On the Data Quality of Remotely Sensed Forest Maps

by

Shadi Ghasemitaheri

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2023

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Accurate forest monitoring data are essential for understanding and conserving forest ecosystems. However, the remoteness of forests and the scarcity of ground truth make it hard to identify data quality issues. We present two state-of-the-art forest monitoring datasets, Annual Forest Change (AFC) and GEDI, and highlight their data quality problems. We then introduce a novel method that leverages GEDI to identify data quality issues in AFC. We show that our approach can identify subsets with three times more errors than a random sample, thus, prioritizing expert resources in validating AFC and allowing for more accurate deforestation detection.

Acknowledgements

I thank my supervisor, Professor Lukasz Golab, for his support, patience, and guidance throughout my research.

I thank Professor Srinivasan Keshav for his invaluable advice and guidance throughout this research.

I extend special thanks to Amelia Holcomb for introducing me to this amazing project and for her insight, enthusiasm, and patience during many meetings.

I also thank Professor Paulo Alencar and Professor Charles Clarke, who served as the readers of this thesis, for their valuable time.

Dedication

I dedicate this thesis to my family.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Data-intensive models are only as good as their training data. As a result, the past two decades have seen a great deal of research and industry effort toward monitoring and improving data quality. Solutions exist for deduplication, missing data imputation, and identifying and repairing incorrect data (e.g., using integrity constraints such as Conditional Functional Dependencies and Denial Constraints as rules that define data correctness) [35]. However, as data-intensive systems gain traction in new application areas, new data quality problems arise, complicating the task of identifying incorrect data.

We present a novel approach to finding data errors in one new and critical application area: forest monitoring. Forests have a significant impact on the Earth's climate and biodiversity [48, 26, 79], but they have been severely damaged by deforestation and climate change [68]. To create effective conservation policies, it is crucial to accurately map forest change (e.g., deforestation or degradation) on a global scale. Forest change maps help scientists understand the impacts of deforestation [29, 82] and are used in preparing government policies and reports [12].

Satellite remote sensing or satellite Earth Observation (EO) has made global forest monitoring possible by enabling reliable, consistent, and long-term data collection. EO technologies can be categorized into two types: passive and active remote sensing. Passive sensors detect the naturally reflected or emitted energy, while active sensors emit signals and measure the return energy from the surface. Optical (passive), LiDAR, and Radar (active) are the three main EO technologies for forest monitoring.

Optical sensors capture different wavelengths of light (e.g., Red, Green, Blue, and Near Infrared) reflected or emitted by the Earth's surface [81] (Figure 1.1). Different vegetation covers (e.g., forests and pastures) are identified based on how they absorb or reflect light

[32]. On the other hand, LiDAR is an active remote sensing technology that emits light pulses and measures the reflected energy [19]. It enables studying the 3D features of the forest, such as height and density and can penetrate the forest canopy and provide information about the layers underneath. Radar, particularly Synthetic Aperture Radar (SAR) [66], emits radio waves and processes the return energy to determine properties such as surface roughness and texture (Figure 1.3).

Optical satellite images are available at a wide range of resolutions from low (over 30 meters) to medium (30 meters) and very high (0.3 meters) resolution. Landsat is one of the most notable satellite EO projects that has been collecting data since 1972 [81]. Landsat images are widely used in monitoring global land cover and land use [9, 63], agriculture [38], forestry [28, 16], and water resources [50]. Landsat has been a primary resource in EO over the past decades, primarily due to its long-term availability, accessibility, global coverage, and reliable calibration [81].

Many forest change maps are created from Landsat images [75, 28, 16], leading to new data quality problems related to sensor limitations, image obstruction due to cloud cover and other weather conditions (Figure 1.1), and medium image resolution. Additionally, these images lack forest height information, which is useful in detecting deforestation [28]. Evaluating the accuracy of these maps is also a complex and costly task due to the limited availability of ground truth data, as collecting forest condition data through field visits is expensive and does not scale. As a result, there is no simple way of identifying errors in forest change maps.

It is worth noting that SAR is also used for deforestation detection [58, 57]. SAR offers certain benefits, such as penetrating clouds and collecting data in all weather conditions [57]. However, SAR images can be challenging to interpret, and various types of Radar noise may introduce additional uncertainties [15]. In this thesis, we focus on a state-of-the art forest change map derived from optical images.

The Annual Forest Change (AFC) dataset (Figure 1.2a) is a widely-used forest change dataset that shows annual changes in tropical moist forests (TMFs) from 1990 to 2022 [75]. AFC classifies different land cover types (e.g., forests, savannahs, and water bodies) and possible changes in forests (e.g., degradation and deforestation). AFC is reported to underestimate forest disturbance by 11.8%, equivalent to over 38 million hectares of land [75]. This number is estimated using random sampling [75]. A more efficient method of identifying errors in AFC would help scientists improve classification accuracy and uncover undetected deforestation.

To identify errors in AFC, we use GEDI [19], a recent spaceborne LiDAR dataset that provides information about the 3D structure of forests (e.g., forest height and vegetation

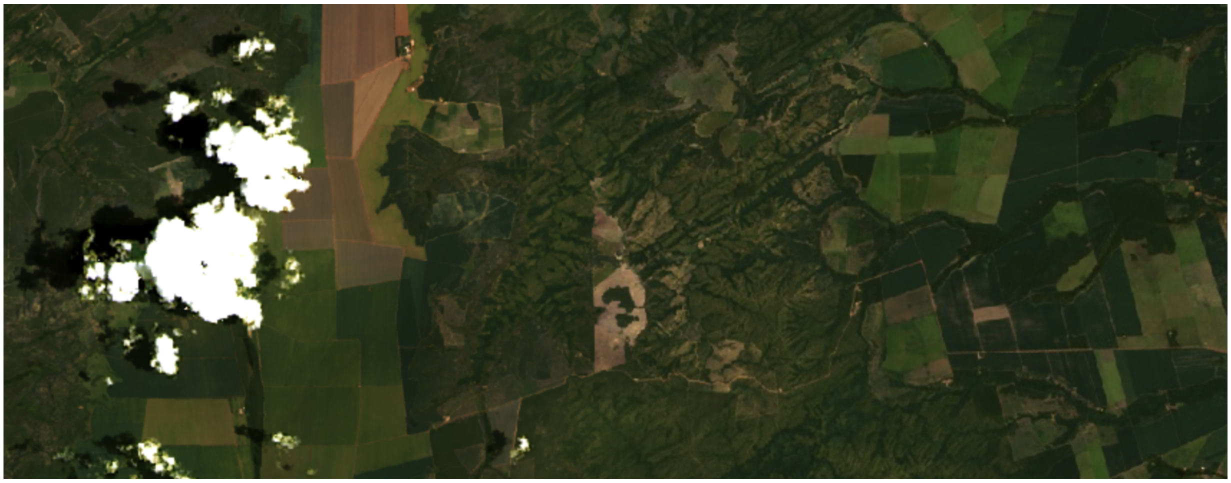
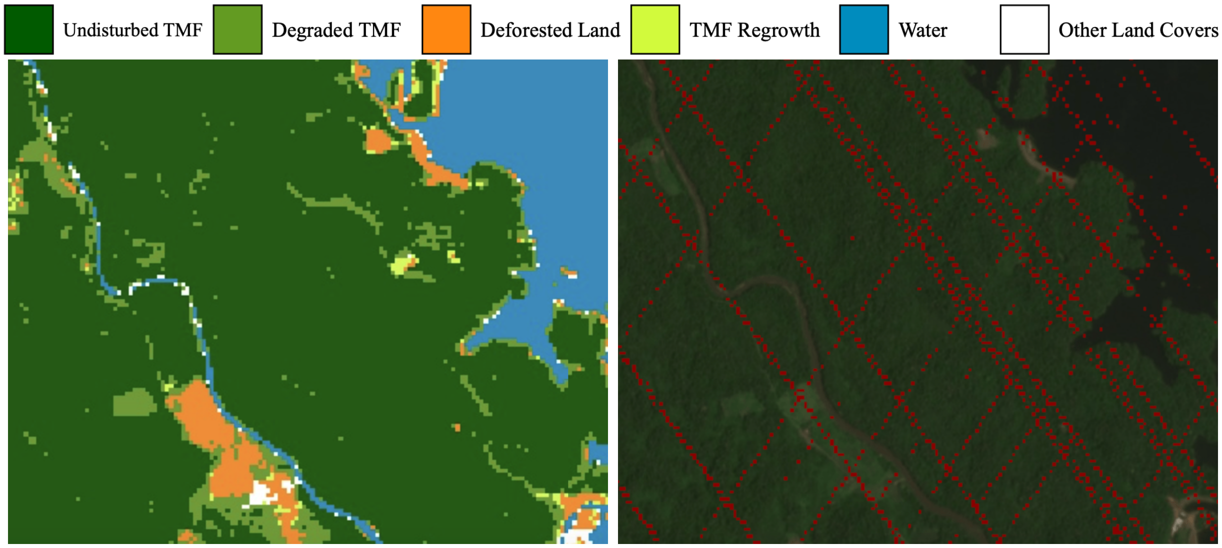


Figure 1.1: An optical satellite image. Landsat L9, January 2022. The left part of the image shows an example of missing data due to cloud cover and shadow.



(a) An illustration of the Annual Forest Change (AFC) map, December 2021.

(b) Available GEDI observations (red dots) in the same area. Background is a satellite image from Planet [49], October 2021.

Figure 1.2: The AFC map and a satellite image from the same location.

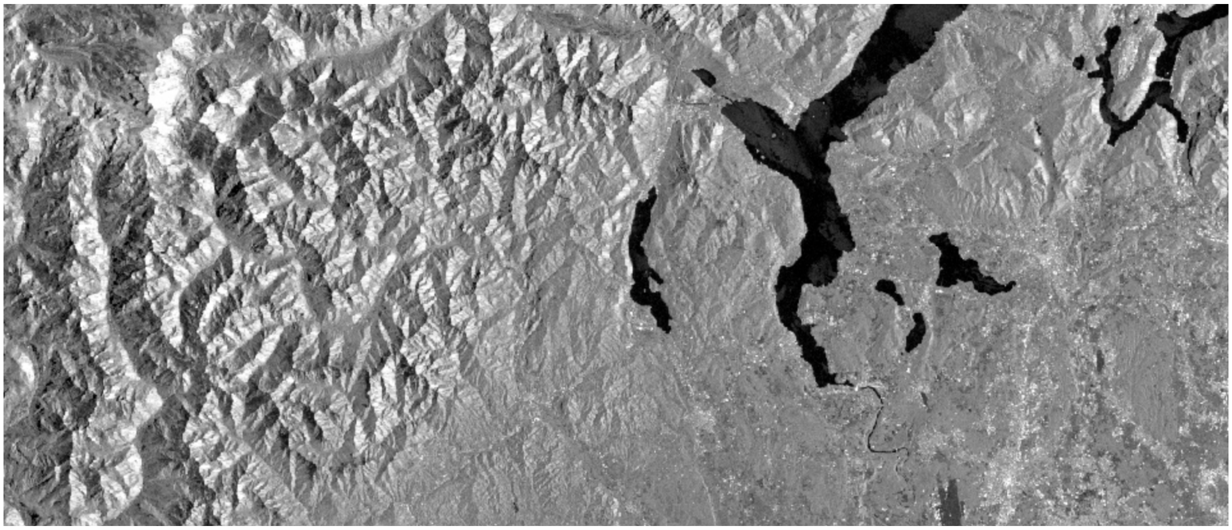


Figure 1.3: A SAR image, Sentinel-1 [66], March 2021.

density) missing from optical images. GEDI helps researchers study and understand the structure of forests, which is essential for monitoring forest health [4], biodiversity [43, 67], and carbon storage [39]. Unlike optical images, GEDI data are spatially discrete observations of the Earth's surface (Figure 1.2b). While some approaches aim to extrapolate GEDI measurements for wall-to-wall spatial coverage [53, 17, 10], our method is solely based on the GEDI observations without using additional EO datasets. Discrepancies between GEDI measurements and AFC were first discovered by Holcomb et al. [30].

We identify non-forested areas, either deforested or non-forest vegetation, that are incorrectly labeled as tropical forests in AFC. Identifying the former reveals undetected deforestation, while the latter corrects the current knowledge of the ecosystems in unreachable areas. Despite GEDI's limited spatial and historical coverage, we show that GEDI's estimates of canopy height (the height of the top of the forest) can identify parts of the forest change map that are three times more likely to contain errors than a random sample. Our approach can be used to prioritize resources for validating a forest change map and assist in more accurate detection of deforestation.

In summary, our contributions are as follows:

- We describe data quality issues associated with the two datasets: AFC [75], and GEDI [19].
- We propose and evaluate a method that leverages GEDI data to identify potential

errors in AFC.

The remainder of this thesis is organized as follows: We present an overview of AFC and its unique data quality issues (Chapter 2), followed by a discussion on GEDI and its associated challenges (Chapter 3). We then introduce our method and evaluate its effectiveness (Chapter 4, 5). Finally, we review related work in Chapter 6 and conclude in Chapter 7.

Chapter 2

Annual Forest Change Data

In this chapter, we introduce a state-of-the-art forest change dataset. We then explain how this dataset was created and point out its data quality issues. Finally, we describe why it is difficult to identify data quality issues in this dataset. Later (Chapter 4), we propose a method that targets this challenging task.

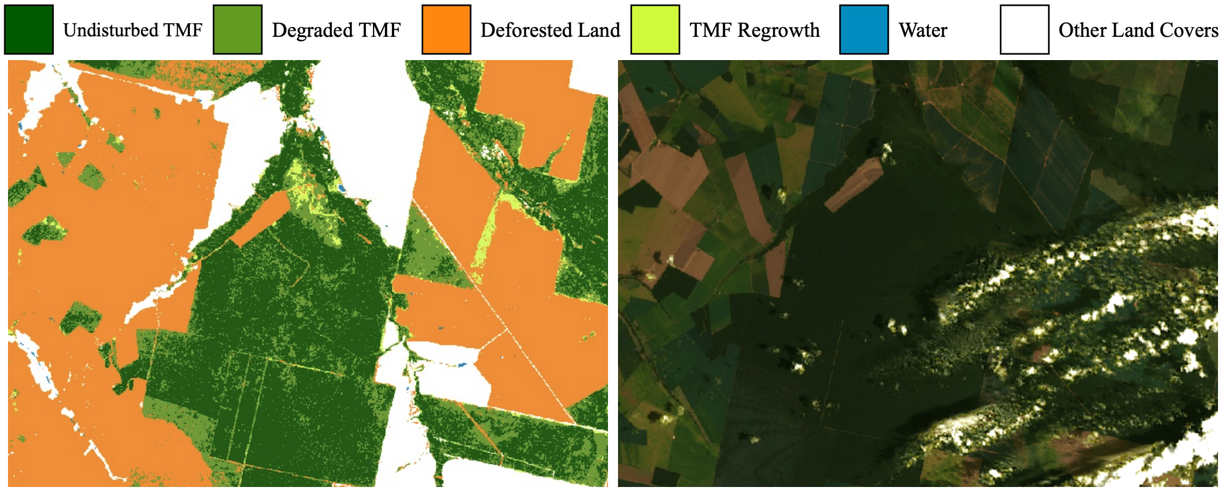
2.1 Introduction

The Annual Forest Change (AFC) dataset (Figure 2.1a) tracks annual changes in tropical moist forests (TMFs) from 1990 to 2022 [75]. It segments different land categories, such as TMFs, water bodies, and grasslands, as well as identifying changes in land cover, such as degradation and deforestation. Forest change datasets are crucial to studying and monitoring forests. These datasets help researchers and policymakers identify deforestation in different forest covers. This information helps in understanding the dynamics of forest ecosystems, global carbon sequestration [29, 82] and government policy making and reporting [12].

AFC maps the annual boundaries and status of TMFs. An AFC map is a 2D grid of pixels, each corresponding to a 30 m × 30 m (0.09 ha) area on the Earth's surface at the equator. AFC classifies each pixel into one of six categories, including *Undisturbed* TMF, *Degraded* TMF, *Deforested* land, TMF *Regrowth*, *Water*, and *Other Land Covers*. TMFs are evergreen or semi-evergreen forest, including mangroves. Other land covers include savannah, deciduous forest, agriculture, evergreen shrubland, non-vegetated cover, and afforestation. Table 2.1 provides a detailed definition of each AFC class.

Table 2.1: Definitions of AFC labels [8].

Value	**Label**	**Description**
1	Undisturbed Tropical Moist Forest	A closed evergreen or semi-evergreen forest that is undisturbed, showing no signs of deforestation or degradation in valid Landsat observations until the year of analysis. This class includes tropical moist forests (e.g., tropical broadleaf forests), mangroves, and bamboo forests.
2	Degraded Tropical Moist Forest	A tropical moist forest (undisturbed or regrowing) that has experienced temporary disturbances for a maximum of 900 consecutive days. Degradation has various causes, including selective logging, fires, hurricane, and drought.
3	Deforested Land	The permanent conversion of a tropical moist forest into non-forested land. Permanent conversion is deforestation observed for over 900 days, with no sign of regrowth. This class includes conversion to water, tree plantation, or other land covers.
4	Tropical Moist Forest Regrowth	A previously deforested tropical moist forest that is regrowing. Forest regrowth is the presence of moist forest cover for at least three years.
5	Permanent and Seasonal Water	Permanent and seasonal water.
6	Other Land Covers	An area that is not considered a tropical moist forest. This class includes deciduous forest, non-forest covers (e.g., savannah, agriculture, evergreen shrubland), non-vegetated cover (e.g., mining), and afforestation.



(a) An illustration of the Annual Forest Change (AFC) map, December 2021.

(b) A Satellite image. Landsat L9, January 2022. The bottom right corner of the image shows an example of missing data due to cloud coverage.

Figure 2.1: The AFC map and a satellite image from the same location.

AFC maps are derived from optical satellite imagery of the Landsat program [81, 80] (Figure 2.1b). The Landsat mission uses a series of satellites that capture images of the Earth's surface from space. These satellites are designed to study and monitor changes in the Earth's land cover. Landsat images are taken by special cameras or sensors onboard the satellites that capture different wavelengths of light, including visible light (Red, Green, and Blue), Near Infrared, and other wavelengths. Different land covers are recognized by how they reflect or emit light [32]. Landsat has been one of the primary data sources in global forest monitoring since 1972.

Nine Landsat satellites (L1-L9) have been launched during the 50 years of the Landsat mission, three of which are active today. Since 1982, Landsat satellites have been capturing images of the Earth's surface from 705 kilometers above at 30 meters resolution, revisiting each location every 16 days. Landsat achieves an 8-day revisit time during most of its mission by having two active satellites in orbit [80].

The AFC dataset is based on per-pixel classifications of Landsat images. Each pixel is classified using expert rules as either potential moist forest, potential non-forest, or invalid (cloud, shadows, noise). Each pixel is then assigned a final class based on the changes in valid observations over time. For instance, a deforested pixel appears as a potential moist forest at first and then changes to a potential non-forest later. AFC is reported to be 91.4% accurate but underestimates forest disturbance by 11.8% [75]. This corresponds to over 38 million hectares of land [75], which is a significant area.

2.2 Data Quality Challenges

Forest maps, including AFC, face several common challenges. Many of these challenges originate from the data source, Landsat. Therefore, datasets created from Landsat images can suffer from similar limitations.

- **Missing Data:** There could be gaps in satellite observations for several reasons, including cloud cover, cloud shadow, and other atmospheric conditions (Figure 2.1b), failures in sensors or other instruments (Figure 2.3) [71], and intentional pauses or stops in data collection. The most significant data loss in the history of Landsat occurred on May 31, 2003, when a technical failure caused a permanent 22% data gap in the Landsat 7 images (Figure 2.3c).

- **Noisy Data:** Satellite imagery is prone to sensor noise, miscalibration, and atmospheric noise, which affects the quality of forest maps. Although some noise can be corrected, the correction process can degrade the overall image quality [69]. In Landsat images, data noise can appear as repeating noise patterns or inaccuracies in specific areas (Figure 2.4). Sensor saturation is an example of noise, where bright light exceeds the 8-bit value capacity of the sensor and is clipped to 255 [73]. Sensor oversaturation is a related issue, which occurs when an object is significantly brighter than the sensor can handle, causing temporary sensor malfunction (Figure 2.4b).

- **Spatial and Temporal Resolution:** The resolution of a forest map is determined by the resolution of the source data, which can impact the accuracy and level of detail provided by the map. For instance, AFC cannot tell the precise location of disruptions or changes smaller than 0.09 hectares. A 30-meter resolution is known as medium resolution [81].

- **Spectral Mixing:** Satellite images often have mixed pixels containing different land cover types (e.g., half forest and half deforested) [62]. These mixed pixels can introduce uncertainty in land cover classification. This issue occurs frequently in complex vegetation covers or at the boundaries between different land cover types.

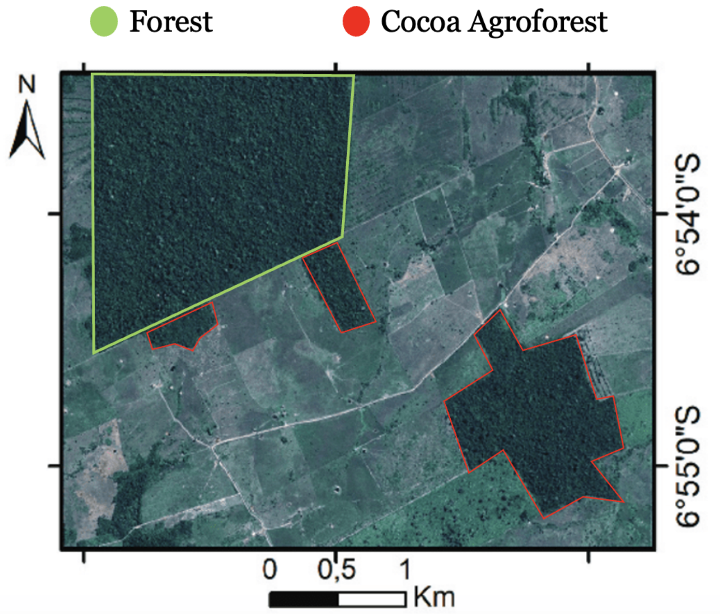
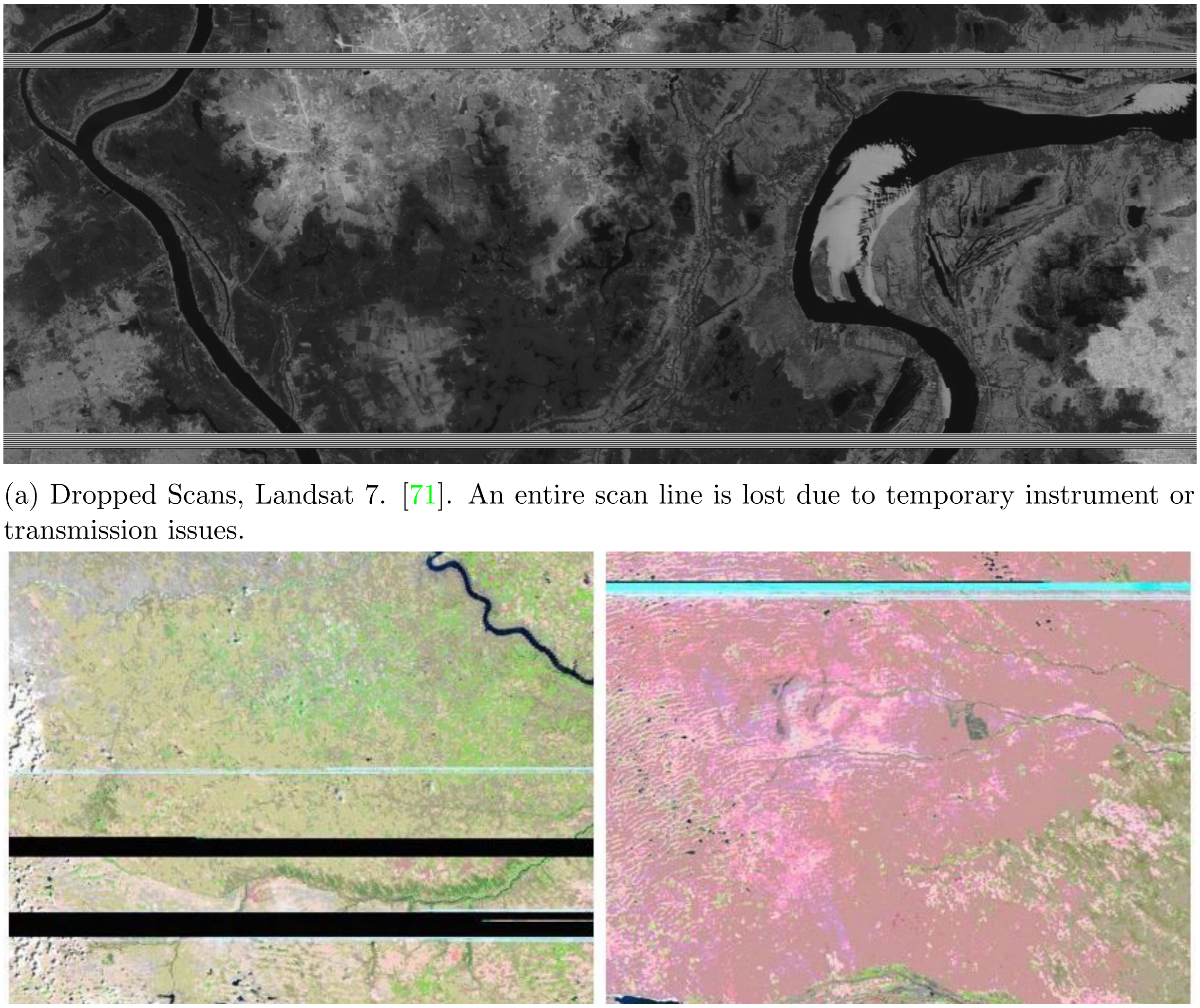


Figure 2.2: Spectral confusion [3]. The cocoa agroforest looks identical to a forest in optical satellite imagery. Areas with green and red outlines are forest and cocoa agroforest, respectively.

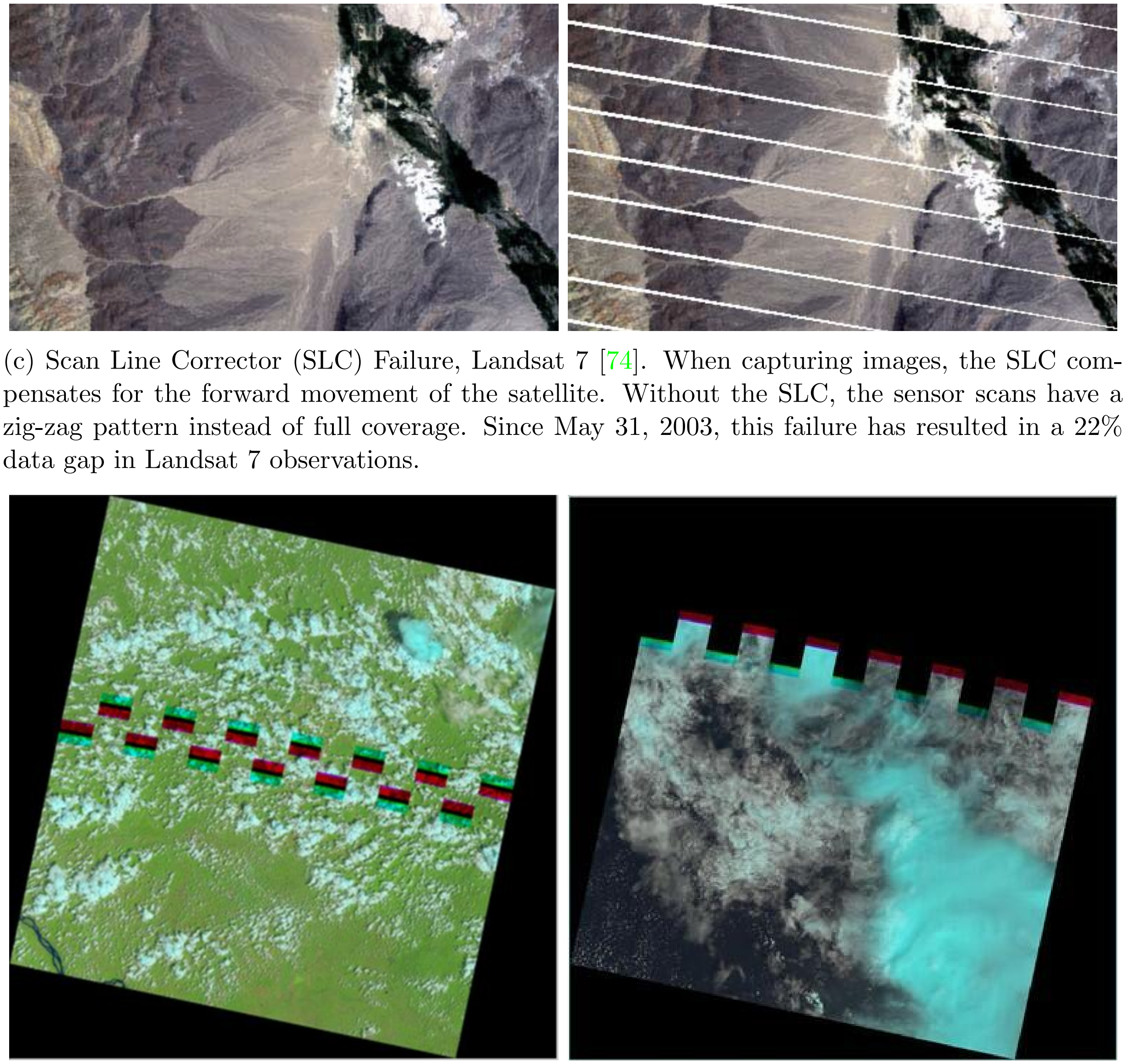
- **Spectral Confusion:** This occurs when different types of land cover have similar appearance when viewed from space. For instance, Figure 2.2 shows how a cocoa agroforest looks similar to a forest in optical satellite imagery [3].
- **Lack of 3D Information:** Optical satellite images lack 3D information such as forest height, limiting their ability to distinguish between some land cover types. For example, height information can accurately distinguish forests from shrublands. Additionally, optical images capture the upper layer of the forest (the canopy) and may not reveal degradation in the lower layers that have not affected the canopy [23].
- **Limited Ground Truth:** Collecting data by visiting a forest ranges from expensive to impossible (for remote and inaccessible locations). As a result, experts rely on remote sensing data to create a reference dataset. Additionally, global maps cover billions of hectares of land, and gathering ground truth for every pixel is impossible. Therefore, only a limited number of forest map pixels are validated.



(a) Dropped Scans, Landsat 7. [71]. An entire scan line is lost due to temporary instrument or transmission issues.

(b) Sun Glint, Landsat 5 [71]. Sunlight reflects off the body of the satellite and causes data loss in transmission. This is an anomaly in Landsat 5 that causes periodic and predictable data loss due to sunlight interference.

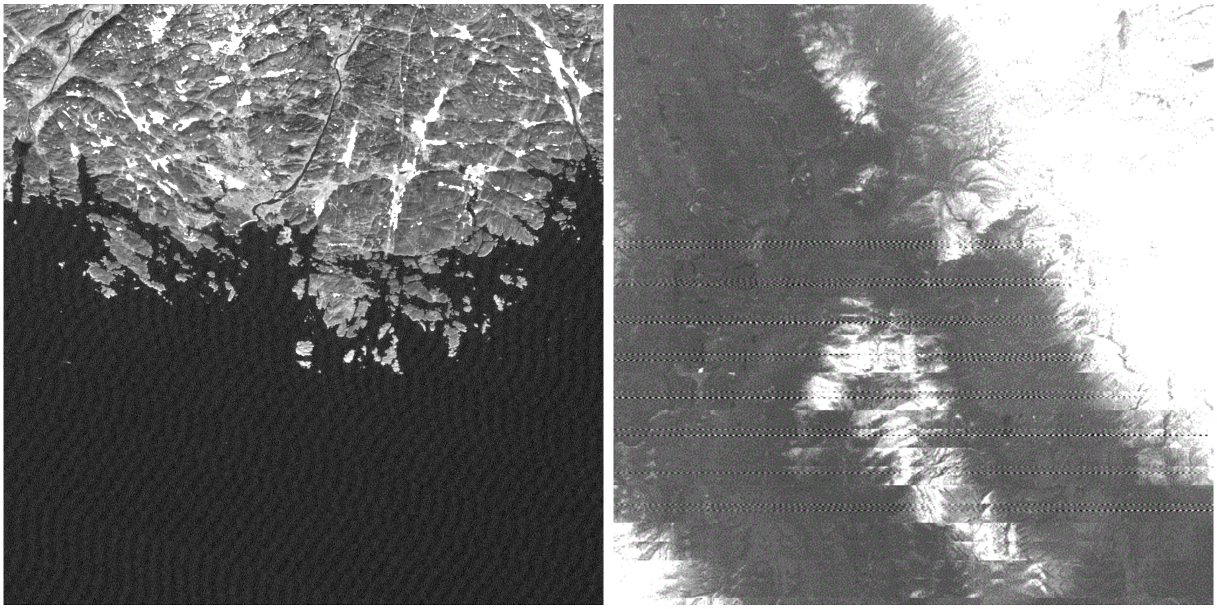
Figure 2.3: Instances of missing data due to instrument, sensor, or transmission failures.



(c) Scan Line Corrector (SLC) Failure, Landsat 7 [74]. When capturing images, the SLC compensates for the forward movement of the satellite. Without the SLC, the sensor scans have a zig-zag pattern instead of full coverage. Since May 31, 2003, this failure has resulted in a 22% data gap in Landsat 7 observations.

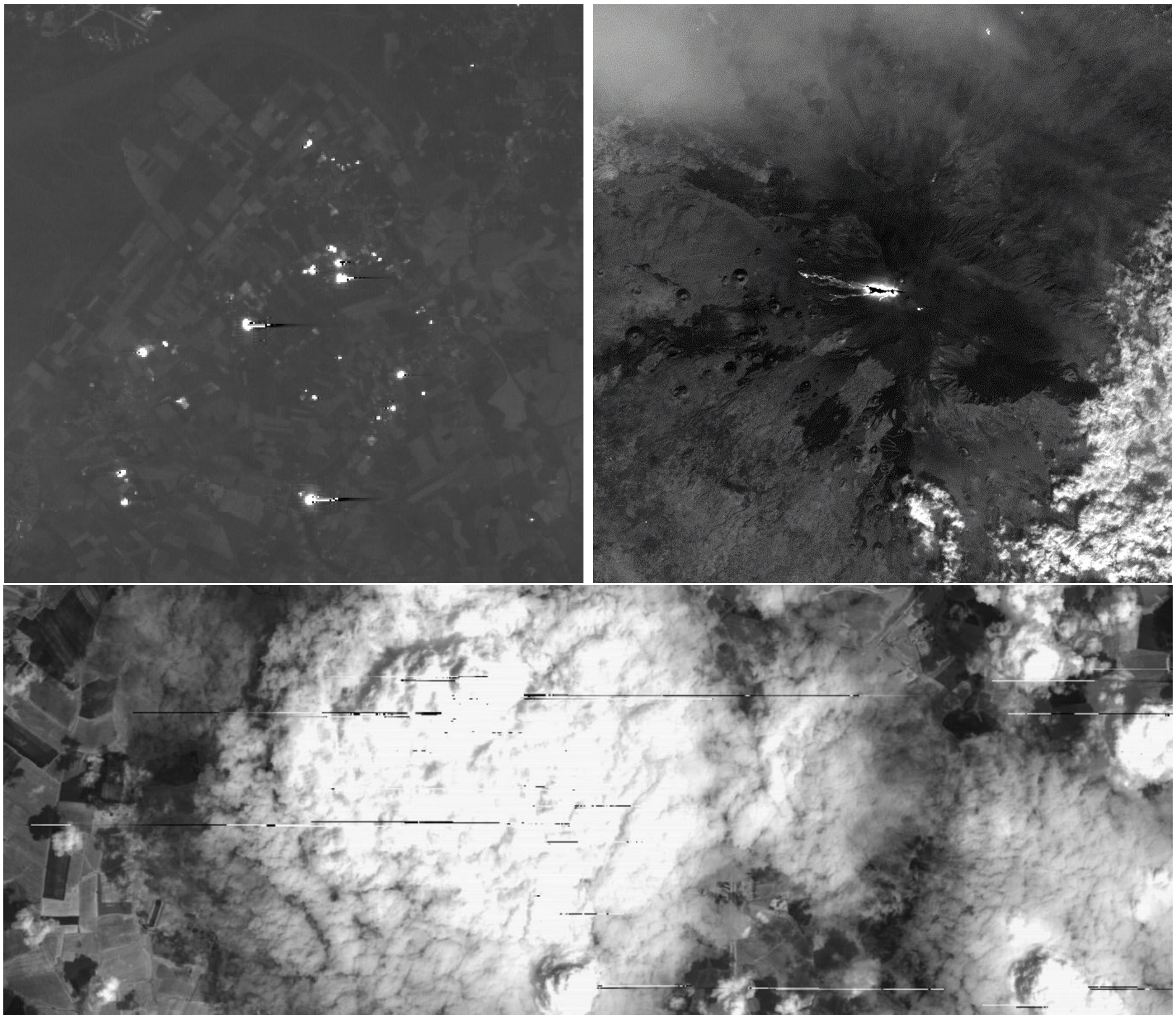
(d) Solid State Recorder (SSR) bad block issue, Landsat 9 [72]. Data is lost due to memory blocks issues. 187 frames are lost in the center of the left figure, and over 2,000 frames are lost in the upper part of the right figure.

Figure 2.3: Instances of missing data due to instrument, sensor, or transmission failures.



(a) Coherent noise, Landsat 7 [69, 70]. This noise appears as a repeating pattern, most visible over uniform dark regions. A coherent noise storm [70] (right) is a sign of a sudden electrical change, possibly due to a serious failure. The bottom frames are also misaligned.

Figure 2.4: Instances of noisy images due to instrument or sensor failures.



(b) Sensor Oversaturation, Landsat 9 [73]. Oversaturation happens when the sensors view an object much brighter than the sensor can tolerate, causing a temporary malfunction in the sensor. Oversaturation is common over reflective surfaces (top left image), fires and active volcanoes (top right image), and bright clouds (bottom image).

Figure 2.4: Instances of noisy images due to instrument or sensor failures.

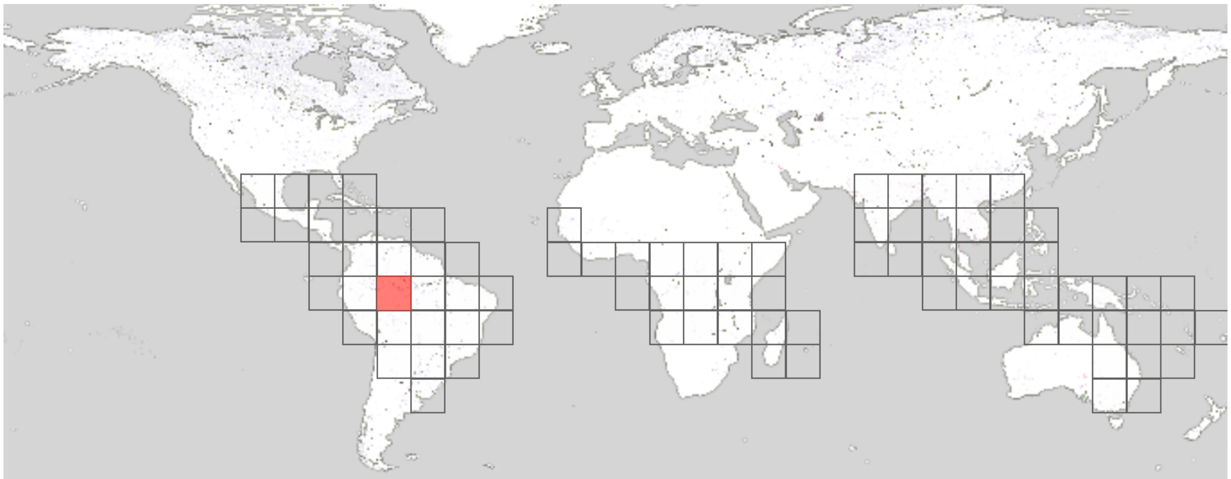


Figure 2.5: AFC maps can be downloaded in several 10° × 10° tiles [7].

Identifying errors in the AFC map is not trivial, as no data or forest map can claim to be 100% accurate. A random sample of the dataset contains only around 10% errors [75]. A more efficient method of identifying errors would help scientists improve classification accuracy. In Chapter 4, we introduce a method that directs experts' attention toward a subset of samples that are more likely to contain errors. To achieve this goal, we use a new forest height dataset, described in Chapter 3, with its own unique data quality issues and challenges.

2.3 Data Acccess

AFC maps are publicly available on the European Commission Joint Research Center website[1]. Each map can be downloaded as several 10° × 10° tiles stored in GeoTIFF format (Figure 2.5). A GeoTIFF image is a matrix where each cell (each pixel) has a geolocation tag. These files can be opened with Python packages such as *Rasterio*. Additionally, AFC maps are available on the Google Earth Engine[2], a cloud platform that hosts petabytes of Earth Observation data and offers computation power to run geospatial analysis [27].

[1] https://forobs.jrc.ec.europa.eu/TMF/

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

Chapter 3

Vegetation Height Data

In this chapter, we introduce GEDI, a new data source for forest monitoring, and provide an overview of its data products. We then discuss GEDI's data quality challenges and limitations and explore its potential for identifying data quality issues in AFC. Furthermore, we discuss various quality metrics available with GEDI products that can be used to filter noisy data to some extent.

3.1 Introduction

Global Ecosystem Dynamics Investigation (GEDI) is a LiDAR (Light Detection and Ranging) instrument that collects data about Earth's forests from space [19]. LiDAR emits laser beams and measures the time it takes for the light to return to the sensor. GEDI LiDAR is designed to penetrate the forest cover, allowing scientists to study the 3D structure of forests. This information helps researchers study the structure of forests, which is essential for monitoring forest health [4], biodiversity [43, 67], and carbon storage [39]. GEDI operated on the International Space Station (ISS) from April 2019 to March 2023. Figure 3.1 illustrates a GEDI return waveform.

GEDI has three lasers, one of which is split into two beams, emitting four beams in total. Each of these four beams is shifted every other shot to create eight tracks on the ground. These tracks are separated by 600 meters, with shots 60 meters apart along the tracks [19]. Figure 3.2 illustrates this pattern. A GEDI shot correspond to a fragment of the Earth's surface with a 25 m diameter called a "footprint" (Figure 3.1).

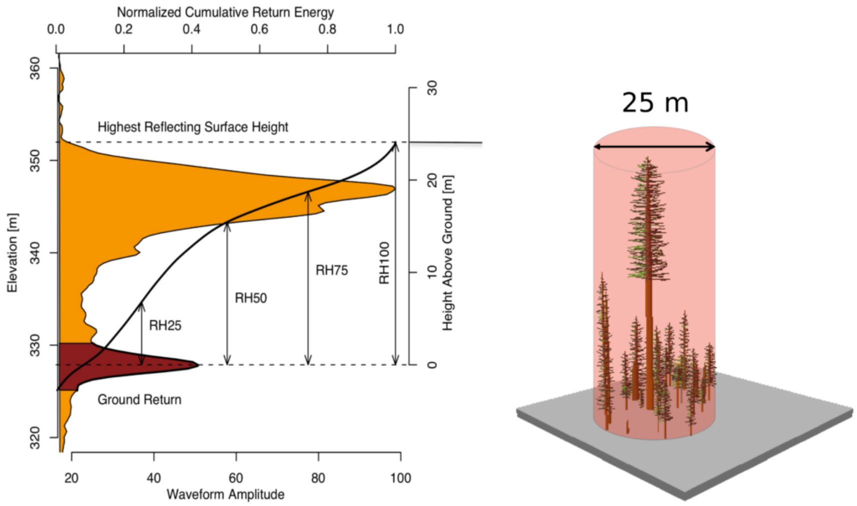


Figure 3.1: GEDI return waveform [19, 25]. The waveform (left) captures the reflected energy at different elevations from the 25 meters diameter footprint (right).

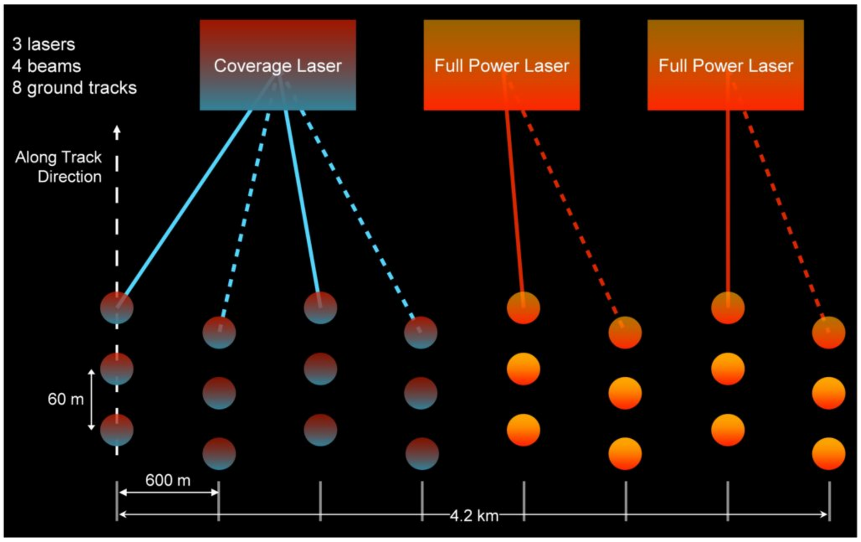


Figure 3.2: GEDI beam pattern [19, 24]. GEDI has three lasers that shoot four beams and create eight tracks on the ground.

3.2 Data Products

Raw GEDI waveforms are processed into higher-level data products that describe the 3D features of forests. GEDI data products include footprint and gridded data. These products are assigned different levels that show the amount of processing on the data:

Level 1: LiDAR waveforms available at footprint level.

Level 2A: Measurements of ground elevation and relative height (RH) [20]. RH is the height above ground at which a certain quantile of cumulative energy was returned (Figure 3.1), and the *RH95* (95% quantile) has been shown to estimate canopy height (height of the top of the forest) [53].

Level 2B: Measurements of the distribution and density of vegetation from top of the canopy to the ground. These measurements include Canopy Cover Fraction (proportion of the ground covered by the canopy projection) and Plant Area Index (total area of canopy elements including leaf, branch, and trunk in the unit ground area) [65].

Level 3A: Level 2 data as 1 km × 1 km grids.

Level 4: Footprint and gridded above-ground biomass density (AGBD). AGBD is the amount of biomass (organic matter) in the above-ground portion of a forest per unit of area. GEDI is optimized to estimate AGBD accurately [18, 22], which is essential for analyzing the amount of carbon stored in forests [39, 40].

Table 3.1 gives an overview of GEDI products in various levels.

Table 3.1: A summary of GEDI data products.

Level	Description	Resolution
L1A	Raw Waveform	25 m diameter
L1B	Geolocated Waveform	25 m diameter
L2A	Elevation and Height Metrics	25 m diameter
L2B	Canopy Cover and Vertical Profile Metrics	25 m diameter
L3	Gridded Level 2 Metrics	1 km grid
L4A	Footprint Above Ground Biomass	25 m diameter
L4B	Gridded Above Ground Biomas	1 km grid

GEDI was calibrated and validated using a ground truth dataset in which the evergreen broadleaf forests of South America were well represented [19]. Studies show that GEDI can accurately estimate ground elevation, $RH95$, and $RH100$ with RMSE of 1.38 m, 2.08 m, and 2.62 m, respectively [78]. Therefore, RH95 is a better estimate for canopy height than RH100 [53].

3.3 Data Quality Challenges

Similar to other data collected from space, GEDI data has the following data quality issues:

- **Noisy Data:** As a laser-based technology, GEDI is sensitive to atmospheric conditions, including cloud cover and moisture. Sensor noise and miscalibration also contribute to errors in the data.
- **Spatial and Temporal Resolution:** GEDI footprints cover a limited portion of the Earth (around 4% in 2 years of operation [19]), and the gridded data has a relatively coarse resolution of 1 km. Additionally, operating from 2019 to 2023, GEDI does not offer extensive historical information. Finally, there are no guaranteed revisits of the same location, making it difficult to monitor for changes.
- **Geolocation Error:** Slight geolocation uncertainties (0 m mean, 10 m standard deviation) exist in the reported coordinates. This uncertainty can significantly affect RH metrics in mixed canopies and forest edges [59].
- **Terrain:** Sloped or complex terrain introduces additional errors in the GEDI data [1, 60, 47]. Such characteristics impact GEDI's ability to detect ground returns and estimate canopy height accurately.

GEDI's spatiotemporal limitations prevent scientists from creating high-resolution forest change maps based solely on GEDI data. Nevertheless, GEDI can help address the *lack of 3D information*, *spectral confusion*, and *limited ground truth* problems in AFC. GEDI offers 3D information for remote unreachable forests. Therefore, GEDI data (like canopy height) can help distinguish forest covers that may look the same in optical satellite imagery. While the spatial limitations of GEDI prevent us from evaluating the entire AFC dataset, we present a novel method to identify data quality issues in AFC more efficiently than random sampling while accounting for geolocation error and noise in GEDI data (Chapter 4).

3.4 Quality Filtering

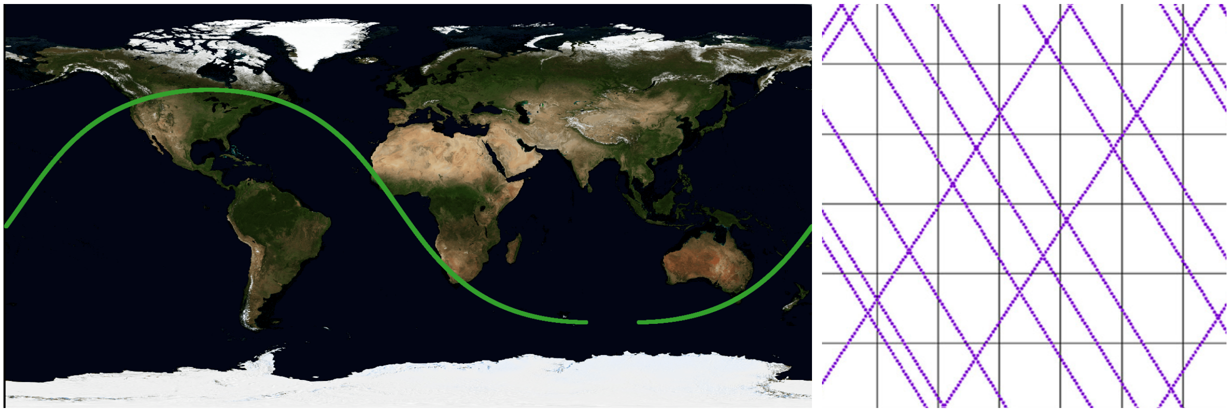
GEDI data products include several quality metrics that can be used to filter noisy data:

- **Beam Sensitivity:** GEDI measurements are most trusted when the beam penetrates the canopy and accurately detects the ground. However, the strength of GEDI return signals depends on atmospheric conditions, impacting their ability to penetrate canopies. Beam sensitivity is a metric to identify waveforms where an accurate ground level is not detected. Beams with low sensitivity cannot penetrate dense vegetation. For high-quality data, it is recommended to use shots with sensitivity greater than 0.98 in tropical evergreen broadleaf forets and greater than 0.95 elsewhere.
- **Degrade Flag:** This is a metric to exclude shots with potential geolocation errors. A non-zero value indicates that the shot was taken during a degraded period when one or more of the star trackers (instruments to determine the orientation of a satellite) were not functioning properly.
- **Quality Flag:** A general quality flag that indicates when multiple quality conditions are met. Each level of GEDI products has its specific flag. However, this flag alone does not mark all data quality issues. For example, GEDI Level 2B measurements may still have negative values even after filtering. To address this issue, we can filter abnormal values when working with higher-level GEDI products.
- **Additional Quality Filters:** Additional filters can be applied to the data depending on the specific use case. For instance, when studying forest structure, we can ensure that the shots are taken during the leaf-on period. We can also exclude shots that fall within seasonal or permanent water bodies.

We use highly-filtered GEDI shots from Burns et al. [6]. However, there may be residual noise and geolocation errors and noise within the data. Therefore, we propose a method in Chapter 4 to find data quality issues in AFC.

3.5 Data Extraction

GEDI data can be downloaded for free from the Land Processes Distributed Active Archive Center (DAAC) or the Oak Ridge National Laboratories DAAC. Located onboard the International Space Station (ISS), GEDI collects data within a latitude range of 51.6° S



(a) (b)

Figure 3.3: GEDI tracks. (a) Visualization of a single GEDI orbit [21]. (b) The simulated pattern of GEDI tracks from multiple orbits in a small region near the equator [19, 24].

to 51.6° N, leaving the same track as the ISS in each orbit (Figure 3.3). The data from each orbit is divided into four granules. The GEDI Finder Service [61] can be used to find all granules overlapping the study area, including those with a limited number of shots within the area. After downloading, the data needs to be extracted from the compressed format and ingested into a spatial dataset such as PostGIS [13], which is the spatial extension of Postgres. Creating a spatial index allows for the parallel loading of all the available GEDI shots in a region. GEDI products are also available on Google Earth Engine.

Chapter 4

Finding Potential Errors in AFC

Forests commonly consist of tall green trees: the formal definition of a forest requires canopies to be at least 5 m tall [51]. Therefore, areas with shorter canopies are more likely to be instances of *deforestation* or *other land covers* such as shrublands. In this chapter, we propose a method to find potential errors in AFC where a non-forest cover is labeled as *undisturbed* TMF. We then discuss the parameters and their associated tradeoffs. In Chapter 5, we apply this method to find potential errors in the AFC map of 2021 and evaluate the results.

4.1 Approach

We define areas with an *undisturbed* label and short canopy height as conflicts or outliers. We suggest that these conflicts can be more effective in identifying errors in the AFC dataset than randomly selected samples. Note that these conflicts represent potential errors that could arise from noise in either the GEDI or AFC data. Thus, several challenges need to be addressed:

- Integrating the two datasets, the AFC map and GEDI footprint data, while accounting for geolocation errors.
- Accounting for the noise in the GEDI data and finding samples that are more trusted.
- Prioritizing some outliers when dealing with thousands of conflicts, as manually examining all of them is too time-consuming.

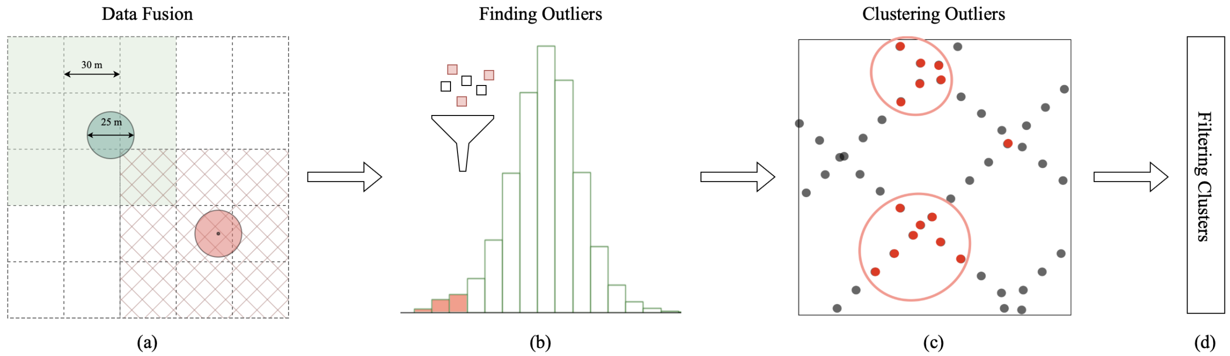


Figure 4.1: An illustration of dataflow. (a) *Data Fusion.* The grid is the AFC map, and the circles are GEDI shots. The nearest 3 × 3 windows is highlighted for each shot. (b) *Finding Outliers.* Samples with $RH95 < h$ are selected. (c) *Clustering Outliers.* Nearby outliers are clustered. (d) *Filtering Clusters.* Smaller clusters are filtered to increase reliability.

We propose a four-step process to utilize GEDI canopy heights (RH95) to identify forests labeled as undisturbed but having conflicting (short) canopy heights. Figure 4.1 shows the dataflow.

4.1.1 Step 1: Data Fusion.

We identify the nine nearest AFC pixels to each GEDI shot. These pixels form a 3 × 3 window on the AFC map, with the center pixel containing the GEDI shot center (Figure 4.1). We only consider GEDI shots within homogeneous windows. This accounts for potential geolocation errors in the GEDI shot: even if the shot has some geolocation error, it still falls within an area classified as *undisturbed* TMF in AFC.

4.1.2 Step 2: Finding Outliers.

We select GEDI shots with $RH95 < h$, where h is a tuneable parameter. These shots are *undisturbed* TMFs with an abnormally short canopy. The parameter h can be selected based on expert knowledge or the RH95 distribution. Note that these anomalies are found in the left tail of the RH95 distribution.

4.1.3 Step 3: Clustering Outliers.

We merge nearby outliers into clusters using hierarchical clustering for two reasons:

- Reducing data noise. Several nearby conflicting observations are more trusted than a single outlier.
- Conflicts occurring close together can belong to the same area and cover type, corresponding to spatially correlated errors [52]. For instance, two consecutive GEDI shots are only 60 m apart, and both may be from a grassland misclassified in AFC. Using clustering, we avoid reporting these points separately.

Hierarchical clustering, also known as agglomerative clustering, is a bottom-up approach where each data point initially represents a separate cluster. The algorithm merges the two nearest clusters in each step, creating a new one. This process continues until a stopping criterion is met (e.g., all the data points are in one cluster, all distances are greater than the threshold, number of clusters reaches the limit).

We use Hierarchical clustering with a distance threshold since it does not require a predefined number of clusters. This approach has two parameters: linkage and distance threshold. Linkage determines how the distance between two clusters is calculated; e.g., single-linkage uses the minimum distance between clusters. The distance threshold determines if clusters should be combined, merging only those closer than the threshold.

4.1.4 Step 4: Filtering Clusters.

Clusters with few conflicts are less likely to represent areas with AFC errors than ones with many conflicts. Hence, we prioritize clusters larger than a certain threshold, c. Additionally, clusters containing GEDI shots from multiple satellite orbits are more reliable and less susceptible to systematic errors. This is because consecutive shots within a single orbit could all be incorrect due to atmospheric conditions or sensor issues. However, there are few clusters with this redundancy due to GEDI coverage limitations.

4.2 Parameters

There are three parameters in this method: height threshold (h), clustering distance (d), and minimum cluster size (c). Each parameter directly impacts the tradeoff between precision and recall:

- **Height Threshold (h):** A lower threshold reduces the number of outliers, which can reduce false positives but may affect recall.
- **Clustering Distance Threshold (d)**: A lower threshold leads to smaller clusters, and several small nearby clusters may represent the same error. A higher distance threshold can merge unrelated clusters or create clusters of multiple noisy samples.
- **Minimum Cluster Size (c):** Although small clusters are more likely to be false positives, choosing a large c affects the recall of small-scale errors.

It is essential to tune these parameters with respect to each other. For example, a higher clustering distance forms larger clusters and may require increasing the minimum cluster size accordingly. In the next chapter, we evaluate the results for one set of carefully selected parameters.

Chapter 5

Evaluation

In this chapter, we evaluate our approach and present the results of two evaluation strategies. We first justify our selection of the study region and parameters. We then introduce two sources for acquiring ground truth data. Finally, we present the results, followed by a discussion about potential errors in the AFC map.

5.1 Overview

We used our method to find data quality issues in the 2021 AFC map of the Brazilian Amazon region. We used RH95 from quality-filtered GEDI shots collected during the second half of 2021. In this section, we describe two evaluation methods. The first uses a highly-validated forest cover map as ground truth, while the second is based on visual interpretation of high-resolution satellite images.

5.1.1 Study Region

We focus on the Brazilian Amazon rainforest, which plays a vital role in global climate stability, and is home to various unique plant and animal species, many of which are found nowhere else on Earth. Additionally, the availability of numerous forest maps and freely accessible satellite data makes the Brazilian Amazon an ideal region for our studies.

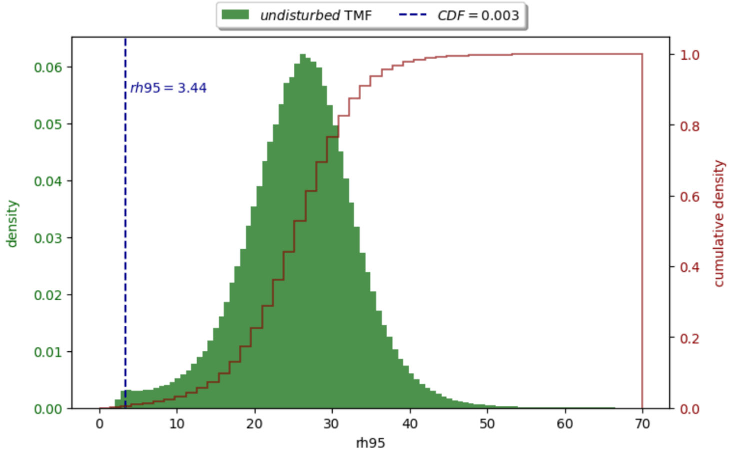


Figure 5.1: Distribution of *undisturbed* TMF height in the second half of 2021.

5.1.2 Parameters

As mentioned in Chapter 4, parameters must be selected with respect to each other, to maintain a balance between precision and recall. Based on empirical fine-tuning, we selected $h = 3.44$ meters to mark 0.3% of the GEDI shots in *undisturbed* TMFs as outliers (Figure 5.1). A lower threshold (e.g., 2 meters) eliminated some evident AFC errors, whereas a higher threshold (e.g., 4 meters) included many shots that were ambiguous as to whether they were AFC errors. We apply single-linkage hierarchical clustering with a distance of $d = 700$ meters to group outliers that are from the same GEDI orbit. We also filter clusters with fewer than 9 shots ($c = 8$).

5.1.3 MapBiomas Evaluation

MapBiomas [63, 55] is an annual dataset of Brazil's land cover maps from 1985 to 2021 at a 30-metre resolution. It uses a hierarchical classification system with four levels to categorize land covers. At Level 1, land covers are classified into six categories: forest, non-forest, farming, non-vegetated, water, and not observed. Level 2 expands Level 1 classes into 22 subcategories [41, 42]. MapBiomas is created from Landsat images, primarily using a Random Forest classifier, and it is validated annually on over 75,000 samples. Level 1 and 2 classification error is estimated to be 7.5% and 9.3%, respectively [41].

In this analysis, if an outlier shot is labeled as *undisturbed* TMF in the AFC map but classified as non-forest in MapBiomas, then we consider MapBiomas to be correct, meaning that the outlier is an error in the AFC map. We report two validation metrics: (1) the percentage of outliers with non-forest Mapbiomas labels and (2) the percentage of clusters with at least one such outlier.

5.1.4 Visual Interpretation

After finding outlier clusters, we randomly select one GEDI shot per cluster. Then, we determine if this represents an AFC error by analyzing the 3 × 3 surrounding AFC pixels in a cloud-free image. We use higher-resolution satellite images with approximately 4 meters per pixel resolution from the Planet NICFI data program [49, 46]. Specifically, we used the last cloud-free Planet images of 2021. Each cluster is assigned one of three validation labels: *Ambiguous* (if no cloud-free observations are available or if it is unclear whether the area is an AFC error), *AFC Error*, or *False Positive* (if the pixels are correctly classified in AFC). Analyzing 3 × 3 windows of the map is similar to AFC's validation method [75].

5.2 Results and Discussion

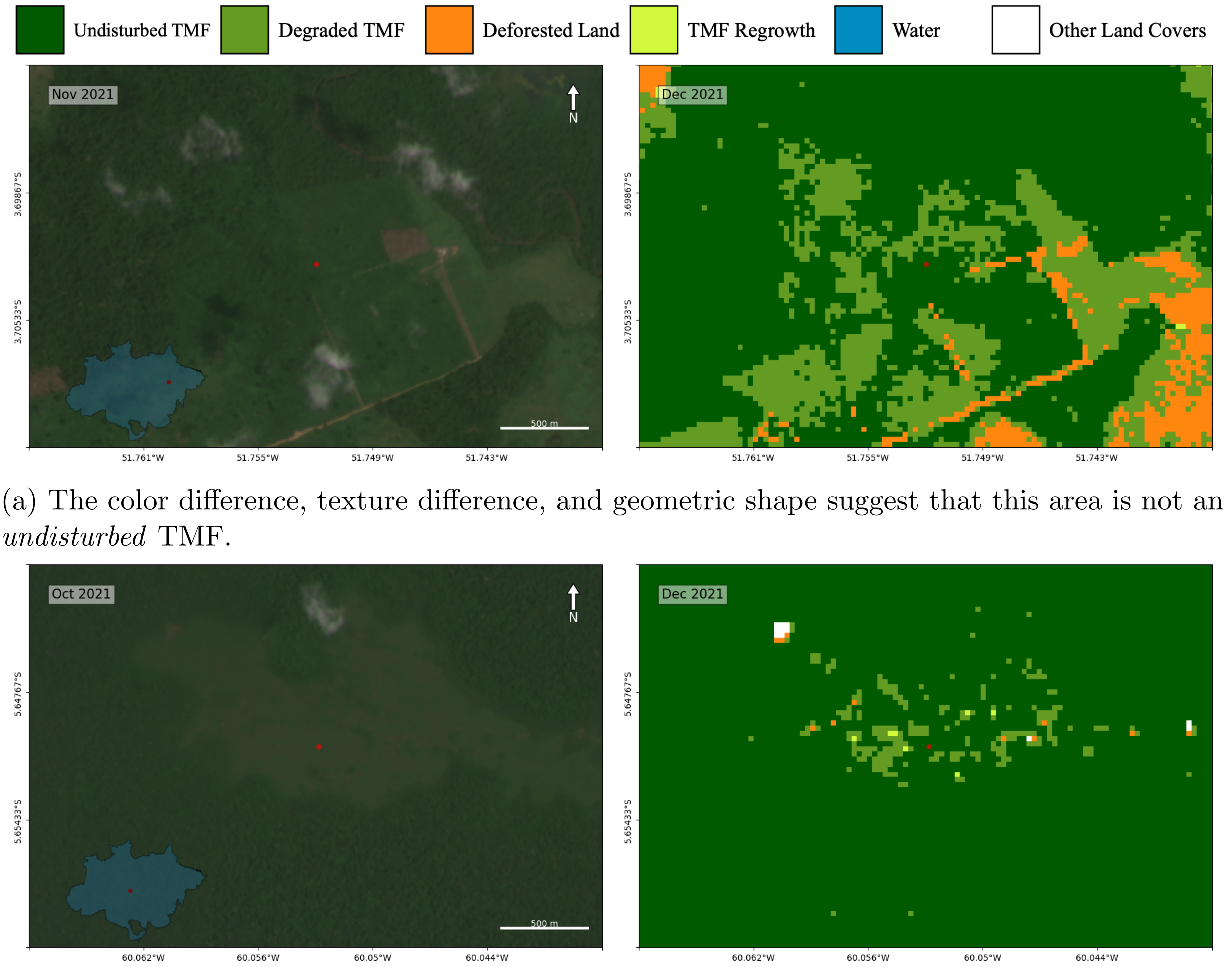
We identified 23,306 conflicts (i.e., marked *undisturbed* forest in AFC with $RH95 < 3.44$ m) in Step 2. After filtering clusters in Step 4, 5,740 samples remain, of which 1.88% are labeled non-forest in MapBiomas. This gives 240 clusters, 12.08% of which have at least one outlier with a non-forest MapBiomas label. Since manual evaluation is time-consuming, we evaluate 100 random clusters out of the 240 clusters using Planet images. Out of the 100 clusters, 33 were found to be AFC errors, 63 were Ambiguous, and 4 were False Positives (see Figure 5.2-5.4 for examples). Assuming that all Ambiguous cases are False Positives, the precision of our method is at least 33%, which is almost three times greater than the precision of random sampling reported by [75].

Visual interpretation revealed cases where both AFC and MapBiomas were inaccurate. This can be because of the MapBiomas limitations due to the lack of 3D information in Landsat images. While MapBiomas has the advantage of evaluating every pixel in the AFC, GEDI, despite its limited coverage, uncovers errors that MapBiomas may not detect. Additionally, there are some vegetation types that should be classified as non-forest covers in AFC but are considered forests in MapBiomas. For instance, MapBiomas assigns wooded savannah and tropical evergreens to the same class, while AFC refers to the former as *other land covers*. Therefore, per-pixel evaluation cannot identify AFC errors in wooded savannahs, but using canopy height can.

It is illuminative to study where AFC made errors. We found many outlier clusters in the Brazilian Amazon's northwest region, with a vegetation cover known as *campinarana* that can be difficult to distinguish in satellite images [14, 56]. This region is remote and challenging to access, making it difficult to obtain field data. Different types of campinarana are forest, wood, shrub, and grass [14, 34]. Forest campinarana appears as islands within open areas. Wood campinarana is a non-forest type with sparse trees and a canopy height of up to 4 meters, while shrub campinarana is shorter and does not exceed 2 meters in height [34]. Therefore, using GEDI canopy height helps distinguish forested and non-forested campinaranas as canopy height in the non-forest classes does not exceed 4 meters [34].

Some *False Positives* were located near shores and water that could potentially be covered by mangroves (Figure 5.4b). Mangroves have a distinct structure that differs from evergreen or semi-evergreen forests. However, all three types are categorized as TMFs in AFC. Excluding such areas from the analysis could improve precision.

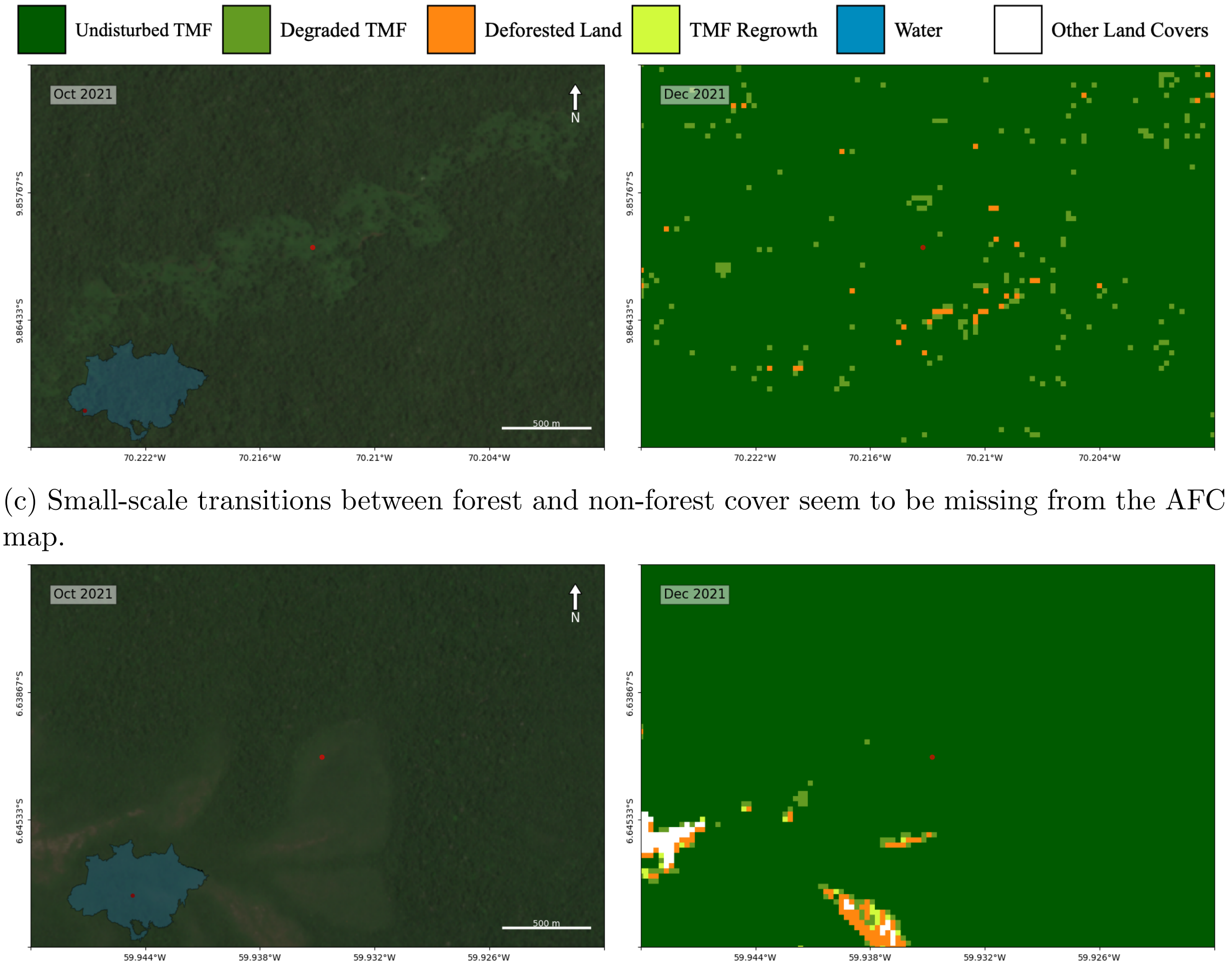
We also attempted to identify *undisturbed* TMFs misclassified as *deforested* land by filtering deforested samples with $RH95 > 20$ m. However, we were unsuccessful for several reasons. First, this approach does not reflect the complex nature of forests. Seeing a few square meters of trees does not indicate the presence of a forest. Second, the lack of historical height data prevents us from analyzing changes to distinguish primary forests from replacing tree plantations. Furthermore, $RH95$ is prone to obstacle noise, such as from a flock of birds. We examined some clusters of this kind and visualized their raw GEDI waveform (Figure 5.6 and 5.7). Waveforms with exceptionally tall RH95 (e.g., $RH95 > 40$ m) seem to be affected by noise in the return waveform and the ground return (Figure 5.6). However, the waveforms with lower RH95 (e.g., $RH95 \approx 20$ m) do not seem noisy (Figure 5.7). Other GEDI data quality issues, such as geolocation errors, could have caused such observations.



(a) The color difference, texture difference, and geometric shape suggest that this area is not an *undisturbed* TMF.

(b) The brown colors and texture difference suggests that this area is not an *undisturbed* TMF.

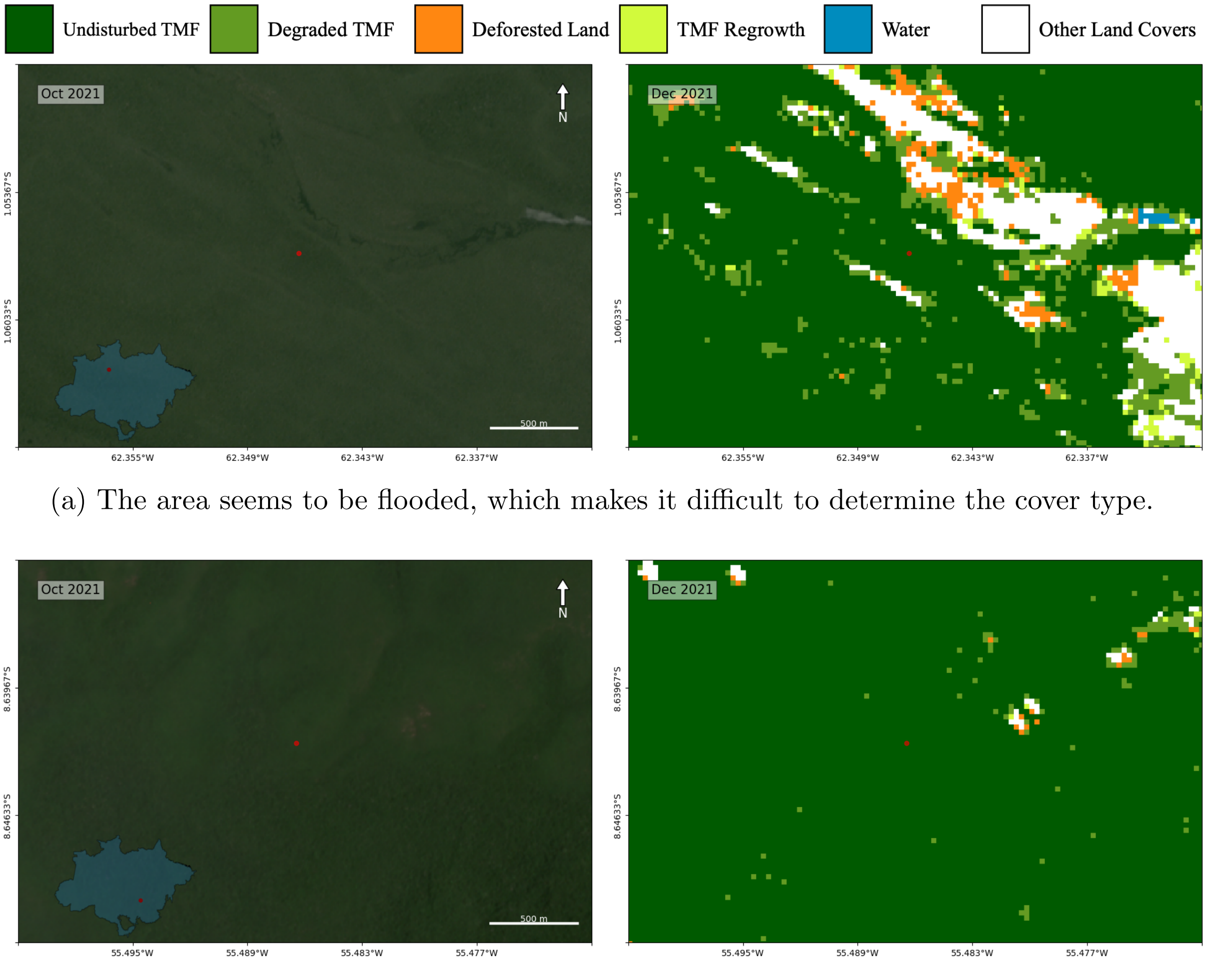
Figure 5.2: *AFC Error* examples. The images on the left are higher-resolution Planet imagery. Images on the right visualize the AFC map in the same location.



(c) Small-scale transitions between forest and non-forest cover seem to be missing from the AFC map.

(d) Although some deforestation was identified, the colour and texture difference suggests inaccuracies in the AFC map.

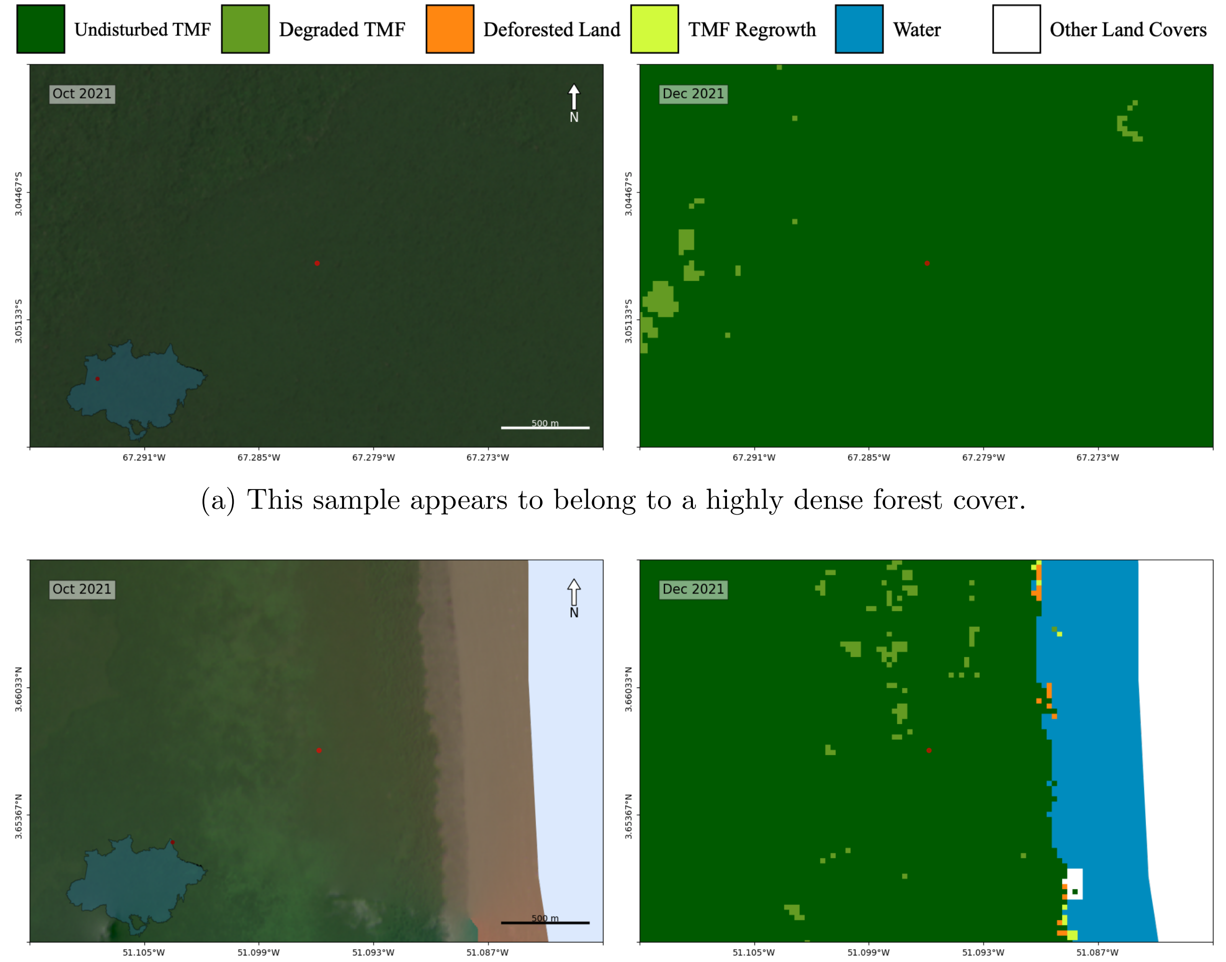
Figure 5.2: *AFC Error* examples. The images on the left are higher-resolution Planet imagery. Images on the right visualize the AFC map in the same location.



(a) The area seems to be flooded, which makes it difficult to determine the cover type.

(b) The available context is insufficient to determine whether this sample is an error.

Figure 5.3: *Ambiguous* examples. The images on the left are higher-resolution Planet imagery. Images on the right visualize the AFC map in the same location.



(a) This sample appears to belong to a highly dense forest cover.

(b) This area could potentially be covered by mangroves, which are classified as TMFs.

Figure 5.4: *False Positive* examples. The images on the left are higher-resolution Planet imagery. Images on the right visualize the AFC map in the same location.

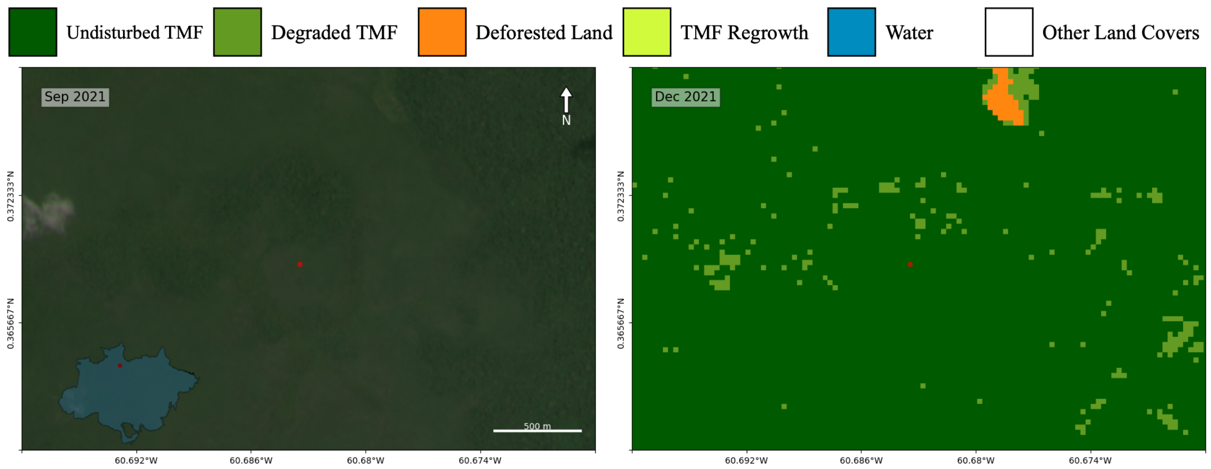
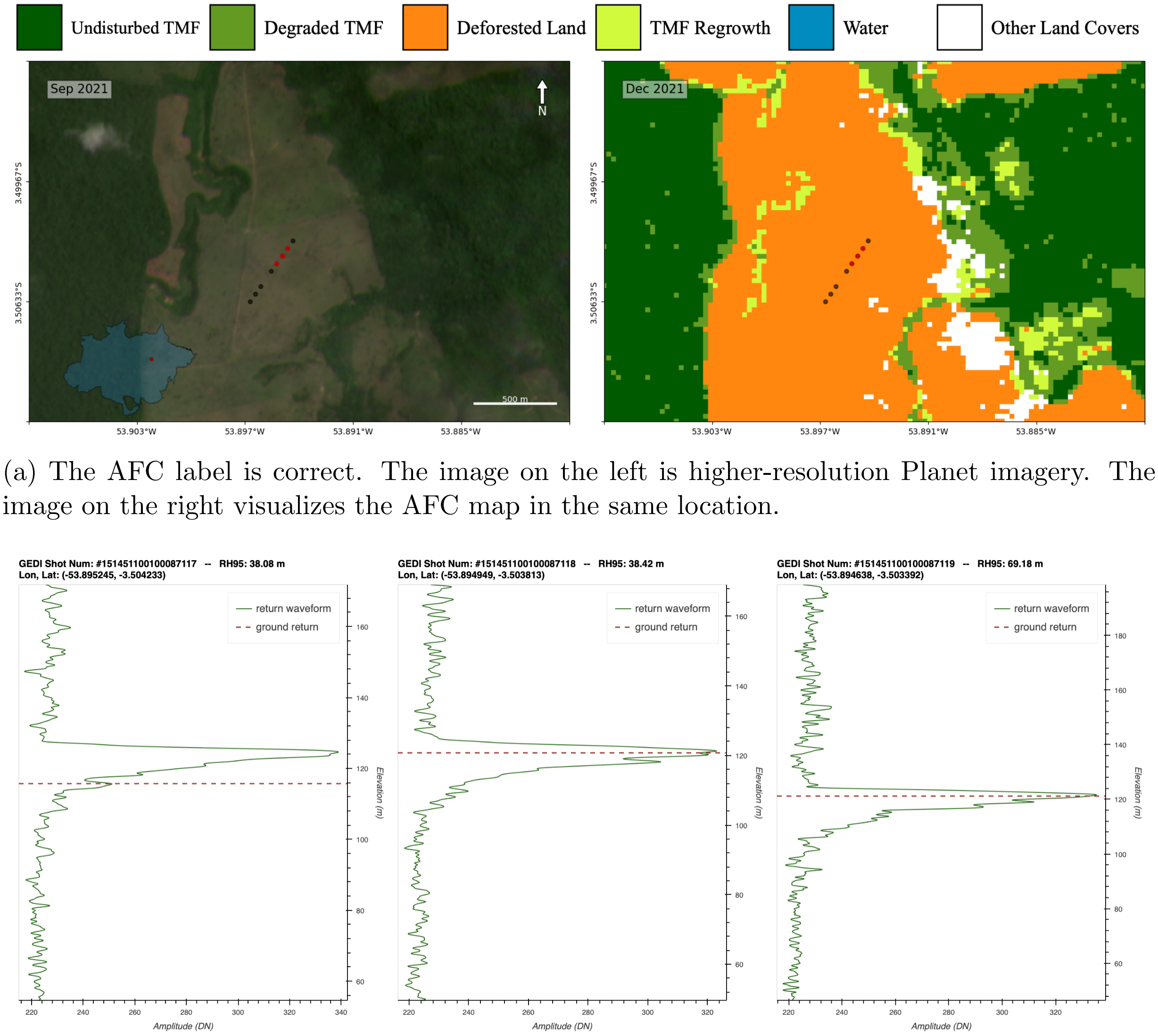


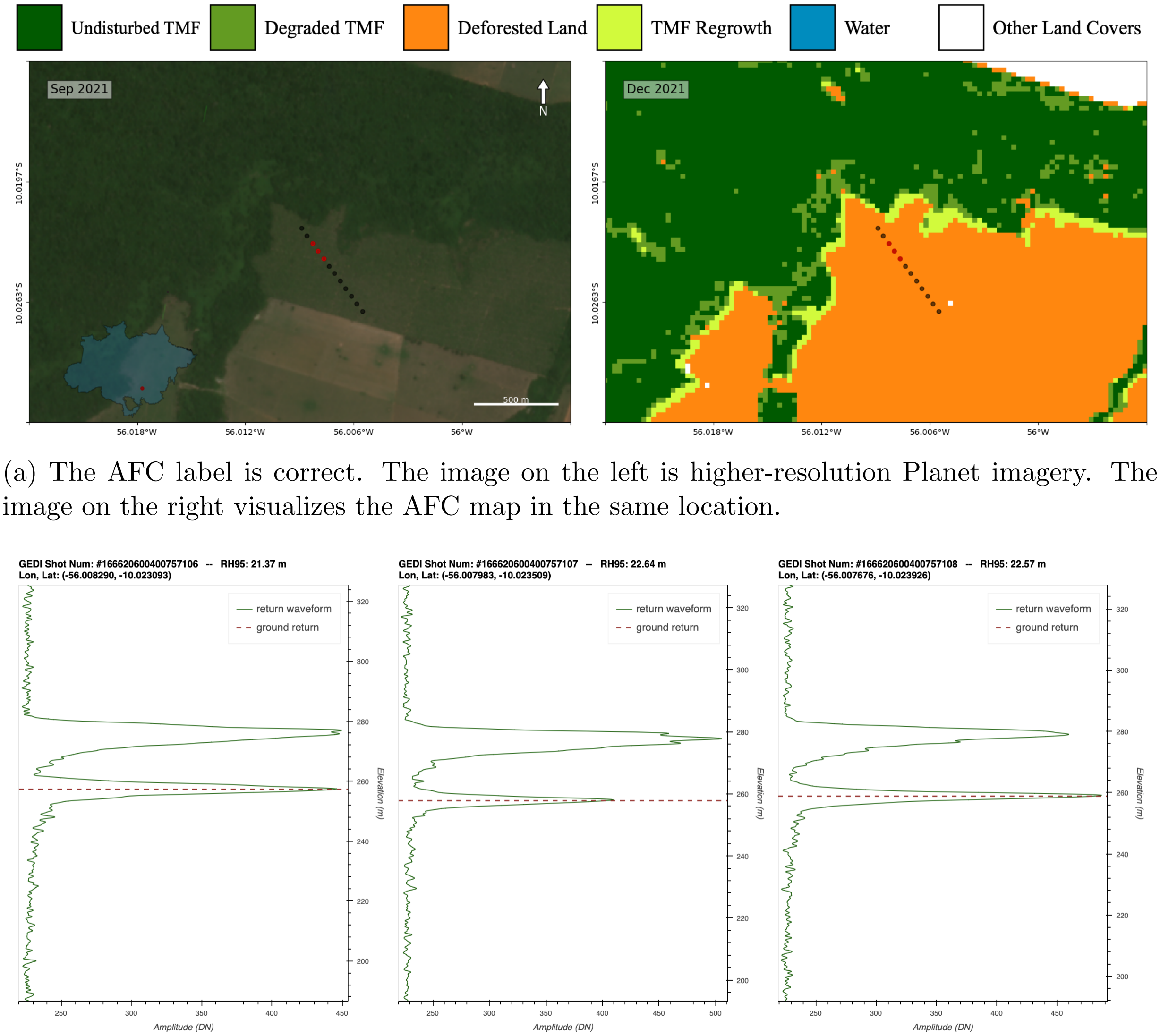
Figure 5.5: The distinction between forested campinaranas and open vegetation appears to be inaccurate in this area. The image on the left is higher-resolution Planet imagery. The image on the right visualizes the AFC map in the same location.



(a) The AFC label is correct. The image on the left is higher-resolution Planet imagery. The image on the right visualizes the AFC map in the same location.

(b) Raw GEDI waveforms of three consecutive shots (red shots) in the area with high RH95. All shots are from the same track, and the waveforms appear to be noisy.

Figure 5.6: A group of GEDI shots with high RH95 estimates in a deforested area.



(a) The AFC label is correct. The image on the left is higher-resolution Planet imagery. The image on the right visualizes the AFC map in the same location.

(b) Raw GEDI waveforms of three consecutive shots (red shots) in the area with high RH95. All shots are from the same track; however, it is not clear what has caused this error.

Figure 5.7: A group of GEDI shots with high RH95 estimates in a deforested area.

Chapter 6

Related Work

In this chapter, we discuss the related work in two parts. We first review previous studies on evaluating forest maps and explain how their findings are consistent with ours. We then review several case studies on using GEDI for forest monitoring and deforestation detection.

6.1 Forest Map Evaluation

In this study, we used canopy height to find data quality issues in a forest change map. A recent study explored the use of canopy height to distinguish forested and unforested tropical wetlands [76]. They used a global canopy height map with a 30 m resolution, created by combining GEDI RH metrics and Landsat images [53]. In contrast, our approach relies solely on raw GEDI height measurements.

In addition to creating a forest change map that estimates the year of forest loss, Hansen et al. [28] studied the relationship between loss year and canopy height using an older spaceborne LiDAR dataset. They observed that samples from undisturbed forests, on average, had greater canopy height than disturbed forests. This finding is consistent with our work.

Assessing the pixel-level agreement of two forest change maps is another way to find errors. However, this can be challenging due to variations in the map legends and differences in resolution. Moreover, two maps based on Landsat optical images can be subject to the same data quality problems. Cohen et al. compared seven forest change maps [5, 37, 77, 31, 36, 83, 33] at pixel level, revealing a low level of agreement [11]. On the other hand, GEDI allows us to leverage 3D information that does not exist in Landsat.

The first step to comparing two forest maps at the pixel level is creating a joint legend. Nevel et al. [45] compared two land cover maps, MapBiomas (collections 2 and 3) [54] and TerraClass [2], by reclassifying the labels to create a joint legend. This process involved identifying equivalent classes and merging some of them, such as merging different pasture types from TerraClass to align with MapBiomas. While the results of the per-pixel evaluation indicated a good level of agreement between them, the differences in the legends caused some uncertainties. For example, TerraClass categorized any area after deforestation as *secondary vegetation*, whereas MapBiomas (collection 3) did not have a separate class for this type. Per-pixel evaluation between AFC and MapBiomas would cause similar uncertainties. For instance, MapBiomas classifies wooded savannah as *forest formation*, whereas AFC classifies it as *other land covers*. Therefore, per-pixel evaluation cannot identify AFC errors in wooded savannahs, but using canopy height can.

6.2 GEDI Case Studies

In this section, we summarize some of the recent studies that used GEDI measurements for forest monitoring.

6.2.1 Tree Species Richness

Marselis et al. explored the relationship between canopy structure and tree species richness (number of unique species in an area) in undisturbed forests [43]. They used several GEDI Level 2 measurements to represent canopy structure, including RH98, the total Plant Area Index (PAI) along the vertical axis, the total PAI in 10-meter increments from 0 to 50 meters (0-10 m, 10-20 m, . . . , 40-50 m), the total PAI above 50 meters, and the ground elevation. The results indicated that GEDI measurements could explain tree diversity species; however, this does not offer an advantage over environmental variables (e.g., average temperature, geographic region) in undisturbed forests.

6.2.2 Forest Regrowth Rate

Milenković et al. studied the regrowth rate in the Amazon rainforest using GEDI canopy height measurements [44]. Since there is no GEDI data from before 2019, they classified the land by forest age to investigate the distribution of forest height across age groups. They also used a ground truth dataset to demonstrate that the RH98 metric overestimates shorter heights and underestimates taller heights. To address this issue, they developed a linear model to correct these errors. The results showed that forest height grows with a median of 20.17 meters over 33 years.

6.2.3 Impacts of Insect Infestation

Insect infestation causes deforestation and degradation. Certain insects initially cause defoliation in the lower layers of the forest before affecting the top canopy. Therefore, in the early stages of infestation, the forest would look undisturbed from satellite images; however, GEDI measurements could reveal disruptions at lower layers. Boucher et al. studied the impact of an insect infestation on northeastern US forests [4]. The infestation started in 2012, and the fields were revisited in 2016 to calculate the mortality rate (the number of dead trees divided by the total number of trees). Since no GEDI data is available for that period, they used airborne LiDAR (collected by aircraft) to simulate GEDI waveforms. The study found notable changes in the mid-canopy region (40-70% height) with slight canopy growth in the upper layers.

6.2.4 Characteristics of Old Forests

Spracklen et al. [64] examined the structural differences between old-growth forests (forests without major disturbance) and other forest types (including younger forests and managed forests) using GEDI Level 2 metrics. Although the study focused on forests in the Ukrainian Mountains characterized by slopes and dense canopy cover, the authors effectively employed GEDI metrics to classify the two forest types with over 70% accuracy using a Random Forest Classifier. The analysis found that old-growth forests are more complex, with layered canopies and more vegetation at the shrub layer.

6.2.5 Effects of Understory Fire

Detecting and assessing forest fires, including their severity, extent, and impacts, is crucial to forest monitoring. However, studying fires that occur underneath the forest canopy (known as the understory layer) is challenging, and the lack of 3D information in optical satellite images, such as Landsat, is one of the reasons. East et al. studied whether GEDI RH measurements could bridge this gap [23]. They use airborne LiDAR data to simulate

GEDI waveforms for 2013 and study two sites that burned in 2005 and 2008. They showed that the burned areas had lower RH than the unburned areas; however, in some cases, the difference is 2 to 3 meters. Thus, considering the geolocation error and RMSE of the actual GEDI data, it might not detect these effects. Other GEDI measures, such as PAI and biomass, are yet to be studied for this purpose.

Chapter 7

Conclusions

We described AFC and GEDI, two important forest monitoring datasets, and their data quality challenges. Although GEDI alone cannot be used to create a forest change map, it provides 3D information about forests missing from optical satellite imagery. We proposed a novel approach to find data quality issues in AFC using GEDI data, specifically, areas marked as TMF in the AFC map but with low RH95. Our findings show that this information can be used to create more accurate forest change maps.

We implemented our method to find potential data quality errors in the Brazilian Amazon's AFC map of 2021 and evaluated the results in two ways. Firstly, we used MapBiomas, a highly-validated land cover classification dataset, as ground truth. Secondly, we conducted visual interpretation of high-resolution Planet images. Our visual interpretation revealed instances where both MapBiomas and AFC were inaccurate. This could be because both maps are derived from Landsat images and are subject to similar data quality issues. Therefore, GEDI metrics could also be used to improve MapBiomas classifications.

The visual evaluation process was limited by the interpretation of a single evaluator, and future studies could benefit from using a voting technique and involving experts. Furthermore, there were many cases where it was ambiguous whether the land was covered with a forest. Using very high resolution images such as commercial Maxar[1] images at 0.3 meters resolution would eliminate such uncertainties.

An immediate next step is to apply the same method to other remote tropical regions such as the TMFs of Africa and Asia. Furthermore, GEDI data products provide additional measurements, such as various RH metrics, Canopy Cover Fraction (CCF), and Plant Area

[1] https://www.maxar.com

Index (PAI). Exploring these metrics in future work can reveal other data quality errors in existing datasets. Additionally, our method could be used to identify errors in other land cover classification maps by finding inconsistencies between GEDI data and the target class. For instance, we can apply this method to find errors in other 30-meter resolution forest change maps.

Another direction for future work is to build an interactive platform for finding potential errors in a forest change map using the GEDI RH95 metric. Data fusion is the only time-consuming step of the algorithm, and the subsequent three steps can be executed fast (Figure 4.1). Therefore, after precomputing this step and storing the results in a spatial database, we could deploy a platform that allows researchers to select an area, configure model parameters, and explore potential data quality issues. A ready-to-use data exploration platform could be helpful in the process of creating and validating forest maps.

References

[1] Markus Adam, Mikhail Urbazaev, Clémence Dubois, and Christiane Schmullius. Accuracy assessment of gedi terrain elevation and canopy height estimates in european temperate forests: Influence of environmental and acquisition parameters. *Remote Sensing*, 12(23):3948, 2020.

[2] Cláudio Aparecido de Almeida, Alexandre Camargo Coutinho, Júlio César Dalla Mora Esquerdo, Marcos Adami, Adriano Venturieri, Cesar Guerreiro Diniz, Nadine Dessay, Laurent Durieux, and Alessandra Rodrigues Gomes. High spatial resolution land use and land cover mapping of the brazilian legal amazon in 2008 using landsat-5/tm and modis data. *Acta Amazonica*, 46:291–302, 2016.

[3] João E Batista, Nuno M Rodrigues, Ana IR Cabral, Maria JP Vasconcelos, Adriano Venturieri, Luiz GT Silva, and Sara Silva. Optical time series for the separation of land cover types with similar spectral signatures: cocoa agroforest and forest. *International Journal of Remote Sensing*, 43(9):3298–3319, 2022.

[4] Peter Brehm Boucher, Steven Hancock, David A Orwig, Laura Duncanson, John Armston, Hao Tang, Keith Krause, Bruce Cook, Ian Paynter, Zhan Li, et al. Detecting change in forest structure with simulated gedi lidar waveforms: A case study of the hemlock woolly adelgid (hwa; adelges tsugae) infestation. *Remote Sensing*, 12(8):1304, 2020.

[5] Evan B Brooks, Randolph H Wynne, Valerie A Thomas, Christine E Blinn, and John W Coulston. On-the-fly massively multitemporal change detection using statistical quality control charts and landsat data. *IEEE Transactions on Geoscience and Remote Sensing*, 52(6):3316–3332, 2013.

[6] Patrick Burns, Chris Hakkenberg, and Scott Goetz. Gedi l2ab + l4a global processing, 2023. Manuscript in Preparation.

[7] European Commission Joint Research Center. Tropical moist forests product - data access, 2023. [Online; accessed Jun 22, 2023].

[8] European Commission Joint Research Centre. Tropical moist forest - data users guide (v1), 2022. https://forobs.jrc.ec.europa.eu/TMF/download/TMF_DataUsersGuide.pdf.

[9] Jun Chen, Jin Chen, Anping Liao, Xin Cao, Lijun Chen, Xuehong Chen, Chaoying He, Gang Han, Shu Peng, Miao Lu, et al. Global land cover mapping at 30 m resolution: A pok-based operational approach. *ISPRS Journal of Photogrammetry and Remote Sensing*, 103:7–27, 2015.

[10] Changhyun Choi, Victor Cazcarra-Bes, Roman Guliaev, Matteo Pardini, Konstantinos P Papathanassiou, Wenlu Qi, John Armston, and Ralph O Dubayah. Large-scale forest height mapping by combining tandem-x and gedi data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16:2374–2385, 2023.

[11] Warren B Cohen, Sean P Healey, Zhiqiang Yang, Stephen V Stehman, C Kenneth Brewer, Evan B Brooks, Noel Gorelick, Chengqaun Huang, M Joseph Hughes, Robert E Kennedy, et al. How similar are forest disturbance maps derived from different landsat time series algorithms? *Forests*, 8(4):98, 2017.

[12] European Commission, Joint Research Centre, R Beuchle, F Achard, C Bourgoin, and C Vancutsem. *Deforestation and forest degradation in the Amazon : status and trends up to year 2021*. Publications Office of the European Union, Luxembourg, 2022.

[13] PostGIS Project Steering Committee et al. Postgis, spatial and geographic objects for postgresql.

[14] Carlos Leandro de Oliveira Cordeiro and Dilce de Fátima Rossetti. Mapping vegetation in a late quaternary landform of the amazonian wetlands using object-based image analysis and decision tree classification. *International Journal of Remote Sensing*, 36(13):3397–3422, 2015.

[15] Ruusa M David, Nick J Rosser, and Daniel NM Donoghue. Remote sensing for monitoring tropical dryland forests: a review of current research, knowledge gaps and future directions for southern africa. *Environmental Research Communications*, 4(4):042001, 2022.

[16] Claudio Aparecido de Almeida, Luis Eduardo Pinheiro Maurano, Dalton de Morisson Valeriano, Gilberto Camara, Lubia Vinhas, Alessandra Rodrigues Gomes, Antonio

Miguel Vieira Monteiro, Arlesson Antonio de Almeida Souza, Camilo Daleles Rennó, Daniel E Silva, et al. Methodology for forest monitoring used in prodes and deter projects. *CEP*, 12(010), 2021.

[17] Stefania Di Tommaso, Sherrie Wang, and David B Lobell. Combining gedi and sentinel-2 for wall-to-wall mapping of tall and short crops. *Environmental Research Letters*, 16(12):125002, 2021.

[18] Ralph Dubayah, John Armston, Sean P Healey, Jamis M Bruening, Paul L Patterson, James R Kellner, Laura Duncanson, Svetlana Saarela, Göran Ståhl, Zhiqiang Yang, et al. Gedi launches a new era of biomass inference from space. *Environmental Research Letters*, 17(9):095001, 2022.

[19] Ralph Dubayah, James Bryan Blair, Scott Goetz, Lola Fatoyinbo, Matthew Hansen, Sean Healey, Michelle Hofton, George Hurtt, James Kellner, Scott Luthcke, et al. The global ecosystem dynamics investigation: High-resolution laser ranging of the earth's forests and topography. *Science of remote sensing*, 1:100002, 2020.

[20] Ralph Dubayah, Michelle Hofton, James Blair, John Armston, Hao Tang, and Scott Luthcke. Gedi l2a elevation and height metrics data global footprint level v002, 2021.

[21] Ralph Dubayah, Scott Luthcke, James Blair, Michelle Hofton, John Armston, and Hao Tang. Gedi l1b geolocated waveform data global footprint level v001, 2020.

[22] Laura Duncanson, James R Kellner, John Armston, Ralph Dubayah, David M Minor, Steven Hancock, Sean P Healey, Paul L Patterson, Svetlana Saarela, Suzanne Marselis, et al. Aboveground biomass density models for nasa's global ecosystem dynamics investigation (gedi) lidar mission. *Remote Sensing of Environment*, 270:112845, 2022.

[23] Alyson East, Andrew Hansen, Dolors Armenteras, Patrick Jantz, and David W. Roberts. Measuring understory fire effects from space: Canopy change in response to tropical understory fire and what this means for applications of gedi to tropical forest fire. *Remote Sensing*, 15(3):696, Jan 2023.

[24] GEDI. Instrument specifications, 2023. [Online; accessed Aug 02, 2023].

[25] GEDI. Products, 2023. [Online; accessed Aug 02, 2023].

[26] Luke Gibson, Tien Ming Lee, Lian Pin Koh, Barry W Brook, Toby A Gardner, Jos Barlow, Carlos A Peres, Corey JA Bradshaw, William F Laurance, Thomas E Lovejoy, et al. Primary forests are irreplaceable for sustaining tropical biodiversity. *Nature*, 478(7369):378–381, 2011.

[27] Noel Gorelick, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore. Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote sensing of Environment*, 202:18–27, 2017.

[28] Matthew C Hansen, Peter V Potapov, Rebecca Moore, Matt Hancher, Svetlana A Turubanova, Alexandra Tyukavina, David Thau, Stephen V Stehman, Scott J Goetz, Thomas R Loveland, et al. High-resolution global maps of 21st-century forest cover change. *science*, 342(6160):850–853, 2013.

[29] Viola HA Heinrich, Christelle Vancutsem, Ricardo Dalagnol, Thais M Rosan, Dominic Fawcett, Celso HL Silva-Junior, Henrique LG Cassol, Frédéric Achard, Tommaso Jucker, Carlos A Silva, et al. The carbon sink of secondary and degraded humid tropical forests. *Nature*, 615(7952):436–442, 2023.

[30] Amelia Holcomb, Simon Mathis, Srinivasan Keshav, and David A. Coomes. Computational tools for assessing forest recovery with gedi and forest change maps., 2023. In review.

[31] Chengquan Huang, Samuel N Goward, Jeffrey G Masek, Nancy Thomas, Zhiliang Zhu, and James E Vogelmann. An automated approach for reconstructing recent forest disturbance history using dense landsat time series stacks. *Remote Sensing of Environment*, 114(1):183–198, 2010.

[32] Alfredo R Huete. Vegetation indices, remote sensing and forest monitoring. *Geography Compass*, 6(9):513–532, 2012.

[33] Michael Joseph Hughes. *New remote sensing methods for detecting and quantifying forest disturbance and regeneration in the Eastern United States.* PhD thesis, University of Tennessee, Knoxville, 2014.

[34] IBGE. Manual técnico da vegetação brasileira, 2012.

[35] Ihab F. Ilyas and Xu Chu. *Data Cleaning.* Association for Computing Machinery, New York, NY, USA, 2019.

[36] Suming Jin, Limin Yang, Patrick Danielson, Collin Homer, Joyce Fry, and George Xian. A comprehensive change detection method for updating the national land cover database to circa 2011. *Remote sensing of environment*, 132:159–175, 2013.

[37] Robert E Kennedy, Zhiqiang Yang, and Warren B Cohen. Detecting trends in forest disturbance and recovery using yearly landsat time series: 1. landtrendr—temporal segmentation algorithms. *Remote Sensing of Environment*, 114(12):2897–2910, 2010.

[38] Colin R Leslie, Larisa O Servina, and Holly M Miller. *Landsat and Agriculture: Case Studies on the Uses and Benefits of Landsat Imagery in Agricultural Monitoring and Production.* US Department of the Interior, US Geological Survey Reston, VA, USA, 2017.

[39] Mengyu Liang, Mariano González-Roglich, Patrick Roehrdanz, Karyn Tabor, Alex Zvoleff, Veronika Leitold, Julie Silva, Temilola Fatoyinbo, Matthew Hansen, and Laura Duncanson. Assessing protected area's carbon stocks and ecological structure at regional-scale using gedi lidar. *Global Environmental Change*, 78:102621, 2023.

[40] Lei Ma, George Hurtt, Hao Tang, Rachel Lamb, Andrew Lister, Louise Chini, Ralph Dubayah, John Armston, Elliott Campbell, Laura Duncanson, et al. Spatial heterogeneity of global forest aboveground carbon stocks and fluxes constrained by spaceborne lidar data and mechanistic modeling. *Global Change Biology*, 2023.

[41] MapBiomas. Algorithm theoretical basis document (atbd): Collection 7, August 2022. https://mapbiomas-br-site.s3.amazonaws.com/ATBD_Collection_7_v2.pdf.

[42] MapBiomas. Mapbiomas legend description: Collection 7, August 2022.

[43] Suzanne M Marselis, Petr Keil, Jonathan M Chase, and Ralph Dubayah. The use of gedi canopy structure for explaining variation in tree species richness in natural forests. *Environmental Research Letters*, 17(4):045003, 2022.

[44] Milutin Milenković, Johannes Reiche, John Armston, Amy Neuenschwander, Wanda De Keersmaecker, Martin Herold, and Jan Verbesselt. Assessing amazon rainforest regrowth with gedi and icesat-2 data. *Science of Remote Sensing*, 5:100051, 2022.

[45] Alana Kasahara Neves, Thales Sehn KÖRTING, Leila Maria Garcia Fonseca, and Maria Isabel Sobral Escada. Assessment of terraclass and mapbiomas data on legend and map agreement for the brazilian amazon biome. *Acta Amazonica*, 50:170–182, 2020.

[46] NICFI. Norway's international climate and forest initiative (nicfi), 2022.

[47] Pedro VC Oliveira, Xiaoyang Zhang, Birgit Peterson, and Jean P Ometto. Using simulated gedi waveforms to evaluate the effects of beam sensitivity and terrain slope on gedi l2a relative height metrics over the brazilian amazon forest. *Science of Remote Sensing*, 7:100083, 2023.

[48] Yude Pan, Richard A. Birdsey, Jingyun Fang, Richard Houghton, Pekka E. Kauppi, Werner A. Kurz, Oliver L. Phillips, Anatoly Shvidenko, Simon L. Lewis, Josep G. Canadell, Philippe Ciais, Robert B. Jackson, Stephen W. Pacala, A. David McGuire, Shilong Piao, Aapo Rautiainen, Stephen Sitch, and Daniel Hayes. A large and persistent carbon sink in the world&#x2019;s forests. *Science*, 333(6045):988–993, 2011.

[49] Planet Labs PBC. Planet application program interface: In space for life on earth, 2018–.

[50] Jean-François Pekel, Andrew Cottam, Noel Gorelick, and Alan S Belward. High-resolution mapping of global surface water and its long-term changes. *Nature*, 540(7633):418–422, 2016.

[51] Anssi Pekkarinen. Global forest resources assessment 2020, 2020.

[52] Pierre Ploton, Frédéric Mortier, Maxime Réjou-Méchain, Nicolas Barbier, Nicolas Picard, Vivien Rossi, Carsten Dormann, Guillaume Cornu, Gaëlle Viennois, Nicolas Bayol, Alexei Lyapustin, Sylvie Gourlet-Fleury, and Raphaël Pélissier. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature Communications*, 11(1):4540, September 2020.

[53] Peter Potapov, Xinyuan Li, Andres Hernandez-Serna, Alexandra Tyukavina, Matthew C Hansen, Anil Kommareddy, Amy Pickens, Svetlana Turubanova, Hao Tang, Carlos Edibaldo Silva, et al. Mapping global forest canopy height through integration of gedi and landsat data. *Remote Sensing of Environment*, 253:112165, 2021.

[54] MapBiomas Amazonía Project. Collection of brazilian land cover & land use map series, 2017. accessed on 25 July 2017.

[55] MapBiomas Amazonía Project. Collection 7.0 of annual land cover and land use maps, 2021. accesed on 05.2023 via link http://amazonia.mapbiomas.org.

[56] Victor Hugo Rohden Prudente, Sergii Skakun, Lucas Volochen Oldoni, Haron AM Xaud, Maristela R Xaud, Marcos Adami, and Ieda Del'Arco Sanches. Multisensor approach to land use and land cover mapping in brazilian amazon. *ISPRS Journal of Photogrammetry and Remote Sensing*, 189:95–109, 2022.

[57] Johannes Reiche, Richard Lucas, Anthea L Mitchell, Jan Verbesselt, Dirk H Hoekman, Jörg Haarpaintner, Josef M Kellndorfer, Ake Rosenqvist, Eric A Lehmann, Curtis E

Woodcock, et al. Combining satellite data for better tropical forest monitoring. *Nature Climate Change*, 6(2):120–122, 2016.

[58] Johannes Reiche, Adugna Mullissa, Bart Slagter, Yaqing Gou, Nandin-Erdene Tsendbazar, Christelle Odongo-Braun, Andreas Vollrath, Mikaela J Weisse, Fred Stolle, Amy Pickens, et al. Forest disturbance alerts for the congo basin using sentinel-1. *Environmental Research Letters*, 16(2):024005, 2021.

[59] David P Roy, Herve B Kashongwe, and John Armston. The impact of geolocation uncertainty on gedi tropical forest canopy height estimation and change monitoring. *Science of Remote Sensing*, 4:100024, 2021.

[60] Fabian D Schneider, António Ferraz, Steven Hancock, Laura I Duncanson, Ralph O Dubayah, Ryan P Pavlick, and David S Schimel. Towards mapping the diversity of canopy structure from space with gedi. *Environmental Research Letters*, 15(11):115006, 2020.

[61] LP DAAC User Services. Lp daac gedi finder service user guide version 1.0, 2020.

[62] Christopher Small. The landsat etm+ spectral mixing space. *Remote sensing of Environment*, 93(1-2):1–17, 2004.

[63] Carlos M Souza Jr, Julia Z. Shimbo, Marcos R Rosa, Leandro L Parente, Ane A. Alencar, Bernardo FT Rudorff, Heinrich Hasenack, Marcelo Matsumoto, Laerte G. Ferreira, Pedro WM Souza-Filho, et al. Reconstructing three decades of land use and land cover changes in brazilian biomes with landsat archive and earth engine. *Remote Sensing*, 12(17):2735, 2020.

[64] Ben Spracklen and Dominick V Spracklen. Determination of structural characteristics of old-growth forest in ukraine using spaceborne lidar. *Remote Sensing*, 13(7):1233, 2021.

[65] Hao Tang and John Armston1. Algorithm theoretical basis document (atbd) for gedi l2b footprint canopy cover and vertical profile metrics. Technical report, Goddard Space Flight Center: Greenbelt, MD, USA, 2019.

[66] Ramon Torres, Paul Snoeij, Dirk Geudtner, David Bibby, Malcolm Davidson, Evert Attema, Pierre Potin, BjÖrn Rommen, Nicolas Floury, Mike Brown, et al. Gmes sentinel-1 mission. *Remote sensing of environment*, 120:9–24, 2012.

[67] Michele Torresani, Duccio Rocchini, Alessandro Alberti, Vítězslav Moudrỳ, Michael Heym, Elisa Thouverai, Patrick Kacic, and Enrico Tomelleri. Lidar gedi derived tree canopy height heterogeneity reveals patterns of biodiversity in forest ecosystems. *Ecological Informatics*, 76:102082, 2023.

[68] UNEP. Unep annual report 2022. Technical report, UN Environment Programme (UNEP), 2023.

[69] United States Geological Survey (USGS). Coherent noise, 2023. [Online; accessed Jun 08, 2023].

[70] United States Geological Survey (USGS). Coherent noise storm, 2023. [Online; accessed Jun 08, 2023].

[71] United States Geological Survey (USGS). Data loss, 2023. [Online; accessed Jun 08, 2023].

[72] United States Geological Survey (USGS). Landsat 9 solid state recorder bad block issue, 2023. [Online; accessed Jun 08, 2023].

[73] United States Geological Survey (USGS). Oversaturation, 2023. [Online; accessed Jun 08, 2023].

[74] United States Geological Survey (USGS). Preliminary assessment of the value of landsat 7 etm+ data following scan line corrector malfunction. Technical report, United States Geological Survey (USGS), 2003.

[75] Christelle Vancutsem, Frédéric Achard, J-F Pekel, Ghislain Vieilledent, S Carboni, Dario Simonetti, Javier Gallego, Luiz EOC Aragao, and Robert Nasi. Long-term (1990–2019) monitoring of forest cover changes in the humid tropics. *Science Advances*, 7(10):eabe1603, 2021.

[76] Kamiel Verhelst, Yaqing Gou, Martin Herold, and Johannes Reiche. Improving forest baseline maps in tropical wetlands using gedi-based forest height information and sentinel-1. *Forests*, 12(10):1374, 2021.

[77] James E Vogelmann, George Xian, Collin Homer, and Brian Tolk. Monitoring gradual ecosystem change using landsat time series analyses: Case studies in selected forest and rangeland ecosystems. *Remote Sensing of Environment*, 122:92–105, 2012.

[78] Cangjiao Wang, Andrew J Elmore, Izaya Numata, Mark A Cochrane, Lei Shaogang, Jiu Huang, Yibo Zhao, and Yuanyuan Li. Factors affecting relative height and ground elevation estimations of gedi among forest types across the conterminous usa. *GIScience & Remote Sensing*, 59(1):975–999, 2022.

[79] James EM Watson, Tom Evans, Oscar Venter, Brooke Williams, Ayesha Tulloch, Claire Stewart, Ian Thompson, Justina C Ray, Kris Murray, Alvaro Salazar, et al. The exceptional value of intact forest ecosystems. *Nature ecology & evolution*, 2(4):599–610, 2018.

[80] Michael A Wulder, Thomas R Loveland, David P Roy, Christopher J Crawford, Jeffrey G Masek, Curtis E Woodcock, Richard G Allen, Martha C Anderson, Alan S Belward, Warren B Cohen, et al. Current status of landsat program, science, and applications. *Remote sensing of environment*, 225:127–147, 2019.

[81] Michael A Wulder, David P Roy, Volker C Radeloff, Thomas R Loveland, Martha C Anderson, David M Johnson, Sean Healey, Zhe Zhu, Theodore A Scambos, Nima Pahlevan, et al. Fifty years of landsat science and impacts. *Remote Sensing of Environment*, 280:113195, 2022.

[82] Lei Zhu, Wei Li, Philippe Ciais, Jiaying He, Alessandro Cescatti, Maurizio Santoro, Katsumasa Tanaka, Oliver Cartus, Zhe Zhao, Yidi Xu, et al. Comparable biophysical and biogeochemical feedbacks on warming from tropical moist forest degradation. *Nature Geoscience*, 16(3):244–249, 2023.

[83] Zhe Zhu and Curtis E Woodcock. Continuous change detection and classification of land cover using all available landsat data. *Remote sensing of Environment*, 144:152–171, 2014.