Apport du lidar spatial pour le développement de méthodes d'inventaire forestier multisource adaptées à la gestion durable des forêts dans un contexte de changement global

Anouk Schleich

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'INSTITUT DES SCIENCES ET INDUSTRIES DU VIVANT ET DE L'ENVIRONNEMENT - AGROPARISTECH

N°: 2024AGPT0002

En Géomatique

École doctorale GAIA - Biodiversité, Agriculture, Alimentation, Environnement, Terre, Eau – n°584
Portée par l'Université de Montpellier

Unité de recherche TETIS - Territoires, Environnement, Télédétection et Information Spatiale

Apport du lidar spatial pour le développement de méthodes d'inventaire forestier multisource adaptées à la gestion durable des forêts dans un contexte de changement global

Présentée par Anouk SCHLEICH

Le 25 avril 2024

Sous la direction de Sylvie DURRIEU
et Cédric VEGA

Devant le jury composé de

Richard FOURNIER, Professor, Université de Sherbrooke, Canada	Rapporteur
Jacqueline ROSETTE, Senior Lecturer, Swansea University, UK	Rapporteur
Philippe LEJEUNE, Professor, University of Liège, Belgium	Membre du jury
Suzanne MARSELIS, Assistant professor, University of Leiden, Netherlands	Membre du jury
Pierre COUTERON, Directeur de Recherche, IRD, France	Président

AgroParisTech

UNIVERSITÉ DE MONTPELLIER

Foreword

This thesis was conducted at the UMR TETIS at « La Maison de la Télédétection », in Montpellier, France. It was co-funded by the ACT department of INRAE and ENSG-IGN.

The thesis was supervised by Sylvie Durrieu and Cédric Vega and conducted within the SLIM project, supported by TOSCA Continental Surface program of the CNES order n° 4500066524.

Publications

Schleich, A., Durrieu, S., Soma, M., Vega, C., 2023. Improving GEDI Footprint Geolocation Using a High Resolution Digital Elevation Model. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 16, 7718–7732, doi: 10.1109/JSTARS.2023.3298991

Schleich, A., Bouriaud, O., Vega, C., Durrieu, S., 2024. Usefulness of GEDI footprints as first-phase sample for forest inventories based on double sampling for post-stratification. *Remote Sensing of Environment.*
In revision.

Schleich, A., Vega, C., Renaud, J.P., Bouriaud, O., Durrieu, S., 2024. kNN - Bagging NFI, GEDI, Sentinel-2 and Sentinel-1 data to produce estimates of forest volumes.
For submission to ISPRS Journal of Photogrammetry and Remote Sensing.

Communications at conferences

Oral presentations

Schleich, A., Durrieu, S., Soma, M., Vega, C., 2021. Improving GEDI Footprint Geolocation Using a High Resolution Digital Elevation Model. *SilviLaser* Vienna, Austria.

Schleich, A., Soma, M., Bouriaud, O., Renaud, J.P., Vega, C., Durrieu, S., 2022. Improving estimations of the French national forest inventory by stratification based on spaceborne LiDAR (GEDI). *ForestSAT* Berlin, Germany.

Poster

Schleich, A., Vega, C., Renaud, J.P., Durrieu, S., 2023. Bagging NFI, GEDI and Sentinel-2 data to produce high-resolution maps of forest volumes. *SilviLaser* London, UK.

Abstract

"Contribution of spaceborne lidar to the development of multisource forest inventory methods adapted to sustainable forest management in a context of global change"

The thesis focuses on the contribution of spaceborne lidar to the development of Multisource Forest Inventory (MFI) methods. In France, the National Forest Inventory (NFI) method addresses the requirements of public policies at regional and national levels. However, on smaller territories, precision is often insufficient to meet the needs of management activities. MFI methods better address these needs by combining inventory data with remote sensing data. This thesis aims to improve NFI accuracy at sub-regional to local scales by integrating data from the spaceborne lidar GEDI into multisource approaches.

Unfortunately, this integration is complicated due to the lack of spatial correspondence between field samples (inventory plots) and GEDI footprints. Additionally, GEDI data are poorly georeferenced, making them difficult to integrate into certain MFI approaches. This thesis focuses on these issues and is divided into three main parts.

As a first step, a method for improving GEDI georeferencing, based on a high-resolution reference digital elevation model (DEM) was developed. This method compares, for a series of positions around the location indicated in the GEDI products, the ground elevations of the GEDI footprints with those of the reference DEM, generating an error map according to X and Y offsets. Using a flow accumulation algorithm on this error map, an improved position minimizing the distance from the DEM is proposed for each GEDI footprint.

Next, two approaches for using GEDI data with NFI data were developed. The study sites are located in the Vosges and use ~ 500 NFI plots and over $100{,}000$ GEDI footprints.

The first approach is a double sampling for post-stratification (DSPS) approach, based on common variables between GEDI and NFI, without requiring spatial correspondence of the two data sources. DSPS approaches are generally based on probabilistic data samples, which is not a priori the case for GEDI's sampling pattern. Thus, a preliminary analysis was required to understand the characteristics of the spatial distribution of the GEDI sample. The relevance of the chosen common variable, i.e. the maximum tree height, was also verified. Compared with estimates based only on NFI data, the DSPS approach improved the variance of growing stock volume estimates by up to 56%.

The second approach is based on a link between GEDI data and NFI data, established indirectly by using spatially exhaustive data sources, the Sentinel-2 and Sentinel-1 images. To establish the model linking the different data sources, we chose to use the k-nearest neighbor (kNN) method combined with bagging (bootstrap aggregation). The aim is to propagate information from field plots to GEDI footprints in order to "densify" NFI plots by taking advantage of GEDI forest structure measurements, which are well correlated

with the forest attributes of interest (e.g. growing stock volume). First, for each NFI plot, we looked for the GEDI footprints with the characteristics of the Sentinel link variables, supplemented or not with a height link variable, that are closest to those of the NFI point. Using a kNN-bagging approach, the set of GEDI variables is therefore estimated for each NFI plot. Next, a regression model is established by kNN-bagging to estimate the volume using the best predicted GEDI variables from the previous step and the Sentinel variables. The volume is estimated at the level of all GEDI footprints. The strategy supplemented by a height link variable performed best and reached a coefficient of determination of 58%. Subsequently, using the resulting dense sample of volume plots, standard methods for small area estimation (scale of the municipality or district) or high-resolution volume mapping can be implemented.

Key words: *Multisource forest inventory, GEDI, spaceborne lidar, georeferencing, stratification, kNN, bagging*.

Résumé

"Apport du lidar spatial pour le développement de méthodes d'inventaire forestier multisource adaptées à la gestion durable des forêts dans un contexte de changement global"

En France, la méthode de l'Inventaire Forestier National (IFN) répond à des besoins de politique publique aux échelles nationales et régionales. Sur des plus petits territoires, la précision est souvent insuffisante pour répondre aux besoins des activités de gestion. Les méthodes IFM peuvent répondre à ce besoin en combinant des données d'inventaire et des données de télédétection. La thèse vise à améliorer la précision de l'IFN à des échelles subrégionales à locales en intégrant les données du système lidar spatial GEDI dans des approches multisources.

Cependant, cette intégration se heurte à un verrou majeur, lié à l'absence de correspondance spatiale entre les échantillons sur le terrain (placettes d'inventaire) et les empreintes GEDI. Par ailleurs, les données GEDI sont mal géoréférencées, ce qui complexifie leur intégration dans certaines approches d'IFM. Cette thèse se concentre sur ces problématiques et est divisée en trois parties principales.

Premièrement, une méthode d'amélioration du géoréférencement de GEDI a été développée en se basant uniquement sur un modèle numérique de terrain (MNT) de référence à haute résolution spatiale. Cette méthode compare, pour une série de positions autour de la localisation indiquée dans les produits GEDI, les élévations du terrain des empreintes GEDI avec celles du MNT de référence, générant une carte d'écarts en fonction des décalages en X et Y. En utilisant un algorithme d'accumulation de flux sur cette carte, une position améliorée qui minimise l'écart avec le MNT est proposée pour chaque empreinte GEDI.

Ensuite, deux approches d'utilisation des données GEDI avec les données de l'IFN ont été élaborées. Les zones d'étude se situent dans les Vosges et utilisent environ 500 placettes IFN et plus de $100{,}000$ empreintes GEDI. La première approche est une approche d'échantillonnage double pour la post-stratification (DSPS), reposant sur des variables communes entre GEDI et IFN, sans nécessiter de coïncidence spatiale entre les deux sources de données. Les approches DSPS reposent généralement sur des échantillons de données probabilistes, ce qui n'est a priori pas le cas de l'échantillonnage de GEDI. Ainsi, une analyse préliminaire a été nécessaire pour comprendre les caractéristiques spécifiques de l'échantillon des mesures GEDI. La pertinence de la variable commune choisie, la hauteur maximale des arbres, a également été vérifiée. Par rapport aux estimations basées uniquement sur les données IFN, l'approche DSPS a amélioré la variance des estimations de volume de 56%.

La deuxième approche utilise un lien entre données GEDI et données IFN établi indirectement en utilisant les images Sentinel-2 et Sentinel-1, avec la méthode des k-plus proches voisins (kNN) combinée avec du bagging (bootstrap aggregation). Il s'agit de propager l'information des placettes terrain au niveau des empreintes GEDI pour densifier les placettes IFN en tirant parti des mesures de structure forestière GEDI,

bien corrélées aux attributs forestiers d'intérêt (ex. le volume de bois). Tout d'abord, en utilisant un kNN-bagging, on cherche pour chaque placette IFN les empreintes GEDI ayant les caractéristiques les plus proches de celles du point IFN pour des variables de lien Sentinel, complétées ou non avec une variable de lien supplémentaire de hauteur. On estime ainsi l'ensemble des variables GEDI pour chaque placette IFN. Ensuite, un modèle de régression est établi par kNN-bagging pour estimer le volume de bois à partir des variables GEDI les mieux prédites à l'étape précédente et les variables Sentinel. Le volume est estimé au niveau de toutes les empreintes GEDI. La stratégie complétée par une variable de lien de hauteur a atteint un coefficient de détermination de 58%. Par la suite, sur la base du réseau dense de placettes avec volume ainsi obtenu, des méthodes standards d'estimation sur de petites surfaces (small area estimation) ou de cartographie haute résolution, pourront être implémentés.

Key words: *Inventaire forestier multisource, GEDI, lidar spatial, géoréférencement, stratification, K plus proches voisins, bagging*.

Acknowledgements

I would like to thank everyone who has participated in any way in my thesis project. The thesis was a very enriching experience, which would not have been possible without all the people who surrounded me.

My first thanks go to my thesis director Sylvie Durrieu and my co-director Cédric Vega. Thank you for supervising my research work, for your constant kindness and availability. Thank you for sharing your knowledge, your invaluable advice, your constructive comments, and follow-up throughout the thesis. It was a great pleasure to complete this work with you. I am very fortunate to have benefited from such good guidance.

I want to express my gratitude to my thesis reviewers Richard Fournier and Jacqueline Rosette, to the jury members Philippe Lejeune and Suzanne Marselis, and to the jury president Pierre Couteron for taking the time to review, evaluate and discuss my work.

I would like to thank all the people who have enriched my thoughts and contributed greatly to the completion of this thesis. First, I would like to thank my co-authors, Olivier Bouriaud, Jean-Pierre Renaud, Maxime Soma and Nikola Besic (and of course Sylvie and Cedric). Thank you for all the work sessions, in Nancy, Paris, Montpellier, Vienna, Berlin, London... but mostly remotely. Thank you for sharing your advice, knowledge and the numerous discussions. Olivier, thank you for the numerous hours spent explaining me statistics and the specificities of the forest inventory. Jean-Pierre, I thank you for the helpful statistical explanations (and the ride on top of the Vienna Wheel). Maxime, thanks for the "insights of a postdoc" and discussions about research. Nikola, I thank you for fruitful conversations and parallel research on GEDI and NFI data. It was a pleasure working with all of you and I am glad you were part of my thesis journey.

I would also like to thank the members of my comité de suivi: Jean-Baptiste Féret, Laurent Gazull, Milena Planells, Christophe Proisy and Jean-Pierre Wigneron. I thank you for your interesting insights and external point of views on the thesis topic. Thank you for your critical and encouraging feedback which greatly contributed to the smooth running of the thesis.

My thanks also go to the UMR TETIS, the ACT department, the entire Maison de la Télédétection, the Laboratoire d'Inventaire Forestier and IGN for the logistics, administrative, and financial support. To all the colleagues that I had the chance to meet during my thesis journey: You are too many to be mentioned here, but you will recognise yourself. Thank you for the (long) lunch breaks, the (no-)coffee breaks, and the extra-work sessions. A special thanks to Karun for your advice and friendship.

Thank you to my ENSG crew for regularly keeping in touch, and for memorable weekends that I eagerly awaited. A special thanks to Marie and Fanny for your daytime and nighttime everyday messaging. I am grateful for your infallible moral support, your help in decision-making; even for rereading, you are always

there.

Thanks to all friends and family who came to visit us in Montpellier, it was fun discovering this beautiful region with you. Thank you to all family and friends who we met halfway, or who welcomed us into their homes: La Buisse, Seyssinet, Walldorf, Paris, les Sables, Geneva, Vienna, Brest, Toulouse, London... it gave me some welcome time-off from my thesis and a feeling of vacation every time I left my "cave".

Last but not least, I would like to thank my sister Rebecca, my parents, and especially Flo. Thank you, for always being there to motivate and encourage me. Thank you for your invaluable support in every aspect of life during these three years and beyond.

Thank you all. Merci à tous. Danke an alle.

Anouk

Table of contents

List of Figures

List of Tables

Chapter 1

Introduction

The objective of this thesis was to estimate the contribution of GEDI spaceborne lidar to the development of multisource forest inventory methods adapted to sustainable forest management in the context of global change. This first chapter introduces the dissertation by putting the research objectives into context.

1.1 Sustainable forest management in the context of global change

Forests play a multifaceted role by delivering essential ecological, economical and societal services. Serving as habitat for diverse biodiversity, they contribute significantly to the global carbon storage, representing $45\,\%$ of terrestrial systems' carbon (Bonan, 2008). Moreover, forests supply valuable wood resources, play a pivotal role in regulating water and soil cycles, e.g. to mitigate erosion, and offer recreational opportunities (IUFRO, 2018; Bonan, 2008).

Forests are deeply and in many ways connected to the global climate change, which is perceived by many as one of the greatest challenge of mankind (UN, 2021; EEA, 2023). Section 1.1.1 asserts the role of forests as carbon sinks, emphasizing their potential to contribute to climate change mitigation. Then, Section 1.1.2 outlines how climate change impacts forests and how those impacts accentuated by forest degradation and deforestation could turn forests into a carbon source. Section 1.1.3 emphasizes on the need for sustainable forest management, and Section 1.1.4 provides a zoomed-in perspective on the French approach of forest management.

1.1.1 Forests: carbon sinks and climate change mitigation

Forests play a crucial role in mitigating climate change, acting as carbon sinks by absorbing and assimilating carbon dioxide (Pan et al., 2011). Globally, estimates reveal that between 2001 and 2019, forests demonstrated a capacity to absorb twice as much carbon as they emitted, absorbing 7.6 billion metric tons of CO_2 per year (Harris et al., 2021). The Intergovernmental Panel on Climate Change (IPCC) underlines that the Agriculture, Forestry, and Other Land Use (AFOLU) sector, including forests, has the potential to contribute up to $30\,\%$ of the reduction of greenhouse gas emissions required by 2050 to limit global warming, aiming to keep the global mean temperature of this century below $2\,^{\circ}\mathrm{C}$ over pre-industrial levels. The IPCC's 6th assessment report emphasizes that the AFOLU sector, responsible for $13\,\%$ to $21\,\%$ of global total anthropogenic greenhouse gas emissions from 2010 to 2019, from which $45\,\%$ come from deforestation, holds significant near-term mitigation potential (Pathak et al., 2022). In the context of France, national strategies envision a substantial $87\,\%$ increase in forestry carbon sinks by 2050, as compared to the business-as-usual scenario (Ministère de la Transition Écologique et Solidaire, 2020). However, the carbon sink in France is supported by forest transition, which allows forests to establish themselves on former agricultural lands. This situation is not tenable in the long term. Aligning with France's commitment to sustainable forestry management, the National Strategy to Combat Deforestation aims to cease the importation of non-sustainable forest or agricultural products contributing to deforestation by 2030 (Ministère de la Transition Écologique et Solidaire, 2018).

1.1.2 Global warming affects forests, prompting adaptation

Forest ecosystems confront a range of challenges emerging from diverse sources, rendering them more susceptible to vulnerability. These sources can be categorized into three primary factors: anthropogenic, climatic, and a combination of both (Prăvălie, 2018). Notably, the climate factors have been largely aggravated as "human activities, principally through emissions of greenhouse gases, have unequivocally caused global

warming" (IPCC, 2023).

Anthropogenic factors, i.e. human activities, including deforestation, fragmentation, and pollution, pose challenges to forest ecosystems. Deforestation, driven by widespread clearing for agriculture or urbanization, is recognized as a significant threat, disrupting biodiversity and essential ecological functions like carbon sequestration and water regulation (Bonan, 2008). Forest fragmentation, resulting from activities like logging and infrastructure development, introduces breaks in the once-uninterrupted forest, dividing it into smaller, isolated patches or fragments. This disruption of ecosystem impacts nutrient and water flow, organism dynamics and species connectivity. Additionally, air pollution contributes to soil acidification and tree productivity decline, which impede photosynthesis and biomass production (Prăvălie, 2018).

Under the climatic factor, various issues such as phenological shifts, range shifts, die-off events, insect infestations, diseases, and severe weather events represent significant threats (Prăvălie, 2018; FAO, 2020; Seidl et al., 2018). Earth is globally warming, and while different scenarios are possible, all scenarios indicate a rise in the mean temperature (IPCC, 2023). As the mean temperature continues to increase, the impact of climatic factors is poised to intensify. The rise of the mean global temperature causes substantial alterations in phenological events, with a notable impact being the lengthening of the growing season, allowing for more growth and more carbon absorption. Conversely, die-off events, defined as large-scale climate-triggered forest mortality events, are predominantly caused by heat and droughts (Hammond et al., 2022; Prăvălie, 2018). Further warming will amplify the occurrence of extreme hotter-drought conditions. For study sites considered worldwide by Hammond et al. (2022), tree-mortality induced by climate conditions are projected to increase by $22\,\%$ and $140\,\%$ for scenarios involving mean global temperature increases of $2\,^{\circ}\mathrm{C}$ and $4\,^{\circ}\mathrm{C}$ above pre-industrial temperatures, respectively.

Moreover, global warming extends insect life cycles, which will allow for some major insect pests to produce two generations a year instead of one, with greater levels of brood survival during winter, consequently amplifying tree damages (Jaime et al., 2023; Fettig et al., 2022). Over the past two decades, large insect infestations have caused extensive tree mortality (Fettig et al., 2022; Kurz et al., 2008). For example, during the 1999 - 2015 period, mountain pine beetle outbreak in British Columbia in Canada, caused the defoliation of 80 million ha, with a maximum of 10 million ha in the year of 2007 (CCFM, 2020). The reduction in carbon storage within these ecosystems is noteworthy. As shown by Kurz et al. (2008) forest ecosystems can turn from net carbon sinks to net carbon sources due to insect outbreaks. The carbon repercussions of insect outbreaks will extend for an extended period, only ceasing when forest regeneration absorbs more carbon than is released from the decomposition of beetle-killed tree biomass and the carbon removed via salvage logging (Kurz et al., 2008).

Unprecedented wildfires further exacerbate challenges faced by forest ecosystems all over the world. The European Union has experienced its second-worst year for wildfires in 2022, with nearly $900{,}000$ ha of natural land affected by fires (San-Miguel-Ayanz et al., 2023), the worst year being 2017 with 1.3 million ha of burnt land. In 2022, burnt area in France presented $513\,\%$ compared to the 2012 - 2021 average: $70{,}301$ ha were affected by fires. Out of this total, $58{,}275$ ha were attributed to forests, while the remaining area consisted of other vegetation fires. The increased fire incidents were due to drought and above seasonal norm temperatures and several heatwaves. Large-scale fires have appeared in regions that are usually unaffected, including Brittany, Vosges and Jura (San-Miguel-Ayanz et al., 2023). The 2023 wildfire in northern Greece marked the largest ever recorded fire in the European Union. The European Forest Fire Information Service

(EFFIS) reports that the cumulative burned area in Greece has exceeded $174{,}000\,\text{ha}$ in 2023 (CAMS, 2023). The International Union of Forest Research Organizations (IUFRO) forecasts that Europe could experience a notable rise in the annual burned area, potentially reaching $120\,\%$ to $270\,\%$ above the average recorded in 2000 – 2010 by the year 2090 (IUFRO, 2018). Similarly, the area burned during the 2023 wildfire season in Canada reached unprecedented levels, marking it as the most extensive in the country's history. By october 6th 2023, $18{,}500{,}000\,\text{ha}$ had burned since January 1st. It multiplies by 2.5 the previous record of 1989 (CIFFC, 2023). Forest fires represent a substantial ecological and economic loss and contribute to extensive carbon emissions (Kurz et al., 2008; Prăvălie, 2018).

All the above-mentioned forest disturbances lead to a decline in habitat quality, adversely affecting plant and animal species. Deforestation has been a significant factor contributing to the extinction of over 300 species of terrestrial vertebrates over the past 500 years (Dirzo et al., 2014). The conversion of natural ecosystems, including forests, into agricultural or artificial areas is a predominant cause, especially in the tropics, significantly contributing to an extinction rate surpassing 100 species per million species per year, as estimated by Rockström et al. (2009). This rate is believed to be 100 to 1000 times higher than the assumed natural baseline (Rockström et al., 2009). The decrease of animals within forests has consequences for the forest ecosystem, particularly regarding pollination and seed dispersion, where animals play an indispensable role (Vidal et al., 2013; Jordano, 2000; Tong et al., 2023).

Climate and human influences also result in changes in forest composition. Human activities involve the substitution of native tree species with exotic species that yield higher timber productions (Remeš et al., 2020). Additionally, changing environmental conditions stimulate the growth of faster-growing trees, leading to a decline of slower-growing trees. The influence of climate change, promoting the spread of non-native plant species in warmer climates, poses a substantial threat to both native species and the overall functionality of forest ecosystems worldwide (Dyderski et al., 2018). The expected impacts by the end of the 21st century primarily affect regions at lower elevations and those already characterized by warmth and dryness. Forest productivity gains are still expected in the short and medium term in Northern Europe and at higher altitudes, provided that water and nutrients are not limiting factors. Conversely, losses are projected in Central and Southern Europe due to the migration of the most productive species (e.g., spruce) towards the North and higher altitudes (Labonne et al., 2019). Currently, mortality is occurring more rapidly in these regions.

All the above-mentioned factors strongly interact with each other, thus creating synergies that intensify challenges. For instance, drought and deforestation exacerbate each other (Desbureaux and Damania, 2018; Bagley et al., 2014) and droughts weaken trees, which can no longer defend themselves against insects and other aggressors (Fettig et al., 2022). Climate change introduces an additional dimension to the challenges by altering precipitation patterns, increasing temperatures, and intensifying extreme weather events. These changes disrupt the delicate balance within forest ecosystems, diminishing ecosystem services, biodiversity values, productivity, health, and carbon stock capacities. Forests have experienced significant increases in disturbances, marking historical records and raising concerns about future projections.

1.1.3 Sustainable forest management: a holistic approach to mitigate disturbances

With regard to the challenges imposed by global changes, sustainable forest management is essential. As defined by IUFRO, 2018 it means "the environmentally appropriate, socially beneficial, and economically viable management of forests for present and future generations". Sustainable forest management involves addressing vulnerabilities to fires, droughts and insects, promoting the wood industry, assessing and managing anthropogenic pressures, and developing precise indicators for a comprehensive understanding of forest conditions and dynamics. The goal is to protect biodiversity, maintain essential ecosystem functions, and ensure the overall health of forest ecosystems. Among others, it requires responsible logging, reforestation efforts, and conservation strategies to balance between human needs and ecosystem preservation.

It encompasses immediate strategies to handle ongoing disturbances, long-term enduring strategies to minimize the likelihood and intensity of future disturbances, and facilitating recovery afterward. As outlined by FAO, 2023, the restoration process begins with an assessment to identify and evaluate the extent and scope of degradation. Following this, planning and design are undertaken to determine appropriate restoration activities. Management strategies are developed, considering short- and long-term site needs. Finally, monitoring and evaluation processes are implemented to measure progress towards recovery. For instance, combating a beetle infestation involves immediate measures such as sanitation harvests, insecticides and semiochemicals (Fettig et al., 2022). The long-term approach focuses on reducing susceptible hosts through activities like thinning, prescribed burning, and adjusting age classes and species compositions, requiring continuous adaptation in response to global warming (Jaime et al., 2023; Fettig et al., 2022). However, evaluating the long-term impact of management decisions is challenging, given that the growth of forest takes several decades. Taking forest fires as another example, human and material resources must be promptly mobilized during fire events for rapid extinguishment. More actions include population awareness, a clear understanding of fires, fire surveillance and early-warning systems, and changes of land use practices for fire-resilience (IUFRO, 2018).

In the context of global change, effective management requires continuous monitoring of forest at local to global scales to assess and address the increasing impact of these disturbances, while reinforcing the regulatory role of forests in global change (European Commission, 2021). Evidence-based policy making is crucial for effective forest protection, necessitating a strong and continuous collaboration between national and international institutions and forest management services (IUFRO, 2018). Addressing these challenges, involves the need of governance with national action plans that lay down the principles of forest management. These plans should address adaptation and substitution mechanisms to adjust forests to global change, data collection to understand disturbances and forest ecosystems, evaluation of aggravating factors, assessment of disturbance intensity, early detection of disturbances, restoration of the forest landscape after disturbances, and the protection of existing forests.

1.1.4 Zoom: French perspective on forest management

French forests, with a wood stock volume of 2.8 billion m^3, stand as the third-largest in Europe, surpassed only by Sweden and Germany (EFA, 2023). The forest area currently covers 31% of the metropolitan territory (IGN, 2023a). The growing stock volume (GSV) has experienced a substantial increase, rising from 1.8 billion m^3 in 1985 to 2.8 billion m^3 today, and continues to increase. This represents a growth of over

50 % in about thirty years (IGN, 2023a). This volume growth is attributed to a significant increase in forest area and wood stock over the past century (IGN, 2023a). Indeed, French forests are undergoing a transition phase, primarily driven by an increase in the abandonment of agricultural lands and mountain pastures from the second half of the 19th century, followed by afforestation, whether occurring naturally or through plantations. The increase in surface area is followed, decades later, by an increase in the wood stock (Bontemps et al., 2020).

Despite its transitional status, this expansion suggests an opportunity to enhance wood harvesting, aligning with proposals dating back to the Grenelle Environment Roundtable in 2007 (Halley Des Fontaines, 2008) and in line with the 2018 update of the European Union (EU) Bioeconomy Strategy (European Commission, 2018). The EU strategy aims to reduce dependence on fossil fuels by promoting the development of a green economy. This strategy has been implemented into France's national forest-wood policies (Ministère de l'agriculture et de l'alimentation, 2017). The national forest-wood plan emphasizes two challenges: an economic challenge to increase the valorization of French forest resources and an environmental challenge to protect the forest, its biodiversity, and renew it to address climate change trough adaptation and mitigation (Ministère de l'agriculture et de l'alimentation, 2017). The shared objective for the forest-wood sector is to actively contribute to reducing greenhouse gas emissions in line with commitments made by the EU and France during COP 21. The European Commission published new guidelines for sustainable forest management with the European Green Deal in 2021, including the EU forest strategy for 2030 (European Commission, 2021). The European Green Deal encompasses a series of policy measures designed to steer the EU toward a sustainable transformation, striving towards achieving climate neutrality by 2050 as its primary objective. As an intermediate target, the EU is committed to reduce greenhouse gas emissions by 55% in 2030 compared to 1990 levels.

A better understanding of forests in general, and in France, is crucial for both assessing the impacts of climate change on forests and fostering the capacity of forests to mitigate climate change. However, successful implementation of forest management strategies requires the availability of relevant and up-to-date information on the forest resource and its dynamics. National Forest Inventories (NFIs) refer to well-established systematic assessments of forests. The primary purpose of a NFI is providing accurate and current information on a country's forest resources, including extent, composition, structure, and health (Tomppo et al., 2010; Vidal et al., 2016b). NFI is an important decision-making tool for all public and private forest actors and policy makers. The main components of NFIs are field plots, the establishment of which is based on statistical principles. Field agents collect data on several plots, and this information is then used to generate regional and national estimates of forest indicators. The French NFI was created in 1958 to assess metropolitan forest resources, the method has been changed in 2004 to better meet new national and international requirements as well as to enhance reactivity and better handle major crises in the forestry sector due to, as previously underlined, an increase in both amplitude and frequency of large disturbances (Vidal et al., 2005). NFIs help estimate forest characteristics such as volume, basal area, dominant heights, and species composition, which are used to inform and control forest policies, as well as management decisions in some countries (Tomppo et al., 2010; Breidenbach et al., 2021). Several forest characteristics can only be assessed from ground measurements; however, most NFIs require multiple years to collect sufficient data to achieve the level of precision required to adapt national policy to local context and to implement sustainable forest management. If the change in method in 2004 enabled to have annual up-to-date accurate resource evaluation from the regional to the national level, the French NFI requires several innovations to extend its scope

and address the need for more frequent updates and higher spatial resolution in forest assessment and monitoring (Hervé et al., 2017; Sagar et al., 2022). The development of Multisource Forest Inventory (MFI), which combines NFI data with auxiliary data, mainly remote sensing data or thematic maps, is considered as a solution to address these new needs.

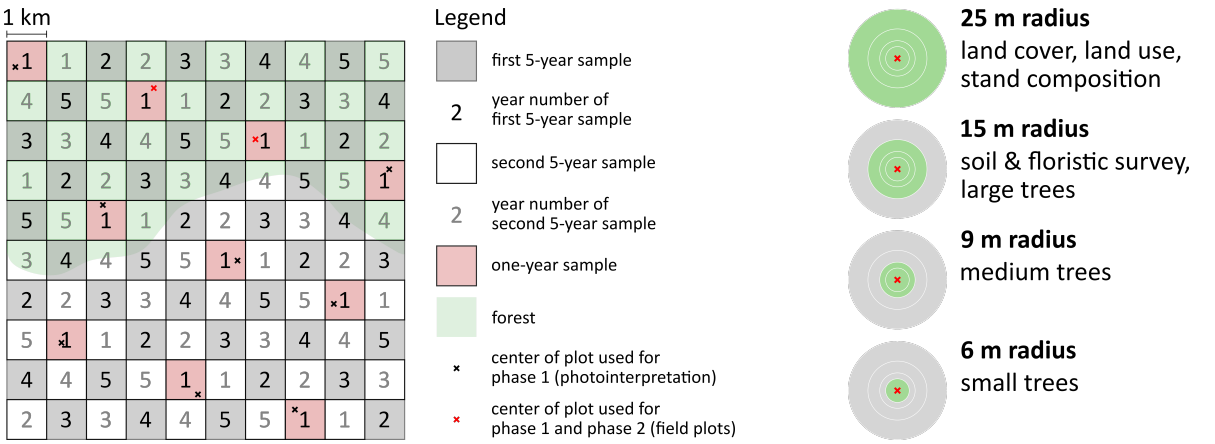
1.2 From national forest inventory to multisource forest inventory

1.2.1 The French national forest inventory

Since 2004, the French NFI employs a sampling strategy that covers the entire metropolitan territory on an annual basis. This strategy relies on a systematic 1 × 1 km grid as the sampling base (Bouriaud et al., 2023). This grid is organized into two 5-year samples, each further subdivided into five systematic sub-samples, with one sub-sample inventoried each year (Fig. 1.1a). This design allows the creation of an annual sample and sets of five or ten consecutive non overlapping annual samples, ensuring systematic coverage of the territory for all three time periods (Hervé, 2016). The grid construction results in a yearly coverage representing approximately one-tenth of the territory's surface. Thus, the remaining nine-tenths of the territory are not sampled in a given year.

The French NFI relies on a two-phase stratified sampling design. In the first phase, a single point is randomly selected within each of the $1\,\text{km}^2$ grid cells of the corresponding yearly sample, resulting in around $100{,}000$ sample points. The land cover and land use type are visually assessed at each of these points through photo interpretation. In the second phase, stratified sub-sample points are selected based on land cover type, defining the NFI field plots. Each year, approximately $7{,}000$ sample plots are surveyed in the field (Bouriaud, 2020; Hervé, 2016; Hervé et al., 2014). Furthermore, the points that were measured 5 years prior are revisited to estimate fluxes, increasing the number of surveyed plots to approximately $14{,}000$ per year.

The field measurements are made in four concentric circular plots (6 , 9 , 15 and $25\,\text{m}$ radii) centered on the sample points (Fig. 1.1b). The largest plot is used to describe the stand structure and includes observations on the stand conditions (i.e. stand structure, composition, topography and soils). Specifically, stand variables document recent disturbances, cutting activities, the age of the dominant stratum, the vertical structure of the stand and its composition. Site variables include details on elevation, slope, soil moisture, soil type, and a survey of plant species present in the plot. Tree measurements are made on the three smaller radii and involve species identification, status (i.e. dead or alive), circumference at breast height, among others (Hervé, 2016). Some measurement, like tree heights are limited to a sub-sample of trees. Small wood trees (from $7.5\,\text{cm}$ to $22.5\,\text{cm}$ in diameter at breast height) are assessed in the $6\,\text{m}$ radius plot, medium wood trees (from $22.5\,\text{cm}$ to $37.5\,\text{cm}$) in the $9\,\text{m}$ radius plot, and large wood trees (over $37.5\,\text{cm}$ in diameter) in the $15\,\text{m}$ radius plot.



(a) Sampling strategy (b) Field plot radii

Figure 1.1: Description of French NFI sampling and field measurement methods. (a) The metropolitan French territory is divided into 1x1km grid cells, forming two systematic five-year samples. Highlighted in red is a one-year subsample. Within each grid cell a randomized point is chosen for phase 1. During this phase land cover and land use are determined on a 25 m radius concentric plot around the point. For phase 2, a subsample of in-forest points from phase 1 is selected and surveyed by field agents. (b) the different concentric radii used by field agents to collect field data. Figure adapted from IGN (2023b).

National forest inventory relies on an optimized national and regional sampling plan to produce estimates for forests at these scales. However, in the context of global change and better resource mobilization, there is a demand for smaller scales, like those at which local public policies are applied, management strategies are developed or industrial stakeholders are prospecting resources. In the absence of systematic, consolidated management forest inventory, as in some Nordic countries (Maltamo and Packalen, 2014), one might consider the NFI as a source of information at those scales. But, the current sampling rate is insufficient to assess up-to-date forest resource at the scale of sub-regional territories (e.g., a supply basin).

The most straightforward solution to improve the precision of NFI data would be to intensify field data sampling, i.e. increasing the size of yearly samples. However, this is very costly and time-consuming. Moreover, the access to some forests can be physically difficult (e.g. no access by car, denied access, dangerous terrain) (Magnussen et al., 2018). The development and continuous refinement of MFI approaches has emerged as the best solution to overcome these practical issues and to meet the growing demand for comprehensive, cost-effective, regularly updated forest statistics with higher resolution.

1.2.2 Multisource forest inventory

Multisource Forest Inventories (MFI) combine NFI field plot data with auxiliary information, such as remote sensing and thematic maps, through statistical frameworks. This fusion of datasets and methodologies in MFIs offers numerous advantages, including cost-effectiveness, faster updated estimates, enhanced precision for larger areas, and acceptable precision for smaller regions (Tomppo et al., 2008; Westfall et al., 2019). The effectiveness of MFI approaches have been extensively tested, consistently demonstrating improved estimation precision across various countries (Tomppo et al., 2008; Saborowski et al., 2010; Westfall et al., 2019).

In MFI, three primary inference frameworks are commonly employed: design-based inference, model-

based inference, and model-assisted inference (Sagar, 2023). Design-based inference, based on statistical principles, relies on the properties of the sampling design to produce unbiased estimates of population parameters (Neyman, 1934). It is the most commonly used framework in NFIs. However, its dependence on a probabilistic sampling scheme may limit its applicability, and in instances of small sample sizes, precision diminishes, resulting in imprecise estimates (Gregoire, 1998). Model-assisted inference combines the robustness of design-based methods, which provide unbiased estimates using observations from the sample, with a precision enhancement offered by models. By linking field measured forest parameters to auxiliary information, models provide a large amount of predictions on a given territory, thus leading to a decrease in variance. This approach still requires NFI plots in each area of interest (Särndal et al., 1992). In Breidenbach and Astrup (2012), authors considered that a minimum of six plots were necessary. However, while well-specified models can significantly enhance precision, a poorly specified model might lead to higher variance compared to design-based methods (Saarela et al., 2015). Model-based inference relies exclusively on a model, allowing for efficient predictions in areas lacking field measurements, yet its accuracy depends on correct model specification, and its reliance on model accuracy may affect the precision of estimates (Gregoire, 1998; Magnussen, 2015). The model-based inference stands out for its capacity to operate without a probability sample from the target area, making it a feasible choice in situations where such samples are not available. Finally, hybrid inference integrates design-based and model-based approaches, accounting for both sampling and model errors (Fortin et al., 2018). It is particularly beneficial when wall-to-wall auxiliary data are unavailable or expensive (Corona et al., 2014; Ståhl et al., 2016). However, its application remains limited, and precise estimation can be challenging, particularly for inventories with diverse species (McRoberts et al., 2016).

The auxiliary data used in MFIs should satisfy the RARE criteria (Related (correlated), Affordable (cost), Renewed (times-series), Exhaustive). Remote sensing data are of particular interest for MFIs because they can provide valuable information about the forest cover and its characteristics over large areas. Three commonly types of Remote Sensing sensors employed in forest studies are optical sensors, Radio Detection and Ranging sensors (radar), and Light Detection And Ranging (lidar) devices (Fernandez-Ordonez et al., 2009; Bouvier et al., 2019; Coops et al., 2021). Optical remote sensing involves capturing sunlight reflected by Earth's surface in different spectral bands, including visible and infrared ranges. Optical images provide information about surface characteristics, and have been widely used to monitor vegetation, using indices derived from combined spectral bands. On the other hand, radar systems emit microwave signals towards the Earth's surface, and measures their reflection to detect surface properties. Radar, unlike optical sensors, can penetrate clouds. It offers insights into topography and vegetation structure. Finally, lidar operates by emitting laser pulses and measuring the time it takes for the reflected light to return to the sensor. This allows for the precise calculation of distances, enabling the creation of highly accurate three-dimensional representations of surfaces. Lidar data acquired from the ground is called Terrestrial Laser Scanning (TLS) and lidar data acquired from aerial platforms is called Aerial Laser Scanning (ALS). Indeed, remote sensing data can be collected from diverse platforms, including aerial platforms such as unmanned aerial vehicles (UAVs), drones, aircraft, and spaceborne platforms like Earth Observation Satellites and including the International Space Station (ISS). The wide range of spectral, spatial, and temporal resolutions, allows for the detection of various forest attributes such as tree species, canopy structure, biomass, and disturbances like deforestation or forest degradation.

While Landsat data are at the core of MFI in Finland, its applicability in France faces limitations due to

higher structural and compositional diversity (Irulappa Pillai Vijayakumar et al., 2019). For diverse forests, 3D data derived from Airborne Laser Scanning (ALS) or Photogrammetry emerges as a more effective solution as those data are very correlated to forest structure and associated characteristics such as basal areas and volume (Zolkos et al., 2013; Gobakken et al., 2012; Lim et al., 2003; Beland et al., 2019). Despite their acquisition costs, ALS coverages have been extensively employed in Nordic European Countries and Switzerland among others, contributing to the development of comprehensive resource maps. In France, a high-density lidar program started in 2020 and should end in 2025. However, during this interim period and because renewal of acquisitions is not guaranteed, alternative solutions are sought, and one promising possibility involves large-scale 3D acquisitions from spaceborne platforms such as ICESat-2 or GEDI.

Indeed, the launch of the two spaceborne lidar systems Ice, Cloud, and land Elevation Satellite-2 (ICESat-2) and Global Ecosystem Dynamics Investigation (GEDI) in 2018 marked a significant breakthrough in spaceborne lidar technology and a great promise for MFIs. They combine advantages from both satellite and lidar data within a single instrument. Spaceborne data has the advantage of being cost-effective as data are often freely available, rapid in acquisition, and capable of covering large areas. In contrast, aerial imagery proves to be costly, involves mostly one-time data acquisition, and is limited to smaller regions. Integration of lidar and satellite capabilities presents a promising synergy, combining the detailed insights on forest structure with lidar-based measurements with the broad coverage and efficiency of satellite data.

ICESat-2 employs a photon-counting lidar system to collect Earth's surface elevation data globally. The primary objective of ICESat-2 is to measure changes in Earth's ice sheets, glaciers, and sea ice, contributing to our understanding of the cryosphere and its impact on global sea level rise (Neuenschwander and Magruder, 2019). However, it can also be used to study other surfaces, including forests. It indeed provides valuable measurements of canopy heights (Neuenschwander and Magruder, 2019; Neuenschwander et al., 2020; Malambo et al., 2023). GEDI, on the other hand, was specifically designed to study forest ecosystems (Dubayah et al., 2020a). Its data provides information about the entire vegetation column, whereas ICESat-2 only provides canopy height. Hence, the choice to study the potential of GEDI in this thesis.

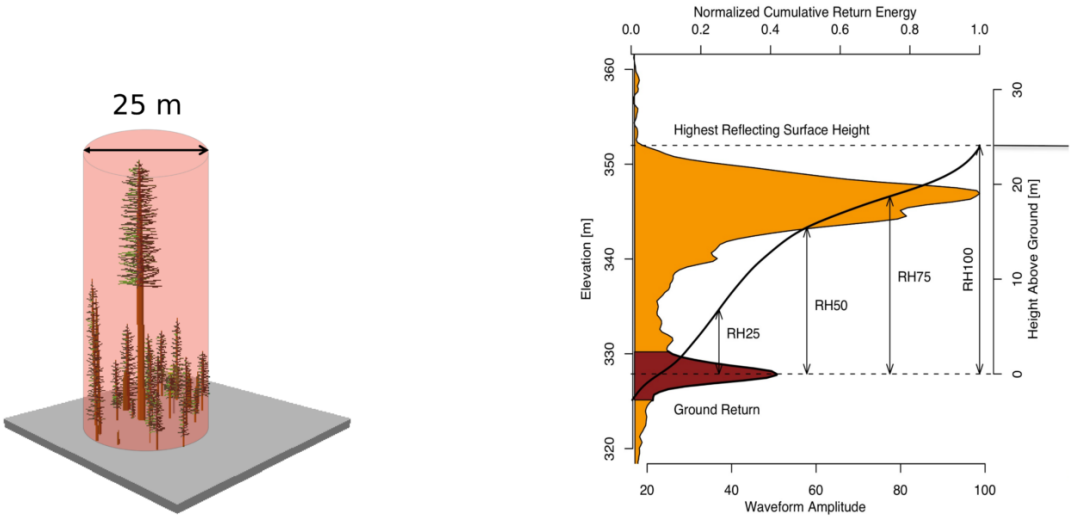
1.3 Integrating GEDI data into multisource forest inventory approaches to improve the monitoring of French forests

In Section 1.3.1, the GEDI mission and the data it collects are first presented to better understand the challenges raised by its integration to MFI approaches. Section 1.3.2 further presents the research questions and objectives addressed during this PhD work, and Section 1.3.3 provides an overview of the thesis.

1.3.1 Assets and challenges raised by GEDI data for its use in MFI approaches

GEDI is a spaceborne lidar instrument installed onboard the ISS, launched by NASA in 2018, specifically developed to monitor forests. It operates at a wavelength of $1{,}064$ nm, emitting laser beams toward the Earth surface to measure vertical vegetation structure. When the laser beam reaches the ground it covers a ~ 25 m diameter area named footprint (Fig. 1.2a and Fig. 1.3). The instrument is equipped with three lasers: two emitting at full power ("power" beams) and the third one being split into two beams of half energy ("coverage" beams). Therefore, at any one time, four beams, each with footprint diameter of ~ 25 m, are incident

on the ground. Each laser fires 242 times per second, and each beam is deflected every other shot by the beam dithering units. This configuration results in eight parallel tracks on the ground, spaced 600 m apart and with a footprint every 60 m along-track (Fig. 1.3). GEDI covers the Earth between the 51.6 °North and South latitudes (Dubayah et al., 2020a). It acquired data during an initial four years phase of the mission, from April 2019 to March 2023. Since then, the instrument has been paused and a new acquisition phase should start in the fall of 2024, and, hopefully, for six additional years (LP DAAC, 2023).



(a) 25-m diameter footprint

(b) GEDI waveform

Figure 1.2: Example of a GEDI waveform and Relative Heights (RHs) in L2A data. From https://gedi.umd.edu/

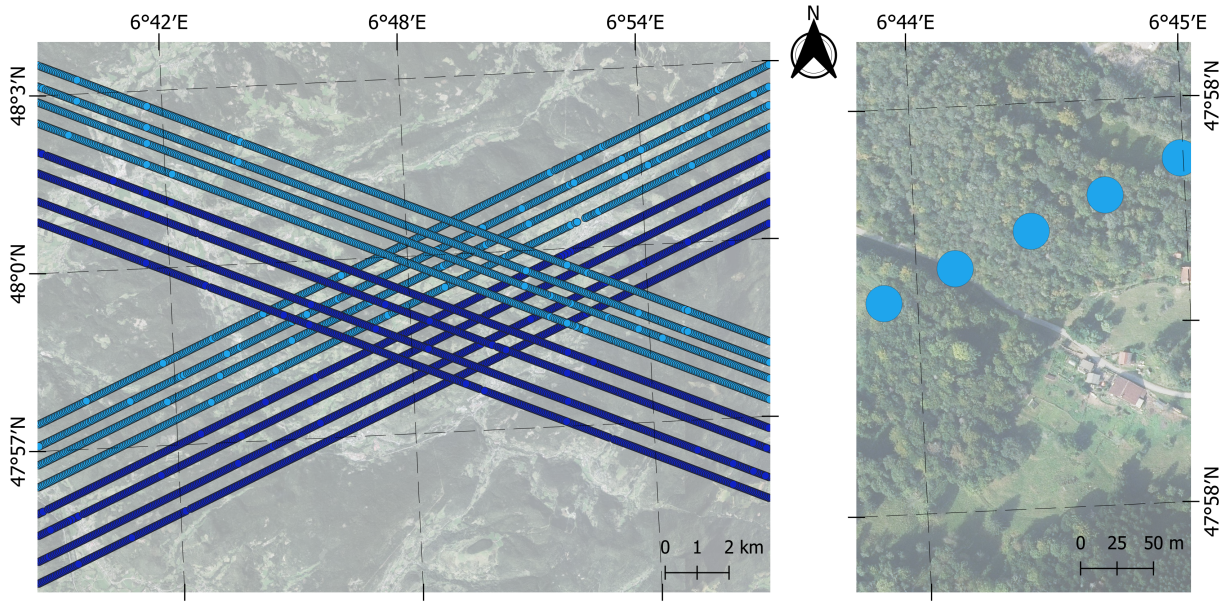


Figure 1.3: Example of an ascending and a descending GEDI orbit. Full-power beams are represented in dark blue and half-power beams are represented in light blue. In the left part, GEDI footprints are not to scale, while the right part displays the GEDI footprints at their actual scale, i.e. with a diameter of 25 m.

The emitted pulses encounter various elements in the forest canopy, such as leaves, branches, and the ground. Part of the laser energy is backscattered towards GEDI instrument and the received signal is digitized

at 1 Gsamp/sec, creating what is known as the "full-waveform". The latter carries information about the positions and the characteristics of encountered objects. The time it takes for the laser pulse to travel to the Earth's surface and back is recorded meticulously and allows to precisely measure the distance between the GEDI sensor and Earth's surface. The last significant return in the full-waveform typically corresponds to the laser pulse hitting the ground (dark red in Fig. 1.2b). It provides information about the elevation of the ground. Subsequent amplitudes in the waveform correspond to reflections from different strata within the forest canopy. The beginning of the backscattered signal (i.e. highest reflecting surface height in Fig. 1.2b) corresponds to the elevation of the top of canopy of the footprint. Relative heights (RH) within the canopy give the height at which a certain quantile of returned energy is reached relative to the ground, i.e. relative to the center of the ground peak. They provide information on the vertical distribution of vegetation within the footprint, offering insights into the forest's three-dimensional structure.

The Science Data Processing System (SDPS) generates the science data products and delivers the Level-0 and higher-Level data products to the NASA DAACs. GEDI's data releases include various levels (L1A, L1B, L2A, L2B, L3, L4A, L4B). L1A data consists in the raw waveforms of $25\,\mathrm{m}$ footprints, L1B in the geolocated waveforms, L2A in ground elevation and relative height metrics, L2B includes more variables such as the canopy cover and the leaf area index, L3 are $1\,\mathrm{km}$ gridded L2 metrics, L4A are footprint level Above Ground Biomass Density estimates (AGBD) and L4B are $1\,\mathrm{km}$ gridded AGBD estimates. GEDI data saw two releases, labeled as version 1 and version 2. Using the GEDI star-tracker system data, the SDPS includes the computation of positioning, pointing and ranging that are required to precisely geolocate the GEDI footprints. However, the ISS presents a challenging environment and footprint geolocation was found to be less precise than expected due to blinding periods of the star trackers and ISS movements (Dubayah et al., 2020a; Roy et al., 2021).

Forest structure measurements provided by GEDI data are of particular interest for the integration of GEDI data in MFI approaches. As underlined in Section 1.2.2, MFIs rely on the statistical combination of inventory data and partially correlated remote sensing data. In the same way as for airborne lidar data, GEDI information is likely to be highly correlated to forest attributes of interest such as volume or basal area. And, contrary to ICESat-2, GEDI mission has been designed to sample forests with an inferential sampling framework which is key for MFI. Therefore, the integration of data from the GEDI mission into MFI is expected to significantly enhance results and address some of the information needs expressed by the forest and wood sector. However, this integration faces a major challenge due to the low spatial density of measurements, resulting in the lack of spatial correspondence between field samples (inventory plots) and GEDI footprints. In addition, geolocation issues might prejudice the joint analyze of GEDI information with any other geolocated datasets, e.g. wall-to-wall remote sensing data like satellite imagery. It has also to be noted that in 2020 the ISS experienced a raise in orbit of around $16\,\mathrm{km}$, which modified the expected sampling, mostly causing clustering of observations along its orbital track and large gaps of data (Dubayah et al., 2022a). This change in the sample properties might cause a decrease in precision compared to expectations.

1.3.2 Research questions and objectives

This thesis is part of the TOSCA SLIM project (Space Lidar for Improved Multisource Forest Inventory), funded by CNES. The project aims to overcome the challenges raised, by integrating spaceborne lidar data, i.e. GEDI and possibly ICESat-2 data, into MFI approaches through the development of various methodological ap-

proaches.

The objective of my thesis is to assess the potential of data acquired by the spaceborne lidar system GEDI to improve forest inventory estimates.

The main research questions are:

1. Since spatial concordance between NFI inventory plots and GEDI measurements is not guaranteed, a first question concerns the ability to establish a link between field surveys (NFI plots with dendrometric measurements) and GEDI signals.
2. If establishing such a link is possible, a second question concerns the integration of GEDI measurements into an MFI approach, using this link. This involves identifying the appropriate statistical framework and estimators in order to study the improvement of forest inventory outputs at different working scales.
3. A third question concerns the impact of geolocation inaccuracy in footprint location when integrating GEDI data into MFI approaches and our capacity to develop strategies to account for this data characteristic. This question will interfere with the two previous ones.

Creating a link between GEDI and NFI data can be envisaged through different strategies relying on different concepts:

- Using a priori common variables between GEDI footprints and NFI plots. Those variables can be identified based on expertise.
- Using an indirect link by relying on continuous "gateway" data, such as wall-to-wall remote sensing data (e.g., optical or radar images).
- Establishing a direct link between NFI and GEDI information. Two main approaches can be considered to that aim. The first one would be to acquire additional NFI field measurements at the level of GEDI footprints. However, this approach is not an option from an operational point of view, as considerable additional fieldwork would be required leading to huge increase in cost. The second one is to use radiative transfer models to simulate GEDI signals at NFI field plot level by leveraging terrestrial lidar data acquired at the level of a significant subset of NFI plots.

For the first strategy, i.e. the identification of common variables, there is in theory no need of spatial intersection between GEDI and NFI data or other auxiliary data. Therefore, the misslocation of GEDI footprints should not be an issue. However, validating the quality of the link, i.e. the quality of the common variable, might require to rely on the second (using for example ALS data) or the third strategy. For the second strategy, using an indirect link, an intersection with independent auxiliary data is required, therefore creating a common variable space. As data are intersected with auxiliary data, the quality of data geolocation is assumed to be important. However, the impact of geolocation issues needs to be evaluated. The third strategy, using a direct link, requires to simulate GEDI signals at NFI plot locations with radiative transfer models. To calibrate the radiative transfer model, calibration plots are required, where GEDI and NFI data overlap. As a perfect overlapping is very unlikely to occur in existing datasets, in the context of the SLIM project, additional field plots were acquired at GEDI footprint positions. To ensure that field measurements are realized at the right location, a good geolocation of GEDI footprint is required.

In this thesis, I developed research to explore the potential of MFI implementing the two first strategies to link NFI and GEDI data. The third strategy is included in the SLIM project, but has not been addressed in this PhD work. However, whatever the investigated strategy, the need for accurate geolocation emerges, at least to quantify the quality of the link between NFI and GEDI data. This is why, focusing on the improvement of GEDI georeferencing emerged as an objective in my PhD and is the key component to address research question 3. Then, MFI approaches belonging to two different families were investigated to progressively downscale NFI results. First, a design-based approach aiming at providing improved up-to-date results up to a sub-regional scale was developed. Second, a modeling approach was proposed that could be further used in a model-assisted or a model-based MFI framework to provide small area or pixel-level estimations, respectively. Specific questions, related working hypotheses, and an overview of the methodological choices to address the three focal points of my PhD are outlined below.

1.3.3 Overview of the thesis

Improving GEDI geolocation with a Digitial Elevation Model

Most MFI methods require spatial coincidence between different types of data for overlay and analysis. Therefore, it should be important to have well-georeferenced data. However, GEDI data has shown horizontal precision below expectations. The horizontal error was estimated to $23.8\,\mathrm{m}$ for GEDI version 1 and to $10.2\,\mathrm{m}$ for GEDI version 2 (Beck et al., 2020, 2021). The first objective of the thesis was to improve the geolocation of GEDI data. Having in mind, that eventually the MFI workflow should be applicable to the entire metropolitan French territory, the geolocation method should also be applicable to the entire territory. Although methods to improve GEDI geolocation using ALS data exist in the literature, ALS data are not yet available everywhere, requires periodic updates, and poses challenges in processing due to its substantial volume. This leads to the formulation of the two following research questions:

1. Can GEDI footprints geolocation be improved on a large scale such as the French metropolitan territory ?
2. To what extend improving the geolocation accuracy of GEDI footprint influences the accuracy of MFI results ?

To address the first question it is hypothesized that using only ground information available through widespread high resolution DEMs can be sufficient to optimize GEDI footprint georeferencing. Indeed, and unlike complete ALS point clouds, Digital Elevation Models (DEMs) from either photogrammetric or ALS point clouds, are easily available at the national level and contain only ground elevations, which are quite stable over time, not subject to major changes in time. To address the second question it is hypothesized that the improvement in GEDI footprint geolocation accuracy will significantly enhance the results of MFI approaches using GEDI data as auxiliary data.

A method was developed to improve GEDI georeferencing using only a reference DEM. Our approach compares GEDI footprint elevations with DEM ground elevations and generates a difference map. An accumulation algorithm is then applied to the difference map, suggesting an improved position for each GEDI footprint. The method was tested on two different study sites: the Landes forest characterized by monocul-

ture of maritime pines planted in the 19th century, situated in a very flat area, and the Vosges forest located in a mountainous area and composed of very diverse coniferous, deciduous and mixed forest stands. While addressing the improvement of georeferencing of GEDI footprints, the article also addresses GEDI data quality. This work will be presented in the second chapter of the thesis (Chapter 2) and has been showcased at the SilviLaser conference in September 2021. An article has been published in the IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (Schleich et al., 2023c).

Subsequently, two MFI methods using GEDI data with NFI data were developed. As an example, my thesis focused on estimating the growing stock volume (GSV). Among the attributes observed by NFIs, GSV holds a key role in providing essential information for policy makers and forest practitioners (Gschwantner et al., 2022). Every year, the French NFI publishes GSV estimates at regional scales. While such information is relevant for national forestry policy purposes, sub-regional to local estimates of GSV are desired for forest management purposes.

Double sampling for post-stratification approach: a design-based MFI approach that can accommodate from geolocation inaccuracy

The first method uses a double sampling for post-stratification approach (DSPS) to estimate GSV. This method uses common variables between GEDI and NFI and does not require spatial coincidence of the two data sources. Geolocation is only considered at the point of determining whether the footprint or plot falls within the study site or not. The approach involves creating strata that group GEDI footprints and NFI plots according to the same criteria. The proportions of the strata, and therefore strata surfaces, are estimated using a large sample, in this case the GEDI footprints, while NFI data are used to calculate the averages of forest variables for each stratum. GSV estimates and their variance over the entire area of interest are then made using estimators designed for DSPS approaches. The development of a design-based MFI approach based on stratification requires answering the following questions:

1. Is the GEDI sample pattern a probabilistic sample pattern ?
2. Can a direct link be established between NFI and GEDI data and be further used as a basis for sample stratification ?

In order to develop and implement a double sampling for post-stratification approach (DSPS) we hypothesised, first, that GEDI's sampling pattern could be considered as a probabilistic sampling pattern and be used as a first-phase sample and, second, that height information, more specifically higher RH values from GEDI L2A products, correlated well with the maximum tree height of the NFI plots and could be used as a link variable.

This approach effectively reduces the width of confidence interval and is compatible with current NFI estimators. However, the two hypothesis mentioned above need to be verified. Prior research was essential to study the sampling plan properties of GEDI. To that aim, the reliability of the strata areas assessed using GEDI sample was evaluated by comparing them to the areas obtained using wall-to-wall ALS data. To verify the existence of at least one variable that is common to GEDI and NFI datasets and can serve as a valuable stratifying variable, we relied on ALS data to check the quality of the link between GEDI and NFI maximum heights. This first approach will be the focus of the third chapter of the thesis (Chapter 3). Preliminary results

were presented at the ForestSAT conference in September 2022 and an article was submitted to the journal Remote Sensing of Environment in 2023. Our manuscript has been accepted with major revisions, and we are currently in the process of addressing the reviewers' comments.

Using wall-to-wall auxiliary datasets in addition to GEDI data to develop a K-Nearest Neighbours model to predict forest attributes: a step towards model-assisted and model-based MFI

The second approach involves establishing an indirect link between GEDI and NFI data by leveraging a third data source accessible at both GEDI footprints and NFI plots. Model development requires to intersect NFI and GEDI data with other auxiliary data source. Given the potential for significant variations of forest characteristics within a few meters, precise geolocation is essential. The objective is to predict GSV and, ultimately, produce spatialized, high-resolution estimates with uncertainty assessments, potentially covering geographical domains of variable size. To establish a link between the different data sources and obtain information on the precision of model estimates, we opted for the k-nearest neighbors (kNN) method combined with bagging. kNN is indeed a simple, non-parametric supervised approach which allows to predict multiple attributes with a single model and which is widely used in MFI studies. Research questions for this approach include:

1. Can the use of additional auxiliary data help to create an indirect link between GEDI and NFI data ?
2. Can the approach be used to compute sub-regional estimates ?

It is hypothesized that Sentinel-1 and Sentinel-2 data, possibly completed with a height information, are relevant candidates to play the role of continuous "gateway" data between NFI and GEDI information. It is also assumed that, through this indirect link, a reliable model can be built to propagate NFI attributes and produce high resolution resource maps.

The method involves two parts. First, employing a kNN-bagging method using Sentinel data, we impute GEDI variables for each NFI plot. Second, using the imputed GEDI variables and Sentinel data, GSV is predicted for each point through another kNN-bagging approach. Three different strategies were tested. Strategy A involved using only Sentinel data to establish the link. Then, considering the hypothesis that Sentinel data alone might be insufficient to impute GEDI variables and predict volumes, auxiliary heights were added. Strategy B used a continuous existing national height map, while strategy C used the maximum height of GEDI and NFI plots as additional linking variables. This work has been presented as a poster at the SilviLaser conference in 2023 and is the focus of the fourth chapter of the thesis (Chapter 4). The manuscript will be submitted to ISPRS Journal of Photogrammetry and Remote Sensing.

To conclude, the last chapter, Chapter 5, provides an overall discussion, gathering key findings from each previous chapters and building on a cross-analysis of the three focal points addressed in my PhD. Several perspectives are also provided regarding future research.

CHAPTER 2

Improving GEDI footprint geolocation using a high resolution digital elevation model

Anouk Schleich[a], Sylvie Durrieu[a], Maxime Soma[a,b], Cédric Vega[c]

[a]UMR TETIS, Univ Montpellier, AgroParisTech, CIRAD, CNRS, INRAE, F-34196 Montpellier, France

[b]UMR RECOVER, INRAE, F-13182 Aix-en-Provence, France

[c]ENSG, IGN, Laboratoire d'inventaire forestier, F-54042 Nancy, France

Abstract

Global Ecosystem Dynamics Investigation (GEDI) is a lidar system on-board the International Space Station designed to study forest ecosystems. However, GEDI footprint low accuracy geolocation is a major impediment to the optimal benefit of the data. We thus proposed a geolocation correction method, GeoGEDI, only based on high-resolution digital elevation models (DEMs) and GEDI derived ground elevations. For each footprint, an error map between GEDI ground estimates and reference DEM was computed, and a flow accumulation algorithm was used to retrieve the optimal footprint position. GeoGEDI was tested on 150 000 footprints in Landes and Vosges, two French forests with various stands and topographic conditions. It was applied to GEDI versions 1 (v1) and 2 (v2), by either a single or four full-power laser beam tracks. GeoGEDI output accuracy was evaluated by analyzing shift distributions and comparing GEDI ground elevations and surface heights to reference data. GeoGEDI corrections were greater for v1 than for v2 and agreed with errors published by NASA. Within forests, GeoGEDI improved the RMSE of ground elevation in Landes by 26.8 % (0.34 m) and by 13.3 % (0.14 m) for v1 and v2, respectively. For Vosges, ground elevation RMSE improved by 59.6 % (3.82 m) and 36.2 % (1.41 m), for v1 and v2, respectively. Regarding surface heights, except for v2 in Landes, where insufficient variations in topography combined to GEDI ground detection issues might have penalized the adjustment, GeoGEDI improved GEDI estimates. Using GeoGEDI showed efficient to improve positioning bias and precision.

2.1 Introduction

The Global Ecosystem Dynamics Investigation (GEDI) instrument has been designed to collect unique data on vegetation structure (Potapov et al., 2021). Launched by NASA in 2018, GEDI is a high-resolution laser system installed onboard the International Space Station (ISS) (Dubayah et al., 2020a). Since March 2019, GEDI has been acquiring high quality 3D observations over non-contiguous 25 m circular footprints on the ground, between 51.6° North and South latitudes, which have proven highly relevant to the study in forest ecosystems on a global scale (Dubayah et al., 2020a; Qi et al., 2019). GEDI footprint geolocations are derived from GEDI's own Inertial Measurement Unit (IMU), Global Positioning Systems (GPS) and star tracker sensors onboard the ISS (Dubayah et al., 2020a; Beck et al., 2020; Luthcke et al., 2019). However, the ISS's low orbit, size and shape result in increased mechanical vibrations and greater variations in orientation and altitude than traditional Earth Observation satellites (Dou et al., 2014). Consequently, the horizontal position precision of GEDI footprints was expected at 10 m after calibration (Dubayah et al., 2020a). For GEDI products' first version (v1), released before in-flight calibration, the mean 1 σ horizontal geolocation error reached 23.8 m. After a calibration process accounting for geolocation biases, a second data (v2) version was released in April 2021 with a positioning error estimated at 10.2 m, with final targeted accuracy at 8 m (Beck et al., 2020, 2021). Assuming as in Roy et al. (2021) that GEDI geolocation errors follow a normal distribution N (μ = 0 m, σ = 10 m), 68.3, 78.9 and 95.4 % of the footprints would have a horizontal location error within 10, 12.5 and 20 m, respectively. Owing to footprint diameter on the ground (i.e., 25 m), more than 20 % of footprints overlap by less than 50 % with the expected footprint. This hampers the comparison and combination between GEDI data and other georeferenced data, such as field measurements and continuous remote sensing data, and therefore GEDI products' qualification and the development of models to predict vegetation attributes from GEDI data (Potapov et al., 2021; Saarela et al., 2016).

Recent studies assessed GEDI data quality to estimate ground elevation, canopy height and above-ground biomass (AGB) through comparison with aerial lidar system (ALS) data (Duncanson et al., 2020, 2021; Lang et al., 2022; Silva et al., 2021). GEDI was found to provide accurate ground elevation and canopy top heights measurements, although errors can reach up to several meters (Adam et al., 2020; Dorado-Roda et al., 2021; Guerra-Hernández and Pascual, 2021; Liu et al., 2021; Urbazaev et al., 2021). A significant part of errors was attributed to low horizontal accuracy (Potapov et al., 2021; Dubayah et al., 2020a; Roy et al., 2021; Lang et al., 2022; Adam et al., 2020; Urbazaev et al., 2021). Based on GEDI data simulations, Milenković et al. (2017) showed that AGB estimation errors increase with increasing geolocation error. The geolocation error has more impact in heterogeneous forests and in fragmented land-covers than in very homogeneous forests (Roy et al., 2021; Milenković et al., 2017). Slope and density of canopy cover have shown to influence GEDI estimations (Adam et al., 2020; Dorado-Roda et al., 2021; Liu et al., 2021; Wang et al., 2022; Quirós et al., 2021), but the link with geolocation error impact has not been tested in these studies. However, as geolocation errors in GEDI coordinates in slope terrain can result in larger elevation differences between the actual and provided coordinates than in flat terrain, it is reasonable to hypothesize that slope terrains will be more impacted by geolocation errors than flat ones. Improving the georeferencing is important and requires specific approaches. The most widespread geolocation improvement method uses ALS data to simulate GEDI-like waveforms around the original footprint location (Lang et al., 2022; Blair and Hofton, 1999; Hancock et al., 2019). The method processes by successive footprint clusters along individual ground tracks and a corrected geolocation is assigned where correlation between simulated and actual GEDI waveforms is

maximized (Lang et al., 2022; Hancock et al., 2019). Different studies used this approach to improve either v1 (Lang et al., 2022; Ilangakoon et al., 2021) or v2 (Liu et al., 2021) data. Lang et al. (2022) compared GEDI derived canopy heights with ALS heights, after geolocation correction, and obtained a 3.6 m root mean square error (RMSE) and a -0.3 m bias, while RMSE dropped to 2.7 m and bias to -0.1 m for 70 % most certain position predictions, i.e., highest correlations between real and simulated waveforms. Liu et al. (2021) compared ground elevation accuracy for v2 with and without geolocation correction and observed that improving geolocation led to a slight decrease in RMSE and in mean absolute error (MAE) of 0.12 m (4.15 m without and 4.03 m with correction) and 0.33 m (2.13 m without and 1.80 m with correction), respectively. Furthermore, Ni et al. (2021) provided a comparison for AGB models based on Relative Height (RH) metrics obtained from v1, v2 and from an optimized geolocation based on waveform matching of v1. When geolocation of v1 data was optimized, the determination coefficient (R^2) of the RH-based AGB model was sharply improved compared to v1 and slightly better than the one obtained with v2 data. Hancock et al. (2019)'s method has been primarily and successfully used to improve GEDI georeferencing. However, it requires waveform simulation from ALS data and is therefore limited to areas surveyed with ALS system, ideally at a time close to GEDI acquisitions. The method also requires downloading GEDI waveforms, a level 1 (L1) product that needs significant storage capacity and is not as user-friendly as higher-level products. To overcome these limitations, the aim of this study is to develop an alternative georeferencing method based on the hypothesis that ground elevation data from reference Digital Elevation Model (DEM) and GEDI level 2 (L2) ground elevation estimates are sufficient to improve the geolocation of GEDI footprints and to assess its performance. The approach, henceforth referred to as GeoGEDI, should benefit from high-resolution DEM increasing availability and temporal stability, thus enabling much broader use. GeoGEDI was tested on v1 and v2 data for different forest and terrain conditions. Its performance was evaluated, by analyzing magnitude and direction of the corrections and the impact on GEDI ground elevation and canopy height errors. The rest of the manuscript is organized as follows. Section 2.2 introduces the data used to test and evaluate GeoGEDI. In Section 2.3, GeoGEDI algorithm is detailed, prior to the presentation of the experimental set-up and statistical analyses. Results are reported and discussed in Sections 2.4 and 2.5, respectively.

2.2 Data

2.2.1 Study sites

Two contrasting French forest environments were considered, the Landes de Gascognes, or Landes' lowland forest, and the Vosges mountainous area. The Landes region is located in south-western France and cover the largest metropolitan French forest. The relief of the Landes is mainly flat, with elevations ranging from 0 to 200 m and mean slope of 2.6 % (± 4.7 %). Forests account for 74 % of the area and are almost entirely composed of maritime pine (Pinus pinaster Ait) plantations (IGN, Sylvoécorégion), with an average canopy cover of 45 % (± 23 %), measured at plot level by the National Forest Inventory. The Vosges site is located in north-eastern France and is much more heterogeneous in terms of topography and forest stands. It covers part of the Vosges forest and the Haguenau forest, a large lowland forest. Elevations range from 100 to 1200 m, with mean terrain slope of 17.8 % (± 17.0 %). Dominant species are European beech (Fagus sylvatica), silver fir (Abies alba) and Norway spruce (Picea abies) (IGN, Sylvoécorégion). The forest cover is dense with mean canopy cover of 78 % (± 21 %). Study sites were bounded by the extents of reference digital height

models (DHMs) (see Section 2.2.3). The Landes study site covers $14{,}051$ km^2 and the Vosges study site covers $6{,}264$ km^2. They will further be referred to as Landes and Vosges.

2.2.2 GEDI L2A data

The GEDI instrument is composed of three lasers emitting 14 ns long near-infrared laser pulses at high frequency (242 Hz). One laser is split into two coverage beams, while the other two lasers produce two full-power beams. Each beam is deflected every other shot by the Beam Dithering Units (BDUs), which results in eight parallel ground tracks. Tracks are spaced 600 m apart and composed of 25 m diameter circular footprints 60 m apart along-track. For each footprint, the lidar waveform backscattered by the Earth's surface is recorded (Dubayah et al., 2020a). The recorded waveforms are processed to provide GEDI data products at footprint level. In GEDI L2A products, ground elevation, top of canopy and relative canopy height (RH) metrics are derived from geolocated waveforms (L1B product). RHs correspond to cumulative waveform energy from bottom (0 %) to top (100 %), in 1 % increments (RH0 to RH100) (Hofton and Blair, 2019). GEDI L2A products over study sites were downloaded from NASA's archive center (Dubayah et al., 2020b, 2021a). A total of 30 and 15 orbits crossing Vosges and Landes sites, respectively, and for which both version 1 (v1) and 2 (v2) GEDI products are available, were selected. Acquisition dates range from May 2019 to May 2020. The latitude, longitude and elevation of the lowest mode (i.e., ground peak) were assimilated to footprint center coordinates and mean ground elevation within the area covered by the footprint, respectively. RH98 was used to assess the maximum height as suggested in Duncanson et al. (2021) and Blair and Hofton (1999). To avoid issues with poor quality data in forest environment, only full-power footprints with good quality flags were used, as recommended in Duncanson et al. (2021). After filtering, Landes and Vosges study sites were sampled with $73{,}280$ and $78{,}719$ footprints, respectively (total: $151{,}999$ footprints, Fig. 2.1).

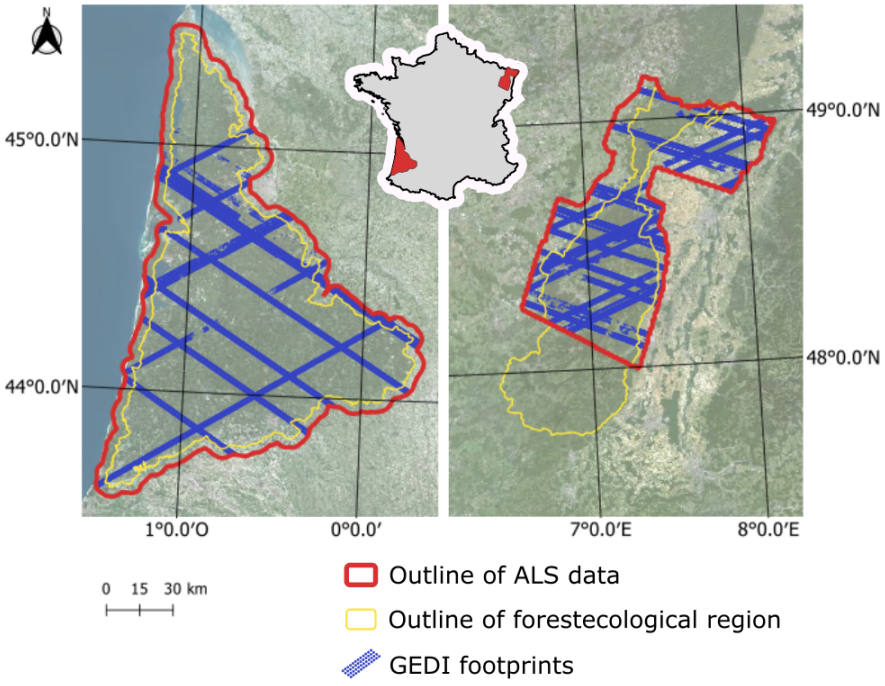


Figure 2.1: Overview of GEDI footprints (in blue) in the two study sites (in red): Landes (left) and Vosges (right). The Landes de Gascognes and Vosges Mountains forest's official ecological border are represented in yellow.

2.2.3 Reference datasets

High resolution DEM and DSM

DEMs at a spatial resolution of 1 m were downloaded from the BD ALTI® product of the National Institute of Geographic and Forest Information (IGN) (IGN, RGE ALTI). For both study sites, the DEMs were derived from ALS acquisitions and delivered with altimetric and planimetric mean quadratic errors within 0.2 m and 0.6 m respectively (IGN, RGE ALTI descriptif). Digital surface models (DSM) representing the top of canopy, top of buildings or other first return objects, were also acquired from IGN, at the same spatial resolution. They were generated from either photogrammetric or ALS point clouds. DSMs were chosen in order to have a minimal temporal acquisition difference with GEDI data. For the Landes, the chosen DSM was produced using a photogrammetric point cloud generated using aerial photographs acquired in summer 2018 at a 35 cm resolution and processed using MicMac dense matching algorithm (Rupnik et al., 2017). For the Vosges, the DSM was computed using ALS data acquired in winter 2020, and characterized by an average first return point density of 4.8 pt/m^2. On both sites, a digital height model (DHM) was obtained by subtracting ALS DEM from DSM. To allow for comparison with GEDI products, DEMref and DHMref, a 1-m resolution focal mean DEM and focal maximum DHM, were computed by using a sliding 25 m diameter circular window at each pixel.

Forest data base

BD Forêt® v2 (IGN, BD Forêt version 2) provides information about the composition and density for forest stands which have areas of greater than 5000 m^2. The open-source database was used to classify footprints as forest or non-forest.

The different datasets are summarized in Table 2.1.

Data	Coordinate System	Source	Processing
GEDI L2A footprints version 1 and version 2	WGS 84	NASA [49; 50]	Filtered on full-power beams, quality flag and availability of version 1 and version 2 Transformation to fit Lambert-93 coordinate system
Vosges DEMref	Lambert-93	IGN [114]	25 m focal mean of aerial lidar DEM
Landes DEMref	Lambert-93	IGN [114]	25 m focal mean of aerial lidar DEM
Vosges DHMref	Lambert-93	IGN	25 m focal maximum of aerial lidar DHM
Landes DHMref	Lambert-93	IGN	25 m focal maximum of photogrammetric DHM
BD Forêt v2	Lambert-93	IGN [111]	

Table 2.1: Data and sources

2.3 Method

In this Section, the GeoGEDI method is presented (in Section 2.3.1) and the experimental setup is designed (in Section 2.3.2). The latter includes parameter settings and filtering criteria used before analyzing algorithm outputs. The statistical analyses used to assess the algorithm performance are presented in Section 2.3.3. The official French coordinate system, Lambert 93, was used during all the processing steps and analyses. While all IGN datasets were given in Lambert 93, GEDI data had to be transformed from WGS84 to Lambert 93. GEDI's latitude and longitude coordinates were transformed to Lambert 93 coordinates and GEDI's ellipsoidal heights were transformed to fit Lambert-93 altitude system by applying an altimetric conversion grid (IGN, RAF18).

2.3.1 GeoGEDI algorithm

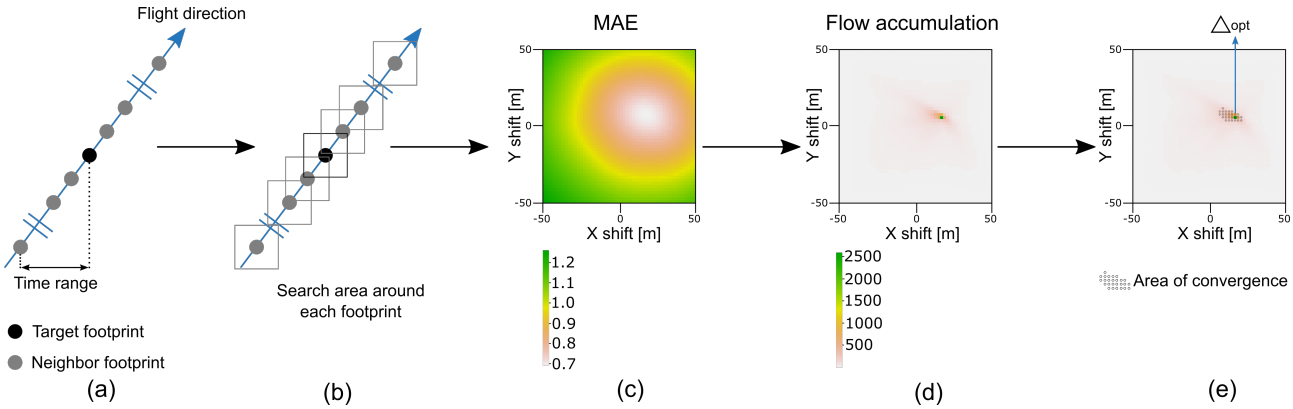


Figure 2.2: (a) and (b) Processing of a given footprint with its neighborhood. (c) Computation of mean absolute error map (MAE). (d) Error flow accumulation. (e) Computation of the optimal position from filtered accumulations barycenter.

GeoGEDI aims to match GEDI ground elevations to a reference DEM. Therefore, two inputs are needed: (1) GEDI footprint positions and ground elevations, and (2) DEMref. Each footprint F_i (with i ranging from 1 to the total number of footprints in the study area) is processed independently. However, co-registration relies on footprints clusters (Fig. 2.2). For each footprint F_i, the cluster C_i is made of n_i footprints acquired in a short time interval ($\delta_t ime$) centered on F_i acquisition time. ISS structural vibration frequency is estimated between 0.1 and 1 Hz (Nelson, 1994; Brown and Engelmann, 2019), which is lower than the GEDI laser emission frequency (242 Hz). Consequently, it can be assumed that position errors of footprints belonging to a small cluster Ci are temporally correlated. During the small amount of time considered for a cluster, the pointing deviations due to ISS movements and vibrations will be similar in direction and magnitude. The lasers will not be randomly pointing in different directions and the cluster mean shift can be used to correct the position of F_i. C_i's optimal position (Δ_{opt}) is searched within a maximal distance of $\pm\, shift_{max}$ (m) in X and Y and with a shift step (δ_{shift}) defined as a multiple of the DEMref resolution (i.e., $k \times r$, with $k \in \mathbb{N}^*$ and r, the resolution of DEMref (i.e. 1 m here)). This results in a $(2 \times shift_{max} + 1)$ wide squared area for the search and in a set of $N_{shift} = ((2 \times shift_{max}/\delta_{shift}) + 1)^2$ positions tested for each footprint. The values selected for $shift_{max}$ and δ_{shift} are presented in Section 2.3.2 focusing on GeoGEDI parametrization. At each

tested position, the Mean Absolute Error (MAEp eq. (1)) between C_i footprint elevations and the underlying DEMref elevations, is computed as follows:

$$MAE_p = \frac{1}{n_i} \sum_{p=1}^{n_i} |z_p - \hat{z}_p| = \frac{1}{n_i} \sum_{p=1}^{n_i} |dz_p| \tag{1}$$

where:

n_i = number of footprints
z_p = DEMref values
$\hat{z}_p$ = GEDI ground elevation
dz_p = difference between z_p and $\hat{z}_p$.

Each MAE_p value is associated to its specific shift in X and Y from the initial footprint position, resulting in a 2D MAE_i map providing a description of spatial error distribution according to shifts (Fig. 2.2(c)). The best shift Δ_{opt}, is computed from the MAEi map using a two-step procedure. First, a divergent flow accumulation algorithm is applied to the MAEi map (Fig. 2.2(d)). The FD8 flow accumulation algorithm Freeman (1991) was used (whitebox R package (Wu and Brown, 2022; Lindsay, 2016)) – a multidirectional flow algorithm commonly used to identify catchment areas and analyze drainage patterns in hydrological studies from raster DEMs. Unlike unidirectional algorithms, multidirectional flow algorithms allow flow dispersion and suit better in flat areas, while results between both types of algorithms are similar in the presence of slope (Schindewolf et al., 2015; Heung et al., 2013). From each DEM cell, the flow is distributed towards the downslope neighboring cells according to proportions depending on the difference in elevation between the starting cell and its neighboring cells, i.e. the higher the difference, the higher the proportion (Schindewolf et al., 2015; Heung et al., 2013; Quinn et al., 1991). The computation continues across grid cells until no more neighboring lower cell is encountered, i.e., once the flow has reached its catchment area. The final highest scores identify cells where flows most often stopped. When applied to the MAEi map, flow accumulation leads to the point with the lowest error. Cells with highest scores highlight the areas corresponding to the shifts minimizing differences between DEMref and GEDI ground elevations. Second step: computing Δ_{opt} from the flow accumulation map. First, a convergence area is defined by selecting a given percentage of cells having the highest accumulation flow values. Then, Δ_{opt} is defined as selected cells' barycenter and computed as the average coordinates weighted by flow accumulation values. The approach integrates information from the entire error map and is relevant to address situations with no clear identified minima, for example when several cells exhibited the same or similar maximum scores.

2.3.2 Experimental Setup

GeoGEDI algorithm's parameter settings

Considering the positional accuracy of GEDI v1 provided in Beck et al. (2021), we used 50 m as a reasonable upper shift limit ($shift_{max}$). Even though the DEMref spatial resolution was 1 m, δ_{shift} was set to 2 m for computational efficiency. This results in $N_{shift} = 2601$ tested positions for each footprint. The convergence area was defined as the 1% cells having the highest accumulation flow value. This choice resulted from an experimental trade-off to include enough pixels to describe the convergence area while limiting the selection of

secondary convergence areas pixels. GEDI laser units are fixed at different positions, with slight orientation differences, and each has its own depointing capacity, resulting in different viewing angles. Consequently, GeoGEDI should theoretically be applied to a cluster of footprints belonging to the same beam track, thus aligned on the ground. However, matching elevations along a single direction could be suboptimal for a robust footprint position adjustment. To overcome this limitation, GeoGEDI can be applied to a cluster including several beam tracks. To analyze the pros and cons of giving priority to the logic of acquisition geometry or 2D spatial distribution of points when co-registering GEDI data and DEMref, GeoGEDI was applied by track or considering the four full-power beam tracks together, using the same time interval (δ_{time}). Selecting δ_{time} lower than the period of structural vibration of the ISS (0.1 to 1 Hz) is recommended. After testing several time intervals, δ_{time} was set to ±0.215 seconds to select a sufficient number of footprints for the adjustment, while avoiding large changes in shifts inside the cluster. This δ_{time} corresponds to a 3-km distance along a track and to 50 and 200 footprints for the single-beam and four-beam approach, respectively. GeoGEDI was initially designed for GEDI v1 release. It was also applied to v2 data to demonstrate its potential for later releases with an improved geolocation. We hypothesize that the algorithm will also improve the later version, as NASA v2 products are said to be corrected for biases only, while GeoGEDI is supposed to improve the precision, i.e. to correct for non-systematic errors due to ISS vibrations, in addition to correcting biases. For each of the 151 999 footprints, GeoGEDI was applied with four configurations. The different GeoGEDI outputs based on v1 or v2, using either the single-beam or four-beam approach, will be referred to as v1_1, v1_4, v2_1 and v2_4.

Data filtering

Once the shifts were computed, several filters were applied. First, footprints associated to too small clusters were discarded. Indeed, cluster size (n_i) can be lowered due to removing low quality footprints (see Section 2.2.2). Threshold value was set to 1/4 of the theoretical maximum number of footprints for the considered time interval, corresponding to 13 and 50 footprints for the single-beam and four-beam approaches, respectively. All footprints that did not meet one of the above-mentioned criteria, with either v1 or v2 dataset, were excluded. From the 151 999 footprints, 150 093 were kept for further analysis. Second, in each dataset i.e., v1_1, v1_4, v2_1 or v2_4, footprints where the shift in X or Y for Δ_{opt} reached $shift_{max}$ (i.e., 50 m) were discarded.

Finally, some footprints were discarded due to issues identified in GEDI ground elevation assessment. Six waveform interpretation algorithms (01 to 06) were defined by the GEDI science team to identify the ground peak from GEDI waveforms, with different thresholds and smoothing settings (Beck et al., 2021; Hofton and Blair, 2019). In GEDI L2A data v1 the default algorithm for all footprints was algorithm 01. In v2, the presumed best ground elevation is provided for each footprint along with the corresponding algorithm. This leads to possible changes in best algorithm choice and in differences in ground detection and elevation between the two GEDI versions. For v1, the default choice is always algorithm 01. For v2, the selected optimal algorithm was either 01 or 02 for our study sites. A comparison with DEMref revealed great ground elevation underestimation for some footprints where algorithm 02 was selected, probably due to faulty ground peak detection (Fig. 2.3 and A.1). To eliminate these misestimations in further analyses, footprints having a ground elevation difference between v1 and v2 of more than 1.5 m were discarded. This concerned 26.9 % and 39.3 % of footprints processed with algorithm 02, corresponding to 3.4 % and 13.8 %

of footprints total number, for Landes and Vosges, respectively.

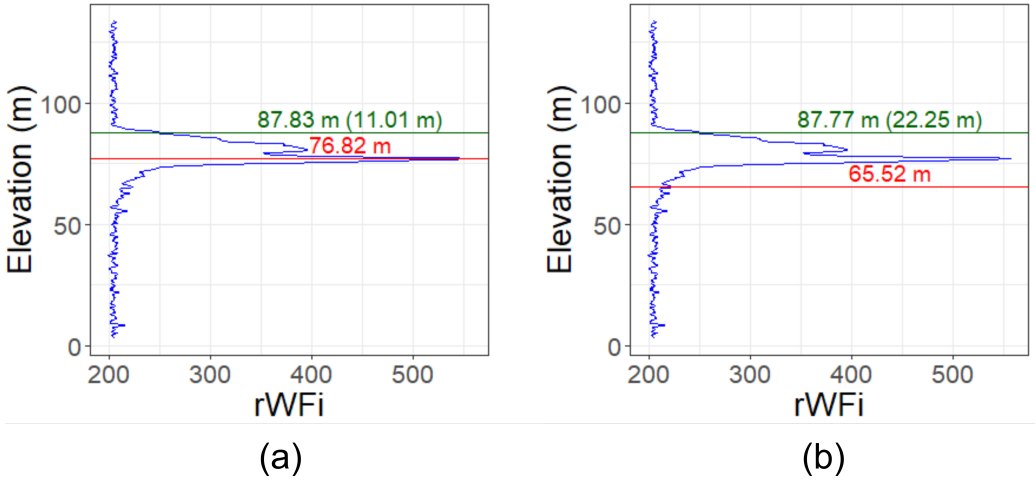


Figure 2.3: Example of a sorted-out GEDI footprint waveform of (a) v1 using algorithm 01 and (b) v2 using algorithm 02. Ground elevation of the variable 'lowest_mode' in red and RH98 transformed to surface elevation (and RH98) in green.

Please note: this source of error was identified after processing footprints; and the footprints were used during the georeferencing process. To limit influence of erroneous ground peak detection when comparing error estimates for different datasets, they were discarded regarding ground elevation and surface height estimation analyses.

2.3.3 Statistical analysis

Analyses were performed on the four GeoGEDI sets, i.e., v1_1, v1_4, v2_1, v2_4. Statistics regarding differences between NASA v1 and v2, further referred to as v1_v2 results, were also reported as baseline for discussion. As effective GEDI footprint positions are unknown, GeoGEDI's performance can only be evaluated indirectly: 1) shifts were analyzed and 2) ground elevation and surface height errors were compared before and after applying GeoGEDI.

GeoGEDI's shift analysis

As GeoGEDI is supposed to correct for geolocation errors, checking whether GeoGEDI positions tend to be in the same direction and shifts of the same magnitude than NASA's is a complementary source of algorithm assessment. Both shift magnitudes and directions were analyzed. In order to analyse mean shift directions while taking into account major differences in orientation between ascending and descending orbits as well as minor differences according to ISS's exact flight path, the coordinate system was changed. X and Y shifts, expressed according to West/East and South/North directions, were transformed into X_T and Y_T considering a coordinate system linked to the local orbit ground track direction. X_T axis follows the orientation of the orbit ground track (i.e. flight path direction relative to the West/East direction assessed by calculating the orientation of the track between the first and last footprint (of a same beam) of each orbit from v2 dataset) and Y_T axis is perpendicular to X_T, forming a local orthonormal coordinate system, centered on the initial footprint position (v1 or v2). Angular deviations can therefore be estimated when transforming new X_T and

Y_T to polar coordinates, i.e. the footprints Euclidean distance to the initial position (0;0) and the shift angle relative to the track direction (X_T).

First, shift magnitudes' mean, median and standard deviations were assessed. Then, mean relative shift distances and directions were used for dataset mean positions inter-comparison. The mean positions were also compared by beam, so as to identify possible beam-dependent behavior. Additionally, the temporal evolution of shift distances and directions was visually analyzed by plotting the positions of successive footprints belonging to the same orbit. For visual simplification, the temporal variability was illustrated for three datasets (v1, v2, v1_1). It was assumed that orbit segments over the study areas can be assimilated to a line, and compared footprint position spread along that line for different datasets. To define the reference track line (Fig. 2.4(a)), we used the first and last v2 footprint of each track. Footprint Euclidean distances to the line were calculated (Fig. 2.4(b)) and reported on the final figure (Fig. 2.4(c)). This highlights differences between ground tracks among the different datasets.

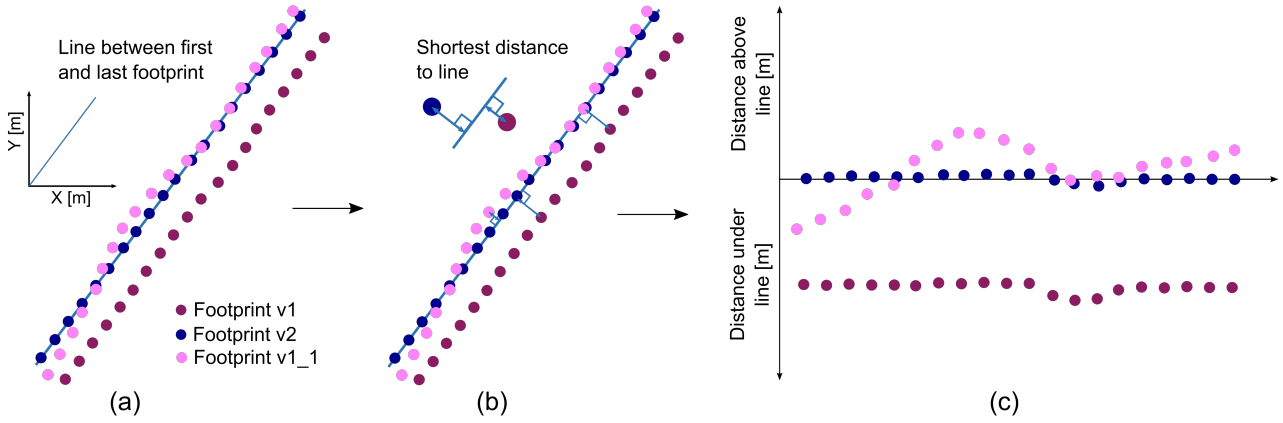


Figure 2.4: (a) Illustrating the ground tracks of GEDI footprints and defining a reference track line. (b) Calculating each footprint's distance to the reference track line. (c) Plotting these distances.

Elevations and heights qualification

GeoGEDI outputs are expected to improve the agreement between GEDI and reference elevations and heights. Therefore, ground elevation and surface height errors were analyzed. Ground elevation errors are expected to diminish as the algorithm is based on minimizing ground elevation errors. However, height errors analysis provides a fully independent evaluation of the algorithm performance. It consists in comparing GEDI RH98 data with DHMref. The evaluation relied on four standard metrics: MAE eq. (2), mean error (ME eq. (3)), error standard deviation (σ eq. (4)) and RMSE eq. (5). These metrics were computed for the six different datasets (v1, v2, v1_1, v1_4, v2_1 and v2_4).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|z_i - \hat{z}_i| = \frac{1}{n}\sum_{i=1}^{n}|dz_i| \tag{2}$$

$$ME = \frac{1}{n}\sum_{i=1}^{n}(z_i - \hat{z}_i) = \frac{1}{n}\sum_{i=1}^{n}dz_i \tag{3}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(dz_i - \overline{dz})^2}{n-1}} \tag{4}$$

$$RMSE = \sqrt{\sum_{i=1}^{n}(z_i - \hat{z}_i)^2} = \sqrt{\sum_{i=1}^{n} dz_i^2} = \sqrt{ME^2 + \sigma^2} \tag{5}$$

where:

n = number of footprints in the dataset
z_i = DEMref values
$\hat{z}_i$ = GEDI ground elevations
dz_i = difference between z_i and $\hat{z}_i$
$\overline{dz}$ = sample's mean difference between z_i and $\hat{z}_i$.

The same statistics were used for height estimations, replacing DEMref by DHMref and z by h.

For each footprint, available auxiliary information included: 1) the study site, 2) forest vs non-forest status, 3) shift magnitude 4) and a local slope indicator. The latter was defined as the ground elevation range at each GEDI footprint level, and was computed from the 1-m raster DEM using v1 footprint positions. Forest vs non forest status was established using both the forest map (see Section 2.2.3) and DHMref. All non-forest footprints of the forest map were assigned the "non-forest" class while forest footprints with a less than 2-m DHMref value were reclassified as "non-forest", in order to remove footprints acquired over clear-cuts or areas that changed from forest to agricultural land between the last forest map update and GEDI data acquisitions. Distributional metrics were compared for several datasets, defined based on auxiliary information. To evaluate the shift magnitude influence, footprints were divided into five classes based on quantiles of shift magnitude distribution, resulting in an equal number of footprints per classes. Classes were noted C_{Q1}, C_{Q2}, C_{Q3}, C_{Q4} and C_{Q5}.

2.4 Results

2.4.1 Shift magnitudes and directions

Table 2.2 shows GeoGEDI shift statistics for the different approaches. For GEDI v1-based approaches, mean shift values were similar across sites and ranged from 23.55 m (v1_1 Vosges) to 23.95 m (v1_4 Landes). Standard deviations proved higher for the Landes, ranging from 9.32 m (v1_4 Vosges) to 14.70 m (v1_1 Landes). As expected, shifts were of lower magnitude for v2-based than for v1-based approaches. For Vosges, mean values were divided by more than two while standard deviations were more stable (10.85 m (± 8.61 m) and 11.84 m (± 9.45 m) for v2_4 and v2_1, respectively). For Landes, the shift magnitudes reduction was reflected by a decrease in medians by at least 4 m rather than by changes in mean and standard deviation underlying the possible presence of outliers in shift distributions. Moreover, mean shifts between v1 and v2 obtained by NASA (v1_v2) were 17.80 m (± 4.52 m) for Landes and 20.60 m (± 3.88 m) for Vosges.

	Landes			Vosges		
	Mean	**Med**	σ	**Mean**	**Med**	σ
v1_1	23.88	20.59	14.70	23.55	22.80	10.07
v1_4	23.95	21.63	13.46	23.64	23.32	9.32
v2_1	22.19	16.12	16.62	11.84	8.94	9.45
v2_4	20.48	14.42	16.72	10.85	8.25	8.61
v1_v2	17.80	17.18	4.52	20.60	20.51	3.88

Table 2.2: Mean, median (Med) and standard deviation (σ) of differences between GeoGEDI and corresponding NASA coordinates

Fig. 2.5 illustrates the relative average positions in X_T and Y_T between the different datasets. As a visual convention, the position of v1 was used as the coordinate system's origin (0;0). In average, all corrections led to positions characterized by both a similar direction (83 to 93°) and magnitude (14.69 to 19.83 m). The NASA correction led to an average position at a distance of 17.59 m from v1 position and in a direction of 93.00° with respect to v1 track direction. Average positions obtained using GeoGEDI on v1 showed distances of 14.69 m and 16.79 m in directions of 85.31° and 87.33° for v1_1 and v1_4, respectively. Average GeoGEDI corrected positions v2_1 and v2_4 are very close to each other, with 19.17 m and 19.84 m at 83.63° and 83.33°, respectively, from v1 positions. If only Vosges footprints are taken into account, all four GeoGEDI average positions, i.e. v1_1, v1_4, v2_1 and v2_4, are grouped within 19.19 m and 22.88 m distances and 86.78° and 87.95° directions.

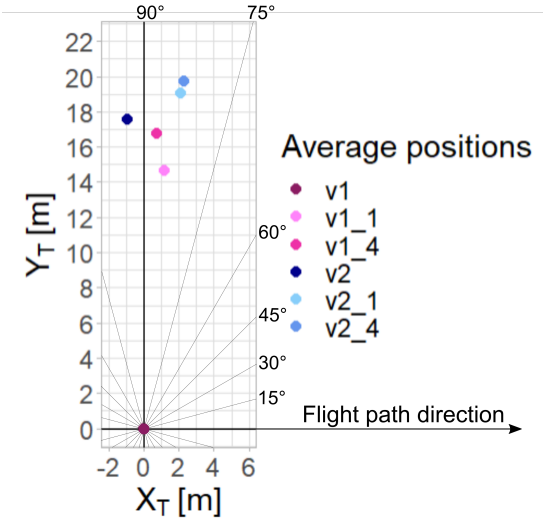


Figure 2.5: Average relative positions between the different approaches (including both study sites). The flight path direction is used as X axis and GEDI v1 is used as coordinate axis origin.

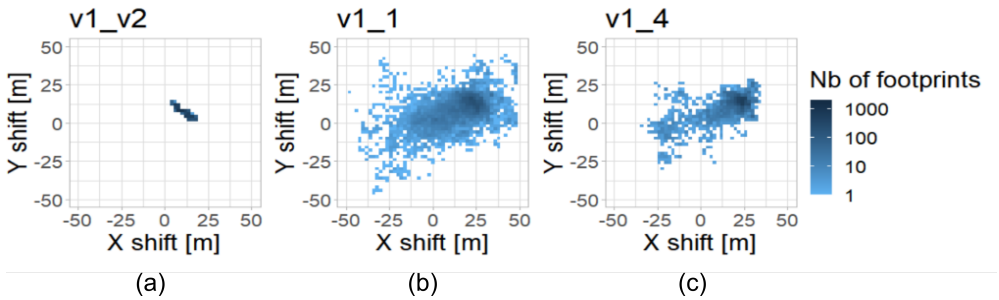


Figure 2.6: All individual shifts applied to footprints from orbit 3144 intersecting Landes. (a) v2 compared to v1. (b) Single-beam approach on v1. (c) Four-beam approach on v1. The original latitude/longitude oriented coordinate system is used for this illustration.

Although footprints average corrected positions were relatively close to each other, there were notable differences among experimental setups. Fig. 2.6 illustrates the spread in shift distributions for an example orbit. Fig. 2.6, unlike Fig. 2.5 and Fig. 2.7, is presenting the applied shifts in X, Y coordinate system, i.e. following the usual West/East and South/North axis, in order to illustrate the shifts with regards to the search window. NASA's shifts (i.e. v1_v2) are concentrated around the mean value with a 18.37 m maximum shift and mean and standard deviation shift magnitude of 13.69 m (± 1.59 m). Shifts are more spread for v1_1 and v1_4 with means (± σ) of 22.96 m (± 9.45 m) and 25.1 m (± 6.39 m), respectively. The global trend in shift corresponds to the bias correction, while the dispersion in shifts around this trend corresponds to the correction of the non-systematic error component and results in an increase in precision.

GeoGEDI average positions according to beam configurations are provided in Fig. 2.7. Mean shifts perpendicular to the flight axis were quite similar whatever the beam and approach, while shifts parallel to the flight path showed greater variations according to the beam and emitting laser. Beams acquired by the same laser, i.e., beams 0101 and 0110, and beams 1000 and 1011, respectively, exhibit similar shifts. For v1_v2, intra-beam pair distances were 1.29 m and 2.83 m for beam pairs (0101, 0110) and (1000, 1011), respectively, and mean distances between the two beam pairs ranged between 7.86 and 11.81 m. Beams 0101 and 0110 were rotated by 107.02° and 111.14° from v1 track direction while beams 1000 and 1011 were rotated by 74.60° and 81.63°.

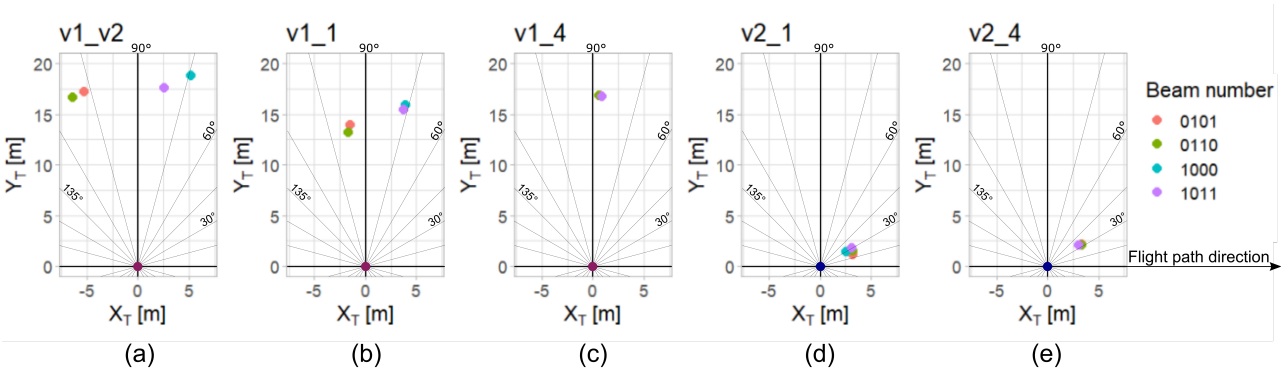


Figure 2.7: Average relative positions by beam, for all footprints between GeoGEDI and corresponding NASA coordinates. (a) v2 compared to v1. (b) v1_1 compared to v1. (c) v1_4 compared to v1. (d) v2_1 compared to v2. (e) v2_4 compared to v2. The flight path direction is used as X axis and GEDI v1 positions were used as coordinate axis origin for (a), (b), (c) and GEDI v2 for (d), (e).

Similar results were obtained for v1_1, with intra-beam pair 0.80 m and 0.46 m distances, respectively, and inter-beam pair distances from 5.49 to 6.21 m. The angles obtained by beam pairs were very close to each other with 96.24° and 97.10° for the first pair, opposed to 76.06° and 76.36° for the second pair. As expected, mean shifts were grouped together using the four-beam approach (Fig. 2.7(c)) with mean positions being 0.07 to 0.28 m apart. The beam pairs are no longer standing out for v2_1 (Fig. 2.7(d)). For v2_1 intra-beam pair distances were 0.30 m and 0.70 m and inter-beam pair distances ranged between 0.38 and 0.66 m. Rotation angles were between 20.39 and 30.46°. For v2_4 average positions are also grouped together, with inter-beam distances ranging from 0.04 to 0.36 m at a maximum distance of 4.00 m from the original v2 position and angles ranging from 31.87 to 34.74°.

Fig. 2.8 illustrates GeoGEDI positions' temporal evolution for an orbit segment and highlights differences between ground tracks corresponding to the various datasets. V1 and v2 tracks are nearly parallel, which

translates the bias correction announced by NASA. Tracks obtained with GeoGEDI wobble around v2 tracks, and may vary quickly over time, as illustrated in Fig. 2.8. In Landes, local variations are greater than in Vosges. Within only three kilometers, the v1_1 track can deviate by more than 50 m from the reference track line (Fig. 2.8(c)).

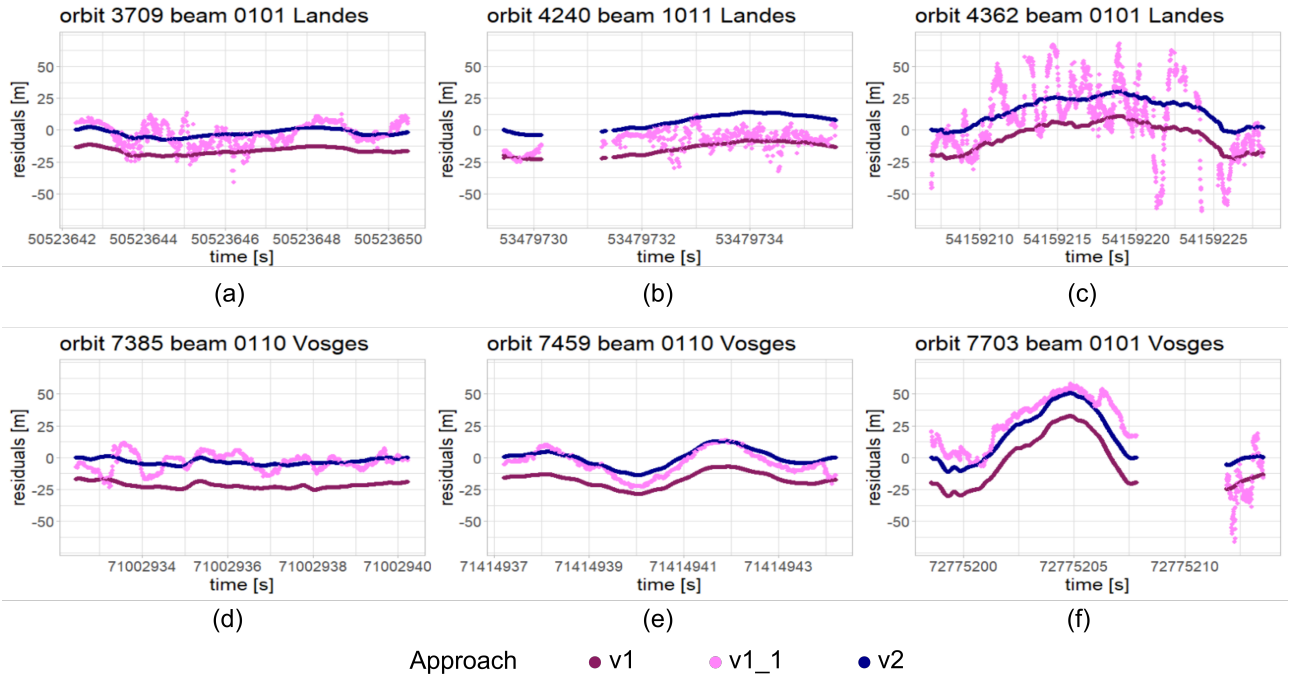


Figure 2.8: Temporal variability of v1, v2 and v1_1 ground tracks for three orbits in Landes (a, b, c) and in Vosges (d, e, f). A reference track line was defined between first and last v2 footprints and plots show the distance of footprints to the reference line. Time corresponds to the delta_time variable of GEDI footprints.

2.4.2 Impact of GeoGEDI corrections on ground elevation and surface height estimates

Next, the differences between DEMref and GEDI ground elevations and between DHMref and GEDI RH98 are referred to as dz ($Z_{DEMref} - Z_{GEDI}$) and dh ($H_{DHMref} - H_{GEDI}$), respectively.

Evaluation of ground elevation and surface height for forest and non-forest areas

Table 2.3 shows ground elevation errors for study sites, by land use (i.e. forest and non-forest). Overall, GEDI overestimated ground elevations. The smallest (-0.2 m) and greatest (-0.63 m) overestimations were observed in the Vosges site, for v1_1 forest footprints and v2 non-forest footprints, respectively. For both land uses, both study sites and both GEDI versions, GeoGEDI outputs systematically decreased ground elevation errors compared to NASA's versions. For Vosges, RMSEs were decreased by 59.6 % and 58.3 % for v1 and by 36.2 % and 30.0 % for v2, for forest and non-forest footprints, respectively. For Landes, RMSEs were decreased by 26.8 % and 28.8 % for v1 and by 13.3 % and 13.3 % for v2, for forest and non-forest footprints, respectively. Best results were achieved with single-beam adjustment. The lowest RMSEs were achieved with v2_1, with 0.91 m and 2.49 m for forest and 0.78 m and 0.98 m for non-forest areas, for Landes and Vosges, respectively. Interestingly, the standard deviations were much smaller for Landes (0.69 – 1.22 m range) than for Vosges (0.87 – 6.40 m range).

	Landes						Vosges					
	Forest (n ≈ 43 900)			Non-Forest (n ≈ 24 000)			Forest (n ≈ 36 800)			Non-Forest (n ≈ 29 500)		
	ME	σ	**RMSE**	**ME**	σ	**RMSE**	**ME**	σ	**RMSE**	**ME**	σ	**RMSE**
v1	-0.36	1.22	1.27	-0.30	1.06	1.11	-0.36	6.40	6.41	-0.50	2.37	2.42
v1_1	**-0.28**	**0.89**	**0.93**	**-0.23**	**0.76**	**0.79**	**-0.20**	**2.58**	**2.59**	**-0.33**	**0.96**	**1.01**
v1_4	-0.30	0.91	0.96	-0.23	0.77	0.81	-0.33	2.86	2.87	-0.42	1.05	1.13
v2	-0.50	0.92	1.05	-0.46	0.77	0.90	-0.48	3.87	3.90	-0.63	1.26	1.40
v2_1	**-0.41**	**0.82**	**0.91**	**-0.37**	**0.69**	**0.78**	**-0.42**	**2.46**	**2.49**	**-0.46**	**0.87**	**0.98**
v2_4	-0.43	0.85	0.95	-0.38	0.69	0.79	-0.43	2.53	2.56	-0.54	0.88	1.04

Table 2.3: Ground elevation errors for all six datasets for forest and non-forest footprints. Best results for v1 and v2-based approaches are highlighted in bold.

Surface height results are presented in Table 2.4. Overall, GEDI heights were closer to reference heights at v2 positions than at v1: ME, σ and RMSE all decreased. The greatest height assessment improvements were achieved with the four-beam approach, except for v2 in Landes; there, GeoGEDI brought no improvement. For Vosges, slightly better performances were observed with v2-based approaches than with v1-based ones. In both sites, mean heights were underestimated for forest footprints – ME ranging from 0.54 to 0.76 m for Landes and from 2.38 to 2.69 m for Vosges – and overestimated for non-forest footprints – ME ranging from -1.12 to -1.41 m for Landes and from -0.84 to -1.10 m for Vosges. RMSEs were similar for both land uses, with values ranging from 4.19 (v2, non-forest) to 5.25 m (v1, forest) and from 5.99 (v2_4, forest) to 7.58 m (v1, non-forest), for Landes and Vosges, respectively. Overall, in Vosges, RMSEs were lower for forest footprints than for non-forest footprints. Opposite results were observed in Landes. As both set-ups (single-beam and four-beam) gave similar results, only single-beam results are reported in Sections 2.4.2 and 2.4.2.

	Landes						Vosges					
	Forest (n ≈ 43 900)			Non-Forest (n ≈ 24 000)			Forest (n ≈ 36 800)			Non-Forest (n ≈ 29 500)		
	ME	σ	**RMSE**	**ME**	σ	**RMSE**	**ME**	σ	**RMSE**	**ME**	σ	**RMSE**
v1	0.76	5.19	5.25	-1.41	4.81	5.01	2.69	6.98	7.48	-1.10	7.50	7.58
v1_1	0.68	4.93	4.98	-1.34	4.49	4.68	2.44	5.65	6.16	-0.86	6.60	6.65
v1_4	**0.64**	**4.69**	**4.74**	**-1.28**	**4.28**	**4.47**	**2.43**	**5.59**	**6.09**	**-0.84**	**6.56**	**6.61**
v2	**0.54**	**4.45**	**4.48**	**-1.12**	**4.04**	**4.19**	2.41	5.82	6.30	-0.87	6.59	6.64
v2_1	0.69	4.94	4.99	-1.35	4.58	4.77	2.38	5.62	6.10	**-0.84**	6.55	6.60
v2_4	0.69	4.76	4.81	-1.29	4.37	4.55	**2.38**	**5.49**	**5.99**	-0.85	**6.44**	**6.50**

Table 2.4: Surface height errors for all six datasets for forest and non-forest footprints. Best results for v1 and v2-based approaches are highlighted in bold.

Shift magnitudes influence

GeoGEDI shift magnitudes impact on ground elevation and height estimates was considered, to evaluate whether large shifts were justified or artifacts. Fig. 2.9 compares dz distributions between v1 and v1_1 (Fig. 2.9(a)) and between v2 and v2_1 (Fig. 2.9(b)) for five shift magnitude classes (see 2.3.3). The improvement in

ground elevation accuracy increases with shift magnitude increase. For v1_1, RMSEs were lowered by 18.8, 39.0, 54.6, 62.0, and 68.1 % for classes C_{Q1}, C_{Q2}, C_{Q3}, C_{Q4} and C_{Q5}, respectively. The same trend, with a decrease in precision and an improvement in bias (Fig. 2.9(b)), was observed for v2, although improvements in accuracy were less pronounced. For v2, ground elevation RMSEs were respectively improved by 5.7, 18.5, 27.5, 39.9 and 61.0 %. As already noticed, shifts applied to v2 were much smaller than those applied to v1 (see class limits, Fig. 2.9). For v1_1, 20 % of footprints were shifted by less than 12.8 m, while for v2, this quantile limit was 6 m. Regarding surface heights (Fig. 2.9(c) and 2.9(d)), compared to v1, v1_1 RMSEs decreased by 4.2, 11.8, 14.7, 20.6 and 6.4 % for classes C_{Q1} to C_{Q5}. Like for ground elevations, the further the shift, the more important the improvement in height estimates in the first four classes. However, RMSEs did not continue to improve for C_{Q5}. Compared to v2, v2_1 height RMSEs were slightly improved by a maximum of 3.2 % for the smallest shift distances (C_{Q1}, C_{Q2}, C_{Q3}). But RMSE improved by only 0.6 % for class C_{Q4}, and even deteriorated by 24 % – from 4.55 to 5.64 m – for footprints belonging to C_{Q5} (Fig. 2.9(d)). Note that C_{Q5} is mainly composed of Landes footprints (79 % Landes against 21 % Vosges).

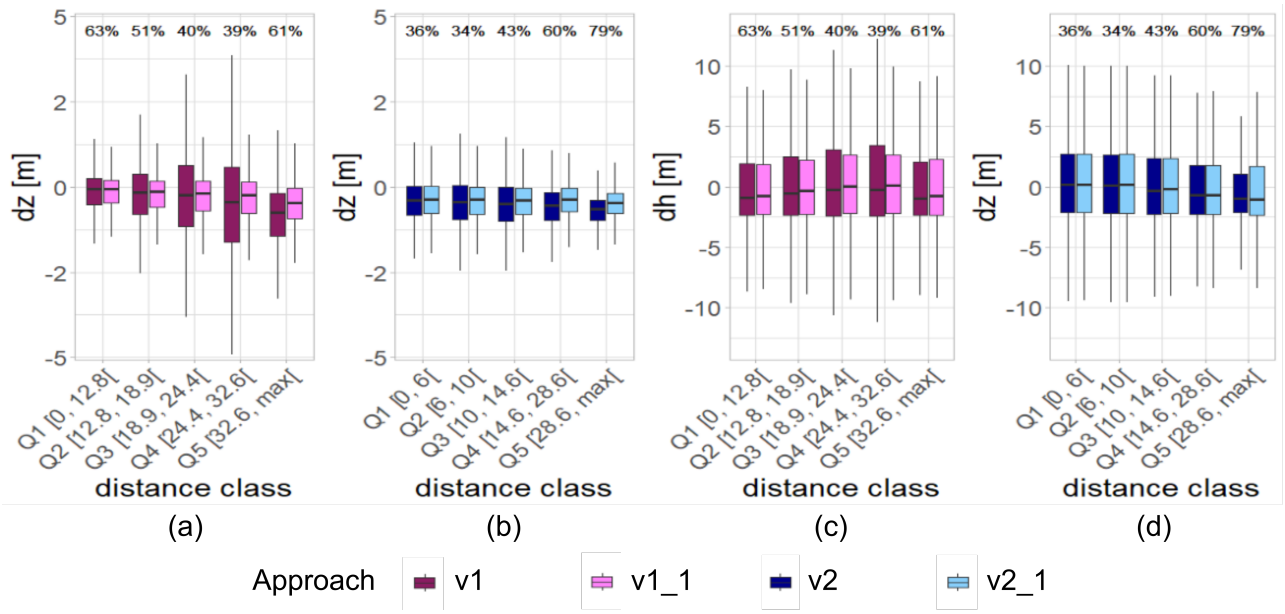


Figure 2.9: (a), (b) Ground elevation errors (dz). (c), (d) Surface height errors (dh) for v1, v1_1, v2 and v2_1 approaches, by footprints shift magnitudes quantiles of v1_1 or v2_1 distances to the initial GEDI version (v1 or v2). Distances are given in meters, e.g. class Q1 for v1 includes all footprints which were moved by 0 to 12.8 m when applying v1_1 GeoGEDI algorithm. The percentage above each class indicates the part of footprints belonging to Landes study site. Remaining footprints belong to Vosges.

Influence of the slope

In sloped terrain, a small error in geolocation results in large ground elevation errors. As expected, the higher the slope indicator, the higher the errors in ground elevations (Fig. 2.10(a) and 2.10(b)). For example, v1 ground elevation RMSEs were 0.98, 1.70, 2.87, 4.65 and 9.05 m for the five slope classes reported in Fig. 2.10. Moreover, the higher the slope indicator, the greater the improvement brought by GeoGEDI, and, compared to v1, v1_1 ground elevation RMSEs were improved by 9.7, 31.3, 48.2, 59.2 and 63.4 %, for classes C_1, C_2, C_3, C_4 and C_5, respectively. Similar results were obtained for v2_1 regarding v2, with improvements of 5.7, 12.7, 21.4, 31.7 and 41.7 % for all five slope classes. The slope effect on height estimates is illustrated in

Fig. 2.10(c) and 2.10(d). For all datasets, the smaller the slope, the better the estimate. For v1, GeoGEDI outputs improved the flattest footprints' height accuracy by 4.9 %. For the other four classes, height RMSEs decreased between 18.1 and 20.8 %. For v2_1, height RMSE increased by 9.6 % for footprints with no slope (C_1) and height RMSE was improved by 1.7 % for footprints with low slope (C_2). Concerning footprints with greater slope (C_3, C_4, C_5), height RMSEs were improved by 5.1, 6.0 and 5.7 %, respectively.

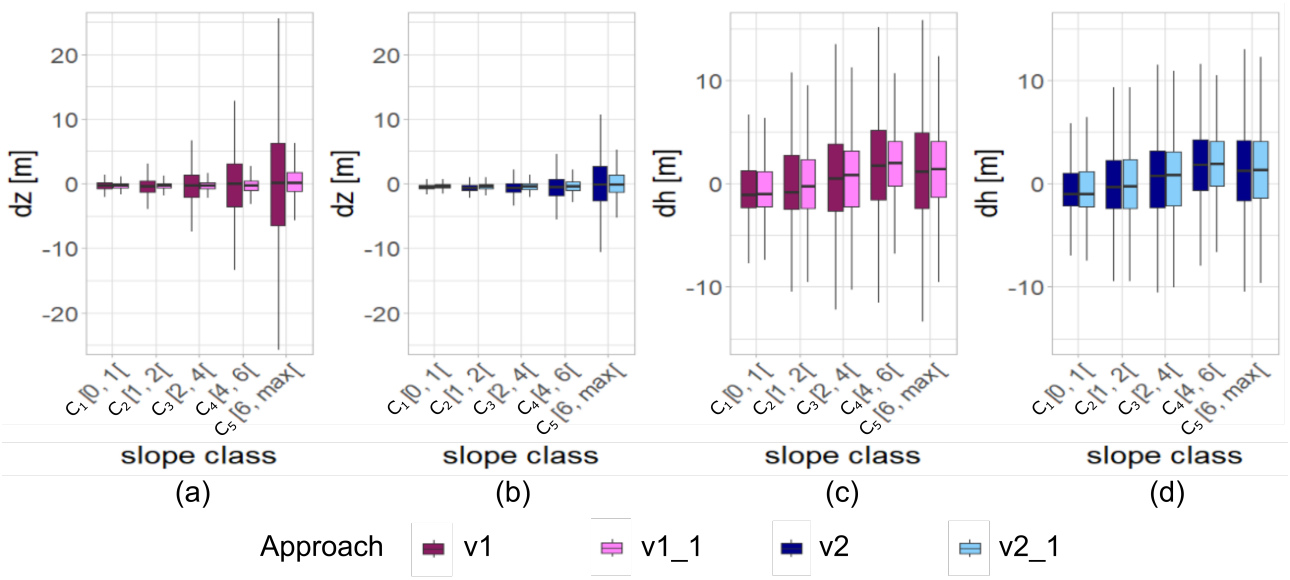


Figure 2.10: (a), (b) Ground elevation errors (dz). (c), (d) surface height errors (dh) for v1, v1_1, v2 and v2_1 approaches, by footprints slope indicator. The slope indicator corresponds to the elevation range in the 25 m circular footprint and is given in meters.

2.5 Discussion

2.5.1 Shift analyses corroborate GeoGEDI's efficiency

GeoGEDI-based mean shifts were in accordance with horizontal geolocation errors announced by NASA's user guide (Beck et al., 2021). Logically, shifts obtained with v1-based approaches were greater than those obtained with v2-based approaches (Table 2.2). Beck et al. (2021) studied GEDI geolocation error over a 30-week time-span. The mean of weekly computed 1 σ errors was 23.8 m with a substantial bias (Beck et al., 2020), and 10.2 m with a limited bias for v1 and v2, respectively. GeoGEDI's v1-based mean shift distances were within this range, with 23.55 to 23.95 m mean shifts. For v2-based approaches, GeoGEDI results varied among sites. Mean shifts for Vosges were close to the 10.2 m geolocation error announced by NASA, with values of 11.84 m (v2_1) and 10.85 m (v2_4). Mean shifts for Landes reached 22.2 m and 20.5 m and were therefore close to the 2 σ mean error announced by NASA, challenging GeoGEDI outputs over large flat areas (see Section 2.5.2). Nevertheless, all GeoGEDI positions converged towards v2 mean positions resulting from the in-flight NASA-operated calibration (Fig. 2.5), thus providing additional GeoGEDI robustness validation. The on v1 applied corrected angles all rotated towards the same direction, above and perpendicular to the flight path direction. Compared to initial v1 positions, v2 positions were moved in a direction of 93° with respect to v1 track direction and GeoGEDI v1_1 and v1_4 in a direction of 85° and 87°, respectively. This is in line with the direction found by Quirós et al. (2021). In order to correct the positions for geolocation

bias, they tested eight directions (0°, 45°, 90°, 135°, 180°, 225°, 270° and 315°) at two distances (5 m and 10 m) from the initial v1 position and defined the best fitting position for each footprint based on the lowest RMSE between GEDI elevation and aerial lidar DEM. Among the 17 tested positions (e.g. eight directions at two distances and the central initial position), the best fitting position was at 10 m and 270° clockwise, corresponding to a 90° angle above the flight path (i.e. standard counter-clockwise angle measurement used in this study). 31.88 % of their footprints had the lowest RMSE for this position.

Moreover, GeoGEDI per beam results complied with theoretical expectations. When the single-beam approach was applied to v1, resulting mean positions were paired according to laser units. Nevertheless, mean positions also exhibited small differences, possibly arising from the difference in pointing direction between the two beams of a beam pair emitted by the same laser unit. Compared to initial v1 positions, v2 positions were rotated by 78° for one beam pair and by 109° for the other pair. GeoGEDI v1_1 average positions were rotated by 76° for beam pairs 1000 and 1011, while beams 0101 and 0110 were rotated by 97°. When applied to v2, GeoGEDI mean shifts were almost identical for all beams regardless of the laser unit, confirming NASA's biases correction on v2 products. Mean positions of v2_1 and v2_4 corrected initial v2 positions with 20 to 35° angles. Finally, we assumed GeoGEDI could correct for geolocation source inaccuracy that cannot be handled from ISS-borne sensors and in-flight calibrations, such as ISS structural vibrations. Beyond the trends provided by mean shifts, the quick shifts temporal changes and their magnitude are worth noticing. For both, the single-beam and the four-beam approaches, two consecutive footprints could have significantly different shifts with respective clusters differing by few footprints. Yet, shift values followed a relatively continuous pattern (Fig. 2.8). This continuity is important, as it is key for our assumption validity, i.e.: using footprints acquired in a shorter time interval than the highest vibration period captures these vibrations impact on the geolocation error. Resulting GeoGEDI tracks have more variable and less "smoothed" track patterns than those observed in NASA footprint positions, highlighting that GeoGEDI succeeded in capturing part of ISS high frequency variations. As a result, computed shifts were observed as spatially correlated (shift continuity). However, we are aware that GeoGEDI tracks are probably still slightly smoothed compared to real tracks, as footprints were corrected for local mean deviations.

2.5.2 GeoGEDI advantages and limitations

The proposed georeferencing method proved efficient and robust for a diversity of environments (Section 2.5.2), even if some limitations (Section 2.5.2) and possible ways of improvements (Section 2.5.2) were identified.

GeoGEDI main strengths

One of GeoGEDI's major assets is it needs only two inputs: 1) coordinates and 'lowest_mode' variable from GEDI L2A footprints and 2) a high-resolution DEM, which are increasingly available worldwide. Additionally, it is simpler than methods based on waveform correlation between GEDI and ALS simulations (Hancock et al., 2019). Results indicated that GeoGEDI greatly improved consistency in ground elevation between GEDI and DEMref (see Section 2.4.2). Height estimates were also improved for most cases, except for v2-based approaches in Landes (see Section 2.5.2). Consistency between GEDI estimations and reference values proved considerably improved in sloped areas where even small geolocation error can lead to high discrepancy.

Note that GeoGEDI results are in the same range as Hancocks' waveform matching approach. After correcting v1 for geolocation, Ilangakoon et al. (2021) and Lang et al. (2022) observed 4.69 m and 3.6 m GEDI surface heights RMSE for their study sites, respectively, while v1-based GeoGEDI reached 4.47 to 6.65 m RMSEs. For ground elevations, after correcting v2, Liu et al. (2021) observed a 4.03 m RMSE value, while GeoGEDIs range from 0.79 (non-forest, Landes) to 2.59 m (forest, Vosges). Relative improvement between v2 and corrected v2 can be computed from results in Liu et al. (2021). MAE was improved by 15.5 % and RMSE by 2.9 %, while GeoGEDI's v2 approaches improved RMSE ground estimations by minimum 13.3 % (forest and non-forest, Landes) and up to 36.2 % (forest, Vosges).

However, results on ground elevation and canopy height accuracy after improving the geolocation were still and inevitably influenced by study site characteristics. The Landes are flat and stands are mainly composed of maritime pine, a species that lets a high proportion of light reach the ground. On the contrary, in the topographically complex area of the Vosges mountains, stands are more dense and are composed of species with higher foliage density. Several studies have reported a link between an increase in RMSEs and an increase in either vegetation density (Dorado-Roda et al., 2021; Liu et al., 2021; Quirós et al., 2021) or terrain slope (Potapov et al., 2021; Liu et al., 2021) for both GEDI ground elevation and vegetation height products. For example, Liu et al. (2021) reported high ground RMSEs (6-7 m) for dense and tall vegetation and a 2.88 m RMSE for areas with slope < 5° compared to 6.70 m for areas with slope > 30°. Similarly, errors for v1 and v2 forest footprints are much higher in the Vosges than in the Landes, and remain higher in the Vosges even after geolocation has been improved, e.g., v2_1 ground elevation RMSEs are 0.91 m and 2.49 m and canopy height RMSEs are 4.99 m and 6.10 m, in Landes and Vosges, respectively. Concerning canopy heights estimations, they are directly impacted by ground estimation accuracy (Liu et al., 2021) and thus by the above mentioned factors. Despite a large geolocation bias correction, improvements in RMSEs between v1 and v2 remain limited (i.e. 5.25 m down to 4.48 m (-17 %) over the Landes and 7.48 m down to 6.30 m (-19 %) over the Vosges). This can be attributed to the relative stability of vegetation height at stand level as both study sites are mainly occupied by even-aged production forests. Even once shifted, a majority of footprints will be located in the same stand and have a similar canopy height value than at their initial location. The uncertainty of reference data may also affect the discrepancy between GEDI and reference data. Most importantly, the time and seasonal differences between the two data acquisitions allow for changes in vegetation heights. The Landes have significant forest dynamics in pine plantations (Guyon et al., 2015), drastically impacting canopy heights.

GeoGEDI limitations in flat areas

Validation highlighted better GeoGEDI performances for Vosges than for Landes. Shift distances v2_1 and v2_4 for Landes were also higher than for Vosges, departing from horizontal geolocation errors announced by the user guide (Beck et al., 2021). Additionally, mean shift distances barely decreased between approaches applied on v1 and on v2. The presence of large flat areas in Landes might explain such results. Typically, DEMref values in Landes optimal position search windows are highly similar, which impedes convergence towards minimal error and finding the optimal position. The error analysis by shift magnitude classes (Fig. 2.9(d)) highlights issues with footprints belonging to C_{Q5} (shift >= 32.6 m) for v1 and to C_{Q4} and C_{Q5} (shift >= 14.6 m) for v2. While all classes' ground elevation estimates improved, surface height estimations of footprints with the largest shifts worsened. These classes may include footprints for which GeoGEDI converged

towards a sub-optimal position. These geolocation errors have more impact on height accuracy than on ground elevation estimates due to the lower variability in elevation compared to surface height variability. It is worth noticing that those very large shifts mainly concern Landes footprints (61 % of the footprints in C_{Q5} in v1 and 79 % of C_{Q5} in v2 belong to Landes). In Section 2.3.2, we also reported that a subset of footprints was removed prior to statistical analyses because the convergence process was interrupted at the search window limit. This mainly concerned Landes footprints, with up to 8.7 % of footprints compared to 1 % in Vosges, suggesting the algorithm had punctually some converging issues in flat areas. The important dispersion of GeoGEDI shifts (e.g. Fig. 2.8(c)) can be explained by ISS large movements and vibrations, or by convergence issues in flat and textureless areas.

Recommendations on the use of GeoGEDI and possible improvements

On the one hand, using the single-beam approach better respects the lidar systems acquisition geometry. On the other hand, using the four-beam approach increases the number of footprints in the cluster and spatial dimension (from 1-D profile to 2-D sampling), which is likely to increase elevation variability within the cluster, especially in low-relief areas. For v1-based approaches, best estimates were observed with the single-beam approach. Therefore, it is more important to respect the acquisition geometry than to build on the beneficial effect of 2-D sampling. To improve georeferencing of v1 data, the single-beam approach should be preferred in all cases. Processing GeoGEDI by beam pair clusters could also be considered in future works, increasing the number of footprints, while respecting the instrument geometry. NASA v2 geolocation was corrected for bias and is less, or even no more impacted by acquisition geometry effects thanks to in-flight calibration. Therefore, the four-beam approach can be considered on v2. Single-beam and four-beam approaches gave very similar GeoGEDI outputs. GeoGEDI v2_1 estimates were slightly better for ground elevations, whereas v2_4 estimates were slightly better for surface height estimates. Both approaches can be used to further improve GEDI v2 geolocation. However, assessing height estimates aimed to provide an independent validation, suggesting that the four-beam approach should be preferred to process v2 data. Furthermore, in low-relief environment, increasing the cluster size would increase heterogeneity in elevations, allowing better convergence of the flow accumulation algorithm. However, it would also result in "smoother" tracks closer to v2 tracks, and thus to lower improvement in geolocation precision with less consideration to errors due to high frequencies vibrations. Moreover, as it is only based on GEDI and DEM ground elevations, GeoGEDI would certainly benefit from improved ground peak detection in L2A data. Indeed, even if GeoGEDI improved estimates, footprints with sharp local ground underestimates were included during adjustment process and might have impacted GeoGEDI's v2-based outputs. Results could also be improved by increasing the search window beyond 50 m and by using a smaller shift step, e.g., equal to the DEMref resolution (1 m), instead of the 2-m δ_{shift} that was used in this study. However, this would result in a sharp computation time increase, and should be accompanied by an optimization strategy, e.g., considering a multiscale approach, using a large step (~ 5 m) to identify the main shift direction, followed with a more local search with a smaller search window and smaller shift step to refine the optimal position. Moreover, as stated in Section 2.3.1, the flow accumulation map value at the optimal Δ_{opt} position can be interpreted as an indicator of GeoGEDI's reliability. The lower the accumulation value of Δ_{opt}, the higher the ambiguity around Δ_{opt}. Examples of low confidence footprints can be found in A.1. A simple threshold could be used and added to each footprint by adding a tag, warning users about possible convergence issues, similarly to quality and degrade flag implemented by NASA.

Conclusion

GEDI footprints provide large scale and high sampling density data about forest structure. But low georeferencing accuracy can be detrimental to their use for predictive models of forest attributes. The proposed method is based on GEDI ground elevations and a high-resolution DEM, to improve geolocation of GEDI footprints. The method was tested on GEDI v1 and v2 for two French forests, broadleaved-dominated forest in a flat area and dense coniferous-dominated forest in a mountainous area. Our results quantified the georeferencing improvements undertaken by NASA between version 1 and 2. Besides, a ground detection issue was identified for GEDI v2 footprints using algorithm 02. However, GeoGEDI successfully improved GEDI v1 and v2 footprints positioning, simultaneously reducing bias and improving precision components. Despite improved footprint geolocation in GEDI v2, already corrected for the systematic error components, there is room for additional improvement. Yet, its performance depends on the topography, with lack of convergence in very flat areas. The method showed efficient to correct for ISS attitude and altitude variations for a diversity of forest environments, and to assess GEDI data quality with more confidence. The methods' relative simplicity allows for fast and efficient large-scale deployment, wherever a high resolution DEM is available. With improvements in the range of those obtained with more complex methods based on waveform processing, the method is a good alternative candidate to process GEDI data prior to implementing methods requiring a precise matching of data sources such as for data fusion purposes.

Chapter 3

Usefulness of GEDI footprints as first-phase sample for forest inventories based on double sampling for post-stratification

Anouk Schleich[a], Olivier Bouriaud[b,c], Cédric Vega[c], Sylvie Durrieu[a]

[a]UMR TETIS, Univ Montpellier, AgroParisTech, CIRAD, CNRS, INRAE, F-34196 Montpellier, France

[b]University Stefan cel Mare of Suceava, RO-720229 Suceava, Romania

[c]ENSG, IGN, Laboratoire d'inventaire forestier, F-54042 Nancy, France

Abstract

The GEDI spaceborne lidar system was specifically designed to study forest ecosystems. Inference on forest attributes using GEDI data was mostly addressed through model-assisted, model-based and hybrid approaches. In this study, we applied a double sampling for post-stratification (DSPS) design-based approach to combine GEDI and national forest inventory data. Although widely used in the field of forest inventories, the use of such a design-based approach relying on GEDI data has not yet been investigated. This method is advantageous because it requires neither precise geolocation nor co-location between GEDI footprints and inventory plots. We evaluated the impact of the bridge variable and the impact of GEDI's spatial sampling scheme on the results of the DSPS approach by comparing our GEDI-based results to reference airborne-laser-based results. We employed maximum tree height as the bridge variable and chose a complex study area in northeastern France with relief and highly diverse forest stands. We used $202{,}808$ GEDI footprints as the first-phase sample and 476 National Forest Inventory (NFI) plots as the second-phase sample to estimate the growing stock volume (GSV). Compared with estimates based solely on NFI field plots, the DSPS approach reduced the GSV variance by up to 54% without any additional cost, aside from the negligible additional time required to download and process the GEDI data.

3.1 Introduction

National forest inventories (NFIs) are essential tools for forest monitoring. They contribute to estimating forest characteristics such as volume, basal area, dominant heights, and species composition, which are used in some countries to inform and control forest policies and management decisions (Tomppo et al., 2010; Breidenbach et al., 2021). Several forest characteristics can only be assessed from field measurements; however NFIs require multiple years to collect sufficient data to achieve the required precision level. To address the need for more frequent updates and finer spatial resolution in forest monitoring, NFIs combine classical field plots with correlated auxiliary data sources, most often remotely sensed data, to enhance the spatial and temporal scales of forest inventories, with minimal development costs (Tomppo et al., 2008; Westfall et al., 2019). These approaches, also referred to as Multi-source Forest Inventories (MFI), have been extensively tested and show improvements in estimation precision across various countries (Tomppo et al., 2008; Saborowski et al., 2010; Westfall et al., 2019). Different data sources such as optical remote sensing satellite data (e.g., Landsat or Sentinel-2 (McRoberts et al., 2007; Puliti et al., 2021), 3D models derived from aerial imagery (Waser et al., 2015; Pulkkinen et al., 2018) or from aerial lidar scanners (ALS) (Næsset, 2004; Ståhl et al., 2011; Asner et al., 2012; Gobakken et al., 2012; Corona et al., 2014; Guerra et al., 2022), or a combination of these data types (Saarela et al., 2015; Irulappa Pillai Vijayakumar et al., 2019) have been utilized. Although using these auxiliary data led to more precise NFI estimates, it also has limitations. Several spaceborne optical sensors provide free near-real-time images with frequent updates (1-2 weeks revisit time). They cover every location on Earth, and the long mission durations make the data particularly valuable for forest dynamics continuous monitoring (Puliti et al., 2021). However, the limited correlation between passive optical signals and forest structure makes it challenging to estimate structure-related variables such as volume (Saarela et al., 2018). ALS data and 3D imagery show a strong correlation with forest attributes related to structure, but their acquisition is costly, time-consuming, and less suitable for large-scale monitoring.

The launch of the Global Ecosystem Dynamics Investigation (GEDI) mission in late 2018 introduced a new data source that overcomes the aforementioned limitations and has the potential to enhance the precision of NFI outputs. The GEDI instrument, mounted onboard the International Space Station (ISS), is an experimental lidar system specifically developed to provide information on forest canopy height and vertical structure at high resolution. GEDI collected data at a fine spatial and temporal scale (Dubayah et al., 2020a) during its first four years in orbit. After a pause of a little over a year, it resumed acquiring data in April 2024 and is expected to continue for six additional years (LP DAAC, 2023). This makes GEDI a milestone mission toward using full-waveform space lidar in MFI approaches and producing more precise information about forest characteristics and their dynamics across a larger range of spatial and temporal scales.

Thus far, GEDI data have been used in model-based approaches, or in a combination of model-based and design-based approaches, such as model-assisted or hybrid inference frameworks (Qi et al., 2019; Potapov et al., 2021; Duncanson et al., 2021; Zhang et al., 2022; Lang et al., 2022). The irregular spatial distribution of GEDI footprints indeed seem to favour model-based approaches, which do not require a particular sampling pattern. A hybrid approach was used for the official GEDI release of Aboveground Biomass Density (AGBD) predictions (L4A). Underlying models were trained using GEDI waveforms simulated from ALS data, and references from field inventories (Patterson et al., 2019; Duncanson et al., 2021). Using hierarchical-model-based estimators, these models enable the creation of global AGBD predictions from GEDI data (Saarela et al., 2022). Zhang et al. (2022) obtained promising results for forest volume estimation using a model-

assisted small-area estimation approach applied in a two-phase estimation procedure. Similarly, Bullock et al. (2023) paired NFI plots and GEDI footprints to calibrate a region-specific field-to-GEDI biomass model and applied the same hybrid inference framework as the L4A data. Bruening et al. (2023) also used this hybrid inference method to improve AGBD estimation using GEDI data and compared the results with the US Forest Inventory estimates.

However, the use of GEDI in MFI under model-based approaches raises three important issues: (1) GEDI acquires data at the level of discrete footprints covering 25 m diameter circular areas on Earths surface. Unlike most other remote sensing data continuously covering Earths surface, GEDI data is not available everywhere. Therefore NFI and GEDI data do not overlap, which makes it challenging to directly combine them to develop predictive models of forest attributes and propagate predictions at every point in space. Additionally, the ISS orbit was raised unexpectedly during GEDI's acquisition phase, changing GEDI's beforehand expected sampling pattern, thus challenging planned hybrid and hierarchical approaches (Saarela et al., 2022; Dubayah et al., 2022a); (2) The geolocation of GEDI is neither very precise nor accurate, with an estimated precision of ~ 10 m (Beck et al., 2021; Roy et al., 2021; Schleich et al., 2023c), which may hamper the development of models carried out through the joint analyis of GEDI information with other geolocated datasets used in MFI approaches, e.g., NFI plots and wall-to-wall remotely sensed data, which violates the requirement of spatial co-location of the auxiliary and NFI-based measurements; (3) Data filtering based on the quality and degrade flags available in GEDI products is not sufficient to screen out problematic waveforms and related higher-level products (Morin et al., 2022; Lang et al., 2022; Bruening et al., 2023), which can result in severe outliers models are particularly sensitive to (Renaud et al., 2022).

Tight relations among field- and remote-sensed data, and a perfect spatial co-location, are prerequisites for approaches using a model, which is not the case with GEDI. The objective of the study is thus to test the potential of a design-based approach built on well-established statistical principles combining GEDI and NFI data to enhance the precision of forest resource estimation at sub-regional level in France. To our knowledge, this is the first study combining GEDI and NFI data in a purely design-based approach. If this approach is validated, GEDI data could enable the straightforward production of new estimations with greater precision, without incurring additional costs. We implemented the Double Sampling for Post-Stratification (DSPS) approach, which proved effective in reducing estimation errors in NFIs (Westfall et al., 2019; Köhl et al., 2006; Westfall et al., 2021). Such a design-based approach can provide unbiased estimates of forest characteristics (Gregoire, 1998; Ståhl et al., 2016; Lister et al., 2020), and is therefore highly valued by policymakers and forest stakeholders. For instance, such an approach has proven successful in achieving small area estimations in the German NFI (Hill et al., 2018).

In DSPS, the first sample comprises a large number of easy-to-assess low-cost sampling units, as provided by GEDI, and is used to estimate strata sizes. The second sample is composed of a smaller subset of higher-cost sampling units such as NFI plots, which provides a mean estimate of the forest attributes of interest for each stratum. While the second sample is often a subsample of the first one, the two samples can be totally independent (Hidiroglou, 2001; Westfall et al., 2019; Haakana et al., 2019). We hypothesized that this DSPS approach could solve most of the issues reported in the literature when integrating GEDI data into MFI approaches, particularly the lack of co-location with field plots that impede model-based or model-assisted approaches, while being more robust regarding estimation errors and outliers. Additionally, geolocation is only used to determine whether the data is situated within the study site. Unlike other MFI approaches, the DSPS approach does not require to extract information from other remote sensed data by intersection; con-

sequently, it is not impacted by geolocation errors in GEDI data. However, the DSPS approach is sensitive to the sampling scheme (Gregoire et al., 2016) and requires a robust link between the field and auxiliary data that allows classification into the same strata. We hypothesized that maximum tree height was a suitable candidate for that purpose and that GEDI sampling characteristics were well-suited for implementing the DSPS approach.

Our goal was to estimate the effectiveness of the DSPS approach using GEDI data for multi-source estimations at a spatial scale beyond the usual NFI capacity. To this end, first we examined the two working hypotheses (1) GEDI's spatial sampling characteristics and (2) the quality of the link variable, relying on an ALS dataset contemporaneous to field and GEDI acquisitions as a highly accurate reference dataset. Second, we implemented the DSPS approach. To assess the efficiency of using GEDI as auxiliary data, DSPS results were compared to estimates obtained from NFI plots only using a simple expansion estimator.

3.2 Data

3.2.1 Study site

The study site is part of the Vosges forest ecoregion, a mountainous forest environment located in North-Eastern France. It has been defined as the area covered by the ALS data recorded in 2020, as described in Section 3.2.2 (see Fig.3.1). The total study site covers a $4{,}777$ km^2 surface area, which represents almost 9% of the French metropolitan area, and is slightly smaller than the mean area of the departmental administrative divisions i.e., $5{,}880$ km^2). The site encompasses very different environmental conditions, with altitudes ranging from 100 to $1{,}300$ m, and slopes reaching up to $60°$. 65% of the study site is covered with forests. The dominant species are Silver Fir (Abies alba), Norway Spruce (Picea abies), European Beech (Fagus sylvatica) and Sessile Oak (Quercus petraea). Lower altitudes are predominantly covered by deciduous forest stands whereas higher altitudes are mainly composed of mixed and coniferous forest stands (IGN, Sylvoécorégion; Cavaignac, 2009).

3.2.2 Reference data: ALS Canopy Surface Model

The National Institute of Geographic and Forest Information (IGN) provided a 1-m resolution Canopy Height Model (CHM), estimated using ALS data acquired during the winter of 2020 with an average 4.8 pt/m² first return point density.

3.2.3 Auxiliary forest database

BD Forêt® v2 (IGN, BD Forêt version 2) is the official national French database outlining forest features and providing information on their composition. The database was intersected with the ALS data extent to compute the forest area of interest, and it was further used to select GEDI footprints inside the forest mask.

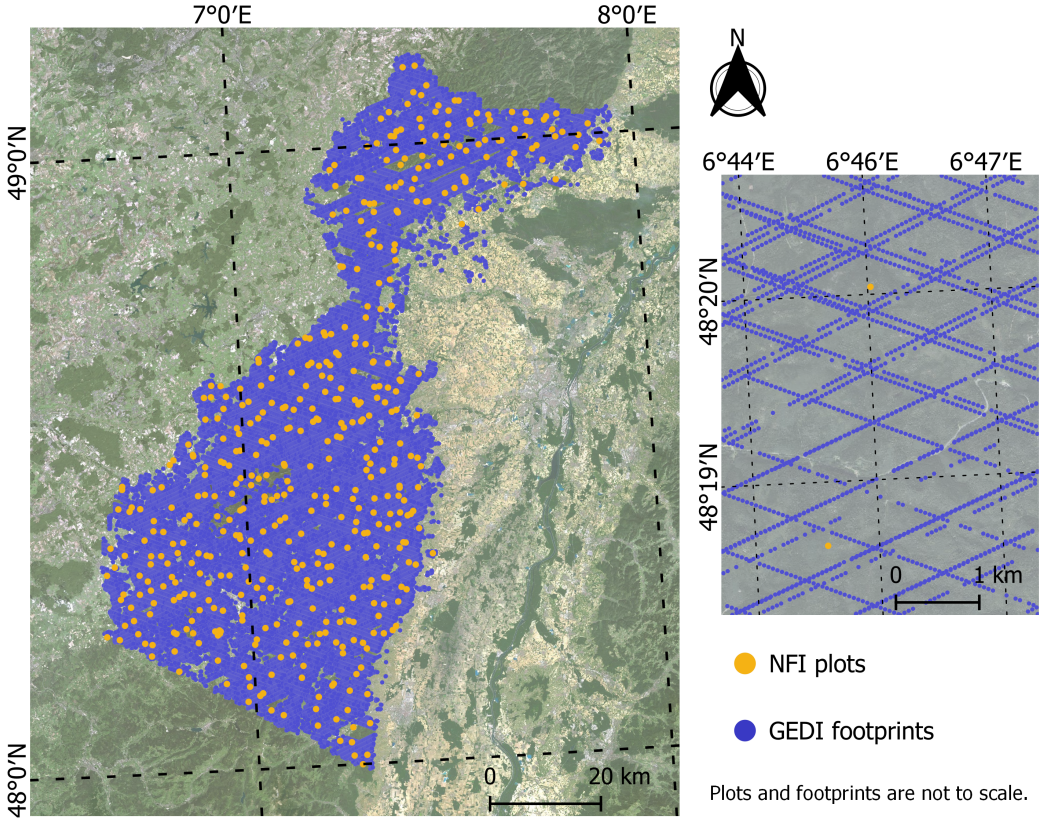


Figure 3.1: NFI plots and forest GEDI footprints over the Vosges study site. Footprints were filtered to intersect with BD Forêt polygons. Moreover, multiple quality-based filters were applied, resulting in 202,808 GEDI footprints. The NFI data consists of 476 plots.

3.2.4 National forest inventory plots

We used data from the NFI database collected between 2017 and 2020 (N = 476) within the ALS data extent. The time period was chosen considering the following three criteria: i) the total number of NFI plots should be sufficient to provide estimates from NFI plots alone (NFI results typically rely on data collected over a five-year period); ii) data availability at the time of the study; and iii) maximizing the overlap with the GEDI acquisition period (2019 to 2022). This represents a sampling effort of around 1 plot per $10\ \text{km}^2$. In the field, trees are inventoried in 3 concentric plots of 6, 9 and 15 m radii according to their circumference at breast height (i.e., 1.3 m). Trees with a circumference of $[23.5 - 70.5\ \text{cm}[$ are assessed in 6 m radius plots; trees in the range of $[70.5 - 117.5\ \text{cm}[$ are assessed in the 9 m radius plots, and those with a circumference larger than 117.5 cm in the 15 m radius plots (Hervé et al., 2017). The target variable was the plot-level growing stock volume (GSV). It is estimated using field measurements, the probability of inclusion of trees, and species-specific volumes models, and converted into a local density in m^3ha^{-1} .

3.2.5 Auxiliary GEDI L2A data

The GEDI instrument is a spaceborne lidar system onboard the ISS, which has been acquiring data since April 2019. It consists of three lasers, one of which is split into two beams. All lasers were deflected between each

laser shot, resulting in two ground tracks per beam. The eight tracks were spaced 600 m apart, perpendicular to the ISS flight direction, and each track consisted of 25 m diameter footprints, 60 m apart (Dubayah et al., 2020a). For this study, GEDI Level 2A and 1B versions 2 were downloaded from NASA's Archive Center (Dubayah et al., 2021a). L2A products provide geolocation, ground elevation, and relative height (RH) metrics estimated for each footprint. RH values were obtained by measuring the cumulative waveform energy from bottom (RH0) to top (RH100). RH elevations are given relatively to the elevation of the lowest detected mode, which is assumed to be the ground level. The canopy top height can be defined by different upper RH values (e.g., RH95, RH98 and RH99 in Dorado-Roda et al. (2021), RH98 in Duncanson et al. (2021), RH100 in Adam et al. (2020), and RH95 and RH100 in Lahssini et al. (2022)). In the present study, RH100 was used. The downloaded data covered acquisition dates from April 2019 to December 2022. To avoid issues with poor-quality data, only full-power beams with appropriate quality and degrade flags were used (Duncanson et al., 2020). The GEDI L1B products were used to identify remaining poor-quality waveforms and develop additional filters based on the variables available in L2A products.

3.3 Methods

Before applying the DSPS approach (Section 3.3.4), some data preprocessing was necessary (Section 3.3.1), and the two working hypotheses were checked. First, the fact that GEDI's sampling scheme can be characterized as a probability-based sampling scheme (Section 3.3.2) and second, the potential of maximum tree height to serve as the "bridge variable" between GEDI and NFI data had to be evaluated (Section 3.3.3).

3.3.1 Data preparation

Imputing NFI maximum heights We assumed that the GEDI canopy height, RH100, represented the height estimation of the tallest tree included in the footprint. It should therefore be equivalent to the maximum tree height in NFI plots. However, for each NFI plot, tree height is only measured for a single tree per species per circumference classes. Therefore, missing heights were imputed. Imputations were conducted nationwide using a random forest MissForest approach (Stekhoven and Bühlmann, 2012), applied per species and ecological regions, and taking into account tree circumferences, plot density and basal area. For the 156 species processed, the overall mean error for imputed heights (observed - predicted) was -0.1 ($\pm$0.4) m (evaluated on a test dataset). Once all the missing tree heights were imputed, the maximum tree height for each NFI plot was assigned as Hmax.

Computing a maximum height raster The 1-m resolution CHM estimated from the ALS data is spatially coincident with both GEDI and NFI datasets and was used to evaluate the degree of matching between the chosen bridge variable, i.e. between RH100 and Hmax. A 1-m resolution focal maximum CHM, called CHMref, was computed by applying a circular 25 m focal maximum filter to the CHM. For all NFI plots and GEDI footprints, ALS heights were extracted from CHMref using the plots and footprint centers coordinates, respectively. Values extracted from CHMref are referred to as HALS.

Excluding non-forest data NFI field measurements focused only on forest-land. To make sure only forest areas were considered, filters were applied to both GEDI and ALS data. The lowest and highest Hmax in our NFI dataset were 7.8 m and 46.5 m, respectively. Applying a 10% margin, GEDI footprints with RH100 < 7 m or HALS < 7 m were excluded. Similarly, and allowing for potentially taller trees not captured by NFI plots while limiting unrealistic heights, all footprints with heights exceeding 60 m were discarded ($\sim 30\%$ margin). Moreover, only footprints identified as forests in the BD Forêt® database were retained for further analysis. The resulting dataset included $277{,}471$ footprints. Similarly, the CHMref raster was masked to exclude pixels below 7 m or above 60 m or outside the BD Forêt®. The corresponding forest area, referred to as Ref_Als, was considered as the estimation domain and its surface area A_T was $3{,}112$ km^2.

GEDI data filtering Some footprints exhibited strong inconsistencies between GEDI and ALS heights. An empirical analysis of inconsistent heights was then conducted to define new filters aimed at excluding the problematic footprints. This definition was supported by a visual analysis of L1B waveforms. The selected filters relied solely on L2A variables, namely RH data, the identifier of the mode selected as the lowest non-noise mode ("selected_mode"), information about the received waveform energy, i.e., the integrated counts in the return waveform ("energy_total") and the maximum amplitude of this return waveform ("rx_maxamp"), both relative to mean noise level (Hofton and Blair, 2019; Dubayah et al., 2021a). Filters are detailed in B.1. The resulting GEDI dataset included $202{,}808$ footprints and was the dataset used for all further analysis.

3.3.2 Verifying if GEDI's sampling scheme has the properties of a probabilistic sampling scheme

The use of DSPS requires a probabilistic sampling scheme. By design, the NFI plots followed a probabilistic sampling scheme. If a sample followed a probabilistic sampling scheme, it correctly reflected the diversity of the entire study area. Random and systematic samples are probabilistic. GEDI footprints are neither regularly nor randomly distributed, and the hypothesis that GEDI sampling has the properties of a probabilistic sampling scheme must be carefully examined. The analysis of GEDI footprint geolocation by Schleich et al. (2023c) showed that the deviation between the announced location and the actual location was highly variable in space and time and exceeded 10 m on average at our study site. The analysis further showed that two consecutive footprints could exhibit significantly different shifts. These elements suggest that the shift between the announced and actual locations could be characterized as a random process, thereby providing randomness to the sample, to some extent. Furthermore, some footprints are filtered owing to the presence of clouds or other sources of signal perturbations, which can also add irregularity to the footprint distribution along the orbit tracks compared to the theoretical distribution. To verify whether the resulting footprint distribution was similar to probabilistic sampling, the following method was applied.

To focus on the spatial sampling properties of GEDI without interference from GEDI height estimation errors, we compared the strata proportions obtained with HALS heights extracted at GEDI locations to other HALS subsets. Differences in stratum proportions were used as diagnostic tools for sampling. To study the strata size distributions, top canopy heights were classified into five classes (i.e. (7,20], (20,25], (25,30], (30,35], (35,60] m) and their proportions were tested among the different layouts listed below.

- Ref_Als: Strata sizes computed using all pixels ($\sim 3 \times 10^9$) constituting CHMref (and inside the study site, intersecting BD Forêt and having heights greater than 7 m and less than 60 m).

- Random_Als and regular_Als: Probability-based sample strata generated by selecting N pixels that are spatially randomly or regularly distributed within Ref_Als and using the extracted HALS values for stratification. N = number of footprints in the corresponding GEDI data-set.
- gedi_Als: HALS extracted at GEDI-locations are used for stratification.
- Shiftedgedi_Als: The HALS extracted at shifted GEDI locations were used for stratification (see Section 3.3.3).
- gedi_Rh: GEDI RH100 values are used for stratification.

Thus, unlike random_Als and regular_Als, gedi_Als and gedi_Rh were both positioned in the GEDI footprints. Ref_Als will be considered as the reference data owing to the complete spatial coverage of the study area and its precision. All three billion pixels were used to create the proportions of the height classes (stratum sizes and proportions). The number of GEDI footprints in the data-set defines the number of points (N) used for the random and regular layouts. For random sampling, 2000 random samples were created, and the mean and 95% confidence intervals were calculated for each height class. If the probability-based samples had the same distribution as Ref_Als, the number of points N was sufficient to correctly sample the study site. If the GEDI data follow a probabilistic sampling scheme, the proportions of gedi_Als should also follow the same distribution and the gedi_Als proportions should fall within the 95% confidence interval of the 2000 random samples. Shiftedgedi_Als layout was added to simulate "real" GEDI spatial distribution, by applying the shift distribution found by Schleich et al. (2023c) and carried out on the same study site.

The proportions of gedi_Rh were also included in the analysis. Gedi_Rh does not inform on GEDI's sampling design alone, but it allows us to see the combined impact of the GEDI sampling design and GEDI and ALS height differences on surface proportions. The gedi_Rh proportions were further used in the DSPS approach based on GEDI.

Analysis was conducted using several sets of GEDI data. Initially, $202{,}808$ footprints were considered. Then, per-year and per-season data subsets were studied to evaluate the evolution of the GEDI sampling scheme characteristics with a decrease in the number of footprints and the possibility of implementing annual estimations.

3.3.3 Verifying the bridge variable between GEDI and NFI data

To verify the link between GEDI and NFI height, we compared both heights to HALS (i.e., HALS to NFI Hmax and HALS to GEDI RH100). Regression line models, R^2 values, and Pearson's correlation coefficients were calculated. Additionally, a paired sample t-test was performed to determine whether the mean difference between two sets of observations was significant (Hsu and Lachenbruch, 2014).

Several factors are likely to influence the quality of the GEDI canopy height assessment and might thus interfere with the relationship between GEDI RH100 and HALS. A more in-depth analysis was conducted to analyze the influence of geolocation accuracy, season, year, forest stand type (extracted from BD Forêt), slope, and the GEDI variable *selected_algorithm*, which corresponds to the algorithm used to detect ground peaks and has been shown to affect height estimation errors (Schleich et al., 2023c).

To study the effects of georeferencing uncertainties on height estimations, we compared the ALS heights at the GEDI footprint location (gedi_Als) with the ALS heights extracted at a shifted location (shiftedgedi_Als).

The direction of the shift was randomized for each footprint and three distance distribution layouts were tested: d following a normal distribution $d \sim \mathcal{N}(\mu = 10, \sigma = 10)$, d following a normal distribution $d \sim \mathcal{N}(\mu = 10, \sigma = 20)$ and d following the empirical distance distribution of Schleich et al. (2023c) estimated in the same study site. $\mu \approx 10m$ has been tested and found by several studies (Roy et al., 2021; Quirós et al., 2021; Schleich et al., 2023c) and we tested two different σ values. For each distance distribution layout, $5{,}000$ simulations were performed using random subsamples of footprints set to 10% of the total number of footprints (i.e., $20{,}281$). The minimum, maximum, mean, and standard deviation of the mean height difference, Pearson correlation coefficient, R^2 and the regression line coefficients were calculated for each run between the original and shifted HALS values.

Regarding other factors, the residuals of the simple linear regression between RH100 and HALS were analyzed to verify whether these factors could influence the relationship.

3.3.4 Double Sampling for Post-Stratification Approach

The French NFI relies on a design-based approach using a two-phase sampling design with post-stratification. This method involves creating a spatially systematic sample on an annual basis to select a representative sample of the French metropolitan territory (Hervé et al., 2014; Vidal et al., 2016a; Bouriaud, 2020). Inventory points were randomly sampled within an annual subset of cells from a hierarchical square grid (Bouriaud et al., 2023). The first phase consists of a photo-interpretation of these points using infrared orthophotographs from the national database BD ORTHO®. This phase provides information on land cover (open or closed woodland, herbaceous, etc.) and land use (agricultural, wood production, etc.) for each 25-meter-radius plot surrounding the inventory points. This allows the estimation of the proportion of each land cover and land use category. The second phase involved drawing a sub-sample from the woodland plots identified in the first phase, which were visited in the field. The field measurements covered over 200 attributes, including stand descriptions and tree measurements. Subsequently, additional variables such as basal area and GSV were estimated for each plot. The post-stratification approach uses the proportions obtained from Phase 1 and plot-level data from Phase 2 to make national or regional estimates of forest attributes. By incorporating this post-stratification approach, the variance of the estimators is considerably reduced compared with the estimators without post-stratification. Each year about $60{,}000$ first-phase plots are photo-interpreted and around $7{,}000$ points are visited in the field.

In this study, we focus on the estimation of the GSV, denoted as the variable of interest Y, in an inventoried territory with a known area A_T. GEDI data were defined as the first-phase sample, denoted as S_1, of size n_1. Second-phase NFI field data are denoted as S_2, with a size of n_2, where n_2 is typically much smaller than n_1. GEDI footprints and NFI plots were classified into strata based on their respective maximum height estimations. The strata were defined in a way to attribute each GEDI footprint and NFI plot to exactly one stratum h. GEDI footprints were used to estimate the surface proportions of each stratum, whereas the NFI plots were used to estimate the mean Y within each stratum. The variable Y measured on NFI plots was expressed as spatial density, that is, GSV per unit area (typically m^3ha^{-1} for volume). The estimate of the total Y in the territory was then obtained by combining the measurements of the two samples using Eq.3.1.

$$\hat{T}_Y = A_T \sum_{h=1}^{H} P_h \overline{Y}_h \tag{3.1}$$

with:

$$\overline{Y}_h = \frac{1}{n_{2_h}} \sum_{j=1}^{n_{2_h}} y_j \tag{3.2}$$

where:

$\hat{T}_Y$ = Total estimation of variable Y for the study area
A_T = Territory area
H = Total number of strata
h = Strata, with $h \in [1, H]$
P_h = Surface proportion of stratum h, often referred to as stratum weight, estimated as the proportion of GEDI footprints belonging to stratum h compared to the total number of footprints
$\overline{Y}_h$ = Average density of Y in stratum h estimated from the NFI field plots
n_{2_h} = Total number of NFI plots in stratum h
y_j = the value of Y measured for NFI point j belonging to strata h

The estimation based on the DSPS approach involves the product of two variables: the area proportion of a stratum and the mean value of the variable of interest, Y, in the stratum. An estimator of variance for the total within DSPS was proposed by Cochran (1977) (equation 12.32) and is widely used in DSPS or is found in a closed form (Scott et al., 2005; Saborowski et al., 2010; McRoberts et al., 2012; Westfall et al., 2021; Bechtold and Patterson, 2015; McRoberts et al., 2013). The first part reflects the variance of the attribute of interest within the strata, while the second term comes from the fact that, in DSPS, strata sizes are not known but estimated, based on the first-phase sample (Eq.3.3 and Eq.3.4). The large size of the first-phase sample and the independence between both samples in our study are particularly favorable factors to using this particular estimator.

$$\hat{Var}(\hat{T}_Y) = A_T^2 \left\{ \sum_h P_h \frac{n_{1_h} - 1}{n_1 - 1} s^2_{\bar{Y}_h} + \frac{1}{n_1 - 1} \sum_h P_h (\bar{Y}_h - \bar{Y})^2 \right\} \tag{3.3}$$

where

$$s^2_{\bar{Y}_h} = \frac{1}{n_{2_h}(n_{2_h} - 1)} \sum_{j=1}^{n_{2_h}} (y_j - \bar{Y}_h)^2 \tag{3.4}$$

is the estimator of variance of the attribute mean in stratum h and $\bar{Y} = \sum_h P_h \bar{Y}_h$ is the estimated mean over the territory of interest.

In this study, the territory area was defined as Ref_Als, which was $3{,}112$ km^2. Three different height stratifications based on the maximum height were tested to estimate Y (i.e., GSV):

- 2 height strata: (7,30], (30,60] (m)
- 3 height strata: (7,20], (20,30], (30,60] (m)
- 5 height strata: (7,20], (20,25], (25,30], (30,35], (35,60] (m)

Here, the 30 m limit was determined based on forest characteristics because it is the usual height at which trees reach maturity. The other height limits were defined to be regularly spaced and to contain sufficient number of points in each stratum.

First, we evaluated the impact of GEDI's spatial sample design using HALS heights for both S_1 and S_2 and testing with random locations (random_Als) and GEDI locations (gedi_Als). Next, to evaluate the impact of height sources on the DSPS estimates, the NFI Hmax was used for S_2, and the results with S_1 based on HALS were compared to the results with S_1 based on RH100. Finally, to assess the impact of using a given subset of the GEDI data on the estimates, the DSPS approach results were compared using yearly (2020, 2021, and 2022) and seasonal subsets (winter: December to March; summer: June to September).

For each stratification, the estimated GSV and 95% confidence intervals were computed. To compare the results of the DSPS approach with the estimates based solely on NFI data, a simple random sampling (SRS) was performed with NFI data. These results were extended to the total area and presented as SRS-GSV with $\pm 95\%$ confidence intervals. Relative efficiency (RE, see Eq.3.5) was used to quantify the improvements achieved by adding GEDI data. An RE greater than 1 indicates that stratification successfully reduced variance and increased precision. RE can also be translated as a factor by which the sample size would need to be increased with the SRS to match the precision obtained using the stratified estimators. In essence, RE serves as a measure of how much improvement is gained through stratification compared to using SRS alone (McRoberts et al., 2012).

$$RE = Var_{SRS}(NFI)/Var(DSPS) \tag{3.5}$$

where:

$Var_{SRS}(NFI)$ = Variance of simple expansion estimator on NFI plots considered to result from a simple random sampling (SRS).
$Var(DSPS)$ = Variance of Double Sampling for Post-Stratification

3.4 Results

3.4.1 Analyzing GEDI's distribution properties

Fig. 3.2 shows the proportions of the five height classes for the tested layouts. The random and regular layouts achieved the same proportions as in Ref_Als. This indicates that the number of GEDI footprints is theoretically sufficient to correctly sample the study site, if it is distributed randomly or systematically. In contrast, the proportions of gedi_Als and shiftedgedi_Als differed slightly from probability-based sampling proportions. The two upper strata (that is, (30,60]) were underestimated with respect to Ref_Als by a total of 2.7 and 2.3 percentage points in gedi_Als and shiftedgedi_Als, respectively, the following two strata (that is, (20,30]) were overestimated by the same proportions. Moreover, the gedi_Als and shiftedgedi_Als proportions do not fall within the random_Als confidence interval (e.g., 15.634 - 15.640% for class (7,20], 18.614 - 18.621% for class (20,25]), 27.322 - 27.330% for class (35,30], 25.262 - 25.270% for class (30,35] and 13.151 -

13.157% for class (35,60]). This indicates that the spatial distribution of the GEDI footprints does not provide an unbiased sampling of the site.

Moreover, gedi_ALS and gedi_Rh comparisons allowed us to consider both working assumptions. This shows the impact of the RH100 - HALS link on the surface proportions. When using RH100 instead of HALS to assign a footprint to a stratum, the upper height strata (35,60] were overestimated, and the intermediate (25,30] class was underestimated. They included 15.0% and 25.8% of the footprints, respectively, compared with 11.0% and 28.7% for gedi_Als. It is worth noting that when grouping the five height classes into the following two classes, (7,30] and (30,60], the resulting class proportions are similar to those obtained with a probability-based sample, with 64.4% and 35.6% for gedi_Rh and 65.7% and 34.3% for random_Als, for classes (7,30] and (30,60], respectively. The RH100 - HALS link is investigated further in the next Section.

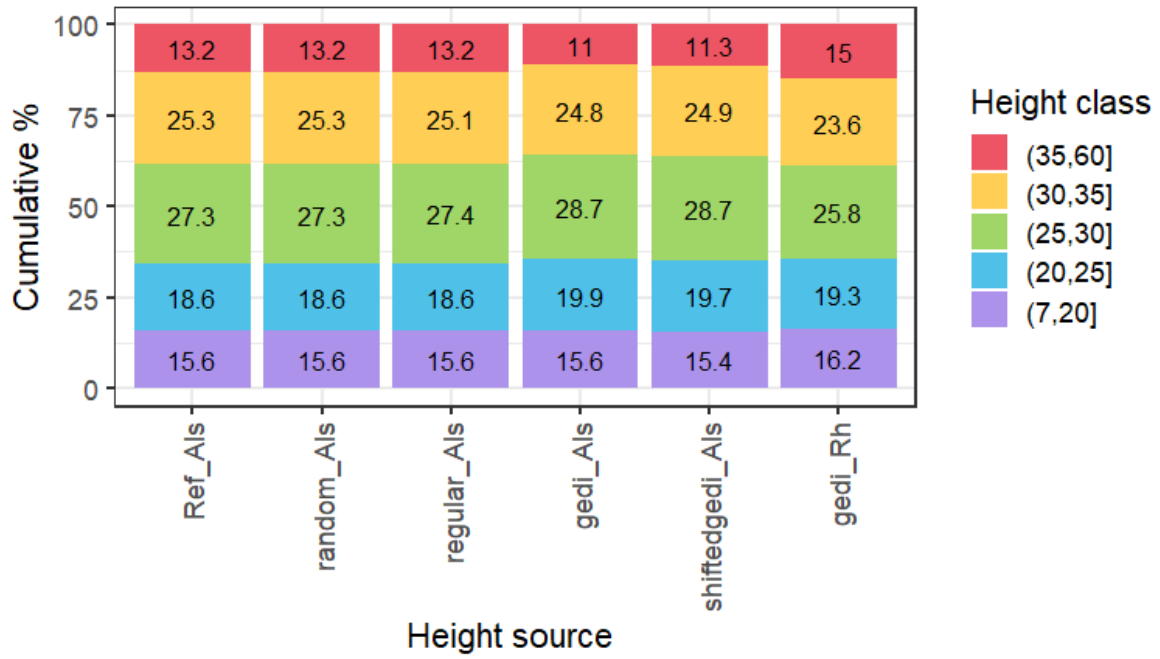


Figure 3.2: Surface proportions of different point layouts. Ref_Als are the reference proportions based on the raster. All other layouts use $\sim 202{,}808$ points. Random_Als and regular_Als used a random and a regular distribution of ALS points, respectively. gedi_Als uses ALS heights extracted for GEDI footprint locations. shiftedgedi_Als uses ALS heights extracted for shifted GEDI footprint locations. gedi_RH uses RH100 values provided in GEDI data. The proportion of each class is marked in %.

3.4.2 Testing the bridge variable

Link between NFI Hmax and ALS heights

NFI Hmax and HALS heights presented a strong correlation of 0.89 ($N = 476$, $p < 2.2 \times 10^{-16}$) (see Fig. 3.3). The regression line was $y = 0.966 + 0.968x$ with $R^2 = 0.80$, $Fvalue = 1{,}888$ and $p < 0.001$. Only a few points strongly depart from the 1:1 line. The few outliers above the 1:1 line with ~ 38 m Hmax heights and < 20 m HALS heights might be clearcuts that occurred between both data acquisitions. However, some of these discrepancies could also be due to errors in the geolocation of the NFI plots in the field. Differences of a few meters may also correspond to the height growth between the NFI and GEDI acquisitions. Deviations

may also be due to inaccuracies in the field height measurements and estimations of Hmax. However, the paired t-test revealed that the difference between the two height variables was not statistically significant ($t = 0.58$, $p = 0.56$; mean difference $d = 0.09$ m).

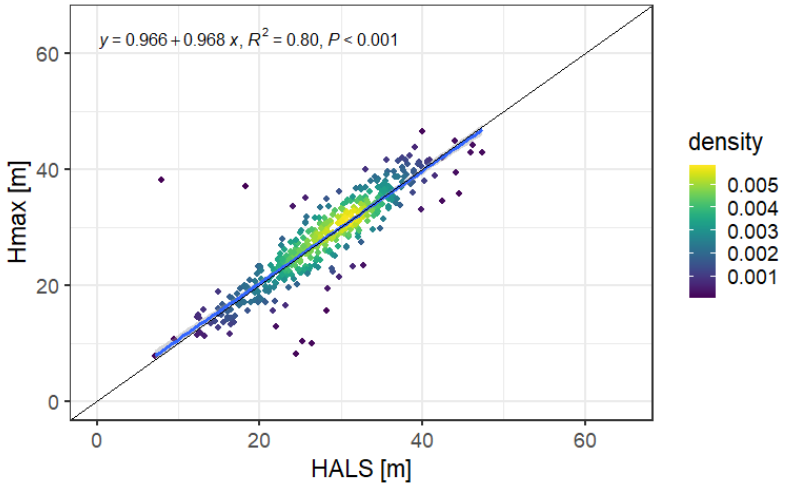
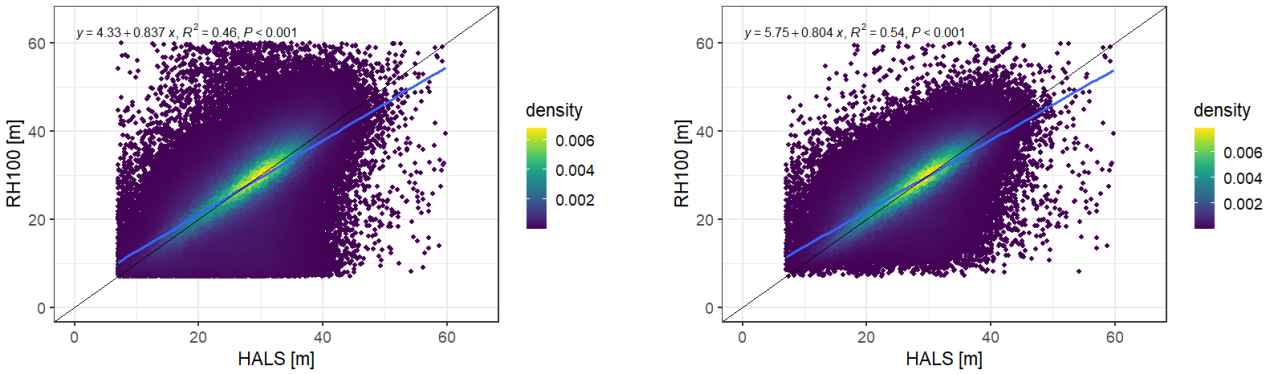


Figure 3.3: Pointcloud of Hmax and HALS heights for NFI plots

Link between GEDI RH100 and ALS heights

Point clouds comparing HALS heights to GEDI RH100 heights showed substantial variability (Fig. 3.4). The potential sources of this dispersion are analyzed in the following subsections.

Filtering additional bad quality GEDI data Fig. 3.4 shows the relationship between HALS and GEDI RH100 before and after applying the additional filters. Before additional filtering, with the standard GEDI filters (Fig. 3.4a), Pearson's correlation coefficient was 0.68 ($N = 277{,}471$, $p < 2.2 \times 10^{-16}$). Applying additional filters, reduced the data-set size by $\sim 27\%$ and the correlation coefficient increased to 0.74 ($N = 202{,}808$, $p < 2.2 \times 10^{-16}$). The paired samples t-test for the first dataset indicated that RH100 (mean μ = 27.20, standard deviation σ = 8.53) underestimated HALS (μ = 27.33, σ = 6.92; t = -10.47, p < 2.2×10^{-16}) with a mean difference d of -0.13 m. After additional filtering, the mean RH100 increased to 27.59 m (σ = 7.22 m) while HALS mean decreased to 27.18 m (σ = 6.62 m), resulting in an overestimation of the heights by 0.42 m ($t = 37.30$, $p < 2.2 \times 10^{-16}$). The p-value achieved for both t-tests (p < 2.2×10^{-16}), indicated that the differences between HALS and RH100 were significant. The regression line model for the standard data-set was $y = 4.33 + 0.837x$ with $R^2 = 0.46$, $Fvalue = 236{,}800$, $p < 0.001$ which became $y = 5.75 + 0.804x$ with $R^2 = 0.54$, $Fvalue = 241{,}000$, and $p < 0.001$ with additional filters. The regression line intersects the 1:1 line at ~ 30 m. Fig. 3.4b further demonstrates that the application of additional filters improves the scatter plot between RH100 and HALS, with a significant reduction in outliers.



(a) Standard GEDI filters

(b) Additional filters

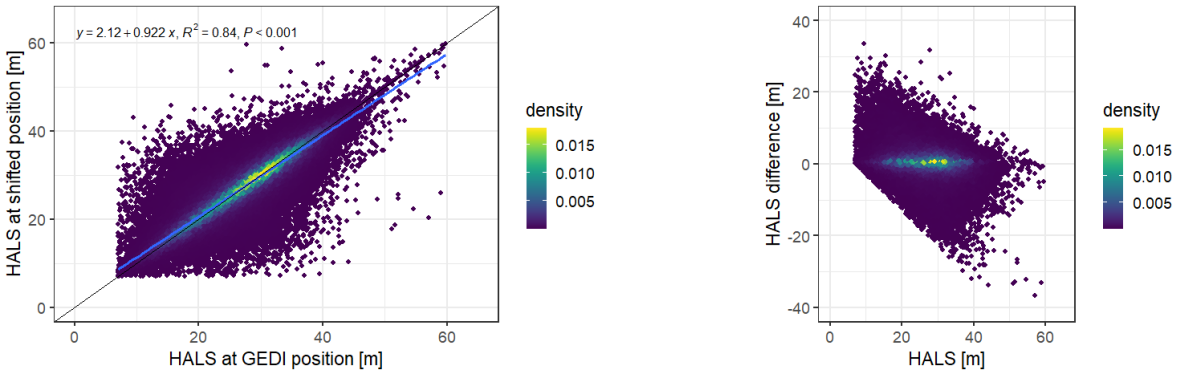
Figure 3.4: Scatterplots of GEDI RH100 and HALS heights, before and after additional filtering.

Impact of poor geolocation accuracy of GEDI footprints Fig. 3.5 illustrates the impact of geolocation inaccuracy. In this figure, the differences in heights result from reproducing a geographical shift using the Schleich et al. (2023c) shift distribution layout on the total data-set (i.e., one iteration). Table 3.1 illustrates the statistical outcomes of the height differences owing to geolocation inaccuracy for all iterations run on the data-set subsets. For the three tested distribution layouts, mean height differences were all zero with standard deviations (sd) ranging from 0.02 m (Schleich et al. (2023c)) to 0.11 m ($\mathcal{N}(10, 20)$). The HALS at GEDI locations and HALS at shifted locations were the most similar when distances were corrected using the distribution proposed by Schleich et al. (2023c), with a mean R^2 of 0.84. Its estimated model intercept and slope were always greater than zero (ranging from 1.76 to 2.53) and lower than one (ranging from 0.91 to 0.93), respectively.

Fig. 3.5a shows that the magnitude and shape of the dispersion observed between the HALS and RH100 (Fig. 3.4) could be partly reproduced by introducing a geographical shift during pairing. Deviations tend to be strongly positive at low height values (heights greater at shifted locations for low height values) and become negative over ~ 30 m (heights smaller at shifted values) (see Fig. 3.5b). Thus, the HALS differences between the announced and unknown real footprint locations resulted in the overestimation of low height values and underestimation of large height values.

Table 3.1: Comparison of ALS heights at the initial GEDI footprint location and at shifted GEDI footprint location (with $5{,}000$ iterations on $20{,}281$ footprint subsets).

	Mean height difference				Correlation				R^2				intercept				slope			
	min	max	mean	sd	min	max	mean	σ	min	max	mean	sd	min	max	mean	sd	min	max	mean	sd
$\mu = 10, \sigma = 10$	-0.32	0.30	0.00	0.09	0.85	0.95	0.91	0.01	0.73	0.91	0.83	0.02	0.84	4.23	2.36	0.48	0.85	0.97	0.91	0.02
$\mu = 10, \sigma = 20$	-0.42	0.46	0.00	0.11	0.78	0.91	0.87	0.02	0.61	0.83	0.75	0.03	1.39	5.65	3.48	0.57	0.79	0.94	0.87	0.02
Schleich et al. (2023c)	-0.07	0.07	0.00	0.02	0.91	0.93	0.92	0.00	0.82	0.86	0.84	0.00	1.76	2.53	2.14	0.10	0.91	0.93	0.92	0.00



(a) (b)

Figure 3.5: Impact of GEDI geolocation inaccuracy on ALS heights by reproducing a geographical shift using Schleich et al. (2023c) distance distribution.

Impact of other factors (season, year, forest stand type, slope, and selected algorithm) We compared the RH100 and HALS data using a simple linear additive model, with both continuous and discrete variables transformed into factors. The residuals exhibited strong skewness, indicating that the model lacked explanatory power. This is most likely mainly due to georeferencing errors, which introduce non-modeling, non-normally distributed noise. This affected the analysis of other influencing factors. In B.4, we explore the residual distributions of the linear regression RH100 $\sim$ HALS according to season, year, forest stand type, slope, and GEDI variable *selected_algorithm*. These factors contribute to the differences between RH100 and HALS; however, quantifying their specific contributions remains a challenge. Trends show that GEDI heights are underestimated in winter compared to summer, and variations occur with forest type, with underestimations in open forests, and the selected algorithm, with a trend to underestimate GEDI heights using Algorithm 1 compared to Algorithm 2. In addition, higher slope percentages correlated with larger residuals.

In short, the NFI Hmax and HALS were strongly correlated and did not differ significantly. Regarding RH100 and HALS, the applied filters successfully improved the correlation. However, the paired t-test showed that GEDI RH100 and HALS differed significantly even after additional filtering. Some of the differences were artifacts explained by georeferencing mismatches between announced and real footprint locations, and not by actual differences. These RH100/HALS artifact differences did not affect the stratification based on RH100 values. However, GEDI data acquisition conditions, i.e. season, forest stand type, slope, and changes in the processing algorithm (ground peak detection) also blurred the expected (1:1) relationship between RH100 and HALS, that is between RH100 and NFI Hmax. Nevertheless, in this study, the maximum heights, RH100 and Hmax, were used as bridge variables for stratification.

3.4.3 Double Sampling for Post-Stratification Approach

Reference stratification using ALS heights for both GEDI footprints and NFI plots for the two positioning scenarios, random and announced GEDI footprint locations, and the three stratifications tested are shown in Fig. 3.6. The estimated mean GSVs underestimated the SRS mean (at $97{,}639{,}473\,\mathrm{m}^3$ for the SRS), but all were within the SRS 95% confidence interval. Stratifications using the GEDI locations had lower GSV estimates than those using random locations. For both positioning scenarios, lower relative efficiencies (REs)

were achieved with stratification of two strata. Stratification into three and five strata showed similar results, with the RE of DSPS varying from 1.46 (i.e., 46% of improvement) with random_3str to 1.55 with gedi_5str.

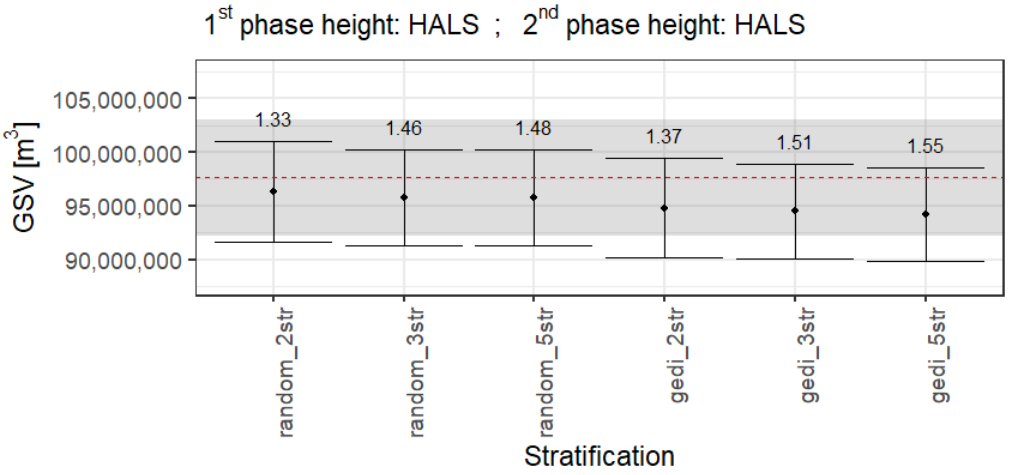


Figure 3.6: Impact of GEDI's spatial sample scheme. Comparison of the growing stock volume (GSV) estimation and variance obtained based on ALS heights using 202,808 footprints and tested with 2, 3, and 5 strata. *random* used Als heights extracted from random locations and *gedi* used Als heights extracted at GEDI locations. *_2str* used 2 strata, *_3str* used 3 strata and *_5str* used 5 strata. SRS volume is presented as a red line and the SRS 95% confidence interval is in dark grey. The relative efficiency (RE) is assessed above each case.

The results obtained using the true NFI heights (Hmax) and both the GEDI and ALS heights at the GEDI locations are shown in Fig. 3.7. As expected, given the strong correlation between Hmax and HALS for the NFI data, the results for the ALS-based stratifications were almost the same as those for the *gedi* stratifications presented in Fig. 3.6. Moreover, the RE achieved with RH100 was in close agreement with that achieved with ALS. Using five strata improved the variance of the GSV estimates by 56% compared with the SRS variance. The RH-based GSV estimations were slightly higher than those of the HALS. The RH GSV estimates were closer to the SRS volume with 95,154,258 m^3, 96,004,131 m^3, and 96,048,008 m^3 for *Rh_2str*, *Rh_3str*, and *Rh_5str*, respectively.

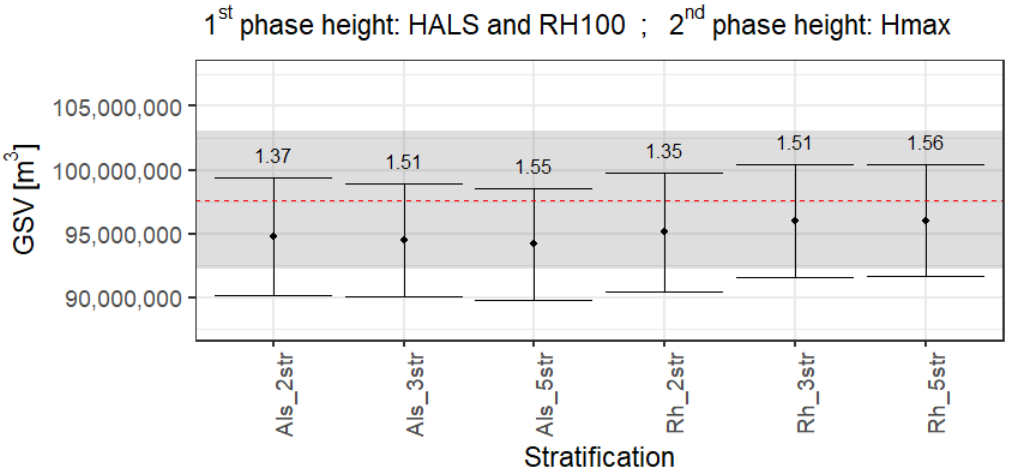


Figure 3.7: Impact of GEDI's height accuracy. Comparison of the growing stock volume (GSV)estimation and variance obtained based on NFI Hmax, and at GEDI footprint locations extracted HALS (*Als*) and GEDI RH100 (*Rh*). Using the filtered GEDI dataset of 202,808 footprints and tested with 2, 3 and 5 strata. SRS volume is presented as a red line and the SRS 95% confidence interval in dark grey. The relative efficiency (RE) is assessed above each case.

We further examined the use of data subsets to analyze the annual and seasonal subsets, as shown in Fig. 3.8. The summer and winter subsets showed significantly different results. The summer estimation approaches the SRS volume estimation with $98{,}240{,}248\ m^3/ha$, whereas the winter estimation largely underestimates the GSV with $93{,}999{,}364\ m^3/ha$ (e.g. for the three strata stratification). The results show fewer differences between years. However, while the 2020 and 2021 stratification GSV estimations, confidence intervals, and RE are very similar to the global data-set (*all*: the same as *Rh* in Fig. 3.7), the stratification for 2022 stands out with a lower estimation, for example, $94{,}981{,}160\ m^3/ha$ with three strata.

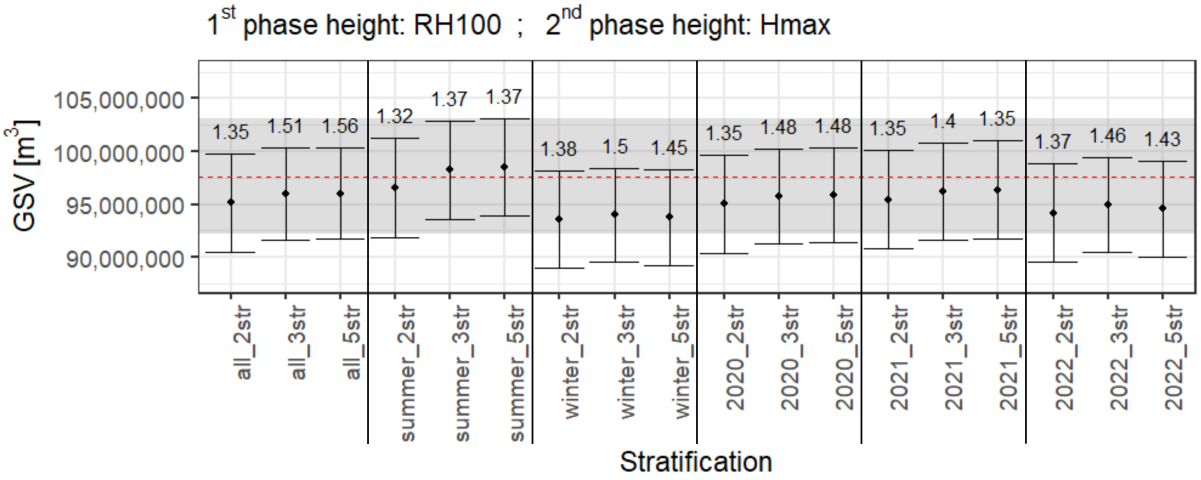


Figure 3.8: Comparison of the growing stock volume (GVS) estimation and variance obtained based on NFI Hmax and GEDI RH100 for the filtered GEDI data-set (all) and yearly and seasonal subsets, using stratifications with 2, 3 and 5 strata. SRS volume is presented as a red line and the SRS 95% confidence interval is in dark grey. The relative efficiency (RE) is assessed above each case.

3.5 Discussion

3.5.1 GEDI sample scheme - neither random nor systematic

The GEDI instrument onboard the ISS captures full-power footprints along four "parallel" ground tracks. The footprints were distributed along the acquisition orbits, creating a form of aggregation by orbit. However, these tracks are not perfectly parallel because of the ISS movements and vibrations (Schleich et al., 2023c). These movements introduce shifts in the GEDI footprint locations, introducing an element of randomness and independence compared to a regular straight-line pattern, which is more akin to clustered sampling. In our study (Section 3.4.1), we compared the surface proportions estimated from the GEDI sampling scheme with those of random and regular sampling schemes using ALS heights. The GEDI-based proportions yielded similar, albeit not identical, proportion estimations to the other schemes. This suggests that, despite its large number of footprints, GEDI's sampling scheme cannot be characterized as a probability-based sampling scheme. Consequently, this introduces a bias in the estimation of the strata proportions. We speculated that these results may be due to quality filtering applied to the footprints. Therefore, we further explored this by estimating the strata proportions of non-filtered footprints and applying different filters (B.3). Using all the full-beam footprints, regardless of quality, brought the proportions closer to those of the reference. While filtering on GEDI quality and degrade flags had a low impact on the total class propor-

tions, applying the additional filters presented in Section 3.3.1 had a more pronounced impact. The filters removed poor-quality footprints, including footprints in sloped terrain, as shown in Fig. B.3b. This, however, resulted in underrepresented sloped areas with the filtered data-set, while sloped areas tended to have a high volume. Applying the filters caused the ALS height proportions to deviate from the reference; however, they brought the RH100-based proportions closer to the reference proportions (Fig. B.3c).

Previous studies have suggested that natural populations in a territory are best sampled based on regular (spatially systematic) sampling, in which the distances between measurement points are constant (Stevens Jr and Olsen, 2004). This type of sampling is superior to uniform random positioning (Christianson and Kaufman, 2016). However, GEDI has no direct control over the location of the footprints, resulting in less-than-optimal sampling. The ISS underwent unexpected changes in its orbital altitude in early 2020, which resulted in nearly 4-day repeating orbiting leading to less uniform coverage than initially planned with the expected randomly processing orbit (Dubayah et al., 2022a). This leads to a pattern with irregular tracks and footprint densities. The ISS orbit was lowered in 2022, resulting in more uniform coverage during the last acquisition period. It is possible that with the planned continuation of the mission from the fall of 2024, the resulting denser and more homogeneous coverage could finally result in a more uniform spatial coverage that is more akin to a probabilistic sampling scheme.

Hence, assimilation to a probability-based sample scheme should be made with great care and cannot be fully confirmed by our analyses. This assimilation may be even more challenging for smaller study sites or smaller GEDI data-sets, where compensation may not occur. However, the smaller yearly and seasonal subsets resulted in similar surface estimations to those of the entire GEDI set, despite their reduced size. The smallest tested data-set included $43{,}869$ footprints acquired over one year at our study site (B.5). This suggests that the geographic positioning distribution, rather than the sampling intensity, is the cause of departure from a probabilistic sampling scheme and remains the largest impediment to the use of GEDI data in design-based approaches. Although the deviations are small (on the order of a few percent for all strata sizes), despite nearly four years of data, GEDI's sampling scheme is insufficient to produce unbiased estimations. Regarding these issues in footprint distribution, a dedicated low Earth orbit satellite embedding the lidar system might offer more suitable conditions than the ISS. This could provide better control of footprint distribution, improve geolocation accuracy, and ultimately lead to improved management of the sampling patterns.

The impact of GEDI's non-probabilistic sampling scheme extends to model-based approaches, necessitating cautious consideration. It might indeed lead to the omission of specific regions, like critical slope areas, and subsequently force models to extrapolate. The risks associated with extrapolation in model-based approaches have been underlined by Renaud et al. (2022). The authors developed a model-based approach using ALS and NFI plots, with $\sim$ 250 calibration plots. Despite the large number of calibration plots, $\sim$ 20% of the pixels in their area of interest were extrapolated with no control over the accuracy of the results. Furthermore, unlike the DSPS design-based approach used in this study, model-based approaches that incorporate GEDI data can be strongly affected by geolocation uncertainties, which introduce substantial errors, as shown in Fig. 3.5. Studies using model-based approaches assume that the height errors between announced and real footprint locations are random; that is, the model is capable of absorbing the consequences of geolocation errors. Our results suggest that these errors introduced biases in height, with a trend of overestimating small heights and underestimating large heights. Owing to the tree height distribution, this trend can be explained by a higher probability of finding a higher or smaller maximum tree height

when shifting the footprint from a plot with a small or high maximum tree height, respectively. Fortunately, the design-based DSPS approach used in this study was not affected by geolocation errors.

3.5.2 Impact of the bridge variable between NFI and GEDI data on the stratification estimations

We developed new filters specifically designed to remove noisy waveforms from the GEDI data. Although the quality_flag and degrade_flag included in the GEDI data provide some level of filtering, we aimed for a stricter selection of good-quality footprints. Other studies directly filtered footprints based on the difference between the GEDI and ALS heights (Morin et al., 2022). We aimed for a purely GEDI-based workflow and did not rely on ALS data for filtering. Fayad et al. (2020) introduced a filter based on the sensitivity variable. We decided not to use this filter because the quality_flag already includes a filter at sensitivity < 0.9 and because our study site has an opener environment than tropical forests, for which a stricter sensitivity was recommended. We tested this additional sensitivity filter and found that it removed too many footprints while maintaining poor-quality ones. Noisy footprints with incorrect RH100 values can significantly affect surface proportion estimations and subsequently impact the results of the DSPS approach. Incomplete filtering of poor-quality GEDI footprints was also identified as an issue by Bruening et al. (2023), who reported reduced bias in GEDI AGBD estimates after additional filtering of footprints. Notably, different system specifications, for example a lidar with a shorter transmitted pulse combined with a higher pulse energy, would probably result in a higher rate of high-quality data. Indeed, the signal-to-noise ratio would be better for ground detection in dense forest stands and ground peak identification on slopes.

In our study, NFI Hmax and GEDI RH100 were chosen as bridge variable for the DSPS approach. Although not perfectly correlated, they exhibited sufficient correlation to implement the DSPS approach. We opted for GEDI RH100 as maximum height. Researchers commonly prefer using RH99, RH98 or RH95 because RH100 may include more outliers (Duncanson et al., 2020; Dorado-Roda et al., 2021). However, RH100 theoretically corresponds to the top of the canopy, and therefore to the maximum tree height. In our study site, all upper RHs were found to underestimate the maximum tree height compared to HALS (see B.2). Thus, RH100 was the most suitable bridge variable available for the maximum height. In the context of stratification, a reduced precision is preferable to a systematic bias. This metric has also been chosen in other studies (Marselis et al., 2019; Adam et al., 2020; Lahssini et al., 2022) and for the production of GEDI level 3 (L3) gridded mean canopy height (Dubayah et al., 2021b).

We believe that putting more effort into poor-quality data filtering and developing an approach to correct the RH100 from the slope effect would solve most of the issues regarding RH100 outliers. Slope was identified as an important factor influencing the quality of the relationship between RH100 and HALS. Several authors (Potapov et al., 2021; Liu et al., 2021; Quirós et al., 2021; Bruening et al., 2023; Schleich et al., 2023c) have highlighted the influence of slope on vegetation height, as assessed from GEDI data. However, as mentioned in Bruening et al. (2023) and Section 3.5.1, applying filters based on topography can breach the random sample scheme approximation of GEDI footprints. When filtering the data, we delete points on steep slopes with a higher canopy height. High canopy heights also indicate large volumes, which may explain why the DSPS estimates were always slightly below the SRS estimates. This limitation affects both design-based (bias in strata proportions) and model-based approaches (extrapolation problems), albeit in different ways. When we are working on height categories, we no longer need a precise estimate of height

using GEDI; we want to classify points, whereas the model-based approach will require a precise estimate. For an area with less relief than our study site, the proportions should be better estimated.

Georeferencing error is the most important factor explaining the differences observed between RH100 and HALS. Our results revealed a correlation coefficient of 0.74 and R^2 of 0.54 between RH100 and HALS. To assess the contribution of georeferencing errors to height differences, we randomized footprint locations and found that $\sim$ 84% (R^2 of HALS and shifted HALS regression in Fig. 3.5a) of the differences could be attributed to geolocation errors. The remaining differences were likely due to measurement effects from either the ALS or GEDI data. Fortunately, geolocation errors do not affect the stratification results or DSPS assessments. However, they significantly penalize the capacity to study the quality of the bridge variable and the influence of other potential sources of differences in RH100 and HALS, such as season, stand type, or slope. This hampers the quantification of the impacts of these other variables and the development of models to correct for RH100 based on their influence. Throughout the study, we were unable to distinguish the effects of geolocation errors from those of other factors to thoroughly evaluate the quality of the chosen bridge variable. Therefore, different strategies can be used for this purpose. The improvement of georeferencing is a complex task. The use of simulations, such as the GEDI simulator (Hancock et al., 2019), could allow for a better study of the impact of other factors.

The difference in plot sizes, with GEDI footprints covering 25 m diameter plots and NFI plots using 30 m diameter plots, may also impact the relationship between both height variables. Additionally, temporal discrepancies between ALS and GEDI acquisitions may account for some differences, as vegetation may have grown or reduced between the two data collections.

Despite the differences in the link variables, the stratifications yielded GSV estimates close to the GSV estimation achieved through SRS, falling within the 95% confidence interval. It is important to highlight that even with a perfect link (Fig. 3.6), differences in the results still arise depending on the chosen number of strata.

3.5.3 Use of DSPS approach with GEDI data

The use of DSPS is a promising way to incorporate GEDI data and help produce multi-source estimations with all the desirable properties of the design-based approach. It is difficult to assess the best stratification results, because there is no perfect reference. NFI plots can be used as a reference, as we did for the SRS estimations and to evaluate the RE, but their low spatial density distinguishes them from all remote-sensed-based estimations. Therefore, the best comparative baseline is the DSPS estimation obtained using the ALS data. Our results showed that the GEDI RH100-based estimations were close to the HALS-based estimations using the same sample size, thus making them directly comparable. Notably, RH100-based estimations were closer to random-sampling-scheme-based HALS estimations than to HALS-based estimations at the GEDI locations. This suggests that the spatial sampling scheme and height accuracy partially compensated for each other. For instance, with three strata, random_str3 with HALS data at random locations yielded 95,737,231 m^3 whereas Rh_3str with RH100 data at GEDI locations yielded 96,004,131 m^3, a difference of less than 0.3% in the total GSV.

These results confirm our assumption that the DSPS approach is fairly robust against multiple errors encountered while using GEDI data. The DSPS is based on estimating and using strata sizes that are not known

prior to sampling but are estimated through sampling. Therefore, some uncertainty exists in the determination of their proportions. Different results were obtained depending on the number of strata used (i.e., 2, 3, or 5). As in McRoberts et al. (2012), the RE tended to be higher when more strata were used. Stratifications with three or five strata yielded higher relative efficiencies than when only two strata were used; that is, variances were reduced by up to 56%. However, NFI users may prefer stratification based only on summer footprints. The RE was lower (variance improved by 37%) but the estimated volume was closer to the SRS volume. The differences observed between the data subsets were due to different surface proportion estimations. For example, winter-based stratification yields a low-volume estimation, which is explained by the overestimation of the (7,20] height class (Fig. B.5b). The differences observed for 2022 cannot be explained by an over-representation of winter footprints because the winter/summer proportions were similar to those of the other years. The differences may be from an unfortunate sampling pattern for this year or may show that GSV decreased, for example, because of bark beetles or cut-downs. Our RE results are consistent with those of other studies and show that the approach has great potential if biases related to GEDI sampling are limited. Roberge et al. (2016) used a two-phase sampling for stratification approach, with existing NFI plots as the first-phase sample and damaged forest plots as the second-phase sample to estimate the damaged area and total number of damaged trees. The approach resulted in an RE of 1.5 compared to SRS estimates. Bullock et al. (2023) estimated AGBD using hybrid inference with GEDI and Paraguay's NFI and found that standard errors were reduced by 47% on average compared with NFI-based estimates alone. McRoberts et al. (2012) assessed the utility of lidar data for post-stratification to increase the precision of mean GSV estimates for the Norwegian NFI. The best RE values achieved were between 2.06 and 3.2. Note that considering the footprints individually instead of as orbit clusters, might have led to a slight overestimation of the SS2 variance estimator, thus to a slight underestimation of the RE in our study.

Although we focused on one strong bridge variable, other variables or combinations of variables could have been tested for stratification. However, finding common variables between the NFI and GEDI plots has proven challenging. Using other variables would most likely require either a model-based link between the NFI and GEDI variables or additional data sources, necessitating an intersection process and therefore rendering the necessity of precise footprint geolocation even more crucial. The DSPS approach could be enhanced by adding auxiliary data. Enriching this approach with spectral data, such as Sentinel-2 images, would supplement information on vegetation composition and density. However, this would require extracting Sentinel-2 data at GEDI footprints, and in a design-based approach, we rely on randomness for compensation, so geolocation errors should compensate for each other on a sample of sufficient size. The method would require a hybrid approach to consider link errors. The method can also be extended to three-phase stratification by adding an additional data source, such as ICESat-2 data. Combining the proposed method with a poststratification method (Roberge et al., 2016), such as the poststratification method used in the French NFI, could also be considered.

Conclusion

Implementing a DSPS design-based multisource inventory approach requires two hypotheses: the existence of a probability-based sampling scheme and a robust link between field and auxiliary data for consistent stratification. An analysis of the GEDI sampling scheme revealed that it differs from probability-based sampling schemes. Applying filters to sort out bad quality GEDI waveforms worsened the probability sample

scheme approximation in the studied mountainous area, because bad quality GEDI data often arise in sloped areas. These areas were sampled less frequently when filtering the data. In this study, maximum height served as a bridge variable between the NFI and GEDI data. However, the analysis showed that NFI Hmax and GEDI RH100 were not perfectly correlated. Slight deviations from the two working hypotheses introduce biases that affect the stratification estimations. Nevertheless, the DSPS approach improved the variance of the GSV estimates by up to 56% compared to the SRS NFI plot-based estimation. GEDI can replace or produce a first-phase sample for NFI, by bringing recent measurements and a high spatial sampling density. Therefore, it could enable the production of new estimations with higher precision without additional costs. The use of the DSPS design-based approach allows accommodation for geolocation errors, because precise geolocation is not considered. However, the selection of the GEDI data subset is very important and can significantly affect the design-based estimations. If no adverse changes in the ISS orbital altitude occur during the second part of the mission, the subsequent sampling scheme should be more suitable for DSPS estimations. Moreover, collocation errors and inaccuracies in height measurements make it challenging to effectively use GEDI in model-based approaches. We drew attention to these limitations, emphasizing the need for careful consideration when incorporating GEDI data into forest modeling workflows.

CHAPTER 4

kNN - Bagging NFI, GEDI, Sentinel-2 and Sentinel-1 data to produce estimates of forest volumes

Anouk Schleich[a], Cédric Vega[b], Jean-Pierre Renaud[b,c], Olivier Bouriaud[b,d], Sylvie Durrieu[a]

[a]UMR TETIS, Univ Montpellier, AgroParisTech, CIRAD, CNRS, INRAE, F-34196 Montpellier, France

[b]ENSG, IGN, Laboratoire d'inventaire forestier, F-54042 Nancy, France

[c]Office National des Forêts, Pôle Recherche et Développement Innovation, F-54600 Villier-les-Nancy, France

[d]University Stefan cel Mare of Suceava, RO-720229 Suceava, Romania

Abstract

This study presents a kNN-bagging approach for predicting forest attributes, specifically growing stock volume (GSV), by integrating optical Sentinel-2, radar Sentinel-1, and spaceborne lidar GEDI datasets following a two-step procedure. First, GEDI variables were imputed to national forest inventory (NFI) plots, then models were developed to predict GSV based on the combined imputed GEDI and Sentinel variables. For GEDI imputation, three strategies were followed, each one using different auxiliary datasets: A) solely relying on Sentinel data, B) using Sentinel data complemented by maximum height values from both GEDI and NFI data sources, C) using Sentinel and height values from an independent canopy height map derived from GEDI and Sentinel through Deep Learning. For each strategy, the models were built using a k-nearest neighbor approach and evaluated on an independent test dataset. The results indicated that the ability of Strategy A to predict forest volume as the percentage of explained variance (R^2) was considerably limited (8%) and the relative root mean squared error (RMSE%) was high (66%). Upon adding heights, Strategy B outperformed the previous strategy, with a R^2 of 58% and a relatively low RMSE% (43%). Compared to Strategy A, the use of estimated heights from a height map improved predictions, but did not outperform Strategy B; R^2 was 34% and RMSE% 53%. Our results suggested that GEDI data could be efficiently combined with Sentinel data to predict NFI volume; however, the model requires well imputed GEDI variables. These imputations yielded better models when GEDI footprint heights were directly used rather than derived from the canopy height map. Therefore, small area model-assisted estimation may be preferred over wall-to-wall maps to avoid local estimation errors.

4.1 Introduction

National forest inventories (NFIs) are designed to produce precise forest resource estimates across vast areas (i.e. national to regional). Various methods for enhancing the precision of estimates over small domains while controlling costs have been developed. Most of these methods rely on remote sensing data or thematic maps. Examples include stratification approaches, where auxiliary data are used to define homogeneous estimation strata (Haakana et al., 2019; Schleich et al., 2024), and multisource forest inventory (MFI) (Tomppo et al., 2008; Saborowski et al., 2010; Westfall et al., 2019), which relies on a model to relate field attributes to auxiliary data, allowing for the derivation of high-resolution maps of forest attributes to support small-area estimation (Guldin, 2021).

Within the multisource estimation framework, moderate-resolution imagery (10-30 m) has been used for its capacity to provide national-scale data on a yearly basis, and its ability to ensure long-term data renewal while having a pixel size in the order of the NFI plots size (Coops et al., 2023; Tomppo et al., 2008). However, optical data have limited correlation with forest attributes in diverse and complex forest structures (Irulappa Pillai Vijayakumar et al., 2019; Saarela et al., 2018). Three-dimensional remote sensing data form airborne laser scanning (ALS) or photogrammetry, which provide key information about forest canopy structure, are more suitable for estimating structure-related forest attributes such as volume, basal area or biomass (Lim et al., 2003). Various European countries, including Finland (Kotivuori et al., 2016), Sweden (Nilsson et al., 2017), Switzerland and France, have initiated nationwide ALS acquisitions to support these needs. A key challenge associated with such aerial surveys is ensuring a consistent and timely update at the national level. Nordic European countries, in particular, have engaged regular ALS acquisitions to that aim (Appiah Mensah et al., 2023). Renewal of 3D data from aerial imagery is also a sustainable solution owing to the regular surveys carried out by most national mapping institutes for several decades (Ginzler and Hobi, 2015). The resulting long time series are also beneficial for the assessment of growth trajectories, which are valuable in monitoring forest dynamics (Véga and St-Onge, 2009). Nevertheless, the time needed to collect data over an entire country continues to be an impediment. While temporal aggregation may be adequate for nationwide estimates of forest resources, it is relatively ineffective in addressing rapid changes in forests, such as those induced by significant disturbances like storms or fires, the frequency and severity of which tend to increase owing to climate change (Forzieri et al., 2021). A potential solution lies in the use of space-borne lidar data, which allows for the monitoring of large areas at a higher temporal frequency, thereby better meeting emerging needs for monitoring changes in forest dynamics.

In 2018, the Global Ecosystem Dynamics Investigation (GEDI) space-borne lidar system was launched to collect forest structure data (Dubayah et al., 2020a). GEDI employs full waveform lidar sensors to capture data within ~ 25 m diameter footprints. Its dense sampling presents potential for enhancing MFIs. In contrast with passive optical systems, GEDI data include several vegetation structural variables such as height profiles and canopy cover, which are highly correlated with forest attributes of interest such as volume or biomass.

Various studies have performed estimations at footprint level or have mapped heights or aboveground biomass density (AGBD) using GEDI (Potapov et al., 2021; Lang et al., 2023; Schwartz et al., 2023). GEDI's L2A and L2B products provide several variables at footprint-level, such as canopy heights and coverage. The L3 product provides 1×1 km gridded mean and standard deviation of canopy height (Dubayah et al., 2021b),

while L4A and L4B products give AGBD respectively at footprint and at 1×1 km grid level (Dubayah et al., 2022b, 2023). Potapov et al. (2021) created a 30 m resolution global canopy height map based on GEDI and Landsat data. Lang et al. (2023) created a similar map, at a finer resolution of 10 m, using GEDI and Sentinel-2. Recently, Schwartz et al. (2023) created a 10 m resolution canopy height map, called FORMS-H, for the French mainland territory based on GEDI, Sentinel-2, and Sentinel-1, using a deep learning approach. Based on this canopy height map and allometric equations, Schwartz et al. (2023) also created an AGBD (FORMS-B) and a wood volume (FORMS-V) map of France at a 30 m resolution and concluded that these maps outperformed existing global and European maps. Their enhanced accuracy was attributed to methodological aspects, and to a more restricted calibration domain, which becomes more diversified from a national to a continental or global scale. Schwartz et al. (2023) have calibrated their model with data obtained specifically from France; therefore, their model is optimized for the characteristics of the French territory. Despite these improvements, the authors highlight the difficulty of deriving an AGBD or volume map from a height product alone.

To take full advantage of the rich structural information of GEDI data to improve biomass or volume predictions, this information must be linked to reference field measurements. Among the attributes observed by NFIs, growing stock volume (GSV) plays a key role in providing essential information for policy decisions and forest management (Gschwantner et al., 2022). GSV, which corresponds to the amount of living standing wood per hectare, indicates the availability of wood resource, an economically and environmentally important variable (Gschwantner et al., 2019). Every year, the French NFI publishes GSV estimates at regional scales. While such information is relevant for national forestry policy purposes, sub-regional to local estimates of GSV are desired for forest management purposes.

The objective of this study is to predict GSV, using GEDI and Sentinel data. Sentinel data have been successfully used to predict a specific GEDI height, such as the 95th or 100th percentiles of relative heights (e.g. RH95 or RH100) measured between the ground and the maximum height (Schwartz et al., 2023; Pereira-Pires et al., 2021)). However, the use of a single height variable is limited in its capability to predict structure-based NFI parameters such as GSV, as height is only a component of GSV. Propagating multiple GEDI heights and other GEDI vegetation derivatives would be more appropriate to that end. However, owing to the tenuous link between Sentinel information and forest structure, we hypothesized that Sentinel data used alone might fail to efficiently bridge this gap and that it needs to be completed with structural information. We assumed that the use of forest height in addition to Sentinel information could constitute an efficient strategy for successfully propagating NFI information at the level of GEDI footprints.

Another specific objective is evaluating different strategies to account for the spatial mismatch of GEDI footprints and NFI plots. To test these hypotheses, three matching strategies were tested. Strategy A only relies on Sentinel information. Strategies B and C rely on both Sentinel data and height information. Two different height sources were compared. The first one relies on information available both in NFI and GEDI datasets. Schleich et al. (2024) found that, among various relative height (RH) candidates, GEDI RH100 and NFI maximum height were correlated best to one another and could be used as a shared bridge variable between the two datasets (Strategy B). However, the use of such height information would limit the prediction of GSV only at the level of GEDI footprints, preventing the realization of wall to wall predictions. Therefore, we ultimately considered using an existing wall-to-wall height map, FORMS-H produced by Schwartz et al. (2023) (Strategy C).

To determine the link between GEDI and NFI and maintain control of the modeling process, we rely on a k-nearest neighbor (kNN) approach (Cover and Hart, 1967). It is a powerful, yet simple, non-parametric supervised approach that can be used for regression, searching the nearest neighbors in the space of variables shared between two datasets, and assigning target variables from the reference dataset to the elements of the other dataset. kNN is widely used to predict forest attributes by combining field and remote sensing data (Tomppo et al., 2008; Chirici et al., 2016; Holmström and Fransson, 2003; Pacheco et al., 2021).

4.2 Study site and data

4.2.1 Study site

Based on bioclimatic and ecological criteria, metropolitan France is divided into 91 sylvoecoregions (Cavaignac, 2009). This geographical subdivision serves as a national reference for forest management. Our study site is located in "Central Vosges Massif", a mountainous sylvoecoregion of North-Eastern France. Elevations range from 200 to $1{,}400$ m. The stand compositions differ considerably, depending on site conditions (altitude, slope, exposure, soil type, available water, etc.) The forest is primarily composed of European beech, silver fir, and Norway spruce trees, either in pure or mixed stands (beech-fir stands) (IGN, Sylvoécorégion). The Vosges study site used to develop the proposed MFI approach is defined by the intersection of the extent of a Sentinel-2 tile with the "Central Vosges Massif" sylvoecoregion (Fig. 4.1). It covers an area of $4{,}634$ km^2.

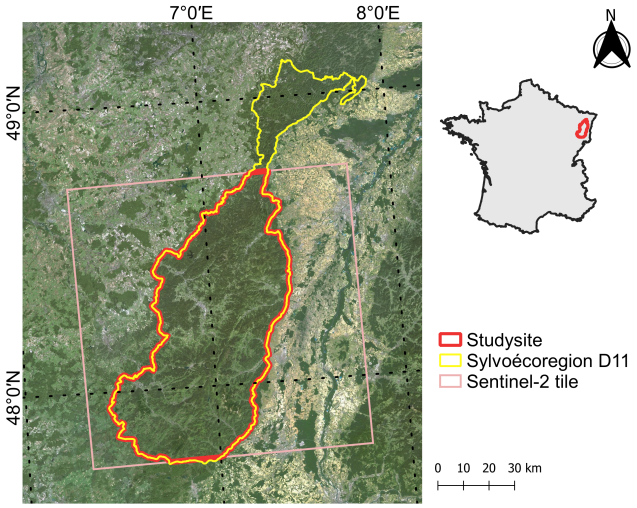


Figure 4.1: Study site in North-Eastern France

4.2.2 Data

Three main datasets were used: GEDI products, NFI field plots and Sentinel (1 & 2) images (Fig. 4.2). We also used a forest map (IGN, BD Forêt version 2) to define a forest mask, and a 1-m digital elevation model (IGN, RGE ALTI) to improve GEDI footprint locations. Furthermore, a 10 m resolution height map (FORMS-H, Schwartz et al., 2023) was used to obtain spatial information on height, a crucial structural variable, which is only available in discrete form with GEDI footprints. Finally, a 30 m resolution wood volume map (FORMS-V, Schwartz et al., 2023) was used to compare final GSV estimates. All data are summarized in Table 4.1.

Table 4.1: Datasets used in this study

Data	Description	Product	Source	References	Variables	Date	Filters or preprocessing	N	Utility in study
RGE Alti	1-m resolution Digital Elevation Model	RGE ALTI 1M	https://geoservices.ign.fr/rgealti	[114]					to apply GeoGEDI
GEDI	full waveform spaceborne lidar 25 m circular footprints	L2A product v2	https://lpdaac.usgs.gov https://search.earthdata.nasa.gov/search	[50]	elev_lowestmode, lat_lowestmode, lon_lowestmode, delta_time	Apr 2019-Mar 2023 Summer: Mai - Sep	quality_flag = 1 degrade_flag = 0 beam = full power (extented study site)	268 109 footprints	to apply GeoGEDI
					Relative heights: RH100, RH90, RH80, RH70, RH60, RH50, RH40, RH30, RH20, RH10		quality_flag = 1 degrade_flag = 0 beam = full power Additional filters [206] RH100 <= 60 RH100 >= 7 in BD Forêt	104 831 footprints	kNN
		L2B product v2		[52]	canopy cover cumulative canopy covers pai vertical pai profile pavd profile fhd_normal		use footprints filtered in L2A	104 831 footprints	kNN
NFI	data from french forest inventory		https://inventaire\-forestier.ign.fr	[112]	growing stock volume (GSV) (i.e. variable named vwac)	2018 - 2022		675 plots	kNN
Sentinel-2 (S2)	optical multispectral spaceborne images spectral bands in the visible, near infrared (NIR), and short-wave infrared (SWIR)	L2A product	https://earthengine.google.com	[36]	bands: B3, B4, B8	12 Aug 2022	choose image with 0% cloud mask	1 image with 3 bands	kNN
		Theia L3A product	https://theia.cnes.fr/	[91]	bands: B2, B3, B4, B5, B6, B7, B8, B8A, B11, B12	Aug 2022 Aug 2017 Jun 2022		3 monthly syntheses images with 10 bands	kNN
Sentinel-1 (S1)	Synthetic Aperture Radar (SAR)	Ground Range Detected (GRD)	https://earthengine.google.com	[35]	Ascending VV, Ascending VH, Descending VV, Descending VH	August 2022	orbits with VV and VH polarizations	total of 37 orbits: 17 ascending and 20 descending; resumed to one monthly mean by polarization and by direction	kNN
BD Forêt	database of polygons providing information about french forest stands	BD Forêt v2	https://geoservices.ign.fr/bdforet	[111]					to filter GEDI footprints
FORMS-H	National Canopy Height Map			[212]					kNN Strategy C
FORMS-V	National Wood Volume Map			[212]					final comparison

National Forest Inventory Plots

The French NFI is a two-phase continuous inventory, based on a 1 km sampling grid (Bouriaud et al., 2023). Each year, one-tenth of the grid cells is covered. The first phase sample is drawn from the yearly grid fraction and constitutes photo-interpreting aerial photographs to estimate forest cover ($\sim$ 100,000 points/year). The second phase sample is drawn from the first phase one classified as forest and comprises field plots measurements ($\sim$ 7,500 points/year) for estimating forest resource attributes. Tree measurements are made in concentric plots of 6 , 9 and 15 m radii according to their diameter at breast height (DBH; [7.5;22.5[, [22.5;37.5[, [37.5;$+\infty$[(in cm), respectively). Along with the species identification and vitality information (dead or

alive), three main variables are measured: DBH, total height (H), and stem height with up to a $7\,\text{cm}$ stem diameter (Hdec, used to determine the solid stem volume). Species, vitality, and DBH information are collected for all trees, and both H and Hdec are measured for a single representative tree within each species and DBH classes.

The GSV is subsequently calculated using species-specific volume equations and imputation methods involving the tree DBH, H, and Hdec. Plot-level attribute density values are then computed by aggregating individual tree data and using tree inclusion probability. Official statistics are consolidated using five-year moving averages (IGN methodology, 2023). For this study, samples from 2018 to 2022 were used. Over the area of interest, 675 plots were inventoried and the mean GSV of the Sylvoecoregion was estimated to be $295 \pm 20\,\text{m}^3\,\text{ha}^{-1}$. For comparison, the mean GSV over France was estimated to be $173 \pm 3\,\text{m}^3\,\text{ha}^{-1}$ over the same period.

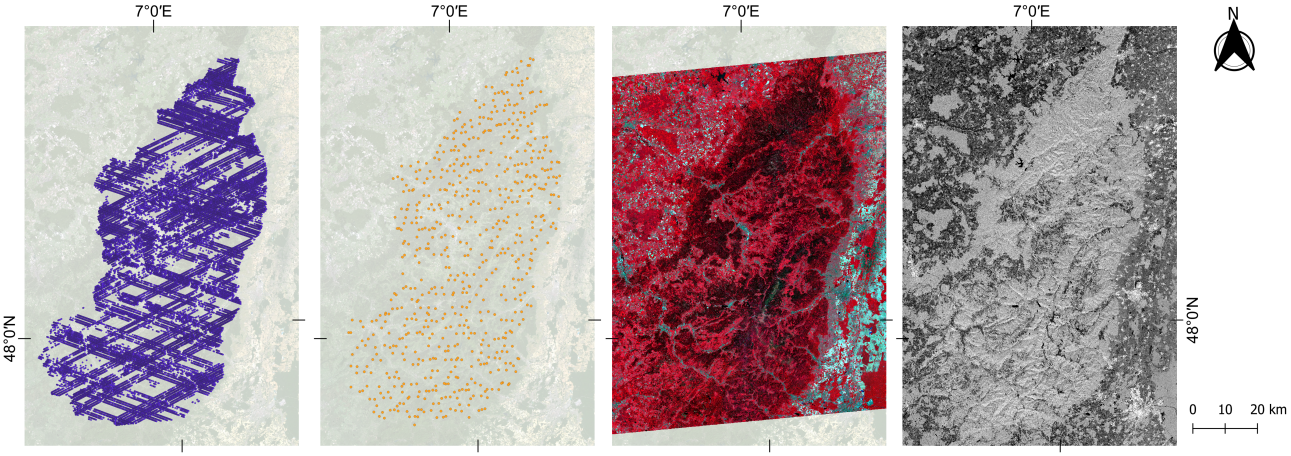


Figure 4.2: GEDI footprints, NFI plots, Sentinel-2, and Sentinel-1 for the Vosges study site. Footprints were filtered to intersect with BD Forêt polygons. Moreover, multiple quality-based filters were applied, resulting in $\sim$105,000 GEDI footprints. The NFI data consist of 675 plots.

GEDI data

GEDI is a spaceborne lidar system mounted on the International Space Station. It is an active sensor, emitting $1{,}064\,\text{nm}$ laser pulses and measuring the backscattered signal over $\sim 25\,\text{m}$ diameter footprints on the ground (Dubayah et al., 2020a). Launched by NASA in 2018, during its initial phase, the GEDI mission collected data from April 2019 to March 2023. The instrument has since been moved and acquisitions have been paused. The resumption of data acquisition is planned for the fall of 2024 (LP DAAC, 2023). This study uses data collected during leaf-on seasons, each spanning from May to September, of the first phase of the mission.

Various filters were applied to ensure footprint data quality. Footprints acquired using full-power lasers with good degrade and quality flags were retained. Additional filters based on GEDI Level 2A (L2A) data proposed by Schleich et al. (2024) were also used. Moreover, GEDI footprints were filtered using a forest mask based on BD Forêt (IGN, BD Forêt version 2) to identify forest footprints and discard non-forest ones. An additional height criterion was introduced to exclude footprints with a maximum height (RH100) below $7\,\text{m}$ or above $60\,\text{m}$. This allows for exclusion of the footprints located in areas where logging has occurred and alleviates unrealistic height estimates. After applying all these filters, $104{,}831$ footprints were retained.

Two GEDI products at the 25-m footprint level were used: L2A data providing relative height (RH) metrics and Level 2B (L2B) data, including estimates of canopy cover (cover), plant area index (pai), plant area volume density (pavd), and foliage height diversity index (fhd_normal). Pavd and both cumulative cover and pai from height z to the ground were available at a vertical step size of $dz = 5$ m; for example, cover_z1 for the forest cover at 5 m over the ground and cover_z2 for the forest cover at 10 m above the ground. Pai is also cumulative, therefore pai_z2 contains all pai between the ground and and a height of 10 m. Pavd, however, constitutes individual layers, pavd_z2 contains pavd from 5 m to 10 m heights.

RH values were transformed to vegetation RH, representing the relative height metrics at 1 % interval for the vegetation component of the backscattered signal. To that aim, for each footprint, a simplified waveform was reconstructed from RH values and a Gaussian function was adjusted to the ground return. Then, the ground component of the signal was removed by subtracting the Gaussian ground peak from the waveform. Finally, the resulting vegetation waveform was transformed back into Relative Height metrics, i.e. heights above the ground at which a certain percentile of the energy returned by the vegetation is reached. These refined vegetation RH values will further be referred to as RHv. Additional variables were derived from the RHv values to provide supplementary information related to the shape of the RHv profile and thus to vertical structure. G_cor represents the correlation between the RHv profile and a straight line connecting RHv_0 (i.e. the ground level) to RHv_100. G_Hr_chgt indicates the first height where the RHv profile intersects the RHv_0-RHv_100 line. This height is expressed as a percentage of the maximum height (RHv_100). G_coeff_var is the coefficient of variation of the RHvs calculated as the ratio of the standard deviation to the mean.

Moreover, cover_z, pai_z, and pavd_z variables were used to estimate additional variables. Depending on already available data, the vertical z step values were used to estimate the variables from different perspectives: cumulative (already given for cover and pai) and by Z step values (already given for pavd), from top to bottom, and from bottom to top. Variables from the top to the bottom were organized in alphabetical order, i.e. the pavd from the highest 5 m layer detected in the footprint, were called pavd_za. Layer wise cover (and pai) variables were named cover_z3z4 when including the cover between 15 and 20 m.

Two different geolocations were used for GEDI: the one provided by GEDI data version 2, and the one obtained after correcting positions available in version 2 using the GeoGEDI algorithm (Schleich et al., 2023c). GeoGEDI corrects GEDI footprint positions using a digital elevation model (DEM). In this study, a publicly available 1-m lidar DEM was obtained through the French mapping agency (IGN, RGE ALTI).

Sentinel-2

The Copernicus Sentinel-2 (S2) mission is composed of two identical satellites acquiring optical multi-spectral imagery. The two satellites share an orbit, phased 180° apart, thereby yielding a revisit time of 5 days. The multi-spectral instrument (MSI) is a passive sensor measuring the Earth's reflected radiance across 13 spectral bands in the visible, near-infrared (NIR), and shortwave infrared (SWIR), ranging from 443 to $2{,}190$ nm. Spatial resolutions vary between 10, 20, or 60 m depending on the band. S2 is widely used for land cover monitoring, including applications in agriculture, forestry, and disaster prevention and management (ESA Sentinel-2).

This study uses Sentinel-2 L2A and Sentinel-2 Theia L3A products. The surface reflectance L2A product

(Copernicus Sentinel-2), downloaded from Google Earth Engine (Gorelick et al., 2017), offers atmospherically corrected data and is available at 110x110 km^2 ortho-image tiles. Our study site corresponds to tile ULU32. The Theia L3A product is a monthly (45 days) cloud-free synthesis of the L2A product using the Weighted Average Synthesis Processor (WASP) (Hagolle et al., 2018). For this study, we downloaded the L3A product for August 2022 as the reference image. L3A images from August 2017 and June 2022 were also downloaded to generate indices related to temporal changes.

For these three L3A images, vegetation indices were calculated using the $10\,m$ resolution bands in the visible (B2 Blue, B3 Green, B4 Red) and NIR (B8). The indices are all defined in the Index DataBase (Henrich et al., 2012).

$$ndvi = (B8 - B4)/(B8 + B4) \tag{4.1}$$

$$msavi = 0.5 * (2 * B8 + 1 - \sqrt{(2 * B8 + 1)^2 - 8 * (B8 - B4)}) \tag{4.2}$$

$$ndwi = (B3 - B8)/(B3 + B8) \tag{4.3}$$

$$gli = (2 * B3 - B4 - B2)/(2 * B3 + B4 + B2) \tag{4.4}$$

The normalized difference vegetation index (ndvi), as introduced by Rouse et al. (1973), measures the distinction between NIR light, reflected by vegetation, and red light, absorbed by vegetation. It is widely used to monitor vegetation density and health. The modified soil adjusted vegetation index (msavi) closely resembles ndvi but mitigates soil noise by accounting for its influence (Qi et al., 1994). The normalized difference water index (ndwi) using NIR and green light, is used to observe water content within vegetation canopies (Gao, 1996). Using only visible bands, the green leaf index (gli) (Louhaichi et al., 2001; Hunt Jr. et al., 2011) has proven effective in vegetation detection (Eng et al., 2019).

Time series indicators were calculated by subtracting indicators from August 2017 to the reference image for an annual change and subtracting June 2022 from the reference image for seasonal change.

In addition, using the R package prosail (Feret and de Boissieu, 2023), we calculated through an hybrid inversion process the leaf area index (lai), the fraction of vegetation cover (fCover), i.e. the gap fraction for nadir direction, and the Fraction of Absorbed Photosynthetically Active Radiation (fAPAR) (Weiss et al., 2016). To run the existing code to calculate these biophysical variables the original L2A data format was needed. Therefore, we used a single cloud-free L2A image acquired on August 12, 2022, i.e. within the time period used to compute the August 2022 L3A product.

Sentinel-1

Sentinel-1 (S1), a synthetic aperture radar (SAR) satellite mission developed by ESA, uses C-band microwave pulses (around 4 to $8\,GHz$) directed towards the Earth's surface. S1 offers the advantage of day-and-night acquisitions unaffected by cloud cover. Two satellites were operated initially; however, a single satellite awaiting a twin is operated currently, resulting in a revisit time of 12 days (ESA Sentinel-1).

S1 provides high-resolution images by capturing backscattered signals. The radar waves' interaction with the Earth's surface, and therefore the backscattered signal, depends on factors such as roughness, moisture content, and geometry of the target. S1 is equipped with different polarization modes (vertical-vertical (VV), vertical-horizontal (VH), horizontal-vertical (HV), and horizontal-horizontal (HH)) that provide

information about the target's characteristics. Using Google Earth Engine (GEE), Level 1 ground range detected interferometric wide-swath (L1 GRD IW) products, with both VV and VH polarizations, acquired during the same period as the 45-day period of Theia L3A of S2 for August 2022 were selected, resulting in a total of 37 orbits, split between 17 ascending and 20 descending orbits. Each point in the study site was observed from 7 to 13 available orbit images (descending and ascending combined). The L1 GRD IW product consists of projected to the ground range SAR data, at 10 m resolution.

To mitigate speckle, mean values for each polarization (VV and VH) and orbit direction (ascending and descending) were retained, i.e. mean values of 3 to 7 images. The four resulting images were corrected from the influence of topography on the backscattered values using an angular-based radiometric slope correction algorithm from Vollrath et al. (2020a) ready to run with GEE and available on github (Vollrath et al., 2020b). We also calculated the average of the corrected ascending and descending mean images. This results in six images: VVasc, VHasc, VVdesc, VHdesc, VVascdesc, and VHascdesc.

Moreover, ratio and radar vegetation index (rvi) were calculated after normalizing the VV and VH values by minimum and maximum. rvi is an alternative to ndvi with optical images (Kim and Zyl, 2009; Charbonneau et al., 2005; Nasirzadehdizaji et al., 2019).

$$Ratio = VV/VH \tag{4.5}$$

$$rvi = (4 * VH)/(VV + VH) \tag{4.6}$$

4.3 Methods

4.3.1 Data preparation

NFI dataset

In contrast with Strategy A, which relies solely on Sentinel information for imputing GEDI values, Strategy B (see Section 4.3.2) takes advantage of the maximum heights to supplement the Sentinel data. In GEDI products, RHv_100 represents the maximum height of a footprint. The default NFI data provide dominant height and not maximum height that could be related to RHv_100. To address this disparity, we augmented the NFI dataset by imputing a maximum height to each NFI plot. This imputation was performed using MissForest on the forest database, as presented in Schleich et al. (2024). The subsequently added maximum height from the NFI data is denoted as Hmax. In addition to Hmax, two other variables from the NFI dataset will be used. The first variable, labeled *vwac* in the official French NFI database, corresponds to the GSV. The second variable, *esspre_fr*, defines whether the NFI plot is dominated by coniferous or deciduous trees. The NFI dataset is composed of 415 coniferous plots with a mean and standard deviation GSV of 360 ± 207 m^3 ha^{-1}, and of 260 deciduous plots with a mean and standard deviation GSV of 227 ± 155 m.

GEDI footprint georeferencing

GEDI footprints have a low geolocation accuracy (~ 10 m for version 2 (Beck et al., 2021)), which may cause problems when coupling GEDI data with other georeferenced datasets. Therefore, we applied the GeoGEDI algorithm outlined by Schleich et al. (2023c) to improve the geolocation of GEDI footprints. In Schleich et al. (2023c), the algorithm was applied by four full-power beams and beam per beam. The latter revealed that horizontal shifts tended to align with beam pairs originating from the same laser. Therefore, as suggested by Schleich et al. (2023c), acknowledging that a beam-pair approach allows for a more optimized topographic characterization compared to a single-beam transect, in this study, we opted for the GeoGEDI algorithm with beam-pair implementation. To identify the optimal location, a search window of 30 m with a step size of 1 m was used around each footprint position. Notably, Schleich et al. (2023c) used a search window of 50 m with a step size of 2 m and observed that, for GEDI v2, most footprints were shifted by less than 30 m. Reducing the search window to 30 m allowed us to downscale the step size to 1 m, otherwise limited by computational execution time.

Mean, median, and standard deviation of the distances between original GEDI v2 positions and corrected GeoGEDI positions were 9.20 , 7.29 , and 7.19 m, respectively.

Auxiliary variables

The following variables were considered for modeling:

- Sentinel-2:
  - L3A data August 2022
    - Bands: B2, B3, B4, B5, B6, B7, B8, B8A, B11, B12
    - Indices: ndvi, msavi, ndwi, gli
    - Monthly differences of indices with June 2022
    - Yearly differences of indices with August 2017
  - L2A data August 12th 2022
    - Biophysical variables: lai, fCover, fAPAR
- Sentinel-1 Mean of August 2022
  - VVasc, VHasc, VVdesc, VVasc, VVascdesc and VHascdesc
  - Indices: Ratio, rvi for asc, desc and ascdesc
- FORMS-H

Data framework for modeling

GEDI and NFI data were intersected with Sentinel and FORMS-H data using spatial intersection, with X and Y coordinates in Lambert 93 as the coordinate reference system. To account for the different spatial resolutions of the data, zonal statistics were used to summarize Sentinel and FORMS-H data at both GEDI footprint and NFI plot centers. Zonal statistics consisted in the mean and standard deviations of all the variables described in Section 4.3.1. The statistics were computed using radii of 15 m and 50 m. The first radius was set

to match the NFI plot radius, and the scale at which dendrometric measurements are made in the field. The second radius was set to consider the spatial context around the plot.

To account for the accuracy of GEDI positions, two versions of GEDI were used: one using initial GEDI v2 positions and another using GeoGEDI-corrected positions.

For clarity, variable names were formed using a prefix characterizing the data source (S1 for Sentinel-1 and S2 for Sentinel-2), followed by the variable name (e.g. B4 for S2 band 4), followed by a suffix made of the aggregation radius (i.e. 15 or 50) and the statistic (mean or sd) when appropriate (for example, S2_B4_15_mean is the mean of Sentinel-2 band 4 computed using a 15 m radius). For the monthly and yearly differences of indices, prefixes MD and YD were added, respectively. Notably, in one modeling strategy (see Section 4.3.2 hereafter), height information was obtained from both GEDI (RHv_100) and NFI (maximum height) as a bridging variable. The variable is denoted Hmax. Similarly, when GEDI data are used as auxiliary data (see Section 4.3.2), a similar naming convention is used with a prefix G.

4.3.2 kNN-Bagging Approach

Modeling is based on random patch kNN bagging regression. It consists in an ensemble of kNN imputations, constructed from a random subset on both instances and features. While kNN has been widely used in MFIs owing to their capacity to predict multiple attributes with a single model, random patched kNN bagging was found to be an efficient approach toward competitive predictive performance with high dimensional data (Louppe and Geurts, 2012; Gomes et al., 2021). In this study, the kNN models are trained using an Euclidean distance, a neighborhood (k) of 1, and an inverse weighted distance. The algorithm is applied n times (n = 1000). For each ensemble, the random patch is formed using sampling with replacement for the instances (i.e. the lines) and sampling without replacement (set to 3 to 6) for the features (i.e. the columns) (Ho, 1998). From the resulting ensemble of predictions, an aggregated predictor is generated for each point (Breiman, 1996).

The modeling framework constitutes two steps described below. The first step involves spatializing GEDI data using Sentinel (Fig. 4.3). The second step involves spatializing NFI attributes using both Sentinel data and GEDI imputed values (Fig. 4.4). In each step, the predictions were evaluated using an independent test dataset.

Step I: Imputing GEDI variables using Sentinel data

For all strategies, a test dataset of 500 footprints was randomly extracted from GEDI footprints for validation purposes. The remaining GEDI footprints (N$\approx 104{,}300$) served as a training dataset for the kNN. Before modeling, dimension-reduction is conducted in the common variable space to create the feature vector, similar to the method mentioned by Sagar et al. (2022). Auxiliary variables (i.e. S1, S2, and possibly auxiliary height depending on the strategy) correlated by more than 0.33 to one of the GEDI variables and correlated by less than 0.85 between each other were selected. The latter criterion was aimed at limiting autocorrelation between the selected variables, which otherwise may pose a risk of overfitting (Moser et al., 2017).

kNN-bagging was performed under R using the package yaImpute (Crookston and Finley, 2008). From the ensemble predictions, for each predicted variable, the median value is calculated and regarded as the

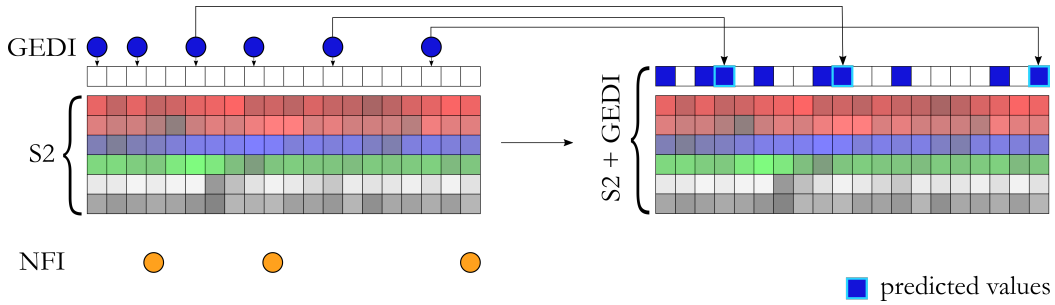


Figure 4.3: Step 1: GEDI variables are imputed over NFI plots using auxiliary data, e.g., Sentinel-2 (S2). Colors are only a figurative way of suggesting the variability in auxiliary data, for instance the different values of S2 bands and indices.

predicted value. The median was selected here rather than the mean due to the risk of outliers caused by issues in lidar signal analysis. Analyzing the test dataset comprising 500 GEDI footprints, theoretically, the imputed median values should align with the observed GEDI values. Nevertheless, a recognized phenomenon in regression analysis is the tendency for values to regress towards the mean, implying that for small values, imputations tend to overestimate the observed values, and for large values, imputations tend to underestimate the observed values. This introduces an additional source of correlated imputation errors, as suggested by Ehlers et al. (2018). To address this issue and enhance the established relationship, we applied a linear regression for each RHv variable on the 500 test footprints. In the context of GSV, this classical calibration procedure, reducing correlated errors, has also been used by Lindgren et al. (2021). First, using the median imputations and observed values, for each variable g, linear regression parameters A_g and B_g are estimated for the GEDI test dataset according to Eq.1:

$$\hat{y}_{i,g} = A_g + B_g y_{i,g} + \varepsilon \tag{1}$$

where:

$\hat{y}_{i,g}$ = median imputation of variable g
A_g = linear regression intercept of variable g
B_g = linear regression slope of variable g
y_g = observed value of variable g
ε = the error term supposed to be normally distributed.

Subsequently, using the linear regression parameters A_g, and B_g, the systematic errors are removed by classical calibration (Osborne, 1991) so that calibrated imputations ($\hat{y}_{i,g,c}$) are obtained as presented in Eq.2:

$$\hat{y}_{i,g,c} = \frac{(\hat{y}_{i,g} - A_g)}{B_g} \tag{2}$$

where:

$\hat{y}_{i,g,c}$ = calibrated median imputations of variable g
$\hat{y}_{i,g}$ = median imputation of variable g
A_g = linear regression intercept estimated in Eq.1
B_g = linear regression slope estimated in Eq.1.

The optimal GEDI footprint is determined by selecting the footprint minimizing residuals of its RHv profile to the calibrated median profile of the 1000 candidates. The different RHv percentiles were scaled, to ensure equal weight.

Step I therefore allows for the matching of an existing GEDI footprint to each NFI plot. Therefore, all GEDI variables, and not only the variables used for the selection of the best imputation, are imputed.

Before moving to Step II, we filtered out poorly imputed GEDI variables. In the assessment of observed and imputed values on the test dataset, variables with a correlation below 0.6 were considered to be poorly imputed and subsequently excluded from consideration in step II.

Step II: Predicting GSV at the level of GEDI footprints

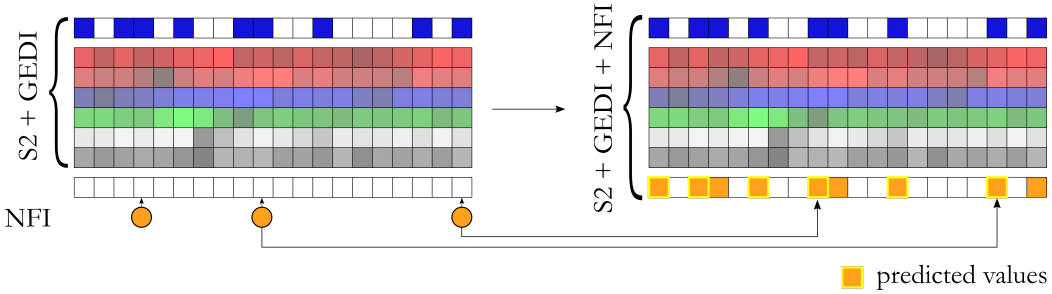


Figure 4.4: Step 2: predicting GSV based on auxiliary data and imputed GEDI data from step 1. Colors are only a figurative way of suggesting the variability in auxiliary data, for instance the different values of S2 bands and indices.

Similar to Step I, the enlarged common variables space is used to develop a model for predicting GSV from GEDI, S1, and S2 features, and from FORMS-H for Strategy C. The model is developed using the GEDI-Sentinel-NFI dataset (Fig. 4.4). To evaluate the method, a 20 % subset of NFI plots was used as test dataset. The remaining GEDI-Sentinel-NFI plots serve as training dataset. Variable reduction is the same as in Step I. Among all the variables, i.e. well imputed GEDI variables from step I (i.e. correlation >= 0.6), S1 and S2 and, for Strategy C, FORMS-H metrics, variables correlated by more than 0.33 to NFI GSV and correlated by less than 0.85 between each other were selected. The kNN algorithm is run 1000 times. Similar to step I, the number of features to be used was randomized to retain 3 to 6 features, if there were more than 3. If there were fewer than 6 features, the number of features to use was randomized between 3 and the number of features, and if there were exactly 3 features, for each ensemble, the number of features to use was randomized to be either 2 or 3.

GSV estimates at NFI plot level are considered accurate. Given the improved accuracy (compared to GEDI signal analysis) and the considerably smaller NFI plot dataset, compared to the GEDI dataset, the mean value of the predictions of the 1000 kNNs was used to predict the GSV. This kNN-bagging approach results in GSV estimates for all plots in the test data and in a model that can be further applied to predict GSV at the level of each GEDI footprint (Strategy B) or possibly at each cell of a 30 m grid, provided that all auxiliary data are available wall-to-wall (Strategies A and C).

4.3.3 Analysis of Results

The different strategies were compared to assess if Strategy A (using only S1 and S2 data) allows for the improvement of GSV estimates, and how the other two strategies performed. For this comparison, we performed the following analysis in particular:

- The variables selected for the first step depending on the strategies and their ability to impute GEDI variables. Their ability to impute GEDI variables is assessed through the analysis of residuals and correlations between observed and imputed values for the test dataset.
- The variables chosen for the second step and their ability to predict GSV at the footprint level. To analyze the ability to predict GSV, using the test dataset, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), MAE%, RMSE%, R^2, Pearson's correlation coefficient and boxplots of errors are calculated between observed and predicted GSV values. The distribution of standard deviation errors of the three strategies will also be compared.
- We compared the GSV predictions generated by our three strategies with those from FORMS-V (Schwartz et al., 2023). The GSV map derived by Schwartz et al. (2023) was derived from their height map (FORMS-H), using two distinct allometric equations – one for deciduous and another for coniferous plots.
- To enhance the comparability of our strategies with FORMS-V, given that the map was created without georeferencing improvement and relied on two allometric equations, we implemented the strategy that performed best using GEDI v2 coordinates and developed separate models for deciduous and coniferous plots. For both v2 and improved footprint coordinates, we will present results using the overall kNN including all data, as well as results from the same overall kNN model categorized by forest stand types. Furthermore, we will present outcomes from the individually conducted coniferous kNN and deciduous kNN. Results will be presented according to stand types and aggregated for direct comparison with the overall kNN regarding prediction accuracy on the entire test dataset.

4.4 Results

This Section is divided into three main subsections. First, results from Step I and II are presented in Sections 4.4.1 and 4.4.2. In addition, alternative setups using stand-specific kNN and offering insights on results without geolocation improvement are presented in Section 4.4.3

4.4.1 Step I: Imputing GEDI variables

Auxiliary variables selected for kNN in step I

The size of the feature space used to impute GEDI variables has been significantly reduced using the variable reduction strategy based on correlations (4.3.2). In Table 4.2 the selected variables are listed in decreasing order of their correlation with GEDI variables.

Auxiliary variable	Most correlated GEDI variable	Corr
Hmax for Strategy B	RHv_100	1.00
FORMS-H_15_mean Strategy C	RHv_90	0.71
S2_greenness_15_mean	cover_zf	0.40
S2_fAPAR_15_mean	pavd_z3	0.40
S2_ndwi_15_mean	cover_zg	-0.35
S2_YD_ndvi_15_sd	cover_z2	-0.35
S2_ndvi_50_mean	cover_z2	0.34
S2_B4_15_sd	cover_z2	-0.34

Table 4.2: Variables selected for step I. The first column contains the selected variables, the second column contains the GEDI variable to which the auxiliary variable is the most correlated, and the third column indicates the correlation between the two. Variables start with "S1" or "S2" if they originate from Sentinel-1 or Sentinel-2 data, followed by the variable name, followed by 15 or 50 depending if the $15\,\mathrm{m}$ or $50\,\mathrm{m}$ radius was used for the extraction of the variable. Variables end with "_mean" or "_sd" depending on the zonal extraction of mean or standard deviation. Auxiliary height for Strategy C is FORMS-H_15_mean. Auxiliary height for Strategy B is Hmax from GEDI and NFI.

In Strategy B, Hmax (i.e. RHv_100) is used as the auxiliary height. Consequently, the maximum correlation between the auxiliary height and the GEDI variables is 1. When using FORMS-H data, i.e. in Strategy C, the strongest correlation is observed with RHv_90, which is 0.71. All other variables are consistent across strategies. Only six variables were retained among the more than 150 initial S1 and S2 variables. Evidently, no S1 variable was selected, all of them exhibited a weak correlation with GEDI variables (-0.33 < Corr < + 0.33), the strongest correlation being equal to -0.22 for S1_vvAscDesc_15_sd.

The S2 variable most correlated to GEDI data is S2_greenness_15_mean, followed by non-autocorrelated variables S2_fAPAR_15_mean, S2_ndwi_15_mean, S2_YD_ndvi_15_sd, S2_ndvi_50_mean and S2_B4_15_sd. Interestingly, S2 variables are more correlated to GEDI cover variables than to the height ones. The most correlated GEDI variables to S2 data are cover_zf, pavd_z3, cover_zg and cover_z2.

Notably, autocorrelated variables were excluded (abs(Cor) > 0.85). Therefore, while other variables (lai_mean, fCover_mean, msavi_mean and ndvi_sd) had a correlation exceeding 0.33 with GEDI variables, they were not selected as their correlation to a previously chosen variable exceeded 0.85.

Evaluation of the kNN regression model used to impute GEDI variables

Fig. 4.5 shows results obtained on the 500 GEDI test footprints dataset, comparing imputed to observed values for G_RHv_100.

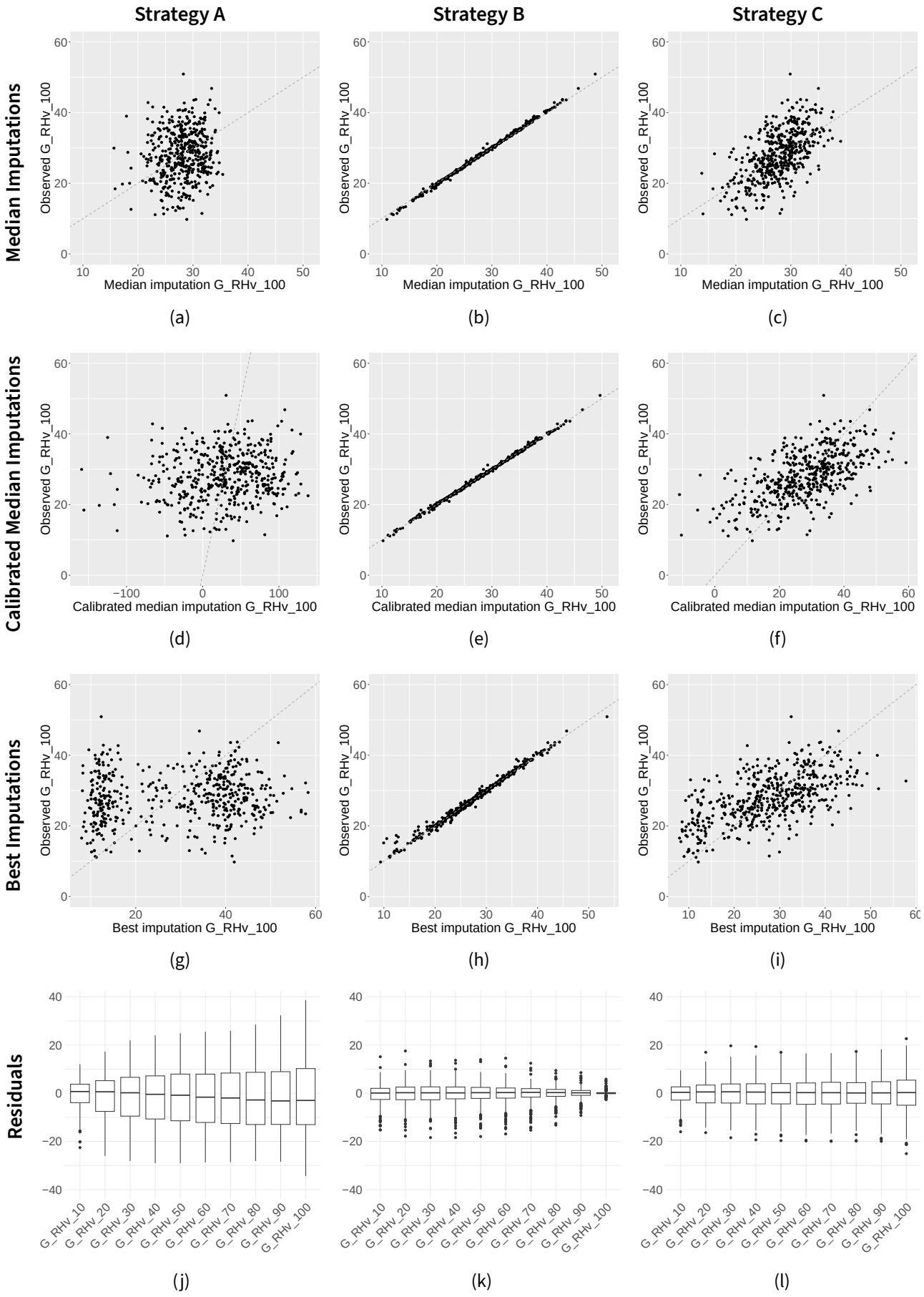


Figure 4.5: G_RHv_100 imputation vs observed data for the 500 test GEDI footprints.

Figure 4.5: (continued) From the 1000 iterations of the bagging process, the median imputation is first selected (first line), then the median predictions are compared to observed values to calibrate the median imputations to account for the regression toward the mean trend (second line). The best imputation is further defined by selecting the real existing GEDI footprint with variables minimizing the distance to the calibrated median imputations of 10 Rhv decile variables (third line). The last line shows the distribution of the residuals between the best imputations and the real Rhv values.

For Strategy A, we observe that median imputations of G_RHv_100 are restricted to heights ranging between 15 and 36 m, whereas the observed G_RHv values range from 9 to 52 m. When a calibration is applied, imputation values are stretched. In Fig. 4.5d, these stretched values even predict negative values, where values should ideally follow a 1:1 trend. The best imputation, then chooses the closest existing GEDI footprint, thereby minimizing the difference to the calibrated median value based on all RHv deciles. For Strategy A, the best imputations for G_RHv_100 have a correlation of only 0.10 to observed G_RHv_100 values. Residual dispersion is vast, increasing from G_RHv_10 to G_RHv_100, as seen in Fig. 4.5j. This increase translates to that in both the mean and range of observed heights from G_RHv_10 to G_RHv_100 combined with a very poor performance of imputations over this range, regardless of the observed value. From G_RHv_40 and onwards, imputations underestimate observed heights.

For Strategy B, G_RHv_100 imputations are highly correlated with the observed values (r = 0.99). Median height imputations are however slightly overestimated for high values and underestimated for low values. The calibration allows for the correction of this slight bias. For this strategy, unlike Strategies A and C, residual distributions tend to narrow from G_RHv_10 to G_RHv_100. This pattern underlines the more optimized predictions for higher deciles compared to lower ones, highly influenced by the use of RHv_100 as an auxiliary variable.

As observed in the third column of Fig. 4.5, imputations with Strategy C are more optimized than with Strategy A. The correlation between best imputations and observed G_RHv_100 values is 0.592, which is similar to the correlation based on median imputations (0.594). Globally, residuals of the imputed values are considerably lower than those for Strategy A. The overall median difference is low (<1 m) for all G_RHvs (Fig. 4.5k). Residual distributions only slightly narrow from G_RHv_100 to G_RHv_10, suggesting an influence of the observed RHv decile distribution on the imputed value distributions.

All GEDI variables were imputed in the GEDI-Sentinel dataset, not only G_RHv_100 and other height deciles. However, we restricted further processing to the ones with a correlation coefficient with observed values that are greater than 0.6 to avoid using wrongly predicted GEDI variables in the subsequent step.

For Strategy A: No GEDI variable reached the correlation threshold of 0.6.

For Strategy B: 18 variables reached the threshold: all the G_RHv deciles except G_RHv_10, G_cover_z3, G_cover_z4, G_cover_z5, G_fhd_normal, G_pai_z4, G_pai_z5, G_pavd_z6, G_cover_z5z6, and G_pai_z5z6.

For Strategy C: G_RHv_90, G_RHv_80, G_RHv_70, and G_RHv_60 reached the threshold.

In summary, Strategy A did not allow for the retention of any GEDI variable for step II, while Strategy B allowed for the retention of 18 and Strategy C the retention of 4 GEDI variables.

4.4.2 Step II: Predicting GSV

GEDI-Sentinel variables for step II

The imputed GEDI variables kept at step I, if any, as well as those of Sentinel (and FORMS-H for Strategy C) were further used to predict GSV. To reduce the number of variables used in the kNN for GSV predictions, and similarly to step I, a variable reduction step based on their correlation with GSV was introduced. The retained variables are displayed in Table 4.3.

Strategy A		Strategy B		Strategy C	
Variable	**Corr**	**Variable**	**Corr**	**Variable**	**Corr**
S2_B6_15_mean	-0.42	G_RHv_60	0.69	FORMS-H_15_mean	0.59
S2_fCover_15_mean	-0.37	G_pai_z5	0.57	S2_B6_15_mean	-0.42
S2_B11_15_mean	-0.34	G_cover_z3	0.54	S2_fCover_15_mean	-0.37
		G_fhd_normal	0.53	S2_B11_15_mean	-0.34
		G_cover_z5z6	0.44		
		S2_B6_15_mean	-0.42		
		S2_fCover_15_mean	-0.37		
		S2_B11_15_mean	-0.34		

Table 4.3: Selected variables for step II. Variables were chosen based on their correlation with GSV.

Common to all strategies, three S2 features computed with a 15 m buffer were selected: two bands (B6 and B11, i.e. vegetation red edge and SWIR, respectively) and fCover. Similar to step I, variables that were autocorrelated with the retained ones were discarded. As no GEDI variables were "correctly" imputed in step I for Strategy A, only S2 data were used in this case.

For Strategy B, 5 among the 18 correctly imputed GEDI variables were chosen. The most correlated variable was G_RHv_60 (r = 0.69), followed closely by other G_RHv variables. However, the latter were discarded based on their correlation with G_RHv_60. Then G_pai_z5, G_cover_z3, G_fhd_normal, and G_cover_z5z6 were also selected.

For Strategy C, the same variables as in Strategy A and the auxiliary height FORMS-H were selected. Notably, FORMS-H height was used because of its correlation (r = 0.59) with GSV. Imputed G_RHv_60 was right behind (r = 0.58), but not selected because of its strong correlation with FORMS-H height.

Evaluation of the kNN regression model used to predict GSV

Results from kNN-bagging using previously selected variables (4.4.2), and applied on the NFI-Sentinel-GEDI dataset, are illustrated for the three strategies in Fig. 4.6, Table 4.4, and Fig. 4.7. The test dataset was also intersected with the volume map FORMS-V. Therefore, we also compared our results to an existing GSV map.

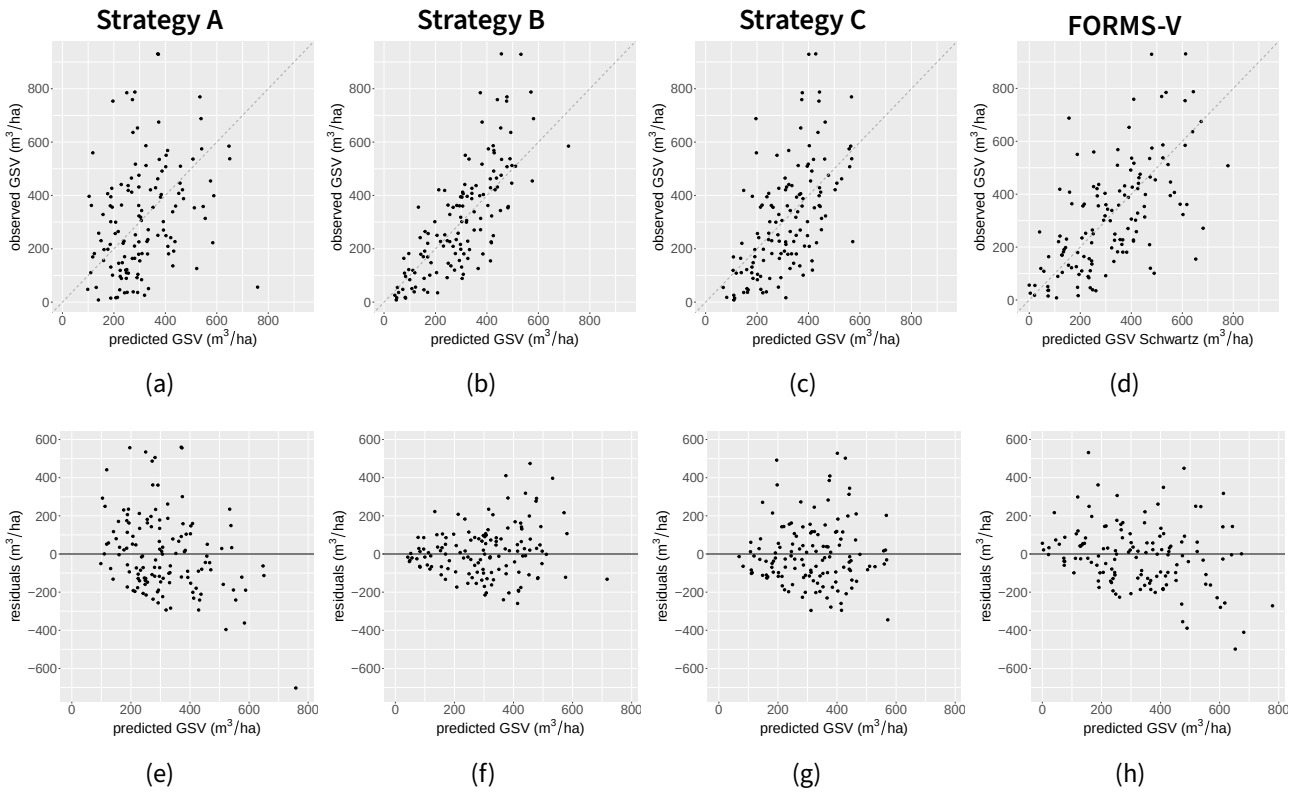


Figure 4.6: GSV predictions diagnosis based on test NFI plots (N = 135 plots) according to the estimation strategy.

Error	Strategy A	Strategy B	Strategy C	FORMS-V
MAE	160.73	101.23	126.51	125.38
MAE%	51.21	32.26	40.31	39.95
RMSE	206.62	133.70	165.49	166.07
RMSE%	65.83	42.60	52.73	52.91
R^2	0.08	0.58	0.34	0.38
Correlation	0.30	0.77	0.59	0.62

Table 4.4: Errors comparing predicted GSV values with observed GSV values on the test dataset. Errors were calculated as observed GSV - predicted GSV.

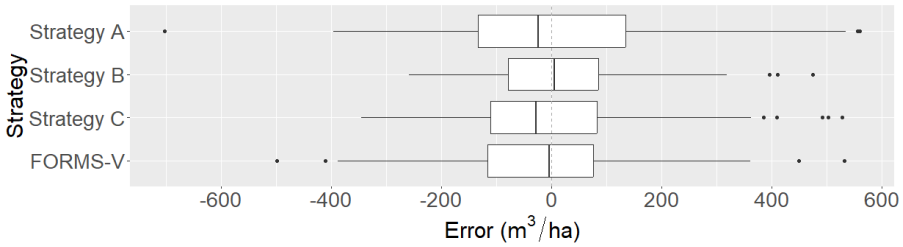


Figure 4.7: Boxplot of errors according to Strategy. The boxes constitute the medians, 1st quartiles (Q1), and 3rd quartiles (Q3) errors and 95% confidence intervals are represented by horizontal lines.

Comparing the observed with the predicted GSV, a gradual improvement is observed from Strategy A to C to B. All strategies have difficulty predicting GSV above $\sim 600\ m^3\ ha^{-1}$. Predictions tend to saturate at high GSV values.

In terms of errors, Strategy C has a small MAE of 101.23 m compared to 160.73 m and 126.51 m for Strategies A and C, respectively. The RMSE improves by 23.23% from Strategy A to B, and by 10.13% from Strategy C to B. R^2 and correlation coefficients also improve from A to C to B. The median value, as observed in Fig. 4.7, is closer to 0 for Strategy B, and the interquartile range also reduces from A to C to B.

When comparing results with estimates from FORMS-V, the MAE for FORMS-V is 125.38 Interestingly, results obtained with our Strategy C yield similar outcomes to estimates from FORMS-V.

All results presented in Sections 4.4.1 and 4.4.2 refer to data used with corrected GEDI positions.

4.4.3 Comparing several setups for the optimal strategy

As Strategy B outperformed the other two strategies in Section 4.4.2, only results using Strategy B are presented for these more specific tests. This Section presents results obtained by implementing Strategy B while using stand-type-specific models, with and without GeoGEDI correction. For both GEDI footprint location versions (GEDI v2 and GEDI v2 corrected by GeoGEDI), we have an overarching model (overall kNN) and individual models for each stand type (Coniferous/Deciduous kNN). For each setup, error metrics are computed using NFI plots of the test dataset (N = 135), both aggregated and separately for each stand type. For the test dataset the mean GSV of coniferous NFI plots was 356.15 m^3 ha^{-1} and the mean GSV of deciduous NFI plots was 234.76 m^3 ha^{-1}.

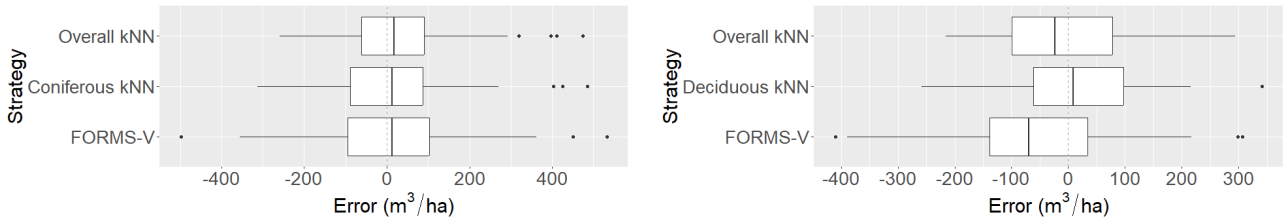
Results of step I and step II without geolocation correction are presented in C.1. For step I, the Sentinel features, extracted at uncorrected GEDI positions, are less correlated with GEDI variables than the ones extracted at corrected GEDI positions, i.e. the greenness indicator correlation coefficient is 0.40 for corrected positions, and 0.37 for uncorrected positions. Consequently, applying the same correlation thresholds compared to the corrected positions, lead to a selection of less Sentinel features for the kNN, i.e. 3 instead of 6.

For step II, with and without GeoGEDI correction, no imputed GEDI variables were retained to predict GSV for Strategies A and C, therefore only Sentinel variables (and FORMS-H) were used. For the test data set these variables were extracted at NFI plot positions, and therefore results do not change between GeoGEDI corrected and uncorrected test NFI datasets. For Strategy B, results on test NFI datasets change, as GEDI variables have been imputed differently. The most correlated variable to GSV for GeoGEDI corrected footprints was G_RHv_60 (r = 0.69) whereas for uncorrected datasets, the highest correlation was for G_RHv_100 (r = 0.68). Five GEDI variables were retained with GeoGEDI correction and four without GeoGEDI correction. The G_fhd_normal had the same correlation without and with GeoGEDI correction (r = 0.54 and 0.53), while the G_cover_z5z6 variable was better predicted with the GeoGEDI correction (r = 0.44 vs 0.37). Final results showed that GeoGEDI corrected datasets performed better to predict GSV than uncorrected data (Table 4.5 and Fig.C.2). RMSE with GeoGEDI was 133.70 m, and without GeoGEDI it was 149.87 m. R^2 was 0.58 and 0.47 and correlation was 0.77 and 0.69, respectively.

Error	With overall kNN			With Conif/Decid kNN			FORMS-V		
	All	Conif	Decid	All	Conif	Decid	All	Conif	Decid
with GeoGEDI correction									
MAE	101.23	106.27	91.82	109.58	119.36	91.30	125.38	127.82	120.82
MAE%	32.26	29.84	39.11	34.92	33.51	38.89	39.95	35.89	51.47
RMSE	133.70	144.18	111.52	142.86	155.25	116.21	166.07	172.44	153.44
RMSE%	42.60	40.48	47.50	45.52	43.59	49.50	52.91	48.42	65.36
$\mathbf{R^2}$	0.58	0.60	0.34	0.52	0.51	0.30	0.38	0.42	0.16
Correlation	0.77	0.78	0.60	0.72	0.72	0.56	0.62	0.65	0.42
without GeoGEDI correction									
MAE	114.52	130.10	85.41	110.67	129.34	75.79	125.38	127.82	120.82
MAE%	36.49	36.53	36.38	35.26	36.31	32.28	39.95	35.89	51.47
RMSE	149.87	168.94	105.31	149.72	170.33	100.44	166.07	172.44	153.44
RMSE%	47.75	47.44	44.86	47.70	47.82	42.78	52.91	48.42	65.36
$\mathbf{R^2}$	0.47	0.44	0.41	0.47	0.41	0.46	0.38	0.42	0.16
Correlation	0.69	0.67	0.65	0.69	0.65	0.69	0.62	0.65	0.42

Table 4.5: Strategy B with and without GeoGEDI correction, for combined (All), coniferous and deciduous test datasets, using overall, and stand-specific models. The "With overall kNN" corresponds to the method described in the methodology. All NFI plots were used to run the kNN, and at the end we split NFI predictions by their dominant stand type. The "With Conif/Decid" kNN columns correspond to two kNNs run distinctly for step II, and aggregating them for "All". The last column is for comparison with estimations from FORMS-V.

With Strategy B of our approach, improving georeferencing had a notable impact on predictions, particularly for coniferous trees. For this stand type, we observed a substantial increase in correlation of 0.11 and 0.07 and a decrease in RMSE% of 6.96 and 4.23, with overall and coniferous-specific kNN, respectively. However, this georeferencing improvement led to a slight degradation in predictions for deciduous trees, with a reduction in correlation of 0.05 to 0.13 and an increase in RMSE% from 2.64 to 6.72, with overall and deciduous-specific kNN, respectively. Surprisingly, using a single model (i.e. overall kNN) amplified the positive impact of georeferencing improvement, enhancing prediction accuracy for coniferous trees while mitigating the degradation of predictions for deciduous trees. Consequently, overall results indicate more significant improvement with a single model compared to a model for each stand type, showing an increase in correlation of 0.08 instead of 0.03 and a decrease in RMSE% of 5.15 instead of 2.18% between non corrected and GeoGEDI corrected datasets for overall kNN and stand-specific kNN, respectively.



(a) Coniferous (b) Deciduous

Figure 4.8: Strategy B without GeoGEDI correction. Boxplot of errors using overall and stand-specific models, compared to FORMS-V estimates. The boxes constitute medians, 1st quartiles (Q1) and 3rd quartiles (Q3) errors and 95% confidence intervals are represented by horizontal lines.

Compared to FORMS-V, our overall-kNN model without GeoGEDI correction exhibited similar perfor-

mance for coniferous trees but demonstrated significantly improved results for deciduous trees (Fig.4.8). Similar trends were observed when comparing our stand-specific kNN model, without improved geolocation, to FORMS-V. Notably, there was an even more substantial improvement for deciduous trees without GeoGEDI correction than seen with the overall kNN model. Consequently, our two models without improved geolocation yielded slightly superior predictions compared to the FORMS-V approach, showing a correlation increase of +0.07 and an RMSE% reduction of -5.2 for both models. In contrast, the overall-KNN model with improved geolocation significantly enhanced predictions, with a correlation increase of +0.15 and an RMSE% reduction of -10.31.

4.5 Discussion

The objective of this study was to assess the potential of GEDI and Sentinel data to estimate the GSV within a MFI framework relying on a model. Given the discrete nature of GEDI data, three matching strategies were tested to relate NFI and GEDI attributes.

4.5.1 Advantages and limitations of Sentinel and GEDI

Results from Strategy A indicate that the sole use of Sentinel data is insufficient for accurately imputing GEDI variables on our study site. We also observed that, whatever the strategy, features coming from Sentinel-1 data were never selected. This result is surprising with respect to various studies. In a similar context of mountainous forest in China, Guo et al. (2023) reported a significant contribution of Sentinel-1 metrics using Recursive Feature Elimination - Support Vector Machine variable selection. While modeling canopy height using IceSat-2, Li et al. (2020) also indicated that backscattering coefficients from Sentinel-1 could positively contribute to the prediction. The results of Kacic et al. (2021) over a tropical forest in Paraguay are more balanced, limiting the interest of Sentinel-1 data to the detection of permanent water bodies and high biomass level. Shendryk (2022) in Australian forests reported that Sentinel-1 performed less than Sentinel-2, but could contribute to improving results when combined. This result was confirmed by Ge et al. (2022), using Sentinel-1 times-series and S2 data to map forest heights. Considering the importance on Sentinel-1 data in the literature cited, the reason for the non-selection in our study may be explained by the selection approach and the overall low correlation of Sentinel-1 data with the attribute of interest, that is more difficult to model than height.

GSV predictions in Strategy B performed better than FORMS-V. A parallel study by Sagar et al. (2022) on a similar study site in the Vosges using S2 and ALS data reported an RMSE% of 41.7%, comparable to our RMSE% of 42.60% with Strategy B. The observed saturation effect of S2 data at approximately $600\,\mathrm{m}^3\,\mathrm{ha}^{-1}$ appeared to be reduced compared to the other tested strategies, but persisted.

Incorporating FORMS-H largely improved imputations (Strategy C) compared to Strategy A, although a regression toward the mean persisted. This was partly mitigated by a calibration process (Lindgren et al., 2021). However, FORMS-H did not efficiently replace GEDI height values and its use led to reduced accuracy in predictions. Notably, FORMS-H has been produced from Sentinel 1 and 2 data using a U-Net deep learning model trained with GEDI RH95 data to map canopy height. Results showed a trend of underestimating height above 25 m (Schwartz et al., 2023). Therefore, the choice of the auxiliary data used to obtain information on

forest structure, thereby complementing Sentinel information, appears to be decisive. In our approach, a height product can be used in addition to Sentinel data as a bridge variable, but this product has to be consistent enough with observed forest height and avoid saturation issues to fully play its role.

The comparison between deciduous and coniferous stands yielded more optimized predictions for coniferous data. Models had more difficulty in predicting GSV for deciduous plots. This trend can also be observed with FORMS-V (Schwartz et al., 2023). Schwartz et al. (2023) used allometric equations based on NFI data to create a GSV map from their height map, having one equation for deciduous and another for coniferous forest stands. Over the French territory and at the sylvoecoregional scale, the authors reported an R^2 of 0.63 with an MAE of $30\ m^3\ ha^{-1}$. However, aggregating results at the sylvoecoregion level masked the high local heterogeneity in plot level prediction accuracy. For our study site, i.e. within an area covering most of the “Central Vosges Massif" sylvoecoregion, FORMS-V showed an R^2 of 0.38 with an MAE of $125.38\ m^3\ ha^{-1}$ compared to NFI test plots. With Strategies B and C, results were significantly improved compared to FORMS-V. Our study site had challenging conditions, mountainous and mixed stand. Other study sites, might have performed better. Previous research by Pereira-Pires et al. (2021) assessed the use of different S2 features to estimate GEDI RH100 and demonstrated strongly varying correlations between S2 and GEDI RH100, depending on the vegetation type.

Interestingly, Strategy C did not select the highest RHv and the overall structure variables, favoring variables reflecting middle layers of the canopy. These layers have more vegetation density than the top layers, and therefore might be better in relation with GSV.

Despite promising results for Strategy B, the method exhibited some instability. Stand-specific kNNs were expected to better fit to their respective datasets, but they did not improve estimates compared to the overall kNN method. When combining results from the two different kNNs, the overall correlation with observed GSV values was 0.72, compared to 0.77 with the overall kNN including both forest stand types. Moreover, predictions are resulting from the aggregation of 1000 kNN runs, and as illustrated in C.2, individually they differ considerably. The mean standard deviation within the 1000 kNN imputations for Strategies A and C is 144.65 m, compared to 126.62 m for Strategy B, as shown in Fig. C.3. Fig. C.4 provides examples of the 1000 predictions of GSVs by NFI plot. It shows a considerable variation in prediction quality and modes. While few were adequately predicted by a numerous kNN runs (C.4a, C.4b, C.4c), some had two modes (C.4i, C.4j), some had very variable predictions resulting in mean estimates (C.4l, C.4k, C.4e), and others were simply inaccurately predicted (C.4f, C.4d, C.4g, C.4h).

4.5.2 Consequences for MFI estimations

Strategy B allows for the creation of a densified GSV point dataset, with GSV estimated at each GEDI footprint. As Strategy B outperformed the other strategies, the results suggest that small area estimation (SAE) approaches, e.g. at the municipality or forest stand level, should be preferred to high resolution maps with pixel-level GSV predictions. SAE creates local predictions using model-assisted approaches, allowing for the correction of bias and improvement of variance. Alternatively, spatialized wall-to-wall maps obtained through model-based MFI approaches present a risk of important local errors, at least requiring accurate quantification of uncertainty at pixel level. Ståhl et al. (2024) also suggests that design-based inference might be a more conceptually appropriate framework, less exposed to biases.

To achieve accurate predictions of GSV, both Sentinel-2 optical data and high-quality height information are necessary. The quality of height data is a critical factor influencing the overall performance of GSV predictions. Improving the height map used (here, FORMS-H) would allow for direct application of the kNN-bagging approach to create a wall-to-wall map. Advanced deep-learning approaches, such as vision transformers (Aleissaee et al., 2023), should be able to learn even better spatial features. Such features have proved useful to mitigating the saturation effect (Lang et al., 2023) to improve wall-to-wall height products. For example, Ge et al. (2022) used an improved semi-supervised deep learning approach to map forest height with S1 and S2 data and the RMSE% of tree heights was estimated to be 24.1% at pixel-level. The integration of multivariate techniques also holds potential for significant improvements.

However, as shown in studies with stands exceeding 20-30 m in height (e.g. Potapov et al. (2021); Lang et al. (2023); Schwartz et al. (2023); Sothe et al. (2022)) and, Gupta and Sharma (2022) revealed that lidar-based information can minimize the saturation issue from optical and SAR C sensors, but cannot solve it entirely. This saturation is indeed inherent to the signal itself. Fully spatializing forest structure features, such as dominant or maximum height, with only high resolution S2 and S1 data even calibrated using dense datasets such as GEDI data continues to be extremely challenging and might not be the optimal solution for obtaining a bridge auxiliary structure variable that is accurate enough.

With the data sources we used and the proposed double kNN-bagging approach, mapping GSV at the pixel level with its associated precision could be counterproductive. Based on standard deviations of plot level GSV predictions (see C.2), 95% confidence intervals could be assessed for each pixel. These confidence intervals would range between 4 and 22 m^3 ha^{-1}, with a mean around 9 m^3 ha^{-1} with Strategy C. However, despite these reasonable precisions, model predictions remain highly inaccurate and local accuracy cannot be assessed. Recently, Ståhl et al. (2024) discussed emerging concerns about model-based wall-to-wall maps and recommended that great caution be exercised in the use of such products. They demonstrated the existence of systematic over- or under-prediction that might call into question the usefulness of the remote-sensing based predictions. Indeed, the use of maps with uncontrolled errors or results from scenario modelling based on such maps may lead to erroneous policy decisions. The risk of negative consequences of the use of remote-sensing based maps is higher when the goodness-of-fit of estimated models is intermediate or poor (Ståhl et al., 2024), which is the case for our GSV models. This is why, even though Strategy C is likely to provide a map with more optimized predictions than those available in FORMS-V, we will not retain this strategy.

Furthermore, we observed that improving the geolocation improved the predictions. Schwartz et al. (2022) also observed notable improvements of their height predictions when changing from GEDI v1 to GEDI v2, which improved footprint geolocation, mainly reducing bias. When integrating GEDI data with auxiliary datasets, improved geolocation is recommended. By ensuring a more accurate geolocation, the integration of GEDI data with other auxiliary sources becomes more robust, contributing to more reliable and coherent predictions.

Moreover, in our kNN-bagging approach, variables representing the middle layers of canopy were chosen over top canopy features. While top of canopy features may capture the maximum height, which is correlated with GSV, the GSV is also influenced by vegetation density present in the middle layers of the canopy. This choice emphasizes the significance of the middle layers in GSV estimation, as they include crucial information related to vegetation density. This highlights an advantage of GEDI data over the utilization of simple

top of canopy height maps derived from photogrammetry.

4.5.3 Possible improvements in terms of modeling and auxiliary data sources

For step I, where the goal is to impute multiple GEDI variables, owing to the large dimension of the data, variable selection is required. The efficiency of the selection is constrained by the large number of predictors in both X and Y dimensions and the relatively low correlation of Sentinel data with the vertical information contained in GEDI. While for this study, simple selection using correlation was found to perform better than more complex methods, additional work would be required to optimize variable selection and associated model performance with such data. Toward that goal, various approaches have been tested in step II, where the goal is to predict a single variable (GSV). These included stepwise regression, partial least squares, or lasso regressions. However, these methods resulted in the selection of numerous variables, which is not adequate for kNN models, as a large number of predictors tends to homogenize distances. Consequently, simple filtering based on correlation was retained.

This approach is more of a first step of variable reduction rather than a final variable selection. It simply relies on correlations. The thresholds of 0.85 and 0.33 are questionable, and changing these thresholds affects final results. Future improvements may involve the use of this correlation method with less restrictive thresholds as a preliminary variable reduction step, followed by a more advanced variable selection.

The correlation-based method may not capture the potential benefits of less correlated variables. For example, Strategy C's retention of FORMS-H could be revisited, with potential exploration of an imputed RHv variable for improved results.

With Strategy C, adding a supplementary common variable of auxiliary height has opened up the possibility to predict GSV at any point of the area of interest. However, the addition of a height map raises concerns about its dependence to Sentinel data and its potential reliability. Moreover, data could be misgeoreferenced, and errors might be transferred to the GSV estimations. Furthermore, FORMS-H shows important errors in mountainous areas (Schwartz et al., 2023), which is the case for our study site. Additionally, Schwartz et al. (2023) attempted to map RH95, considered as a measure of dominant height, introducing a slight departure from Strategy B, where the maximum height is considered. Improving existing height maps could enhance the predictions. However, even with a more optimized height map, results are not expected to be better than those obtained with Strategy B; to improve steps I and II models, other levels, rather than the production of a more optimized height map, should initially be the focus. For example, introducing additional features could present a possible source of improvement. Calculating variables over an extended time span or S1 time series (Ge et al., 2022), could offer insights into temporal dynamics. Extracting textural features from optical images, as described by Couteron et al. (2015), or exploring other GEDI waveform derived indicators may enhance predictions. Furthermore, using a radiative transfer model to simulate GEDI data at the level of NFI plots could provide insight into the link between forest characteristics and GEDI measurements.

Conclusion

In this study, we applied a kNN-bagging approach using Sentinel and GEDI datasets to predict forest GSV. The approach involved imputing GEDI variables using Sentinel auxiliary variables, to subsequently predict GSV. Three strategies were tested to impute GEDI data using different sets of auxiliary data: A) using only Sentinel data, B) using Sentinel and a maximum height variable available on GEDI footprints and on NFI plots, and C) using Sentinel and a national height map. The outcomes emphasized the inadequacy of relying only on Sentinel data for imputing GEDI variables. The use of maximum height variables largely improved estimations. However, the height map used in Strategy C seems to contain uncertainties that tend to reduce its ability to support GEDI imputations. GSV is influenced by more than just height; therefore, the inclusion of various structure-related variables, as available in GEDI data, contribute to improving predictive performance.

Strategy B outperformed all others tested in this study, and appeared to be the most promising strategy. However, it does not allow for pixel-wise mapping of GSV. Even so, the densified GSV points estimated at each GEDI position hold potential for small area estimation, thereby offering insights into local assessments.

In summary, our study highlights the complexities involved in predicting NFI volumes using a kNN-bagging approach with Sentinel and GEDI datasets. The need for careful consideration of auxiliary data, geolocation refinement, and methodological stability is emphasized, thereby providing valuable insights for future research and applications in forest volume estimation.

CHAPTER 5

General discussion and perspectives

Chapter 2 focused on improving GEDI footprints geolocation using a high-resolution DEM. In Chapter 3, a design-based stratification approach to use GEDI footprints for forest inventories estimates was introduced. Chapter 4 presented a kNN-bagging approach using GEDI and Sentinel data to produce augmented forest attribute information which may subsequently be used for model-assisted or model-based estimations. Finally, this last chapter, Chapter 5, aims to provide a general discussion and perspectives. It covers the advantages and limitations of GEDI data and discusses insights into how NFI estimations can be enhanced with GEDI data.

5.1 Advantages and limitations of GEDI data

GEDI, as the first spaceborne lidar system explicitly designed for forest monitoring, offers a range of advantages but is not without limitations.

5.1.1 Advantages of GEDI data

GEDI laser beams penetrate the canopy, allowing assessment of canopy height and vertical distribution of vegetation. In comparison to airborne systems, spaceborne lidar offers repeatability and large-scale coverage capabilities, making it a powerful tool for extensive forest monitoring. GEDI was designed to infer forest attributes at a $1\ km^2$ resolution by the end of its originally planned 24-month mission. In France, for example, GEDI provides the capacity to have national-level up-to-date consolidated information.

For the first time in France, a national-wide high-density aerial lidar survey is currently ongoing. It will provide much higher resolution (i.e. 10 points per m^2 at the ground level) and a wall-to-wall coverage of the country. However, the acquisitions span over 5 years to cover the metropolitan territory and its renewal is not guaranteed yet. Aerial photographs, acquired continuously over a 3-year period to cover the entire French mainland, can be used to create photogrammetric height models. While renewed every 3 years at the department level, these photographs provide the capability to regularly update the country's digital surface model.

Aerial lidar and photogrammetric technologies are useful for estimations at a finer scale (i.e. wall-to-wall maps), while GEDI contributes to improve estimations at a regional scale (i.e. Sylvoecoregions), for an almost-global coverage, with annual renewal capacity. Although national photogrammetric and lidar coverages are highly valuable, they introduce challenges with national estimations due to their departmental scale or tile-by-tile acquisitions. Spaceborne lidar systems like GEDI are not confronted by these problems and, provided that the continuity of such space missions is ensured, should allow to have nationwide fine-scale estimates over relatively short time scales. For example, acquisitions at the department level are not the most appropriate for large scale disturbance assessment (i.e. large storms of 1999 or 2009, large summer wildfires, ongoing bark beetle attacks in the North-East region since 2018). GEDI data should enable easier assessments of these disturbances, by limiting temporal mismatches. A key issue would be to evaluate how the improved temporal resolution of GEDI might compensate the lower spatial resolution compared to aerial and photogrammetric data.

Moreover, GEDI data are free, open and easily accessible to both the scientific community and the forest stackholders, facilitating straightforward data download - a significant advantage over expensive airborne acquisitions (with the exception of national open-source programs, such as the Lidar HD program currently ongoing in France).

Although GEDI's initial acquisition period spanned only over 4 years, it successfully demonstrated that approaches based on spaceborne lidar data are efficient when tackling large-scale problems related to forest environments. Its ability to cover extensive areas with high density and regular renewal, has led to the extension of its mission, emphasizing its significance. As GEDI enters its second acquisition period, it presents an opportunity to study medium-term trends and understand the impact of disturbances on forests, with the promise of future spaceborne lidar satellites providing even more comprehensive capabilities.

5.1.2 Limitations of GEDI data

Georeferencing

One of the primary challenges associated with the use of GEDI data in MFI approaches, as emphasized in this thesis, is its poor georeferencing quality. Positioned on the ISS, GEDI's instrument is prone to movement, and the georeferencing challenges of its footprints are extensively discussed in Chapter 2. Supplied footprint coordinates can easily be 10 m off the real footprint position, which is problematic when intersected with other data. The developed GeoGEDI method's main strength lies in its simplicity, requiring only GEDI L2A footprints coordinates and the 'lowest_mode' variable, along with a high-resolution DEM. GeoGEDI improved consistency in ground elevation and canopy height between GEDI and reference data, especially in sloped areas. However, limitations were observed in flat areas, notably in Landes, where convergence of the flow accumulation algorithm were encountered. While Chapter 2 tested the use of a single-beam and four-beam approach, it showed that beam corrections were grouped by pair-beams, suggesting the use of an in-pair beams approach. Such an approach was therefore used in Chapter 4. To improve georeferencing accuracy in very flat areas, relying solely on ground elevations is inadequate. Alternatively, the GeoGEDI algorithm could be run on a high-resolution canopy height model, or more complicated methods, such as those based on waveform correlation between GEDI and ALS simulations (Hancock et al., 2019), could be used for improved results. The Lidar HD program will enable to implement such approaches over the entire country, thus allowing to evaluate the performance of GEDI in a more optimal context.

GEDI's georeferencing issue arises because GEDI is situated on the ISS. Satellites, being more stable than ISS, encounter fewer problems in this regard. We applied the GeoGEDI algorithm to ICESat-2 data and results analyses revealed that ICESat-2 data are very well georeferenced, with minimal optimal shifts (0 in X and -1 m in Y for our test dataset). This validated the effectiveness of the GeoGEDI methodology and confirmed the accurate georeferencing of ICESat-2 (Soma et al., 2022). Moreover, an approach to improve georeferencing for ISS instruments could involve coupling the lidar with a co-aligned imager. This concept is planned to be used with the Multi-footprint Observation Lidar and Imager (MOLI) mission (Imai et al., 2019). Although GEDI and ICESat-2 have different footprint sizes (100×14 m or 20×14 m segments for ICESat-2 and circular 25 m diameter footprints for GEDI) and are therefore not directly comparable, canopy height estimation performance with ICESat-2 is comparable to GEDI (Urbazaev et al., 2022; Soma et al., 2022). For ground elevations, ICESat-2 is found to outperform GEDI, especially in presence of slope, where GEDI is hindered by its georeferencing issues (Pronk et al., 2023). The precise georeferencing of ICESat-2 enhances its utility in model-based approaches. Exploring the potential of using ICESat-2 either independently or in conjunction with GEDI for MFI approaches presents an interesting perspective.

Concerning the impact of GEDI's georeferencing on MFI approaches, it remains limited on the DSPS approach. However, it hinders the qualification of data and makes it difficult, if not impossible, to quantify the importance of the various factors that influence GEDI canopy height quality (e.g. slope, season, year, forest stand type) and it hampers the verification of the link variable's quality, i.e. the maximum height, used to stratify the data (see Chapter 3).

For approaches involving a model, the models quality depends on the quality of the co-location between reference and lidar data and thus on georeferencing (Milenković et al., 2017; Bouvier et al., 2019). The georeferencing was improved between GEDI v1 and v2, resulting in improved ground elevation and canopy height

estimations at the footprint position (Chapter 2), and improving canopy height maps obtained using models combining GEDI and wall-to-wall remote sensing data such as Sentinel images (Schwartz et al., 2022). Chapter 4 showed that improving the georeferencing also improves links with auxiliary data and therefore the GSV model estimates.

Spatial distribution

Additionally, GEDI's spatial sample pattern cannot be considered as a a probability-based sample, as shown in Chapter 3. Some areas are covered by several footprints, while others lack coverage. This should be improved in the second acquisition phase starting in 2024. Indeed, GEDI was developed to create an optimized spatial distribution at the planned ISS height. However, the orbit of the ISS was unexpectedly raised, resulting in a change in the spatial distribution of GEDI measurements. The higher ISS altitudes led to orbital resonance and reduced coverage (Dubayah et al., 2022a). ISS is now back at the initially planned height, so the spatial distribution of GEDI footprints should be improved for the next acquisition phase. Hopefully no unscheduled changes will occur. The impact of this spatial distribution on models remains unknown, but the new acquisition phase may be an opportunity to evaluate it.

Healey et al. (2012) used ICESat-1 data, where the spatial sample pattern was also identified as neither random nor systematic, posing challenges in the use of the data. They proposed an approach using "equal-area (but not equal-shape) tessellation" and retaining one random footprint in each area to create a dataset of footprints, which can be used as a simple random sample. Running preliminary stratification tests on the GEDI dataset, presented in Schleich et al. (2022), we explored the use of Voronoi tessellation polygons to weight each footprint by its associated Voronoi area. This solution was ultimately set aside as it required recalculating the entire Voronoi tessellation for each case and dataset. While the weighted approach performed better in terms of relative efficiency for small datasets, the equal-weight approach performed better when a large dataset was used.

In my PhD, footprints were considered as independent entities (equal-weight approach). This hypothesis is supported by the irregular shift pattern brought by ISS movements and georeferencing, and the disruption in the sampling introduced by clouds and the applied filters. This hypothesis is on the safe side for variance estimations. However, an alternative approach could consider treating footprints from a same orbit as a cluster. This would imply considering a variance estimator accounting for dependencies arising from cluster sampling. The L4B $1\ km^2$ gridded AGBD estimates used footprint-level L4A data and considered each beam path as a cluster sample (Dubayah et al., 2020a; Ståhl et al., 2011).

Moreover, GEDI does not cover latitudes above 51.6° North or below 51.6° South, excluding significant forest areas like Canada and Sweden from using GEDI data. The ISS orbit restrains observations beyond these longitudes; a dedicated satellite would be required to sample those areas. Meanwhile, extrapolation using a combination of data from various sources (i.e. ALS, ICESat-2, Sentinel, PALSAR) and models have been applied (Sothe et al., 2022; Morin et al., 2022; Potapov et al., 2021).

Waveform processing

GEDI's effectiveness depends on its ability to penetrate the canopy. In regions with dense vegetation or complex canopy structures, ground elevation accuracy may be compromised, potentially leading to an underestimation of canopy height.

While GEDI's full waveform lidar technology yields rich information, translating these waveforms to variables can be intricated, as demonstrated by the erroneous ground peak detection in 2.3.2. Numerous existing and supplementary filters presented in 3.3.1 were applied to discard data with waveform analysis issues. Some of the out-filtered footprints result from acquisition difficulties, while others may stand out due to their environmental context. As a result, especially in mountainous or densely vegetated areas, where waveform processing is more challenging and leads to a higher rate of ground peak miss-detection, filtering alters the spatial distribution of GEDI footprints, with possible impacts on sample representativeness and inference.

For example, we noted that topographical features, such as steep slopes, show spatial aggregation of discarded footprints. Thus, excluding these footprints from estimates, impacts the consideration of certain areas in the overall analysis. In B.3, the comparison of stratum proportions based on height classes, with and without filtering of footprints, suggests a better fit to continuous reference data without filtering. However, the specific effects of discarding footprints on the accuracy of models, remain a subject for further investigations. Rather than discarding them, an improvement in the waveform-to-variable conversion would be more suitable.

In East et al. (2023), the use of spaceborne data simulated from ALS data resulted in better performance than the use of real GEDI data for assessing fire effects on understory structure in tropical forests. Improving both data georeferencing and the translation of waveforms into variables is essential to fully exploit the potential of large-footprint spaceborne lidar data. These improvements are essential to improve models, and ideally, approach the performance levels achieved using simulated data.

Moreover, the GEDI instrument may face limitations in capturing specific horizontal structure components. GEDI measurements integrate information at the footprint level. More detailed characteristics about the 3D distribution of the vegetation within a footprint remain invisible from above, emphasizing the importance of NFIs and the synergy between remote sensing data and on-site field data. Regarding an improved characterization of the 3D structure, it would be worth exploring the synergy between GEDI data and very-high resolution data such as the national coverage by airborne lidar (Lidar HD), currently being acquired in France, or very-high resolution optical images to analyse texture (Couteron et al., 2015).

5.2 Improving NFI estimations with GEDI data

5.2.1 Different methods to link GEDI and NFI data

The spatial concordance between the NFI inventory plots and GEDI measurements not being guaranteed, a primary question arised regarding the ability to establish a link between field surveys (NFI plots with dendrometric measurements) and GEDI signals. Several possibilities exist for this purpose. One can use common variables between GEDI data and variables derived from field surveys at the level of NFI plots, or alterna-

tively, use an indirect link by relying on "gateway" data, such as Sentinel-2 and Sentinel-1 images. I investigated both approaches in my PhD. Another approach is to establish a direct link by using radiative transfer models to simulate GEDI signals at the level of field plots.

Common variables

We established a link based on common variables in Chapter 3, using maximum heights. To evaluate the quality of the link, we used continuous ALS data. The maximum height of GEDI and the maximum height of NFI, are very similar. Despite the general recommendation not to use RH100 to filter outliers, as followed in Chapter 2 using RH98, in Chapter 3 various relative height (RH) metrics were tested. RH100 proved to be the most effective height for our study site, aligning with findings by Zhang et al. (2022). The decision to exclude RH100 and the recommendation to favor lower RH values as maximum height (Duncanson et al., 2021), may be influenced by habits established when working on ALS data. For GEDI, the use of a wide footprint, and full-waveform lidar might make RH100 more robust to outliers. While maximum height from ALS data was found to perfectly match with the maximum tree height of NFI field plots, RH98 and RH95 were found to systematically underestimate the maximum height assessed from ALS data. RH100 was therefore preferred as the counterpart of the maximum tree height at NFI plot level. Also, the calculation of ground elevation and canopy height variables are likely influenced by the pulse width (Potapov et al., 2021). Indeed, GEDI fires near-Gaussian light pulses with a 14 ns full-width at half-maximum, which is equivalent to a 2.1 m width at half-maximum (light celerity equal to 3.10×10^8 m s^{-1}, the distance is divided by two to account for the round-trip travel time of the light). For GEDI footprints on bare ground, the relative heights only refer to the ground return and RH98 (and RH100) overestimate the actual height. This probably explains the systematic bias with an overestimation of heights by around 2 m observed for non-forest footprints in Chapter 2. This same explanation holds for the maximum heights, which may be overestimated due to this vertical resolution. At the same time, GEDI tends to underestimate canopy heights, partly due to the fact that a minimal amount of vegetation have to be intercepted by the laser beam before obtaining a signal exceeding the noise level. RH100 underestimates maximum canopy heights less compared to RH99 or lower RHs. With these two opposite trends, RH100 was found to be a good candidate to assess maximum tree height at plot level.

Other studies have used dominant height from NFI plots and GEDI RH95(Schwartz et al., 2022). NFI's dominant height variable represents the mean height estimation of the 100 tallest trees within a 1-hectare area, calculated based on tree diameters measured in the plot. Dominant height is widely used in forest studies to calibrate or validate models (Morin et al., 2022; Schwartz et al., 2023; Chen et al., 2023). We opted not to use the dominant height, as this variable is not translated in the GEDI variables and extracting dominant height from a waveform in a heterogeneous forest stand poses considerable challenges. Dominant height is typically aiming at primarily characterizing regular stands. In regular stands, it can be assimilated to canopy top height and be related to upper RH values. However, its interpretation becomes more complex in uneven-aged irregular stands (Pardé, 1965). Depending to the level of heterogeneity, the optimal RH associated to the dominant height is likely to change. Therefore, we opted to use the maximum tree height as a more suitable alternative.

Moreover, we also considered using canopy cover as a link variable, as this variable is given in NFI plots and in GEDI L2B product footprints. However the canopy cover variable of the French NFI is not a measurement, but a percentage of visible sky, estimated in the field in 10t% increments based on expert judgement

(IGN, 2023b). No significant correlation could be found between NFI and GEDI canopy cover, leading us to exclude it. We also calculated canopy cover from ALS data, but could not find a correlation between GEDI and ALS canopy cover.

We did not identify any other shared variables between GEDI and NFI datasets. Further investigation is needed to identify potential variables that could establish a meaningful link. Although GEDI variables like PAI and PAVD hold promise, they are not estimated in NFI plots. One approach could involve estimating these variables for NFI plots using NFI variables (and maybe auxiliary data such as ALS), or linking them to an existing highly correlated NFI variable and consider these GEDI and NFI variables as common.

The idea is to understand how the information contained in the signal reflects structural characteristics similar to those obtained from ground measurements. Auxiliary data such as ALS or Sentinel could be used to create an additional common variable (see the next "indirect link" Section), or ground surveys at the GEDI footprint level, but it would require a large number of surveys to study a direct link (see the "direct link" Section).

Indirect link

In Chapter 4, we established an indirect link using Sentinel-2 data.

Sole reliance of Sentinel-2 spectral bands proved inadequate for accurate prediction of GEDI variables. Nonetheless, certain variables derived from Sentinel-2 bands, including vegetation indices (such as greenness indice, NDVI, NDWI, MSAVI) and vegetation biophysical variables (fCOVER, LAI, fAPAR), exhibited correlations of up to 0.4 with GEDI variables. However, these correlations are very site-dependent, as shown by Pereira-Pires et al. (2021).

Strategy A in Chapter 4, which relied solely on Sentinel-2 data, did not yield accurate imputation of the set of GEDI variables. Improved results were achieved by incorporating an additional variable of maximum height (RH100 for GEDI and maximum height for NFI, or an existing canopy height map). Using the maximum height to constrain the model enabled to better predict other GEDI variables. The resulting Sentinel-GEDI dataset was then used to predict NFI GSV, revealing highest correlations to GSV with GEDI variables such as relative vegetation height (RHv) deciles 20 to 100, canopy cover, plant area index (PAI), and foliage height diversity index (fhd_normal). With Strategy B (using RH100 for GEDI and maximum height for NFI), all imputed GEDI variables exhibited a minimum correlation of 0.44 with GSV. Some Sentinel-2 variables also showed correlations with GSV, including B6 band (-0.42), B8A, B8, B7, B11, fCover and fAPAR with negative correlations below -0.3.

Improving the current approach by using additional data and time series could enhance prediction accuracy. Sentinel-1 data was not retained in our models, because it was not sufficiently correlated with GEDI variables, but it has shown valuable to predict canopy heights in other studies (Ge et al., 2022; Guo et al., 2023; Shendryk, 2022; Li et al., 2020). Ge et al. (2022) demonstrated that times series of Sentinel-1 data performed similarly to single-date Sentinel-2 data in predicting canopy height in a boreal area, reporting that the combination of both provided the best results. Morin et al. (2022) reported that texture indices from Sentinel-1 and Sentinel-2 data contributed to improve the RMSE of predictions for various forest attributes.

In Chapter 4, we used a kNN-bagging approach. While several parameters of this approach could be

modified, other approaches may also be used to create an indirect link. Kacic et al. (2023) used a random forest regression with Sentinel-1, Sentinel-2 and GEDI data to create several GEDI-derived attribute maps (i.e. canopy height, canopy cover, PAI, and foliage height diversity index). Ge et al. (2022) compared three machine learning models (Multiple Linear Regression, Random Forest, and Light Gradient Boosting Machine) to deep learning approaches to model canopy height. In the case of machine learning methods, variable selection is necessary and was found challenging in Chapter 4. Some machine learning methods have integrated variable selection (e.g. Random Forest and kNN), but in our case, the integrated kNN variable selection (Crookston and Finley, 2008), did not allow to properly select variables. Too many features were kept. For methods without integrated variable selection, Ge et al. (2022) used principal component analysis to select a maximum of 10 features. They also reported that deep learning approaches perform significantly better than machine learning approaches to model canopy height. On the one hand, machine learning is effective in understanding the relationships between variables and aiding in modeling. On the other hand, deep learning has gained attention in remote sensing due to its automatic feature extraction, high-level semantic segmentation, and effective modeling and mapping in complex environments (Kaselimi et al., 2023; Zhu et al., 2017).

However, deep learning approaches lack transparency and a first key aspect with such approaches will be to understand how the models process information, so that we can improve our knowledge on the relationship between forest attributes and remote-sensing based auxiliary information. Transformed architectures show promise in better understanding of the models operations, offering a potential integration of the advantages of both machine learning and deep learning (Kaselimi et al., 2023).

Additionally, while deep learning has proven effective in mapping forest height by using dense height data from GEDI, the assessment of more intricate forest parameters, such as GSV, faces challenges due to the limited availability of dense reference datasets. While GSV maps can be derived from height maps using allometric equations (e.g. Schwartz et al. (2023)), relying only on height to predict volume is sub-optimal for complex stands. Furthermore, few studies attempted to predict more than one variable (Kacic et al., 2023), and methods predicting multiple RH values, essential for a more comprehensive characterization of forest structure, are still to be developed. Regarding the potential to predict several variables, kNN remains a powerful method.

The kNN-bagging model performed best when using maximum heights additionally to Sentinel data, but is therefore limited to predicting GSV at GEDI plot positions. This approach offers the opportunity to densify GSV points, transitioning from a sparse 675 NFI plots to $100{,}000$ points in the case of our study site. There are multiple possibilities to the use of this data, discussed in 5.2.2.

We implemented the kNN-bagging approach in two steps, as we were exploring the possibility of spatializing the GEDI variables. Alternative strategies, such as directly imputing GSV and other attributes from NFI plots to GEDI footprints, could also be assessed.

Direct link

To establish a direct link between NFI plots and GEDI footprints, we require NFI and GEDI data at the same location. In the context of the SLIM project, our goal was to use the simulation of GEDI signals through radiative transfer modeling (using discrete anisotropic radiative transfer - DART) on NFI plots, using Terrestrial

Laser Scanner (TLS) data to produce forest scenes.

For this purpose, we established a ground dataset of $\sim$ 100 plots covering $\sim$ 100 GEDI footprints. Data collection occurred in winter and summer, spanning across Landes, Vosges, and Sologne forests. When field teams were available, I provided them with corrected positions for a sample of GEDI plots to visit, ensuring close alignment (temporally or seasonally) with the planned field dates. For these $\sim$ 100 SLIM plots, TLS data along with other NFI field data were collected.

Within the SLIM project, colleagues developed a method involving the creation of forest scenes from TLS data acquired at SLIM plots, including the generation of a Digital Terrain Model, separation of wood and foliage, reconstruction of trunks and large branches, and voxelization. They also worked on automating the production of these scenes and the GEDI signal simulations with DART.

The ongoing work has shown promising results, and the next step involves using the $\sim$ 100 GEDI-TLS footprints to calibrate a model bridging GEDI and NFI plots. Once established, this model could be applied to other TLS acquisitions, enabling the simulation of GEDI footprints over large datasets. Between 2010 and 2015, the French NFI collected TLS data alongside typical field data collections for 1338 plots. This extensive dataset, covering a national scale, forms a valuable resource for the calibration and application of the model. However, it would provide estimates of forest parameters only at the GEDI-footprint-level. The subsequent task involves to determine how to integrate these data within a MFI framework, discussed in the following section.

With the extent of the mission and the densification of GEDI footprints, one might also expect that the number of GEDI footprints overlapping NFI plots will increase in the future. If a sufficient number of NFI plots are covered at the end of the mission, direct relationships between NFI and GEDI could be investigated with a statistical significance.

5.2.2 Integration of GEDI data in MFI approaches

The GEDI mission is the first spaceborne lidar mission dedicated to forests. The layout of the mission incorporates a robust statistical approach aimed at providing, for a $1\ \mathrm{km}^2$ grid, estimates of forest attribute means and associated variance (Dubayah et al., 2022a). Attributes in question, provided by GEDI L3 and L4 gridded products, are canopy height, canopy cover, leaf area index (LAI), and above ground biomass density (AGBD). Model-based approaches have been proposed to produce estimates in cells sparsely or not sampled by GEDI data (Saarela et al., 2018). However, the change in the ISS orbit prevented the originally planned sampling. Nevertheless, data analysis demonstrated the potential of GEDI measurements for forest characterization and monitoring at a global scale, advocating to continue efforts for the development of forest-dedicated lidar solutions.

At the beginning of my PhD, these gridded products were not yet available. It is also worth noting that the attributes provided do not correspond to the typical outputs of NFIs, which inform forest policies and typically focus on timber resources and their evolution at national and regional scale. To adapt forest inventory methods to meet local management needs within the context of global change, the use of MFI approaches is widely advocated (Tomppo et al., 2008; Saborowski et al., 2010; Westfall et al., 2019).

The integration of GEDI data into MFI approaches can be tackled through various methods, serving dif-

ferent purposes, and operating at different scales. Based on the results of my PhD, several strategies are suggested to illustrate the offered possibilities of integrating GEDI data into MFI approaches across various spatial scales. For some of these strategies the results from the developed kNN-bagging approach can be used, but many other modeling approaches (e.g. direct link based on DART and other kinds of regression models mentioned in Section 5.2.1) could have been used to establish this GEDI-NFI link.

Design-based solutions, like the DSPS approach, can reduce the variance of estimates, thus enabling to provide inventory results with either an improved confidence level or, for a targeted confidence level, to provide results at a finer scale than the one achievable with NFI data alone. The implemented DSPS design-based approach, despite GEDI's measurement pattern not strictly adhering to a probability-based sample scheme, effectively reduces the variance of GSV estimates. This DSPS approach is appropriate for regional or sub-regional scales, such as departments (on an administrative level) or sylvoecoregions (closer to the reality of forest stands). Further testing is necessary to explore how far downscaling is possible and determine the minimum area size achievable through DSPS.

The DSPS stratification approach, presented in Chapter 3 and evaluated at the level of a complex sylvoecoregion, the Vosges, could easily be extended to a national scale. Relying only on GEDI and NFI data, including the BD Foret national product, it offers a significant advantage. It does not require any other auxiliary data and is not reliant on well-georeferenced GEDI data, facilitating a quick execution of the approach. This could prove advantageous for generating periodic updates, i.e. annual or seasonal, of forest attribute estimates.

A potential framework involves downloading all GEDI footprints in a period of interest, adding a stratification based on structural variables, compared to the current NFI method using only forest and non-forest status for stratification. While DSPS estimations belong to the realm of pure estimation and are highly valuable and reliable. However, focusing on smaller areas and more local predictions, for example at the scale of municipalities or forest patches, or even fully spatializing estimates to provide high-resolution maps, would require the development of model-assisted or model-based approaches.

The kNN-bagging approach allowed to create a link between GEDI and NFI data and resulted in a densified dataset of GSV predicted at GEDI footprints, which can be used in various ways.

First, the kNN-bagging outputs could be incorporated into a DSPS approach, stratifying GEDI footprints and NFI plots based on the GSV, either independently or in addition to height. With a stratification based on the variable of interest, an increase in relative efficiency is expected, thus enabling to reduce the variance of the estimates or to focus on smaller areas (Haakana et al., 2019). However, this combined solution may be influenced by georeferencing errors and modeling biases introduced through the kNN-bagging approach. Besic and Vega (2023) proposed an ensemble modeling approach combining partitioning, classification and regression to match GEDI profiles with NFI plots. Partitioning relied on GEDI profile characteristics, and both classification and regression relied on auxiliary data including height, Sentinel and topographic variables. The resulting classification can then also be used as post-stratification criterion.

The kNN-bagging output, could also be used for small area estimations (SAE) assisted by a model, or even for model-based SAE, with the risk of bias. This would allow to spatialize predictions at smaller scales. SAE, like the DSPS approach, lacks visual appeal, but relies on robust and proven statistical frameworks. Techniques such as the kNN-bagging, Random Forest or Deep Learning's Vision Transformer may be used in this context. SAE proves valuable for generating local predictions using model-assisted methodologies

(Zhang et al., 2022; Breidenbach and Astrup, 2012; Mauro et al., 2017). The SAE approach allows for accurate estimation of forest parameters for small subgroups of domains, enabling precise estimations in specific, smaller areas. This method allows estimations at a finer spatial resolution, providing insights useful for local-scale decision-making. SAE accounts for heterogeneity across the subgroups and offers the flexibility of using various modeling approaches, including both model-based and design-based methods.

When the variables used to link GEDI and NFI data are derived from wall-to-wall high-resolution remote sensing products, like optical or radar images, it offers the possibility to spatialize estimates at the level of high-resolution grid cells resulting in wall-to-wall forest attribute mapping. Given the continuous nature of both Sentinel images and canopy height maps available at the national level Strategy C of the kNN-bagging approach could easily be propagated to predict GSV at every pixel. However, achieving accurate GSV maps at a 30 m pixel scale, for example, requires improvements in both height maps and models.

Various deep learning approaches have combined GEDI with optical and/or radar data (Lang et al., 2023; Potapov et al., 2021; Schwartz et al., 2023) to create wall-to-wall maps. While these maps provide information at the NFI ground inventory points for developing predictive models of attributes like volume or biomass, they often lack associated uncertainty maps to address local biases. In the case of poorly specified models, their use for management purposes could lead to poor choices or decisions. As the title of McRoberts (2011) article "Satellite image-based maps: Scientific inference or pretty pictures ?" attests, this problem has been known for a long time and the community of researchers working in the field of forest inventory is sensitive to it. However, it is not the case for all the scientific and user communities who may take for granted maps that are strongly biased.

Furthermore, current methods for spatializing GEDI data involve a significant loss in vertical profile information. In all above-mentioned studies, only one of the upper GEDI RH values, considered as a measurement of the dominant or maximum height, has been propagated. Moreover, while the performance of deep learning predictions better addresses certain challenges than traditional machine learning approaches, it does not overcome the limitations of many optical and radar data, notably signal saturation in mid-biomass levels (Shendryk, 2022; Ge et al., 2022). While high-resolution maps are of great scientific interest, their usefulness for forest management and public decision-making is undoubtedly less so. Forest managers prefer to use stand-level data, while public decisions are mainly based on administrative boundaries (e.g. departments, municipalities). At these scales, stratified estimates and model-assisted estimates of small geographical domains offer better guarantees against bias and should be preferred (Ståhl et al., 2024; Zhang et al., 2022; Breidenbach and Astrup, 2012). They are also in line with the GEDI mission objectives in terms of statistical inference. As mentioned by Ståhl et al. (2024) adopting a design-based perspective, such as the DSPS stratification approach presented in Chapter 3, can lead to more realistic expectations.

Creating unbiased wall-to-wall maps is challenging with current available information. Overcoming the barrier linked to the saturation of Sentinel signals with forest age, and thus with both height and volumes, remains challenging and will probably require considering using new data sets, including very high resolution data.

Improving predictive models and the underlying methods remains an important area of research. Better-specified models provide better control over biases and offer greater flexibility in estimation procedures. Better models would contribute to increased trust and confidence when using model-based approaches, which enable estimation in areas devoid of field plots. Such approaches could also be useful for estimating

disturbed domains. Several studies have used GEDI for monitoring disturbances like fires and insect infestations (East et al., 2023; Boucher et al., 2020; Sanchez-Lopez et al., 2020). In this context, the hybrid approach integrating stratification and kNN-bagging could also be evaluated. Integrating other data alongside GEDI and Sentinel, such as ICESat-2, high-density airborne lidar data and high resolution images should also enable us to improve the estimation and mapping of forest attributes at national, regional and sub-regional levels.

Conclusion

Methods involving data fusion require a strong spatial agreement between data sources, and spatial mismatch between different data sources may lead to biased estimations and increased variances. While the development of a method improving the georeferencing of GEDI data was not initially planned to be part of the thesis, it became a significant aspect of my PhD work to facilitate the subsequent integration of GEDI data into MFI approaches. Only requiring a DEM, with sufficient computational resources, the developed method can be applied nationwide. The kNN-bagging approach in Chapter 4 showed the positive impact of improved georeferencing on MFI estimates, emphasizing the need to address georeferencing inaccuracies for a robust link between NFI and GEDI data. Mitigating the impact of inaccuracies is essential for the reliability of MFI results and contributes to a more in-depth understanding of GEDI data characteristics.

In this thesis, we reveal the challenges of establishing a reliable link between NFI plots and GEDI signals due to the lack of spatial correspondence. Two approaches to integrate GEDI data into MFI were tested. Chapter 3 provides a solution with a design-based stratification approach and proved that GEDI data can be linked to NFI data through maximum canopy height. However, caution is recommended as GEDI's sample scheme could not be perfectly characterized as probability-based. This was partly due to a change in the ISS altitude that deeply impacted the sampling design of GEDI. This phenomenon is amplified by the spatial arrangement of footprints with respect to their quality. As a result, while some areas were densely sampled, some areas (e.g. slopes) were totally excluded, impacting the sampling scheme and estimates.

Chapter 4 explored different strategies to predict GSV for GEDI footprints using a kNN-bagging approach. Strategy B, incorporating both Sentinel and a maximum height variable, outperformed other strategies, suggesting the use of SAE for improved forest inventory results. The thesis underlines the complexities involved in integrating GEDI data into MFI and recommends cautious consideration of model-based wall-to-wall maps.

The French NFI relies on an annual sampling design that is consolidated using a 5-year moving window to achieve the required level of precision for resource estimations at national down to regional levels. MFI approaches integrating GEDI data, alone or along with a set of well-chosen remote sensing data, hold the promise to better address the need for more frequent updates and higher spatial resolution in forest assessment and monitoring. This need is underpinned by the necessity to quantify the effects of climate change on forest dynamics and mitigate their negative impacts. A key issue will be to evaluate if the achieved precision gain will meet the scales at which appropriate decision-making processes and management activities operate.

For its next acquisition phase, starting in the fall of 2024, GEDI is expected to provide data with both

a nominal spatial distribution and an improved geolocation, thus mitigating some of the limitations faced during my PhD work. Using the developed DSPS approach, estimates at the sylvoecoregion scale will be improved. If efforts are being made to develop long-term forest-dedicated lidar missions, such MFI approaches could become operational. Regarding approaches relying on models, deep learning approaches hold promise to further improve model performances and downscale estimations within a robust and reproducible framework. In this regard, the kNN-bagging approach developed in this thesis could be used to augment NFI data to the GEDI population, allowing for the prediction of attributes other than height.

Appendix A

Improving GEDI footprint geolocation using a high resolution digital elevation model

	Landes						Vosges					
Group	**% of version**	**ME**	**sdE**	**MAE**	**sdAE**	**RMSE**	**% of version**	**ME**	**sdE**	**MAE**	**sdAE**	**RMSE**
v1	100	-0.41	1.44	0.81	1.26	1.50	100	-1.19	6.34	3.90	5.14	6.45
v2 algo 01	87.4	-0.48	0.78	0.62	0.68	0.92	65.0	-0.69	2.44	1.42	2.11	2.54
v2 algo 02	12.6	0.80	3.11	2.06	2.47	3.22	35.0	0.75	5.10	3.54	3.74	5.15
v2 algo 02 pb	3.40	4.42	3.50	4.97	2.68	5.64	13.7	2.07	5.96	4.87	4.01	6.31
v2 algo 02 valid	9.20	-0.52	1.48	0.99	1.22	1.57	21.3	-0.10	4.24	2.68	3.29	4.24

Table A.1: GEDI ground elevation errors for five footprint groups, by study site

The groups are: (v1) v1 footprints, (v2 algo 01) v2 footprints using ground peak algorithm 01, (v2 algo 02) v2 footprints using ground peak algorithm 02, (v2 algo 02 pb) v2 footprints using ground peak algorithm 02 where ground elevation difference between v1 and v2 is greater than 1.5 m, (v2 algo 02 valid) v2 footprints using ground peak algorithm 02 where ground elevation difference between v1 and v2 is lower or equal to 1.5m. Group v2 algo 02 pb refers to footprints for which a bias was identified using algorithm 02. Group v2 algo 02 valid is the complement to the biased v2 algo 02 pb. For each group the percentage of concerned footprints is noted and GEDI ground elevation is compared to MNTref. Mean Error (ME), standard deviation of error (sdE), Mean Absolute Error (MAE), standard deviation of absolute error (sdAE) and Root Mean Square Error (RMSE) of ground elevation are shown.

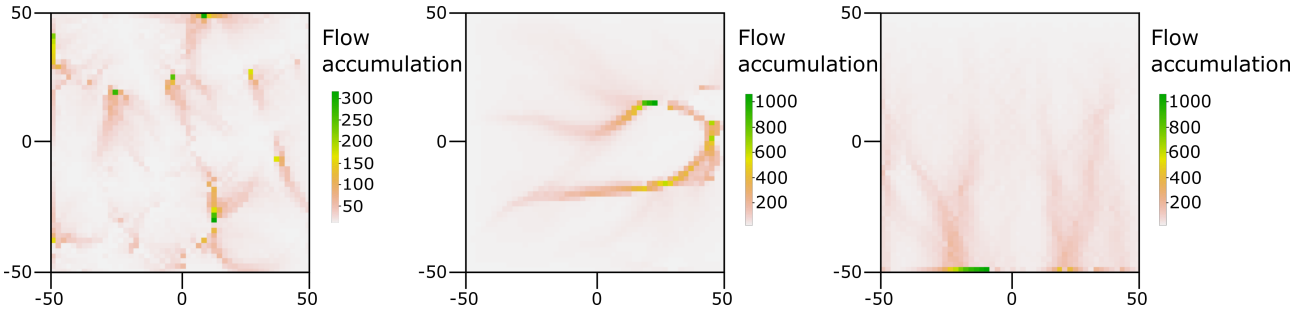


Figure A.1: Flow accumulation error maps with low maximum flow accumulation values.

Available dataset and script

The dataset processed for this article is available online (Schleich et al., 2023a). GEDI footprints with coordinates of unchanged GEDI v1 and v2 releases and the coordinates calculated with GeoGEDI algorithm, as well as all variables used for this article, are included. GeoGEDI R script is available on Github (Schleich et al., 2023b).

Appendix B

Potential and limits of GEDI footprints for forest inventories estimations based on a double sampling for stratification approach

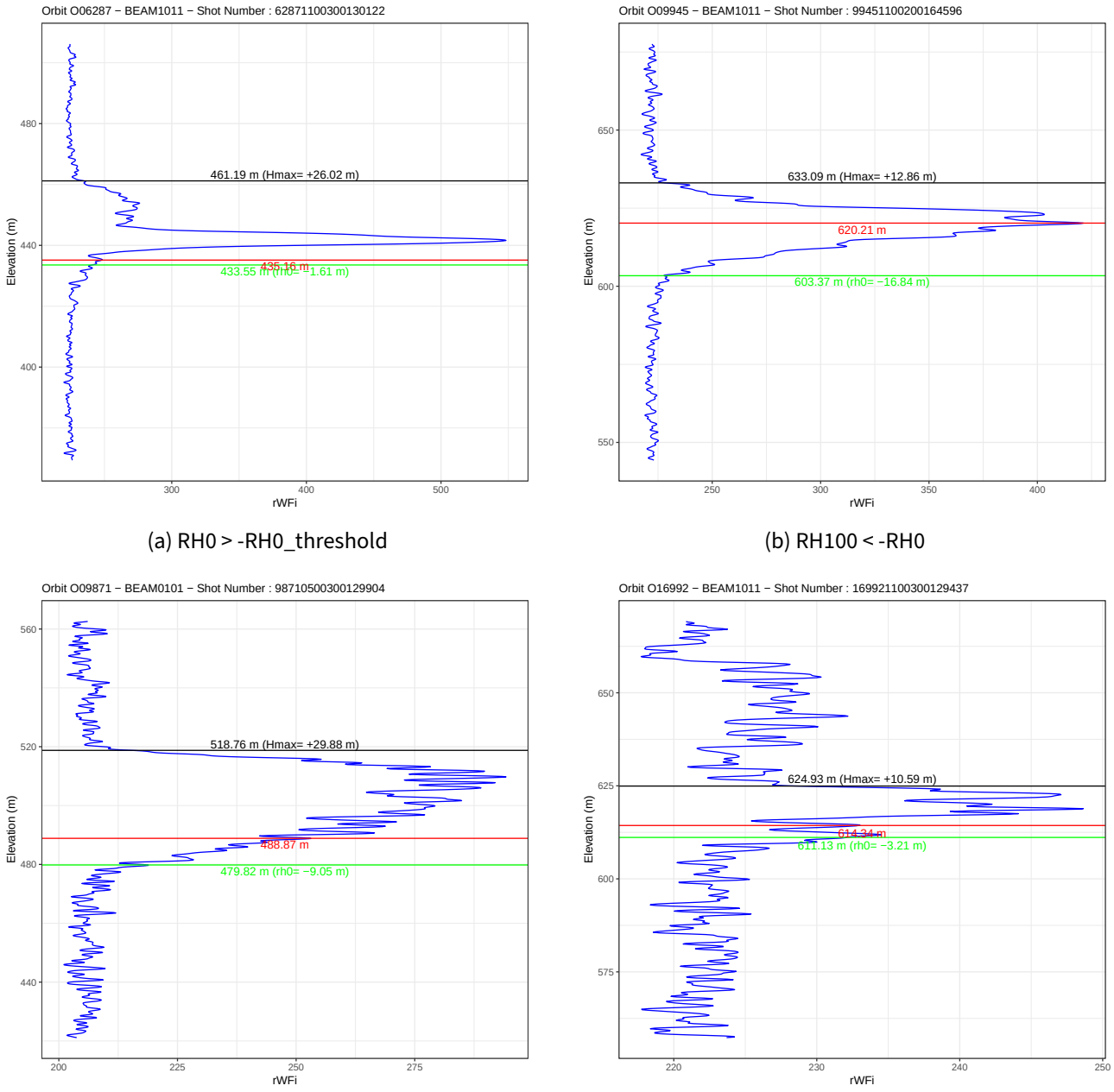
B.1 Additional waveform filters

The filters were defined to sort out the footprints which follow :

- RH0 > -RH0_threshold. The Rh0_threshold was computed as half the maximum width of the ground peak (i.e., the width at the bottom of the peak). The full width at half maximum (FWHM) of the emitted pulse was approximately 15 ns (Dubayah et al., 2020a). The standard deviation (σ) of the Gaussian pulse was linked to the FWHM as $\sigma = FWHM/2.355$ (Roncat et al., 2014). The value of two σ was chosen to fix the RH0_threshold, i.e., 12.74 ns ≈ 1.91 m. RH0 provides the position of the end of the waveform relative to the ground peak. The latter is assumed to have a Gaussian shape similar to or larger than that of the emitted pulse (Jutzi and Stilla, 2006). If RH0 is smaller than the -RH0_threshold, then the ground peak or the end of the waveform is expected to be inadequately detected, leading to biased RH values.
- RH100 < -RH0. -RH0 corresponds to the "mirrored end" of the detected ground peak. If RH100 is smaller than -RH0; then, RH100 was contained in the ground peak. Either the latter was poorly detected, or there was no vegetation.
- selected_mode starting with "0". The selected_mode variable indicates the index of the lowest non-noise mode found in the waveform. If it starts with zero, the algorithms could only clearly identify a single mode, whereas typical vegetation waveforms usually exhibit at least two modes, which can overlap: one for the ground component and at least one for the vegetation component. Thus, it is possible to obtain vegetation waveforms with a single mode in some situations, for example low vegetation or vegetation on very steep slopes. Waveforms with a single mode were discarded because they revealed the absence of vegetation or issues with mode detection.
- rx_maxamp <= 100 or energy_total <= $10{,}000$ If the maximum amplitude of the waveform relative to the mean noise level (rx_maxamp) is below 100, or if the integrated counts in the return waveform relative to the mean noise level (energy_total) are below $10{,}000$, then the footprint is discarded. The

waveform is likely to be highly noisy, with an important risk of misdetection at both the front and back ends of the signal and ground peak.

Examples of outsorted footprints are illustrated in B.1.



(a) RH0 > -RH0_threshold

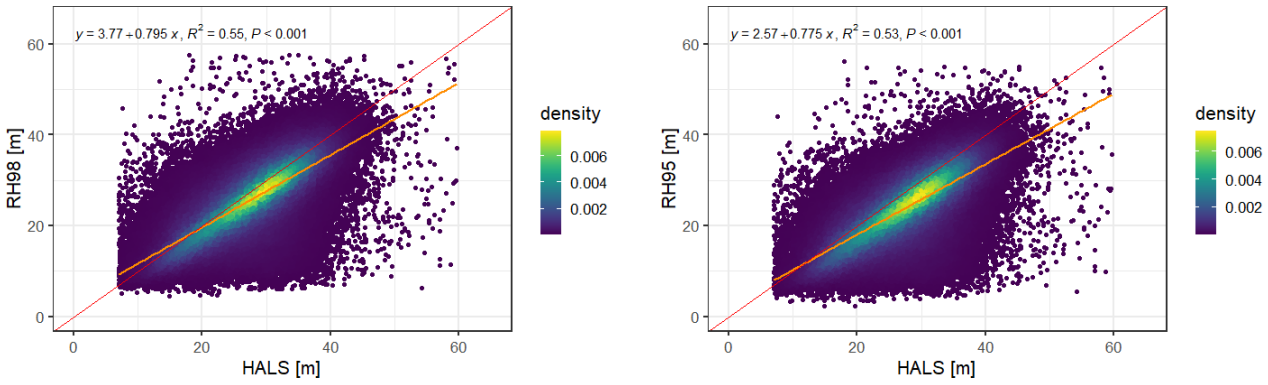
(b) RH100 < -RH0

(c) *selected_mode* stating with "0"

(d) energy filter

Figure B.1: Examples of outsorted footprints based on filters presented in 3.3.1. RH0 is presented in green, RH100 in black and the ground elevation of the *lowest_mode* variable in red.

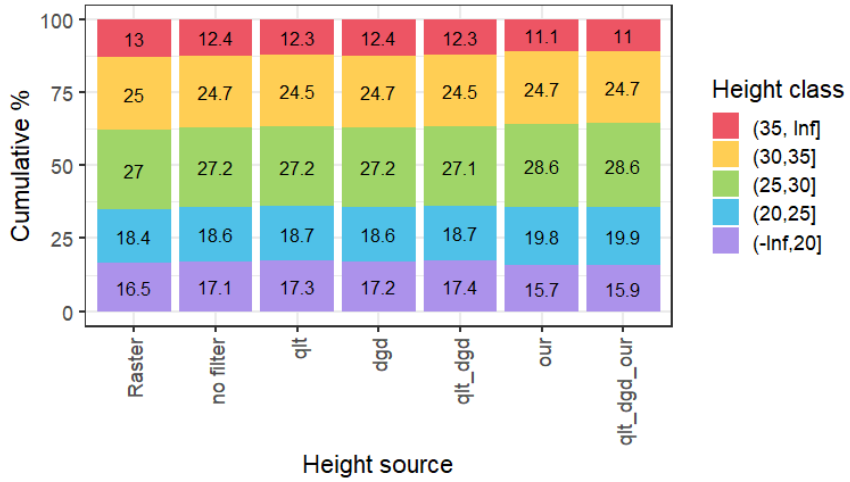
B.2 Additional GEDI - ALS height scatter plots



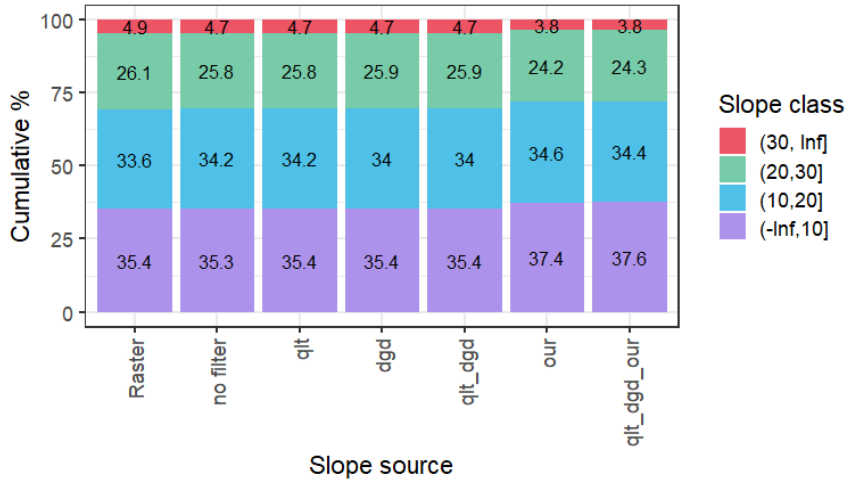
(a) (b)

Figure B.2: Scatter plot of H_{ALS} and different GEDI RH

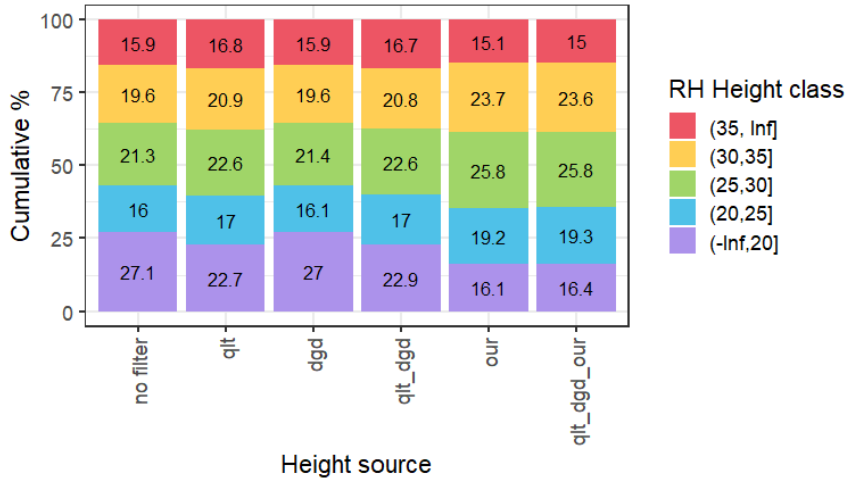
B.3 Surface proportions impacted by quality filters



(a) Proportions of ALS heights at GEDI locations compared to reference raster proportions



(b) Proportions of slope classes [degrees] at GEDI locations compared to reference raster proportions



(c) Proportions of RH100 heights

Figure B.3: Surface proportions of different subsets of GEDI footprints. *Raster* presents the reference raster proportions (in BD Forêt) and *no filter* corresponds to all full-beam GEDI footprints. Other data subsets correspond to the *no filter* dataset with applied filters. *qlt* was filtered by the GEDI quality_flag, *dgd* was filtered by the GEDI degrade_flag, *qlt_dgd* was filtered by the GEDI degrade and quality flags, *our* was filtered by our additional filters presented in 3.3.1 and *qlt_dgd_our* was filtered by GEDI degrade and quality flags and our additional filters.

B.4 Impact of other variables on the GEDI - ALS height relation

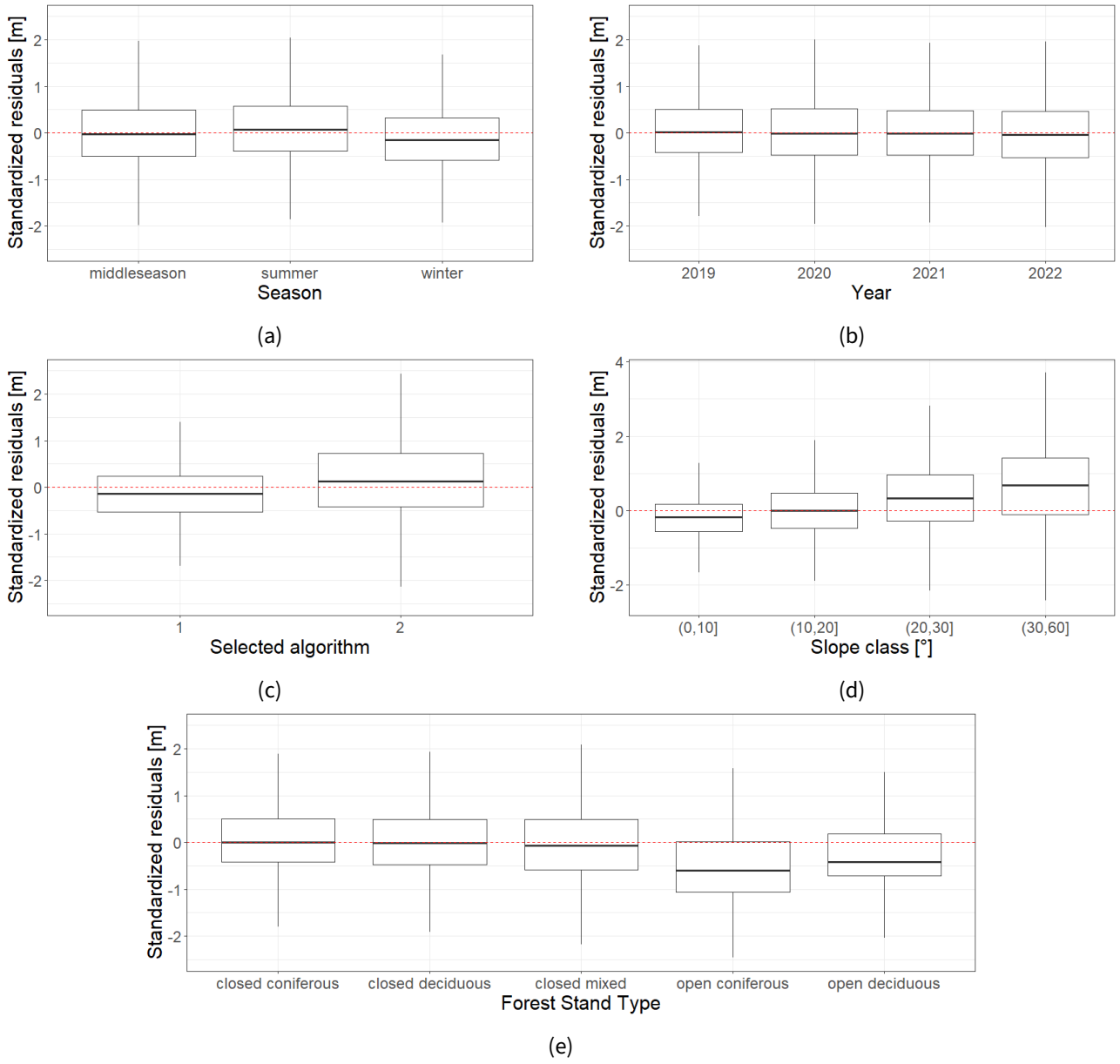
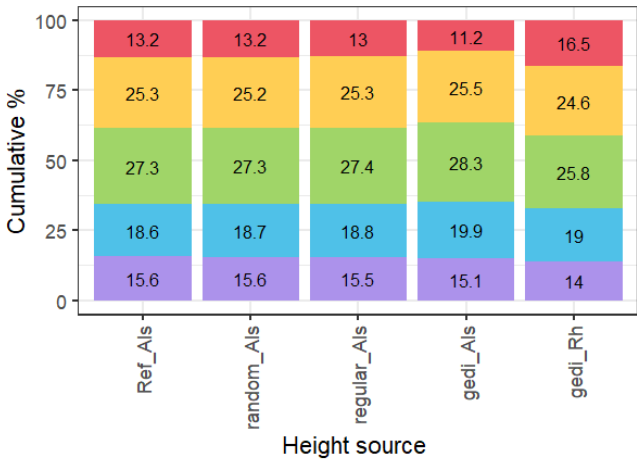
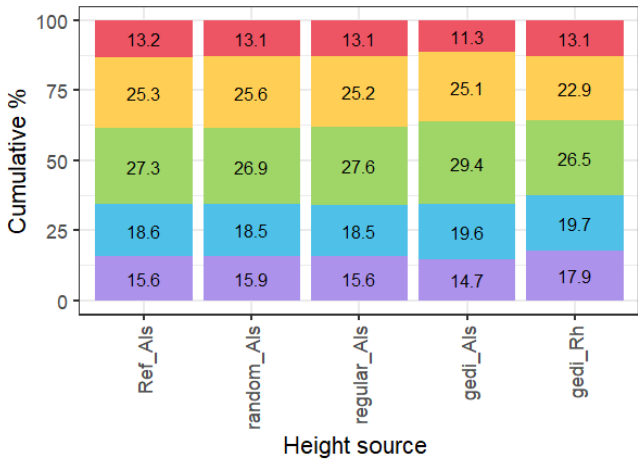


Figure B.4: Boxplots of standardized residuals of GEDI RH100 and ALS heights linear regression by factor classes

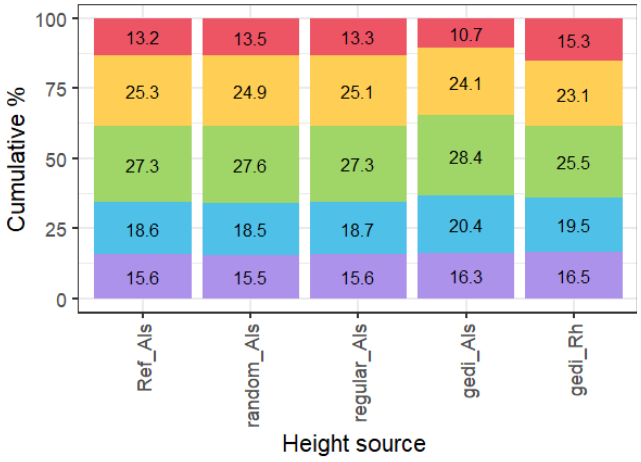
B.5 Surface proportions by season and year



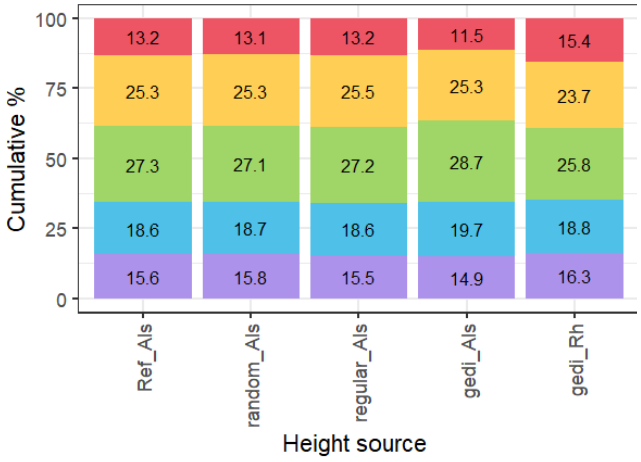
(a) Summer N=80,784



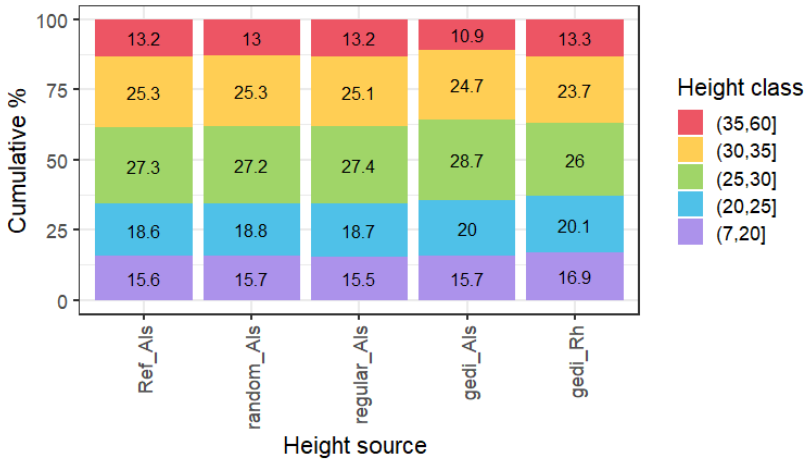
(b) Winter N=47,754



(c) 2020 N=86,305



(d) 2021 N=43,869



(e) 2022 N=47,241

Figure B.5: Surface proportions of different point layouts. Ref_Als are the reference proportions based on the raster. The proportion of each class is marked in %.

Appendix C

kNN - Bagging NFI, GEDI, Sentinel-2 and Sentinel-1 data to produce estimates of forest volumes

C.1 Results using datasets without GeoGEDI correction

C.1.1 Step I

Auxiliary variable	Most correlated GEDI variable	Corr
Hmax for Strategy B	RHv_100	1.00
FORMS-H_15_mean for Strategy C	RHv_90	0.70
S2_fAPAR_15_mean	G_pavd_z3	0.37
S2_greenness_15_mean	G_cover_zf	0.37
S2_ndvi_50_mean	G_cover_z2	0.33

Table C.1: Variables selected for step I. The first column contains the selected variables, the second column contains the GEDI variable to which the auxiliary variable is the most correlated, and the third column indicates the correlation between the two. Variables start with "S1" or "S2" if they come from Sentinel-1 or Sentinel-2 data, followed by the variable name, followed by 15 or 50 depending if the 15 m or 50 m radius was used for the extraction of the variable. Variables end with "_mean" or "_sd" depending on the zonal extraction of mean or standard deviation. Auxiliary height for Strategy C is FORMS-H_15_mean. Auxiliary height for Strategy B is RHv_100 from GEDI

.

Other S2 variables show correlations above 0.33 with one or more GEDI variables (i.e. lai_mean, fCover_mean, ndvi_mean, msavi_mean) but are not selected because they are correlated by more than 0.85 to a previously chosen variable. Fig. C.1 shows results on the 500 GEDI footprint test dataset, comparing imputed to observed values for G_RHv_100.

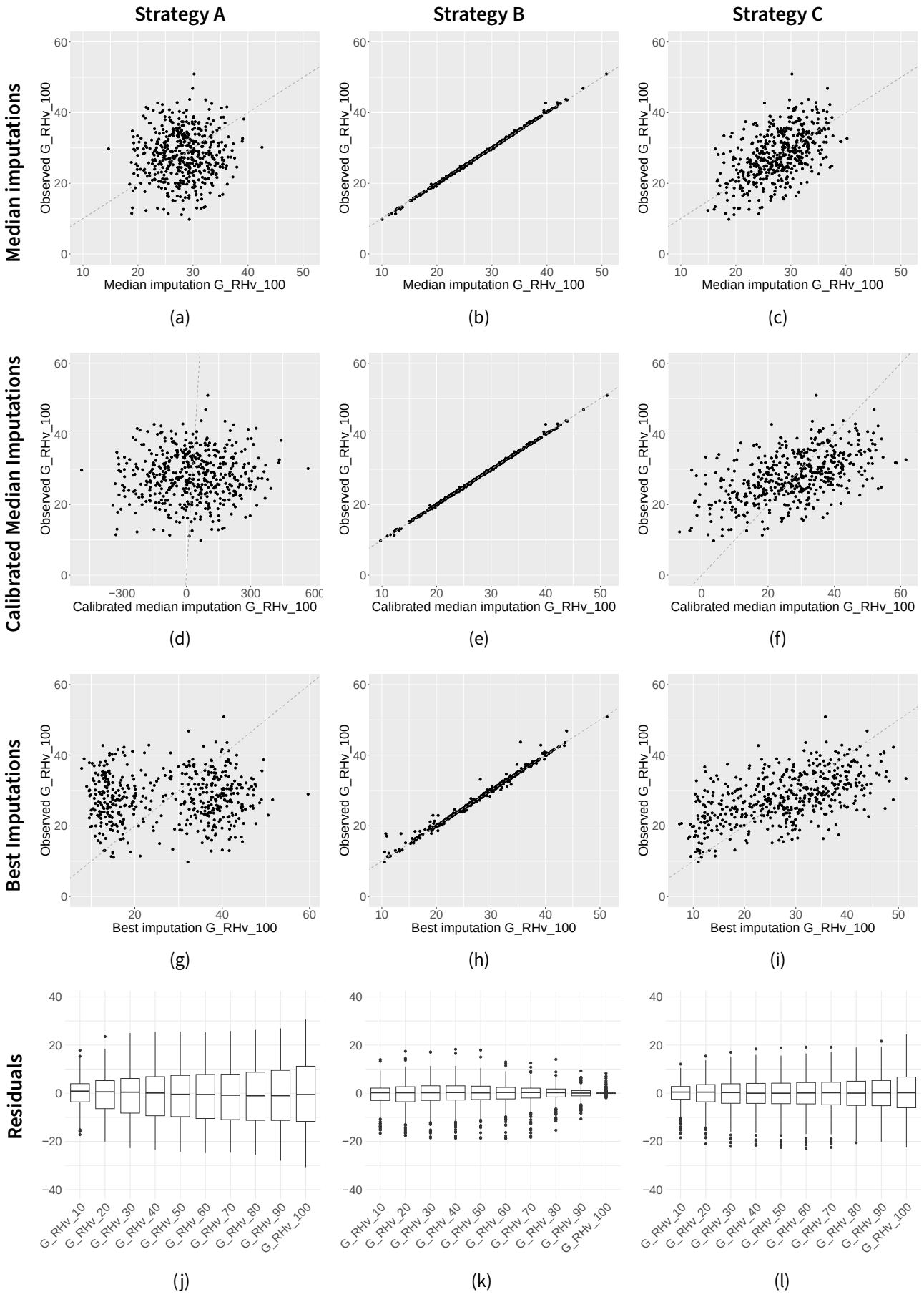


Figure C.1: G_RHv_100 imputations for 500 test GEDI footprints - Without GeoGEDI correction

Subsequently, for further processing only the GEDI variables which were considered well imputed (correlation between observed and best imputations on the test dataset > 0.6), were kept.

For Strategy A : /

For Strategy B : G_RHv_100, G_RHv_90, G_RHv_80, G_RHv_70, G_RHv_60, G_RHv_50, G_RHv_40, G_cover_z4, G_cover_z5, G_fhd_normal, G_pai_z5, G_pavd_z6, and G_cover_z5z6.

For Strategy C : /

C.1.2 Step II

Strategy A		Strategy B		Strategy C	
Variable	**Corr**	**Variable**	**Corr**	**Variable**	**Corr**
S2_B6_15_mean	-0.42	G_RHv_100	0.68	FORMS-H_15_mean	0.59
S2_fCover_15_mean	-0.37	G_fhd_normal	0.54	S2_B6_15_mean	-0.42
S2_B11_15_mean	-0.34	G_cover_z5	0.53	S2_fCover_15_mean	-0.37
		S2_B6_15_mean	-0.42	S2_B11_15_mean	-0.34
		S2_fCover_15_mean	-0.37		
		G_cover_z5z6	0.37		
		S2_B11_15_mean	-0.34		

Table C.2: Selected variables for step II - Without GeoGEDI correction

This results in graphics in Fig. C.2.

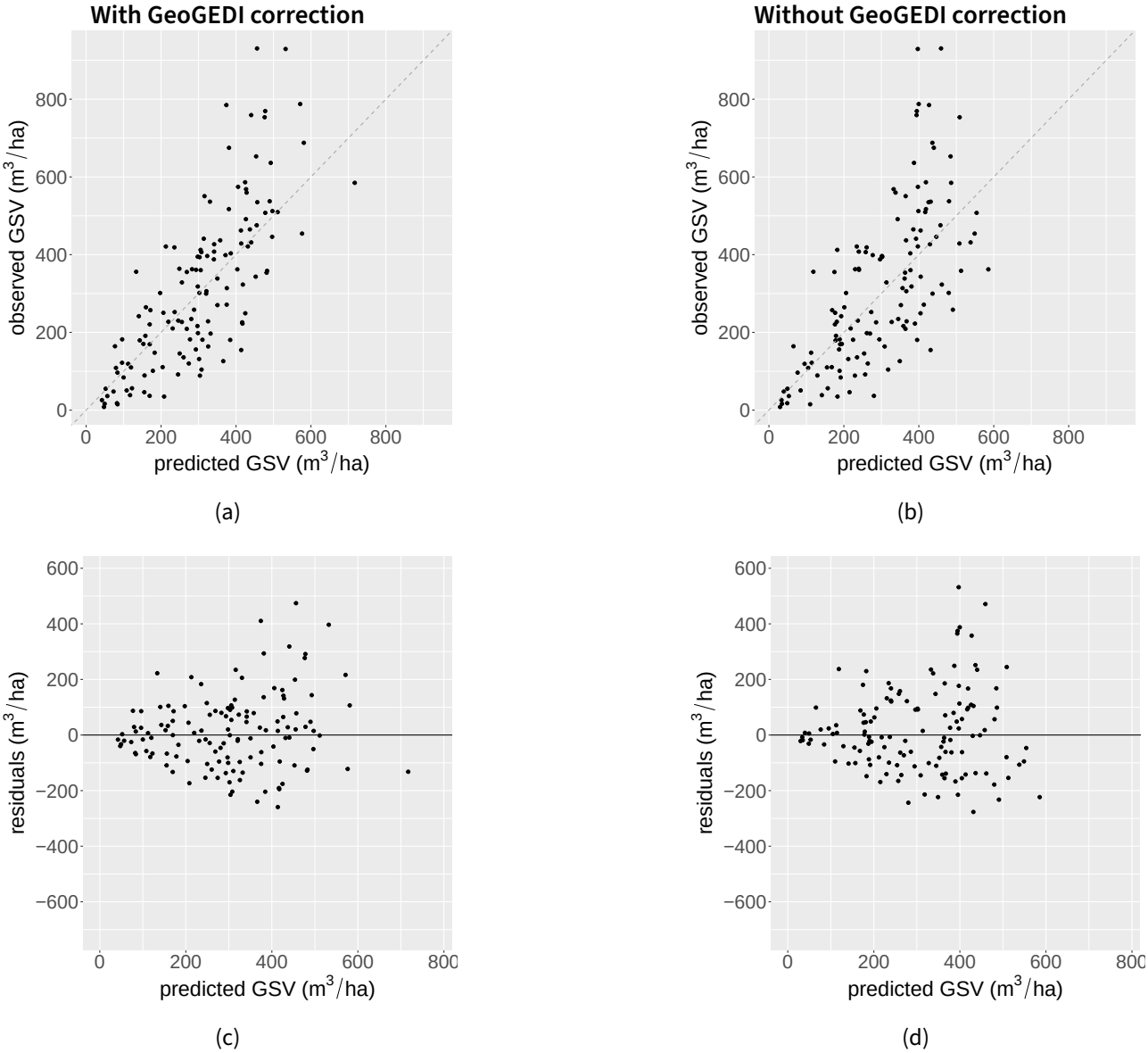
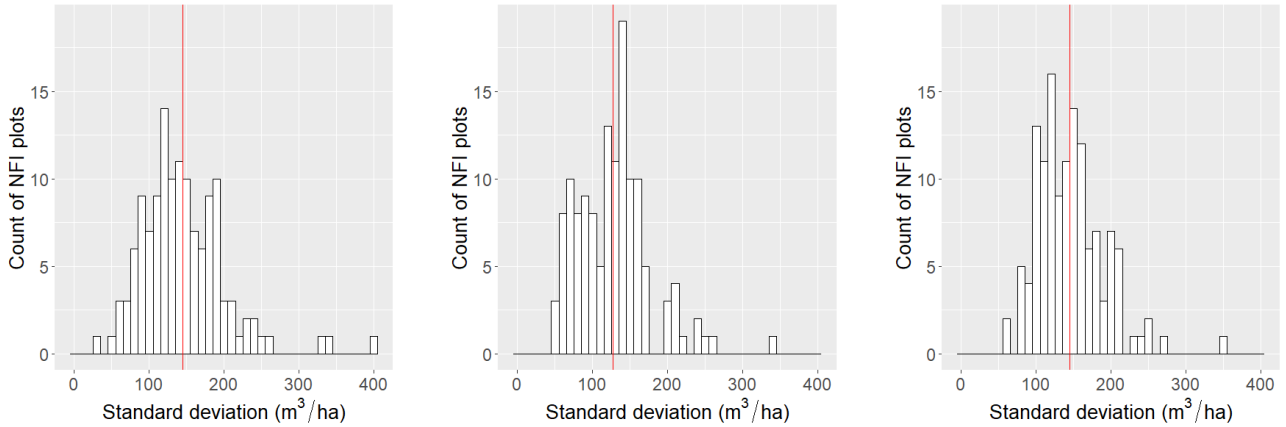


Figure C.2: GSV predictions for test NFI plots with Strategy B.

For errors of Strategy B using uncorrected GEDI positions, refer to Table 4.5.

C.2 Step II: the 1000 predictions from the 1000 kNN runs

This Section shows results of the 1000 kNN runs, before aggregating the single predictions into mean values. Fig. C.3 shows the distribution of standard deviations within the 1000 predictions for all NFI plots.



(a) Strategy A (b) Strategy B (c) Strategy C

Figure C.3: Distribution of standard deviations within the 1000 kNN predictions. The red line presents the mean.

Fig. C.4 shows examples of the 1000 kNN predicted GSVs for some NFI plots.

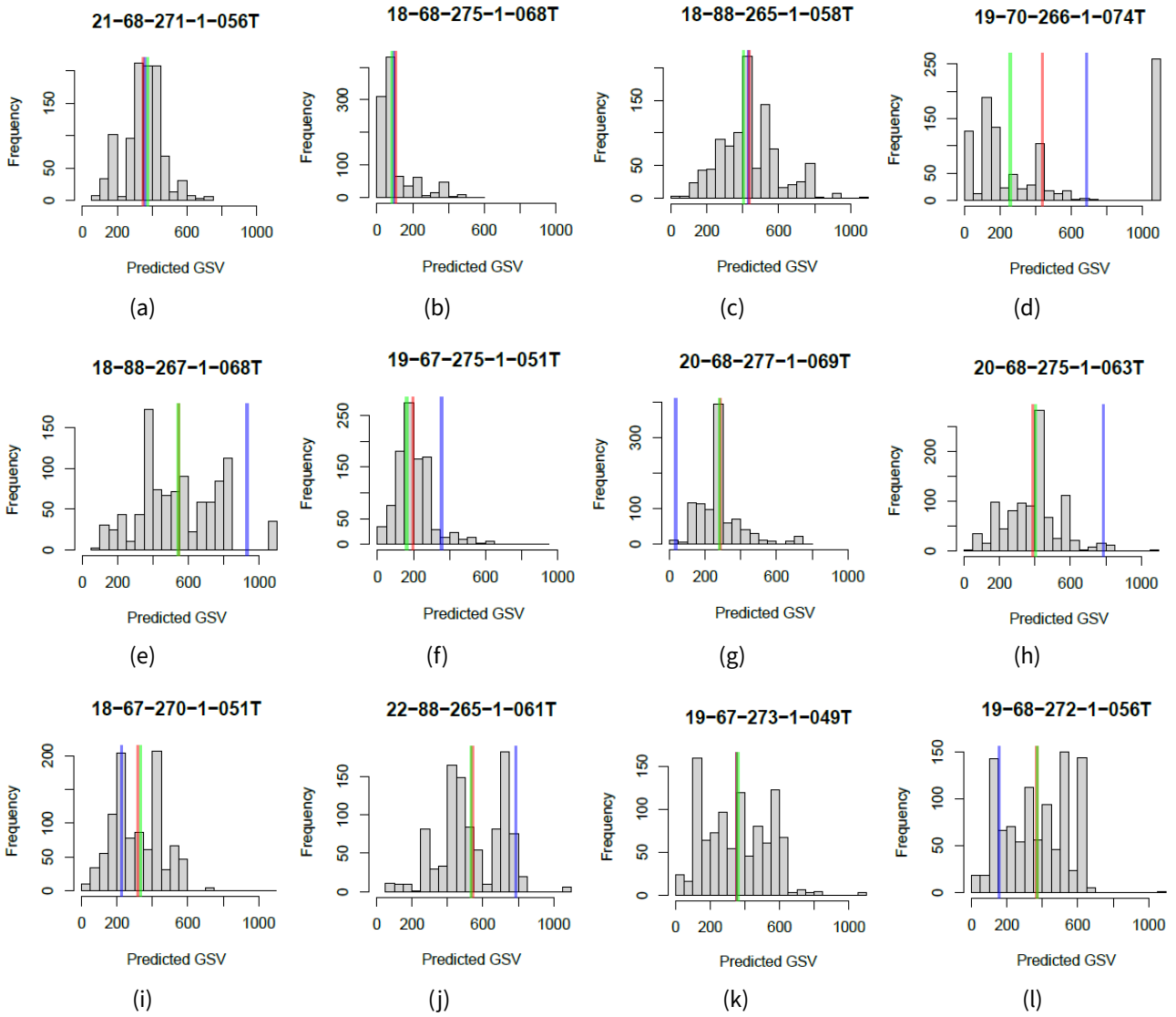


Figure C.4: A thousand predictions by NFI plot. The blue line represents the observed GSV value, the red line the mean of the 1000 predicted GSVs used in our methodology, and the green line represents the median of the 1000 predicted GSVs.

While few were adequately predicted and by numerous kNN runs (C.4a, C.4b, C.4c), some exhibtited two modes (C.4i, C.4j), some yielded very variable predictions resulting in mean estimates (C.4l, C.4k, C.4e), and others were simply erroneously predicted (C.4f, C.4d, C.4g, C.4h). Figs. C.4k and C.4l have very similar-looking prediction histograms, resulting in similar aggregated final predictions. However, their observed GSV values (in blue) are quite different.

Résumé long en français

Apport du lidar spatial pour le développement de méthodes d'inventaire forestier multisource adaptées à la gestion durable des forêts dans un contexte de changement global

L'objectif central de cette thèse est d'évaluer le potentiel des données acquises par le système lidar spatial GEDI pour améliorer l'inventaire de la ressource forestière en France.

Chapitre 1 : Contexte et objectifs

Contexte

Les forêts jouent un rôle essentiel en fournissant des services écologiques, économiques et sociétaux ainsi qu'en contribuant de manière significative au stockage du carbone et au maintien de la biodiversité au niveau mondial (IUFRO, 2018; Bonan, 2008). Le changement global exerce des pressions sur les forêts, nécessitant une adaptation face aux perturbations d'origine anthropique et climatique, telles que la déforestation, la fragmentation des forêts, les infestations d'insectes, la sécheresse et les feux de forêt (IPCC, 2023; Prăvălie, 2018). Ces dernières années les forêts ont connu une augmentation significative de ces perturbations, avec l'enregistrement de records historiques quant à l'ampleur de certaines perturbations, suscitant des inquiétudes concernant les projections futures (IUFRO, 2018). Face à ces perturbations, la gestion durable des forêts est essentielle. Elle implique une adaptation aux perturbations, la promotion de l'industrie du bois, l'évaluation et la gestion des pressions anthropiques ainsi que le développement d'indicateurs précis pour une meilleure compréhension de l'état et de la dynamique des forêts. Une compréhension approfondie du fonctionnement des écosystèmes forestiers est essentielle pour évaluer et atténuer l'impact croissant de ces perturbations. Cela nécessite une surveillance continue des forêts de l'échelle locale à mondiale (European Commission, 2021).

L'Inventaire Forestier National (IFN) joue un rôle crucial dans l'évaluation et la gestion des ressources forestières (Tomppo et al., 2010; Vidal et al., 2016b). En France, le plan de sondage de l'IFN est dimensionné pour répondre à des besoins de politique publique de l'échelle nationale à régionale. Sur des plus petits territoires, où s'exercent les activités de gestion, la précision est souvent insuffisante au regard des besoins. Les méthodes d'Inventaire Forestier Multisource (IFM) permettent d'effectuer une descente d'échelle en préservant la précision. Elles reposent sur la combinaison statistique des données d'inventaire et de données auxiliaires, souvent des données de télédétection, partiellement corrélées aux attributs forestiers. Ces données auxiliaires permettent de densifier les données IFN et d'améliorer la précision des estimations à effort de sondage constant (Westfall et al., 2019).

Jusqu'à présent, les IFM faisaient plutôt appel à des images optiques ou radar (Fernandez-Ordonez et al., 2009), qui permettent de couvrir les échelles nationales à la fois à haute résolution spatiale (< 50 m) et temporelle (i.e. synthèses mensuelles). Cependant, en 2018 deux missions exploratoires de lidar spatiaux ont été lancées : IceSat-2 et GEDI. Ces systèmes permettent de mesurer la structure 3D de la végétation en couvrant de vastes territoires à une haute densité spatiale et avec une bonne répétitivité temporelle (Dubayah et al., 2020a; Neuenschwander and Magruder, 2019). Ces mesures lidar sont particulièrement prometteuses car fortement corrélées aux attributs forestiers tels que la hauteur, le volume et la surface terrière (Lim et al., 2003; Beland et al., 2019).

Le lidar spatial Global Ecosystem Dynamics Investigation (GEDI), installé à bord de la Station Spatiale Internationale (ISS) fin 2018, a été spécifiquement conçu pour étudier les écosystèmes forestiers et a acquis des données d'avril 2019 à mars 2023. Depuis, l'instrument a été mis en pause et une nouvelle phase d'acquisition devrait débuter en automne 2024 (LP DAAC, 2023). GEDI émet de brèves impulsions laser à une longueur d'onde de $1{,}064$ nm vers la surface de la Terre pour mesurer la structure verticale de la végétation. Lorsque le faisceau laser atteint le sol, il couvre une zone de ~ 25 m de diamètre appelée empreinte. Le signal rétrodiffusé par les différentes cibles interceptées au sein d'une empreinte (e.g., végétation, bâti, sol), appelé forme d'onde, est enregistré à haute fréquence temporelle. Les formes d'ondes sont géoréférencées et analysées pour fournir divers produits contenant des informations sur le terrain et sur la structure de la végétation. L'instrument est équipé de trois lasers : deux émettant à pleine puissance (faisceaux "power") et le troisième étant divisé en deux faisceaux de demi-énergie (faisceaux "coverage"). Par conséquent, à tout moment, quatre faisceaux, chacun ayant un diamètre d'empreinte de ~ 25 m, sont incidents sur le sol. Chaque laser émet 242 fois par seconde et chaque faisceau est dévié tous les deux tirs. Cette configuration permet d'obtenir huit traces parallèles au sol, espacées de 600 m, avec une empreinte tous les 60 m le long de la trace. Il est à noter que l'échantillonnage GEDI a été pensé dans une logique inférentielle, afin que les estimations d'attributs forestiers issus de ces mesures puissent être assorties d'une estimation de variance (Dubayah et al., 2020a).

Objectifs de la thèse

L'objectif de cette thèse est d'évaluer le potentiel des données acquises par le système lidar spatial GEDI pour améliorer les estimations de l'inventaire forestier. Les principales questions de recherche sont les suivantes:

1. Étant donné que les données GEDI ne sont pas spatialement continues et que la concordance spatiale entre les placettes IFN et les empreintes GEDI n'est pas assurée, une première question concerne la capacité à établir un lien entre les données de terrain (placettes IFN avec mesures dendrométriques) et les signaux GEDI.

2. Une deuxième question porte sur l'intégration des mesures GEDI dans une approche IFM, en utilisant le lien établi en 1. Il s'agit d'identifier le cadre statistique et les estimateurs appropriés pour étudier l'amélioration des résultats de l'inventaire forestier à différentes échelles de travail.

3. Une troisième question concerne l'impact de l'imprécision du géoréférencement des empreintes GEDI lors de l'intégration des données GEDI dans les approches IFM et notre capacité à développer des stratégies pour prendre en compte cette caractéristique des données. Cette question interférera avec les deux précédentes.

Pour établir un lien entre les données GEDI et celles de l'IFN, différentes stratégies peuvent être envisagées :

- Utilisation de variables a priori communes entre les empreintes GEDI et les placettes IFN (Str.1). Ces variables peuvent être identifiées sur la base de l'expertise.
- Utilisation d'un lien indirect en s'appuyant sur des données "passerelles" continues, telles que des données de télédétection continues (par exemple, des images optiques ou radar) (Str.2).
- Établir un lien direct entre les empreintes GEDI et les placettes IFN en créant une concordance spatiale. Deux stratégies principales peuvent être envisagées à cette fin. La première consisterait à acquérir des mesures supplémentaires de l'IFN sur le terrain au niveau des empreintes GEDI. Toutefois, cette stratégie n'est pas envisageable d'un point de vue opérationnel, car un travail de terrain supplémentaire considérable serait nécessaire, ce qui entraînerait une augmentation importante des coûts. La seconde consiste à utiliser des modèles de transfert radiatif pour simuler les signaux GEDI au niveau des placettes de terrain de l'IFN en exploitant les données lidar terrestres acquises au niveau d'un sous-ensemble significatif de placettes de l'IFN (Str.3).

Pour la première stratégie, i.e. l'identification de variables communes, une intersection spatiale entre les empreintes GEDI et les placettes IFN ou d'autres données auxiliaires n'est théoriquement pas nécessaire. Par conséquent, le mauvais géoréférencement des empreintes GEDI ne devrait pas poser de problème. Toutefois, la validation de la qualité du lien, i.e. de la qualité de la variable commune, peut nécessiter le recours à une intersection avec d'autres données (en utilisant par exemple les données ALS). La deuxième stratégie, qui utilise un lien indirect, nécessite une intersection avec des données auxiliaires indépendantes, créant ainsi un espace de variables communes. Comme les données sont croisées avec des données auxiliaires, la qualité du géoréférencement des données GEDI est a priori importante. Toutefois, en phase de développement d'une approche d'IFM basée sur ce lien, l'impact du géoréférencement doit être évalué. La troisième stratégie, utilisant un lien direct, nécessite la simulation des signaux GEDI aux emplacements des placettes IFN à l'aide de modèles de transfert radiatif. Pour calibrer le modèle de transfert radiatif, il faut des placettes de calibration au niveau desquelles les données GEDI et de l'IFN se chevauchent. Comme il est très peu probable qu'un chevauchement parfait se produise dans les ensembles de données existants, des placettes de terrain supplémentaires ont été acquises aux positions d'empreintes GEDI. Pour s'assurer que les mesures sur le terrain sont réalisées au bon endroit, un bon géoréférencement des empreintes GEDI est nécessaire.

Dans le cadre de cette thèse, les deux premières stratégies pour relier les empreintes GEDI et les placettes IFN, ont été explorées. La troisième stratégie est incluse dans le projet dans lequel la thèse s'inscrit (Projet SLIM), mais n'a pas été abordée dans ce travail de thèse. Toutefois, quelle que soit la stratégie étudiée, un géoréférencement précis est important, au moins pour quantifier la qualité du lien entre les empreintes GEDI et les placettes IFN. Ainsi, l'amélioration du géoréférencement de GEDI est devenue un objectif de mon travail de thèse et constitue l'élément clé pour répondre à la question de recherche n°3. Ensuite, des approches d'IFM appartenant à deux familles différentes ont été développées et étudiées. Tout d'abord, une approche basée sur le plan de sondage (design-based) visant à fournir des résultats actualisés améliorés à une échelle sous-régionale, a été développée. Deuxièmement, une approche de modélisation du lien GEDI-IFN a été proposée, qui pourrait être utilisée dans un cadre d'IFM assisté par un modèle ou basé sur un modèle, afin de fournir des estimations au niveau de petites zones (small area estimation) ou de pixels. Les questions de recherche spécifiques, les hypothèses de travail liées à ces questions et les principaux résultats sont décrits

ci-dessous.

Chapitre 2 : Améliorer le géoréférencement des empreintes GEDI en utilisant un Modèle Numérique de Terrain (MNT) à haute résolution

Les données GEDI ont montré une précision horizontale inférieure aux attentes. L'erreur planimétrique de géoréférencement est estimée à 23.8 m pour la version 1 de GEDI et à 10.2 m pour la version 2 (Beck et al., 2020, 2021). Le premier objectif de la thèse était d'améliorer le géoréférencement des données GEDI. Dans la littérature il existe des méthodes permettant d'atteindre cet objetif à l'aide des données ALS, mais ce type de données n'est pas disponible partout, nécessite des mises à jour et pose des problèmes de traitement en raison de leur volume important. En revanche, les modèles numériques de terrain (MNT) sont souvent disponibles au niveau national et ne contiennent que les élévations du sol, qui sont assez stables dans le temps et ne sont pas sujettes à des changements majeurs. Cela conduit à la formulation des deux questions de recherche suivantes :

1. La qualité du géoréférencement des empreintes GEDI peut-elle être améliorée sur une grande échelle telle que le territoire métropolitain français ?
2. Dans quelle mesure l'amélioration de la précision du géoréférencement de l'empreinte GEDI influence-t-elle la précision des résultats de l'IFM ?

Pour répondre à la première question, nous émettons l'hypothèse que l'utilisation de l'information au sol disponible grâce aux MNT à haute résolution peut être suffisante pour optimiser le géoréférencement des empreintes GEDI. Pour répondre à la deuxième question, on suppose que l'amélioration de la précision du géoréférencement de l'empreinte GEDI améliorera considérablement les résultats des approches IFM utilisant les données GEDI comme données auxiliaires.

Nous avons donc proposé une méthode de correction du géoréférencement, GeoGEDI, uniquement basée sur un MNT à haute résolution et sur les élévations du sol dérivées des données GEDI. Pour chaque empreinte, une carte d'erreur entre les estimations au sol GEDI et le MNT de référence a été calculée, et un algorithme d'accumulation de flux a été utilisé pour retrouver la position optimale de l'empreinte. La méthode GeoGEDI a été testée sur $150{,}000$ empreintes, extraites de 45 orbites, sur deux sites forestiers en France : 1) la plaine des Landes dominée par des plantations de pins et 2) le massif des Vosges composé de structures et compositions forestières diverses. L'algorithme a été appliqué aux versions 1 et 2 des données GEDI en utilisant des empreintes voisines provenant soit d'un seul faisceau laser, soit des quatre faisceaux laser de pleine puissance. La précision des résultats de GeoGEDI a été évaluée en analysant les distributions de décalage et en comparant les élévations du sol et les hauteurs de la canopée de GEDI à des valeurs de référence extraites de MNT et de modèles numériques de hauteur de la canopée (MNH) à haute résolution dérivés de données lidar et photogrammétriques aériennes.

Globalement, selon la méthode et le site d'étude, les décalages moyens entre les positions avant et après optimisation varient de 23.55 m à 23.95 m pour la version 1 et de 10.85 m à 22.2 m pour la version 2. Comme prévu, les décalages se sont révélés plus importants pour la version 1 de GEDI que pour la version 2 et ils

correspondent aux erreurs de géoréférencement annoncées par le guide d'utilisation de GEDI (Beck et al., 2021).

Pour tous les cas d'étude, notre méthode améliore l'estimation de l'élévation du sol. GeoGEDI a amélioré la RMSE de l'élévation du sol dans les Landes de 26.8% (0.34 m) pour la v1 et de 13.3% (0.14 m) pour la v2. Pour les Vosges, la RMSE de l'élévation du sol a été améliorée de 59.6% (3.82 m) pour la v1 et de 36.2% (1.41 m) pour la v2. Concernant la hauteur de la canopée, à l'exception de la v2 dans les Landes où des variations insuffisantes de la topographie combinées à des problèmes de détection du sol par la chaîne de traitement des données GEDI auraient pu pénaliser l'ajustement, GeoGEDI a amélioré la concordance entre élévations du sol et hauteurs de canopée issues de GEDI et des données de référence, preuve d'une meilleure correspondance spatiale entre les deux jeux de données.

Nous avons également étudié l'influence de la magnitude du décalage appliqué sur les écarts entre élévations du sol et de hauteurs de canopée issus de GEDI et des données de référence. Globalement, plus le décalage appliqué horizontalement est grand, plus l'amélioration des estimations verticales est importante, confirmant ainsi l'intérêt d'améliorer le géoréférencement. De plus, nous avons étudié l'influence de la pente sur les estimations. Les résultats montrent que plus le terrain est pentu, plus les écarts d'estimations sont grands et plus notre méthode améliore les estimations.

GeoGEDI a permis d'améliorer à la fois le biais et la précision du positionnement. Nos résultats ont également démontré l'intérêt de corriger le géoréférencement GEDI au niveau de l'empreinte (plutôt qu'au niveau d'un bloc d'empreintes). La méthode s'est avérée plus efficace dans les topographies contrastées que dans les zones plates, où la faible variabilité de l'élévation du sol a pénalisé l'ajustement. Par ailleurs, on a pu tester la méthode sur des données ICESat-2. Cela a montré l'absence de problème de géoréférencement sur ces données, confirmant ainsi la validité de l'approche.

Par la suite, deux méthodes utilisant les données GEDI avec les données IFN ont été développées. À titre d'exemple, ma thèse s'est concentrée sur l'estimation du volume bois sur pieds (growing stock volume en anglais : GSV). Parmi les attributs étudiés par l'IFN, le volume joue un rôle clé en fournissant des informations essentielles aux décideurs publics et aux gestionnaires forestiers (Gschwantner et al., 2022).

Chapitre 3 : Double échantillonnage pour la stratification

Dans cette étude, nous avons cherché à évaluer si les données GEDI se prêtent à une approche basée sur un double échantillonnage pour la post-stratification (DSPS). Cette méthode nécessite deux échantillons, dont l'un pour estimer les surfaces de strates et l'autre pour le calcul des attributs. Cette méthode présente l'avantage de ne pas nécessiter de géoréférencement précis, ni de co-localisation entre les empreintes GEDI et les placettes d'inventaire. La méthode DSPS repose sur deux hypothèses : 1) les plans d'échantillonnage spatiaux sont probabilistes et 2) il existe une variable de lien direct entre les deux échantillons, i.e. entre les données de terrain et les données auxiliaires GEDI. Dans notre cas d'étude la hauteur maximale a été choisie en tant que variable de lien. Pour GEDI, il s'agit de la variable RH100.

Nous avons examiné en détail le plan d'échantillonnage de GEDI et la précision des mesures de la hauteur maximale des arbres, i.e. la variable de liaison choisie, dans la zone d'étude des Vosges. Pour évaluer la qualité de la variable de lien, nous avons comparé successivement la hauteur maximale des données

terrain et le RH100 GEDI à la hauteur maximale issue de données ALS. Nos résultats ont révélé que le plan d'échantillonnage de GEDI s'écarte légèrement d'un plan d'échantillonnage probabiliste, ce qui entraîne un biais dans les estimations de proportion. Ce biais est aggravé lorsque des filtres de qualité sont appliqués. De plus, la corrélation entre hauteur maximale GEDI et ALS est assez faible (0.74). Cependant, nous avons démontré que l'analyse de cette relation est entravée par les importantes erreurs de géoréférencement dans les données GEDI.

L'objectif est d'effectuer des estimations au niveau de strates dont la surface est estimée en s'appuyant sur un échantillon de grande taille, ici celui constitué des empreintes GEDI. Il s'agit ainsi de créer des strates, c'est-à-dire des classes, qui regroupent les empreintes GEDI et les placettes IFN répondant aux mêmes critères fixés sur la base de leurs variables communes. Ainsi, par exemple, une strate va regrouper l'ensemble des données ayant une hauteur maximale de moins de 20 m. L'estimation des surfaces des strates est le résultat simple de comptage du nombre d'empreintes et n'est donc pas impactée par la précision du géoréférencement des données GEDI. Les données IFN et GEDI sont donc rattachées à des strates homogènes, puis les surfaces des strates sont estimées avec les données GEDI, tandis que les données IFN permettent de calculer les moyennes des variables forestières de chaque strate. Finalement, on peut calculer l'attribut forestier et son intervalle de confiance pour l'ensemble de la zone, puis évaluer les résultats en les comparant à une approche classique de simple échantillonnage aléatoire utilisant uniquement les données IFN, sans stratification. Nous avons utilisé $202{,}808$ empreintes GEDI comme échantillon de première phase et 482 placettes de l'IFN comme échantillon de deuxième phase pour estimer le volume sur pieds global (GSV). Par rapport aux estimations basées uniquement sur les données de l'IFN, l'approche DSPS a amélioré la variance du GSV de 56%. Bien que ce résultat soit prometteur, le schéma d'échantillonnage non probabiliste et la difficulté d'évaluer rigoureusement les imprécisions des mesures de hauteur de GEDI nous ont amenés à recommander l'utilisation des données GEDI avec prudence, que ce soit avec des approches basées sur un plan de sondage ou basées sur un modèle.

Chapitre 4 : Utilisation de données auxiliaires continues pour développer un modèle k-plus proches voisins pour prédire les attributs forestiers

Cette méthode consiste à établir un lien indirect entre les données GEDI et IFN en exploitant une troisième source de données accessible à la fois au niveau des empreintes GEDI et des placettes IFN. L'objectif est de prédire le volume et, en dernier lieu, de produire des estimations spatialisées à haute résolution avec des évaluations de l'incertitude, couvrant potentiellement des domaines géographiques de taille variable. Pour établir un lien entre les différentes sources de données et obtenir des informations sur la précision des estimations du modèle, nous avons opté pour la méthode des k-plus proches voisins (kNN) combinée avec le bagging (bootstrap aggregation). Le kNN est une approche supervisée simple et non paramétrique qui permet de prédire plusieurs attributs avec un seul modèle et qui est largement utilisée dans les études sur les IFM. Les questions de recherche relatives à cette partie de la thèse sont les suivantes :

1. L'utilisation de données auxiliaires supplémentaires peut-elle contribuer à créer un lien indirect entre les données GEDI et les données IFN ?

2. L'approche peut-elle être utilisée pour calculer des estimations sub-régionales ?

On suppose que les données Sentinel-1 et Sentinel-2, éventuellement complétées par une information sur la hauteur, sont des candidats pertinents pour jouer le rôle de données "passerelles" continues entre les placettes IFN et les empreintes GEDI. On suppose également que, grâce à ce lien indirect, un modèle peut être construit pour propager les attributs de l'IFN et produire des cartes de ressources à haute résolution.

Cette étude présente une approche combinant kNN et bagging pour prédire les attributs forestiers, en particulier le volume de bois sur pieds (GSV), en intégrant les données issues des capteurs optiques Sentinel-2, radar Sentinel-1 et lidar spatial GEDI. Tout d'abord, nous imputons les variables GEDI aux placettes de l'IFN, puis, sur la base des variables GEDI imputées et de Sentinel, nous élaborons un modèle pour prédire le GSV. Trois stratégies utilisant différents ensembles de données auxiliaires ont été utilisées : A) en s'appuyant uniquement sur les données Sentinel, B) en utilisant les données Sentinel avec une variable de hauteur maximale exclusivement disponible sur les empreintes GEDI et les placettes de l'IFN, et C) en utilisant les données Sentinel et une carte nationale des hauteurs (Schwartz et al., 2023). L'ensemble de la méthode a été appliqué sur le même jeu de données GEDI, une fois sans et une fois avec l'application de l'algorithme GeoGEDI pour améliorer le géoréférencement.

Les variables Sentinel-2 incluent les bandes spectrales, des indices de végétation (NDVI, MSAVI, NDWI, GLI) et des variables biophysiques de la végétation (LAI, fCover, fAPAR) issues d'images de synthèse mensuelle d'août 2022, ainsi que les différences avec les images de synthèses mensuelles d'août 2017 et de juin 2022. Pour Sentinel-1, des variables extraites des polarisations verticale-verticale et verticale-horizontale, acquises en août 2022, sont utilisées. Les variables GEDI comprennent le taux de couvert, l'indice de surface foliaire (PAI), la densité volumique de surface foliaire (PAVD), l'indice de diversité de la hauteur du feuillage (fhd_normal) et les métriques de hauteur relative (corrigées pour inclure uniquement les composants végétaux (RHv)).

Avant d'appliquer le kNN-bagging une réduction du nombre de variables est réalisée. La méthode utilisée ici est basée uniquement sur la corrélation entre les variables d'entrée et les variables cibles.

Pour la première étape (prédire les variables GEDI au niveau des placettes d'inventaire), les caractéristiques Sentinel extraites aux positions corrigées de GEDI, présentent une corrélation plus marquée avec les variables de GEDI que celles extraites au niveau des positions non corrigées. Par exemple, le coefficient de corrélation de l'indicateur de verdure (greenness) est de 0.40 pour les positions corrigées et de 0.37 pour les positions non corrigées. Pour la deuxième étape (prédire le volume), la variable la plus corrélée au volume (pour les empreintes corrigées avec GeoGEDI) est RHv_60 (r = 0.69).

L'utilisation des seules données Sentinel s'est révélée insuffisante pour imputer les variables GEDI. Davantage de données auxiliaires sont nécessaires. La stratégie C (avec carte de hauteurs) permet d'améliorer les estimations par rapport à la stratégie A (sans hauteurs), et la stratégie B améliore encore ces estimations. La RMSE est de $206.62\ m^3\ ha^{-1}$ pour la stratégie A, de $165.49\ m^3\ ha^{-1}$ pour la stratégie C et de $133.70\ m^3\ ha^{-1}$ pour la stratégie B. Néanmoins, toutes les stratégies montrent un effet de saturation pour les volumes élevés qui sont ainsi sous-estimés.

Les résultats finaux ont montré que les jeux de données corrigés par GeoGEDI étaient plus performants pour prédire le volume de bois que les données non corrigées. Pour la stratégie B, la RMSE pour les volumes prédits avec GeoGEDI était de $133.70\ m^3\ ha^{-1}$, tandis qu'elle était de $149.87\ m^3\ ha^{-1}$ sans GeoGEDI. Les valeurs de R^2 étaient respectivement de 0.58 et 0.47, et la corrélation était de 0.77 et 0.69.

La stratégie B a surpassé toutes les autres testées dans cette étude et s'est révélée être la stratégie la plus prometteuse. Cependant, elle ne permet pas de cartographier pixel par pixel le volume sur pied (GSV). Néanmoins, les points de GSV densifiés estimés à chaque position GEDI présentent un potentiel pour développer des approches d'Inventaire Forestier Multisource assistées d'un modèle pour des estimations à l'échelle de petites zones, comme les communes, et offrant également des perspectives pour les estimations locales avec des approches d'estimation de petites surfaces (SAE, Small Area Estimation en anglais). En revanche, il est important de souligner que les cartes de volumes à haute résolution (par exemple avec des prédictions au niveau de chaque pixel Sentinel, soit avec une résolution de $\sim 30\,\mathrm{m}$, telles que celles présentées dans Potapov et al. (2021); Lang et al. (2023), ou encore Schwartz et al. (2023), doivent être utilisées avec précaution car elles présentent des erreurs locales parfois importantes et non quantifiées.

Chapitre 5 : Conclusion et perspectives

Nous avons présenté l'amélioration du géoréférencement des empreintes GEDI en utilisant un MNT à haute résolution au chapitre 2, une approche basée sur le plan de sondage et faisant appel à une stratification pour utiliser les données GEDI dans les estimations des inventaires forestiers au chapitre 3, et une approche kNN-bagging utilisant les données GEDI et Sentinel pour développer un modèle permettant de produire des estimations des paramètres forestiers, ici le volume de bois (growing stock volume (GSV)), par des approches d'IFM assistées d'un modèle ou basées sur un modèle au chapitre 4.

GEDI présente de nombreux avantages...

Les faisceaux laser de GEDI pénètrent le couvert végétal, permettant l'évaluation de la hauteur de la canopée et de la distribution verticale de la végétation. Bien que la période initiale d'acquisition de GEDI ne s'étend que sur quatre ans, elle a démontré avec succès l'intérêt de ce type de données pour répondre à des besoins en informations sur les écosystèmes forestiers. Sa capacité à couvrir des zones étendues avec une densité élevée et un renouvellement régulier a conduit à la prolongation de sa mission, soulignant son importance. Cette deuxième période d'acquisition offrira l'opportunité d'étudier les tendances à long terme et de comprendre l'impact des perturbations sur les forêts. De plus, les données GEDI sont accessibles gratuitement, permettant une large utilisation des données auprès de l'ensemble des acteurs.

... mais aussi des limitations

L'une des principales difficultés associées à l'utilisation des données GEDI, comme soulignée dans cette thèse, réside dans la mauvaise qualité de son géoréférencement. Positionné sur l'ISS, l'instrument de GEDI est susceptible de mouvements, et les défis de géoréférencement de ses empreintes sont largement discutés dans le chapitre 2. Les coordonnées des empreintes calculées en tenant compte des informations de trajectographie et d'orientation du système enregistrées à bord, qui sont communiquées avec les données, peuvent facilement être décalées de $10\,\mathrm{m}$ par rapport à leur position réelle, ce qui pose problème lors de leur intersection avec d'autres données géoréférencées.

Bien que la technologie lidar full-waveform de GEDI fournisse des informations riches, la traduction de

ces formes d'onde en variables peut être complexe, comme démontré par la détection erronée du pic au sol dans la section 2.3.2. De nombreux filtres existants et supplémentaires présentés dans la section 3 ont été appliqués. Certaines des empreintes filtrées résultent de difficultés d'acquisition, tandis que d'autres se démarquent du fait de leur contexte environnemental (pente, végétation dense). Plutôt que de les rejeter, une amélioration de la conversion de la forme d'onde en variables serait plus appropriée.

L'efficacité de GEDI dépend de sa capacité à pénétrer dans le couvert végétal. Dans les régions avec une végétation dense ou des structures de couvert complexes, la précision peut être compromise, entraînant potentiellement une sous-estimation de la hauteur du couvert végétal.

De plus, la distribution spatiale des empreintes GEDI ne peut pas être assimilée à un échantillonnage probabiliste, comme indiqué dans la section 3. Certaines zones sont plus densément couvertes que d'autres. Cela devrait être amélioré lors de la deuxième phase d'acquisition débutant à l'automne 2024. En effet, GEDI a été développé pour créer un échantillonnage spatial optimisé à la hauteur planifiée de l'ISS. Cependant, l'orbite de l'ISS a été élevée de manière inattendue, entraînant un changement dans la distribution spatiale des mesures GEDI. Les altitudes plus élevées de l'ISS entraînent une résonance orbitale et une couverture réduite (Dubayah et al., 2022a). L'ISS est maintenant de retour à la hauteur initialement prévue, donc la distribution spatiale des empreintes GEDI devrait être améliorée pour la prochaine phase d'acquisition. Espérons qu'aucun changement imprévu ne se produira.

De plus, GEDI ne couvre pas les latitudes au-dessus de 51.6° Nord ou en dessous de 51.6° Sud, excluant des zones forestières importantes au niveau mondial telles que le Canada ou les pays européens nordiques. L'orbite de l'ISS restreint les observations au-delà de ces latitudes ; un satellite dédié serait nécessaire pour une couverture dans ces zones.

Intégration des données GEDI dans l'IFN

La difficulté principale d'intégration des données GEDI dans l'IFN réside dans la non-concordance spatiale de ces deux jeux de données non continues.

Nous avons exploré deux liens dans le cadre de cette thèse : l'utilisation de variables communes telles que la hauteur maximale dans une approche de stratification (Chapitre 3), et l'utilisation de données auxiliaires indirectes telles que Sentinel dans une approche kNN-bagging (Chapitre 4). L'approche de stratification s'est montrée très prometteuse, malgré l'écart entre les caractéristiques de l'échantillonnage GEDI et celles d'un échantillonnage probabiliste. Le modèle de kNN-bagging utilisant uniquement des données Sentinel comme lien indirect s'est montré très limité pour prédire le volume. En rajoutant une hauteur comme variable explicative supplémentaire, une amélioration significative des résultats a été obtenue, au détriment de la capacité de spatialisation. Cette capacité de spatialisation peut-être obtenue via l'exploitation de cartes de hauteurs globales ou nationales obtenues par des méthodes d'apprentissage profond. Mais la qualité locale de ces cartes doit être améliorée et qualifiée.

L'utilisation d'un lien direct en utilisant un modèle de transfert radiatif pour simuler les signaux GEDI sur des placettes IFN, est également envisageable. Pour cela, 100 placettes terrain ont été visitées au niveau d'empreintes GEDI (après amélioration de leur géoréférencement). Des données IFN ainsi que des données TLS ont été acquises sur ces placettes. Dans le cadre du projet dans lequel s'insère cette thèse, ce jeu de don-

nées GEDI-TLS de 100 points doit être utilisé pour calibrer un modèle reliant l'information contenue dans le signal GEDI à celle des placettes IFN. Une fois établi, ce modèle pourrait être appliqué à d'autres acquisitions TLS, permettant la simulation des empreintes GEDI au niveau de grands ensembles de données de référence terrain. Entre 2010 et 2015, l'IFN français a collecté des données TLS en parallèle des collectes de données sur le terrain pour 1338 placettes. Ce vaste ensemble de données, couvrant une échelle nationale, constitue une ressource précieuse pour l'étalonnage et l'application du modèle.

La mission GEDI est la première mission lidar spatiale dédiée aux forêts. La conception de la mission a intégré une approche statistique robuste visant à fournir, pour une grille kilométrique, des estimations de moyennes d'attributs forestiers ainsi que de la variance associée (Dubayah et al., 2022a). Des approches basées sur des modèles ont été proposées pour produire des estimations dans les mailles peu à pas échantillonnées (Saarela et al., 2018). Malheureusement, le changement d'orbite de l'ISS n'a pas permis de réaliser l'échantillonnage initialement prévu. Néanmoins, l'analyse des données a démontré le potentiel de ces mesures pour la caractérisation et le suivi des forêts, soulignant ainsi la nécéssité de poursuivre les efforts de développement des solutions lidar spatiaux dédiées aux forêts.

Le développement des approches d'apprentissage profond a également apporté une dimension cartographique aux données GEDI, en les combinant avec des données optiques et/ou radar (Potapov et al., 2021; Lang et al., 2023; Schwartz et al., 2023; Morin et al., 2022). Ces cartographies fournissent des informations au niveau des points d'inventaire de terrain, s'avérant utile pour le développement de modèles prédictifs d'attributs d'intérêt tels que le volume ou la biomasse. Cependant, ces cartes sont rarement accompagnées de cartes d'incertitudes, à même de renseigner sur des biais locaux. En cas de modèle mal spécifié, leur exploitation à des fins de gestion pourrait conduire à des mauvais choix ou décisions. Comme l'atteste le titre de l'article de McRoberts (2011) « Satellite image-based maps: Scientific inference or pretty pictures? » (Cartes basées sur des images satellites : inférences scientifiques ou belles images ?), ce problème est connu depuis longtemps et la communauté des chercheurs travaillant dans le domaine de l'inventaire forestier y est sensible, mais ce n'est pas le cas de toutes les communautés scientifiques et d'utilisateurs qui peuvent prendre pour acquis des cartes pourtant biaisées.

Par ailleurs, les méthodes actuelles de spatialisation des données GEDI s'opèrent au prix d'une perte d'information importante sur le profil vertical. De plus, la performance des prédictions par apprentissage profond ne parvient pas (encore) à surmonter les limites imposées par les données optiques et radar, notamment la saturation des signaux dans les niveaux moyens de biomasse. Bien que les cartes haute résolution présentent de nombreux intérêts sur le plan scientifique, leur utilité pour la gestion forestière et les décisions publiques est probablement limitée.

Les gestionnaires forestiers privilégient les données à l'échelle des parcelles, tandis que les décisions publiques reposent principalement sur des découpages administratifs (e.g. les départements ou communes). À ces échelles, les estimations par stratification et les estimations de petits domaines géographiques assistées d'un modèle, offrent de meilleures garanties concenrant les biais et doivent être privilégiées (Ståhl et al., 2024; Zhang et al., 2022; Breidenbach and Astrup, 2012). Ces approches sont également en ligne avec les objectifs de la mission GEDI en matières d'inférence statistique. L'approche de stratification présentée dans le Chapitre 3 pour une partie du massif des Vosges peut être facilement étendue à l'échelle nationale en utilisant les données GEDI et la BD Forêt, offrant une flexibilité d'échelle d'estimation.

L'amélioration des modèles de prédictions et des méthodes sous-jacentes reste un enjeu important en

recherche. Des modèles mieux spécifiés permettent une meilleure maîtrise des biais et offrent davantage de flexibilité dans les procédures d'estimation. De meilleurs modèles permettraient de recourir avec plus de confiance aux approches basées sur des modèles et permettraient d'estimer des attributs forestiers dans des zones dépourvues de points terrain. De telles approches pourraient être bénéfiques pour effectuer des estimations dans les zones ayant subi des perturbations. Dans ce contexte, une approche hybride intégrant le jeu de données densifié par le kNN-bagging dans une stratification, pourrait également être mise en place. L'exploration de sources de données complémentaires ou alternatives, telles que les données acquises par ICESat-2 ou la couverture nationale par lidar aéroporté (lidar HD de l'IGN) actuellement en cours d'acquisition en France, devrait permettre d'améliorer les capacités d'estimation et de cartographie des attributs forestiers à l'échelle nationale, régionale et sub-régionale.

Bibliography

[1] Adam, M., Urbazaev, M., Dubois, C., Schmullius, C., 2020. Accuracy Assessment of GEDI Terrain Elevation and Canopy Height Estimates in European Temperate Forests: Influence of Environmental and Acquisition Parameters. Remote Sensing 12, 3948. doi:10/gpfhdv.

[2] Aleissaee, A.A., Kumar, A., Anwer, R.M., Khan, S., Cholakkal, H., Xia, G.S., Khan, F.S., 2023. Transformers in Remote Sensing: A Survey. Remote Sensing 15, 1860. doi:10.3390/rs15071860.

[3] Appiah Mensah, A., Jonzén, J., Nyström, K., Wallerman, J., Nilsson, M., 2023. Mapping site index in coniferous forests using bi-temporal airborne laser scanning data and field data from the Swedish national forest inventory. Forest Ecology and Management 547, 121395. doi:10.1016/j.foreco.2023.121395.

[4] Asner, G.P., Mascaro, J., Muller-Landau, H.C., Vieilledent, G., Vaudry, R., Rasamoelina, M., Hall, J.S., van Breugel, M., 2012. A universal airborne LiDAR approach for tropical forest carbon mapping. Oecologia 168, 1147–1160. doi:10.1007/s00442-011-2165-z.

[5] Bagley, J.E., Desai, A.R., Harding, K.J., Snyder, P.K., Foley, J.A., 2014. Drought and Deforestation: Has Land Cover Change Influenced Recent Precipitation Extremes in the Amazon? Journal of Climate 27, 345–361. doi:10.1175/JCLI-D-12-00369.1.

[6] Bechtold, W.A., Patterson, P.L., 2015. The Enhanced Forest Inventory and Analysis Program - National Sampling Design and Estimation Procedures. URL: http://dx.doi.org/10.2737/SRS-GTR-80, doi:10.2737/srs-gtr-80.

[7] Beck, J., Luthcke, S., Hofton, M., Armston, J., 2020. Global Ecosystem Dynamics Investigation (GEDI) level 1B user guide. Document version 1.0. USGS Earth Resources Observation and Science (EROS) Center: NASA's Land Processes Distributed Active Archive Center (LP DAAC) URL: https://lpdaac.usgs.gov/documents/987/GEDI01B_User_Guide_V1.pdf.

[8] Beck, J., Wirt, B., Luthcke, S., Hofton, M., Armston, J., 2021. Global Ecosystem Dynamics Investigation (GEDI) level 1B user guide. Document version 2.0. USGS Earth Resources Observation and Science (EROS) Center: NASA's Land Processes Distributed Active Archive Center (LP DAAC) URL: https://lpdaac.usgs.gov/documents/987/GEDI01B_User_Guide_V2.pdf.

[9] Beland, M., Parker, G., Sparrow, B., Harding, D., Chasmer, L., Phinn, S., Antonarakis, A., Strahler, A., 2019. On promoting the use of lidar systems in forest ecosystem research. Forest Ecology and Management 450. doi:10.1016/j.foreco.2019.117484.

[10] Besic, N., Vega, C., 2023. An example of the conjoint use of different remote sensing sensors and the NFI in france, in: SilviLaser, London.

[11] Blair, J., Hofton, M., 1999. Modeling Laser Altimeter Return Waveforms Over Complex Vegetation Using High-Resolution Elevation Data. Geophysical Research Letters 26. doi:10.1029/1999GL010484.

[12] Bonan, G.B., 2008. Forests and Climate Change: Forcings, Feedbacks, and the Climate Benefits of Forests. Science 320, 1444–1449. doi:10.1126/science.1155121.

[13] Bontemps, J.D., Denardou, A., Hervé, J.C., Bir, J., Dupouey, J.L., 2020. Unprecedented pluri-decennial increase in the growing stock of French forests is persistent and dominated by private broadleaved forests. Annals of Forest Science 77, 98. doi:10.1007/s13595-020-01003-6.

[14] Boucher, P.B., Hancock, S., Orwig, D.A., Duncanson, L., Armston, J., Tang, H., Krause, K., Cook, B., Paynter, I., Li, Z., Elmes, A., Schaaf, C., 2020. 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, 1304. doi:10.3390/rs12081304.

[15] Bouriaud, O., 2020. Échantillonnage et estimation dans l'Inventaire Forestier National. Essai de reconstruction et formalisation. Research Report. Institut National de l'Information Géographique et Forestière ; Laboratoire d'Inventaire Forestier. URL: https://hal.science/hal-03039886.

[16] Bouriaud, O., Morneau, F., Bontemps, J.D., 2023. Square-grid sampling support to reconcile systematicity and adaptivity in periodic spatial surveys of natural populations. Journal of Vegetation Science 34. doi:10.1111/jvs.13195.

[17] Bouvier, M., Durrieu, S., Fournier, R., Saint-Geours, N., Guyon, D., Grau, E., De Boissieu, F., 2019. Influence of Sampling Design Parameters on Biomass Predictions Derived from Airborne LiDAR Data. Canadian Journal of Remote Sensing 45, 1–23. doi:10.1080/07038992.2019.1669013.

[18] Breidenbach, J., Astrup, R., 2012. Small area estimation of forest attributes in the Norwegian National Forest Inventory. European Journal of Forest Research 131, 1255–1267. doi:10.1007/s10342-012-0596-7.

[19] Breidenbach, J., McRoberts, R.E., Alberdi, I., Antón-Fernández, C., Tomppo, E., 2021. A century of national forest inventories – informing past, present and future decisions. Forest Ecosystems 8, 36. doi:10.1186/s40663-021-00315-x.

[20] Breiman, L., 1996. Bagging predictors. Machine Learning 24, 123–140. doi:10.1007/BF00058655.

[21] Brown, P., Engelmann, A., 2019. TSIS Experiences with ISS Jitter from Inception to On-Orbit Operation - NASA Technical Reports Server (NTRS). URL: https://ntrs.nasa.gov/citations/20190000637.

[22] Bruening, J., May, P., Armston, J., Dubayah, R., 2023. Precise and unbiased biomass estimation from GEDI data and the US Forest Inventory. Frontiers in Forests and Global Change 6. doi:10.3389/ffgc.2023.1149153.

[23] Bullock, E.L., Healey, S.P., Yang, Z., Acosta, R., Villalba, H., Insfrán, K.P., Melo, J., Wilson, S., Duncanson, L.I., Næsset, E., Armston, J., Saarela, S., Ståhl, G., Patterson, P.L., Dubayah, R., 2023. Estimating

aboveground biomass density using hybrid statistical inference with GEDI lidar data and Paraguay's national forest inventory. Environmental Research Letters doi:10.1088/1748-9326/acdf03.

[24] CAMS (Copernicus Atmosphere Monitoring Service), 2023. Northern hemisphere wildfires: A summer of extremes. https://atmosphere.copernicus.eu/northern-hemisphere-wildfires-summer-extremes. (accessed Dec. 2, 2023).

[25] Cavaignac, S., 2009. Les sylvoécorégions (SER) de France métropolitaine : Étude de définition. URL: https://inventaire-forestier.ign.fr/IMG/pdf/Part1_rapport_ser.pdf.

[26] CCFM (Canadian Council of Forest Ministers), 2020. National forestry database - base de données nationales des forêts - canada (version 2.0.0) [data set].natural resources canada – ressources naturelles canada. URL: http://nfdp.ccfm.org/en/data/insects.php, doi:http://doi.org/10.5281/zenodo.3690046.

[27] Charbonneau, F., Trudel, M., Fernandes, R., 2005. Ruse of dual polarization and multi-incidence sar for soil permeability mapping, in: In Proceedings of the 2005 Advanced Synthetic Aperture Radar (ASAR) Workshop, St-Hubert, QC, Canada.

[28] Chen, M., Dong, W., Yu, H., Woodhouse, I., Ryan, C.M., Liu, H., Georgiou, S., Mitchard, E.T.A., 2023. Multimodal deep learning for mapping forest dominant height by fusing GEDI with earth observation data. URL: https://arxiv.org/abs/2311.11777v1.

[29] Chirici, G., Mura, M., McInerney, D., Py, N., Tomppo, E., Waser, L., Travaglini, D., Mcroberts, R., 2016. A meta-analysis and review of the literature on the k-Nearest Neighbors technique for forestry applications that use remotely sensed data. Remote Sensing of Environment 176, 282–294. doi:10.1016/j.rse.2016.02.001.

[30] Christianson, D.S., Kaufman, C.G., 2016. Effects of sample design and landscape features on a measure of environmental heterogeneity. Methods in Ecology and Evolution 7, 770–782. doi:https://doi.org/10.1111/2041-210X.12539.

[31] CIFFC (Canadian Interagency Forest Fire Center), 2023. Fire situation report - rapport national sur les incendies de forêt. URL: https://ciffc.net/. (accessed Dec. 2, 2023).

[32] Cochran, W.G., 1977. Sampling techniques. John Wiley & Sons.

[33] Coops, N.C., Tompalski, P., Goodbody, T.R.H., Achim, A., Mulverhill, C., 2023. Framework for near real-time forest inventory using multi source remote sensing data. Forestry: An International Journal of Forest Research 96, 1–19. doi:10.1093/forestry/cpac015.

[34] Coops, N.C., Tompalski, P., Goodbody, T.R.H., Queinnec, M., Luther, J.E., Bolton, D.K., White, J.C., Wulder, M.A., van Lier, O.R., Hermosilla, T., 2021. Modelling lidar-derived estimates of forest attributes over space and time: A review of approaches and future trends. Remote Sensing of Environment 260, 112477. doi:10.1016/j.rse.2021.112477.

[35] Copernicus Sentinel-1, (processed by ESA), Retrieved from Google Earth Engine .

[36] Copernicus Sentinel-2, 2021. (processed by ESA), MSI Level-2A BOA Reflectance Product. Collection 1. doi:10.5270/S2_-znk9xsj.

[37] Corona, P., Fattorini, L., Franceschi, S., Scrinzi, G., Torresan, C., 2014. Estimation of standing wood volume in forest compartments by exploiting airborne laser scanning information: Model-based, design-based, and hybrid perspectives. Canadian Journal of Forest Research 44, 1303–1311. doi:10.1139/cjfr-2014-0203.

[38] Couteron, P., Barbier, N., Deblauwe, V., Pelissier, R., Ploton, P., 2015. Texture analysis of very high spatial resolution optical images as a way to monitor vegetation and forest biomass in the tropics, in: Murthy, M., Wesselman, S., Gilani, H. (Eds.), Multi-Scale Forest Biomass Assessment and Monitoring in the Hindu Kush Himalayan Region: A Geospatial Perspective. International Centre for Integrated Mountain Development (ICIMOD). ICIMOD special publication, pp. 157–164. URL: https://hal.umontpellier.fr/hal-02303735.

[39] Cover, T., Hart, P., 1967. Nearest neighbor pattern classification. IEEE Transactions on Information Theory 13, 21–27. doi:10.1109/TIT.1967.1053964.

[40] Crookston, N.L., Finley, A.O., 2008. yaImpute: An R Package for kNN Imputation. Journal of Statistical Software 23, 1–16. doi:10.18637/jss.v023.i10.

[41] Desbureaux, S., Damania, R., 2018. Rain, forests and farmers: Evidence of drought induced deforestation in Madagascar and its consequences for biodiversity conservation. Biological Conservation 221, 357–364. doi:10.1016/j.biocon.2018.03.005.

[42] Dirzo, R., Young, H., Galetti, M., Ceballos, G., Isaac, N., Collen, B., 2014. Defaunation in the Anthropocene. Science (New York, N.Y.) 345, 401–6. doi:10.1126/science.1251817.

[43] Dorado-Roda, I., Pascual, A., Godinho, S., Silva, C., Botequim, B., Rodríguez-Gonzálvez, P., González-Ferreiro, E., Guerra, J., 2021. Assessing the Accuracy of GEDI Data for Canopy Height and Aboveground Biomass Estimates in Mediterranean Forests. Remote Sensing 13, 2279. doi:10/gkjf2c.

[44] Dou, C., Zhang, X., Guo, H., Han, C., Liu, M., 2014. Improving the Geolocation Algorithm for Sensors Onboard the ISS: Effect of Drift Angle. Remote Sensing 6, 4647–4659. doi:10.3390/rs6064647.

[45] Dubayah, R., Armston, J., Healey, S., Yang, Z., Patterson, P., Saarela, S., Stahl, G., Duncanson, L., Kellner, J., Bruening, J., Pascual, A., 2023. GEDI L4B Gridded Aboveground Biomass Density, Version 2.1. URL: https://daac.ornl.gov/cgi-bin/dsviewer.pl?ds_id=2299, doi:10.3334/ORNLDAAC/2299. oRNL Distributed Active Archive Center.

[46] Dubayah, R., Armston, J., Healey, S.P., Bruening, J.M., Patterson, P.L., Kellner, J.R., Duncanson, L., Saarela, S., Ståhl, G., Yang, Z., Tang, H., Blair, J.B., Fatoyinbo, L., Goetz, S., Hancock, S., Hansen, M., Hofton, M., Hurtt, G., Luthcke, S., 2022a. GEDI launches a new era of biomass inference from space 17, 095001. doi:10.1088/1748-9326/ac8694.

[47] Dubayah, R., Armston, J., Kellner, J., Duncanson, L., Healey, S., Patterson, P., Hancock, S., Tang, H., Bruening, J., Hofton, M., Blair, J., Luthcke, S., 2022b. GEDI L4A footprint level aboveground biomass density, version 2.1. URL: https://daac.ornl.gov/cgi-bin/dsviewer.pl?ds_id=2056, doi:10.3334/ORNLDAAC/2056. oRNL Distributed Active Archive Center.

[48] Dubayah, R., Blair, J.B., Goetz, S., Fatoyinbo, L., Hansen, M., Healey, S., Hofton, M., Hurtt, G., Kellner, J., Luthcke, S., Armston, J., Tang, H., Duncanson, L., Hancock, S., Jantz, P., Marselis, S., Patterson, P.L.,

Qi, W., Silva, C., 2020a. The Global Ecosystem Dynamics Investigation: High-resolution laser ranging of the Earth's forests and topography. Science of Remote Sensing 1, 100002. doi:10.1016/j.srs.2020.100002.

[49] Dubayah, R., Hofton, M., Blair, J., Armston, J., Tang, H., Luthcke, S., 2020b. GEDI L2A Elevation and Height Metrics Data Global Footprint Level V001, distributed by NASA EOSDIS Land Processes DAAC doi:https://doi.org/10.5067/GEDI/GEDI02_A.001.

[50] Dubayah, R., Hofton, M., Blair, J., Armston, J., Tang, H., Luthcke, S., 2021a. GEDI L2A Elevation and Height Metrics Data Global Footprint Level V002, distributed by NASA EOSDIS Land Processes DAAC doi:https://doi.org/10.5067/GEDI/GEDI02_A.002.

[51] Dubayah, R., Luthcke, S., Sabaka, T., Nicholas, J., Preaux, S., Hofton, M., 2021b. GEDI L3 Gridded Land Surface Metrics, Version 2. URL: https://daac.ornl.gov/cgi-bin/dsviewer.pl?ds_id=1952, doi:10.3334/ORNLDAAC/1952. oRNL Distributed Active Archive Center.

[52] Dubayah, R., Tang, H., Armston, J., Luthcke, S., Hofton, M., Blair, J., 2021c. GEDI L2B Canopy Cover and Vertical Profile Metrics Data Global Footprint Level V002, distributed by NASA EOSDIS Land Processes DAAC doi:10.5067/GEDI/GEDI02_B.002.

[53] Duncanson, L., Kellner, J., Armston, J., Dubayah, R., Minor, D., Hancock, S., Healey, S., Patterson, P., Saarela, S., Marselis, S., Silva, C., Bruening, J., Goetz, S., Tang, H., Hofton, M., Blair, B., Luthcke, S., Fatoyinbo, L., Abernethy, K., Zgraggen, C., 2021. Aboveground Biomass Density Models for NASA's Global Ecosystem Dynamics Investigation (GEDI) Lidar Mission. Remote Sensing of Environment doi:10/gn3jrm.

[54] Duncanson, L., Neuenschwander, A., Hancock, S., Thomas, N., Fatoyinbo, T., Simard, M., Silva, C.A., Armston, J., Luthcke, S.B., Hofton, M., Kellner, J.R., Dubayah, R., 2020. Biomass estimation from simulated GEDI, ICESat-2 and NISAR across environmental gradients in Sonoma County, California. Remote Sensing of Environment 242, 111779. doi:10.1016/j.rse.2020.111779.

[55] Dyderski, M.K., Paź, S., Frelich, L.E., Jagodziński, A.M., 2018. How much does climate change threaten European forest tree species distributions? Global Change Biology 24, 1150–1163. doi:10.1111/gcb.13925.

[56] East, A., Hansen, A., Armenteras, D., Jantz, P., Roberts, D.W., 2023. 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, 696. doi:10.3390/rs15030696.

[57] EEA (European Environment Agency), Climate change is one of the biggest challenges of our times. URL: https://www.eea.europa.eu/themes/climate/climate-change-is-one-of. (accessed Feb. 03, 2024).

[58] EFA (European Forest Accounts), 2023. Forests, forestry and logging. URL: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Forests,_forestry_and_logging. (accessed Dec. 11, 2023).

[59] Ehlers, S., Saarela, S., Lindgren, N., Lindberg, E., Nyström, M., Persson, H., Olsson, H., Ståhl, G., 2018. Assessing Error Correlations in Remote Sensing-Based Estimates of Forest Attributes for Improved Composite Estimation. Remote Sensing 10, 667. doi:10.3390/rs10050667.

[60] Eng, L., Ismail, R., Hashim, W., Baharum, A., 2019. The Use of VARI, GLI, and VIgreen Formulas in Detecting Vegetation In aerial Images. International Journal of Technology 10, 1385. doi:10.14716/ijtech.v10i7.3275.

[61] ESA (European Space Agency), a. URL: https://sentinels.copernicus.eu/web/sentinel/missions/sentinel-2. (accessed Nov. 14, 2023).

[62] ESA (European Space Agency), b. URL: https://sentinels.copernicus.eu/web/sentinel/technical-guides/sentinel-1-sar. (accessed Nov. 14, 2023).

[63] European Commission, 2018. A sustainable bioeconomy for Europe: strengthening the connection between economy, society and the environment. doi:doi/10.2777/792130.

[64] European Commission, 2021. Communication from the commission to the european parliament, the council, the european economic and social committee and the committee of the regions : New EU Forest Strategy for 2030 .

[65] FAO, SER, IUCN/CEM, 2023. Standards of Practice to Guide Ecosystem Restoration: A Contribution to the United Nations Decade on Ecosystem Restoration: Summary Report. FAO, Rome, Italy. doi:10.4060/cc5223en.

[66] FAO (Food and Agriculture Organization), 2020. Global Forest Resources Assessment 2020: Main Report. FAO, Rome, Italy. doi:10.4060/ca9825en.

[67] Fayad, I., Baghdadi, N., Bailly, J.S., Frappart, F., Zribi, M., 2020. Analysis of GEDI Elevation Data Accuracy for Inland Waterbodies Altimetry. Remote Sensing 12, 23. doi:10.3390/rs12172714.

[68] Feret, J.B., de Boissieu, F., 2023. prosail: PROSAIL leaf and canopy radiative transfer model and inversion routines. URL: https://gitlab.com/jbferet/prosail. R package version 2.2.2.

[69] Fernandez-Ordonez, Y., Soria-Ruiz, J., Leblon, B., Fernandez-Ordonez, Y., Soria-Ruiz, J., Leblon, B., 2009. Forest Inventory using Optical and Radar Remote Sensing, in: Advances in Geoscience and Remote Sensing. IntechOpen. doi:10.5772/8330.

[70] Fettig, C.J., Egan, J.M., Delb, H., Hilszczański, J., Kautz, M., Munson, A.S., Nowak, J.T., Negrón, J.F., 2022. 11 - Management tactics to reduce bark beetle impacts in North America and Europe under altered forest and climatic conditions, in: Gandhi, K.J.K., Hofstetter, R.W. (Eds.), Bark Beetle Management, Ecology, and Climate Change. Academic Press, pp. 345–394. doi:10.1016/B978-0-12-822145-7.00006-4.

[71] Fortin, M., Manso, R., Schneider, R., 2018. Parametric bootstrap estimators for hybrid inference in forest inventories. Forestry: An International Journal of Forest Research 91, 354–365. doi:10.1093/forestry/cpx048.

[72] Forzieri, G., Girardello, M., Ceccherini, G., Spinoni, J., Feyen, L., Hartmann, H., Beck, P.S.A., Camps-Valls, G., Chirici, G., Mauri, A., Cescatti, A., 2021. Emergent vulnerability to climate-driven disturbances in European forests. Nature Communications 12, 1081. doi:10.1038/s41467-021-21399-7.

[73] Freeman, T., 1991. Calculating Catchment Area with Divergent Flow Based on a Regular Grid. Computers & Geosciences 17, 413–422. doi:10.1016/0098-3004(91)90048-I.

[74] Gao, B.c., 1996. NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. Remote Sensing of Environment 58, 257–266. doi:10.1016/S0034-4257(96)00067-3.

[75] Ge, S., Gu, H., Su, W., Praks, J., Antropov, O., 2022. Improved semisupervised unet deep learning model for forest height mapping with satellite sar and optical data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 15, 5776–5787. doi:10.1109/JSTARS.2022.3188201.

[76] Ginzler, C., Hobi, M.L., 2015. Countrywide Stereo-Image Matching for Updating Digital Surface Models in the Framework of the Swiss National Forest Inventory. Remote Sensing 7, 4343–4370. doi:10.3390/rs70404343.

[77] Gobakken, T., Næsset, E., Nelson, R., Bollandsås, O.M., Gregoire, T.G., Ståhl, G., Holm, S., Ørka, H.O., Astrup, R., 2012. Estimating biomass in Hedmark County, Norway using national forest inventory field plots and airborne laser scanning. Remote Sensing of Environment 123, 443–456. doi:10.1016/j.rse.2012.01.025.

[78] Gomes, H.M., Read, J., Bifet, A., Durrant, R.J., 2021. Learning from evolving data streams through ensembles of random patches doi:10.25455/wgtn.20104412.

[79] Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google earth engine: Planetary-scale geospatial analysis for everyone. Remote Sensing of Environment doi:10.1016/j.rse.2017.06.031.

[80] Gregoire, T.G., 1998. Design-based and model-based inference in survey sampling: Appreciating the difference. Canadian Journal of Forest Research 28, 1429–1447. doi:10.1139/x98-166.

[81] Gregoire, T.G., Næsset, E., McRoberts, R.E., Ståhl, G., Andersen, H.E., Gobakken, T., Ene, L., Nelson, R., 2016. Statistical rigor in LiDAR-assisted estimation of aboveground forest biomass. Remote Sensing of Environment 173, 98–108. doi:10.1016/j.rse.2015.11.012.

[82] Gschwantner, T., Alberdi, I., Balázs, A., Bauwens, S., Bender, S., Borota, D., Bosela, M., Bouriaud, O., Cañellas, I., Donis, J., Freudenschuß, A., Hervé, J.C., Hladnik, D., Jansons, J., Kolozs, L., Korhonen, K.T., Kucera, M., Kulbokas, G., Kuliešis, A., Lanz, A., Lejeune, P., Lind, T., Marin, G., Morneau, F., Nagy, D., Nord-Larsen, T., Nunes, L., Pantić, D., Paulo, J.A., Pikula, T., Redmond, J., Rego, F.C., Riedel, T., Saint-André, L., Šebeň, V., Sims, A., Skudnik, M., Solti, G., Tomter, S.M., Twomey, M., Westerlund, B., Zell, J., 2019. Harmonisation of stem volume estimates in European National Forest Inventories. Annals of Forest Science 76, 1–23. doi:10.1007/s13595-019-0800-8.

[83] Gschwantner, T., Alberdi, I., Bauwens, S., Bender, S., Borota, D., Bosela, M., Bouriaud, O., Breidenbach, J., Donis, J., Fischer, C., Gasparini, P., Heffernan, L., Hervé, J.C., Kolozs, L., Korhonen, K.T., Koutsias, N.,

Kovácsevics, P., Kučera, M., Kulbokas, G., Kuliešis, A., Lanz, A., Lejeune, P., Lind, T., Marin, G., Morneau, F., Nord-Larsen, T., Nunes, L., Pantić, D., Redmond, J., Rego, F.C., Riedel, T., Šebeň, V., Sims, A., Skudnik, M., Tomter, S.M., 2022. Growing stock monitoring by European National Forest Inventories: Historical origins, current methods and harmonisation. Forest Ecology and Management 505, 119868. doi:10.1016/j.foreco.2021.119868.

[84] Guerra, J., Botequim, B., Buján, S., Jurado-Varela, A., Molina-Valero, J., Martínez-Calvo, A., Pérez Cruzado, C., 2022. Interpreting the uncertainty of model-based and design-based estimation in downscaling estimates from NFI data: A case-study in Extremadura (Spain). GIScience & Remote Sensing 59, 686–704. doi:10.1080/15481603.2022.2051383.

[85] Guerra-Hernández, J., Pascual, A., 2021. Using GEDI lidar data and airborne laser scanning to assess height growth dynamics in fast-growing species: A showcase in Spain. Forest Ecosystems 8, 14. doi:10/gpfhdx.

[86] Guldin, R.W., 2021. A Systematic Review of Small Domain Estimation Research in Forestry During the Twenty-First Century From Outside the United States. Frontiers in Forests and Global Change 4. doi:10.3389/ffgc.2021.695929.

[87] Guo, Q., Du, S., Jiang, J., Guo, W., Zhao, H., Yan, X., Zhao, Y., Xiao, W., 2023. Combining GEDI and sentinel data to estimate forest canopy mean height and aboveground biomass. Ecological Informatics 78, 102348. doi:10.1016/j.ecoinf.2023.102348.

[88] Gupta, R., Sharma, L., 2022. Mixed tropical forests canopy height mapping from spaceborne LiDAR GEDI and multisensor imagery using machine learning models. Remote Sensing Applications: Society and Environment 27, 100817. doi:10.1016/j.rsase.2022.100817.

[89] Guyon, D., Laventure, S., Belouard, T., Samalens, J.C., Wigneron, J.P., 2015. Retrieving the stand age from a retrospective detection of multinannual forest changes using Landsat data. Application on the heavily managed maritime pine forest in Southwestern France from a 30-year Landsat time-series (1984–2014), in: 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), IEEE, Milan, Italy. pp. 1968–1971. doi:10.1109/IGARSS.2015.7326182.

[90] Haakana, H., Heikkinen, J., Katila, M., Kangas, A., 2019. Efficiency of post-stratification for a large-scale forest inventory—case Finnish NFI. Annals of Forest Science 76. doi:10.1007/s13595-018-0795-6.

[91] Hagolle, O., Morin, D., Kadiri, M., 2018. Detailed Processing Model for the Weighted Average Synthesis Processor (WASP) for Sentinel-2 URL: https://doi.org/10.5281/zenodo.1401360, doi:10.5281/zenodo.1401360.

[92] Halley Des Fontaines, S., 2008. Grenelle de l'environnement et Assises de la forêt. Plan d'actions pour la forêt. Revue forestière française 60, 7–12. doi:10.4267/2042/17237.

[93] Hammond, W.M., Williams, A.P., Abatzoglou, J.T., Adams, H.D., Klein, T., López, R., Sáenz-Romero, C., Hartmann, H., Breshears, D.D., Allen, C.D., 2022. Global field observations of tree die-off reveal hotter-drought fingerprint for Earth's forests. Nature Communications 13, 1761. doi:10.1038/s41467-022-29289-2.

[94] Hancock, S., Armston, J., Hofton, M., Sun, X., Tang, H., Duncanson, L.I., Kellner, J.R., Dubayah, R., 2019. The GEDIsimulator: A large-footprint waveform lidar simulator for calibration and validation of spaceborne missions. Earth and Space Science 6, 294–310. doi:10.1029/2018EA000506.

[95] Harris, N.L., Gibbs, D.A., Baccini, A., Birdsey, R.A., De Bruin, S., Farina, M., Fatoyinbo, L., Hansen, M.C., Herold, M., Houghton, R.A., Potapov, P.V., Suarez, D.R., Roman-Cuesta, R.M., Saatchi, S.S., Slay, C.M., Turubanova, S.A., Tyukavina, A., 2021. Global maps of twenty-first century forest carbon fluxes. Nature Climate Change 11, 234–240. doi:10.1038/s41558-020-00976-6.

[96] Healey, S.P., Patterson, P.L., Saatchi, S., Lefsky, M.A., Lister, A.J., Freeman, E.A., 2012. A sample design for globally consistent biomass estimation using lidar data from the Geoscience Laser Altimeter System (GLAS). Carbon Balance and Management 7, 10. doi:10.1186/1750-0680-7-10.

[97] Henrich, V., Krauss, G., Götze, C., Sandow, C., 2012. Idb - www.indexdatabase.de, entwicklung einer datenbank für fernerkundungsindizes. URL: https://www.indexdatabase.de/info/credits.php.

[98] Hervé, J.C., Morneau, F., Véga, C., Leban, J.M., Saint-André, L., Bontemps, J.D., 2017. Evaluation des ressources forestières pour la bioéconomie : quels nouveaux besoins et comment y répondre ? Innovations Agronomiques 56, 71–80. URL: https://hal.science/hal-01608001, doi:10.15454/1.5137802056816558E12.

[99] Hervé, J.C., 2016. France, in: Vidal, C., Alberdi, I.A., Hernández Mateo, L., Redmond, J.J. (Eds.), National Forest Inventories: Assessment of Wood Availability and Use. Springer International Publishing, Cham, pp. 385–404. doi:10.1007/978-3-319-44015-6_1.

[100] Hervé, J.C., Wurpillot, S., Vidal, C., Roman-Amat, B., 2014. L'inventaire des ressources forestières en France : un nouveau regard sur de nouvelles forêts. Revue Forestière Française doi:10.4267/2042/56055.

[101] Heung, B., Bakker, L., Schmidt, M., Dragicevic, S., 2013. Modelling the dynamics of soil redistribution induced by sheet erosion using the Universal Soil Loss Equation and cellular automata. Geoderma doi:10.1016/j.geoderma.2013.03.019.

[102] Hidiroglou, M., 2001. Double sampling. Survey methodology 27, 143–154. URL: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2001002/article/6091-eng.pdf?st=i_Ou2kzo.

[103] Hill, A., Mandallaz, D., Langshausen, J., 2018. A Double-Sampling Extension of the German National Forest Inventory for Design-Based Small Area Estimation on Forest District Levels. Remote Sensing 10, 1052. doi:10.3390/rs10071052.

[104] Ho, T.K., 1998. Nearest neighbors in random subspaces, in: Amin, A., Dori, D., Pudil, P., Freeman, H. (Eds.), Advances in Pattern Recognition, Springer, Berlin, Heidelberg. pp. 640–648. doi:10.1007/BFb0033288.

[105] Hofton, M., Blair, J.B., 2019. Algorithm Theoretical Basis Document (ATBD) for GEDI Transmit and Receive Waveform Processing for L1 and L2 Products .

[106] Holmström, H., Fransson, J., 2003. Combining Remotely Sensed Optical and Radar Data in kNN-Estimation of Forest Variables. Forest Science 49, 409–418. doi:10.1093/forestscience/49.3.409.

[107] Hsu, H., Lachenbruch, P.A., 2014. Paired t Test, in: Wiley StatsRef: Statistics Reference Online. John Wiley & Sons, Ltd. doi:10.1002/9781118445112.stat05929.

[108] Hunt Jr., E.R., Daughtry, C.S.T., Eitel, J.U.H., Long, D.S., 2011. Remote sensing leaf chlorophyll content using a visible band index. Agronomy Journal 103, 1090–1099. doi:10.2134/agronj2010.0395.

[109] IGN, 2023a. Mémento 2023. URL: https://inventaire-forestier.ign.fr/IMG/pdf/memento_2023.pdf. (accessed Dec. 11, 2023).

[110] IGN, 2023b. Méthodologie : Résultats d'inventataire forestier – pour bien comprendre les résultats publiés. URL: https://inventaire-forestier.ign.fr/IMG/pdf/methodologie-2023.pdf. (accessed Dec. 11, 2023).

[111] IGN, BD Forêt version 2. URL: https://geoservices.ign.fr/bdforet. (accessed Jun. 21, 2023).

[112] IGN, IFN website. URL: https://inventaire-forestier.ign.fr. (accessed Nov. 13, 2023).

[113] IGN, RAF18. URL: https://geodesie.ign.fr/index.php?page=grilles. (accessed Mar. 06, 2022).

[114] IGN, RGE ALTI. https://geoservices.ign.fr/rgealti. (accessed Mar. 06, 2022).

[115] IGN, RGE ALTI version 2 Descriptif de contenu. URL: https://geoservices.ign.fr/sites/default/files/2021-07/DC_RGEALTI_2-0.pdf. (accessed Mar. 06, 2022).

[116] IGN, Sylvoécorégion. URL: https://inventaire-forestier.ign.fr/spip.php?article773. (accessed Jun. 21, 2023).

[117] Ilangakoon, N., Glenn, N., Schneider, F.D., Dashti, H., Hancock, S., Spaete, L., Goulden, T., 2021. Airborne and Spaceborne Lidar Reveal Trends and Patterns of Functional Diversity in a Semi-Arid Ecosystem. Frontiers in Remote Sensing 2. doi:10/gnkn5j.

[118] Imai, T., Hayashi, M., Sakaizawa, D., Mitsuhashi, R., Kimura, T., 2019. An overview of vegetation lidar MOLI, in: AGU Fall Meeting Abstracts.

[119] IPCC (Intergovernmental Panel on Climate Change), 2023. Summary for Policymakers. In: Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, H. Lee and J. Romero (eds.)]. IPCC, Geneva, Switzerland doi:10.59327/IPCC/AR6-9789291691647.001.

[120] Irulappa Pillai Vijayakumar, D.B., Renaud, J.P., Morneau, F., Mcroberts, R., Vega, C., 2019. Increasing Precision for French Forest Inventory Estimates using the k-NN Technique with Optical and Photogrammetric Data and Model-Assisted Estimators. Remote Sensing 11, 991. doi:10.3390/rs11080991.

[121] IUFRO (International Union of Forest Research Organizations), 2018. Global fire challenges in a warming world. Occasional Paper No. 32. IUFRO, Vienna URL: https://www.iufro.org/uploads/media/op32.pdf.

[122] Jaime, L., Batllori, E., Lloret, F., 2023. Bark beetle outbreaks in coniferous forests: A review of climate change effects. European Journal of Forest Research , 1–17doi:10.1007/s10342-023-01623-3.

[123] Jordano, P., 2000. Fruits and frugivory, in: Seeds: The Ecology of Regeneration in Plant Communities, pp. 125–166. doi:10.1079/9780851994321.0125.

[124] Jutzi, B., Stilla, U., 2006. Precise range estimation on known surfaces by analysis of fullwaveform laser, in: Proceedings of Phtogrammetric Computer Vision PCV. doi:10.5445/IR/1000073787.

[125] Kacic, P., Hirner, A., Canova, E., 2021. Fusing Sentinel-1 and -2 to Model GEDI-Derived Vegetation Structure Characteristics in GEE for the Paraguayan Chaco. Remote Sensing 13, 5105. doi:10.3390/rs13245105.

[126] Kacic, P., Thonfeld, F., Gessner, U., Kuenzer, C., 2023. Forest Structure Characterization in Germany: Novel Products and Analysis Based on GEDI, Sentinel-1 and Sentinel-2 Data. Remote Sensing 15, 1969. doi:10.3390/rs15081969.

[127] Kaselimi, M., Voulodimos, A., Daskalopoulos, I., Doulamis, N., Doulamis, A., 2023. A Vision Transformer Model for Convolution-Free Multilabel Classification of Satellite Imagery in Deforestation Monitoring. IEEE Transactions on Neural Networks and Learning Systems 34, 3299–3307. doi:10.1109/TNNLS.2022.3144791.

[128] Kim, Y., Zyl, J., 2009. A Time-Series Approach to Estimate Soil Moisture Using Polarimetric Radar Data. Geoscience and Remote Sensing, IEEE Transactions on 47, 2519–2527. doi:10.1109/TGRS.2009.2014944.

[129] Kotivuori, E., Korhonen, L., Packalen, P., 2016. Nationwide airborne laser scanning based models for volume, biomass and dominant height in Finland. Silva Fennica 50. doi:10.14214/sf.1567.

[130] Kurz, W.A., Stinson, G., Rampley, G.J., Dymond, C.C., Neilson, E.T., 2008. Risk of natural disturbances makes future contribution of Canada's forests to the global carbon cycle highly uncertain. Proceedings of the National Academy of Sciences of the United States of America 105, 1551–1555. doi:10.1073/pnas.0708133105.

[131] Köhl, M., Magnussen, S., Marchetti, M., 2006. Sampling in forest surveys, in: Sampling Methods, Remote Sensing and GIS Multiresource Forest Inventory, Springer, Berlin, Heidelberg. pp. 71–196. doi:10.1007/978-3-540-32572-7-3.

[132] Labonne, S., Cordonnier, T., Kunstler, G., Fuhr, M., 2019. Forêts de montagne et changement climatique : impacts et adaptations. Sciences Eaux & Territoires Numéro 28, 38–43. doi:10.3917/set.028.0038.

[133] Lahssini, K., Baghdadi, N., le Maire, G., Fayad, I., 2022. Influence of GEDI Acquisition and Processing Parameters on Canopy Height Estimates over Tropical Forests. Remote Sensing 14, 6264. doi:10.3390/rs14246264.

[134] Lang, N., Jetz, W., Schindler, K., Wegner, J.D., 2023. A high-resolution canopy height model of the Earth. Nature Ecology & Evolution 7, 1778–1789. doi:10.1038/s41559-023-02206-6.

[135] Lang, N., Kalischek, N., Armston, J., Schindler, K., Dubayah, R., Wegner, J.D., 2022. Global canopy height regression and uncertainty estimation from GEDI LIDAR waveforms with deep ensembles. Remote Sensing of Environment 268, 112760. doi:10/gnkn4c, arXiv:2103.03975.

[136] Li, W., Niu, Z., Shang, R., Qin, Y., Wang, L., Chen, H., 2020. High-resolution mapping of forest canopy height using machine learning by coupling ICESat-2 LiDAR with Sentinel-1, Sentinel-2 and Landsat-8 data. International Journal of Applied Earth Observation and Geoinformation 92, 102163. doi:10.1016/j.jag.2020.102163.

[137] Lim, K., Treitz, P., Wulder, M., St-Onge, B., Flood, M., 2003. LIDAR remote sensing of forest structure. Progress in Physical Geography 27, 88–106. doi:10.1191/0309133303pp360ra.

[138] Lindgren, N., Olsson, H., Nyström, K., Nyström, M., Ståhl, G., 2021. Data assimilation of growing stock volume using a sequence of remote sensing data from different sensors. Canadian Journal of Remote Sensing 48, 127–143. doi:10.1080/07038992.2021.1988542.

[139] Lindsay, J.B., 2016. Whitebox GAT: A case study in geomorphometric analysis. Computers & Geosciences 95, 75–84. doi:10.1016/j.cageo.2016.07.003.

[140] Lister, A.J., Andersen, H., Frescino, T., Gatziolis, D., Healey, S., Heath, L.S., Liknes, G.C., McRoberts, R., Moisen, G.G., Nelson, M., Riemann, R., Schleeweis, K., Schroeder, T.A., Westfall, J., Wilson, B.T., 2020. Use of Remote Sensing Data to Improve the Efficiency of National Forest Inventories: A Case Study from the United States National Forest Inventory. Forests 11, 1364. doi:10.3390/f11121364.

[141] Liu, A., Cheng, X., Chen, Z., 2021. Performance evaluation of GEDI and ICESat-2 laser altimeter data for terrain and canopy height retrievals. Remote Sensing of Environment 264, 112571. doi:10/gkzw4v.

[142] Louhaichi, M., Borman, M., Johnson, D., 2001. Spatially Located Platform and Aerial Photography for Documentation of Grazing Impacts on Wheat. Geocarto International 16. doi:10.1080/10106040108542184.

[143] Louppe, G., Geurts, P., 2012. Ensembles on Random Patches, in: Flach, P.A., De Bie, T., Cristianini, N. (Eds.), Machine Learning and Knowledge Discovery in Databases, Springer, Berlin, Heidelberg. pp. 346–361. doi:10.1007/978-3-642-33460-3_28.

[144] LP DAAC, 2023. Nasa announces pause in gedi mission. URL: https://www.earthdata.nasa.gov/news/nasa-announces-pause-gedi-mission?. (accessed Mar. 27, 2023).

[145] Luthcke, S., Rebold, T., Thomas, T., Pennington, R., 2019. Algorithm Theoretical Basis Document (ATBD) for gedi waveform geolocation for l1 and l2 products. document version 1.0. Goddard Space Flight Center, Greenbelt,MD: NASA's Land Processes Distributed Active Archive Center (LP DAAC) URL: https://lpdaac.usgs.gov/documents/579/GEDI__WFGEO_ATBD_v1.0.pdf.

[146] Magnussen, S., 2015. Arguments for a model-dependent inference? Forestry: An International Journal of Forest Research 88, 317–325. doi:10.1093/forestry/cpv002.

[147] Magnussen, S., Nord-Larsen, T., Riis-Nielsen, T., 2018. Lidar supported estimators of wood volume and aboveground biomass from the Danish national forest inventory (2012–2016). Remote Sensing of Environment 211, 146–153. doi:10.1016/j.rse.2018.04.015.

[148] Malambo, L., Popescu, S., Liu, M., 2023. Landsat-Scale Regional Forest Canopy Height Mapping Using ICESat-2 Along-Track Heights: Case Study of Eastern Texas. Remote Sensing 15, 1. doi:10.3390/rs15010001.

[149] Maltamo, M., Packalen, P., 2014. Species-Specific Management Inventory in Finland, in: Maltamo, M., Næsset, E., Vauhkonen, J. (Eds.), Forestry Applications of Airborne Laser Scanning: Concepts and Case Studies. Springer Netherlands, Dordrecht. Managing Forest Ecosystems, pp. 241–252. doi:10.1007/978-94-017-8663-8_12.

[150] Marselis, S.M., Tang, H., Armston, J., Abernethy, K., Alonso, A., Barbier, N., Bissiengou, P., Jeffery, K., Kenfack, D., Labrière, N., Lee, S.K., Lewis, S.L., Memiaghe, H., Poulsen, J.R., White, L., Dubayah, R., 2019. Exploring the relation between remotely sensed vertical canopy structure and tree species diversity in Gabon. Environmental Research Letters 14, 094013. doi:10.1088/1748-9326/ab2dcd.

[151] Mauro, F., Monleon, V.J., Temesgen, H., Ford, K.R., 2017. Analysis of area level and unit level models for small area estimation in forest inventories assisted with LiDAR auxiliary information. PLOS ONE 12, e0189401. doi:10.1371/journal.pone.0189401.

[152] McRoberts, R.E., 2011. Satellite image-based maps: Scientific inference or pretty pictures? Remote Sensing of Environment 115, 715–724. doi:10.1016/j.rse.2010.10.013.

[153] McRoberts, R.E., Chen, Q., Domke, G.M., Ståhl, G., Saarela, S., Westfall, J.A., 2016. Hybrid estimators for mean aboveground carbon per unit area. Forest Ecology and Management 378, 44–56. doi:10.1016/j.foreco.2016.07.007.

[154] McRoberts, R.E., Gobakken, T., Næsset, E., 2012. Post-stratified estimation of forest area and growing stock volume using lidar-based stratifications. Remote Sensing of Environment 125, 157–166. doi:10.1016/j.rse.2012.07.002.

[155] McRoberts, R.E., Næsset, E., Gobakken, T., 2013. Inference for lidar-assisted estimation of forest growing stock volume. Remote Sensing of Environment 128, 268–275. doi:10.1016/j.rse.2012.10.007.

[156] McRoberts, R.E., Tomppo, E.O., Finley, A.O., Heikkinen, J., 2007. Estimating areal means and variances of forest attributes using the k-Nearest Neighbors technique and satellite imagery. Remote Sensing of Environment 111, 466–480. doi:10.1016/j.rse.2007.04.002.

[157] Milenković, M., Schnell, S., Holmgren, J., Ressl, C., Lindberg, E., Hollaus, M., Pfeifer, N., Olsson, H., 2017. Influence of footprint size and geolocation error on the precision of forest biomass estimates from space-borne waveform LiDAR. Remote Sensing of Environment 200, 74–88. doi:10/gb4qsb.

[158] Ministère de la Transition Écologique et Solidaire, 2018. Stratégie nationale de lutte contre la déforestation importée (2018-2030) URL: https://www.deforestationimportee.ecologie.gouv.fr/la-sndi/article/sndi.

[159] Ministère de la Transition Écologique et Solidaire, 2020. National low carbon strategy : The ecological and inclusive transition towards carbon neutrality URL: https://www.ecologie.gouv.fr/sites/default/files/en_SNBC-2_summary.pdf.

[160] Ministère de l'agriculture et de l'alimentation, 2017. Programme national de la forêt et du bois 2016-2026 URL: https://agriculture.gouv.fr/le-programme-national-de-la-foret-et-du-bois-2016-2026.

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

[162] Moser, P., Vibrans, A.C., McRoberts, R.E., Næsset, E., Gobakken, T., Chirici, G., Mura, M., Marchetti, M., 2017. Methods for variable selection in LiDAR-assisted forest inventories. Forestry: An International Journal of Forest Research 90, 112–124. doi:10.1093/forestry/cpw041.

[163] Næsset, E., 2004. Accuracy of forest inventory using airborne laser scanning: Evaluating the first nordic full-scale operational project. Scandinavian Journal of Forest Research 19, 554–557. doi:10.1080/02827580410019544.

[164] Nasirzadehdizaji, R., Balik Sanli, F., Abdikan, S., Cakir, Z., Sekertekin, A., Ustuner, M., 2019. Sensitivity Analysis of Multi-Temporal Sentinel-1 SAR Parameters to Crop Height and Canopy Coverage. Applied Sciences 9, 655. doi:10.3390/app9040655.

[165] Nelson, E., 1994. An examination of anticipated g-jitter on Space Station and its effects on materials processes. URL: https://ntrs.nasa.gov/citations/19950006290.

[166] Neuenschwander, A., Guenther, E., White, J.C., Duncanson, L., Montesano, P., 2020. Validation of ICESat-2 terrain and canopy heights in boreal forests. Remote Sensing of Environment 251, 112110. doi:10.1016/j.rse.2020.112110.

[167] Neuenschwander, A.L., Magruder, L.A., 2019. Canopy and Terrain Height Retrievals with ICESat-2: A First Look. Remote Sensing 11, 1721. doi:10.3390/rs11141721.

[168] Neyman, J., 1934. On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. Journal of the Royal Statistical Society 97, 558–625. doi:10.2307/2342192, arXiv:2342192.

[169] Ni, W., Zhang, Z., Sun, G., 2021. Assessment of Slope-Adaptive Metrics of GEDI Waveforms for Estimations of Forest Aboveground Biomass over Mountainous Areas. Journal of Remote Sensing 2021, 1–17. doi:10/gpcfzb.

[170] Nilsson, M., Nordkvist, K., Jonzén, J., Lindgren, N., Axensten, P., Wallerman, J., Egberth, M., Larsson, S., Nilsson, L., Eriksson, J., Olsson, H., 2017. A nationwide forest attribute map of Sweden predicted using airborne laser scanning data and field data from the National Forest Inventory. Remote Sensing of Environment 194, 447–454. doi:10.1016/j.rse.2016.10.022.

[171] Osborne, C., 1991. Statistical Calibration: A Review. International Statistical Review / Revue Internationale de Statistique 59, 309–336. doi:10.2307/1403690, arXiv:1403690.

[172] Pacheco, A.d.P., Junior, J.A.d.S., Ruiz-Armenteros, A.M., Henriques, R.F.F., 2021. Assessment of k-Nearest Neighbor and Random Forest Classifiers for Mapping Forest Fire Areas in Central Portugal Using Landsat-8, Sentinel-2, and Terra Imagery. Remote Sensing 13, 1345. doi:10.3390/rs13071345.

[173] Pan, Y., Birdsey, R.A., Fang, J., Houghton, R., Kauppi, P.E., Kurz, W.A., Phillips, O.L., Shvidenko, A., Lewis, S.L., Canadell, J.G., Ciais, P., Jackson, R.B., Pacala, S.W., McGuire, A.D., Piao, S., Rautiainen, A., Sitch, S., Hayes, D., 2011. A Large and Persistent Carbon Sink in the World's Forests. Science 333, 988–993. doi:10.1126/science.1201609.

[174] Pardé, J., 1965. A useful concept: dominant height of forest stands. Revue forestière française 8, 850–856. doi:10.4267/2042/27262.

[175] Pathak, M., Slade, R., Shukla, P., Skea, J., Pichs-Madruga, R., Ürge Vorsatz, D., 2022. Technical summary. in: Climate change 2022: Mitigation of climate change. contribution of working group iii to the sixth assessment report of the intergovernmental panel on climate change [p.r. shukla, j. skea, r. slade, a. al khourdajie, r. van diemen, d. mccollum, m. pathak, s. some, p. vyas, r. fradera, m. belkacemi, a. hasija, g. lisboa, s. luz, j. malley, (eds.)]. Cambridge University Press, Cambridge, UK and New York, NY, USA doi:10.1017/9781009157926.002.

[176] Patterson, P.L., Healey, S.P., Staahl, G., Saarela, S., Holm, S., Andersen, H.E., Dubayah, R.O., Duncanson, L., Hancock, S., Armston, J., Kellner, J.R., Cohen, W.B., Yang, Z., 2019. Statistical properties of hybrid estimators proposed for GEDI—NASA's Global Ecosystem Dynamics Investigation. Environmental Research Letters 14, 065007. doi:10.1088/1748-9326/ab18df.

[177] Pereira-Pires, J.E., Mora, A., Aubard, V., Silva, J.M.N., Fonseca, J.M., 2021. Assessment of Sentinel-2 Spectral Features to Estimate Forest Height with the New GEDI Data, in: Camarinha-Matos, L.M., Ferreira, P., Brito, G. (Eds.), Technological Innovation for Applied AI Systems, Springer International Publishing, Cham. pp. 123–131. doi:10.1007/978-3-030-78288-7_12.

[178] Potapov, P., Li, X., Hernandez-Serna, A., Tyukavina, A., Hansen, M.C., Kommareddy, A., Pickens, A., Turubanova, S., Tang, H., Silva, C.E., Armston, J., Dubayah, R., Blair, J.B., Hofton, M., 2021. Mapping global forest canopy height through integration of GEDI and Landsat data. Remote Sensing of Environment doi:10.1016/j.rse.2020.112165.

[179] Prăvălie, R., 2018. Major perturbations in the Earth's forest ecosystems. Possible implications for global warming. Earth-Science Reviews 185, 544–571. doi:10.1016/j.earscirev.2018.06.010.

[180] Pronk, M., Eleveld, M., Ledoux, H., 2023. Assessing vertical accuracy and spatial coverage of ICESat-2 and GEDI spaceborne lidar for creating global terrain models doi:10.31223/X5309R.

[181] Puliti, S., Breidenbach, J., Schumacher, J., Hauglin, M., Klingenberg, T.F., Astrup, R., 2021. Above-ground biomass change estimation using national forest inventory data with Sentinel-2 and Landsat. Remote Sensing of Environment 265, 112644. doi:10.1016/j.rse.2021.112644.

[182] Pulkkinen, M., Ginzler, C., Traub, B., Lanz, A., 2018. Stereo-imagery-based post-stratification by regression-tree modelling in Swiss National Forest Inventory. Remote Sensing of Environment 213, 182–194. doi:10.1016/j.rse.2018.04.052.

[183] Qi, J., Chehbouni, A., Huete, A.R., Kerr, Y.H., Sorooshian, S., 1994. A modified soil adjusted vegetation index. Remote Sensing of Environment 48, 119–126. doi:10.1016/0034-4257(94)90134-1.

[184] Qi, W., Saarela, S., Armston, J., Ståhl, G., Dubayah, R., 2019. Forest biomass estimation over three distinct forest types using TanDEM-X InSAR data and simulated GEDI lidar data. Remote Sensing of Environment 232, 111283. doi:10.1016/j.rse.2019.111283.

[185] Quinn, P., Beven, K., Chevallier, P., Planchon, O., 1991. The prediction of hillslope flow paths for distributed hydrological modelling using digital terrain models. Hydrological Processes 5, 59–79. doi:10.1002/hyp.3360050106.

[186] Quirós, E., Polo, M.E., Fragoso-Campón, L., 2021. GEDI Elevation Accuracy Assessment: A Case Study of Southwest Spain 14, 5285–5299. doi:10.1109/JSTARS.2021.3080711.

[187] Remeš, J., Pulkrab, K., Bílek, L., Podrázský, V., 2020. Economic and Production Effect of Tree Species Change as a Result of Adaptation to Climate Change. Forests 11, 431. doi:10.3390/f11040431.

[188] Renaud, J.P., Sagar, A., Barbillon, P., Bouriaud, O., Deleuze, C., Vega, C., 2022. Characterizing the calibration domain of remote sensing models using convex hulls 112, 102939. doi:10.1016/j.jag.2022.102939.

[189] Roberge, C., Wulff, S., Reese, H., Ståhl, G., 2016. Improving the precision of sample-based forest damage inventories through two-phase sampling and post-stratification using remotely sensed auxiliary information. Environmental Monitoring and Assessment 188. doi:10.1007/s10661-016-5208-4.

[190] Rockström, J., Steffen, W., Noone, K., Persson, Å., Chapin III, F.S., Lambin, E., Lenton, T., Scheffer, M., Folke, C., Schellnhuber, H., Nykvist, B., de Wit, C., Hughes, T., Van der Leeuw, S., Rodhe, H., Sörlin, S., Snyder, P., Costanza, R., Svedin, U., Foley, J., 2009. A safe operating space for humanity. Nature 46, 472–475. doi:10.1038/461472a.

[191] Roncat, A., Morsdorf, F., Briese, C., Wagner, W., Pfeifer, N., 2014. Laser Pulse Interaction with Forest Canopy: Geometric and Radiometric Issues, in: For. Appl. Airborne Laser Scanning Concepts Case Stud.. volume 27, pp. 19–41. doi:10.1007/978-94-017-8663-8_2.

[192] Rouse, J., Haas, R.H., Schell, J.A., Deering, D., 1973. Monitoring vegetation systems in the great plains with ERTS.

[193] Roy, D.P., Kashongwe, H.B., Armston, J., 2021. The impact of geolocation uncertainty on GEDI tropical forest canopy height estimation and change monitoring. Science of Remote Sensing 4, 100024. doi:10/gktzv5.

[194] Rupnik, E., Daakir, M., Pierrot Deseilligny, M., 2017. MicMac – a free, open-source solution for photogrammetry. Open Geospatial Data, Software and Standards 2, 14. doi:10.1186/s40965-017-0027-2.

[195] Saarela, S., Grafström, A., Ståhl, G., Kangas, A., Holopainen, M., Tuominen, S., Nordkvist, K., Hyyppä, J., 2015. Model-assisted estimation of growing stock volume using different combinations of LiDAR and Landsat data as auxiliary information. Remote Sensing of Environment 158, 431–440. doi:10.1016/j.rse.2014.11.020.

[196] Saarela, S., Holm, S., Healey, S.P., Andersen, H.E., Petersson, H., Prentius, W., Patterson, P.L., Naesset, E., Gregoire, T.G., Ståhl, G., 2018. Generalized hierarchical model-based estimation for aboveground biomass assessment using gedi and Landsat data. Remote Sensing 10, 1832. doi:10.3390/rs10111832.

[197] Saarela, S., Holm, S., Healey, S.P., Patterson, P.L., Yang, Z., Andersen, H.E., Dubayah, R.O., Qi, W., Duncanson, L.I., Armston, J.D., Gobakken, T., Næsset, E., Ekström, M., Ståhl, G., 2022. Comparing frameworks for biomass prediction for the Global Ecosystem Dynamics Investigation. Remote Sensing of Environment 278, 113074. doi:10.1016/j.rse.2022.113074.

[198] Saarela, S., Schnell, S., Tuominen, S., Balazs, A., Hyyppä, J., Grafström, A., Stahl, G., 2016. Effects of positional errors in model-assisted and model-based estimation of growing stock volume. Remote Sensing of Environment 172, 101–108. doi:10.1016/j.rse.2015.11.002.

[199] Saborowski, J., Marx, A., Nagel, J., Böckmann, T., 2010. Double sampling for stratification in periodic inventories—infinite population approach. Forest ecology and management 260, 1886–1895. doi:10.1016/j.foreco.2010.08.035.

[200] Sagar, A., 2023. Multisource Forest Inventory: a Generic and Flexible Tool for Forest Resource Estimation and Mapping at a Fine Scale. PhD thesis. Institut National de l'Information Géographique et Forestière ; Laboratoire d'Inventaire Forestier.

[201] Sagar, A., Vega, C., Bouriaud, O., Piedallu, C., Renaud, J.P., 2022. Multisource forest inventories: A model-based approach using k-NN to reconcile forest attributes statistics and map products. ISPRS Journal of Photogrammetry and Remote Sensing 192, 175–188. doi:10.1016/j.isprsjprs.2022.08.016.

[202] San-Miguel-Ayanz, J., Durrant, T., Boca, R., Maianti, P., Liberta', G., Jacome Felix Oom, D., Branco, A., De Rigo, D., Suarez-Moreno, M., Ferrari, D., Roglia, E., Scionti, N., Broglia, M., Onida, M., Tistan, A., Loffler, P., 2023. Forest fires in europe, middle east and north africa 2022. Publications Office of the European Union, Luxembourg, doi:10.2760/871593.

[203] Sanchez-Lopez, N., Boschetti, L., Hudak, A.T., Hancock, S., Duncanson, L.I., 2020. Estimating Time Since the Last Stand-Replacing Disturbance (TSD) from Spaceborne Simulated GEDI Data: A Feasibility Study. Remote Sensing 12, 3506. doi:10.3390/rs12213506.

[204] Särndal, C.E., Swensson, B., Wretman, J., 1992. Model Assisted Survey Sampling. Model Assisted Survey Sampling, Springer-Verlag Publishing, New York, NY, US. doi:10.1007/978-1-4612-4378-6.

[205] Schindewolf, M., Bornkampf, C., Schmidt, M.v.W.a.J., Schindewolf, M., Bornkampf, C., Schmidt, M.v.W.a.J., 2015. Simulation of Reservoir Siltation with a Process-based Soil Loss and Deposition Model, in: Effects of Sediment Transport on Hydraulic Structures, pp. 51–57. doi:10.5772/61576.

[206] Schleich, A., Bouriaud, O., Durrieu, S., Vega, C., 2024. Usefulness of gedi footprints as first-phase sample for forest inventories based on double sampling for post-stratification. in Review .

[207] Schleich, A., Durrieu, S., Soma, M., Vega, C., 2023a. GEDI footprints with corrected geolocation with DEM. URL: https://geoservices.ign.fr/bdforet, doi:10.57745/EJ4CI3. recherche Data Gouv.

[208] Schleich, A., Durrieu, S., Soma, M., Vega, C., 2023b. GeoGEDI code. URL: https://github.com/aschleich/GeoGEDI.

[209] Schleich, A., Durrieu, S., Soma, M., Vega, C., 2023c. Improving GEDI Footprint Geolocation Using a High Resolution Digital Elevation Model. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 16, 7718–7732. doi:10.1109/JSTARS.2023.3298991.

[210] Schleich, A., Soma, M., Bouriaud, O., Vega, C., Durrieu, S., 2022. Improving estimations of the French national forest inventory by stratification based on spaceborne LiDAR (GEDI), in: ForestSAT, Berlin, Germany.

[211] Schwartz, M., Ciais, P., Ottlé, C., De Truchis, A., Vega, C., Fayad, I., Brandt, M., Fensholt, R., Baghdadi, N., Morneau, F., Morin, D., Guyon, D., Dayau, S., Wigneron, J.P., 2022. High-resolution canopy height map in the Landes forest (France) based on GEDI, Sentinel-1, and Sentinel-2 data with a deep learning approach doi:10.48550/arXiv.2212.10265, arXiv:arXiv:2212.10265.

[212] Schwartz, M., Ciais, P., Truchis, A., Chave, J., Ottlé, C., Vega, C., Wigneron, J.P., Nicolas, M., Jouaber, S., Liu, S., Brandt, M., Fayad, I., 2023. FORMS: Forest Multiple Source height, wood volume, and biomass maps in France at 10 to 30 m resolution based on Sentinel-1, Sentinel-2, and GEDI data with a deep learning approach. Earth System Science Data doi:10.5194/essd-2023-196.

[213] Scott, C.T., Bechtold, W.A., Reams, G.A., Smith, W.D., Westfall, J.A., Hansen, M.H., Moisen, G.G., 2005. Sample-based estimators used by the forest inventory and analysis national information management system. Research Report. URL: https://www.fs.usda.gov/research/treesearch/20379.

[214] Seidl, R., Thom, D., Kautz, M., Martin-Benito, D., Peltoniemi, M., Vacchiano, G., Wild, J., Ascoli, D., Petr, M., Honkaniemi, J., Lexer, M.J., Trotsiuk, V., Mairota, P., Svoboda, M., Fabrika, M., Nagel, T.A., Reyer, C.P.O., 2017. Forest disturbances under climate change 7, 395–402. URL: https://www.nature.com/articles/nclimate3303, doi:10.1038/nclimate3303.

[215] Shendryk, Y., 2022. Fusing GEDI with earth observation data for large area aboveground biomass mapping. International Journal of Applied Earth Observation and Geoinformation 115, 103108. doi:10.1016/j.jag.2022.103108.

[216] Silva, C., Duncanson, L., Hancock, S., Neuenschwander, A., Thomas, N., Hofton, M., Simard, M., Marshak, C., Armston, J., Lutchke, S., Dubayah, R., 2021. Fusing simulated GEDI, ICESat-2 and NISAR data for regional aboveground biomass mapping. Remote Sensing of Environment 253, 112234. doi:10/ghpm88.

[217] Soma, M., Schleich, A., Vega, C., Durrieu, S., 2022. Qualification of ICESat-2 ground elevation and canopy height products: evaluating the potential of 100 m and 20 m resolution products for forest inventory in temperate heterogeneous forests, in: ForestSAT, Berlin, Germany.

[218] Sothe, C., Gonsamo, A., Lourenço, R.B., Kurz, W.A., Snider, J., 2022. Spatially Continuous Mapping of Forest Canopy Height in Canada by Combining GEDI and ICESat-2 with PALSAR and Sentinel. Remote Sensing 14, 5158. doi:10.3390/rs14205158.

[219] Ståhl, G., Gobakken, T., Saarela, S., Persson, H.J., Ekström, M., Healey, S.P., Yang, Z., Holmgren, J., Lindberg, E., Nyström, K., Papucci, E., Ulvdal, P., Ørka, H.O., Næsset, E., Hou, Z., Olsson, H., McRoberts,

R.E., 2024. Why ecosystem characteristics predicted from remotely sensed data are unbiased and biased at the same time – and how this affects applications. Forest Ecosystems 11, 100164. doi:10.1016/j.fecs.2023.100164.

[220] Ståhl, G., Holm, S., Gregoire, T.G., Gobakken, T., Næsset, E., Nelson, R., 2011. Model-based inference for biomass estimation in a LiDAR sample survey in Hedmark County, Norway 41. doi:10.1139/X10-161.

[221] Ståhl, G., Saarela, S., Schnell, S., Holm, S., Breidenbach, J., Healey, S.P., Patterson, P.L., Magnussen, S., Næsset, E., McRoberts, R.E., Gregoire, T.G., 2016. Use of models in large-area forest surveys: Comparing model-assisted, model-based and hybrid estimation. Forest Ecosystems 3, 5. doi:10.1186/s40663-016-0064-9.

[222] Stekhoven, D.J., Bühlmann, P., 2012. Missforest non-parametric missing value imputation for mixed-type data. Bioinformatics 28, 112–118. doi:10.1093/bioinformatics/btr597.

[223] Stevens Jr, D.L., Olsen, A.R., 2004. Spatially balanced sampling of natural resources. Journal of the American statistical Association 99, 262–278. doi:10.1198/016214504000000250.

[224] Tomppo, E., Gschwantner, T., Lawrence, M., Mcroberts, R., Godinho-Ferreira, P., 2010. National Forest Inventories: Pathways for Common Reporting. doi:10.1007/978-90-481-3233-1.

[225] Tomppo, E., Olsson, H., Ståhl, G., Nilsson, M., Hagner, O., Katila, M., 2008. Combining national forest inventory field plots and remote sensing data for forest databases. Remote Sensing of Environment 112, 1982–1999. doi:10.1016/j.rse.2007.03.032.

[226] Tong, Z.Y., Wu, L.Y., Feng, H.H., Zhang, M., Armbruster, W.S., Renner, S.S., Huang, S.Q., 2023. New calculations indicate that 90 % of flowering plant species are animal-pollinated. National Science Review 10. doi:10.1093/nsr/nwad219.

[227] UN (United Nations), Climate Change 'Biggest Threat Modern Humans Have Ever Faced'. URL: https://press.un.org/en/2021/sc14445.doc.htm. (accessed Feb. 03, 2024).

[228] Urbazaev, M., Hess, L., Hancock, S., Ometto, J., Thiel, C., Dubois, C., Adam, M., Schmullius, C., 2021. Accuracy Assessment of Terrain and Canopy Height Estimates from ICESat-2 and GEDI LiDAR Missions in Temperate and Tropical Forests: First Results.

[229] Urbazaev, M., Hess, L., Hancock, S., Sato, L., Ometto, J., Thiel, C., Dubois, C., Heckel, K., Urban, M., Adam, M., Schmullius, C., 2022. Assessment of terrain elevation estimates from ICESat-2 and GEDI spaceborne LiDAR missions across different land cover and forest types 6, 100067. doi:10.1016/j.srs.2022.100067.

[230] Véga, C., St-Onge, B., 2009. Mapping site index and age by linking a time series of canopy height models with growth curves. Forest Ecology and Management 257, 951–959. doi:10.1016/j.foreco.2008.10.029.

[231] Vidal, C., Alberdi, I.A., Hernández Mateo, L., Redmond, J.J. (Eds.), 2016a. National Forest Inventories. Springer International Publishing, Cham. doi:10.1007/978-3-319-44015-6.

[232] Vidal, C., Bélouard, T., Hervé, J.C., Robert, N., Wolsack, J., 2005. A new flexible forest inventory in France, 7th annual forest inventory and analysis symposium. pp. 3–6.

[233] Vidal, C., Sallnäs, O., Redmond, J., Alberdi, I., Barreiro, S., Hernández, L., Schadauer, K., 2016b. Introduction, in: Vidal, C., Alberdi, I.A., Hernández Mateo, L., Redmond, J.J. (Eds.), National Forest Inventories: Assessment of Wood Availability and Use. Springer International Publishing, Cham, pp. 1–23. doi:10.1007/978-3-319-44015-6_1.

[234] Vidal, M.M., Pires, M.M., Guimarães, P.R., 2013. Large vertebrates as the missing components of seed-dispersal networks. Biological Conservation 163, 42–48. doi:10.1016/j.biocon.2013.03.025.

[235] Vollrath, A., Mullissa, A., Reiche, J., 2020a. Angular-Based Radiometric Slope Correction for Sentinel-1 on Google Earth Engine. Remote Sensing 12, 1867. doi:10.3390/rs12111867.

[236] Vollrath, A., Mullissa, A., Reiche, J., 2020b. Angular-based radiometric slope correction of Sentinel-1 on Google Earth Engine. URL: https://github.com/ESA-PhiLab/radiometric-slope-correction. google Colab python script.

[237] Wang, C., Elmore, A.J., Numata, I., Cochrane, M.A., Shaogang, L., Huang, J., Zhao, Y., Li, Y., 2022. Factors affecting relative height and ground elevation estimations of GEDI among forest types across the conterminous USA. GIScience & Remote Sensing 59, 975–999. doi:10.1080/15481603.2022.2085354.

[238] Waser, L., Fischer, C., Wang, Z., Ginzler, C., 2015. Wall-to-Wall Forest Mapping Based on Digital Surface Models from Image-Based Point Clouds and a NFI Forest Definition. Forests 6, 4510–4528. doi:10.3390/f6124386.

[239] Weiss, M., Baret, F., Jay, S., 2016. S2ToolBox Level 2 products:LAI, FAPAR, FCOVER. URL: https://step.esa.int/docs/extra/ATBD_S2ToolBox_V2.0.pdf. (accessed Oct. 25, 2023).

[240] Westfall, J., Lister, A., Scott, C., Weber, T., 2019. Double sampling for post-stratification in forest inventory. European Journal of Forest Research 138. doi:10.1007/s10342-019-01171-9.

[241] Westfall, J.A., Lister, A.J., Coulston, J.W., McRoberts, R.E., 2021. Realized and potential efficiency for post-stratified estimation in a national forest inventory. Canadian Journal of Forest Research 51, 1450–1457. doi:10.1139/cjfr-2020-0379.

[242] Wu, Q., Brown, A., 2022. "whitebox': 'WhiteboxTools' R frontend",R package version 2.2.0. URL: https://CRAN.R-project.org/package=whitebox.

[243] Zhang, S., Vega, C., Deleuze, C., Durrieu, S., Barbillon, P., Bouriaud, O., Renaud, J.P., 2022. Modelling forest volume with small area estimation of forest inventory using GEDI footprints as auxiliary information. International Journal of Applied Earth Observation and Geoinformation 114. doi:10.1016/j.jag.2022.103072.

[244] Zhu, X.X., Tuia, D., Mou, L., Xia, G.S., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. IEEE Geoscience and Remote Sensing Magazine 5, 8–36. doi:10.1109/MGRS.2017.2762307.

[245] Zolkos, S., Goetz, S., Dubayah, R., 2013. A meta-analysis of terrestrial aboveground biomass estimation using lidar remote sensing. Remote Sensing of Environment 128, 289–298. doi:10.1016/j.rse.2012.10.017.

Acronyms

AGB Above Ground Biomass.

AGBD Above Ground Biomass Density.

ALS Aerial Laser Scanner.

CHM Canopy Height Model.

DEM Digital Elevation Model.

DSPS Double Sampling for Post-Stratification.

fAPAR fraction of Absorbed Photosynthetically Active Radiation.

fCover fraction of vegetation cover.

GEDI Global Ecosystem Dynamics Investigation. A full-waveform spaceborne lidar.

GEE Google Earth Engine.

GeoGEDI Method developed to improve georeferencing of GEDI footprints.

GLI Green Leaf Index.

GSV Growing Stock Volume.

IGN Institut National de l'information Géographique et forestière.

INRAE Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement.

ISS International Space Station.

kNN k-Nearest Neighbors.

LAI Leaf Area Index.

LIF Laboratoire d'Inventaire Forestier, IGN.

MFI Multisource Forest Inventory.

MSAVI Modified Soil Adjusted Vegetation Index.

NDVI Normalized Difference Vegetation Index.

NDWI Normalized Difference Water Index.

NFI National Forest Inventory.

NIR Near Infra Red.

PAI Plant Area Index.

PAVD Plant Area Volume Density.

RH Relative Height. Variable in GEDI data.

RVI Radar Vegetation Index.

S1 Sentinel-1.

S2 Sentinel-2.

SAE Small Area Estimation.

SAR Synthetic Aperture Radar.

SLIM Space Lidar for Improved Multisource Forest Inventory. Project encompassing the thesis.

SRS Simple Random Sampling.

TLS Terrestrial Laser Scanner.