



COMPUTER SCIENCE MSc

MASTER'S THESIS

GLOBAL BIOMASS ESTIMATION  
AND UNCERTAINTY QUANTIFICATION  
WITH MULTI-TASK BAYESIAN  
DEEP ENSEMBLES

Supervisors: Prof. Dr. Konrad Schindler & Prof. Dr. Jan Dirk Wegner

Advisors: Nikolai Kalischek & Yuchang Jiang

## Abstract

Every year, 15 billion trees are lost, 25% of which to deforestation, primarily in Latin America and Southeast Asia. This has significant impacts on biodiversity loss and the climate crisis, as tropical forests are home to over half of the world's species and act as a major carbon sink. Carbon offsetting, which allows landowners to monetize the carbon stored in their trees and soils, is currently the main strategy for incentivizing and financing conservation and restoration efforts. Carbon stocks are derived from above-ground biomass density (AGBD), which raises the question of how to accurately estimate biomass. Traditional methods, such as manually measuring sample trees and using allometric equations, are time-consuming, expensive, and prone to human error. Machine learning (ML) methods, on the other hand, can leverage openly accessible satellite imagery available globally and frequently. They have the potential to reduce the time and cost needed to produce biomass maps, and to do so more accurately. This work evaluates the performance of ML methods to estimate AGBD from optical satellite imagery, leveraging Multi-task learning (MTL), Bayesian learning, and deep ensembles. Consistent with prior results, this work highlights that satellite imagery can be leveraged to estimate AGBD globally. It also shows that leveraging MTL can improve the estimation, at least in low-data regime. We find the quality of the estimates to be highly dependent on the input data, encouraging further data collection. Both in terms of number of data samples, and in terms of leveraging additional data sources with different modalities.

## Acknowledgments

I would like to express my sincere gratitude to my supervisors, Prof. Dr. Konrad Schindler and Prof. Dr. Jan Dirk Wegner, for their guidance, support, and encouragement throughout the course of this research. Their expertise and insights have been instrumental in shaping the direction and outcome of this thesis. Thank you for the opportunity to work on such an interesting and challenging topic; I look forward to continue working with you!

I would also like to thank Nikolai Kalischek, for his contributions and support. His insights, feedbacks, and bug-detection skills have been essential in helping me focus on the key issues and to improve the quality of this work. He also made me the coffee-drinker I am today. Many thanks to Yuchang for her support and insights. I am so grateful to everyone at PRS for the warm welcome, numerous coffee breaks, occasional beers, daily lunches, and this *one* foosball game. Those past six months wouldn't have been the same without you.

Bonus thanks to Thomas Gazel-Anthoine for showing me the ropes upon my arrival, and proof-reading this work. Missed typos are on him.

# Contents

<b>1</b>	<b>Introduction</b>	<b>7</b>
1.1	Organizational details . . . . .	7
1.2	Motivation . . . . .	7
1.3	Problem statement . . . . .	8
1.4	Outline . . . . .	9
<b>2</b>	<b>Related Work</b>	<b>10</b>
2.1	Machine learning and remote sensing for biomass estimation . . . . .	10
2.2	Uncertainty Estimation in Deep Learning . . . . .	11
2.3	Multi-task Learning . . . . .	12
<b>3</b>	<b>Data</b>	<b>15</b>
3.1	Data sources . . . . .	15
3.1.1	GEDI data products . . . . .	15
3.1.2	Sentinel-2 imagery . . . . .	17
3.1.3	Canopy height map . . . . .	18
3.1.4	ESA CCI . . . . .	18
3.1.5	CMS FIA . . . . .	18
3.2	Datasets . . . . .	18
3.2.1	Clemence . . . . .	18
3.2.2	Yuchang . . . . .	19
3.3	Variable importance analysis . . . . .	19
3.3.1	Dataset . . . . .	19
3.3.2	Method . . . . .	20
3.3.3	Results . . . . .	20
3.4	Dataset for MTL . . . . .	23
3.4.1	Dataset creation . . . . .	23
3.4.2	Geographical split . . . . .	23
3.4.3	Comparison with the other datasets . . . . .	23
<b>4</b>	<b>Methods</b>	<b>27</b>
4.1	Models . . . . .	27
4.1.1	Heads . . . . .	27
4.1.2	Fully Convolutional Neural network . . . . .	28
4.1.3	ResNeXt . . . . .	28
4.1.4	Xception . . . . .	29
4.2	Training procedure . . . . .	30
4.2.1	Training losses . . . . .	31
4.2.2	MTL Weighting schemes . . . . .	32
4.3	Evaluation scheme . . . . .	32
4.3.1	Evaluation metrics . . . . .	32
4.3.2	Comparison with other biomass maps . . . . .	33
<b>5</b>	<b>Results and Discussion</b>	<b>34</b>
5.1	Experiments . . . . .	34
5.2	Cross-datasets evaluation . . . . .	37
5.3	Visual assessment . . . . .	37
5.4	Comparison with other biomass maps . . . . .	40

<b>6</b>	<b>Conclusion</b>	<b>41</b>
6.1	Outcomes . . . . .	41
6.2	Limitations . . . . .	41
6.2.1	Measuring biomass . . . . .	41
6.2.2	Dataset . . . . .	42
6.2.3	Methods . . . . .	42
6.3	Future Steps . . . . .	43
6.3.1	Working with big data . . . . .	43
6.3.2	Biomass prediction . . . . .	43
6.3.3	Biomass change detection . . . . .	43

## List of Figures

1	Annual Deforestation, 2015. Our World in Data [1]. . . . .	7
2	Procedure for calculating the amount of Carbon Stock of a forest [2]. . . . .	8
3	Frequency of CNN-based studies in the context of agriculture, forestry, and conservation [3] . . . . .	10
4	Example of hard parameter sharing architecture [4] . . . . .	13
5	Example of soft parameter sharing architecture [4] . . . . .	13
6	Example subset of aboveground biomass density predictions from the GEDI Level-4A footprint product over Northern California, U.S., spanning April to July 2019 [5] . . . . .	15
7	Locations at which LiDAR and field data are available for calibration of GEDI biomass equation. Blue indicates publicly available datasets, orange indicates private datasets. The size of each marker (circle) indicates the number of GEDI footprint size field plots that are available at the calibration location [5]. . . . .	16
8	Sample Sentinel-2 UTM Tiling Grid, ESA . . . . .	17
9	RGB bands from a Sentinel-2 tile (59GLL) . . . . .	17
10	Global canopy top height map for the year 2020, visualized in Equal-Earth projection [6] . . . . .	18
11	Dataset AGBD values distribution. AGBD values up to $500 \text{ Mg} \cdot \text{ha}^{-1}$ are binned in 25 $\text{Mg} \cdot \text{ha}^{-1}$ bins; AGBD values large than $500 \text{ Mg} \cdot \text{ha}^{-1}$ are binned together. . . . .	21
12	Pearson correlation coefficients for a subset of the variables. . . . .	21
13	Variable importance analysis for the best performing LightGBM model. . . . .	21
14	Variable importance analysis for the second best performing LightGBM model. . . . .	22
15	Geographical distribution of the data points. The red, green, and blue colors respectively correspond to the train, test, and validation data. . . . .	24
16	Distribution of Clémence’s dataset AGBD values, constrained to the May-September range. . . . .	25
17	My train dataset AGBD values distribution ( $5 \text{ Mg} \cdot \text{ha}^{-1}$ bins) . . . . .	26
18	My test dataset AGBD values distribution ( $5 \text{ Mg} \cdot \text{ha}^{-1}$ bins) . . . . .	26
19	Clemence’s train dataset AGBD values distribution ( $5 \text{ Mg} \cdot \text{ha}^{-1}$ bins) . . . . .	26
20	Clémence’s test dataset AGBD values distribution ( $5 \text{ Mg} \cdot \text{ha}^{-1}$ bins) . . . . .	26
21	Yuchang’s train (sensitive) dataset AGBD values distribution ( $5 \text{ Mg} \cdot \text{ha}^{-1}$ bins) . . . . .	26
22	Yuchang’s test (sensitive) dataset AGBD values distribution ( $5 \text{ Mg} \cdot \text{ha}^{-1}$ bins) . . . . .	26
23	Architecture of the Multi Task Fully Convolutional Neural network (FCN). . . . .	28
24	Architecture of the Multi Task ResNeXt network. . . . .	29
25	Architecture of the Multi Task Xception network. . . . .	30
26	Predictive uncertainty vs. Empirical RMSE for an uncertainty-weighted MTL ResNeXt model (left), and a GNLL single-task ResNeXt model (right). . . . .	35
27	Sample prediction displaying the cell-like pattern (tile 59GLL). Right: zoomed. . . . .	36
28	FIA AGB values distribution, for both scenarios. . . . .	40
29	Binned residuals for the average-pooled model, for the FIA optimistic (top) and pessimistic (bottom) scenarios. . . . .	51
30	Binned residuals for the max-pooled model, for the FIA optimistic (top) and pessimistic (bottom) scenarios. . . . .	51
31	Schematic of the framework developed by [7]. . . . .	52
32	The GEDI04_A [5] global stratification of plant functional types (PFT) . . . . .	52
33	Intersection procedure pseudo-code. . . . .	53

## List of Tables

1	Sentinel-2 Level-2A Bands, Google Earth Engine . . . . .	17
2	LightGBM models description and performance metrics. A ticked box indicates the model was trained with the variable at hand. <i>lat &amp; lon</i> indicates that both latitude and longitude were used, as opposed to only latitude; <i>canopy</i> includes <i>canopy_cover</i> , <i>pai</i> , and <i>fhd_normal</i> ; <i>dem</i> is the <i>digital_elevation_model</i> ; and <i>rh</i> indicates which relative height metrics were used. . . . .	22
3	Number of data points per dataset, and their respective train/test/splits. <i>Yuchang</i> refers to the data from [8] and <i>Clemence</i> refers to the data from [9]. . . . .	24
4	Number of trainable parameters per model architecture. . . . .	27
5	The models' test performance metrics. <i>CH</i> indicates that Canopy Height was an input feature. <i>S2</i> indicates that the Sentinel-2 bands were input features. <i>MTL</i> describes the MTL weighting strategy, where a blank cell indicates single-task learning. . . . .	34
6	Cross-datasets test metrics. . . . .	37

# 1 Introduction

## 1.1 Organizational details

This Master thesis was undertaken at ETH Zürich, within the Photogrammetry and Remote Sensing group and EcoVision Lab, under the supervision of Prof. Dr. Konrad Schindler and Prof. Dr. Jan Dirk Wegner. It builds on the work of Nico Lang, Yuchang Jiang, and Clémence Lanfranchi.

## 1.2 Motivation

Fifteen billion trees are lost every year [10], of which one-quarter to deforestation [11]. Ninety-five percent of that deforestation occurs in the tropics, mainly in Latin America and Southeast Asia (Fig 1). The resulting impact on biodiversity loss and the climate crisis is unspeakable: not only does over half of the world's species reside in tropical forests [12] but tropical forests are among the biggest carbon sinks to exist. As a consequence, when deforestation happens, almost all of the carbon stored in the trees and vegetation is lost. At a time where carbon emissions must be reduced, the potential of massive forest conservation and restoration for climate mitigation and the protection of biodiversity is clear [13].

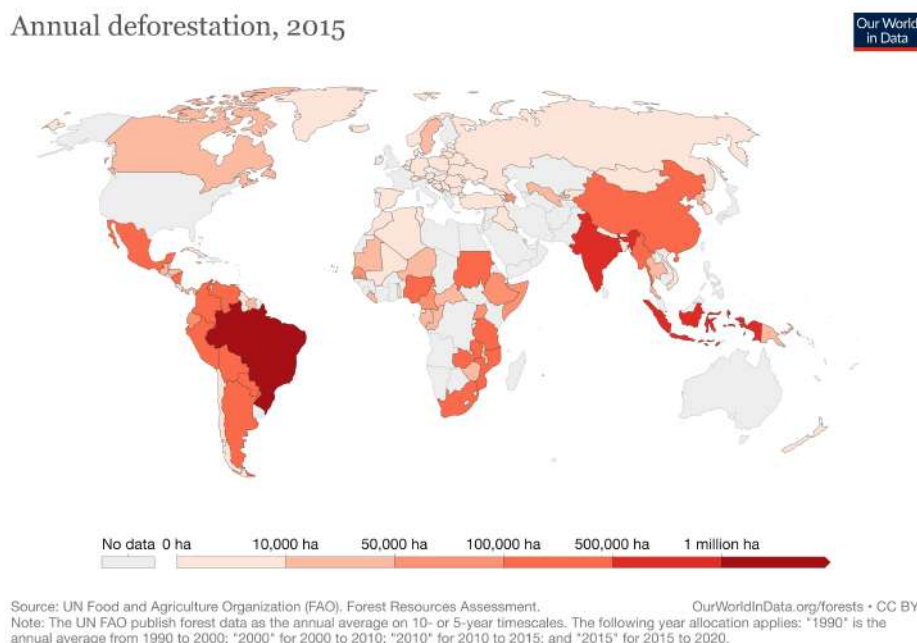


Figure 1: Annual Deforestation, 2015. Our World in Data [1].

The current strategy to incentivize and finance the needed conservation and restoration is through carbon offsetting, although it comes with its share of limitations. A discussion of the ethics and effectiveness of carbon offsetting is out of the scope of this work, however, such a discussion can be found in [14]. Forest carbon offsets allow landowners to monetize the carbon being stored by their trees and soils, and sell it. The effectiveness of this financial tool is based on the accurate and consistent estimation of the additional carbon storage that should be expected under an offset deal. This estimation thus mainly relies on measuring how much carbon is stored at a given time.



### 1.3 Problem statement

The standardized procedure for forest carbon stock inventory consists of manually registering sample trees of the project site, and measuring tree parameters (e.g., diameter at breast height (DBH), height, species). Those parameters are then put through regression models called *allometric equations* which are tailored to the forest type and region, to estimate the living vegetation above the soil, or Above Ground Biomass (AGB), of the project site and the corresponding carbon stock (as seen in Figure 2). Those allometric models are usually constructed using calibration data where the tree parameters are measured concurrently with the AGB. For the latter, it consists in a destructive measurement process which relies on harvesting trees, drying them, and weighing the biomass.

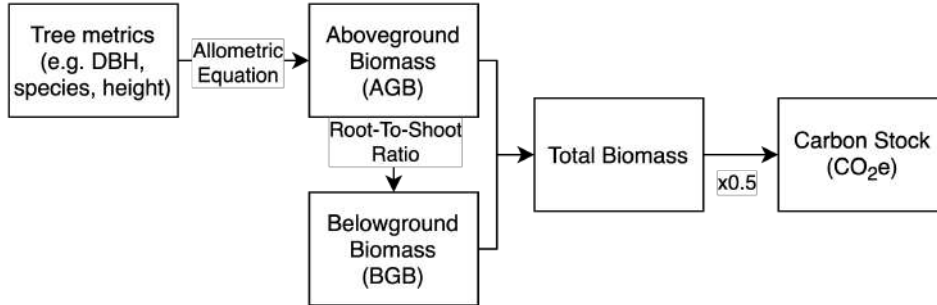


Figure 2: Procedure for calculating the amount of Carbon Stock of a forest [2].

Another approach consists of leveraging remote sensing (RS) technologies to estimate the tree parameters used as input for the allometric equations. Such technologies can be separated into optical (passive) and radar (active) sensors [15].

- **Optical** RS imagery gives spectral information on the horizontal forest structure [16]. Such sensors are commonly used in space-borne platforms such as Landsat [17] or Sentinel satellites.
- **Radar** RS imagery gives information on the vertical forest structure, due to its ability to penetrate the forest canopy [18]. Common active sensors include SAR (Synthetic Aperture Radar) and LiDAR (Light Detection and Ranging) [9]. Because SAR sensors saturate at lower biomass levels than LiDAR sensors [19], LiDAR data is more prevalently used to estimate forest AGB [20, 21]. This is done either via space-borne platforms, airborne laser scanning, or terrestrial laser scanning (TLS).

Such practices, however, are highly sensitive to the choice of allometric equation [22], and recent research [23] suggests that questioning the underlying assumptions of allometric models is an urgent priority. Indeed, allometric estimates of biomass are usually biased. Various studies [24, 25] have experimented with terrestrial laser scanning (TLS) as an alternative to directly estimate AGB, and have reported that TLS estimates of AGB were less biased and more accurate than those from allometric models, especially for larger trees.

For those reasons, machine learning (ML) methods are becoming more prevalent for performing biomass estimation [15]. As openly accessible satellite imagery that offers longer-term global coverage with high spatial and temporal resolution is becoming increasingly available, it is being widely combined with ML methods for various applications. Such practices [26, 27] reduce the time and cost needed to produce biomass maps, and circumvent their reliance on allometric equations. Previous studies [28] have even shown that LiDAR-derived AGB estimates can provide up to a 15-fold increase in accuracy over allometric counterparts. However, assessing the quality of the predictions is challenging due a lack of appropriate reference data and evaluation framework; we elaborate on those limitations in Section 6. This work focuses on estimating above-ground biomass based on optical satellite imagery,

and LiDAR-based reference data. We build on previous works [8, 9] with the novelty of attempting to learn solely from satellite data, by leveraging Multi-Task Learning (MTL). We quantify the uncertainty linked with the estimation by using Bayesian deep ensembles, to make the predictions more interpretable.

## 1.4 Outline

The remainder of this work is organized as follows: Section 2 presents the latest literature on the topics at hand; Section 3 describes the data sources and datasets used throughout this work; Section 4 defines the proposed methods and corresponding experiments; Section 5 presents the results of said experiments; and finally, in Section 6 we discuss our results' impact and limitations.

## 2 Related Work

In this chapter, we provide an overview of the existing work in terms of biomass mapping using machine learning methods (Section 2.1), uncertainty estimation in deep learning (Section 2.2), and multi-task learning (Section 2.3).

### 2.1 Machine learning and remote sensing for biomass estimation

Forests are very diverse ecological systems that display complex behavior across different temporal and spatial scales. As discussed in Section 1.2, non-parametric machine learning algorithms make fewer assumptions on the shape and distribution of the reference data, and tend to outperform parametric methods. Machine learning models can thus be used to estimate the amount and spatial distribution of biomass. Furthermore, the increasing availability of high spatial and temporal resolution remote sensing data has revolutionized the way research is being conducted. RS-based ML methods are applied to a wide spectrum of thematic categories (Figure 3). We hereby summarize the work most closely related to ours, and refer the interested readers to the review by [3] for additional background.

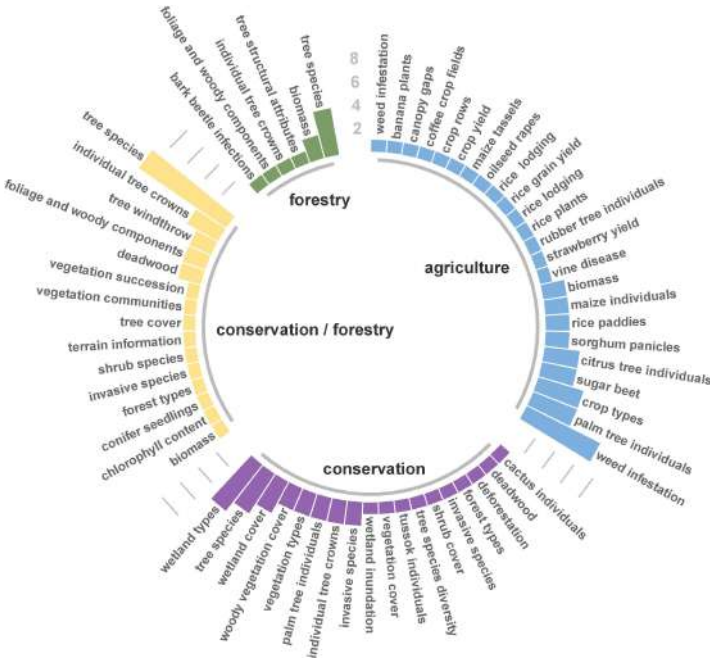


Figure 3: Frequency of CNN-based studies in the context of agriculture, forestry, and conservation [3]

In recent years, several maps of AGB have been produced by RS-based ML methods:

- The Global Forest Watch (GFW) Aboveground Live Woody Biomass Density dataset [27] is a global map of AGBD at 0.00025-degree (approximately 30-meter) resolution for the year 2000. Its authors first calculated AGBD at more than seven hundred thousand points with LiDAR and regional allometric equations, then used those data points to train a wall-to-wall ML model based on Landsat imagery. However, such a high-resolution, global map is yet to be produced on a yearly basis, although there is a clear need to monitor biomass loss globally at a high temporal resolution.
- The authors of [6] proposed a probabilistic deep learning approach to regress canopy top height globally (described in Section 3.1.3). Building on this work and leveraging the fact that canopy height metrics derived from LIDAR data are strongly correlated with AGBD [29], the authors

of [9] developed a probabilistic deep learning model to estimate GEDI-derived AGBD [5] from the canopy top height predictions, and quantify the uncertainty in those estimates. As this approach is not wall-to-wall, however, it risks accumulating uncertainties.

- Research scientists at Sylvera<sup>1</sup> have built their very-own AGB reference dataset. Using multi-scale LiDAR (MSL) methods, they reconstruct and model tree-scale parameters such as AGB. With this data, they calibrate GEDI footprints [5] and enhance their reference dataset. They subsequently rely on ML methods, SAR, and multispectral optical satellite imagery, to upscale their biomass measurements to other time periods and over large areas. However, as a company, Sylvera does not make any of this data openly accessible.

Plenty of scientists are working towards creating global biomass maps, each with different assumptions, goals, and methods. Our approach attempts to do so by relying solely on, and directly from, satellite imagery; making it as interpretable as possible; and with the goals of producing it on a yearly basis, and providing it openly to the scientific community.

## 2.2 Uncertainty Estimation in Deep Learning

Measuring the uncertainty of DL models is crucial, as it allows for a better understanding of the model’s predictions and its confidence in those predictions. This can help to increase the transparency and interpretability of models, which can be beneficial for a wide range of applications. For example, in safety-critical applications such as self-driving cars or medical diagnosis, understanding the model’s uncertainty can help to ensure that the model only takes actions when it is highly confident in its predictions, and that it does not take dangerous or harmful actions when it is uncertain. In situations where the model’s predictions are used to make important decisions, such as financial or legal decisions, understanding the model’s uncertainty can help to ensure that the right level of caution is exercised when making decisions based on the model’s predictions. In the context of this work, we want our global AGBD map to be as accurate and interpretable as possible, to eventual inform decision-making and serve forest conservation efforts.

In this section, we explain how uncertainty can be estimated in DL models.

**Aleatoric uncertainty estimation** The first type of uncertainty that can be captured is *aleatoric* uncertainty. It refers to the noise inherent to the observations. This type of uncertainty cannot be reduced by collecting more data or increasing the complexity of the model. For example, in image classification, there may be aleatoric uncertainty in the output of a model due to variations in lighting or camera angle.

Aleatoric uncertainty can be further divided into two sub-categories: *heteroscedastic* uncertainty, which depends on the input data and can be predicted as a model output; and *homoscedastic* uncertainty, which is constant regardless of the input data and is not a model output. Homoscedastic regression assumes constant noise distribution across the dataset, while heteroscedastic regression assumes that observation noise can vary with the input [30]. This is particularly useful in cases where parts of the observation space might have higher noise levels than others, for example in a study of climate change, where the variability of temperature might vary across different regions and over different time periods. Accounting for this heteroscedasticity would be important to ensure that the model accurately captures the patterns across different regions and time periods, and avoids underestimating the uncertainty of the predictions in high-variability areas and periods.

In practice, we model the aleatoric noise by a Gaussian defined by its mean  $\hat{\mu}$  and variance  $\hat{\sigma}^2$ , as an approximation of the true conditional distribution  $p(y|x) \approx \mathcal{N}(\hat{\mu}(x), \hat{\sigma}^2(x))$ . Then, for a given input  $x$ ,

---

<sup>1</sup><https://www.sylvera.com/blog/mapping-forest-structure-across-the-landscape?>

we can estimate the aleatoric uncertainty in the prediction by the variance of the model,  $\hat{\sigma}^2(x)$ . This is done via minimizing the negative log-likelihood:

$$\mathcal{L}_{NN} = \frac{1}{N} \sum_{i=1}^N \frac{1}{2\sigma^2(x_i)} \|y_i - \mu(x_i)\|^2 + \frac{1}{2} \log \sigma^2(x_i)$$

**Epistemic uncertainty estimation** The second type of uncertainty that can be captured is *epistemic* uncertainty. It captures what our model does not know due to lack of training data. For example, in image classification, a model may be uncertain about the output of an image if it has not seen enough examples of that particular class during training.

Researchers from Google and DeepMind [31] have found that deep ensembles [32] seem to outperform other more widespread approaches to measuring epistemic uncertainty, like Monte Carlo dropout [33]. The idea is to construct an ensemble of  $M$  Bayesian neural networks: we replace the deterministic networks' weights parameters with Gaussian distributions over these parameters, and instead of optimising the networks' weights directly we average over all possible weights. More concretely, we put a prior distribution over each network's weights [30], e.g., a Gaussian prior:  $W \sim \mathcal{N}(0, I)$ . Then, the  $M$  networks are trained independently on the entire dataset using random initialization. Given an input, we thus obtain  $\{\hat{\mu}_m\}_{m=1}^M$  estimates of the mean, and  $\{\hat{\sigma}_m^2\}_{m=1}^M$  estimates of the variance. The epistemic uncertainty can then be quantified by the variance in the predictions of the mean:

$$\frac{1}{M} \sum_{m=1}^M (\hat{\mu}_m - \frac{1}{M} \sum_{m=1}^M \hat{\mu}_m)^2$$

**Predictive uncertainty estimation** Predictive uncertainty is the overall uncertainty in the model's predictions. With deep ensembles, it can be quantified as:

$$\hat{\sigma}^2 = \underbrace{\frac{1}{M} \sum_{m=1}^M (\hat{\mu}_m - \frac{1}{M} \sum_{m=1}^M \hat{\mu}_m)^2}_{\text{epistemic uncertainty}} + \underbrace{\frac{1}{M} \sum_{m=1}^M \hat{\sigma}_m^2}_{\text{aleatoric uncertainty}}$$

## 2.3 Multi-task Learning

We hereby provide some background knowledge on MTL, and refer the interested readers to [34] for an exhaustive overview.

**Definition** Formally, multi-task learning [35] aims to learn  $m$  functions  $\{f_i(x)\}_{i=1}^m$  for  $m$  tasks where all the tasks (or a subset of them) are related but not identical. The motivation behind this approach stems from Stein's paradox [36], which states that it is better to estimate the means of three or more Gaussian random variables using samples from all of them rather than estimating them separately, even when the Gaussians are independent. Several works [37, 38] have proved that combining multiple tasks to form a MTL model is effective for each single task learning and will help the model converge faster. Additionally, MTL effectively increases the size of the dataset, by combining datasets from each task. MTL can thus exploit useful information from related learning tasks to help alleviate data sparsity problems [34].

**Parameter sharing** More specifically, in the context of Deep Learning, MTL is typically done with either hard or soft parameter sharing of hidden layers:

- Hard parameter sharing (see Figure 4) consists in learning a common space representation for all tasks while keeping additional task-specific parameters that are learned independently for

each task. Hard parameter sharing acts as regularization and reduces the risk of overfitting, as the model learns a representation that will (hopefully) generalize well for all tasks. It also allows to implement simpler solutions to reduce computational cost and inference latency; at the same time, it potentially improves generalization by implicit bias induction. However, it may suffer from optimization conflicts caused by task differences because tasks may compete for the same set of parameters in the shared layers. Hard parameter sharing is the most widely used approach, and most works in the literature share bottom layers of a deep neural network while using separate top layers for each task. This is seemingly motivated by the fact that researchers [41, 39, 24, 29] believe that bottom layers serve as low-level feature extractors (which should be able to be shared across multiple tasks) while top layers are more sensitive to the input data (so that different tasks should possess distinct top layers to generate diverse high-level features). However, a recent publication [39] is questioning this assumption. Its authors found that using separate bottom-layer parameters for each task and sharing a small proportion of task-specific top-layer ones could achieve significantly better performance than the common practice.

- In soft parameter sharing (see Figure 5), each task has its own model with its own set of parameters, while some constraints are applied to the parameters to encourage them to have similar values. This approach is typically used when the tasks are less related and it is expected that each task will have its own set of features, by allowing some degree of specialization for each task while still sharing some information across tasks. It is usually done via the introduction of a regularization term in the loss function:

$$\mathcal{L} = l + f(\{W^{(j)}\}_{j=1}^m)$$

where  $l$  is the original loss function for all tasks, and  $f$  is a regularization function on the tasks' parameters  $\{W^{(j)}\}_{j=1}^m$ . This function can be a simple  $L_2$  regularization, the squared Frobenius norm [40], or the nuclear or trace norm [41].

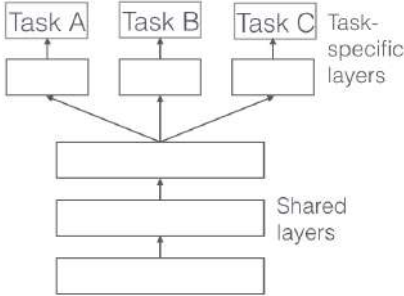


Figure 4: Example of hard parameter sharing architecture [4]

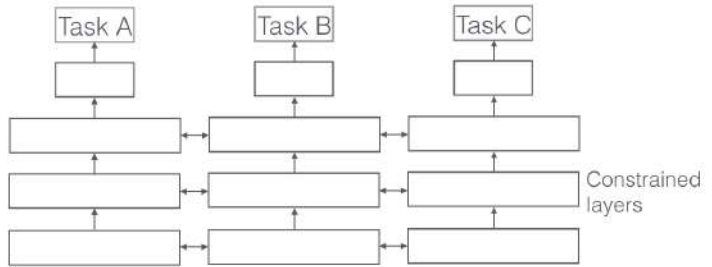


Figure 5: Example of soft parameter sharing architecture [4]

**Multi-objective optimization** In most of the literature, MTL models are optimized using a weighted linear combination of per-task losses. The main challenge of this approach is to define how the gradients of different tasks should be combined. We can formally formulate this problem as:

$$\mathcal{L} = \sum_{i=1}^m w_i \mathcal{L}_i, \quad \sum_{i=1}^m w_i = 1$$

where  $w_i$  is the weight assigned to the  $i$ -th task, and  $\mathcal{L}_i$  is the loss of the  $i$ -th task. Most approaches rely on a naïve weighted sum of losses, some by fixing  $w_i = w_j \forall i, j$  [42–46], and others by setting differentiated weights through "trials and experiments" or by leveraging prior knowledge on the tasks [47–49]. However, the authors of [50] show that performance is highly dependent on an

appropriate choice of weighting between each task’s loss. They propose a different approach: to use homoscedastic uncertainty (described in Section 2.2) to weight the losses. The overall loss thus becomes:

$$\mathcal{L} = \sum_{i=1}^m \frac{1}{h_i \sigma_i^2} \mathcal{L}_i + \log(\sigma_i)$$

At this point, one has to distinguish between regression tasks and classification tasks. In the case of a regression task,  $p(y|\mu(x)) = \mathcal{N}(\mu(x), \sigma_i(x))$  where  $\mu(x)$  is the mean variable predicted by the  $i$ -th task,  $\sigma_i(x)$  is the variance for the variable predicted by the  $i$ -th task, and  $h_i = 2$ . In this case,  $\mathcal{L}_i = \|y - \mu(x)\|^2$ , with  $y$  the target variable. In the case of a classification task,  $p(y|\mu(x), \sigma_i) = \text{Softmax}(\frac{1}{\sigma_i} \mu(x))$  where  $\mu(x)$  is the mean variable predicted by the  $i$ -th task,  $\sigma_i$  is a positive scalar (the *temperature*) that determines how flat the discrete distribution for the variable predicted by the  $i$ -th task is, and  $h_i = 1$ . In this case,  $\mathcal{L}_i = \|y - \mu(x)\|^2$ , with  $y$  the target variable.

**Applications** Many real-world applications have successfully applied MTL: the authors of [42] developed a deep multi-task encoder-transformer-decoder architecture for semantic change detection; the authors of [48] mapped greenhouses across China from high resolution remote sensing imagery using a dual-task DL framework; the authors of [47] performed relative height estimation and semantic segmentation from single airborne RGB images using MTL; the authors of [45] leveraged physics-informed MTL to simultaneously estimate multiple air pollutants from satellite imagery; the authors of [46] simultaneously detected plant species and predicted plant disease with a DL-based MTL model; the authors of [43] leveraged MTL for remote sensing scene classification when annotated data is limited by using related tasks as a regularizer; the authors of [49] used DL-based MTL to predict ship emissions and fuel sulfur content based on imaging; and the authors of [44] simultaneously tackled land use land cover classification and wastewater treatment plants detection using a dual-task DL framework and remote sensing data.

To the best of our knowledge, MTL has not yet been applied to the high-resolution estimation of AGBD.

## 3 Data

### 3.1 Data sources

In the present section, we describe the various data sources used throughout this work.

#### 3.1.1 GEDI data products

The Global Ecosystem Dynamics Investigation (GEDI) produces high resolution laser ranging observations of the 3D structure of the Earth. It is the first satellite-based LiDAR system specifically designed to study vegetation structure. The GEDI instrument is comprised of 3 lasers: one coverage laser optically split into two beams (“coverage” beams) and two lasers operating at full power (“power” beams). Each laser fires 242 times per second and illuminates a 25m spot (a footprint) on the surface over which 3D structure is measured. The coverage laser produces four tracks of footprints, and each power laser produces two tracks of footprints; resulting in 8 parallel tracks of observations. Each footprint is separated by 60 m along track, with an across-track distance of about 600 m between each of the 8 tracks. An example can be found in Fig 6. GEDI produced billions of cloud-free observations during its mission length (Dec. 2018 - Jan. 2023), spanning the Earth’s surface between 51.6° N and 51.6° S. It is now heading into storage on the International Space Station, where it will remain in hibernation for approximately 13-18 months. GEDI will be back at its original location in the fall of 2024 and collecting data again, potentially through 2030.

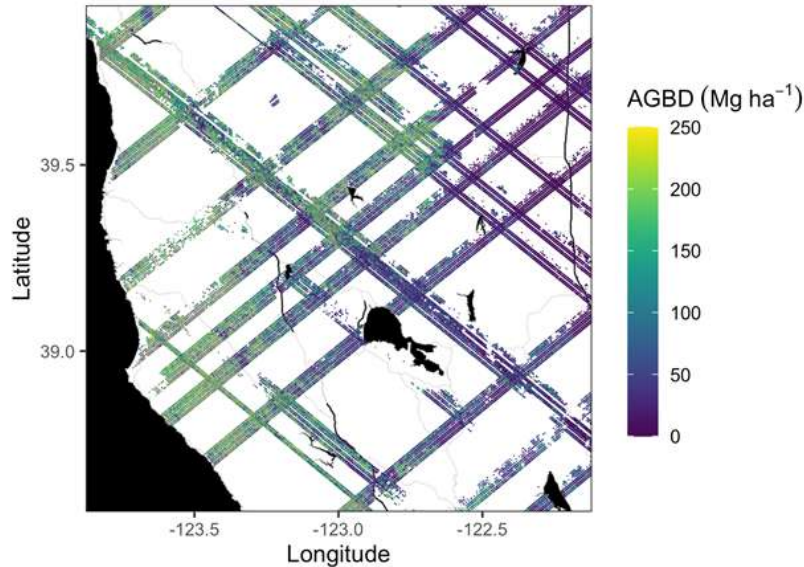


Figure 6: Example subset of aboveground biomass density predictions from the GEDI Level-4A footprint product over Northern California, U.S., spanning April to July 2019 [5]

GEDI data products include footprint and gridded datasets that describe the 3D features of the Earth. These data products are assigned different levels, which indicate the amount of processing that the data has undergone after collection. In the following paragraphs, a *waveform* designates a digitally-recorded return laser pulse that represents the time history of the laser pulse as it interacts with the reflecting surfaces.

**L1 - Waveforms** The GEDI Level 1 data products are developed in two separate products, a Level 1A (L1A) and a Level 1B (L1B) product. The L1A data product contains the raw



waveform and geolocation information used to compute higher level data products. The L1B data product includes the corrected receive waveform, as well as the receive waveform geolocation information.

**L2 - Footprint level canopy height and profile metrics** The waveforms are processed to provide canopy height (L2A) and profile (L2B) metrics. These values are calculated directly from the waveform return for each footprint and provide easy-to-use and interpret information about the vertical distribution of the canopy material.

**L3 - Gridded canopy height metrics and variabilities** Level 3 products are gridded by spatially interpolating the Level 2 footprint estimates of canopy height and profile metrics, and their uncertainties.

**L4 - Aboveground carbon estimates** Footprint metrics derived from the L2 data products are converted to footprint estimates of aboveground biomass density (L4A) using calibration equations. Subsequently, these footprints are used to produce mean biomass and its uncertainty in cells of 1 km (L4B) using statistical theory.

**Calibration and Validation** GEDI relies on world-wide crowd-sourced *in situ* and airborne datasets to develop representative pre-launch calibration equations for predicting AGB across the GEDI observation domain. These LiDAR data, collected with an airborne/drone LiDAR system or a terrestrial laser scanning (TLS) instrument, are used to simulate GEDI-like waveforms and derived metrics. These simulated waveform metrics are then used to calibrate the biomass equations with the field biomass values as reference data (see Figure 7 for an overview of the available calibration datasets). Such forest inventory plots are usually measured over long periods of time, using different methodologies, sampling designs, sizes, and shapes. This can lead to large discrepancies when assessing different projects and regions. Additionally, we generally lack reference data from many areas of the world’s forests due to inaccessibility and/or very high cost. This makes access to good-quality reference data one of the main challenges when monitoring biomass stocks. Therefore we avoid the term *ground truth data* and prefer instead the term *reference data*.



Figure 7: Locations at which LiDAR and field data are available for calibration of GEDI biomass equation. Blue indicates publicly available datasets, orange indicates private datasets. The size of each marker (circle) indicates the number of GEDI footprint size field plots that are available at the calibration location [5].



Figure 8: Sample Sentinel-2 UTM Tiling Grid, ESA<sup>3</sup>



Figure 9: RGB bands from a Sentinel-2 tile (59GLL)

Band	Resolution	Description
B1	60 m	Aerosols
B2	10 m	Blue
B3	10 m	Green
B4	10 m	Red
B5	20 m	Red Edge 1
B6	20 m	Red Edge 2
B7	20 m	Red Edge 3
B8	10 m	Near Infrared (NIR)
B8a	20 m	Red Edge 4
B9	60 m	Water vapor
B11	20 m	Short Wave Infrared (SWIR) 1
B12	20 m	SWIR 2

Table 1: Sentinel-2 Level-2A Bands, Google Earth Engine<sup>4</sup>

### 3.1.2 Sentinel-2 imagery

Sentinel-2 is a European, high-resolution, multi-spectral imaging mission with a revisit frequency of 5 days at the Equator. The resulting tiles are  $100 \times 100 \text{ km}^2$  orthoimages<sup>2</sup> in UTM/WGS84 projection, a sample of which can be found in Figure 8. Sentinel-2 provides users with two products: Level-1C (Top-of-atmosphere reflectances in cartographic geometry) and Level-2A (Bottom-of-atmosphere reflectance in cartographic geometry). The latter comprises 12 spectral bands ranging from the visible to the shortwave infrared spectrum, with a spatial resolution of up to 10 meters per pixel, as described in Table 1. Additional outputs are an Aerosol Optical Thickness (AOT) map, a Water Vapour (WV) map and a Scene Classification (SCL) map together with Quality Indicators (QI) for cloud and snow probabilities at 60 m resolution. With its frequent revisit time, Sentinel-2 data enables the monitoring of environmental changes on a global scale, and has become a valuable resource for the remote sensing community.

<sup>2</sup>computer-generated images of an aerial photograph in which displacements (distortions) caused by terrain relief and camera tilts have been removed, <https://www.usgs.gov/faqs/what-digital-orthophoto-quadrangle-dog-or-orthoimage>

<sup>3</sup><https://eatlas.org.au/data/uuid/f7468d15-12be-4e3f-a246-b2882a324f59>

<sup>4</sup>[https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS\\_2SR](https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_2SR)

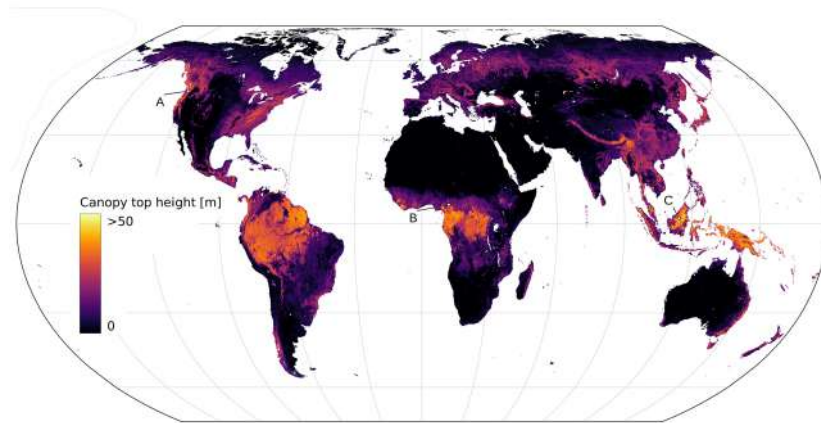


Figure 10: Global canopy top height map for the year 2020, visualized in Equal-Earth projection [6]

### 3.1.3 Canopy height map

The authors of [6] used a probabilistic deep-learning approach to leverage the geolocated GEDI waveforms (L1B) and regress canopy top height (the relative canopy height at which 98% of the energy has been returned, i.e.,  $rh_{98}$ ) at the footprint-level. They subsequently used the canopy height footprints as sparse reference data to train an ensemble of neural networks, each taking Sentinel-2 imagery as input and regressing canopy height values. They thus provide a global canopy height map (with well-calibrated uncertainties) at a 10m resolution for the year 2020 (see Figure 10). An interactive version of the map can be found [here](#).

### 3.1.4 ESA CCI

The Climate Change Initiative (CCI) Biomass project provides global maps of above-ground biomass for four epochs (mid 1990s, 2010, 2017 and 2018), at a 100m per pixel resolution. They are derived from a combination of Earth observation data, depending on the year, from the Copernicus Sentinel-1 mission, Envisat’s ASAR instrument and JAXA’s Advanced Land Observing Satellite (ALOS-1 and ALOS-2), along with additional information from Earth observation sources.

### 3.1.5 CMS FIA

The NASA Carbon Monitoring System (CMS) program used field data obtained through the USDA Forest Service’s Forest Inventory and Analysis (FIA) program to derive biomass estimates across the Conterminous USA, at a spatial resolution of 640km<sup>2</sup> hexagonal cells. The data is derived from field surveys conducted from 2009 to 2019.

## 3.2 Datasets

We now describe an additional two datasets that will be referenced throughout this work.

### 3.2.1 Clemence

The authors of [9] downloaded a random subset (approximately 25%, at the file level) of the L4A GEDI footprints for the year 2020 (January - December). They processed the data by applying a beam sensitivity threshold of 0.95, which is an estimate of the maximum canopy cover that can be penetrated considering the signal-to-noise ratio of the waveform. They then extracted patches of 15 × 15

pixels (corresponding to  $150 \times 150 \text{ m}^2$  on the ground) from the canopy height map (described in 3.1.3), centered around those footprints.

### 3.2.2 Yuchang

The authors of [8] first created a world-wide mosaic made of the least cloudy Sentinel-2 L1C tile during what they defined as the *leaf-on season* (May - September) for the year 2020, before processing the mosaic into the L2A data product. Subsequently, they downloaded a random subset (approximately 25%, at the file level) of the L4A GEDI footprints for the year 2020 (January - December).

On the one hand, they applied a beam sensitivity threshold of 0.95 to the GEDI footprints. They further ignored non-vegetated areas: locations where the canopy height map indicated 0 height, and where the Sentinel-2 SCL map indicated *Bare Soils*, *Water*, or *Snow / Ice*. They then extracted patches of  $15 \times 15$  pixels (corresponding to  $150 \times 150 \text{ m}^2$  on the ground) both from the Sentinel-2 mosaic and the canopy height map, centered around those footprints. We call this version of the dataset, **no vegetation**.

On the other hand, they applied a beam sensitivity threshold of 0.98 to the GEDI footprints. They then also extracted patches of  $15 \times 15$  pixels both from the Sentinel-2 mosaic and the canopy height map, centered around those footprints. We call this version of the dataset, **sensitive**.

## 3.3 Variable importance analysis

In this section, we describe the variable importance analysis that was conducted in order to identify which GEDI LiDAR-derived variables are most helpful in predicting biomass. The procedure is quite straightforward: building on [9], who estimated biomass solely from canopy height (and geographical coordinates), we gradually provide a toy-model with additional GEDI variables as input (which we call *auxiliary variables*).

### 3.3.1 Dataset

To conduct the variable importance analysis, we first need to identify the GEDI variables of interest, and subsequently compile a dataset putting them all together.

**Variables of interest** After a careful examination of all GEDI data products' descriptions, the following variables were retained.

- *rh* (GEDI L2A) - Canopy Relative Height metric at 1% interval (from *rh0* to *rh100*).
- *digital\_elevation\_model* (GEDI L2A) - TanDEM-X [51] elevation at GEDI footprint location.
- *canopy\_cover* (GEDI L2B) - Percent of the ground covered by the vertical projection of canopy material (i.e. leaves, branches and stems only).
- *fhd\_normal* (GEDI L2B) - Foliage height diversity (FHD) is a canopy structural index describing the vertical heterogeneity of foliage profile [52]. It measures the complexity of canopy structure.
- *pai* (GEDI L2B) - The plant area index (PAI) is defined as one half of the total plant area (all canopy structural elements, e.g. leaf, branch, and trunk) per unit ground surface.
- *pft\_class* (GEDI L2B) - In terrestrial ecology, plant functional types (PFTs) are sets of species with similar responses to the environment and with similar effects on ecosystem functioning [53]. PFTs are provided by GEDI as a 1-km grid derived from the MODIS MCD12Q1v006 Product [54]. Values follow the Land Cover Type 5 Classification scheme: DBT (deciduous broadleaf trees), DNT (deciduous needleleaf trees), EBT (evergreen broadleaf trees), ENT (evergreen needleleaf trees), GSW (grasses, shrubs, and woodlands).

Other variables were considered for quality-filtering and matching purposes. Those variables include:

- *quality\_flag* (GEDI L2A, L2B) - Flag provided to allow users to easily remove erroneous and/or lower quality returns.
- *l4\_quality\_flag* (GEDI L4A, L4B) - Flag identifying shots that may be considered as samples of the population of which the applied models are representative.
- *shot\_number* (all GEDI products) - Unique identifier used to link observations between groups and between data products. The shot number format is OOOOBBRRGNNNNNNNN, where OOOOO is the orbit number, BB is the beam number, RR is reserved for the future and G is the sub-orbit number, and NNNNNNN is the shot number within the beam.

An additional variable is defined for matching purposes: *pattern*. For a given data point, the corresponding *pattern* references the file of origin. It is made of: the date and time of acquisition (YYYYDDDDHHMMSS); the orbit number; the sub-orbit granule number; the track number; and the positioning and pointing determination system (PPDS) type.

**Data Processing** We downloaded GEDI L2A/L2B/L4A data from May 2020 to September 2020, to align with the temporal range of Sentinel-2 data (as described in Section 3.1.2). We filter for quality data using the *quality\_flag* and *l4\_quality\_flag*. We only keep data coming from power beams as previous work [9] has shown that they provide more reliable estimates than coverage beams. We do not consider AGBD values that are bigger than 500, as the literature [55] suggests, to maintain a representative sample of values. The footprints across GEDI data products were then matched, based on their latitude, longitude, *shot\_number*, and *pattern*. We thus obtain a dataset made of 150,174,906 GEDI footprints, with AGBD values and the afore-mentioned *auxiliary variables*.

**Insights** From the AGBD values distribution (Figure 11) we observe that the data is positively skewed: low biomass values are much more prevalent. This long-tail distribution is consistent with high biomass values being concentrated in a few regions on Earth and most of the land surface having low biomass density [55]. We also compute the pairwise Pearson correlation [56] coefficients between a subset of the variables (Figure 12); where  $-1$  means perfect negative correlation,  $+1$  means perfect positive correlation, and zero being no correlation between X and Y [57]. Results indicate that the auxiliary variables are sufficiently correlated to eventually perform multi-task learning.

### 3.3.2 Method

In order to conduct the variable importance analysis, we leverage LightGBM [58], a gradient boosting framework that uses tree-based learning algorithms. It was developed to solve the training speed and memory consumption issues associated with the conventional implementations of gradient boosted decision trees when working with large datasets. We leverage its built-in *feature\_importance* method, which describes the number of times each feature is used in a given model. As decision trees are built by splitting observations based on feature values, this inherently quantifies each feature’s importance in the model’s prediction. We also conducted an ablation study by training various LightGBM models, each on a different subset of the variables.

### 3.3.3 Results

We report each model’s performance metrics in Table 2. Additionally, we plot the best two models’ respective features’ importance in Figures 13 and 14. Unsurprisingly, the best performing model is the one having access to all available variables; and the worst performing model is the one having access to the smallest subset of variables, i.e., the latitude, longitude, and *rh98*. Notably, there is a huge performance gap between the two models. As you increasingly make more variables available to the models, their respective performance improve, which shows that having access to more variables

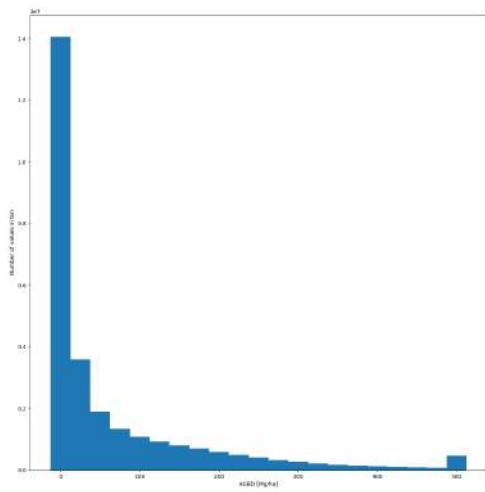


Figure 11: Dataset AGBD values distribution. AGBD values up to  $500 \text{ Mg} \cdot \text{ha}^{-1}$  are binned in  $25 \text{ Mg} \cdot \text{ha}^{-1}$  bins; AGBD values large than  $500 \text{ Mg} \cdot \text{ha}^{-1}$  are binned together.

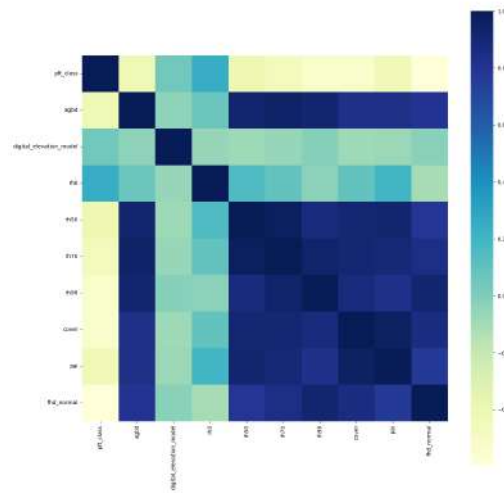


Figure 12: Pearson correlation coefficients for a subset of the variables.

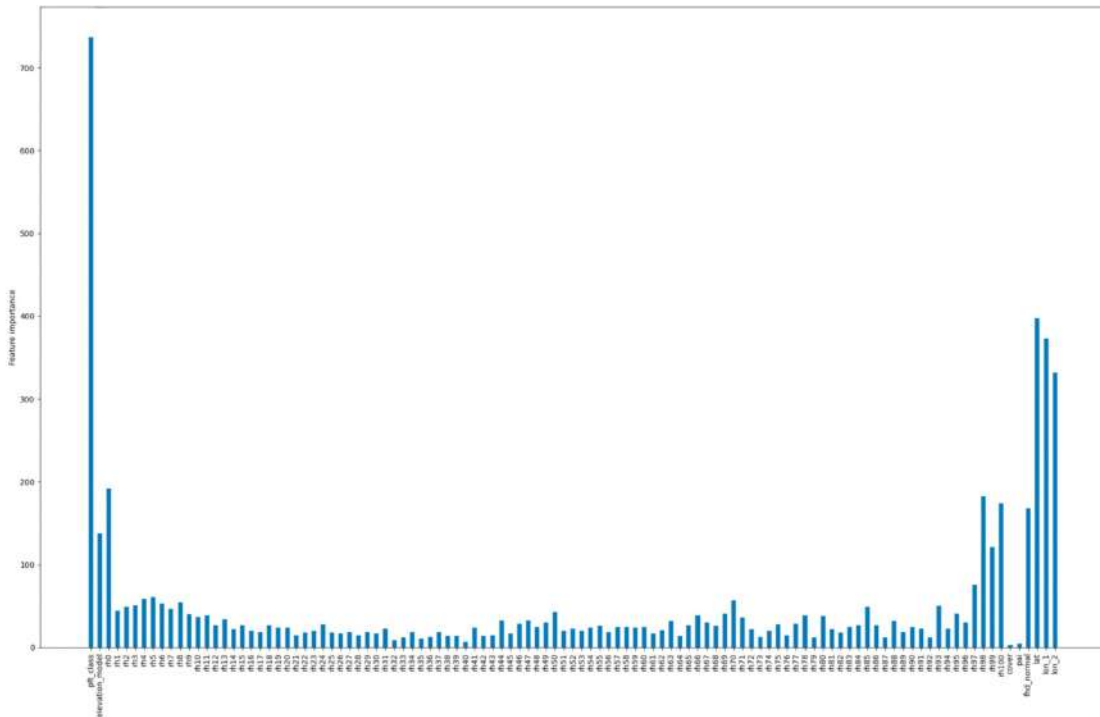


Figure 13: Variable importance analysis for the best performing LightGBM model.

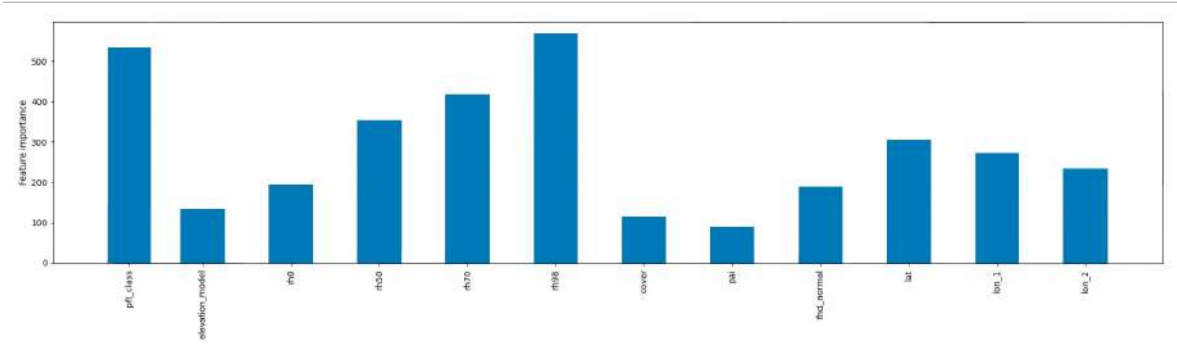


Figure 14: Variable importance analysis for the second best performing LightGBM model.

ultimately does help the prediction of AGBD. The main take-away from the ablation study is the identification of the most relevant RH metrics: while making all RH metrics available to the model yields the best performance, it is not feasible in practice, as it represents a huge amount of data; however, the analysis shows that using as few as 4 well-chosen RH metrics already yields adequate performance metrics. We choose those metrics based on the best model’s features importance plot, and Pearson correlation between RH metrics. Additionally, the feature importance plots show that the models do not rely on the different *canopy* variables equally: the *fhd\_normal* appears to be much more important than either the *canopy\_cover* or the *pai*; we still include them, but keep it in mind for the remainder of this work.

RMSE	MAE	ME	lat & lon	canopy	dem	pft_class	rh
20.7	6.7	0.03	✓	✓	✓	✓	all
22.8	7.4	0.03	✓	✓	✓	✓	0-50-70-98
23.9	8.1	0.04	✓	✓	✓	✓	0-70-98
24.5	8.3	0.03		✓	✓	✓	0-50-70-98
25.7	8.5	0.03	✓		✓	✓	0-70-98
26.8	9.3	0.04	✓				0-50-60-70-98-100
29.72	10.1	0.03	✓	✓	✓	✓	0-98-100
30.86	10.5	0.02	✓	✓	✓	✓	98
32.28	11.62	0.02	✓	✓	✓		98
37.45	15.7	0.02	✓		✓	✓	98
39.2	16.7	0.02	✓				98

Table 2: LightGBM models description and performance metrics. A ticked box indicates the model was trained with the variable at hand. *lat & lon* indicates that both latitude and longitude were used, as opposed to only latitude; *canopy* includes *canopy\_cover*, *pai*, and *fhd\_normal*; *dem* is the *digital\_elevation\_model*; and *rh* indicates which relative height metrics were used.

From those analyses, we can draw the following conclusions:

- having access to more variables (as opposed to only *rh98*) helps the prediction of AGBD;
- the most helpful variables are *rh0*, *rh50*, *rh70*, *rh98*, *fhd\_normal*, *digital\_elevation\_model*, and the *pft\_class*.

At this point, we have identified the *auxiliary variables* that should prove most helpful for estimating biomass, when provided as additional inputs. However, the ultimate goal is to produce a global fine-grained AGBD map, and we only have *auxiliary variables* values at the sparse GEDI footprint level. The question we now want to answer is: with only Sentinel-2 data (and geographical coordinates) as input, could predicting not only biomass but the *auxiliary variables* as well, help guiding the biomass estimation? In an effort to answer this question, we turn to Multi-task learning (MTL).

### 3.4 Dataset for MTL

In this section, we describe the dataset used for Multi-task learning, as well as how it was curated. As a reminder, we leverage MTL, i.e. predicting biomass along with the auxiliary variables (*rh0*, *rh50*, *rh70*, *rh98*, *fhd\_normal*, *pai*, *cover*, *digital\_elevation\_model*, and the *pft\_class*), to attempt to learn solely from Sentinel-2 data.

#### 3.4.1 Dataset creation

Relying on already curated Sentinel-2 imagery (Yuchang’s **sensitive** dataset) was both a blessing and a curse. It exempted us from fetching and processing the imagery ourselves, which represents a huge gain of time. But on the other hand, we only have Sentinel-2 patches for  $\sim 25\%$  of the GEDI footprints in 2020; whereas the present work relies on all of the GEDI footprints from May to September 2020. By intersecting our GEDI footprints with the ones for which a Sentinel-2 patch is available, we hereby reduce the size of our dataset.

The challenge was to actually perform this intersection, as the authors of [8] did not store a reference to the GEDI file from which the footprint was from for each patch. Consequently, we had to develop a matching strategy, which we describe in Figure ?? (Appendix). It basically boils down to iterating over the patches, extracting the central footprint’s AGBD/latitude/longitude values, and finding the matching GEDI footprint in my extended dataset.

#### 3.4.2 Geographical split

The data we subsample from, Yuchang’s **sensitive** dataset, was separated into train/validation/test sets by simply randomly splitting the data points (i.e. the GEDI footprints). On the other hand, the authors of [9] used what we call a *geographical* split: they considered all Sentinel-2 tiles, discarded overlapping tiles, and then randomly split them. As a consequence, all GEDI footprints that belong to the same tile, belong to the same set (train/validation/test). This approach reduces data leakage, as neighbouring GEDI footprints will remain within the same split. We therefore chose to follow this approach, and re-shuffled our dataset. We used the exact same Sentinel-2 tiles split as in [9]. The resulting geographical distribution of the data points is depicted in Figure 15.

#### 3.4.3 Comparison with the other datasets

We hereby provide insights on the obtained dataset, and compare it with Yuchang’s [8] and Clémence’s [9]. Table 3 displays the number of data points per dataset, as well as their respective train/test/validation splits. The present dataset is very much reduced in terms of number of data points, compared to the other datasets: there is a ratio of 1 : 40 between my dataset, and Clémence’s dataset. Figures 17-22 further show the data distributions in terms of AGBD values. Evidently, the present dataset displays a widely different data distribution from Yuchang’s and Clémence’s, which are similar. This data distribution shift it is not due to the fact that the data is constrained to the May-September range, as the distribution of Clémence’s dataset when constrained to May-September range is similar to its non-constrained counterpart (see Figure 16). It is also not due to the additional *quality\_flag*, as the distribution in Figure 11 is similar to that of Yuchang and Clémence’s datasets. It is, seemingly, due to a shift in the geographical distribution of the data, that appeared after the matching procedure.



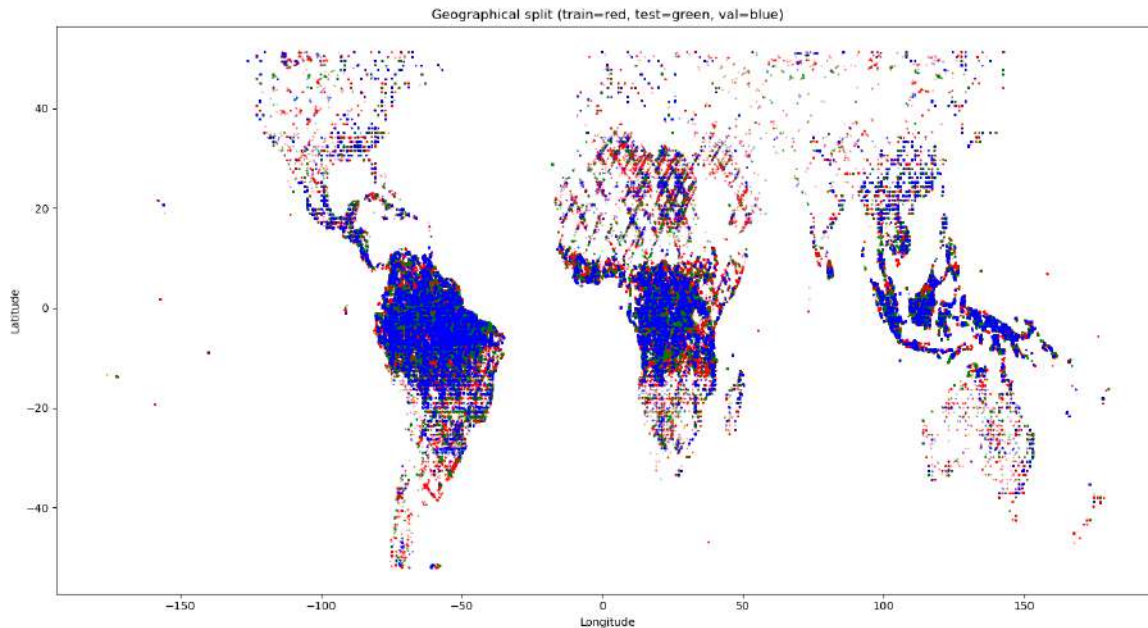


Figure 15: Geographical distribution of the data points. The red, green, and blue colors respectively correspond to the train, test, and validation data.

Dataset	Train	Validation	Test	Total
Mine (pre reshuffling)	5,801,220 (61%)	1,753,521 (18%)	1,996,634 (21%)	9,551,375
Mine (post reshuffling)	6,187,548 (65%)	1,363,698 (14%)	2,000,129 (21%)	9,551,375
Yuchang (sensitive)	105,874,949 (63%)	26,261,569 (16%)	34,718,285 (21%)	166,854,803
Yuchang (no vegetation)	117,802,785 (64%)	28,110,626 (15%)	37,713,955 (21%)	183,627,366
Clemence	256,874,374 (64%)	65,684,201 (16%)	77,549,389 (20%)	400,107,964

Table 3: Number of data points per dataset, and their respective train/test/splits.

*Yuchang* refers to the data from [8] and *Clemence* refers to the data from [9].

Indeed, the data points are more densely located around the equator (see 15, where biomass values tend to be higher). However, the bug responsible for this geographical shift is yet to be found.

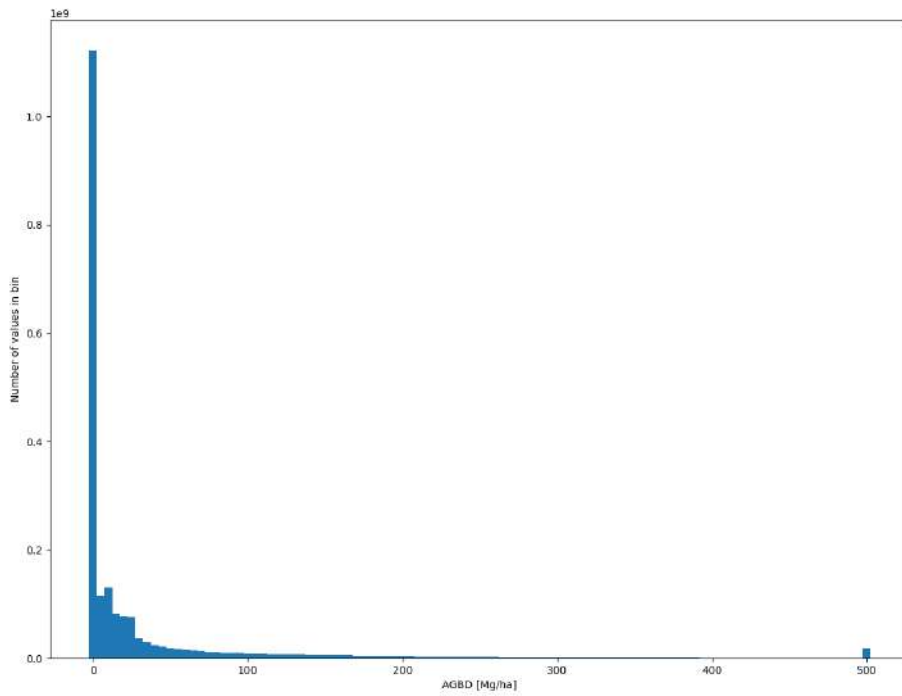


Figure 16: Distribution of Clémence's dataset AGBD values, constrained to the May-September range.

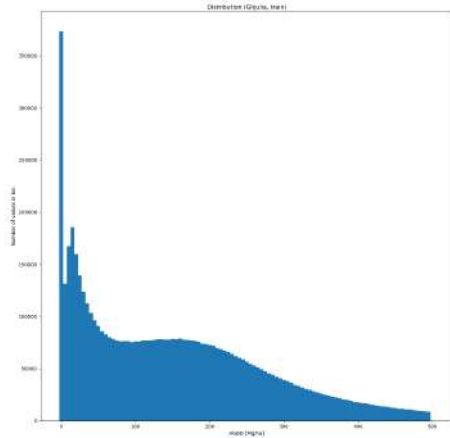


Figure 17: My train dataset AGBD values distribution ( $5 \text{ Mg} \cdot \text{ha}^{-1}$  bins)

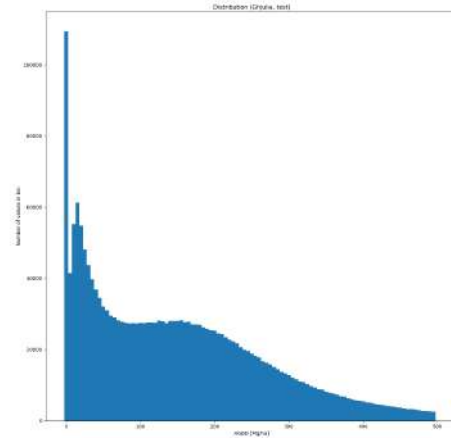


Figure 18: My test dataset AGBD values distribution ( $5 \text{ Mg} \cdot \text{ha}^{-1}$  bins)

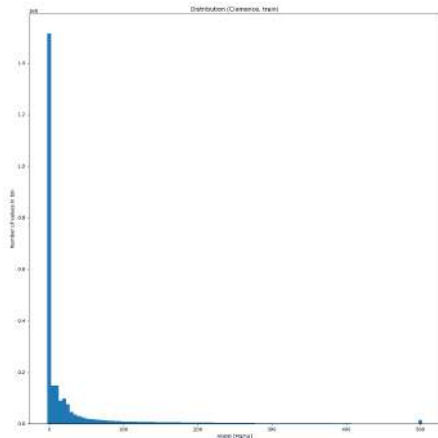


Figure 19: Clemence's train dataset AGBD values distribution ( $5 \text{ Mg} \cdot \text{ha}^{-1}$  bins)

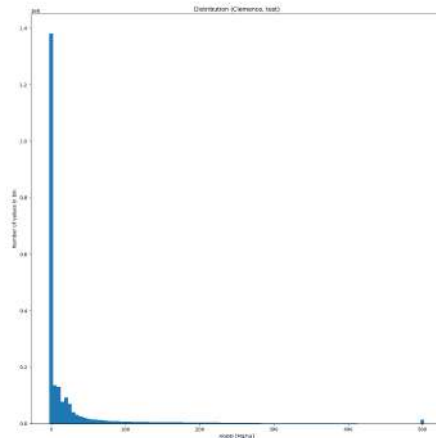


Figure 20: Clémence's test dataset AGBD values distribution ( $5 \text{ Mg} \cdot \text{ha}^{-1}$  bins)

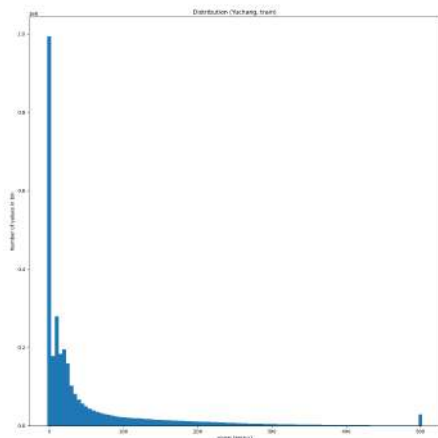


Figure 21: Yuchang's train (sensitive) dataset AGBD values distribution ( $5 \text{ Mg} \cdot \text{ha}^{-1}$  bins)

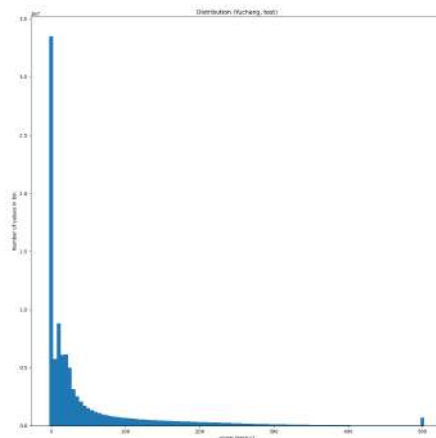


Figure 22: Yuchang's test (sensitive) dataset AGBD values distribution ( $5 \text{ Mg} \cdot \text{ha}^{-1}$  bins)

## 4 Methods

We have outlined the various data sources and datasets used throughout this work. Notably, the dataset curated for MTL. We now turn to the actual methods and describe *how* MTL is conducted. This chapter describes the models that were trained, the training procedure, and the evaluation framework.

### 4.1 Models

To recapitulate, we perform MTL by providing the models with Sentinel-2 data and geographical coordinates (in accordance with previous work [9], we only provide the latitude). In turn, the models predict biomass, along with the auxiliary variables (*rh0*, *rh50*, *rh70*, *rh98*, *fhd\_normal*, *pai*, *cover*, *digital\_elevation\_model*, and the *pft\_class*).

As per Section 2.3, a model performing MTL is (typically) made of: a *body*, whose parameters are shared among all tasks; and multiple *heads*, whose parameters are task specific. We hereby describe the various architectures that were experimented with, both in terms of *heads* and *bodies*. The total number of trainable parameters per model can be found in Table 4.

Model	Number of trainable parameters
FCN	351K
ResNeXt	5.5M
Xception	15.0M

Table 4: Number of trainable parameters per model architecture.

#### 4.1.1 Heads

Each *head* is defined as a Convolutional Neural Network (CNN). This section describes the high-level architecture of the heads; we refer the reader to the specific models’ architectures for implementation details.

In order to model the aleatoric uncertainty, as described in Section 2.2, we design a two-branch CNN: both branches are identical in terms of components, but one outputs the mean of the predicted variable, while the other one outputs the log-variance. We hence define the *head*’s architecture as a *Gaussian* CNN. Note that outputting log-variance and subsequently applying the exponential function guarantees that the estimated variance is positive [59].

Each branch consists of:

- a various number of stacked convolutional blocks (**Conv Block**), each made of a 2d convolution (with *kernel\_size* = 3, *stride* = 1, and *padding* = 1), a batch normalization layer, a ReLU activation function, and a 2d max pooling layer (with *kernel\_size* = 3, *stride* = 1, and *padding* = 1).
- a downsampling layer (**Downsampling**), that can either be a 2d convolution with *kernel\_size* = 1, *stride* = 5, and *padding* = 0; or a 2d average pooling layer with *kernel\_size* = 5, *stride* = 5, and *padding* = 0.
- a 2d convolution with *kernel\_size* = 1, *stride* = 1, and *padding* = 0.

The downsampling component lowers the resolution from  $15 \times 15$  pixels for the input, to  $3 \times 3$  pixels for the output, which amounts to a  $50m \times 50m$  resolution on the ground. This is motivated by suggestions [9] made by experts from ESA and NASA.

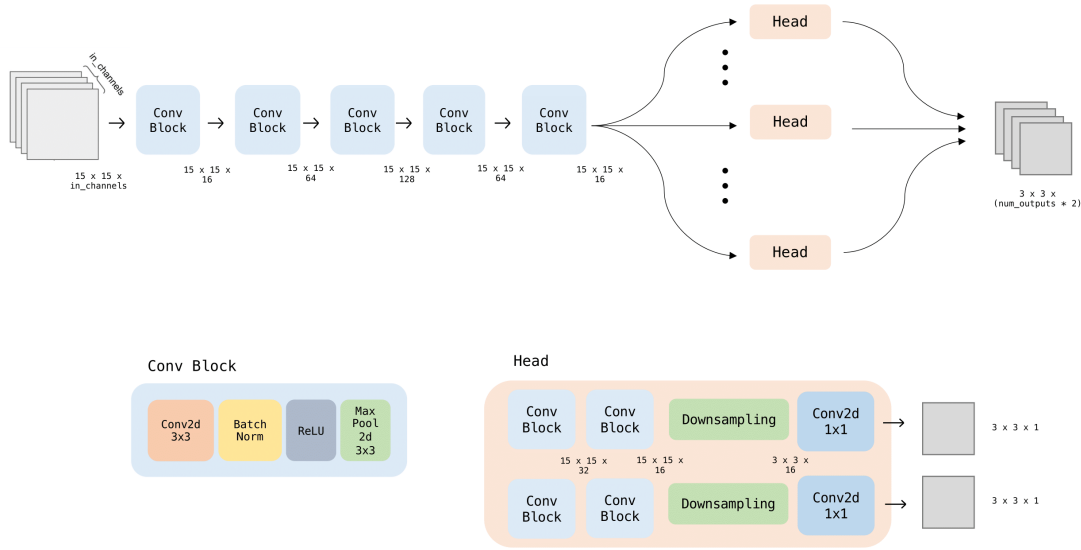


Figure 23: Architecture of the Multi Task Fully Convolutional Neural network (FCN).

#### 4.1.2 Fully Convolutional Neural network

The first (and less complex) model architecture that was experimented with is an adaptation of [9]. The body is a Fully Convolutional Network (FCN), which is a particular type of CNNs that takes inputs of arbitrary size and produces outputs of the same size. This is particularly relevant for dense pixel-level prediction such as pixel-wise regression, which is precisely the focus of this project. The FCN at hand is made of 5 stacked Conv Blocks. The output is then fed into as many *heads* as there are tasks. The overall architecture, and its details, are described in Figure 23.

#### 4.1.3 ResNeXt

The second (and more complex) model architecture that was experimented with is an adaptation of [60], whose authors modified the design principles of ResNeXt to address the specific challenges of forest structure mapping. The *body's* architecture is composed of:

- an **Entry block** made of a 2d convolutional layer (with  $kernel\_size = 1$ ), followed by a batch normalization layer, and a ReLU activation function.
- a succession of **ResNeXt blocks**, made of a 2d convolution layer (with  $kernel\_size = 1$ ), another 2d convolution layer (with  $kernel\_size = 3$ ,  $groups = 32$ ), and another 2d convolution layer (with  $kernel\_size = 1$ ). The former two are each followed by a batch normalization layer and a ReLU activation function, while the latter is followed only by a batch normalization layer. Following the ResNet principle, a skip connection bypasses each block, after which point a ReLU activation function is applied, such that the block learns an additive residual update to its input. Due to the grouped convolution, each block effectively implements a multi-branch computational graph, where all branches are eventually combined by summation.

The model at hand is set with  $N_{blocks} = [2, 5, 3]$  and  $N_{channels} = [64, 128, 256]$ . The output is then fed into as many *heads* as there are tasks. The overall architecture, and its details, are described in Figure 24.

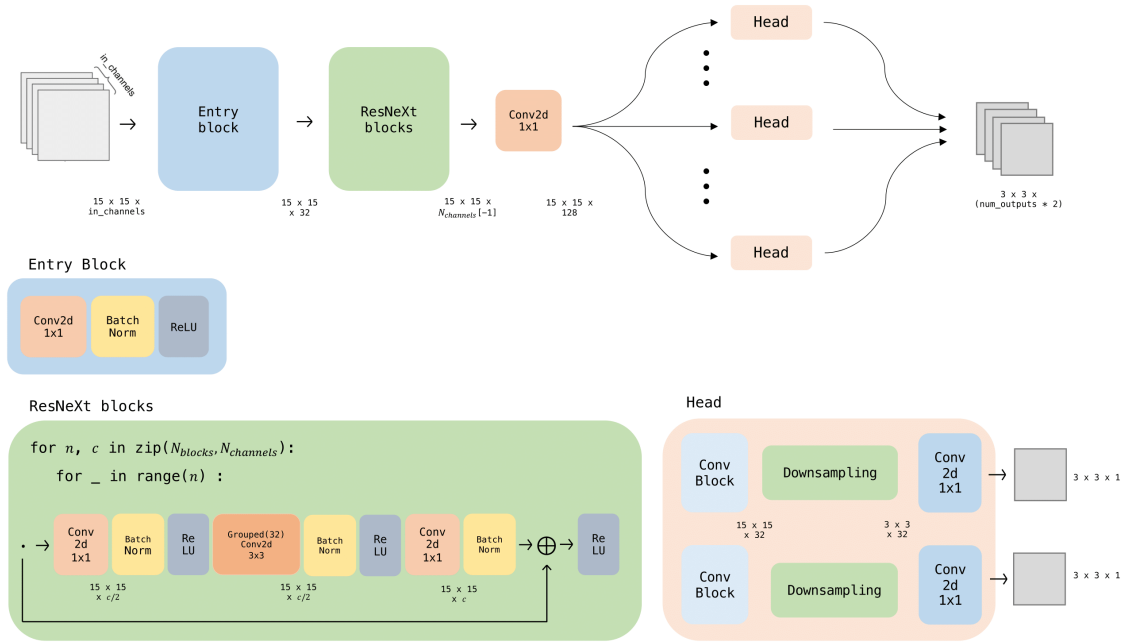


Figure 24: Architecture of the Multi Task ResNeXt network.

#### 4.1.4 Xception

The last (and most complex) model architecture that was experimented with is an adaptation of [61], whose authors modified the design principles of Xception [62] to perform vegetation height mapping. The *body*'s architecture is composed of an **Entry Block**, and a series of 8 **Middle Blocks**. Each of these blocks are themselves composed as follows:

- the **Entry Block** is made of a succession of three 2d convolutional layers (with  $kernel\_size = 1$ ); the first two are followed by a batch normalization layer and a ReLU activation function; the latter is followed by a batch normalization layer only. At this point, a "shortcut" is added (point-wise) to the output, before being put through a ReLU activation function. The shortcut consists of applying a 2d convolutional layer (with  $kernel\_size = 1$ ) and a batch normalization layer to the block's input.
- the **Middle Block** is made of a ReLU activation function, a **SepConv Block**, a batch normalization layer, a ReLU activation function, another **SepConv Block** followed by a batch normalization layer. At this point, as in the **Entry Block**, the same "shortcut" is added (point-wise) to the output, before being put through a ReLU activation function.
- the **SepConv Block** is at the core of the Xception architecture. It consists in a spatial convolution (a 2d convolutional layer with  $kernel\_size = 3$  and  $groups = 728$ ) being performed independently over each channel of the input, followed by a point-wise convolution (a 2d convolutional layer with  $kernel\_size = 1$ ) projecting the channels output by the depthwise convolution onto a new channel space.

The output of the blocks is then put through a 2d convolutional layer with  $kernel\_size = 1$ , and fed into as many *heads* as there are tasks. The overall architecture, and its details, are described in Figure 25.

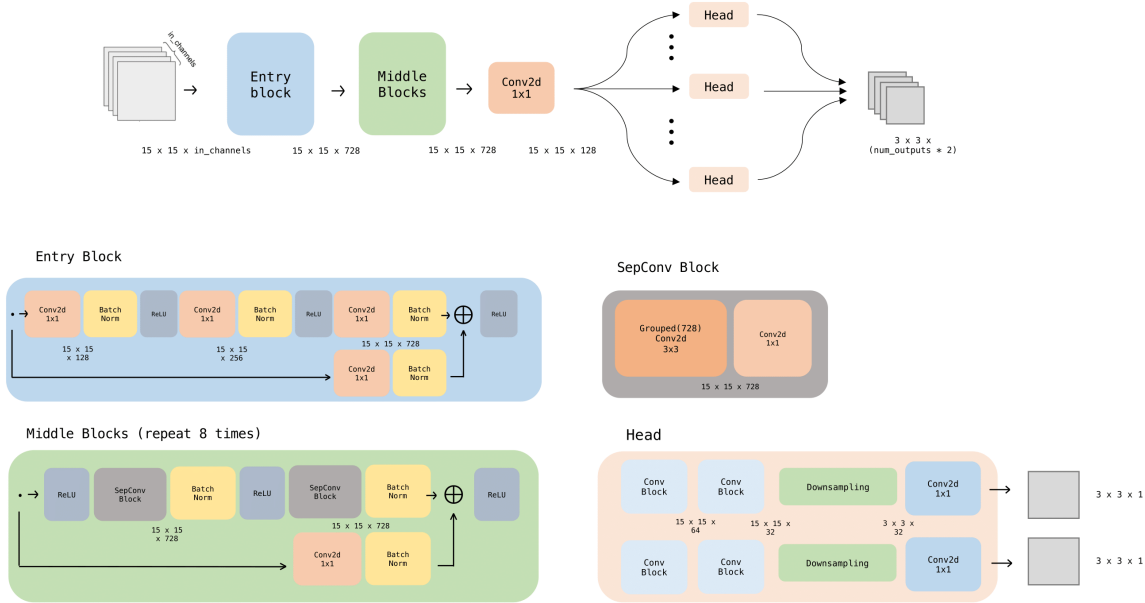


Figure 25: Architecture of the Multi Task Xception network.

## 4.2 Training procedure

We use a patchwise training procedure with patches of size  $15 \times 15$  pixels, corresponding to  $150 \times 150$  m<sup>2</sup> on the ground. Each patch has one valid ground truth pixel at its center. Pixels with missing ground truth do not contribute to the per-patch loss, such that they do not affect the training procedure.

The optimisation of the parameters of the networks is done by mini-batch stochastic gradient descent (SGD). We use a popular variant of SGD called ADAM [63] that adaptively adjusts the learning rate for each trainable parameter by normalizing the global learning rate with the running average of the gradient. This has the effect of amplifying the step size along low gradients and attenuating it for high gradients. In this way, the solver is less sensitive to the chosen base learning rate, and there is no need to design a careful learning schedule [6].

For the experimental evaluation of our method, the base learning rate is set to 0.0001, and the batch size to 64 patches (dictated by memory constraints of the GPU hardware). Each epoch, the network sees 5000 batches of the training set, and evaluates on 1000 batches of the validation set (as in [8] and [9]).

To quantify the epistemic uncertainty, as described in Section 2.2, we train ensembles of models for each experiment. Because of GPU hardware constraints, and motivated by the fact that even a small number of models in an ensemble can effectively monitor the epistemic uncertainty [31, 64], we use 3 models per ensemble.

As mentioned in Section 3, the available data is first split into two geographically separated parts, a training set and a test set that is never seen during training and only used to evaluate the performance of the trained model. The training set is, in turn, split into a training part and a (smaller) validation part, where the latter is never used to compute gradients for back-propagation. The learning process is monitored by observing both the training loss and the validation loss.

### 4.2.1 Training losses

We define multiple loss functions, for various purposes. For single-task learning (i.e. our baseline models), any typical regression loss can be applied. For MTL, one might use different losses depending on the weighting scheme, and variables to be predicted. Indeed, with all but one of our target variables, one can apply any regression loss. But one of the target variables (*pft\_class*) is categorical, so we also define a classification loss.

For all of the losses listed below,  $N$  is the number of samples,  $y_i$  is the target value,  $x_i$  is the input data point,  $\mu(x_i)$  is the mean predicted value,  $\sigma(x_i)$  is the predicted variance. As explained in Section ??, the model outputs the *log*-variance. To obtain the variance, we apply the exponential function. But, when the log-variance tends to  $-\infty$ , the variance will tend to 0, which we do not want as we are dividing by the variance in the GNLL and LNLL losses. For numerical stability, we hence add a small number ( $\epsilon = 1e-6$ ) to the log-variance. We hence obtain:  $\sigma(x_i) = \exp(\log\_sigma(x_i) + \epsilon)$ .

The Root Mean Squared Error (RMSE) is a commonly used loss function for regression tasks in neural networks. Mathematically, it is defined as:

$$(RMSE) \text{ Loss} = \frac{1}{N} \sum_{i=1}^N (\mu(x_i) - y_i)^2$$

The use of the square function in the loss function ensures that the loss increases quadratically as the predicted value deviates further from the target value. The square root function in the definition of RMSE then scales the loss back to the original units of the target variable, making it easier to interpret. The goal of the RMSE loss function is to minimize the mean squared error between the predicted values and the target values. This makes it well-suited for regression tasks where the target variable is a continuous variable that does not necessarily follow a specific probability distribution.

The Gaussian Negative Log Likelihood (GNLL) is a commonly used loss function for regression tasks in neural networks where the target variable is assumed to have a Gaussian distribution. Mathematically, it is defined as:

$$(GNLL) \text{ Loss} = \frac{1}{N} \sum_{i=1}^N \left[ \log \sigma(x_i) + \frac{(y_i - \mu(x_i))^2}{\sigma(x_i)} \right]$$

The goal of the GNLL loss function is to minimize the mean squared error between the predicted values and the target values, while taking into account the uncertainty or variability of the target values. This makes it well-suited for regression tasks where the target variable is assumed to have a Gaussian distribution.

The Laplace Negative Log Likelihood (LNLL) is a loss function commonly used in neural networks for regression tasks where the target variable has a Laplace distribution. Mathematically, it is defined as:

$$(LNLL) \text{ Loss} = \frac{1}{N} \sum_{i=1}^N \left[ \log \frac{\sigma(x_i)}{2} + \frac{|y_i - \mu(x_i)|}{\sigma(x_i)} \right]$$

The goal of the LNLL loss function is to minimize the average distance between the predicted values and the target values, while taking into account the uncertainty or variability of the target values. This makes it well-suited for regression tasks where the target variable is known to have a Laplace distribution, such as in image denoising, audio signal processing, and other types of signal processing.

The cross-entropy (CE) loss is a commonly used loss function for classification tasks in neural networks. Mathematically, it is defined as:



$$(CE) \text{ Loss} = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^C y_{i,j} \log(\mu(x_{i,j}))$$

where  $C$  is the number of classes,  $y_{i,j}$  is the ground truth label for the  $i$ -th sample and  $j$ -th class (it is 1 if the sample belongs to class  $j$  and 0 otherwise), and  $\mu(x_{i,j})$  is the predicted probability of the  $i$ -th sample belonging to the  $j$ -th class. The use of the logarithm function ensures that the loss increases exponentially as the predicted probability deviates further from the ground truth label.

#### 4.2.2 MTL Weighting schemes

When performing MTL, one needs to combine in some way the losses of the individual tasks into a global loss for the network, i.e., setting  $\{w_i\}_{i=1,\dots,M}$  such that

$$\mathcal{L} = \sum_{i=1}^M w_i \mathcal{L}_i$$

where  $\mathcal{L}_i$  is the loss of the  $i$ -th task, and  $M$  is the total number of tasks.

As explained in Section 2.3, several weighting schemes exist. We experimented with the following ones:

**Equal weights** – this approach consists in setting  $w_i = 1 \forall i \in \{1, \dots, M\}$ , not taking into account the potential differences in magnitude in the values predicted for each task;

**Constant task-specific weights** – this approach consists in setting  $w_i = \text{feature\_importance}_i \forall i \in \{1, \dots, M\}$ , where  $\text{feature\_importance}_i$  is the feature importance of the  $i$ -th task as calculated in Section 3.3.2. This approach does not take into account the potential differences in magnitude in the values predicted for each task, but assigns relative importance factors to each task;

**Uncertainty weighting** – this approach, described in Section 2.3, consists in setting  $w_i = \sigma_i \forall i \in \{1, \dots, M\}$  where  $\sigma_i$  is the predicted variance of the  $i$ -th task, i.e., the output of the second branch of each *head*. In the case of *pft\_class*, a categorical variable, we fix the *temperature* parameter  $\sigma_i = 1$ .

### 4.3 Evaluation scheme

#### 4.3.1 Evaluation metrics

We rely on several metrics to evaluate the quality of our predictions. The focus of this work is on the prediction of AGBD, so the evaluation metrics hereby described evaluate the performance of regression models only. In the following,  $\hat{y}_i$  is the predicted value, while  $y_i$  is the target value.

The Root Mean Squared Error (RMSE), as described previously, measures the average distance between the predicted and true values, taking into account the magnitude of the errors. RMSE is useful when we want to penalize larger errors more heavily than smaller errors.

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

The Mean Error (ME) measures the average difference between the predicted and true values. A positive ME indicates that the model tends to overestimate the true values, while a negative ME

indicates that the model tends to underestimate the true values.

$$ME = \frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)$$

The Mean Absolute Error (MAE) measures the average absolute difference between the predicted and true values, regardless of their direction. It provides a measure of the magnitude of the errors made by the model.

$$MAE = \frac{1}{N} \sum_{i=1}^N |\hat{y}_i - y_i|$$

### 4.3.2 Comparison with other biomass maps

Following the approach defined in [9], we benchmark our AGBD predictions against those of ESA CCI (3.1.4) and CMS FIA (3.1.5). The authors of [9] used forest growth data to adjust the existing maps' AGBD values to the year 2020. They derived two trajectories: a *pessimistic* one, with growth data worst case scenarios, and an *optimistic* one, with growth data from best case scenarios. Then, we can simply aggregate and average the predictions at the hexagon cell level defined by the FIA.

## 5 Results and Discussion

### 5.1 Experiments

We display the results of the experiments in Table 5, and hereby discuss their implications.

Architecture	CH	S2	Loss	MTL	RMSE	MAE	ME
Xception		✓	RMSE	uncertainty	78.60	55.97	-7.15
Xception		✓	GNLL		79.61	56.87	-6.97
Xception	✓	✓	RMSE	uncertainty	75.52	53.83	-4.80
Xception	✓	✓	GNLL	constant	77.14	55.03	-6.21
Xception	✓	✓	LNLL	constant	78.90	54.56	-13.04
Xception	✓	✓	GNLL		77.58	55.85	-3.95
Xception	✓	✓	LNLL		78.50	54.12	-13.96
Xception	✓		RMSE	uncertainty	81.74	60.38	-3.30
Xception	✓		GNLL		81.61	60.20	-3.88
ResNeXt		✓	RMSE	uncertainty	79.93	57.16	-6.22
ResNeXt		✓	GNLL		84.29	60.98	-7.26
ResNeXt	✓	✓	RMSE	uncertainty	76.77	54.88	-5.66
ResNeXt	✓	✓	GNLL		79.15	57.44	-4.50
ResNeXt	✓		RMSE	uncertainty	81.97	60.61	-4.09
ResNeXt	✓		GNLL		82.03	60.69	-3.38
FCN		✓	RMSE	uncertainty	86.64	63.58	-4.38
FCN		✓	GNLL		86.61	63.80	-3.01
FCN	✓	✓	RMSE	uncertainty	80.01	58.19	-5.58
FCN	✓	✓	GNLL		79.78	58.12	-2.52
FCN	✓		RMSE	uncertainty	82.16	61.18	-1.39
FCN	✓		GNLL		82.07	60.50	-4.42

Table 5: The models’ test performance metrics. *CH* indicates that Canopy Height was an input feature. *S2* indicates that the Sentinel-2 bands were input features. *MTL* describes the MTL weighting strategy, where a blank cell indicates single-task learning.

**Impact of model complexity** As one might expect, models with more parameters tend to outperform those with fewer parameters. In terms of RMSE, the gap between the best performing Xception model and the best performing FCN model is 4.26; while the gap between the former and the best performing resNeXt model is 1.25. Moving forward, it is important to note that this increase in performance comes at the cost of a 10-fold increase in the number of parameters. As such, the added complexity and resource requirements of the Xception models may need to be carefully considered when deciding whether the marginal improvement in performance is worth the additional cost.

**Impact of loss function** We can only infer a potential impact of the choice of loss function from comparing identical models that were trained with different losses. We therefore consider the 4-th

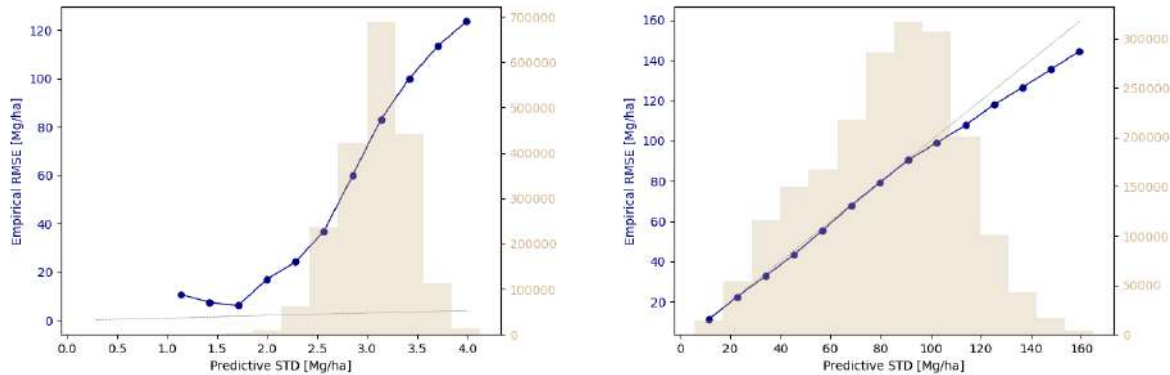


Figure 26: Predictive uncertainty vs. Empirical RMSE for an uncertainty-weighted MTL ResNeXt model (left), and a GNLL single-task ResNeXt model (right).

and 5-th line of Table 5: Xception models with Sentinel-2 data and canopy height data as input and trained using, respectively, the GNLL and the LNLL as constant-weighted loss for MTL. While the GNLL-based model performs better than the LNLL-one with respect to RMSE, it performs marginally worse in terms of MAE. However, when it comes to the ME, that of the LNLL-based model is slightly more than twice that of the GNLL-based model. We thus retain that the GNLL loss tend to outperform the LNLL, for this use case.

**Impact of the MTL weighting schemes** The *equal* weighting scheme was quickly dismissed in early trials, as it clearly reported under-performing results; we do not even include it here. Regarding the other two weighting schemes (we refer the reader back to Section 2.3 for a description), the *uncertainty*-based one outperforms the one based on *constant* weights. This is true for both the GNLL and the LNLL, and is consistent with the existing literature on MTL. On another note, using the model’s predicted variance (or log-variance) to weight the losses terms has the impact that the variances are eventually parametrized, and updated accordingly. As a result, the variances are not well-calibrated anymore, and cannot be interpreted as the predictive uncertainty. We illustrate this behavior in Figure 26.

**MTL vs. single-task learning** In the case of the FCN models, it seems that MTL does not help at all in the estimation of biomass, it even hinders it to some extent. This might be due to the fact that the model is not complex enough to learn how to properly update the parameters to perform MTL effectively. In the case of the ResNeXt and Xception models, MTL consistently out-performs single-task learning; except in the notable case where the Xception model was provided as sole input the canopy height map. This might suggest that the features learned by the Xception model from the canopy height map are highly informative for estimating biomass, and that the additional tasks added in MTL may have interfered with the learning of these important features, resulting in a drop in performance. Whereas, when the models are provided with the canopy height map, and the Sentinel-2 data, they are able to leverage the complementary information from both sources, leading to improved performance through MTL. The combination of the two sources of data allows the model to better capture the complexity of the biomass estimation task and effectively leverage the learned features to improve accuracy.

**Impact of the input data sources** Although the goal of this work was to leverage MTL to learn solely from Sentinel-2 data, we realized that our low number of data samples would likely not make it possible. In an effort to palliate that, and to assess how the models would perform with

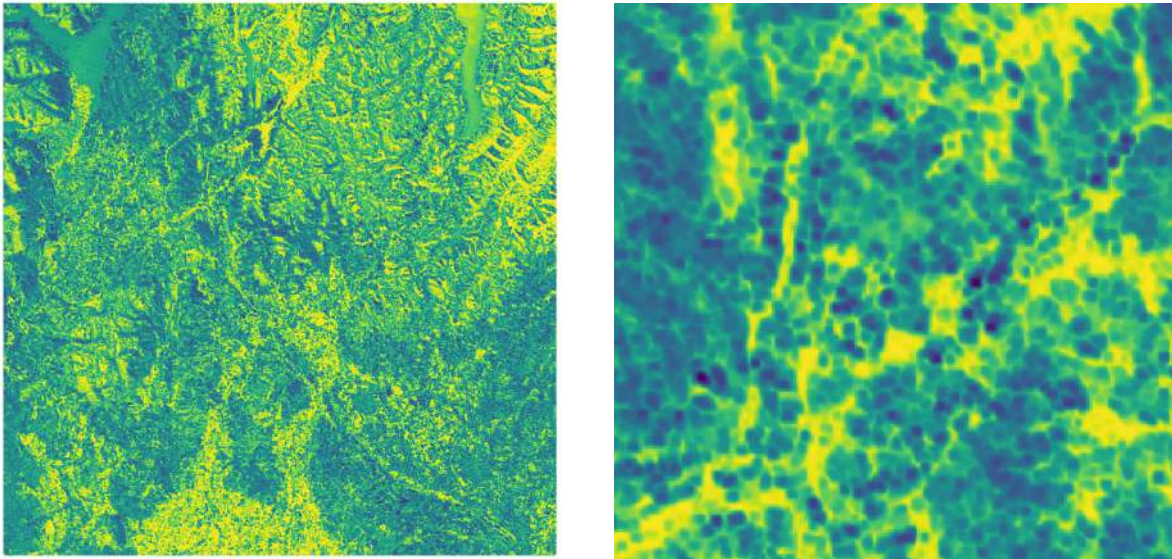


Figure 27: Sample prediction displaying the cell-like pattern (tile 59GLL). Right: zoomed.

an additional data source, we also provide canopy height data as an input. For benchmarking purposes, i.e. to compare with Cl mence’s work, we also train models with only canopy height data as input.

Across all architectures, losses, MTL or not, the best performing models are consistently those that are provided canopy height data and Sentinel-2 data as input. Unsurprisingly, as both types of data contain important information that is relevant for estimating biomass: the canopy height data provides information about the vertical structure of the forest, while the Sentinel-2 data provides information about the spectral characteristics of the forest. By combining these two sources of information, the models are able to better capture the complexity of the relationship between these variables, resulting in improved performance. Now, as we look further into canopy height data versus Sentinel-2 data as input, we treat FCN as a special case, as it does not align with the conclusions that can be drawn from the other models. In the case of FCN, the models provided with canopy height data outperform those provided with Sentinel-2 data. In the case of ResNeXt and Xception, the models provided with Sentinel-2 data outperform those provided with canopy height data. This discrepancy may be due to the fact that FCN has significantly fewer parameters than ResNeXt and Xception, and as a result may be less capable of effectively learning from the more complex Sentinel-2 data. In general, more complex architectures like ResNeXt and Xception may be better suited to learn from the high-dimensional Sentinel-2 data, whereas simpler models like FCN may benefit more from the additional information provided by the canopy height data.

**Choice of Downsampling layer** While there was no quantitative difference between the two proposed downsampling methods (using a convolutional layer vs. using an average pooling layer), we did notice an important qualitative difference. More specifically, models trained with the Conv2d downsampling layer displayed an undesirable behavior at inference time, where cell-like patterns could be seen in the tile-level predictions (see Figure 27). Switching to the AvgPool downsampling layer got rid of this artifact. As this switch was applied early-on, we do not report metrics of models trained with the Conv2d downsampling layer.

## 5.2 Cross-datasets evaluation

In an effort to understand more about the impact of the data distribution shift, we perform cross-datasets evaluation of the best performing model, as well as Clémence’s best model. Simply put, we evaluate each model on: Clémence’s test set, Yuchang’s test set, and my test set. While Clémence’s and Yuchang’s datasets have a similar biomass values distribution (cf. Figures 20 and 22), my dataset has much fewer datapoints, but with more medium-to-high biomass values (cf. Figure 18).

Model	Dataset	RMSE	MAE	ME
Clémence	Clémence	64.87	20.29	-3.21
	Yuchang	77.12	32.30	-5.11
	Mine	85.80	61.25	-6.40
Mine	Clémence	69.65	30.75	11.05
	Yuchang	79.60	39.65	10.04
	Mine	75.52	53.83	-4.80

Table 6: Cross-datasets test metrics.

*Model* describes whose model was evaluated and *Dataset* describes on whose dataset it was evaluated.

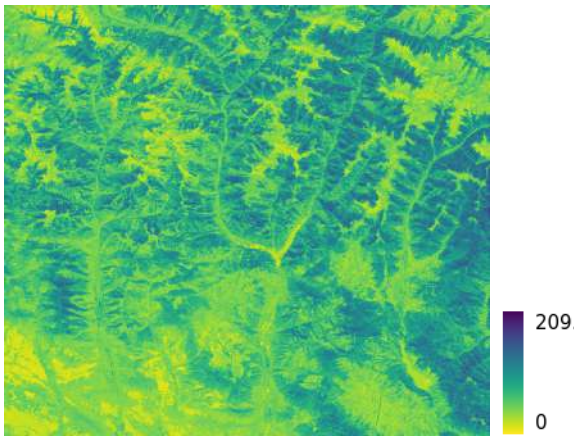
Both models are relatively sensitive to the data distribution shift: there is a consistent performance drop when the models are evaluated on a dataset they have not been trained on. Future work could explore techniques such as domain adaptation or transfer learning to mitigate the impact of data distribution shift on model performance.

## 5.3 Visual assessment

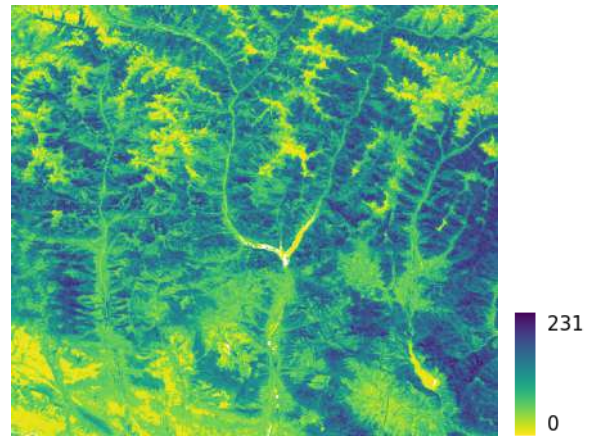
We finally visualize a few selected models’ predictions on a couple of selected Sentinel-2 tiles. Regarding the models, we compare the following ones: the best model with GNLL (CH, S2, MTL with constant GNLL), the best model with LNLL (CH, S2, single-task learning with LNLL), the best model with only Sentinel-2 data as input (S2, MTL with uncertainty RMSE), and the best model overall (CH, S2, MTL with RMSE uncertainty). We also include predictions from Clémence’s model, and the GEDI L4B data product. As for the tiles, the next page displays predictions for Sentinel-2 tile 38TMM, located in Georgia; the following page displays predictions for Sentinel-2 tile 22MDE, located in Brazil.

From this (limited) qualitative assessment, we can infer the following:

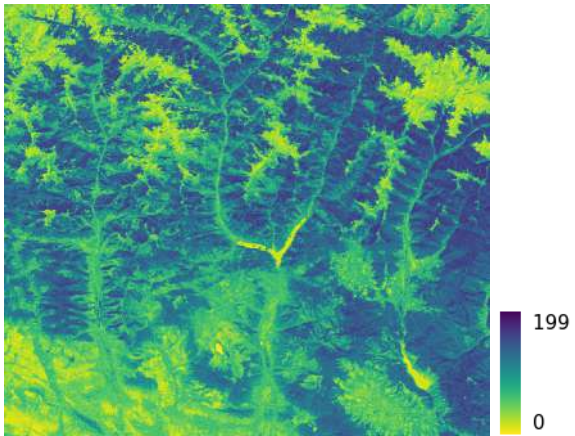
- models that rely solely on Sentinel-2 data as input are more sensitive to the visual artifacts of Sentinel-2: one can see clouds in prediction [22MDE - 3], which are present in the input Sentinel-2 image, while they are not present in either of the other predictions, which also relied on canopy height data;
- models that were trained with LNLL ([38TMM-4] and [22MDE-4]) are the ones that under-predict biomass the most;
- models that manage to learn the highest biomass values are those trained using uncertainty-based MTL ([38TMM-2] and [22MDE-2]);
- overall, the models extract the appropriate features from the input, but consistently under-predict the corresponding biomass values.



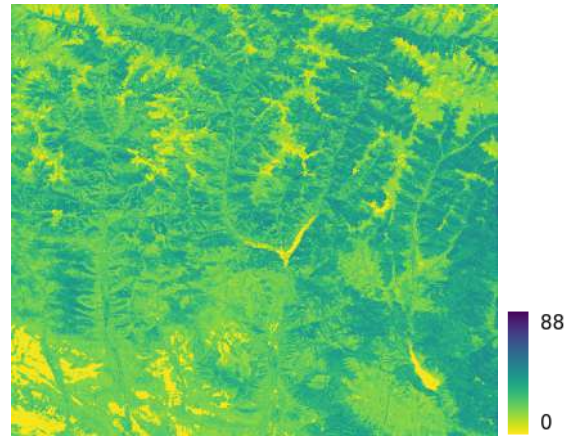
[38TMM - 1] CH, S2, MTL  
(GNLL constant)



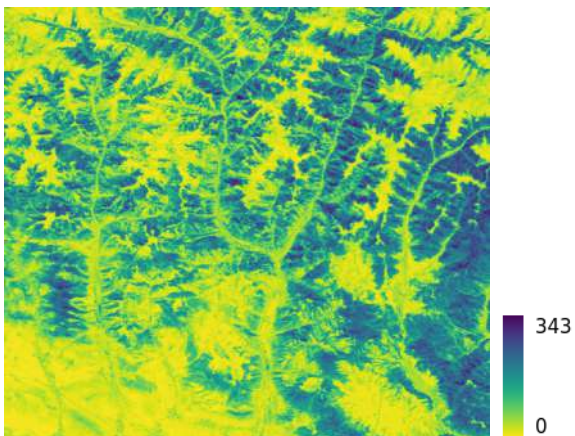
[38TMM - 2] CH, S2, MTL  
(RMSE uncertainty)



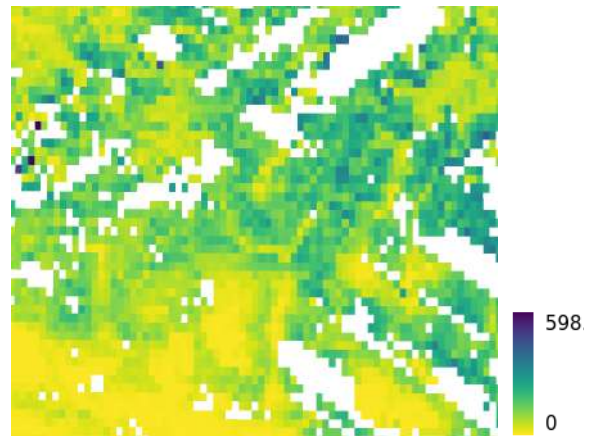
[38TMM - 3] S2, MTL  
(RMSE uncertainty)



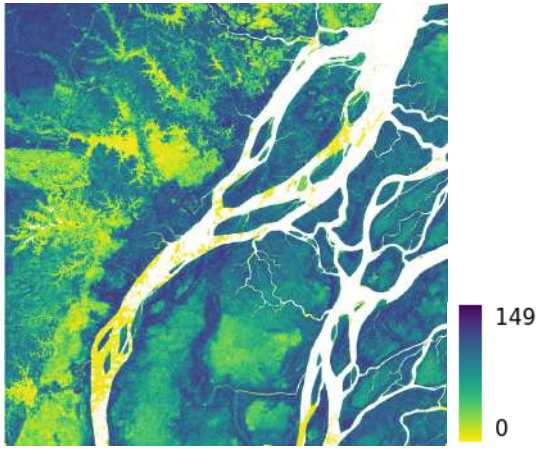
[38TMM - 4] CH, S2, LNLL, STL



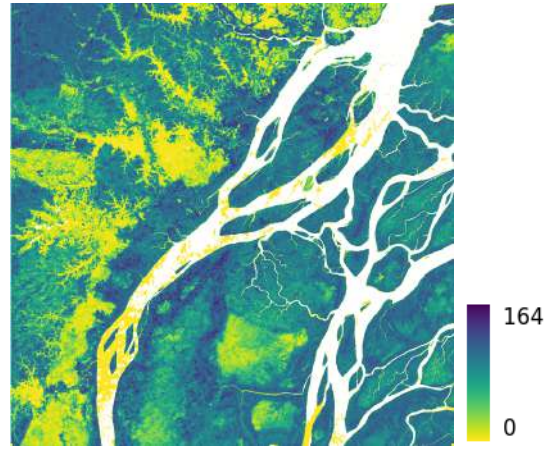
[38TMM - 5] Clémence



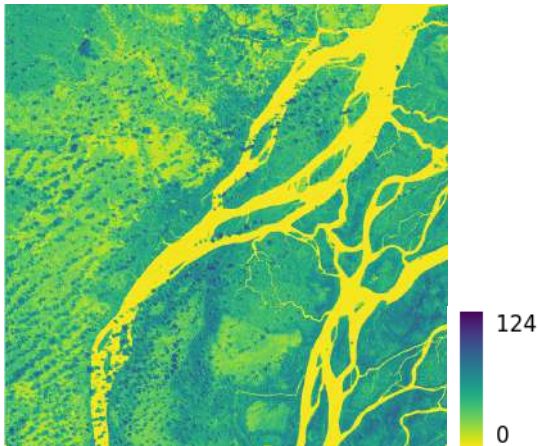
[38TMM - 6] GEDI L4B



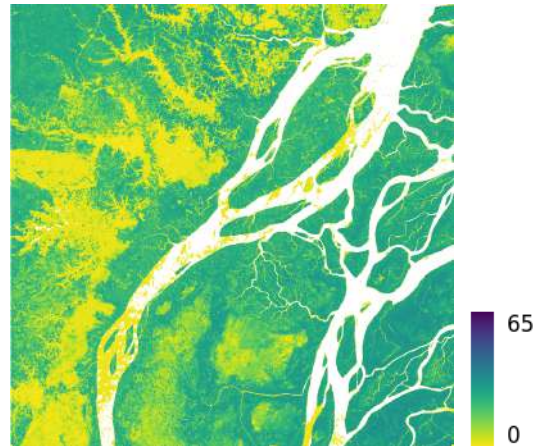
[22MDE - 1] CH, S2, MTL  
(GNLL constant)



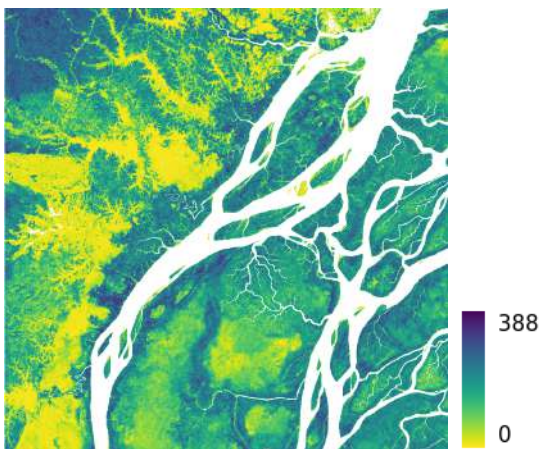
[22MDE - 2] CH, S2, MTL  
(RMSE uncertainty)



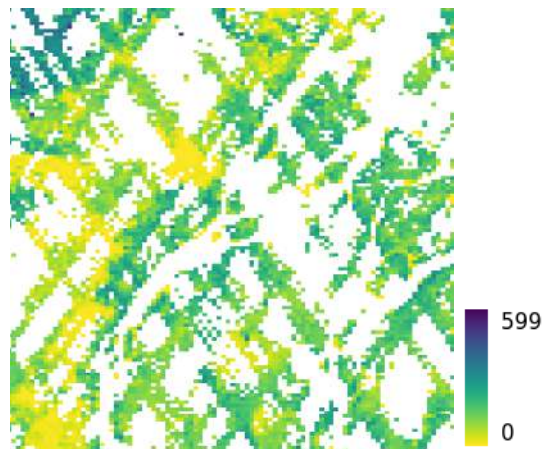
[22MDE - 3] S2, MTL  
(RMSE uncertainty)



[22MDE - 4] CH, S2, LNLL, STL



[22MDE - 5] Clémence



[22MDE - 6] GEDI L4B



## 5.4 Comparison with other biomass maps

In order to benchmark our best model’s performance on FIA reference data (cf. Section 3.1.5), one has to aggregate the biomass predictions at the hexagon-level. This is performed in two steps: we first downsample the predictions from a  $50m \times 50m$  pixel resolution to a  $1km \times 1km$  pixel resolution, and average the predictions to reach the hexagon-level resolution. We experiment with two ways of performing the downsampling step:

- the first method, *average*, consists in taking the average of the  $50m \times 50m$  predictions;
- the second method, *max*, consists in taking the maximum of the  $50m \times 50m$  predictions.

In Figures 29 and 30 we report the binned residuals for, respectively, the average-pooled method, and the max-pooled method. Additionally, Figure 28 displays the distribution of biomass values in the FIA dataset.

In the average downsampling case, the performance of the model is clearly sub-optimal: it consistently underpredicts biomass, by a very large amount compared to the other models. The only positive outcome is that it outperforms FIA in the  $0 - 25$  Mg/ha bin.

In the maximum downsampling case, the model performs slightly better. It is rather accurate in the  $0 - 25$  Mg/ha bin. In some cases, e.g. in the optimistic case for the  $25 - 50$  Mg/ha and  $50 - 75$  Mg/ha bins, or in the pessimistic case for the  $50 - 75$  Mg/ha bin, it even outperforms the other models. However, as soon as we reach the  $75$  Mg/ha threshold, the model goes back to consistently underpredicting biomass.

As the FIA data distribution closely follows a long-tail distribution, similar to that of Yuchang’s and Clémence’s datasets, it was expected that Clémence’s model would overall achieve a better performance on the FIA data. However, the consistent underestimation of medium-to-high biomass values was not expected, as the data distribution shift would suggest that such values would be more accurately estimated using my model than using Clémence’s model. We blame most of this behaviour on the low number of training samples, and elaborate on other contributing reasons in Section 6.2.3.

It is worth noting that these results should be taken with a grain of salt: the difference in data quantity and data distribution across datasets is significant, and the FIA data distribution is very different to that of the training dataset.

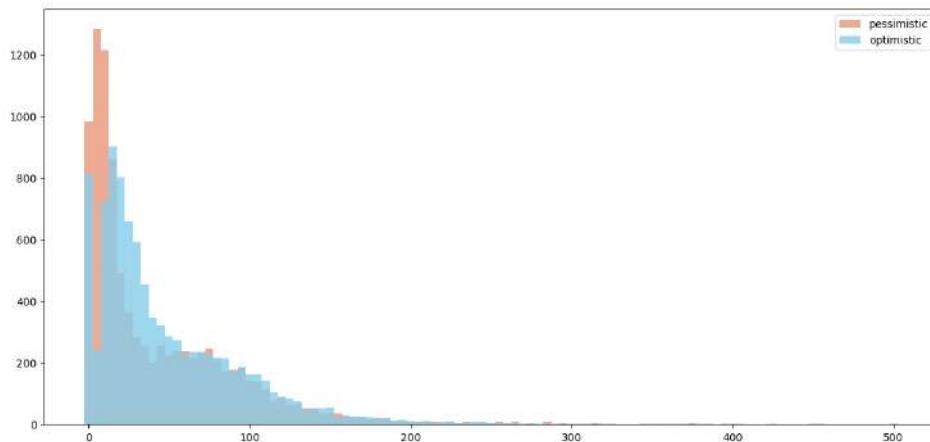


Figure 28: FIA AGB values distribution, for both scenarios.

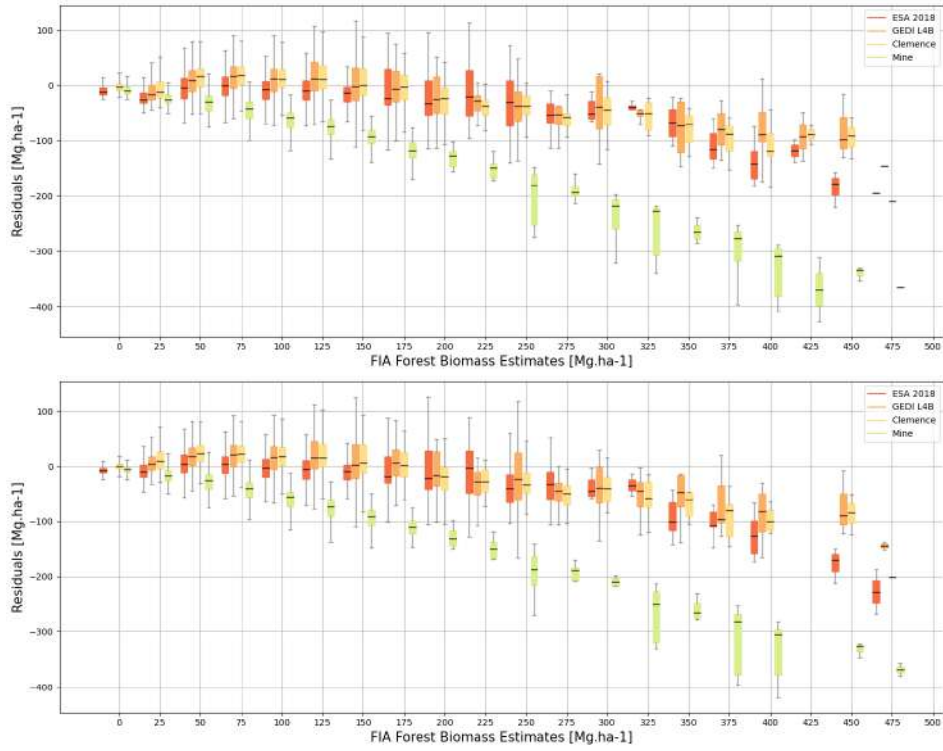


Figure 29: Binned residuals for the average-pooled model, for the FIA optimistic (top) and pessimistic (bottom) scenarios.

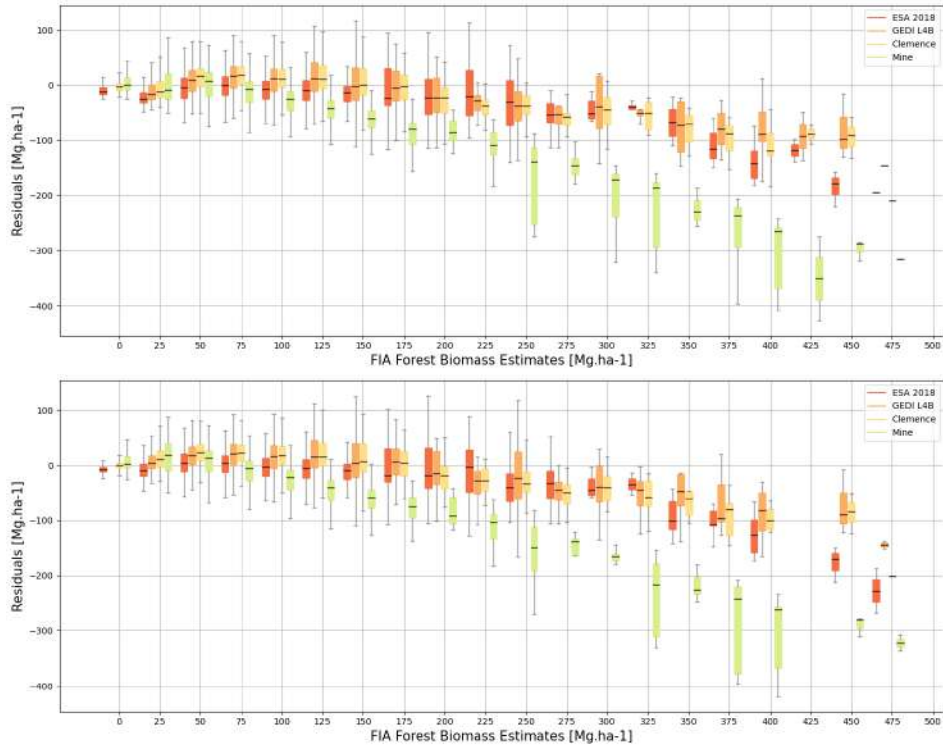


Figure 30: Binned residuals for the max-pooled model, for the FIA optimistic (top) and pessimistic (bottom) scenarios.

## 6 Conclusion

In this chapter, we take a step back, and consider the outcomes and limitations of this work.

### 6.1 Outcomes

First of all, the use of Sentinel-2 data is a valuable approach for estimating biomass. However, the incorporation of canopy height data can further improve the accuracy of the biomass estimation model. Using both Sentinel-2 and canopy height data outperformed using only one of these sources. While not surprising per se, one could have expected models with only Sentinel-2 data as input to learn as much as models with Sentinel-2 data and canopy height data, as the latter was itself extracted from the former. One possible explanation for this finding is that the canopy height data used in the study was obtained from multiple Sentinel-2 images at different timestamps. Therefore, combining the information from multiple timestamps led to a more accurate estimation of the canopy height, which, in turn, improved the accuracy of the biomass estimation model.

The study also found that multi-task learning (MTL) can help with biomass estimation when Sentinel-2 data is available (either as the sole input, or alongside canopy height data). However, MTL does not provide much added benefit when only canopy height data is present. The reason for this might be that the signals provided by canopy height data alone seem insufficient, and MTL tends to confuse the task.

Lastly, the study raises a crucial question: Would providing more data cancel out the added benefit of MTL? The answer to this question is not straightforward since it depends on various factors such as the quality and quantity of the data provided. One possibility is that MTL is only useful in low-data regimes since it helps to leverage information across different tasks. However, the study found that MTL can improve the model’s performance even when more data is available. In particular, MTL provides more added benefit when both Sentinel-2 data and canopy height data are present than when only Sentinel-2 data is used as input. Therefore, more studies are required to determine the optimal combination of data and techniques that can enhance the accuracy of biomass estimation models.

As a final word, the outcome of this work indicates that there’s only so much you can learn from Sentinel-2 data only. Increasing model complexity, experimenting with various MTL approaches, helped the regression task to some extent. But providing additional signals, like canopy height, helped the task even more. This suggests that, while engineering state-of-the-art networks can get you quite far, there exists a certain threshold that, to be overtaken, requires more data. Both in terms of data samples, and by leveraging additional data sources with different modalities.

### 6.2 Limitations

We identify several limitations pertaining to this work, which we organize thematically.

#### 6.2.1 Measuring biomass

**Reference data** When dealing with any type of learning, the quality of the ground-truth data is a crucial factor to ensure the subsequent quality and robustness of the model. With regards to the *biomass* reference data at hand, while it has been processed by GEDI, the underlying data relies on “field” biomass values. There are many potential sources of bias when estimating those biomass values, including methodological, human and equipment biases when measuring tree dimensions, and the use of incorrect allometric models when estimating tree biomass from these measurements [23]. As a result, the consistent accuracy and uncertainty assessment of continental and global scale AGB maps are hampered by the lack of a global reference dataset [65, 66]. Additionally, the lack of availability of reference biomass data from Majority World countries [27, 67, 68] raises ethical concerns.

**Earth Observations** Additionally, previous works [69, 70] have shown that many factors can negatively affect the potential of Earth observations (EO) to accurately estimate biomass. Those factors include saturation at high biomass, mixed soil and vegetation components influencing signals from low biomass areas, and variations in the EO signal due to environmental effects such as rain and snow. And while leveraging EO is crucial to monitor biomass stocks, as it allows us to do it more frequently, across longer periods of time, and over larger spatial scales when compared to forest inventories, forest inventory plots were never designed to be used in combination with pixels from EO. Data manually collected on the ground can hugely differ from remote satellite measurements in terms of spatial resolution and coverage, so discrepancies are usually introduced when trying to generate wall-to-wall remote sensing-derived products. As a result, a consistently observed pattern in current AGB maps derived from space data is overestimation of low biomass and underestimation of high biomass [65, 69]. This is consistent with what we observe in this work, as well as in [9]’s and [8]’s works. Pointers for a somewhat harmonized evaluation procedure could include a framework (see Figure 31) developed by [7] to compare AGB maps with AGB estimates from a global collection of National Forest Inventories and research plots that accounts for the uncertainty of plot AGB errors.

### 6.2.2 Dataset

**Quantity** One cannot cast aside the possibility that the reduced size of the dataset (compared to that of Clémence and Yuchang, for example) might be at fault regarding the models’ inability to learn more, at least on a global scale. Recovering additional data points, notably by correcting the matching procedure, could have changed the outcome of this work to some extent.

**Temporal range** Restricting the temporal range of Sentinel-2 images from May to September, for the *leaf-on season*, might be an issue for the Southern hemisphere, e.g., Australia where their winter spans June through August. Such seasonal changes can impact the quality of the features that can be extracted from the Sentinel-2 images: should they be covered in snow, the models will struggle to learn anything, and will most likely heavily rely on the geographical prior. Additionally, this results in a temporal offset between the Sentinel-2 image and the reference GEDI biomass values: a GEDI footprint might have been sampled in June and the corresponding Sentinel-2 image might have been shot in September. A better way to consistently match Sentinel-2 images and GEDI footprints could be to leverage the date at which the footprint was obtained; or to leverage the *leaf\_off* flag in the GEDI L2A data product, which indicates whether the observation was recorded during leaf-off conditions in DBT and DNT prediction strata (see Figure 32). It is worth noting that taking such steps might prove superfluous. While biomass production does vary over the course of the growing season, it may not change significantly over a period of a few months. In some forests, biomass production may be relatively constant throughout the year due to the presence of evergreen trees that retain their leaves year-round. In contrast, in grasslands, biomass production may be more strongly tied to seasonal changes in rainfall and temperature. Additionally, one might view this temporal mismatch as introducing noise in the dataset, which can help the model generalize better.

### 6.2.3 Methods

**MTL** With respect to Multi-Task Learning, many promising methods were not investigated, for lack of time. Those include using a bottom- vs. top-layer approach, as explained in 2.3. Other weighting schemes, particularly those taking into account the history of the gradients, could have been experimented with. Additionally, for a given architecture (e.g., Xception), all the *heads* were identical, independently of the task being performed. While this remains the most widespread approach, one might experiment with each *head* being tailored to the task at hand. Finally, the *heads* were also identical (modulo the number of input features, and the number of hidden layers) across architectures, making the ratio of *head*-parameters vs. *body*-parameters highly inconsistent.

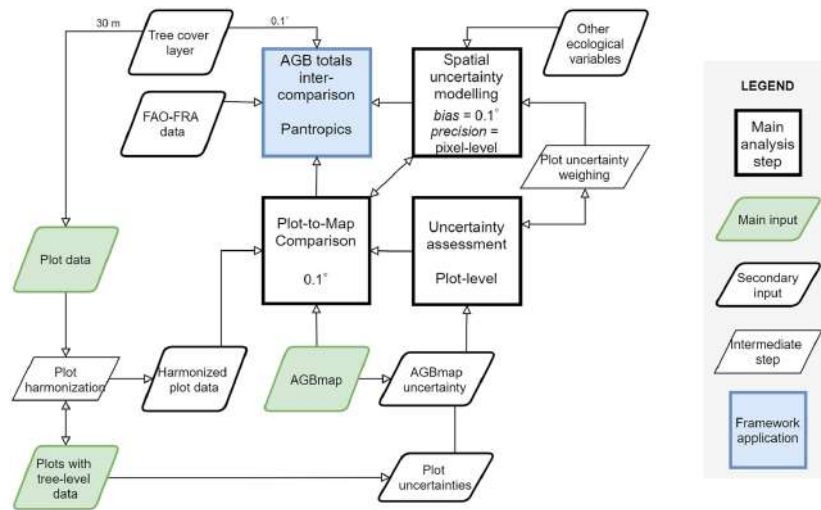


Figure 31: Schematic of the framework developed by [7].

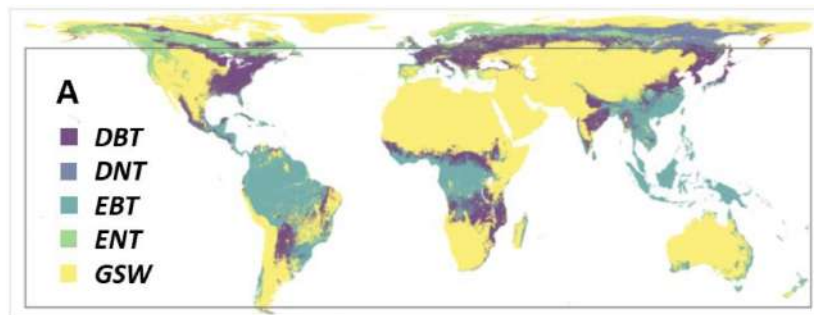


Figure 32: The GEDI04\_A [5] global stratification of plant functional types (PFT)

**Under-estimation** The consistent under-estimation of biomass is the aspect of the model that should be fixed in priority. It might be due to the fact that the best performing model, benchmarked against FIA, was trained using RMSE. Considering that the training dataset has a wide range of values, not balanced equally, this loss may not be the best choice for guiding the performance of the model. Other losses, like the GNLL, that take into account the uncertainty linked with the prediction, might be more suited in another framework. Additionally, it may be necessary to employ techniques such as resampling or weighting.

**Cloud mask** In previous works [6, 9], the authors used the Level 2A cloud probability mask to exclude cloudy patches: pixels with  $>10\%$  cloud probability are considered as cloudy, and any patch with  $\geq 10\%$  cloudy pixels is discarded. They suggest that this approach avoids showing the network patches carrying too little signal, but enables it to learn textural features that are robust near the cloud borders. We did not proceed to mask patches this way.

### 6.3 Future Steps

Now that we have put things into perspective, it is time to reflect on the future steps that could be taken.

#### 6.3.1 Working with big data

The size of this dataset (although much smaller compared to Yuchang’s and Clémence’s) presented a number of challenges, specifically, the slightest of bugs resulted in an exorbitant amount of time being spent re-running the code on the entire dataset. In the future, I would rather start on a geographical subset of the data. Once the methods and framework have been finalized, they can then be extended to the global scale.

#### 6.3.2 Biomass prediction

The present approach to biomass prediction could be meliorated through a number of steps. First of all, the creation of a reference dataset, on which all methods (notably Clémence’s and Yuchang’s) would be trained and tested. This would render the, currently difficultly comparable methods comparable. Bench-marking the methods on the FIA data is a first step in this direction, but the GEDI footprints from the continental US are not removed from the training data, so the metrics can be biased to some extent. Additionally, looking for additional data sources might be what would help the learning most. In addition to Sentinel-2 imagery, there are several data sources that can be used to regress biomass density: LiDAR (Light Detection and Ranging) data can provide information about the height and structure of vegetation; SAR (Synthetic Aperture Radar) data can penetrate through clouds and provide information about the structure and moisture content of vegetation; other remote sensing data, such as hyperspectral imagery and thermal imagery, can provide additional information about vegetation properties.

#### 6.3.3 Biomass change detection

Biomass change detection plays a critical role in environmental monitoring and management. With the growing concern about climate change, the ability to accurately quantify and monitor changes in vegetation biomass is becoming increasingly important. For example, changes in biomass can indicate deforestation or forest degradation, which can lead to loss of habitat, soil erosion, and carbon emissions. However, biomass change detection is a complex and challenging task. Directly estimating change in biomass raises the issue of which reference data to use, as there is currently no large-scale biomass change dataset. Another approach would be to compare predicted AGB maps across time-periods (e.g. annually). One such statistically-robust method could be Bayesian hierarchical modeling [71]: one specifies a probability distribution for the biomass values in each map, including the uncertainties of

the individual biomass values; the model would then estimate the parameters of these distributions, as well as the difference between the two maps, and provide a posterior distribution for the difference that takes into account the uncertainties associated with each biomass value. An additional issue is to distinguish between natural variability and actual changes in biomass. One could leverage climate data such as temperature and precipitation, which can influence the growth and productivity of vegetation, to somewhat teach the model how the biomass should be evolving. This could contribute to the explainability of the model.

## References

- [1] H. Ritchie and M. Roser, “Forests and deforestation,” *Our World in Data*, 2021, <https://ourworldindata.org/forests-and-deforestation>.
- [2] G. Reiersen, D. Dao, B. Lütjens, K. Klemmer, X. Zhu, and C. Zhang, “Tackling the overestimation of forest carbon with deep learning and aerial imagery,” *arXiv preprint arXiv:2107.11320*, 2021.
- [3] T. Kattenborn, J. Leitloff, F. Schiefer, and S. Hinz, “Review on convolutional neural networks (CNN) in vegetation remote sensing,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 173, pp. 24–49, Mar. 2021. [Online]. Available: <https://doi.org/10.1016/j.isprsjprs.2020.12.010>
- [4] S. Ruder, “An overview of multi-task learning in deep neural networks,” 2017. [Online]. Available: <https://arxiv.org/abs/1706.05098>
- [5] R. Dubayah, J. Armston, J. Kellner, L. Duncanson, S. Healey, P. Patterson, S. Hancock, H. Tang, J. Bruening, M. Hofton, J. Blair, and S. Luthcke, “Gedi 14a footprint level aboveground biomass density, version 2.1,” 2022. [Online]. Available: [https://daac.ornl.gov/cgi-bin/dsvviewer.pl?ds\\_id=2056](https://daac.ornl.gov/cgi-bin/dsvviewer.pl?ds_id=2056)
- [6] N. Lang, N. Kalischek, J. Armston, K. Schindler, R. Dubayah, and J. D. Wegner, “Global canopy height regression and uncertainty estimation from GEDI LIDAR waveforms with deep ensembles,” *Remote Sensing of Environment*, vol. 268, p. 112760, Jan. 2022. [Online]. Available: <https://doi.org/10.1016/j.rse.2021.112760>
- [7] A. Araza, S. de Bruin, M. Herold, S. Quegan, N. Labriere, P. Rodriguez-Veiga, V. Avitabile, M. Santoro, E. T. Mitchard, C. M. Ryan, O. L. Phillips, S. Willcock, H. Verbeeck, J. Carreiras, L. Hein, M.-J. Schelhaas, A. M. Pacheco-Pascagaza, P. da Conceição Bispo, G. V. Laurin, G. Vieilledent, F. Slik, A. Wijaya, S. L. Lewis, A. Morel, J. Liang, H. Sukhdeo, D. Schepaschenko, J. Cavlovic, H. Gilani, and R. Lucas, “A comprehensive framework for assessing the accuracy and uncertainty of global above-ground biomass maps,” *Remote Sensing of Environment*, vol. 272, p. 112917, Apr. 2022. [Online]. Available: <https://doi.org/10.1016/j.rse.2022.112917>
- [8] M. Rüetschi, Y. Jiang, N. Lang, A. Becker, L. T. Waser, M. Marty, K. Schindler, J. D. Wegner, and C. Ginzler, “Annual vegetation height maps based on sentinel-2 data – potential applications for the swiss national forest inventory,” 2022. [Online]. Available: [https://scholar.google.com/citations?view\\_op=view\\_citation&hl=fr&user=u9k4HecAAAAJ&citation\\_for\\_view=u9k4HecAAAAJ:9yKSN-GCB0IC](https://scholar.google.com/citations?view_op=view_citation&hl=fr&user=u9k4HecAAAAJ&citation_for_view=u9k4HecAAAAJ:9yKSN-GCB0IC)
- [9] C. Lanfranchi, “Global biomass mapping and uncertainty estimation from GEDI LIDAR data using bayesian deep learning,” 2022. [Online]. Available: [https://github.com/ClemenceLanfranchi/MasterThesis/blob/main/Master\\_Thesis\\_PRS\\_report.pdf](https://github.com/ClemenceLanfranchi/MasterThesis/blob/main/Master_Thesis_PRS_report.pdf)
- [10] T. W. Crowther, H. B. Glick, K. R. Covey, C. Bettigole, D. S. Maynard, S. M. Thomas, J. R. Smith, G. Hintler, M. C. Duguid, G. Amatulli, M.-N. Tuanmu, W. Jetz, C. Salas, C. Stam, D. Piotto, R. Tavani, S. Green, G. Bruce, S. J. Williams, S. K. Wiser, M. O. Huber, G. M. Hengeveld, G.-J. Nabuurs, E. Tikhonova, P. Borchardt, C.-F. Li, L. W. Powrie, M. Fischer, A. Hemp, J. Homeier, P. Cho, A. C. Vibrans, P. M. Umunay, S. L. Piao, C. W. Rowe, M. S. Ashton, P. R. Crane, and M. A. Bradford, “Mapping tree density at a global scale,” *Nature*, vol. 525, no. 7568, pp. 201–205, Sep. 2015. [Online]. Available: <https://doi.org/10.1038/nature14967>
- [11] P. G. Curtis, C. M. Slay, N. L. Harris, A. Tyukavina, and M. C. Hansen, “Classifying drivers of global forest loss,” *Science*, vol. 361, no. 6407, pp. 1108–1111, Sep. 2018. [Online]. Available: <https://doi.org/10.1126/science.aau3445>



- [12] B. R. Scheffers, L. N. Joppa, S. L. Pimm, and W. F. Laurance, “What we know and don’t know about earth’s missing biodiversity,” *Trends in Ecology & Evolution*, vol. 27, no. 9, pp. 501–510, Sep. 2012. [Online]. Available: <https://doi.org/10.1016/j.tree.2012.05.008>
- [13] P. Goymer, “A trillion trees,” *Nature Ecology & Evolution*, vol. 2, no. 2, pp. 208–209, Jan. 2018. [Online]. Available: <https://doi.org/10.1038/s41559-018-0464-z>
- [14] K. Hyams and T. Fawcett, “The ethics of carbon offsetting,” *Wiley Interdisciplinary Reviews: Climate Change*, vol. 4, pp. 91–98, 2013.
- [15] P. Lourenço, “Biomass estimation using satellite-based data,” in *Forest Biomass*, A. C. Gonçalves, A. Sousa, and I. Malico, Eds. Rijeka: IntechOpen, 2021, ch. 3. [Online]. Available: <https://doi.org/10.5772/intechopen.93603>
- [16] J. Blackard, M. Finco, E. Helmer, G. Holden, M. Hoppus, D. Jacobs, A. Lister, G. Moisen, M. Nelson, R. Riemann, B. Ruefenacht, D. Salajanu, D. Weyermann, K. Winterberger, T. Brandeis, R. Czaplowski, R. McRoberts, P. Patterson, and R. Tymcio, “Mapping u.s. forest biomass using nationwide forest inventory data and moderate resolution information,” *Remote Sensing of Environment*, vol. 112, no. 4, pp. 1658–1677, 2008, remote Sensing Data Assimilation Special Issue. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0034425707004269>
- [17] R. Booyesen, R. Gloaguen, S. Lorenz, R. Zimmermann, and P. A. Nex, “Geological remote sensing,” in *Encyclopedia of Geology (Second Edition)*, second edition ed., D. Alderton and S. A. Elias, Eds. Oxford: Academic Press, 2021, pp. 301–314. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/B978012409548912127X>
- [18] M. García, D. Riaño, E. Chuvieco, and F. M. Danson, “Estimating biomass carbon stocks for a mediterranean forest in central spain using lidar height and intensity data,” *Remote Sensing of Environment*, vol. 114, no. 4, pp. 816–830, 2010. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0034425709003551>
- [19] M. Hayashi, N. Saigusa, Y. Yamagata, and T. Hirano, “Regional forest biomass estimation using ICESat/GLAS spaceborne LiDAR over borneo,” *Carbon Management*, vol. 6, no. 1-2, pp. 19–33, Mar. 2015. [Online]. Available: <https://doi.org/10.1080/17583004.2015.1066638>
- [20] W. Ni-Meister, A. Rojas, and S. Lee, “Direct use of large-footprint lidar waveforms to estimate aboveground biomass,” *Remote Sensing of Environment*, vol. 280, p. 113147, 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0034425722002619>
- [21] S. Qin, S. Nie, Y. Guan, D. Zhang, C. Wang, and X. Zhang, “Forest emissions reduction assessment using airborne lidar for biomass estimation,” *Resources, Conservation and Recycling*, vol. 181, p. 106224, 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0921344922000726>
- [22] F. Zhao, Q. Guo, and M. Kelly, “Allometric equation choice impacts lidar-based forest biomass estimates: A case study from the sierra national forest, ca,” *Agricultural and Forest Meteorology*, vol. 165, pp. 64–72, 2012. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0168192312001967>
- [23] K. Calders, H. Verbeeck, A. Burt, N. Origo, J. Nightingale, Y. Malhi, P. Wilkes, P. Raunonen, R. G. H. Bunce, and M. Disney, “Laser scanning reveals potential underestimation of biomass carbon in temperate forest,” *Ecological Solutions and Evidence*, vol. 3, no. 4, Oct. 2022. [Online]. Available: <https://doi.org/10.1002/2688-8319.12197>

- [24] M. Demol, H. Verbeeck, B. Gielen, J. Armston, A. Burt, M. Disney, L. Duncanson, J. Hackenberg, D. Kükenbrink, A. Lau, P. Ploton, A. Sevdien, A. Stovall, S. M. Takoudjou, L. Volkova, C. Weston, V. Wortel, and K. Calders, “Estimating forest above-ground biomass with terrestrial laser scanning: Current status and future directions,” *Methods in Ecology and Evolution*, vol. 13, no. 8, pp. 1628–1639, Jun. 2022. [Online]. Available: <https://doi.org/10.1111/2041-210x.13906>
- [25] K. Calders, G. Newnham, A. Burt, S. Murphy, P. Raunonen, M. Herold, D. Culvenor, V. Avitabile, M. Disney, J. Armston, and M. Kaasalainen, “Nondestructive estimates of above-ground biomass using terrestrial laser scanning,” *Methods in Ecology and Evolution*, vol. 6, no. 2, pp. 198–208, Nov. 2014. [Online]. Available: <https://doi.org/10.1111/2041-210x.12301>
- [26] S. S. Saatchi, N. L. Harris, S. Brown, M. Lefsky, E. T. A. Mitchard, W. Salas, B. R. Zutta, W. Buermann, S. L. Lewis, S. Hagen, S. Petrova, L. White, M. Silman, and A. Morel, “Benchmark map of forest carbon stocks in tropical regions across three continents,” *Proceedings of the National Academy of Sciences*, vol. 108, no. 24, pp. 9899–9904, 2011. [Online]. Available: <https://www.pnas.org/content/108/24/9899>
- [27] Global Forest Watch, “Aboveground live woody biomass density,” 2019. [Online]. Available: <https://www.globalforestwatch.org>
- [28] A. Burt, M. B. Vicari, A. C. L. da Costa, I. Coughlin, P. Meir, L. Rowland, and M. Disney, “New insights into large tropical tree mass and structure from direct harvest and terrestrial lidar,” *Royal Society Open Science*, vol. 8, no. 2, Feb. 2021. [Online]. Available: <https://doi.org/10.1098/rsos.201458>
- [29] S. Zolkos, S. Goetz, and R. Dubayah, “A meta-analysis of terrestrial aboveground biomass estimation using lidar remote sensing,” *Remote Sensing of Environment*, vol. 128, pp. 289–298, 2013. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0034425712004051>
- [30] A. Kendall and Y. Gal, “What uncertainties do we need in bayesian deep learning for computer vision?” 2017. [Online]. Available: <https://arxiv.org/abs/1703.04977>
- [31] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. V. Dillon, B. Lakshminarayanan, and J. Snoek, “Can you trust your model’s uncertainty? evaluating predictive uncertainty under dataset shift,” 2019. [Online]. Available: <https://arxiv.org/abs/1906.02530>
- [32] B. Lakshminarayanan, A. Pritzel, and C. Blundell, “Simple and scalable predictive uncertainty estimation using deep ensembles,” 2016. [Online]. Available: <https://arxiv.org/abs/1612.01474>
- [33] Y. Gal and Z. Ghahramani, “Dropout as a bayesian approximation: Representing model uncertainty in deep learning,” in *Proceedings of The 33rd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. F. Balcan and K. Q. Weinberger, Eds., vol. 48. New York, New York, USA: PMLR, 20–22 Jun 2016, pp. 1050–1059. [Online]. Available: <https://proceedings.mlr.press/v48/gal16.html>
- [34] Y. Zhang and Q. Yang, “An overview of multi-task learning,” *National Science Review*, vol. 5, no. 1, pp. 30–43, Sep. 2017. [Online]. Available: <https://doi.org/10.1093/nsr/nwx105>
- [35] R. Caruana, *Machine Learning*, vol. 28, no. 1, pp. 41–75, 1997. [Online]. Available: <https://doi.org/10.1023/a:1007379606734>
- [36] C. Stein, “Inadmissibility of the usual estimator for the variance of a normal distribution with unknown mean,” *Annals of the Institute of Statistical Mathematics*, vol. 16, no. 1, pp. 155–160, Dec. 1964. [Online]. Available: <https://doi.org/10.1007/bf02868569>
- [37] L. Liebel and M. Körner, “Auxiliary tasks in multi-task learning,” 2018. [Online]. Available: <https://arxiv.org/abs/1805.06334>

- [38] M. Islam, V. S. Vibashan, and H. Ren, “Ap-mtl: Attention pruned multi-task learning model for real-time instrument detection and segmentation in robot-assisted surgery,” in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 8433–8439.
- [39] L. Zhang, Q. Yang, X. Liu, and H. Guan, “Rethinking hard-parameter sharing in multi-domain learning,” 2021. [Online]. Available: <https://arxiv.org/abs/2107.11359>
- [40] L. Duong, T. Cohn, S. Bird, and P. Cook, “Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser,” in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Beijing, China: Association for Computational Linguistics, Jul. 2015, pp. 845–850. [Online]. Available: <https://aclanthology.org/P15-2139>
- [41] M. Yuan, A. Ekici, Z. Lu, and R. Monteiro, “Dimension reduction and coefficient estimation in multivariate linear regression,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 69, no. 3, pp. 329–346, Jun. 2007. [Online]. Available: <https://doi.org/10.1111/j.1467-9868.2007.00591.x>
- [42] Z. Zheng, Y. Zhong, S. Tian, A. Ma, and L. Zhang, “ChangeMask: Deep multi-task encoder-transformer-decoder architecture for semantic change detection,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 183, pp. 228–239, Jan. 2022. [Online]. Available: <https://doi.org/10.1016/j.isprsjprs.2021.10.015>
- [43] T. Gong, X. Zheng, and X. Lu, “Remote sensing scene classification with multi-task learning,” in *Proceedings of the 7th China High Resolution Earth Observation Conference (CHREOC 2020)*. Springer Nature Singapore, 2022, pp. 403–418. [Online]. Available: [https://doi.org/10.1007/978-981-16-5735-1\\_30](https://doi.org/10.1007/978-981-16-5735-1_30)
- [44] H. Li, J. Zech, D. Hong, P. Ghamisi, M. Schultz, and A. Zipf, “Leveraging OpenStreetMap and multimodal remote sensing data with joint deep learning for wastewater treatment plants detection,” *International Journal of Applied Earth Observation and Geoinformation*, vol. 110, p. 102804, Jun. 2022. [Online]. Available: <https://doi.org/10.1016/j.jag.2022.102804>
- [45] Q. Yang, Q. Yuan, M. Gao, and T. Li, “A new perspective to satellite-based retrieval of ground-level air pollution: Simultaneous estimation of multiple pollutants based on physics-informed multi-task learning,” *Science of The Total Environment*, vol. 857, p. 159542, Jan. 2023. [Online]. Available: <https://doi.org/10.1016/j.scitotenv.2022.159542>
- [46] A. S. Keceli, A. Kaya, C. Catal, and B. Tekinerdogan, “Deep learning-based multi-task prediction system for plant disease and species detection,” *Ecological Informatics*, vol. 69, p. 101679, Jul. 2022. [Online]. Available: <https://doi.org/10.1016/j.ecoinf.2022.101679>
- [47] M. Lu, J. Liu, F. Wang, and Y. Xiang, “Multi-task learning of relative height estimation and semantic segmentation from single airborne RGB images,” *Remote Sensing*, vol. 14, no. 14, p. 3450, Jul. 2022. [Online]. Available: <https://doi.org/10.3390/rs14143450>
- [48] A. Ma, D. Chen, Y. Zhong, Z. Zheng, and L. Zhang, “National-scale greenhouse mapping for high spatial resolution remote sensing imagery using a dense object dual-task deep learning framework: A case study of china,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 181, pp. 279–294, Nov. 2021. [Online]. Available: <https://doi.org/10.1016/j.isprsjprs.2021.08.024>
- [49] K. Cao, Z. Zhang, Y. Li, M. Xie, and W. Zheng, “Surveillance of ship emissions and fuel sulfur content based on imaging detection and multi-task deep learning,” *Environmental Pollution*, vol. 288, p. 117698, Nov. 2021. [Online]. Available: <https://doi.org/10.1016/j.envpol.2021.117698>
- [50] A. Kendall, Y. Gal, and R. Cipolla, “Multi-task learning using uncertainty to weigh losses for scene geometry and semantics,” 2017. [Online]. Available: <https://arxiv.org/abs/1705.07115>

- [51] P. Rizzoli, M. Martone, C. González, C. Wecklich, D. B. Tridon, B. Bräutigam, M. Bachmann, D. Schulze, T. Fritz, M. Huber, B. Wessel, G. Krieger, M. Zink, and A. Moreira, “Generation and performance assessment of the global tandem-x digital elevation model,” *Isprs Journal of Photogrammetry and Remote Sensing*, vol. 132, pp. 119–139, 2017.
- [52] R. H. Macarthur and H. S. Horn, “Foliage profile by vertical measurements,” *Ecology*, vol. 50, pp. 802–804, 1969.
- [53] A. López-Urrutia, “Macroscopic patterns in marine plankton,” in *Encyclopedia of Biodiversity*. Elsevier, 2013, pp. 667–680. [Online]. Available: <https://doi.org/10.1016/b978-0-12-384719-5.00291-4>
- [54] M. Friedl and D. Sulla-Menashe, “Mcd12q1 modis/terra+aqua land cover type yearly l3 global 500m sin grid v006,” 2019. [Online]. Available: <https://lpdaac.usgs.gov/products/mcd12q1v006/>
- [55] J. Carreiras, S. Quegan, T. Le Toan, D. Ho, D. HO TONG MINH, S. Saatchi, N. Carvalhais, M. Reichstein, and K. Scipal, “Coverage of high biomass forests by the esa biomass mission under defense restrictions,” *Remote Sensing of Environment*, vol. 196, pp. 154–162, 05 2017.
- [56] D. G. Bonett and T. A. Wright, “Sample size requirements for estimating pearson, kendall and spearman correlations,” *Psychometrika*, vol. 65, no. 1, pp. 23–28, Mar. 2000. [Online]. Available: <https://doi.org/10.1007/bf02294183>
- [57] V. Kotu and B. Deshpande, “Classification,” in *Data Science*. Elsevier, 2019, pp. 65–163. [Online]. Available: <https://doi.org/10.1016/b978-0-12-814761-0.00004-6>
- [58] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, “Lightgbm: A highly efficient gradient boosting decision tree,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS’17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 3149–3157.
- [59] J. Gast and S. Roth, “Lightweight probabilistic deep networks,” 2018. [Online]. Available: <https://arxiv.org/abs/1805.11327>
- [60] A. Becker, S. Russo, S. Puliti, N. Lang, K. Schindler, and J. D. Wegner, “Country-wide retrieval of forest structure from optical and sar satellite imagery with deep ensembles,” 2021. [Online]. Available: <https://arxiv.org/abs/2111.13154>
- [61] N. Lang, K. Schindler, and J. D. Wegner, “Country-wide high-resolution vegetation height mapping with sentinel-2,” *Remote Sensing of Environment*, vol. 233, p. 111347, Nov. 2019. [Online]. Available: <https://doi.org/10.1016/j.rse.2019.111347>
- [62] F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” 2016. [Online]. Available: <https://arxiv.org/abs/1610.02357>
- [63] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014. [Online]. Available: <https://arxiv.org/abs/1412.6980>
- [64] X. Wang, D. Kondratyuk, E. Christiansen, K. M. Kitani, Y. Alon, and E. Eban, “Wisdom of committees: An overlooked approach to faster and more accurate models,” 2020. [Online]. Available: <https://arxiv.org/abs/2012.01988>
- [65] P. Rodríguez-Veiga, S. Quegan, J. Carreiras, H. J. Persson, J. E. Fransson, A. Hoscilo, D. Ziólkowski, K. Stereńczak, S. Lohberger, M. Stängel, A. Berninger, F. Siegert, V. Avitabile, M. Herold, S. Mermoz, A. Bouvet, T. L. Toan, N. Carvalhais, M. Santoro, O. Cartus, Y. Rauste, R. Mathieu, G. P. Asner, C. Thiel, C. Pathe, C. Schmullius, F. M. Seifert, K. Tansey, and H. Balzter, “Forest biomass retrieval approaches from earth observation in different biomes,”

- International Journal of Applied Earth Observation and Geoinformation*, vol. 77, pp. 53–68, May 2019. [Online]. Available: <https://doi.org/10.1016/j.jag.2018.12.008>
- [66] D. Schimel, R. Pavlick, J. B. Fisher, G. P. Asner, S. Saatchi, P. Townsend, C. Miller, C. Frankenberg, K. Hibbard, and P. Cox, “Observing terrestrial ecosystems and the carbon cycle from space,” *Global Change Biology*, vol. 21, no. 5, pp. 1762–1776, Feb. 2015. [Online]. Available: <https://doi.org/10.1111/gcb.12822>
- [67] A. E. White, D. A. Lutz, R. B. Howarth, and J. R. Soto, “Small-scale forestry and carbon offset markets: An empirical study of vermont current use forest landowner willingness to accept carbon credit programs,” *PLOS ONE*, vol. 13, no. 8, pp. 1–24, 08 2018. [Online]. Available: <https://doi.org/10.1371/journal.pone.0201967>
- [68] L. Duncanson, J. Armston, M. Disney, V. Avitabile, N. Barbier, K. Calders, S. Carter, J. Chave, M. Herold, T. W. Crowther, M. Walkowski, J. Kellner, N. Labrière, R. Lucas, N. MacBean, R. E. McRoberts, V. Meyer, E. Naeset, J. E. Nickeson, K. I. Paul, O. L. Phillips, M. Réjou-Méchain, M. Román, S. Roxburgh, S. Saatchi, D. Schepaschenko, K. Scipal, P. R. Siqueira, A. Whitehurst, and M. Williams, “The importance of consistent global forest aboveground biomass product validation,” *Surveys in Geophysics*, vol. 40, pp. 979–999, 2019. [Online]. Available: <https://doi.org/10.1007/s10712-019-09538-8>
- [69] D. Schepaschenko, J. Chave, O. L. Phillips, S. L. Lewis, S. J. Davies, M. Réjou-Méchain, P. Sist, K. Scipal, C. Perger, B. Herault, N. Labrière, F. Hofhansl, K. Affum-Baffoe, A. Aleinikov, A. Alonso, C. Amani, A. Araujo-Murakami, J. Armston, L. Arroyo, N. Ascarrunz, C. Azevedo, T. Baker, R. Bałazy, C. Bedeau, N. Berry, A. M. Bilous, S. Y. Bilous, P. Bissiengou, L. Blanc, K. S. Bobkova, T. Braslavskaya, R. Brienen, D. F. R. P. Burslem, R. Condit, A. Cuni-Sanchez, D. Danilina, D. del Castillo Torres, G. Derroire, L. Descroix, E. D. Sotta, M. V. N. d’Oliveira, C. Dresel, T. Erwin, M. D. Evdokimenko, J. Falek, T. R. Feldpausch, E. G. Foli, R. Foster, S. Fritz, A. D. Garcia-Abril, A. Gornov, M. Gornova, E. Gothard-Bassébé, S. Gourlet-Fleury, M. Guedes, K. C. Hamer, F. H. Susanty, N. Higuchi, E. N. H. Coronado, W. Hubau, S. Hubbell, U. Ilstedt, V. V. Ivanov, M. Kanashiro, A. Karlsson, V. N. Karminov, T. Killeen, J.-C. K. Koffi, M. Konovalova, F. Kraxner, J. Krejza, H. Krisnawati, L. V. Krivobokov, M. A. Kuznetsov, I. Lakyda, P. I. Lakyda, J. C. Licona, R. M. Lucas, N. Lukina, D. Lussetti, Y. Malhi, J. A. Manzanera, B. Marimon, B. H. M. Junior, R. V. Martinez, O. V. Martynenko, M. Matsala, R. K. Matyashuk, L. Mazzei, H. Memiaghe, C. Mendoza, A. M. Mendoza, O. V. Moroziuk, L. Mukhortova, S. Musa, D. I. Nazimova, T. Okuda, L. C. Oliveira, P. V. Ontikov, A. F. Osipov, S. Pietsch, M. Playfair, J. Poulsen, V. G. Radchenko, K. Rodney, A. H. Rozak, A. Ruschel, E. Rutishauser, L. See, M. Shchepashchenko, N. Shevchenko, A. Shvidenko, M. Silveira, J. Singh, B. Sonké, C. Souza, K. Stereńczak, L. Stonozhenko, M. J. P. Sullivan, J. Szatniewska, H. Taedoumg, H. ter Steege, E. Tikhonova, M. Toledo, O. V. Trefilova, R. Valbuena, L. V. Gamarra, S. Vasiliev, E. F. Vedrova, S. V. Verhovets, E. Vidal, N. A. Vladimirova, J. Vleminckx, V. A. Vos, F. K. Vozmitel, W. Wanek, T. A. P. West, H. Woell, J. T. Woods, V. Wortel, T. Yamada, Z. S. N. Hajar, and I. C. Zo-Bi, “The forest observation system, building a global reference dataset for remote sensing of forest biomass,” *Scientific Data*, vol. 6, no. 1, Oct. 2019. [Online]. Available: <https://doi.org/10.1038/s41597-019-0196-1>
- [70] M. Santoro, A. Beaudoin, C. Beer, O. Cartus, J. E. Fransson, R. J. Hall, C. Pathe, C. Schmullius, D. Schepaschenko, A. Shvidenko, M. Thurner, and U. Wegmüller, “Forest growing stock volume of the northern hemisphere: Spatially explicit estimates for 2010 derived from envisat ASAR,” *Remote Sensing of Environment*, vol. 168, pp. 316–334, Oct. 2015. [Online]. Available: <https://doi.org/10.1016/j.rse.2015.07.005>
- [71] C. H. Schmid and E. N. Brown, “Bayesian hierarchical models,” in *Numerical Computer Methods, Part C*. Elsevier, 2000, pp. 305–330. [Online]. Available: [https://doi.org/10.1016/s0076-6879\(00\)21200-7](https://doi.org/10.1016/s0076-6879(00)21200-7)

## Appendix

---

Algorithm: Intersection procedure

inputs:

- patches, an array of size (num\_patches x 15 x 15 x 3) and whose channels correspond to the AGBD, latitude, and longitude;
- GEDI, our reference dataset with auxiliary variables.

output: array of size (num\_patches x 15 x 15 x num\_auxiliary\_variables)

---

```
auxiliary_patches = []

for patch in patches :

    # result placeholder
    auxiliary_patch = array(15, 15, num_auxiliary_variables) filled with NaN

    # take the GEDI footprint at the center of the patch
    center_agbd, center_lat, center_lon = patch[7, 7, :]

    # get all data points in GEDI within a 0.001° radius of the footprint
    area = buffer(center_lat, center_lon, 0.001)
    potential_matches = GEDI.in(area)

    # for each potential match, further match on AGBD values
    for m in potential_matches:
        if m.lat == center_lat and m.lon == center_lon and m.agbd == center_agbd:
            auxiliary_patch[7, 7, :] = m.auxiliary_variables
        else:
            auxiliary_patch[7, 7, :] = [NaN for _ in m.auxiliary_variables]

    auxiliary_patches += auxiliary_patch

return auxiliary_patches
```

Figure 33: Intersection procedure pseudo-code.

## Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

**Title of work** (in block letters):

GLOBAL BIOMASS ESTIMATION AND UNCERTAINTY QUANTIFICATION  
WITH MULTI-TASK BAYESIAN DEEP ENSEMBLES

**Authored by** (in block letters):

*For papers written by groups the names of all authors are required.*

**Name(s):**

SIALELLI

**First name(s):**

GHJULIA

With my signature I confirm that

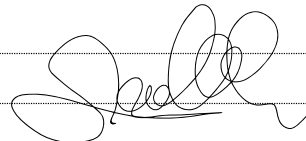
- I have committed none of the forms of plagiarism described in the '[Citation etiquette](#)' information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

**Place, date**

Zürich, 11/04/2023

**Signature(s)**



*For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.*