study of Lang et al., 2023a, LVIS data were obtained from the aforementioned study's validation data (Lang et al., 2023b) without any modification. The results presented in Fig. K1 show that qualitatively, the map qualities from Hy-TeC over Gabon are in line with those observed over Ghana. Specifically, over the four Sentinel-2 tile regions, Hy-TeC had a wider range of possible height estimate values between the three compared products (evidenced by the higher contrast in the color range), produced higher estimation values for tall trees, and had overall fewer artifacts than Lang's product. Quantitavley, the results in Table K1 show that Hy-TeC outperforms the two previous canopy height models, namely that of Lang et al., 2023a and Potapov et al., 2021, on all error metrics and by a significant margin over the four Sentinel-2 tile regions. In terms of RMSE (MAE) (Table K1) the difference between Hy-TeC and the product of Lang over tile 32NNF (average LVIS height of ~21 m) is 0.5 m (0.6 m) and 2.5 m (1.5 m) in comparison to that of Potapov. However, the difference in RMSE (MAE) between the different products increased with increasing average LVIS canopy height estimates. Indeed, for tile 32MPE (Table K1) where the LVIS average canopy height is ~32 m, Hy-TeC showed an RMSE (MAE) of 7.9 m (6.1 m) while the products of Lang and Potapov showed an RMSE (MAE) of respectively 8.8 m (6.8 m) and 10.8 m (8.7 m). For tile 32MQE (Table K1, average LVIS canopy height of ~36 m), the RMSE (MAE) of Hy-TeC is 6.5 m (5.0 m), which is 2.4 m (2.2 m) lower than that of Lang, and 4.9 m (4.2 m) lower than Potapov's. For tile 32MPC (Table K1, LVIS canopy height of ~35 m), the differences are less drastic than tile 32MQE, with an RMSE (MAE) of 6.6 m (5.1 m) with Hy-TeC, which is 0.6 m (0.4 m) lower than Lang's product, and 3.4 m (4.2

m) lower than Potapov's product. However, over tile 32MPC, Hy-TeC overestimated the canopy height by 1.0 m in comparison to LVIS (bias, Table K1), while Lang and Potapov, both underestimated the canopy height by 1.9 m and 7.1 m respectively. Overall, Hy-TeC showed on average, lower and more consistent bias than the two previous canopy height products..



1146

1147 Figure K1. Illustration of the generated canopy height maps from Hy-TeC, Lang, and Potapov over Gabon. a, b, c, and d represent four
1148 Sentinel-2 tiles that intersect with the GEDI-like canopy heights derived from LVIS in Gabon.

Tile name	Lat.	Lon.	RMSE (m)			MAE (m)			Bias (m)			r		
			Hy-TeC	Lang	Potapov	Hy-TeC	Lang	Potapov	Hy-TeC	Lang	Potapov	Hy-TeC	Lang	Potapov
32NNF	0.40	9.49	**7.6**	8.1	10.1	**6.1**	6.7	7.6	**-0.3**	-1.5	5.9	**0.85**	0.85	0.82
32MPE	-0.50	10.39	**7.9**	8.8	10.8	**6.1**	6.8	8.7	**-1.1**	1.4	6.2	**0.77**	0.60	0.56
32MQE	-0.50	11.28	**6.5**	8.9	11.4	**5.0**	7.2	9.2	**0.7**	4.8	7.9	**0.86**	0.71	0.74
32MPC	-2.30	10.39	**6.6**	7.2	10.0	**5.1**	5.5	8.3	**-1.0**	1.9	7.1	**0.47**	0.35	0.27

Table K1. Error metrics between LVIS and the different generated canopy height maps over Gabon per Sentinel-2 Tile. Best results are highlighted in bold.

References

Alexander, C., Korstjens, A.H., Hill, R.A., 2018. Influence of micro-topography and crown characteristics on tree height estimations in tropical forests based on LiDAR canopy height models. Int. J. Appl. Earth Obs. Geoinformation 65, 105–113. https://doi.org/10.1016/j.jag.2017.10.009

Baghdadi, N., Bernier, M., Gauthier, R., Neeson, I., 2001. Evaluation of C-band SAR data for wetlands mapping. Int. J. Remote Sens. 22, 71–88.

Baghdadi, N., Cerdan, O., Zribi, M., Auzet, V., Darboux, F., El Hajj, M., Kheir, R.B., 2008. Operational performance of current synthetic aperture radar sensors in mapping soil surface characteristics in agricultural environments: application to hydrological and erosion modelling. Hydrol. Process. Int. J. 22, 9–20.

Baghdadi, N., El Hajj, M., Zribi, M., Bousbih, S., 2017. Calibration of the water cloud model at C-band for winter crop fields and grasslands. Remote Sens. 9, 969.

Baghdadi, N., Le Maire, G., Bailly, J.-S., Ose, K., Nouvellon, Y., Zribi, M., Lemos, C., Hakamada, R., 2015. Evaluation of ALOS/PALSAR L-Band Data for the Estimation of Eucalyptus Plantations Aboveground Biomass in Brazil. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 8, 3802–3811. https://doi.org/10.1109/JSTARS.2014.2353661

Barbier, N., Couteron, P., Proisy, C., Malhi, Y., Gastellu-Etchegorry, J.-P., 2010. The variation of apparent crown size and canopy heterogeneity across lowland Amazonian forests: Amazon forest canopy properties. Glob. Ecol. Biogeogr. 19, 72–84. https://doi.org/10.1111/j.1466-8238.2009.00493.x

Barron, J.T., 2017. A General and Adaptive Robust Loss Function. https://doi.org/10.48550/ARXIV.1701.03077

Bazzi, H., Baghdadi, N., Amin, G., Fayad, I., Zribi, M., Demarez, V., Belhouchette, H., 2021. An Operational Framework for Mapping Irrigated Areas at Plot Scale Using Sentinel-1 and Sentinel-2 Data. Remote Sens. 13, 2584. https://doi.org/10.3390/rs13132584

Blair, J.B., Rabine, D.L., Hofton, M.A., 1999. The Laser Vegetation Imaging Sensor: a medium-altitude, digitisation-only, airborne laser altimeter for mapping vegetation and topography. ISPRS J. Photogramm. Remote Sens. 54, 115–122. https://doi.org/10.1016/S0924-2716(99)00002-7

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A., 2021. Emerging Properties in Self-Supervised Vision Transformers.

Chen, B., Liu, Y., Zhang, Z., Lu, G., Kong, A.W.K., 2022. TransAttUnet: Multi-level Attention-guided U-Net with Transformer for Medical Image Segmentation.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y., 2021. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation.

Chen, Q., McRoberts, R.E., Wang, C., Radtke, P.J., 2016. Forest aboveground biomass mapping and estimation across multiple spatial scales using model-based inference. Remote Sens. Environ. 184, 350–360. https://doi.org/10.1016/j.rse.2016.07.023

Couteron, P., Pelissier, R., Nicolini, E.A., Paget, D., 2005. Predicting tropical forest stand structure parameters from Fourier transform of very high-resolution remotely sensed canopy images: Predicting tropical forest

stand structure. J. Appl. Ecol. 42, 1121–1128. https://doi.org/10.1111/j.1365-2664.2005.01097.x
Crete, F., Dolmiere, T., Ladret, P., Nicolas, M., 2007. The blur effect: perception and estimation with a new no-reference perceptual blur metric, in: Rogowitz, B.E., Pappas, T.N., Daly, S.J. (Eds.), . Presented at the Electronic Imaging 2007, San Jose, CA, USA, p. 64920I. https://doi.org/10.1117/12.702790
de Thoisy, B., Fayad, I., Clément, L., Barrioz, S., Poirier, E., Gond, V., 2016. Predators, Prey and Habitat Structure: Can Key Conservation Areas and Early Signs of Population Collapse Be Detected in Neotropical Forests? PLOS ONE 11, e0165362. https://doi.org/10.1371/journal.pone.0165362
Di Tommaso, S., Wang, S., Lobell, D.B., 2021. Combining GEDI and Sentinel-2 for wall-to-wall mapping of tall and short crops. Environ. Res. Lett. 16, 125002. https://doi.org/10.1088/1748-9326/ac358c
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. https://doi.org/10.48550/ARXIV.2010.11929
Dubayah, Ralph, Hofton, Michelle, Blair, James, Armston, John, Tang, Hao, Luthcke, Scott, 2021. GEDI L2A Elevation and Height Metrics Data Global Footprint Level V002. https://doi.org/10.5067/GEDI/GEDI02_A.002
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. Sci. Remote Sens. 1, 100002. https://doi.org/10.1016/j.srs.2020.100002
Dubayah, R., Luthcke, S., Blair, J., Hofton, M., Armston, J., Tang, H., 2021a. GEDI L1B Geolocated Waveform Data Global Footprint Level V002. https://doi.org/10.5067/GEDI/GEDI01_B.002
Dubayah, R., Tang, H., Armston, J., Luthcke, S., Hofton, M., Blair, J., 2021b. GEDI L2B Canopy Cover and Vertical Profile Metrics Data Global Footprint Level V002. https://doi.org/10.5067/GEDI/GEDI02_B.002
Fayad, I., Baghdadi, N., Frappart, F., 2022. Comparative Analysis of GEDI's Elevation Accuracy from the First and Second Data Product Releases over Inland Waterbodies. Remote Sens. 14, 340. https://doi.org/10.3390/rs14020340
Fayad, I., Baghdadi, N., Riedi, J., 2021a. Quality Assessment of Acquired GEDI Waveforms: Case Study over France, Tunisia and French Guiana. Remote Sens. 13, 3144. https://doi.org/10.3390/rs13163144
Fayad, I., Ienco, D., Baghdadi, N., Gaetano, R., Alvares, C.A., Stape, J.L., Ferraço Scolforo, H., Le Maire, G., 2021b. A CNN-based approach for the estimation of canopy heights and wood volume from GEDI waveforms. Remote Sens. Environ. 265, 112652. https://doi.org/10.1016/j.rse.2021.112652
Flood, N., 2013. Seasonal Composite Landsat TM/ETM+ Images Using the Medoid (a Multi-Dimensional Median). Remote Sens. 5, 6481–6500. https://doi.org/10.3390/rs5126481
Frappart, F., Wigneron, J.-P., Li, X., Liu, X., Al-Yaari, A., Fan, L., Wang, M., Moisy, C., Le Masson, E., Aoulad Lafkih, Z., Vallé, C., Ygorra, B., Baghdadi, N., 2020. Global Monitoring of the Vegetation Dynamics from the Vegetation Optical Depth (VOD): A Review. Remote Sens. 12, 2915. https://doi.org/10.3390/rs12182915
Giri, C., Ochieng, E., Tieszen, L.L., Zhu, Z., Singh, A., Loveland, T., Masek, J., Duke, N., 2011. Status and distribution of mangrove forests of the world using earth observation satellite data: Status and distributions of global mangroves. Glob. Ecol. Biogeogr. 20, 154–159. https://doi.org/10.1111/j.1466-8238.2010.00584.x
Goetz, S.J., Steinberg, D., Betts, M.G., Holmes, R.T., Doran, P.J., Dubayah, R., Hofton, M., 2010. Lidar remote sensing variables predict breeding habitat of a Neotropical migrant bird. Ecology 91, 1569–1576. https://doi.org/10.1890/09-1670.1
Goodwin, N.R., Collett, L.J., Denham, R.J., Flood, N., Tindall, D., 2013. Cloud and cloud shadow screening across Queensland, Australia: An automated method for Landsat TM/ETM+ time series. Remote Sens. Environ. 134, 50–65. https://doi.org/10.1016/j.rse.2013.02.019
Hancock, S., Armston, J., Hofton, M., Sun, X., Tang, H., Duncanson, L.I., Kellner, J.R., Dubayah, R., 2019. The GEDI Simulator: A Large-Footprint Waveform Lidar Simulator for Calibration and Validation of Spaceborne Missions. Earth Space Sci. 6, 294–310. https://doi.org/10.1029/2018EA000506
He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R., 2021. Masked Autoencoders Are Scalable Vision Learners. https://doi.org/10.48550/ARXIV.2111.06377
Hese, S., Lucht, W., Schmullius, C., Barnsley, M., Dubayah, R., Knorr, D., Neumann, K., Riedel, T., Schröter, K., 2005. Global biomass mapping for an improved understanding of the CO2 balance—the Earth

observation mission Carbon-3D. Remote Sens. Environ. 94, 94–104. https://doi.org/10.1016/j.rse.2004.09.006
Hinton, G., Vinyals, O., Dean, J., 2015. Distilling the Knowledge in a Neural Network.
Holben, B.N., 1986. Characteristics of maximum-value composite images from temporal AVHRR data. Int. J. Remote Sens. 7, 1417–1434. https://doi.org/10.1080/01431168608948945
Huber, P.J., 1964. Robust Estimation of a Location Parameter. Ann. Math. Stat. 35, 73–101. https://doi.org/10.1214/aoms/1177703732
Huete, A., Didan, K., Miura, T., Rodriguez, E.P., Gao, X., Ferreira, L.G., 2002. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. Remote Sens. Environ. 83, 195–213. https://doi.org/10.1016/S0034-4257(02)00096-2
Imhoff, M.L., 1995. Radar backscatter and biomass saturation: ramifications for global biomass inventory. IEEE Trans. Geosci. Remote Sens. 33, 511–518. https://doi.org/10.1109/36.377953
Jiang, K., Peng, P., Lian, Y., Xu, W., 2022. The encoding method of position embeddings in vision transformer. J. Vis. Commun. Image Represent. 89, 103664. https://doi.org/10.1016/j.jvcir.2022.103664
Kennedy, R., Yang, Z., Gorelick, N., Braaten, J., Cavalcante, L., Cohen, W., Healey, S., 2018. Implementation of the LandTrendr Algorithm on Google Earth Engine. Remote Sens. 10, 691. https://doi.org/10.3390/rs10050691
Kobayashi, K., Salam, M.U., 2000. Comparing Simulated and Measured Values Using Mean Squared Deviation and its Components. Agron. J. 92, 345–352. https://doi.org/10.2134/agronj2000.922345x
Kuga, R., Kanezaki, A., Samejima, M., Sugano, Y., Matsushita, Y., 2017. Multi-task Learning Using Multi-modal Encoder-Decoder Networks with Shared Skip Connections, in: 2017 IEEE International Conference on Computer Vision Workshops (ICCVW). Presented at the 2017 IEEE International Conference on Computer Vision Workshop (ICCVW), IEEE, Venice, Italy, pp. 403–411. https://doi.org/10.1109/ICCVW.2017.54
Kundu, A., Li, Y., Rehg, J.M., 2018. 3D-RCNN: Instance-Level 3D Object Reconstruction via Render-and-Compare, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Presented at the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Salt Lake City, UT, USA, pp. 3559–3568. https://doi.org/10.1109/CVPR.2018.00375
Lahssini, K., Baghdadi, N., le Maire, G., Fayad, I., 2022. Influence of GEDI Acquisition and Processing Parameters on Canopy Height Estimates over Tropical Forests. Remote Sens. 14, 6264. https://doi.org/10.3390/rs14246264
Lang, N., Jetz, W., Schindler, K., Wegner, J.D., 2023a. A high-resolution canopy height model of the Earth. Nat. Ecol. Evol. https://doi.org/10.1038/s41559-023-02206-6
Lang, N., Schindler, K., Wegner, J.D., 2023b. Canopy top height models at 10m GSD from airborne LIDAR (derived from LVIS and small-footprint ALS). https://doi.org/10.5281/ZENODO.7885699
Lang, N., Schindler, K., Wegner, J.D., 2019. Country-wide high-resolution vegetation height mapping with Sentinel-2. Remote Sens. Environ. 233, 111347. https://doi.org/10.1016/j.rse.2019.111347
Le Toan, T., Quegan, S., Davidson, M.W.J., Balzter, H., Paillou, P., Papathanassiou, K., Plummer, S., Rocca, F., Saatchi, S., Shugart, H., Ulander, L., 2011. The BIOMASS mission: Mapping global forest biomass to better understand the terrestrial carbon cycle. Remote Sens. Environ. 115, 2850–2860. https://doi.org/10.1016/j.rse.2011.03.020
Lefsky, M.A., Cohen, W.B., Harding, D.J., Parker, G.G., Acker, S.A., Gower, S.T., 2002. Lidar remote sensing of above-ground biomass in three biomes: *Biomass estimation by LIDAR*. Glob. Ecol. Biogeogr. 11, 393–399. https://doi.org/10.1046/j.1466-822x.2002.00303.x
Lefsky, M.A., Harding, D.J., Keller, M., Cohen, W.B., Carabajal, C.C., Del Bom Espirito-Santo, F., Hunter, M.O., de Oliveira, R., 2005. Estimates of forest canopy height and aboveground biomass using ICESat: ICESAT ESTIMATES OF CANOPY HEIGHT. Geophys. Res. Lett. 32, n/a-n/a. https://doi.org/10.1029/2005GL023971
Li, S., Brandt, M., Fensholt, R., Kariryaa, A., Igel, C., Gieseke, F., Nord-Larsen, T., Oehmcke, S., Carlsen, A.H., Junttila, S., Tong, X., d'Aspremont, A., Ciais, P., 2023. Deep learning enables image-based tree counting, crown segmentation and height prediction at national scale. PNAS Nexus pgad076. https://doi.org/10.1093/pnasnexus/pgad076
Li, W., Niu, Z., Shang, R., Qin, Y., Wang, L., Chen, H., 2020. High-resolution mapping of forest canopy height using machine learning by coupling ICESat-2 LiDAR with Sentinel-1, Sentinel-2 and Landsat-8 data. Int. J. Appl. Earth Obs. Geoinformation 92, 102163. https://doi.org/10.1016/j.jag.2020.102163

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 Sens. Environ. 284, 113367. https://doi.org/10.1016/j.rse.2022.113367

Liu, A., Cheng, X., Chen, Z., 2021. Performance evaluation of GEDI and ICESat-2 laser altimeter data for terrain and canopy height retrievals. Remote Sens. Environ. 264, 112571. https://doi.org/10.1016/j.rse.2021.112571

Liu, S., Brandt, M., Nord-Larsen, T., Chave, J., Reiner, F., Lang, N., Tong, X., Ciais, P., Igel, C., 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 (preprint). In Review. https://doi.org/10.21203/rs.3.rs-2573442/v1

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Boston, MA, USA, pp. 3431–3440. https://doi.org/10.1109/CVPR.2015.7298965

Loshchilov, I., Hutter, F., 2019. Decoupled Weight Decay Regularization.

Lu, D., Chen, Q., Wang, G., Moran, E., Batistella, M., Zhang, M., Vaglio Laurin, G., Saah, D., 2012. Aboveground Forest Biomass Estimation with Landsat and LiDAR Data and Uncertainty Analysis of the Estimates. Int. J. For. Res. 2012, 1–16. https://doi.org/10.1155/2012/436537

Ma, X., Zheng, G., Chi, X., Yang, L., Geng, Q., Li, J., Qiao, Y., 2023. Mapping fine-scale building heights in urban agglomeration with spaceborne lidar. Remote Sens. Environ. 285, 113392. https://doi.org/10.1016/j.rse.2022.113392

Massa, F., Marlet, R., Aubry, M., 2016. Crafting a multi-task CNN for viewpoint estimation.

Miles, L., Kapos, V., 2008. Reducing Greenhouse Gas Emissions from Deforestation and Forest Degradation: Global Land-Use Implications. Science 320, 1454–1455. https://doi.org/10.1126/science.1155358

Mitchard, E.T.A., Saatchi, S.S., White, L.J.T., Abernethy, K.A., Jeffery, K.J., Lewis, S.L., Collins, M., Lefsky, M.A., Leal, M.E., Woodhouse, I.H., Meir, P., 2012. Mapping tropical forest biomass with radar and spaceborne LiDAR in Lopé National Park, Gabon: overcoming problems of high biomass and persistent cloud. Biogeosciences 9, 179–191. https://doi.org/10.5194/bg-9-179-2012

Morin, D., Planells, M., Baghdadi, N., Bouvet, A., Fayad, I., Le Toan, T., Mermoz, S., Villard, L., 2022. Improving Heterogeneous Forest Height Maps by Integrating GEDI-Based Forest Height Information in a Multi-Sensor Mapping Process. Remote Sens. 14, 2079. https://doi.org/10.3390/rs14092079

Mousavian, A., Anguelov, D., Flynn, J., Kosecka, J., 2016. 3D Bounding Box Estimation Using Deep Learning and Geometry. https://doi.org/10.48550/ARXIV.1612.00496

Mugabowindekwe, M., Brandt, M., Chave, J., Reiner, F., Skole, D.L., Kariryaa, A., Igel, C., Hiernaux, P., Ciais, P., Mertz, O., Tong, X., Li, S., Rwanyiziri, G., Dushimiyimana, T., Ndoli, A., Uwizeyimana, V., Lillesø, J.-P.B., Gieseke, F., Tucker, C.J., Saatchi, S., Fensholt, R., 2023. Nation-wide mapping of tree-level aboveground carbon stocks in Rwanda. Nat. Clim. Change 13, 91–97. https://doi.org/10.1038/s41558-022-01544-w

Neuenschwander, A., Pitts, K., Jelley, B., Robbins, J., Klotz, B., Popescu, S., Nelson, R., Harding, D., Pederson, D., Sheridan, R., 2022. ATLAS/ICESat-2 L3A Land and Vegetation Height. Version 4, 17.

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 Sens. 15, 975. https://doi.org/10.3390/rs15040975

Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., Glocker, B., Rueckert, D., 2018. Attention U-Net: Learning Where to Look for the Pancreas.

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A., 2016. Context Encoders: Feature Learning by Inpainting. https://doi.org/10.48550/ARXIV.1604.07379

Peel, M.C., Finlayson, B.L., McMahon, T.A., 2007. Updated world map of the Köppen-Geiger climate classification. Hydrol. Earth Syst. Sci. 11, 1633–1644. https://doi.org/10.5194/hess-11-1633-2007

Pingel, T.J., Clarke, K.C., McBride, W.A., 2013. An improved simple morphological filter for the terrain classification of airborne LIDAR data. ISPRS J. Photogramm. Remote Sens. 77, 21–30. https://doi.org/10.1016/j.isprsjprs.2012.12.002

Ploton, P., Pélissier, R., Barbier, N., Proisy, C., Ramesh, B.R., Couteron, P., 2013. Canopy Texture Analysis for Large-Scale Assessments of Tropical Forest Stand Structure and Biomass, in: Lowman, M., Devy, S., Ganesh, T. (Eds.), Treetops at Risk. Springer New York, New York, NY, pp. 237–245.

https://doi.org/10.1007/978-1-4614-7161-5_24
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 Sens. Environ. 253, 112165. https://doi.org/10.1016/j.rse.2020.112165
Pourshamsi, M., Garcia, M., Lavalle, M., Balzter, H., 2018. A Machine-Learning Approach to PolInSAR and LiDAR Data Fusion for Improved Tropical Forest Canopy Height Estimation Using NASA AfriSAR Campaign Data. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 11, 3453–3463. https://doi.org/10.1109/JSTARS.2018.2868119
Queinnec, M., White, J.C., Coops, N.C., 2021. Comparing airborne and spaceborne photon-counting LiDAR canopy structural estimates across different boreal forest types. Remote Sens. Environ. 262, 112510. https://doi.org/10.1016/j.rse.2021.112510
Ranftl, R., Bochkovskiy, A., Koltun, V., 2021. Vision Transformers for Dense Prediction.
Reiner, F., Brandt, M., Tong, X., Kariryaa, A., Fensholt, R., 2023. An African-wide map of tree cover at individual tree level (other). display. https://doi.org/10.5194/egusphere-egu23-15488
Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. https://doi.org/10.48550/ARXIV.1505.04597
Roy, D.P., Ju, J., Kline, K., Scaramuzza, P.L., Kovalskyy, V., Hansen, M., Loveland, T.R., Vermote, E., Zhang, C., 2010. Web-enabled Landsat Data (WELD): Landsat ETM+ composited mosaics of the conterminous United States. Remote Sens. Environ. 114, 35–49. https://doi.org/10.1016/j.rse.2009.08.011
Sandberg, G., Ulander, L.M.H., Fransson, J.E.S., Holmgren, J., Le Toan, T., 2011. L- and P-band backscatter intensity for biomass retrieval in hemiboreal forest. Remote Sens. Environ. 115, 2874–2886. https://doi.org/10.1016/j.rse.2010.03.018
Schwartz, M., Ciais, P., Ottlé, C., De Truchis, A., Vega, C., Fayad, I., Brandt, M., Fensholt, R., Baghdadi, N., Morneau, F., Morin, D., Guyon, D., Dayau, S., Wigneron, J.-P., 2022. High-resolution canopy height map in the Landes forest (France) based on GEDI, Sentinel-1, and Sentinel-2 data with a deep learning approach. https://doi.org/10.48550/ARXIV.2212.10265
Schwerdt, M., Schmidt, K., Tous Ramon, N., Klenk, P., Yague-Martinez, N., Prats-Iraola, P., Zink, M., Geudtner, D., 2017. Independent system calibration of Sentinel-1B. Remote Sens. 9, 511.
Stahlschmidt, S.R., Ulfenborg, B., Synnergren, J., 2022. Multimodal deep learning for biomedical data fusion: a review. Brief. Bioinform. 23, bbab569. https://doi.org/10.1093/bib/bbab569
Stereńczak, K., Laurin, G.V., Chirici, G., Coomes, D.A., Dalponte, M., Latifi, H., Puletti, N., 2020. Global Airborne Laser Scanning Data Providers Database (GlobALS)—A New Tool for Monitoring Ecosystems and Biodiversity. Remote Sens. 12, 1877. https://doi.org/10.3390/rs12111877
Tolan, J., Yang, H.-I., Nosarzewski, B., Couairon, G., Vo, H., Brandt, J., Spore, J., Majumdar, S., Haziza, D., Vamaraju, J., Moutakanni, T., Bojanowski, P., Johns, T., White, B., Tiecke, T., Couprie, C., 2023. Sub-meter resolution canopy height maps using self-supervised learning and a vision transformer trained on Aerial and GEDI Lidar.
Topouzelis, K., Singha, S., Kitsiou, D., 2016. Incidence angle normalization of Wide Swath SAR data for oceanographic applications. Open Geosci. 8, 450–464.
Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H., 2021. Training data-efficient image transformers & distillation through attention.
Turner, D.P., Cohen, W.B., Kennedy, R.E., Fassnacht, K.S., Briggs, J.M., 1999. Relationships between Leaf Area Index and Landsat TM Spectral Vegetation Indices across Three Temperate Zone Sites. Remote Sens. Environ. 70, 52–68. https://doi.org/10.1016/S0034-4257(99)00057-7
Vaswani, A., Ramachandran, P., Srinivas, A., Parmar, N., Hechtman, B., Shlens, J., 2021. Scaling Local Self-Attention for Parameter Efficient Visual Backbones.
Wigneron, J.-P., Li, X., Frappart, F., Fan, L., Al-Yaari, A., De Lannoy, G., Liu, X., Wang, M., Le Masson, E., Moisy, C., 2021. SMOS-IC data record of soil moisture and L-VOD: Historical development, applications and perspectives. Remote Sens. Environ. 254, 112238. https://doi.org/10.1016/j.rse.2020.112238
Woodcock, C.E., Collins, J.B., Gopal, S., Jakabhazy, V.D., Li, X., Macomber, S., Ryherd, S., Judson Harward, V., Levitan, J., Wu, Y., Warbington, R., 1994. Mapping forest vegetation using Landsat TM imagery and a canopy reflectance model. Remote Sens. Environ. 50, 240–254. https://doi.org/10.1016/0034-4257(94)90074-4
Yu, F., Koltun, V., 2016. Multi-Scale Context Aggregation by Dilated Convolutions.

Yuan, K., Guo, S., Liu, Z., Zhou, A., Yu, F., Wu, W., 2021. Incorporating Convolution Designs into Visual Transformers.

Zhang, J., Nielsen, S.E., Mao, L., Chen, S., Svenning, J.-C., 2016. Regional and historical factors supplement current climate in shaping global forest canopy height. J. Ecol. 104, 469–478. https://doi.org/10.1111/1365-2745.12510

Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T., 2022. iBOT: Image BERT Pre-Training with Online Tokenizer.