DISSERTATION

LIDAR REMOTE SENSING OF SAVANNA BIOPHYSICAL ATTRIBUTES

Submitted by

David Gwenzi

Graduate Degree Program in Ecology

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Summer 2015

Doctoral Committee:

Advisor: Michael Andrew Lefsky

Stephen Leisz
Michael Ryan
Jason Sibold

ABSTRACT

LIDAR REMOTE SENSING OF SAVANNA BIOPHYSICAL ATTRIBUTES

Although savanna ecosystems cover approximately 20 % of the terrestrial land surface and can have productivity equal to some closed forests, their role in the global carbon cycle is poorly understood. Studies using Light Detection And Ranging (Lidar) have demonstrated the sensor's ability to measure canopy height, that is in turn strongly related to biophysical attributes such as aboveground carbon storage, but most of this work has focused on closed canopy forests. The sparse observation network in savannas means they remain one of the weak links in our understanding of the global carbon cycle. This study explored the applicability of a past spaceborne Lidar mission and the potential of future missions to estimate canopy height and carbon storage in these biomes.

The research used data from two Oak savannas in California, USA: the Tejon Ranch Conservancy in Kern County and the Tonzi Ranch in Santa Clara County. In the first paper we used non-parametric regression techniques to estimate canopy height from waveform parameters derived from the Ice Cloud and land Elevation Satellite's Geoscience Laser Altimeter System (ICESat-GLAS) data. Merely adopting the methods derived for forests did not produce adequate results but the modeling was significantly improved by incorporating canopy cover information and interaction terms to address the high structural heterogeneity inherent to savannas.

Paper 2 explored the relationship between canopy height and aboveground biomass. To accomplish this we developed generalized models using the classical least squares regression modeling approach to relate canopy height to above ground woody biomass and then employed Hierarchical Bayesian Analysis (HBA) to explore the implications of using generalized instead of species composition-specific models. Models that incorporated canopy cover proxies performed better than those that did not. Although the model parameters indicated interspecific variability, the distribution of the posterior densities of the differences between composition level and global level parameter values showed a high support for the use of global parameters, suggesting that these canopy height-biomass models are universally (large scale) applicable.

As the spatial coverage of spaceborne lidar will remain limited for the immediate future, our objective in paper 3 was to explore the best means of extrapolating plot level biomass into wall-to-wall maps that provide more ecological information. We evaluated the utility of three spatial modeling approaches to address this problem: deterministic methods, geostatistical methods and an image segmentation approach. Overall, the mean pixel biomass estimated by the 3 approaches did not differ significantly but the output maps showed marked differences in the estimation precision and ability of each model to mimic the primary variable's trend across the landscape. The results emphasized the need for future satellite lidar missions to consider increasing the sampling intensity across track so that biomass observations are made and characterized at the scale at which they vary.

With ICESat-GLAS having been decommissioned in 2010, the earliest planned spaceborne lidar mission is ICESat-2, which will use the Advanced Topography Laser Altimeter System (ATLAS) sensor, which uses a photon counting technique. In paper 4 we explore the capability of this mission for studying three dimensional vegetation structure in savannas. We used data from the Multiple Altimeter Beam Experimental Lidar (MABEL), an airborne photon counting lidar sensor developed by NASA Goddard to simulate ICESat-2 data. We segmented each transect into different block sizes and calculated canopy top and mean ground elevation based on the structure of the histogram of the block's aggregated photons. Our algorithm was able to compute canopy height and generate visually meaningful vegetation profiles at MABEL's signal and noise levels but a simulation of the expected performance of ICESat-2 by adjusting MABEL data's detected number of signal and noise photons to that predicted using ATLAS instrument model design cases indicated that signal photons will be substantially lower. The lower data resolution reduces canopy height estimation precision especially in areas of low density vegetation cover.

Given the clear difficulties in processing simulated ATLAS data, it appears unlikely that it will provide the kind of data required for mapping of the biophysical properties of savanna vegetation. Rather, resources are better concentrated on preparing for the Global Ecosystem Dynamics Investigation (GEDI) mission, a waveform lidar mission scheduled to launch by the end of this decade. In addition to the full waveform technique, GEDI will collect data from 25 m diameter contiguous footprints with a high across track density, a requirement that we identified as critically necessary in paper 3.

ACKNOWLEDGEMENTS

I would like to express my deep gratitude to Michael Andrew Lefsky for helping me get funding for my PhD and for being an excellent advisor. Special thanks go to NASA grant numbers NNX12AH19G, NNX12AK07G and NNH11ZDA001N-ICESAT2 for funding this research project. I also thank Jason Sibold, Stephen Leisz and Michael Ryan for serving on my committee. My appreciation also goes to the following people and organizations:

- The National Snow and Ice Data Center (NSIDC) provided the raw ICESat-GLAS data used in chapters 2 and 4
- The Center for Lidar Information Coordination and Knowledge (CLICK) and the United States Geological Survey (USGS) provided the lidar data used in paper 1 and the Landsat 5 TM and SRTM data used in paper 1,2 and 3
- Robin Reich (Colorado State University) introduced me to spatial statistics and provided spatial analysis packages to work with R
- Thompson Hobbs (Colorado State University) introduced me to Bayesian Inference and provided advice on models used in chapter 3
- Dr D., Baldocchi and Dr Q., Chen (University of California, Berkley) provided all the necessary background information and guidelines about the Tonzi Ranch study area
- The Tejon Conservancy authorities allowed me access to do field work in their conservancy
- Jason D. McClure assisted me with fieldwork.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

1. Background

Savanna ecosystems account for approximately 20% of the land area of terrestrial vegetation and are a highly productive ecosystem that plays an important role in the global carbon cycle (Lucas *et al*., 2011). In seasonally warm biomes, such as subtropical regions of Africa, savannas are undergoing rapid conversions to other land uses at rates exceeding 1% per year which is likely to contribute a significant flux of greenhouse gases, mainly CO_2 (Potter, 2011). Despite the increasingly acknowledged importance of these ecosystems, their carbon cycle is relatively understudied compared to other biomes (Williams et al., 2008). A lack of representative studies, coupled with a sparse observation network in these biomes, mean they remain one of the weak links in our understanding of the global carbon cycle (Bombelli et al., 2009; Williams et al., 2008). Some savannas like those in tropical regions are as productive as some closed forests, they are often rich in biodiversity and with optimal management they are relatively resilient to anthropogenic disturbance hence they remain a great hope for maintaining carbon sinks. Estimating structural attributes such as aboveground woody biomass is an important step towards reliable monitoring of the carbon pools in these ecosystems to help us better understand the global carbon cycle. Since conventional field based biomass assessment is tedious and time consuming (especially at large scales), remote sensing combined with limited ground truth data has been proposed to efficiently monitor terrestrial ecosystems at various temporal and spatial scales.

Because of the signal saturation problem, optical passive or radar sensors have proven to be more useful in describing the canopy's two dimensional aspects but inaccurate and at times inapplicable to the three dimensional aspects that explain biomass storage, particularly in complex, heterogeneous vegetation systems such as mixed deciduous woodlands (Patenaude *et al.* 2005). Light Detection and Ranging (Lidar) can accurately measure the three dimensional aspects of a vegetation canopy and have been shown to accurately estimate Leaf Area Index (LAI) and aboveground biomass even in those high biomass ecosystems where passive optical and active Radar sensors typically fail to do so (Boudreau *et al.*, 2008; Drake *et al.*, 2002; Lefsky *et al*., 1999 ; Lefsky *et al*., 2002; Patenaude *et al*., 2005) . The main disadvantage of airborne Lidar is its high cost, and its relatively low horizontal coverage, preventing it from global use. As such, attention has shifted to spaceborne Lidar applications (Lefsky, 2010, Hall *et al*., 2011).

Data products from multiple remote sensing techniques can be combined to improve the accuracy of modeling 3 dimensional vegetation biophysical parameters. Lidar measurements provide the most direct estimates of canopy height and the vertical structure of canopy foliage, thus enables ecologists to quantify the 3D distribution of vegetation and understand processes such as carbon accumulation and forest succession thereby improving the state of ecosystem models (Chambers *et al*., 2007). Radar backscatter enables direct measurements of live aboveground woody biomass (carbon stocks) and structural attributes such as volume, basal area and crown mass. Combining Lidar and Radar remote sensing data with moderate resolution multispectral remote sensing data such as Landsat and Moderate Resolution Imaging Spectroradiometer (MODIS) imagery has proven to be efficient in measuring and mapping canopy height and woody biomass at large scale landscapes either as a combination of variables in one model or the other sensors providing auxiliary variables to extend the plot level estimates of Lidar to landscape level (Chambers *et al*., 2007; Hudak *et al.*, 2002; Hyde *et al*., 2007; Saatchi *et al*., 2007; Slatton *et al*., 2001; Wulder & Seemann, 2003). These kinds of maps also provide information about how basin-wide gradients, such as total precipitation drive regional biomass distribution, which significantly improves our ability to estimate the carbon flux resulting from land-use change (Chambers *et al*, 2007).

So far, the only used spaceborne Lidar data has been the Geoscience Laser Altimeter System (GLAS) waveforms from the Ice Cloud and Land Elevation Satellite (ICESat). GLAS was developed by NASA-Goddard for the ICESat mission (Abshire *et al.*, 2005; Zwally *et al.*, 2005). The ICESat mission started in 2003 and officially ended in 2010. A number of biomass related studies have used its data but mostly in closed forests, with an emphasis on characterizing the forest canopy parameters and addressing the problem of slope at plot level (Chen, 2010; Lefsky, 2010; Lefsky *et al*., 2007; Lefsky *et al.*, 2005; Xing *et al*., 2010) and often using a combination of airborne and spaceborne data (Boudreau *et al*., 2008; Duncanson *et al*., 2010; Hilker *et al.*, 2010).

A second generation spaceborne Lidar platform (ICESat-2) is expected to launch in 2016. Building on the lessons from its predecessor, ICESat-2 is expected to provide observations with much greater spatial and temporal resolution, and accuracy through the use of a micro-pulse multi beam high repetition photon counting approach (Abdalati et al., 2010; NASA, 2011). The primary objective of ICESat-2 will be the quantification of ice sheets and sea ice but just like ICESat-GLAS, vegetation height retrieval for biomass assessment is a science objective, although not a mission requirement. With this approach, data will be available in smaller footprints (10m) with a dense along track sampling of about 70 cm which should minimize the problem of surface slope and sparse coverage encountered with ICESat-GLAS data. Simulation data is currently being made available via airborne sensors (MABEL and Sigma Space MPL) flown in selected areas of the United States.

Change detection is important for monitoring the relationships between land use dynamics, climate change and carbon storage and fluxes. For this purpose combined use of ICESat-1 and ICESat-2 data will enable monitoring of canopy height and aboveground biomass using spaceborne Lidar from as far back as 2003 into the future. The use of multi-temporal Lidar data has proven to be useful in monitoring a number of ecological processes such as species invasion (Rosso *et al.*, 2006), forest gap dynamics (Vepakomma *et al.*, 2008), forest disturbance (Dolan et al., 2011) and forest growth and Net Primary Production (NPP) / biomass dynamics (Lefsky *et al*., 2005; Wulder *et al*., 2007). Given the ground breaking capabilities of ICESat-GLAS, developing relationships between its data and any future spaceborne Lidar system will ensure a continuous monitoring of vegetation canopy and the associated ecological processes.

2. Problem identification and justification of this research

A demonstration of the role of vegetation in carbon sequestration requires reliable estimates of the actual amount of biomass stored by the vegetation system of interest. Since spaceborne Lidar is a sampling instrument, we also require reliable models that can extrapolate footprint level biomass to the landscape level. Derivation of the appropriate means of modeling biomass from ICESat-GLAS footprint data and linking it with future NASA missions such as ICESat-2 is an important step in ensuring continuous monitoring of carbon fluxes. When reliable methods are developed, public institutions and land managers can also easily participate in carbon quantification related projects such as Reducing Emissions from Deforestation and Degradation (REDD) or the joint implementation, Clean Development Mechanisms (CDM) and emissions trading under the Kyoto protocol.

Most Lidar related canopy height/biomass studies have been done in closed canopy forests (e.g boreal forests of North America and Eurasia, temperate forests of China and North America, and tropical forests of South America and Africa) but very little has been done in savannas. As a result the current global canopy cover maps such as the first ever by Lefsky (2010) focus on forests only and subsequent attempts have had inconsistencies mainly resulting from differences in vegetation systems boundaries. As an example Simard *et al.* (2011) reported that their map and that of Lefsky (2010) had large differences in vegetation cover, which they attributed partly to the fact that they mapped more open systems not covered by Lefsky (2010). Nonetheless, Simard *et al.* (2011) used a subjective slope bias correction method hence the issue of slope and open canopy cover was not adequately addressed. Savannas therefore have potential issues that serve as both advantageous and disadvantageous to monitor with Lidar sensors. The main disadvantage of low canopy cover is that it reduces the power of the canopy return and increases the potential contribution of sloped terrain. One advantage of such openness would be that it eliminates the problem of occluded understorey and terrain surface experienced in some dense forests. Exploring these factors in detail will help in developing correction factors for global use of spaceborne Lidar data.

3. Research Aim

Assess the applicability of spaceborne Lidar and auxiliary remote sensing data in estimating, mapping and monitoring canopy height and carbon storage in savanna woodlands.

4. Structure of this dissertation

This dissertation is a compilation of the manuscripts we developed to address the aforementioned issues. The research questions we sought to answer in each chapter are outlined below

Chapter 1: To what extent can slope effect be removed from Lidar waveforms to reliably estimate canopy height indices in savanna woodlands? How much does canopy cover data contribute in this effort?

Chapter 2: To what extent are the estimated canopy height indices correlated to footprint level aboveground woody biomass? How does this compare with efforts reported in other biomes? How do generalized models compare with species class specific models?

Chapter 3: How reliably can the non-contiguous footprints of Lidar be used in combination with Radar, passive remote sensing imagery and relevant auxiliary data to extrapolate and map the spatial distribution of the biomass at landscape level?

Chapter 4: What are the prospects of a photon counting lidar approach, as simulated by an airborne sensor in estimating canopy height in savannas?

CHAPTER 1 REFERENCES

Abdalati, W., Zwally, H. J., Bindschadler, R., Csatho, B., Farrell, S. L., Fricker, H. A., … Webb, C. (2010). The ICESat-2 Laser Altimetry Mission. *Proceedings of the IEEE*, *98*(5), 735–751.

Abshire, J. B., Sun, X., Riris, H., Sirota, M. J., McGarry, J. F., Palm, S., … Liiva, P. (2005). Geoscience Laser Altimeter System (GLAS) on the ICESat mission: On-orbit measurement performance. *Geophysical Research Letters*, *32*(21), 1–4.

Bombelli, A., Henry, M., Castaldi, S., Arneth, A., Grandcourt, A. De, Grieco, E., … Cedex, M. (2009). An outlook on the Sub-Saharan Africa carbon balance. *Biogeosciences*, *6*(10), 2193–2205.

Boudreau, J., Nelson, R., Margolis, H., Beaudoin, A., Guindon, L., & Kimes, D. (2008). Regional aboveground forest biomass using airborne and spaceborne LiDAR in Québec. *Remote Sensing of Environment*, *112*(10), 3876–3890.

Chambers, J. Q., Asner, G. P., Morton, D. C., Anderson, L. O., Saatchi, S. S., Espírito-Santo, F. D. B., … Souza, C. (2007). Regional ecosystem structure and function: ecological insights from remote sensing of tropical forests. *Trends in Ecology & Evolution*, *22*(8), 414–23.

Chen, Q. (2010). Retrieving vegetation height of forests and woodlands over mountainous areas in the Pacific Coast region using satellite laser altimetry. *Remote Sensing of Environment*, *114*(7), 1610–1627.

Dolan, K. a., Hurtt, G. C., Chambers, J. Q., Dubayah, R. O., Frolking, S., & Masek, J. G. (2011). Using ICESat's Geoscience Laser Altimeter System (GLAS) to assess large-scale forest disturbance caused by hurricane Katrina. *Remote Sensing of Environment*, *115*(1), 86–96.

Drake, J. B., Dubayah, R. O., Clark, D. B., Knox, R. G., Blair, J. B., Hofton, M. A., … Prince, S. D. (2002). Estimation of tropical forest structural characteristics using large-footprint lidar. *Remote Sensing of Environment*, *79*(2-3), 305–319.

Duncanson, L. I., Niemann, K. O., & Wulder, M. A. (2010). Estimating forest canopy height and terrain relief from GLAS waveform metrics. *Remote Sensing of Environment*, *114*(1), 138–154.

Hilker, T., Leeuwen, M., Coops, N. C., Wulder, M. a., Newnham, G. J., Jupp, D. L. B., & Culvenor, D. S. (2010). Comparing canopy metrics derived from terrestrial and airborne laser scanning in a Douglas-fir dominated forest stand. *Trees*, *24*(5), 819–832.

Hudak, A. T., Lefsky, M. A., Cohen, W. B., & Berterretche, M. (2002). Integration of lidar and Landsat ETM+ data for estimating and mapping forest canopy height. *Remote Sensing of Environment*, *82*(2-3), 397–416.

Hyde, P., Nelson, R., Kimes, D., & Levine, E. (2007). Exploring LiDAR–RaDAR synergy—predicting aboveground biomass in a southwestern ponderosa pine forest using LiDAR, SAR and InSAR. *Remote Sensing of Environment*, *106*(1), 28–38. http://doi.org/10.1016/j.rse.2006.07.017

Lefsky, M. A. (2010). A global forest canopy height map from the Moderate Resolution Imaging Spectroradiometer and the Geoscience Laser Altimeter System. *Geophysical Research Letters, 37*(15), 1–5.

Lefsky, M. A., Cohen, W. B., Acker, S. A., Parker, G. G., Spies, T. A., & Harding, D. (1999). Lidar Remote Sensing of the Canopy Structure and Biophysical Properties of Douglas-Fir Western Hemlock Forests. *Remote Sensing of Environment*, *70*(3), 339–361.

Lefsky, M. A., Harding, D. J., Keller, M., Cohen, W. B., Carabajal, C. C., Del Bom Espirito-Santo, F., … de Oliveira Jr, R. (2005). Estimates of forest canopy height and aboveground biomass using ICESat. *Geophysical Research Letters*, *32*(22), 1–4. Retrieved from http://dx.doi.org/10.1029/2005GL023971

Lefsky, M. A., Keller, A. M., Pang, Y., Camargo, P. B. de, & Hunter, M. O. (2007). Revised method for forest canopy height estimation from Geoscience Laser Altimeter System waveforms. *Journal of Applied Remote Sensing*, *1*.

Lefsky, M. A., Turner, D. P., Guzy, M., & Cohen, W. B. (2005). Combining lidar estimates of aboveground biomass and Landsat estimates of stand age for spatially extensive validation of modeled forest productivity. *Remote Sensing of Environment*, *95*(4), 549–558.

Lucas, R. M., Lee, A. C., Amston, J., Carreiras, J. M. B., Viergever, K. M., Bunting, P., … Woodhouse, I. (2011). Quantifying carbon in Savannas: The role of active sensors in measurements of tree structure and biomass. In M. J. Hill & N. P. Hanan (Eds.), *Ecosystem function in Savannas*. Florida: Tylor and Francis Group.

Means, J. E., Acker, S. A., Harding, D. J., Blair, J. B., Lefsky, M. A., Cohen, W. B., … McKee, W. A. (1999). Use of Large-Footprint Scanning Airborne Lidar To Estimate Forest Stand Characteristics in the Western Cascades of Oregon. *Remote Sensing of Environment, 67*(3), 298–308.

NASA. (2011). ICESAT Home Page. Retrieved from http://icesat.gsfc.nasa.gov/

Patenaude, G., Milne, R., & Dawson, T. (2005). Synthesis of remote sensing approaches for forest carbon estimation: reporting to the Kyoto Protocol. *Environmental Science & Policy*, *8*(2), 161–178.

Potter, C. (2011). Carbon cycle and vegetation dynamics of Savannas based on Global satellite data products. In M. J. Hill & N. P. Hanan (Eds.), *Ecosystem function in Savannas: Measurement and Modeling at Landscape to Global Scales*. Florida: Tylor and Francis Group.

Rosso, P., Ustin, S., & Hastings, a. (2006). Use of lidar to study changes associated with Spartina invasion in San Francisco Bay marshes. *Remote Sensing of Environment*, *100*(3), 295–306. http://doi.org/10.1016/j.rse.2005.10.012

Saatchi, S. S., Houghton, R. a., Dos Santos Alvalá, R. C., Soares, J. V., & Yu, Y. (2007). Distribution of aboveground live biomass in the Amazon basin. *Global Change Biology*, *13*(4), 816–837. Retrieved from http://doi.wiley.com/10.1111/j.1365-2486.2007.01323.x

Simard, M., Pinto, N., Fisher, J. B., & Baccini, A. (2011). Mapping forest canopy height globally with spaceborne lidar. *Journal of Geophysical Research*, *116*(G04021), 1–12. Retrieved from http://www.agu.org/pubs/crossref/2011/2011JG001708.shtml

Slatton, K. C., Member, S., Crawford, M. M., Member, S., & Evans, B. L. (2001). Fusing Interferometric Radar and Laser Altimeter Data to Estimate Surface Topography and Vegetation Heights, *39*(11), 2470–2482.

Vepakomma, U., St-Onge, B., & Kneeshaw, D. (2008). Spatially explicit characterization of boreal forest gap dynamics using multi-temporal lidar data. *Remote Sensing of Environment*, *112*(5), 2326–2340. http://doi.org/10.1016/j.rse.2007.10.001

Williams, M., Ryan, C., Rees, R., Sambane, E., Fernando, J., & Grace, J. (2008). Carbon sequestration and biodiversity of re-growing miombo woodlands in Mozambique. *Forest Ecology and Management*, *254*(2), 145–155.

Wulder, M. a., Han, T., White, J. C., Sweda, T., & Tsuzuki, H. (2007). Integrating profiling LIDAR with Landsat data for regional boreal forest canopy attribute estimation and change characterization. *Remote Sensing of Environment*, *110*(1), 123–137. http://doi.org/10.1016/j.rse.2007.02.002

Wulder, M. A., & Seemann, D. (2003). Forest inventory height update through the integration of lidar data with segmented Landsat imagery. *Methods*, *29*(5), 536–543.

Xing, Y., de Gier, A., Zhang, J., & Wang, L. (2010). An improved method for estimating forest canopy height using ICESat-GLAS full waveform data over sloping terrain: A case study in Changbai mountains, China. *International Journal of Applied Earth Observation and Geoinformation*, *12*(5), 385–392.

Zwally, H. J., Shuman, C. A., Hancock, D., & DiMarzio, J. P. (2005). Overview of the ICESat Mission. *Geophysical Research Letters*, *32*.

CHAPTER 2: MODELING CANOPY HEIGHT IN A SAVANNA ECOSYSTEM USING SPACEBORNE LIDAR WAVEFORMS[1]

Synopsis

Although savanna ecosystems cover about 20% of the terrestrial land surface and can have productivity equal to some closed forests, their role in the global carbon cycle is poorly understood. As a result, these ecosystems are globally more important than generally appreciated in the earth observation and modeling communities. Remote sensing has been proposed as an efficient tool in assessing the physical structure of an ecosystem which in turn is closely related to its ecological functionality such as carbon storage. Studies using Light Detection and Ranging (lidar) have demonstrated the technology's ability to measure canopy height and the strong relationship between canopy height and structural attributes such as aboveground biomass, but most of this work has focused on closed canopy forests. This study explored the applicability of spaceborne lidar to estimate canopy height as a pre-requisite for aboveground biomass and carbon storage assessment in savannas. The research used a case study of the Oak Savannas of Santa Clara in California, USA. Discrete return airborne lidar data was used to extract height metrics in plots coincident with waveform data from the Ice Cloud and land Elevation Satellite (ICESat)'s Geoscience Laser Altimeter System (GLAS). Detailed analysis of GLAS waveforms was followed by non-parametric regression modeling to estimate maximum canopy height and

[1] David Gwenzi and Michael Andrew Lefsky, 2014
Remote sensing of Environment, Volume 154, 338-344

80th and 90th percentile vegetation heights. Existing methods were adapted with the inclusion of NDVI (as a canopy cover proxy) and interaction terms to increase utility in savanna ecosystems. Our main findings were that merely adopting the methods derived for forests would not produce adequate results. Maximum canopy height was estimated with better accuracy compared to percentile height metrics. The inclusion of NDVI and interaction terms improved maximum canopy height modeling much more than it did for the 80th and 90th percentile height modeling. Taller stands on flat terrain had the best results while shorter stands on steep terrain had the worst. Our work has demonstrated the capability of waveform lidar to assess vegetation structural attributes in savannas. The challenge in canopy height modeling using this technique in such ecosystems is not limited to terrain slope but also includes the interacting influence of low canopy cover and short height. As such, we need special models for savanna areas in an effort to do global assessments of terrestrial vegetation structure using lidar. For future studies we recommend a closer look at the non-significant influence of canopy cover on the percentile canopy height models especially its implication on the subsequent biomass modeling.

Key words: Savanna; canopy height; canopy cover; lidar; ICESat-GLAS

1. Introduction

Savannas account for approximately 20% of the land area of terrestrial vegetation and are highly productive ecosystems that play an important role in the global carbon cycle (Lucas *et al*., 2011). Globally, savannas and other open woodlands are experiencing changes in the balance between woody and herbaceous cover (Hill & Hanan, 2010). In seasonally warm biomes, such as the subtropical regions of Africa, they are undergoing rapid conversion to other land uses at rates exceeding 1% per year (Potter, 2011). Savannas and open canopy ecosystems in tropical regions and other dryland areas can be as productive as some closed forests (Rotenberg & Yakir, 2010; Schimel, 2010), they are often rich in biodiversity and can be major stores of carbon in woody biomass and soils (Scholes & Hall, 1996) hence they may be a good carbon sink.

Despite the increasingly acknowledged importance of these ecosystems, like open dryland forests the carbon cycle of savannas is relatively understudied in the earth observation and modeling communities compared to other biomes (Williams *et al*., 2008; Schimel, 2010; Hill and Hanan, 2010). This sparse observation network means their role in the climate system and feedbacks with the atmosphere are not well understood and they are a weak link in our understanding of the global carbon cycle (Bombelli *et al*., 2009; Hill & Hanan, 2010; Williams *et al*., 2008). Estimating structural attributes such as aboveground woody biomass is an important step towards reliable monitoring of the carbon pools in these ecosystems.

The carbon storage capacity and related ecological functionality of an ecosystem is largely represented by its physical structure (Wulder *et al*., 2004). Remote sensing combined with field data has been employed as a tool to measure and monitor the structure of terrestrial ecosystems at various temporal and spatial scales (Lefsky *et al*., 2002; Patenaude *et al*., 2005). Most remote sensing studies use empirical relationships between structural properties of vegetation such as biomass and the intensity of electromagnetic energy (or the ratio of energy at different wavelengths) that is received and recorded by optical passive or microwave sensors (Patenaude *et al*., 2005). However, these relationships are often useful in describing the canopy's two dimensional aspects but imprecise and at times inapplicable to its three dimensional aspects, particularly in complex, heterogeneous vegetation systems such as mixed deciduous woodlands (Ranson *et al*., 1997; Austin *et al*., 2003; Patenaude *et al*., 2005).

Light Detection and Ranging (lidar) can directly measure the three dimensional aspects of a vegetation canopy and has been shown to accurately estimate Leaf Area Index (LAI) and aboveground biomass even in those high biomass ecosystems where passive optical and active microwave sensors typically fail (Lefsky *et al*., 1999; Drake *et al.*, 2002; Patenaude *et al*., 2005; Boudreau *et al.*, 2008). A prerequisite to biomass modelling using lidar is a reliable estimation of the vegetation canopy height (Means *et al*., 1999; Drake *et al*., 2002, Lefsky *et al*. 2002; 2005; 2007; 2010).The main disadvantage of airborne lidar is its high cost for relatively low horizontal coverage. As such, for regional and global extents, attention has shifted to spaceborne lidar applications (Lefsky, 2010).

Currently, most spaceborne applications of lidar have used waveforms from the Geoscience Laser Altimeter System (GLAS) on the Ice Cloud and Land Elevation Satellite (ICESat) developed by the National Aeronautics and Space Administration (NASA). The sensor used 1064 nm laser pulses to illuminate an elliptical area (footprint) and record the returned laser energy from these footprints. The footprint size was nominally 65m in diameter but varied between the mission's 3 operation periods. The spacing between footprint centroids was about 175 m (Brenner *et al*., 2003). Details of the sensor specifications and methods of data collection can be found in Zwally *et al*. (2002) , Abshire *et al*. (2005) and Schutz *et al*. (2005).

A number of vegetation structural studies have used ICESat-GLAS data with an emphasis on characterizing forest canopy parameters and addressing the problem of slope at plot level (Lefsky *et al*., 2005; Lefsky *et al*., 2007; Chen, 2010; Xing *et al*., 2010; Lefsky, 2010; Duncason *et al*., 2010), and often using combinations of airborne and spaceborne data (Boudreau *et al.*, 2008; Hilker *et al.*, 2010). A common problem cited by these studies is that in steep terrain, the total length of the waveform from signal start to signal end (waveform extent) increases as a function of the product of the slope and the footprint size, and returns from both canopy and ground surfaces can occur at the same elevation thereby complicating waveform interpretation. Short height stands on steep terrain are the main problem (Lefsky *et al*., 2007; Gwenzi, 2008). This problem can be so significant that most studies have even considered discarding waveforms from footprints that are in high relief areas e.g. Sun *et al*. (2008) only considered flat areas; Baccini *et al.* (2008) and Dolan *et al*. (2009) only used waveforms in areas of at most 5 degrees; Xing *et al.* (2010) only used those waveforms in areas not exceeding 30 degrees; Simard *et al*. (2011) only used waveforms on slopes below 5 degrees and for which the original height was above another threshold level.

The correction of the slope problem and subsequent canopy height modeling has been done in various ways, providing varying levels of success. Lefsky *et al*. (2005) used the Shuttle Radar Topography Mission (SRTM) digital elevation model (DEM) data to calculate terrain slope indices subsequently used in combination with the waveform extent to estimate canopy height in 3 forest sites, (R^2 = 0.48 - 0.68). Subsequent work demonstrated the sufficiency of waveform information only, eliminating the need for DEM data. In 2007, Lefsky *et al.* developed a revised model that estimated mean canopy height using the waveform extent, leading edge extent and trailing edge extent (R^2 = 0.83, see section 2.3 for definitions). In 2010, Lefsky used full waveform extent as well as the 10th and 90th percentile of waveform energy to estimate Lorey's height, the basal area weighted height of all trees, (R^2= 0.67). Duncannon *et al*. (2010) used several other waveform parameters and a dummy variables model based on relief classes to estimate 85th percentile canopy height, (R^2 = 0.75 - 0.88). These studies have been conducted in forests (e.g. boreal forests of North America and Eurasia, temperate forests of China and North America, and tropical forests of South America and Africa) but very little has been done in savannas. Direct application of these methods to savannas may not yield accurate results because of the structural complexities of savannas.

Savannas have structural characteristics that are both advantageous and disadvantageous to monitor with lidar sensors. The main disadvantage of low canopy cover is that it reduces the power of the canopy return and increases the potential contribution of sloped terrain. This is especially so on short height stands where the relative contribution of vegetation to the waveform is much lower than that of the ground resulting in the waveform explaining more of ground than vegetation information. One advantage of such openness would be that it minimizes the problem of occluded understory and terrain surface experienced in some dense forests. As a result, some global canopy cover maps such as Lefsky (2010) for global and Saatchi *et al.* (2011) for the tropical latitudes focused on forests only and subsequent attempts have had inconsistencies partially due to differences in the interpretation of lidar data collected from open forests and savannas. As an example, Simard *et al.* (2011) reported that their map and that of Lefsky (2010) had large differences in vegetation cover and height, which they attributed partly to the fact that they mapped more open systems not covered by Lefsky (2010). Nonetheless, Simard *et al.*

(2011) used an untested slope bias correction method, hence the issue of slope and open canopy cover was not adequately addressed. Exploring these factors in detail will help in developing correction factors for global use of spaceborne lidar data.

Chen (2010) adopted Lefsky *et al.*'s (2005; 2007) methods in a savanna area in the Pacific coast. A number of linear and nonlinear models were developed to estimate maximum canopy height but the correlations between observed and modelled canopy height were very low, i.e. maximum r^2 value of 0.34 and a root mean square error (RMSE) of as high as 5m (which is about 40% of the mean stand maximum canopy height). In this paper we demonstrate that unique models should be developed for savannas that take into account the structural challenges posed by such systems. The short height, open canopies, multi-story arrangement, and terrain relief interactively influence the manner in which energy is reflected back to the sensor and must be considered in statistical modelling of canopy height. Based on this argument, we propose the inclusion of a canopy cover proxy to address the heterogeneity in canopy openness. Additionally, since the predictors act in an interactive manner, we also went further and tested the importance of interaction terms in developing the final canopy height models. Our second objective was to demonstrate the challenge of slope and short height combination that characterizes most savannas in mountainous areas. Our hypothesis was that better results are expected in patches with taller stands on flat terrain while short height stands on steep terrain are the most problematic.

2. Methods and materials

2.1. Study area

This research was done in the Oak Savannas of Santa Clara in California, USA. The site's extent is about 13650 Ha centred at 37.10° N, 121.25° W. The primary reasons for selecting this area were that it had available high accuracy airborne discrete lidar data from 2006 (coinciding with ICESat operation time) for validation and it is also highly heterogeneous in tree density and topography to allow for the investigation of slope and canopy cover. We believed that a success in this highly rugged terrain would indicate more likelihood of success in other relatively homogenous and flat savanna ecosystems. According to Chen (2010) and Baldocchi *et al.* (2011), this savanna ecosystem consists of a mix of herbaceous and woody, evergreen and deciduous and annual and perennial species. The co-dominant tree species are Blue Oak (*Quecus douglasii*), Coast Live Oak (*Quercus agrifolia*), Valley Oak (*Quercus Lobata*) and Buckeye (*Aesculus californica*) intermixed with Diablan sage scrub that is comprised of California Sagebrush (*Artemisia californica*) and non-native annual grassland. The average tree height is 11 m. Topography is highly heterogeneous, with mean slopes of 20 degrees. The mean annual air temperature is 15 °C and the annual precipitation range is 400 to 800 mm. Substantial work has been done to quantify structural features of this oak savannah with both direct and remote sensing methods (Baldocchi *et al*., 2011) but a number of questions still remain unanswered as far as lidar estimation of canopy height and biomass is concerned. Chen (2010) looked at canopy height derivation using ICESat - GLAS waveforms and concluded that terrain slope and the large diameter footprint of GLAS waveforms are the main limiting factors.

2.2. Data

The study used cloud free geolocated waveforms from ICESat - GLAS laser operation periods 2 and 3, acquired from 2003 to 2006. Processing algorithms from Lefsky *et al*. (2007; 2010) and Miller *et al*. (2011), along with tools developed by NASA Goddard were used to model waveforms. Data product GLA01 provided the raw waveforms while the GLA06/GLA14 data products provided surface elevations including the laser footprint geolocation and reflectance, as well as atmospheric corrections for range measurements. GLAS waveforms were filtered for atmospheric conditions using flag information (FRir_qaFlag = 15 and satNdx = 0) from the GLA14 data products following Chen (2010) and Duncanson (2010). As an additional filter, we only used those waveforms whose elevation was no more than 100 m below or above the SRTM elevation (Chen, 2010). High accuracy airborne discrete return lidar data was obtained from the United States Geological Survey (USGS) Centre for Lidar Information Coordination and Knowledge (CLICK). This point cloud data set has a density of 1 pulse per square meter and was acquired by an Optech ALTM 3100 lidar system by Optimal Geomatics in April and May 2006. The airborne lidar data was used for validating the ICESat-GLAS height models. Height indices from airborne lidar data are highly correlated to field measured height but have advantages over field estimates because of their accurate geolocation, high sampling density and correspondence with the three dimensional geometry of canopies (Lefsky *et al*., 2002; 2007).

2.3. GLAS waveforms processing and parameter extraction

Waveform modeling involved a series of steps including converting the original 0-255 values of the waveforms into voltage units and Gaussian decomposition. The transmitted and received waveforms were modeled and smoothed using algorithms developed by Brenner *et al*., 2003 and the detailed steps and equations are provided by (Harding & Carabajal, 2005; Duong *et al.*, 2006). Three main parameters (waveform extent, leading edge and trailing edge) were extracted from the waveforms (Figure 2.1) following the method developed by Lefsky *et al.*, 2007. Waveform extent is the distance between the point of signal start and signal end. Signal start is defined as the point when the increasing waveform intensity first crosses the background noise threshold level, corresponding to canopy tops. Signal end occurs where the decreasing waveform intensity cross the same threshold, corresponding to the last ground returns. The noise threshold level is calculated as mean background noise plus *n* times the standard deviation. Previous studies have used different values of *n* ranging from 3 to 4.5. Chen (2010) found that in this area the value of *n* for optimal threshold is 3.5 for signal start and 5 for signal end therefore an average value of 4.5 was used in this study. The leading edge is a function of canopy variability and is calculated as the distance between the elevation of signal start and the first elevation at which the waveform is half of the maximum signal above the background noise value. Trailing edge is the distance between elevation of signal end and lowest elevation at which the signal strength of the waveform is half of the maximum signal above the background noise value, and is a function of terrain slope.

2.4. Airborne lidar data processing

2.4.1. Bare earth modeling

Airborne lidar data points were first classified into ground return and non-ground return using MCC-LIDAR. MCC-LIDAR is a command line tool that uses an automated approach to iteratively identify non-ground (bare earth) points that exceed positive curvature thresholds at multiple scales using the Multiscale Curvature Classification algorithm (Evans & Hudak, 2007). The main advantage of the MCC algorithm is that it uses a thin-plate spline (TPS) which allows for adjustment of tension between points and integrates a multiscale approach where the surface is interpolated at different resolutions hence addressing topological relationships of non-ground objects at variable scales. The details of the parameters and how the algorithm further works are found in Evans & Hudak (2007).

Ground return points were then used to create a digital elevation model (bare earth surface) using the *GridSurfaceCreate* algorithm of FUSION software developed by the United States Department of Agriculture, Forest Service (McGaughey, 2012). *GridSurfaceCreate* uses points filtered as bare earth to compute the elevation of each grid cell using the average elevation of all points within the cell. The algorithm also has a spike option that works well to remove spikes that may have resulted from residual returns from vegetation. This was necessary in our study area since very short herbaceous plants characterize the understory of savannas and these are likely to be confused as ground returns in the first iterations.

2.4.2. Plot level canopy height metrics

We used the *Cloudmetrics* algorithm of FUSION to extract all airborne data points above the earlier generated bare earth surface and within circular plots centered at the waveform coordinates given by the GLA14 data product. Plot diameters were matched with footprint size of each waveform's laser operation period, i.e 70 m for laser 2 shots and 55 m for laser 3. These dimensions gave an area that is equal to the average area of the ecliptical footprint for each relative observation period. The *Cloudmetrics* algorithim of FUSION computes a variety of statistical parameters describing a lidar dataset using elevation and intensity values. From this processing we obtained the mean, maximum and n^{th} percentile heights for each plot.

2.4.3. NDVI as an index of canopy cover

A Landsat 5 Thematic Mapper (TM) image from summer (9 June, 2006) was acquired from the USGS site (http://glovis.usgs.gov/). The year 2006 was chosen to coincide with the year the airborne lidar data (used for validation) was acquired. Atmospheric correction was done on the image using the Quick Atmospheric Correction (QUAC) module of the Environment for Visualizing Images (ENVI) software (Bernstein *et al*., 2005). The NDVI index (Myneni *et al.*, 1995) was calculated and scaled from 0 to 255 for use as an index of canopy cover. Previous studies have shown a linear relationship between NDVI and percentage vegetation cover (Gamon *et al*., 1995; Todd & Hoffer, 1998). Since trees green up in summer, while grasses are gray/dead for this biome, the June NDVI index was considered to be a good variable to indicate crown cover in data analysis as explained in section 2.5. The aboveground biomass value for this area averages about 120 Mg/ Ha (Battles *et al.*, 2008) therefore it is highly unlikely that this NDVI-cover relationship saturates. The NDVI value for each plot was extracted using values of a 3 x 3 cell window around its center geolocation through bilinear interpolation.

2.5. Data Analysis

A non-parametric stochastic gradient boosting software (TreeNet) was used to relate waveform parameters and NDVI to airborne lidar data derived canopy height metrics. TreeNet (Salford Systems, 2001) uses an algorithm that generates thousands of small decision trees built in sequential error-correcting process to converge to an accurate model. This gives it an edge over other statistical methods like random forests where on the tree nodes the splitting attribute is selected from a randomly chosen sample of attributes. For consistency and simplicity of terms, we refer to the waveform extent as width, leading edge extent as lead, trailing edge extent as trail, maximum canopy height as H_{max} and any n^{th} percentile height as H_n. We evaluated the ability of waveform parameters and NDVI to estimate H_{max}, H_{80} and H_{90}. In this paper we present all the height metrics models, but focus on H_{max}. The percentile heights will mainly be used in the subsequent biomass modeling work since they have been proven to emphasize more the importance of large trees (Duncanson *et al*., 2010), a situation that is ideal for computing plot level tree biomass, especially in open canopy areas.

We firstly developed and validated canopy height models with the commonly used waveform parameters (width, lead and trail) and then developed subsequent models adding NDVI and interaction terms to get the following 10 candidate variables:

width (*w*); lead (*l*); trail (*t*); *NDVI*; width*lead (*wl*), width*trail (*wt*) ;

width *NDVI (*w*NDVI*); lead*trail (*lt*); lead*NDVI (*l*NDVI*); trail* NDVI (*t*NDVI*)

We ran the TreeNet model with all 10 variables and determined the order of their importance. The model was then run 9 times, dropping one variable (starting with least important) on each run so that we could trace the change in goodness of fit statistics in relation to number of model parameters. The goodness of fit statistics calculated and traced were Training R^2 value, Validation R^2 value, AIC and BIC and these were used as the cut off criterion in determining optimum number of parameters to use in the final model. We also tested the results of estimating all the 3 canopy height metrics with interaction terms but excluding NDVI so we could roughly judge the statistical worthiness of the extra effort required in its inclusion.

A tenfold cross validation approach was used to assess the accuracy of each model. With this approach, the original data set is randomly partitioned into 10 subsamples. Of these 10 subsamples, 9 are used as training data while the other one is retained as the validation data for testing the model. This process is repeated 10 times and each of the 10 subsamples is used exactly once for validation. In the end the 10 results from the 10 folds are averaged to produce a single estimation. This approach is useful in work like ours where the ultimate purpose of modeling is prediction (Kohavi, 1995).

To test our hypothesis that short height stands at steep slopes are the most problematic we divided our data set into 4 terrain-height classes using natural breaks in the data identified though agglomerative hierarchical cluster analysis. The four classes and the terms we use later to refer to them (in parentheses) are shown in Table 2.1. For each class we calibrated the best canopy height model identified in earlier steps and judged its accuracy in that class using the RMSE expressed as a percentage of the class' mean H_{max}. Two regression models were then developed to show the trend of modeling accuracy versus the slope-height interaction. Since this resulted in only one observation as a measure of accuracy in each class, we generated pseudo-replicates to obtain more observations to use in the regression model. This was done by a bootstrapping approach where a random sample of 75% of the data points in each class was drawn and used to run the model and save the RMSE. This would be repeated by replacing the sample and resampling over 1000 iterations. The resulting 1000 observation points for each class were then used in a dummy variable regression model of mean RMSE versus height-slope interaction class, and a linear model of each class' mean height/slope ratio versus mean RMSE.

3. Results

3.1. Maximum canopy height

Modeling accuracy (Table 2.2) was low using width, lead and trail only and slightly improved when NDVI was added. The model became much better when interaction terms were used. For all the models the training R^2 was always higher than the validation R^2 value. After including interaction terms, the order of decreasing variable importance picked by TreeNet was *w, l*NDVI, w*NDVI, wl, lt, t, t*NDVI, NDVI, l, wt.* As Figure 2.2 shows, using the earlier mentioned goodness of fit statistics, the optimum number of parameters (excluding intercept) was the first 6 ($R^2_{training}$ = 0.89; $R^2_{validation}$ = 63 ; RMSE = 1.6 m, i.e 12%). Modeling with interaction terms but excluding NDVI reduced the model's $R^2_{training}$ to 0.83 and the $R^2_{validation}$ to 0.59 while increasing the RMSE from 12% to 15%.

3.2. Percentile canopy height

As shown in Tables 2.2, the important variables for percentile heights differed from those identified in the maximum canopy height models. Overall the models for percentile heights did not perform as well as those for maximum canopy height. Between the two percentiles, H_{90} models were better than those for H_{80}. For both percentiles, adding *NDVI* to *width, lead,* and *trail* without interaction did not improve the model at all. After including interaction terms, the H_{80} model had 8 optimum variables while the H_{90} model had 7. *NDVI* was not an important variable for the percentile height models since the 2 models with interaction terms including and excluding *NDVI* performed just about the same. This suggests that cover may not have an important influence when the percentile canopy height metric is used. However, as with maximum canopy height, the models with interaction terms were better than those without.

3.3. Slope-height interaction

As we had hypothesized, the *short-steep* class had the highest RMSE values. *Short-flat* and *tall-steep* classes had better results and the *tall-flat* class had the best (Figure 2.3A). This relationship was significant ($p < 0.05$) as verified by both the dummy variables model and the mean height/slope ratio model. The dummy variables model had an R^2 value as high as 0.94, $p<0.001$. The second model (Figure 2.3B) shows the significant inverse relationship ($R^2 = 0.87$, $p = 0.042$) between height/slope ratio and RMSE. The *short-steep* class has the lowest height/slope ratio and the highest RMSE while *tall-flat* class has the highest height/slope ratio and the lowest RMSE, therefore we failed to reject our hypothesis. The mean RMSE for the 4 height-slope classes was slightly higher than the best model’s RMSE (presented in table 2.2) because the former was heavily distorted by the low accuracy associated with steep terrain areas when they are modeled separately (especially those plots in the short-steep class).

4. Discussion

In forest based studies such as Lefsky *et al.* (2005; 2007; 2010), models developed using the variables *w*,*l* and *t* gave good results. In this work they gave less accurate results suggesting that we need special models for savanna ecosystems. Interaction terms are a better representation of the interactive nature of the factors that contribute to the complex structure of savannas. The *l*NDVI* and *w*NDVI* terms demonstrate the importance of stem density. Since *l* is a function of canopy variability, it is very much influenced by the density of stems within the plot, hence the significance of *l* in combination with *NDVI* (our surrogate measure for crown cover). The *w*NDVI* term is important in the sense that canopy cover determines the power of ground returns recorded by the sensor. Holding terrain slope constant, dense canopies can result in relatively shorter *w* because there would be insufficient energy returned from the ground and waveform extent will not capture the full range of footprint elevations (Lefsky *et al*., 2007). The first 3 variables (*w, l*NDVI, w*NDVI*) in our best model therefore explain much of the vegetation physical realism in savannas. The other variables like *lt* explain more of the terrain issues. The interactive effect of height and terrain slope was also demonstrated by the height-slope class results. Modeling accuracy decreases as height reduces or as slope increases.

Our work has demonstrated that waveform information alone (*t*) is sufficient to account for slope terrain, although it is less accurate at higher terrain slopes. Duncanson *et al*. (2010) suggested another way of using only waveform information to correct for terrain. They developed a model that used 4 waveform parameters (3 of them different from the ones we used) to calculate footprint maximum relief. Footprints were then grouped into relief classes and a model with dummy variables for each class finally developed to estimate 85th percentile canopy height. Although the results were good ($R^2 = 0.81$) in their forested area, the final model required as many as 10 variables computed separately. Our simpler model gave good results with only 6 terms. Moreover, as we have shown earlier with the Lefsky *et al*. (2007) and Chen (2010) models, merely adopting this method to savanna ecosystems is not likely to give good results, given the complex structure of savannas, especially when terrain slope is high.

In this work we used NDVI as a surrogate variable but a more accurate way of directly computing canopy cover would be better. We selected NDVI because it is easily derived from freely available data sources such as Landsat and the Moderate Resolution Imaging Spectroradiometer (MODIS). Moreover, optical passive remote sensing imagery has generally been proven to be good at delineating the two dimensional structural attributes of vegetation (Patenaude *et al*., 2005). This will be a very important wall-to-wall data source in the next steps of this kind of work i.e. modeling biomass distribution across landscapes. However, the density and phenology of the vegetation has a big impact on the usefulness of NDVI as a proxy for canopy cover. In our study site, it worked well because the trees and grass green up in different seasons. This empirical relationship may not work well in areas where at one point in time there is greenness from both trees and grasses or when stem density and biomass values are so high that the NDVI-cover relationship saturates. We therefore recommend this as one area of improvement in subsequent studies. Alternatives include using NDVI and texture to discriminate between vegetation and grass or directly computing canopy cover from freely available optical passive remote sensing imagery such as Landsat TM alongside freely available software such as Forest Canopy Density Mapper (Rikimaru, 2002) but these methods were not tried in our work and we leave it as a recommendation to others who may do similar work in future. It may also be possible to compute canopy cover from the lidar waveforms themselves.

Dropping NDVI and using only *w,l,t,wl*, *lt* and *wt* gave a model that was still better than the one developed using *w,l* and *t* alone but the amount of variability explained became much lower hence we strongly recommend the inclusion of NDVI. We plan to test and compare the relationships of percentile heights and Lorey's height to aboveground biomass in future work. These height metrics are related to basal area which is a measure of the area occupied by the cross section of tree trunks hence they may be sufficient for modeling biomass when canopy cover data is not available. The percentile height model results were poorer than those of H_{max} most probably because of the way H_{90} and H_{80} were computed from the point cloud lidar data. A cut off point of 2 m was used to separate tree returns from non-tree returns but such a cut off is not applied in determining waveform width where the presence of dense understory vegetation such as herbs complicates the discrimination of background noise and signal end. This is an inevitable problem in savannas since such short understory vegetation can always be expected.

This work has emphasized that in savannas, the challenge of canopy height modeling with waveform lidar is not only slope, but the short heights of stands in these biomes as well. It is because of these complex interactions that models developed for forests will fail when merely adopted to savannas. Future missions of spaceborne lidar such as ICESat-2 remain untested in savanna ecosystems, and should be included in mission designs due to the observed differences between savanna and forest structure.

5. Conclusion

Our work has demonstrated the capability of lidar waveforms to predict canopy height in savannas. However, the complex structures of savannas require different models from those developed for forests. Canopy cover and interaction terms are a very important input into savannas specific models. These results suggest a possibility for height change detection using spaceborne lidar data back to at least 2003 when ICESat-GLAS data was first available. If ICESat-2 data will be usable in savannas as well then continuity is ensured in the global availability of multi-temporal spaceborne lidar data and hence improving our knowledge about savanna ecosystems' contribution to important global ecological processes such as the carbon cycle.

6. Tables and Figures

Table 2.1: Height-slope classes used to test the interactive influence of height and slope in canopy height modeling

Class	**Height range**	**Slope range**
Tall stands on flat terrain (*Tall -flat*)	Above 11m	Below 8 degrees
Tall stands on steep terrain (*Tall-steep*)	Above 11m	Above 8 degrees
Short stands on flat terrain (*Short-flat*)	Below 11m	Below 8 degrees
Short stands on steep terrain (S*hort-steep*)	Below 11m	Above 8 degrees

Table 2.2: Modeling results

Height Index	**Model and important variables**	**R^2**		**RMSE (% of mean H_{max})**
		Training	**Validation**	
H_{max}	**A.Width, lead and trail only** $H_{max} \sim w+l+t$	0.73	0.55	19
	B. NDVI added to model 1 $H_{max} \sim w+l+t+NDVI$	0.78	0.57	17
	C. Interaction terms added to model 2 $H_{max} \sim w+lNDVI+wNDVI+wl+lt+t$	0.89	0.63	12
	D.Interaction terms without NDVI $H_{max} \sim width+wl+lt+t$	0.83	0.59	15
H_{80}	**A.Width, lead and trail only** $H_{80} \sim w+l+t$	0.68	0.40	19
	B. NDVI and interactions terms added to model 1 $H_{80} \sim w+ lNDVI+wNDVI+wl+lt+tNDVI+wt+t$	0.80	0.42	16
	C. Interaction terms; without NDVI $H_{80} \sim w+wl+lt+wt+t$	0.74	0.40	17
H_{90}	**A.Width, lead and trail only** $H_{90} \sim w+l+t$	0.70	0.42	19
	B. NDVI and interactions terms added to model 1 $H_{90} \sim w+ lNDVI+wNDVI+wl+lt+tNDVI+t$	0.76	0.46	17
	C. Interaction terms; without NDVI $H_{90} \sim w+wl+lt+t$	0.77	0.48	17

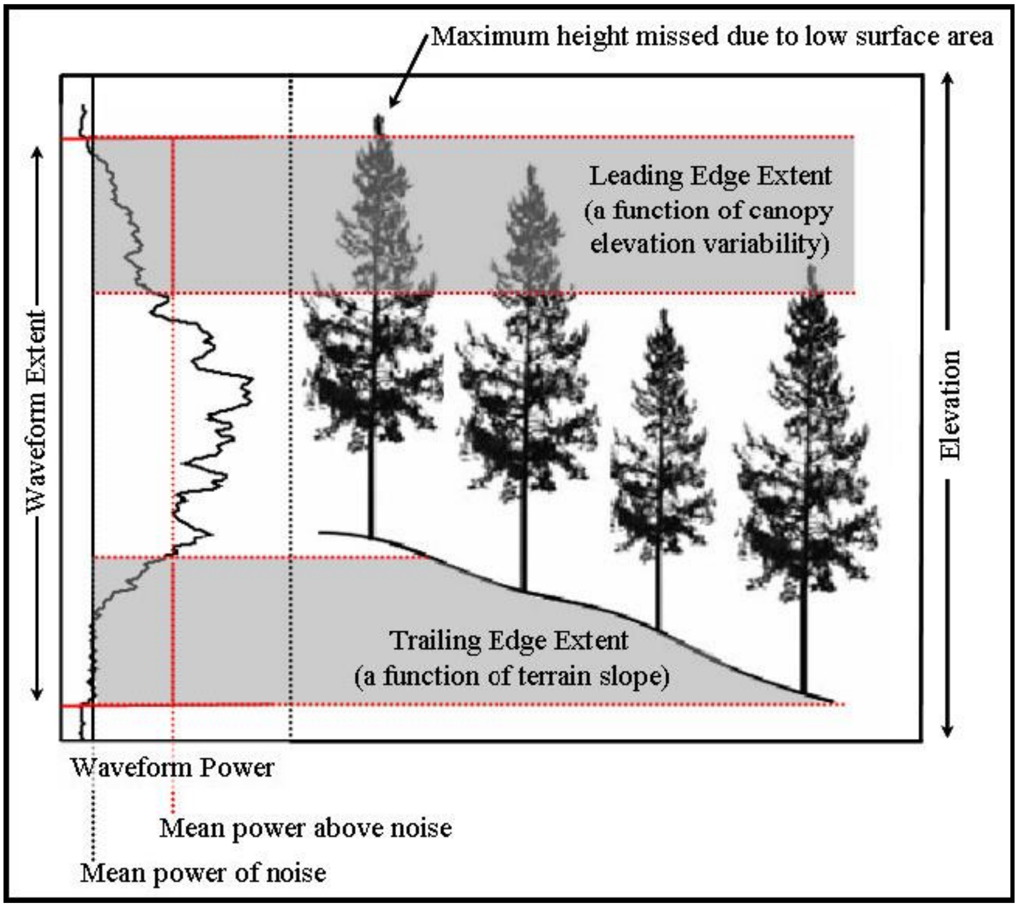


Figure 2.1: Diagrammatic representation of the waveform parameters used. Adopted from Lefsky *et al*., (2007)

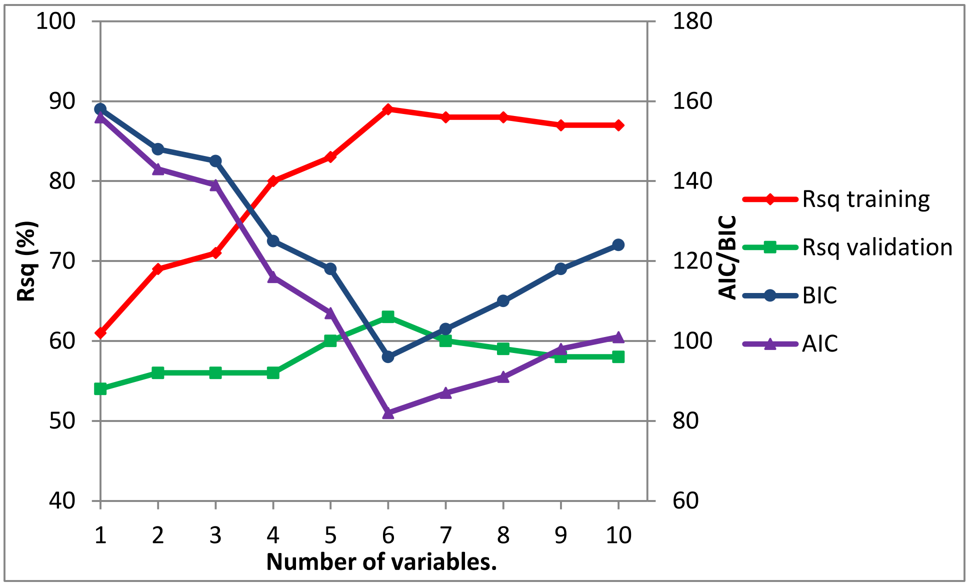


Figure 2.2: Modeling assessment criterion trends. The number of variables on the x-axis corresponds to the cumulative variables in the following order: *w*, *l* $*$ *NDVI*, *w* $*$ *NDVI*, *wl*, *lt*, *t*, *t* $*$ *NDVI*, *NDVI*, *l*, *and wt*.

CHAPTER 2 REFERENCES

Abshire, J. B., Sun, X., Riris, H., Sirota, M. J., McGarry, J. F., Palm, S., … Liiva, P. (2005). Geoscience Laser Altimeter System (GLAS) on the ICESat mission: On-orbit measurement performance. *Geophysical Research Letters*, *32*(21), 1–4.

Austin, J. M., Mackey, B. G., & Van Niel, K. P. (2003). Estimating forest biomass using satellite radar: an exploratory study in a temperate Australian Eucalyptus forest. *Forest Ecology and Management*, *176*(1-3), 575–583.

Baccini, A., Laporte, N., Goetz, S. J., Sun, M., & Dong, H. (2008). A first map of tropical Africa’s above-ground biomass derived from satellite imagery. *Environmental Research Letters*, *3*(4), 045011.

Baldocchi, D.D., Chen, Q., Chen, X., Ma, S., Miller, G., Ryu, Y., Xiao, J., Wenk, R., & Battles, J. (2011). The dynamics of energy, water and carbon fluxes in a Blue Oak (Quercus douglasii) savanna in California, USA. In M. . Hill & N. . Hanan (Eds.), *Ecosystem Function in Savannas: Measurement and Modeling at Landscape to Global Scales*. Florida: Tylor and Francis Group.

Battles, J. J., Jackson, R. D., Shlisky, A., & Bartolome, J. W. (2008). Net Primary Production and Biomass Distribution in the Blue Oak Savanna 1. In A. Merenlender, D. McCreary, & K. L. Purcell (Eds.), *Proceedings of the Sixth Symposium on Oak Woodlands : Today’s Challenges, Tomorrow's Opprtunities* (pp. 511–524). USDA Forest Service, General Technical Report PSW-GTR-217.

Bernstein, L. S., Adler-Golden, S. M., Sundberg, R. L., Levine, R. Y., Perkins, T. C., Berk, A., … Hoke, M. (2005). Validation of the QUick Atmospheric Correction (QUAC) algorithm for VNIR-SWIR multi- and hyperspectral imagery. In *SPIE Proceedings, Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XI Vol. 5806* (pp. 668–678). SPIE Digital Library.

Bombelli, A., Henry, M., Castaldi, S., Arneth, A., Grandcourt, A. De, Grieco, E., … Cedex, M. (2009). An outlook on the Sub-Saharan Africa carbon balance. *Biogeosciences*, *6*(10), 2193–2205.

Boudreau, J., Nelson, R., Margolis, H., Beaudoin, A., Guindon, L., & Kimes, D. (2008). Regional aboveground forest biomass using airborne and spaceborne LiDAR in Québec. *Remote Sensing of Environment*, *112*(10), 3876–3890.

Brenner, D., Brenner, M., Brenner, A., Harding, D., & Zwally, H. (2003). Derivation of Range and Range Distributions from Laser Pulse Waveform Analysis for Surface Elevations, Roughness, Slope and Vegetation Heights. Algorithm Theoretical Basis Document. Volume 4.1. http://www.csr.utexas.edu/glas/pdf/Atbd_20031224.pdf.

Chen, Q. (2010). Retrieving vegetation height of forests and woodlands over mountainous areas in the Pacific Coast region using satellite laser altimetry. *Remote Sensing of Environment, 114*(7), 1610–1627.

Dolan, K., Masek, J. G., Huang, C., & Sun, G. (2009). Regional forest growth rates measured by combining ICESat GLAS and Landsat data. *Journal of Geophysical Research, 114*(G00E05), 1–7.

Drake, J. B., Dubayah, R. O., Clark, D. B., Knox, R. G., Blair, J. B., Hofton, M. A., … Prince, S. D. (2002). Estimation of tropical forest structural characteristics using large-footprint lidar. *Remote Sensing of Environment*, *79*(2-3), 305–319.

Duncanson, L. I., Niemann, K. O., & Wulder, M. A. (2010). Estimating forest canopy height and terrain relief from GLAS waveform metrics. *Remote Sensing of Environment*, *114*(1), 138–154.

Duong, H., Pfeifer, N., & Lindenbergh, R. (2006). Analysis of repeated ICESat full waveform data: methodology and leaf-on/leaf-off comparison. In *Workshop on 3D Remote Sensing in Forestry*. Vienna, Austria.

Evans, J. S., & Hudak, A. T. (2007). A Multiscale Curvature Algorithm for Classifying Discrete Return LiDAR in Forested Environments. *IEEE Transactions on Geoscience and Remote Sensing*, *45*(4), 1029–1038.

Gamon, J. A., Field, C. B., Goulden, M. L., Griffin, K. L., Hartley, E., Joel, G., … Valentini, R. (1995). Relationships between NDVI, canopy structure and photosythesis in three Californian vegetation types. *Ecological Applications*, *5*(1), 28–41.

Gwenzi, D. (2008). *Spaceborne Lidar canopy height estimation for aboveground forest biomass assessment in the cool montane area of North East China*. University of Twente: http://www.itc.nl/Pub/Home/library/Academic_output/AcademicOutput.html?p=11&y=8&l=20.

Harding, D. J., & Carabajal, C. C. (2005). ICESat waveform measurements of within-footprint topographic relief and vegetation vertical structure. *Geophysical Research Letters*, *32*(21), 1–4.

Hilker, T., Leeuwen, M., Coops, N. C., Wulder, M. a., Newnham, G. J., Jupp, D. L. B., & Culvenor, D. S. (2010). Comparing canopy metrics derived from terrestrial and airborne laser scanning in a Douglas-fir dominated forest stand. *Trees*, *24*(5), 819–832.

Hill, M. J., & Hanan, N. P. (2010). *Ecosystem Function in Savannas: Measurement and Modeling at Landscape to Global Scales*. (M. J. Hill & N. P. Hanan, Eds.). CRC Press.

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *International Joint Conference on Artificial Intelligence* (pp. 1137–1143). San Mateo, CA.

Lefsky, M. A. (2010). A global forest canopy height map from the Moderate Resolution Imaging Spectroradiometer and the Geoscience Laser Altimeter System. *Geophysical Research Letters*, *37*(15), 1–5.

Lefsky, M. A., Cohen, W. B., Acker, S. A., Parker, G. G., Spies, T. A., & Harding, D. (1999). Lidar Remote Sensing of the Canopy Structure and Biophysical Properties of Douglas-Fir Western Hemlock Forests. *Remote Sensing of Environment*, *70*(3), 339–361.

Lefsky, M. A., Harding, D. J., Keller, M., Cohen, W. B., Carabajal, C. C., Del Bom Espirito-Santo, F., … de Oliveira Jr, R. (2005). Estimates of forest canopy height and aboveground biomass using ICESat. *Geophysical Research Letters*, *32*(22), 1–4.

Lefsky, M. A., Keller, A. M., Pang, Y., Camargo, P. B. de, & Hunter, M. O. (2007). Revised method for forest canopy height estimation from Geoscience Laser Altimeter System waveforms. *Journal of Applied Remote Sensing*, *1*.

Lefsky, M., Cohen, W., Parker, G., & David Harding. (2002). Lidar Remote Sensing for Ecosystem Studies. *BioScience*, *52*(1), 19–30.

Lucas, R. M., Lee, A. C., Amston, J., Carreiras, J. M. B., Viergever, K. M., Bunting, P., … Woodhouse, I. (2011). Quantifying carbon in Savannas: The role of active sensors in measurements of tree structure and biomass. In M. J. Hill & N. P. Hanan (Eds.), *Ecosystem function in Savannas*. Florida: Tylor and Francis Group.

McGaughey, R. J. (2012). FUSION/LDV: Software for LIDAR Data Analysis and Visualization. http://forsys.cfr.washington.edu/fusion/fusionlatest.html. United States Department of Agriculture, Forest Service.

Means, J. E., Acker, S. A., Harding, D. J., Blair, J. B., Lefsky, M. A., Cohen, W. B., … McKee, W. A. (1999). Use of Large-Footprint Scanning Airborne Lidar To Estimate Forest Stand Characteristics in the Western Cascades of Oregon. *Remote Sensing of Environment*, *67*(3), 298–308.

Miller, M. E., Lefsky, M., & Pang, Y. (2011). Optimization of Geoscience Laser Altimeter System waveform metrics to support vegetation measurements. *Remote Sensing of Environment*, *115*(2), 298–305.

Myneni, R. B., Hall, F. G., Sellers, P. J., & Marshak, A. L. (1995). The Interpretation of Spectral Vegetation Indexes. *IEEE Transactions on Geoscience and Remote Sensing*, *33*(2), 481–486.

Patenaude, G., Milne, R., & Dawson, T. (2005). Synthesis of remote sensing approaches for forest carbon estimation: reporting to the Kyoto Protocol. *Environmental Science & Policy*, *8*(2), 161–178.

Potter, C. (2011). Carbon cycle and vegetation dynamics of Savannas based on Global satellite data products. In M. J. Hill & N. P. Hanan (Eds.), *Ecosystem function in Savannas: Measurement and Modeling at Landscape to Global Scales*. Florida: Tylor and Francis Group.

Ranson, K. J., Lang, R. H., Chauhan, N. S., Cacciola, R. J., Kilic, O., & Guoqing, S. (1997). Mapping of boreal forest biomass from spaceborne Synthetic Aperture Radar. *Journal of Geophysical Research*, *102*(D24), 29599–29610.

Rikimaru, A. (2002). Tropical forest cover density mapping. *Tropical Ecology*, *43*(1), 39–47.

Rotenberg, E., & Yakir, D. (2010). Contribution of semi-arid forests to the climate system. *Science (New York, N.Y.)*, *327*(5964), 451–4.

Saatchi, S. S., Harris, N. L., Brown, S., Lefsky, M., Mitchard, E. T. a, Salas, W., … Morel, A. (2011). Benchmark map of forest carbon stocks in tropical regions across three continents. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(24), 9899–904.

Salford Systems. (2001). TreeNet Stochastic Gradient Boosting : An implementation of the MART methodology. San Diego, California, USA: Salford Systems.

Schimel, D. S. (2010). Drylands in the Earth system. *Science*, *327*(5964), 418–9.

Scholes, R. J., & Hall, D. O. (1996). The Carbon Budget of Tropical Savannas, Woodlands and Grasslands. In A. I. Breymeyer, D. O. Hall, J. M. Melillo, & G. I. Agren (Eds.), *Global Change: Effects on Coniferous Forests and Grasslands* (pp. 69–100). Wiley, Chichester: House and Hall 2001.

Schutz H. J. Zwally, C. A.Shuman, D. Hancock, and J. P. DiMarzio, B. E. (2005). Overview of the ICESat Mission. *Geophysical Research Letters*, *32*(L21S01), 1–4.

Simard, M., Pinto, N., Fisher, J. B., & Baccini, A. (2011). Mapping forest canopy height globally with spaceborne lidar. *Journal of Geophysical Research*, *116*(G04021), 1–12.

Sun, G., Ranson, K., Kimes, D., Blair, J., & Kovacs, K. (2008). Forest vertical structure from GLAS: An evaluation using LVIS and SRTM data. *Remote Sensing of Environment*, *112*(1), 107–117.

Todd, S. W., & Hoffer, R. M. (1998). Responses of Spectral Indices to Variations in Vegetation Cover and Soil Background. *Photogrammetric Engineering and Remote Sensing*, *64*(9), 915–921.

Williams, M., Ryan, C., Rees, R., Sambane, E., Fernando, J., & Grace, J. (2008). Carbon sequestration and biodiversity of re-growing miombo woodlands in Mozambique. *Forest Ecology and Management*, *254*(2), 145–155.

Wulder, M. A., Hall, R. J., Coops, N. C., Steven, E., & Franklin, S. E. (2004). High Spatial Resolution Remotely Sensed Data for Ecosystem Characterization. *BioScience*, *54*(6), 511–521.

Xing, Y., de Gier, A., Zhang, J., & Wang, L. (2010). An improved method for estimating forest canopy height using ICESat-GLAS full waveform data over sloping terrain: A case study in Changbai mountains, China. *International Journal of Applied Earth Observation and Geoinformation*, *12*(5), 385–392.

Zwally, H. J., Schutz, B., Abdalati, W., Abshire, J., Bentley, C., Brenner, A., … Thomas, R. (2002). ICESat's laser measurements of polar ice, atmosphere, ocean, and land. *Journal of Geodynamics*, *34*(3-4), 405–445.

CHAPTER 3: PLOT LEVEL ABOVEGROUND WOODY BIOMASS MODELING USING CANOPY HEIGHT AND AUXILIARY REMOTE SENSING DATA IN A HETEROGENEOUS SAVANNA[2]

Synopsis

Remote sensing studies aimed at assessing woody biomass have demonstrated a strong relationship between canopy height and plot level aboveground biomass, but these studies have mostly been performed in forests. To date, few studies have examined the limitations and challenges of this relationship using large footprint lidar in savannas. Furthermore, methods for the comparison of generalized versus species composition or vegetation type-specific models have not been adequately explored at the plot level. In this work, we developed generalized models using the classical least squares regression modeling approach to relate selected height metrics to above ground woody biomass and then employed a Hierarchical Bayesian Analysis (HBA) to explore the implications of using generalized instead of composition-specific models. Our study used field data, airborne discrete return lidar and Landsat 5 TM data collected from the oak savannas of Tejon Ranch Conservancy in Kern County, California. Model parameters were developed and analyzed at the level of 50 m diameter plots, comparable to the resolution of large footprint lidar waveforms. The three generalized models that incorporated canopy cover proxies performed better than one model that did not use canopy cover information. From the HBA, we found out that for all the models, both the intercept and slope have interspecific

[2] David Gwenzi and Michael Andrew Lefsky, 2015
Submitted to the Journal of Applied Remote Sensing

variability. The valley oak dominated plots consistently had higher slopes and intercepts while the plots dominated by blue oaks had the lowest. However, the intercept and slope values of the composition-specific models did not differ much from the global (overall) values and their 95% Credible Intervals (CIs) overlapped the global mean values. We conclude that the narrow range of the distribution and the overlap of the CIs of the composition-specific and global parameters suggest that scaling rules do exist for savannas. The distribution of the posterior densities of the differences between composition level and global level parameter values showed a high support for the use of global parameters suggesting that all of the 4 models are universally (large scale) applicable. Therefore, in this case the choice of method depends more on secondary considerations.

Key Words: Lidar, Canopy height, aboveground biomass, canopy cover, Hierarchical Bayesian

1. Introduction

Large footprint lidar has a demonstrated capability to measure canopy height in both forests and savannas (Duncanson et al. 2010; Xing et al. 2010; Lefsky, 2010; Gwenzi & Lefsky 2014) thus it is recognized as a valuable technique for large scale assessment of vegetation structure and function. Canopy height is often measured as part of aboveground biomass assessment since the two have been found to be highly correlated for forest plots (Lefsky et al. 1999; Means et al. 1999; Drake et al. 2002; Lefsky et al. 2002; 2005; 2007; Boudreau et al. 2008; Zolkos et al. 2013). Little work has been done on lidar remote sensing of biomass in savanna landscapes or other open canopy ecosystems, and the published work has often used small footprint lidar (Colgan *et al*., 2012; Nyström *et al*., 2012; Colgan *et al*., 2013; McGlinchy *et al*., 2014) or the analyses were done at the individual tree level (Wu *et al*., 2009). The only available satellite based lidar data to date was collected by the Ice Cloud and land Elevation Satellite's Geoscience Laser Altimeter System (ICESat-GLAS), operational between 2003 and 2010.

Large scale assessment of aboveground biomass in savannas is important for determining aspects of ecological function such as carbon sequestration, habitat availability, bio-energy production, water flow, herbivory, and fuelwood supply. For large scale applications, large footprint spaceborne lidar data sources are often preferable to small footprint airborne lidar because of the former's lower cost and increased availability. The selection of height metrics like maximum canopy height in savannas is complicated by the structural heterogeneity of these ecosystems. Capturing this heterogeneity may require the use of extra variables to account for differences in stem density and the vertical structure of vegetation observed among plots, even in the same locality. Canopy cover directly estimates the amount of woody cover in a plot, making it an important variable to consider in modeling biomass in such heterogeneous vegetation systems.

Canopy cover can be accurately derived using high point density discrete return lidar data (Colgan *et al*., 2012) but in most systems it can also be estimated by multispectral vegetation indices like NDVI to which it is highly correlated (Gamon *et al*., 1995; Todd & Hoffer, 1998). Another approach to incorporate canopy cover information is to weight the canopy height by some variables that are directly related to stem density. Lorey's height (VanLaar & Akca, 2007) is a commonly used metric that uses basal area or crown area weighting (Pang *et al*., 2008). Percentiles such as the 90th percentile height metric emphasize the importance of larger trees making them another good alternative.

The selection of methods to compare the utility of generalized versus species or composition or vegetation type-specific regression equations has been a problem for studies of individual tree allometry relating diameter, height and biomass in both forests and savannas. Previous studies have demonstrated that at the individual tree level the use of generalized allometric models (with Diameter at Breast Height (DBH) and total height) can lead to bias in estimating the biomass of particular tree species (Zianis & Mencuccini, 2004; Chave et al., 2005; Litton & Kauffman, 2008; Mwakalukwa *et al*., 2014). However, some of these studies have also noted an allometric convergence of the scaling exponents despite the multitude of site-specific factors affecting tree growth (Zianis & Mencuccini, 2004; Pilli *et al*., 2006; Tredennick *et al*., 2013). Thus, despite the better performance of species/composition/vegetation type or site-specific models, generalized models can be applied to achieve comparable results with less time and effort. At the plot level, these trends may be different since errors in model initialization tend to compensate at larger spatial scales (Hurtt *et al*., 2010). Large scale remote sensing estimation of biomass often use plot parameters thus an investigation of the differences between generalized and species/composition-specific models at the plot level is necessary for evaluating large scale efforts. The inherent heterogeneity of savanna ecosystems complicates the use of species/composition-specific models and yet the errors resulting from using generalized models are not well investigated at the plot level.

In this work we investigated the utility of selected canopy height metrics for estimating aboveground woody biomass at the plot level in a typical savanna landscape. For our first objective, we developed and tested generalized models that related canopy height to biomass using the empirical frequentist statistical approach, which has been the standard in previous work

(Lefsky *et al*. 1999; 2002; 2005; Means *et al*. 1999; Drake *et al*. 2002). The second objective was to investigate the influence of using generalized models instead of species-specific models. Because of the mixed vegetation characteristics of savannas, our plots rarely comprised of single tree species thus our classification was defined by the dominant species in each plot. As a result, in this paper we refer to the different groups of plots as composition classes. Comparisons of species or composition-specific and generalized models for estimating biomass whether at tree level or plot level has been traditionally done by calculating and comparing the relative bias of each model's estimates to the validation data set or testing of differences in slope and intercept parameters of the respective classical regression models. These approaches are mostly useful with large enough data sets but can be difficult or even impossible when the sample size is too small to obtain significant sub-models.

In contrast, bayesian inference (section 1.1) is unbiased with respect to sample size. In their work in forests, Zapata-Cuartas *et al*. (2012) were able to create and evaluate Bayesian models with a sample size of 6 on a task that would require a sample size of at least 40 for a classical statistical method. Bayesian analysis encorporates prior information about model parameters to produce an updated distribution (posterior) and a metric of estimate reliability (Robert, 2007; Hall, 2012; Zapata-Cuartas *et al*., 2012). The prior distribution for a parameter (θ) is updated after accounting for observed data (y) to yield the posterior distribution. On the contrary, the frequentist approach does not condition on the observed data but rather the accuracy of the evidence from an experiment is restricted to statements about long run averages from hypothetical replicates of sampled data, were the experiment repeatedly performed (Jaynes, 2003; Wagenmakers *et al*., 2008).

Because of the low sample size of our data, and the additional advantages mentioned above, we chose to use a Bayesian approach to create hierarchical models so that the posterior distributions of composition-specific parameters could be compared to those of the global parameters. This allowed us to investigate the effect of using general models over composition-specific models in estimating plot level above ground biomass in savannas. Bayesian analysis is a relatively unexplored area in large scale lidar remote sensing of vegetation structure although some studies e.g Zapata-Cuartas *et al*. (2012) in forests and Tredennick *et al*. (2013) in savannas have investigated the usefulness of a Bayesian approach for the creation of tree level allometric equations. These studies demonstrated the importance of considering allometric scaling coefficients in the framework of probability distributions rather than as fixed parameter values, as explained in the following sections.

1.1. Bayesian Analysis

Bayesian methods are based on Bayes' rule (Carlin & Thomas, 2000; Ghosh *et al,* 2006) which breaks down knowledge into 4 components: prior knowledge (1) and new data (2) are combined by a model (3) to produce posterior/updated knowledge (4). For estimating parameters, the prior distribution is the probability distribution of the parameter that we have before observing the data. When the prior has minimal impact on the posterior distribution, it is said to be objective, sometimes called non-informative. On the other hand, the prior is subjective/informative if it expresses specific, definitive information about a variable. The subjectivity can be based on for example information gathered from a previous study, past experience or expert opinion. The posterior distribution represents our updated beliefs about the parameter after observing the data. Our new knowledge of the parameter is therefore contained in the posterior and statistical inferences are made by summarizing its distribution. The posterior distribution depends on the weight placed on the prior compared to the new data and the magnitude of the difference between the two. Thus, while a frequentist approach investigates the probability of observing the data, given that the hypothesis is true a Bayesian approach investigates the probability of the hypothesis being true, given the observed data (McCarthy, 2007; Robert, 2007).

Bayes' rule is based on conditional probability and for a finite number of hypotheses, it states that the probability of the hypothesis given the data is calculated by:

$$\Pr(H_\mathrm{i}|D) = \frac{\Pr(H_i) * \Pr(D \mid H_i)}{\sum_j \Pr(H_j) * \Pr(D \mid H_j)} \tag{1}$$

where $\Pr(H_\mathrm{j})$ is the prior probability of the different hypotheses and $\Pr(D|H_\mathrm{j})$ is the probability of obtaining the data given the hypotheses.

For continuous hypotheses, Bayes' rule is expressed as:

$$\Pr(H|D) = \frac{\Pr(H) * \Pr(D \mid H)}{\int_0^\infty \Pr(x) * \Pr(D \mid x) dx} \tag{2}$$

where H represents a particular value for the parameter and the limits of the integration are over all the possible vales of the parameter x. To summarize the above equations, the posterior probability equals the prior multiplied by the likelihood of the data and a scaling constant. The scaling constant is the denominator in both of the above cases. In Hierarchical Bayesian Analysis (HBA), models are written modularly, i.e. in terms of sub-models. The sub-models then combine to form the hierarchical model, and Bayes theorem is used to integrate the pieces together. The challenge of estimating the scaling constant analytically has been overcome by the development of software such as the Microsoft windows version of Bayesian inference Using Gibbs Sampling

(WinBUGS) (Spiegelhalter *et al*., 2005) and Just Another Gibbs Sampler (JAGS) (Plummer, 2003). These programs draw samples from the posterior distribution using Markov Chain Monte Carlo (MCMC) (Brooks *et al*., 2011; Robert & Casella, 2004). MCMC is a general purpose technique for generating samples from a probability in high (e.g. millions) dimensional state space, using random numbers drawn from uniform probability in a certain range.

2. Methods and materials

2.1. Study area

This research was conducted in the oak savannas of Tejon Ranch Conservancy (centered roughly at 34.85° N, 118.86° W). A cooperative agreement between the Tejon Ranch Company and a group of conservation organizations in 2008 resulted in the creation of this 72 000 Ha conservancy. The conservancy was created to protect and implement science based stewardship, thus preserving, enhancing and restoring the native biodiversity and ecosystem values of the Tejon Ranch and Tahachapi Range for the benefit of California's future generations (Tejon Ranch Conservancy, 2011). These oak savannas comprise mainly of blue oaks (*Quercus douglasii*), black oaks (*Quercus kelloggii*) and valley oaks (*Quercus lobata*). Other non-dominant tree species found in this ecosystem include canyon live oak (*Quercus chrysolepis*), interior live oak (*Quercus wislizeni*), the California buckeye (*Aesculus californica*) and a few conifers. Blue oak woodlands are dominant at the lower elevations (between 500 and 1 000 m), black oak woodlands are dominant in higher elevation areas (> 1 200 m) while valley oak woodlands are found on both lower (400- 600 m) and higher (1400- 1800 m) elevations. Grass dominates the understories of blue and valley oaks while shrubs are found in combination with grass in the understory of black oaks.

2.2. Data

Field work was conducted in August and September of 2012. Inventory was carried out on circular plots of 50 m diameter to approximate the nominal diameter of satellite based large footprint lidar (ICESat–GLAS). Plot center coordinates were recorded using a Trimble Juno 3C handheld GPS within an accuracy of 0.5 - 2 m. Diameter at Breast Height (DBH) and total height were measured for every live tree taller than 2 m and greater than 10 cm in DBH in each plot. A total of 26 plots were enumerated and later grouped into four main composition classes determined by the dominant tree species (i.e blue oak plots, n = 11; black oak plots, n = 6; valley oak plots, n = 6 and mixed plots, n = 3). The biomass for each tree was calculated using the respective species' equations from Jenkins *et al*. (2004).

Airborne discrete return lidar data was collected in July 2012 by a commercial lidar vendor at an average density of 1 point per m^2. The *las2dem* algorithm of LAStools (Isenburg, 2013) was used to create a digital elevation model, above which non-ground points were normalized to determine canopy height using the *lasheight* algorithm. As in fieldwork, a cut off height of 2 m was used to minimize the influence of non-tree vegetation. The plot center coordinates recorded during fieldwork were used to extract 50 m diameter plot level statistics from the point cloud lidar data. Plot statistics, specifically the maximum height metric and canopy cover were calculated using the *lascanopy* algorithm of LAStools and the *cover* algorithm of FUSION software (McGaughey, 2010) respectively. The *cover* algorithm calculates canopy cover as the number of first returns over a specified height threshold divided by the total number of first returns within each cell. The purpose of the lidar data was to demonstrate the ability of lidar to represent field conditions in this vegetation system and calibrate canopy cover as explained below. Airborne lidar data was an accurate representation of field conditions as evidenced by the significant ($p<0.001$) and very high correlation ($R^2 = 0.95$) between field measured and lidar derived maximum canopy height (Figure 3.1). Such a high correlation between variables that can be directly validated indicates the reliability of other lidar derived variables that could be used in similar work

We selected height metrics that have been identified as readily derivable from spaceborne lidar data based on our previous work. These included Lorey's height (Lefsky, 2010) and the maximum and 90th percentile canopy height (Gwenzi and Lefsky, 2014). Lorey's height was calculated as the mean height of all trees weighted by their basal area:

$$H_{lorey} = \frac{\sum G_i H_i}{\sum G_i} \qquad (3)$$

where G_i and H_i are the basal area and height of tree *i* respectively.

We used field measurements of these height indices to predict aboveground biomass. An alternative approach would have been to estimate the height indices from lidar remote sensing using the field measurements as a calibration dataset but we did not have large footprint lidar data available for this area. Since our main aim was to investigate the canopy height-biomass relationship at a large scale and the implications of using generalized instead of composition-specific models, our discussion mainly focused on plot level model performance treating the 50 m plot as analogous to an ICESat-GLAS footprint. We have confidence that we can estimate these height indices from spaceborne lidar as we have demonstrated in the above mentioned previous work thus conclusions about height-biomass relationships reached using field data should be equally valid for height metrics derived from lidar. The idea is that the height-biomass relationship derived here can be taken and used in areas that have ICESat-GLAS footprints or any future large footprint spaceborne lidar satellite data in a global/larger scale effort.

Small scale studies have used discrete return lidar data to calculate canopy cover (Colgan *et al*., 2012; Colgan *et al*., 2013; Nyström *et al*., 2012). However with our aim of using more affordable and globally available data sets, we opted for a proxy of canopy cover derivable from Landsat TM imagery. Previous studies have demonstrated a strong relationship between Normalized Difference Vegetation Index (NDVI) and canopy cover (Gamon *et al*., 1995; Todd & Hoffer, 1998). To confirm this relationship in our study side we compared NDVI derived from Landsat TM imagery to canopy cover derived from discrete return lidar data. In the study site, trees green up in summer, while grasses are gray/dead so a summer image was most ideal for computing NDVI to be used as an indicator of tree crown cover. A July 27, 2011 Landsat 5 TM image was acquired from the United States Geological Survey (USGS) distribution site (http://glovis.usgs.gov/). Atmospheric correction was done using the Quick Atmospheric Correction (QUAC) module of the Environment for Visualizing Images (ENVI) software (Bernstein *et al*., 2005). We then computed NDVI on the atmospherically corrected image, resampled to 50 m to match the plot size. Canopy cover was found to be significantly correlated to NDVI ($p<0.001$; $r^2 = 0.65$, see Figure 3.2) derived from the 2011 Landsat TM image mentioned above. We therefore relied on this empirical relationship and used NDVI as a proxy for canopy cover in subsequent modeling.

2.3. Data Analysis

The aim of this work was to derive composition-specific and global models to estimate biomass from canopy height for a plot comparable in size to the resolution of large footprint spaceborne lidar waveforms. We assigned each plot to one of the four main composition classes (blue oaks, black oaks, valley oaks and mixed) so as to investigate the potential loss of accuracy when generalized instead of composition-specific models are used. The Hierarchical Bayesian modeling involved the derivation of parameters at both the composition and global levels. The performance of the generalized model compared to the different composition-specific models was used as an indicator of how generalizable it can be. Such a detailed analysis at a scale where direct measurements are possible can develop principles to be used at regional to global scales where direct validation is difficult (Waring & Running, 2007).

2.3.1. Frequentist generalized modeling

We did simple least squares regression modeling to relate the field measurements of the height metrics to aboveground biomass at the plot level. We developed a global model for each height metric by fitting linear, power, exponential, logarithmic and polynomial equations to the data and picked the best for each (equations 4 - 7). The best model was determined by the training and leave-one-out cross validation R^2 value (R. Kohavi, 1995) as well as the Root Mean Square Error and bias. We initially tested the relationship between maximum canopy height (H_{max}) and aboveground biomass. Next, we adjusted the model to include NDVI (as a surrogate for canopy cover) to account for differences in plot stem density. We tried both the model with single variables and one with interaction terms. The interaction terms did not significantly improve the model (additional terms were not significant at the $\alpha = 0.05$ level and the adjusted R^2 became lower) therefore we dropped them. Finally we also tested two other models, one that estimated biomass from Lorey's height (H_{lorey}) and another that used the 90th percentile height (H_{90}). The final four models were:

$$Y = a + bx^2 \qquad \text{where } x = H_{max} \tag{4}$$

$$Y = a + bx^2 + cz \qquad \text{where } x = H_{max;}\ z = NDVI \tag{5}$$

$$Y = a + bx \qquad \text{where } x = H_{lorey} \tag{6}$$

$$Y = a + bx \qquad \text{where } x = H_{90} \tag{7}$$

For a better comparison with other studies, relative RMSE and bias were also calculated and reported as percentages of the mean observed biomass. We also partitioned the bias to determine if it was consistent at both low and high biomass plots. This was achieved by calculating the bias on plots below and above the median plot biomass and then comparing the two.

2.3.2. Hierarchical Bayesian Analysis (HBA): Generalized vs composition-specific modeling

We implemented a HBA to fit the four different models (equations 4-7), each to the same data set. MCMC methods as implemented in the JAGS (Plummer, 2003) program within R were used to estimate the posterior distributions of each parameter. From each chain we obtained 1 000 000 iterations (samples) after discarding the initial 200 000 as burn-in. Convergence of chains was assessed using the Heidelberger diagnostic tool within the "coda" package of R (Heidelberger & Welch 1983; Plummer *et al*. 2006). Since the 4 composition classes have different physical structures, both the intercepts and the slopes were varied and our JAGS model considered the possible correlation between the intercept and the slope of each model as explained by Gelman & Hill (2009, Chapter 17). We used objective (non-informative) priors since we didn't have any previous similar work from which we could derive subjective priors.

Non-informative priors have minimal impact on the posterior distribution and are preferred as a more objective starting point when there is minimal or no knowledge about the prior conditions (Berger, 2006; Goldstein, 2006; Gelman & Hill, 2009).

To compare composition-specific and global model parameters, we modeled composition-specific parameters as coming from a global (overall) population that is defined by population level parameters:

$$\bar{a}_{Y,s} \sim N(A_Y,\sigma^2_{AY}) \; ; \; \bar{b}_{Y,s} \sim N(B_Y,\sigma^2_{BY}) \; ; \; \bar{c}_{Y,s} \sim N(C_Y,\sigma^2_{CY}) \qquad (8)$$

where $\bar{a}_{Y,s}$; $\bar{b}_{Y,s}$ and $\bar{c}_{Y,s}$ are composition-specific parameters and the variance terms σ^2_{AY}; σ^2_{BY} and σ^2_{CY} describe the variability among the composition classes for the three parameters in the models to which they are applicable. We used posterior predictive checks to assess model goodness of fit for both composition-specific and global models. A posterior predictive check is a comparison between the replicated dataset as simulated from the model and the dataset that was used to estimate parameters (Gelman & Hill, 2009). We used a statistic from the replicated data (T^{rep}) and an identical test statistic from the observed data (T^{obs}) to test for lack of fit by calculating P_B, the probability that the replicated data is more extreme than the real data (Gelman & Hill, 2009; Hobbs *et al*. 2012). A P_B value close to 0 or 1 indicates a failure of the distribution of simulated data to mimic the distribution of the observed data (lack of fit) and values close to 0.5 indicate a strong fit (Gelman *et al*. 2004). P_B was broken down into two, P_B^{mean} that measures the ability of the model to capture the mean tendency of the data and P_B^{var} that measures the ability of the model to portray the variation in the data (Hobbs *et al*. 2012). We used equation 9 and 10 to compute the test statistics for P_B^{mean} and P_B^{var} respectively

$$T^{obs} = \frac{\sum_{i=1}^{N} Y_i^{obs}}{N} \quad , \quad T^{rep} = \frac{\sum_{i=1}^{N} Y_i^{rep}}{N} \tag{9}$$

$$T^{obs} = \sum_{i=1}^{N} \frac{\left(Y_i^{obs} - \mu_i\right)^2}{\mu_i}, \quad T^{rep} = \sum_{i=1}^{N} \frac{\left(Y_i^{rep} - \mu_i\right)^2}{\mu_i} \tag{10}$$

where Y_i^{obs} is the observed data, Y_i^{rep} is the replicated data and μ_i is the model estimation for biomass.

To assess the generalizability of a model, we compared the parameters of each model among the four different composition classes by examining the overlap of their 95% Credible Intervals (CIs) and their mean values. Credible intervals are analogous to confidence intervals in frequentist statistics but they incorporate information from the prior distribution into the estimate. A 95% credible interval will therefore be one in which given the data and the model, there is a 95% chance the unknown parameter lies in that interval (Robert, 2007). Additionally we calculated the densities of the differences in the posterior distributions of the composition-specific and global parameters. The distribution of the posterior densities of these differences among composition classes and between each composition class and the global value were then used as an indicator of the generalizability of the model. For two parameters being compared, a density distribution with a mean difference of zero means no difference while a distribution that does not capture zero at all means a complete difference. To aid in interpretation, we also calculated P_{diff}, the two tailed probability that the difference between the values of two parameters being compared is greater than zero. If we are comparing a global parameter *G* and 3 composition-specific parameters (*Sp1*, *Sp2* and *Sp3*), as shown in figure 3.3, the P_{diff} concerning *Sp1* (50%) means there is a strong support for the use of the global parameter. The P_{diff}s concerning *Sp2* and *Sp3* are close to 0 and 100 percent respectively meaning there is a very weak support for the use of the global parameter. In this case, there is always a high probability that *Sp2* is less than the global parameter and *Sp3* is greater than the global parameter.

3. Results

3.1. Frequentist generalized modeling

All four models gave good estimations of biomass, with the H_{max}+NDVI model (equation 6) being the best ($R^2 = 0.75$; RMSE = 30% of the mean), see Table 3.1. Overall, all the models had a mean bias of 0 Mg/Ha. However, using the median as a cut-off point to separate the observations into two groups showed that the bias was not random but all of the models had a tendency to overestimate the low biomass plots while underestimating the higher biomass plots. The magnitude of the relative bias was higher for low biomass plots than for high biomass plots mainly because of a higher number of low biomass plots characterizing this area.

3.2. Bayesian modeling

Posterior predictive checks for each model (equation 4-7), whether composition-specific or overall gave P_B^{mean} and P_B^{var} ranging from 0.43 to 0.58, meaning the models were not only capable of replicating the mean tendency of the data but they also replicated well the variability of the data. The posterior distributions of the parameters for the 4 models are shown in figures 3.4 – 3.7. Part A of the figures show the mean values for the parameters and the associated 95% credible intervals. Part B shows how these parameters differ between composition classes while part C shows how the composition-specific parameters differ from the global parameters. There was a consistent pattern of intercept and slope values for each composition class. The valley oak class consistently had the highest values while the blue oak class consistently had the lowest. However, for all the models, the 95% CIs overlapped the mean values, and the composition-specific mean values were not very different from the global mean values (especially for the H_{max} + NDVI model). This suggests that although composition-specific models would be more accurate, there is a high support for the use of global parameters as shown by the posterior differences curves with P_{diff} values of mostly between 25 and 75 %, implying that the models are generalizable. This trend was however less pronounced for the H_{max} model (equation 4) compared with the rest.

4. Discussion and conclusions

The heterogeneity associated with open canopy systems suggests the use of structure parameters that combine height and at least one gap fraction measure to reliably estimate volumetric attributes such as aboveground biomass. The importance of a canopy cover proxy demonstrated in this work is a confirmation of previous studies in savannas or other open canopy areas (Colgan *et al*., 2012; Nyström *et al*., 2012). Whereas those studies used canopy cover as derived by high density airborne discrete return lidar data, we used NDVI as a proxy so we could evaluate the utility of using data that we can easily obtain over larger scales from space. In our study area, trees and grasses green up in different seasons hence an image captured in the trees' green season will be a good indication of tree cover. Most savannas, especially in northern temperate areas have this type of phenology which makes our method applicable to other areas. However, in areas where trees and grasses green up at the same time, this empirical relationship may not work well. Where NDVI is not adequate, other indices that describe vegetation type can also be derived from passive optical remote sensing imagery and used to improve canopy height-biomass modeling (Ni-Meister *et al*., 2010). Another alternative will be to use NDVI and texture to derive an index that identifies well the contribution of trees.

Canopy cover information can also be incorporated using height metrics that consider stem density such as Lorey's height or percentile height metrics. The 90th percentile canopy height and Lorey's height allow the larger trees to contribute more to the mean (Naesset, 1997). Basal area or crown area defines the area of the stand that is occupied by trees. Larger stem or crown diameter trees contribute more to the total plot biomass thus the use of Lorey's height in estimating biomass also takes advantage of this fact. The problem is that these metrics are not readily obtainable from spaceborne lidar data alone. There are models developed to derive Lorey's height from spaceborne lidar waveforms but these are based on field plots from forest sites (Lefsky, 2010) and have been shown to reduce accuracy of biomass estimation even when applied to related forests (Mitchard *et al*., 2011). Although there are other several metrics previously derived and used e.g quadratic mean canopy height, height of median energy, canopy height profile, canopy volume method , vertical canopy profile (Lefsky *et al*. 1999; Means *et al*. 1999; Drake *et al*. 2002; Boudreau *et al*. 2008), maximum canopy height is most easily defined in terms of geometry and more easily identified in lidar data, especially in regions with variable terrain and high slopes. Broadleaved savanna ecosystem trees mostly have flat and wide canopies therefore the probability of identifying the canopy top is always high as compared to for example coniferous forests with narrow canopy tips. Since maximum canopy height and NDVI performed as well as the more complicated height indices, the H_{max}+NDVI model (Equation 5) would be best for regional to global applications, provided the tree and grass NDVI contribution can be separated.

Composition-specific parameters were distinctively different with the valley oak class consistently having the highest values while the blue oak class had the lowest. This was consistent with field observations that valley oaks have big stem and crown diameters therefore even a few trees in a valley oak plot would have wide spreading and interlocking braches, giving a high canopy cover and biomass from its crown compared to many small diameter trees in blue oak stands whose biomass is concentrated in the stems. Fitting a generalized model averaged out these differences. The distribution of the posterior densities of the differences between the composition level and global level parameter values coupled with the overlap of the parameters' 95% CIs and the mean values for the composition level and global level models show a universal applicability (generalizability) of the models. The higher differences between composition-specific and global parameters for equation 4 are most likely to be a result of the fact that the model did not have any information on the horizontal structure of the plots. With the other models, NDVI (equation 5), basal area weighting (equation 6) and use of the 90th percentile height (equation 7) provided extra information about structure and hence adjusted the parameters accordingly when a generalized model is fit. Fitting a composition-specific model would be best for a local and homogeneous scale but in large scale studies, fitting nested composition-specific models within a large landscape is not practically meaningful.

We recommend the use of the global parameters of the intercept and slope coefficients in large scale studies regardless of composition since the HBA results showed a logical generalizability of the models. The narrow range of the distribution and the overlap of the CIs and mean values of the composition-specific and global parameters suggest that scaling rules do exist for savannas. This has also been demonstrated at tree level (Tredennick *et al*., 2013; Zapata-Cuartas *et al*., 2012). As Zapata-Cuartas *et al*., (2012) suggest, and from the results of our Bayesian analysis, we also conclude that the model parameters are better represented as probability distributions rather than as constant values. For small scale studies in sites that do not have local equations, this distribution can be applied as priors to develop new models.

5. Tables and Figures

Table 3.1: Generalized frequentist regression modeling: leave one out cross validation results

Model	**Goodness of fit statistics**						
	R^2	**RMSE**		**Bias**			
		Absolute Mg/Ha	Relative (%)	Absolute, Below median (Mg/Ha)	Relative, below median (%)	Absolute, Above median (Mg/Ha)	Relative, above median (%)
H_{max}	0.68	81.10	34	24.51	21	-20.60	6
$H_{max}+NDVI$	0.75	70.85	30	12.08	10	-9.01	3
H_{lorey}	0.73	73.65	31	16.26	14	-15.18	4
H_{90}	0.71	77.11	33	14.56	12	-12.46	3

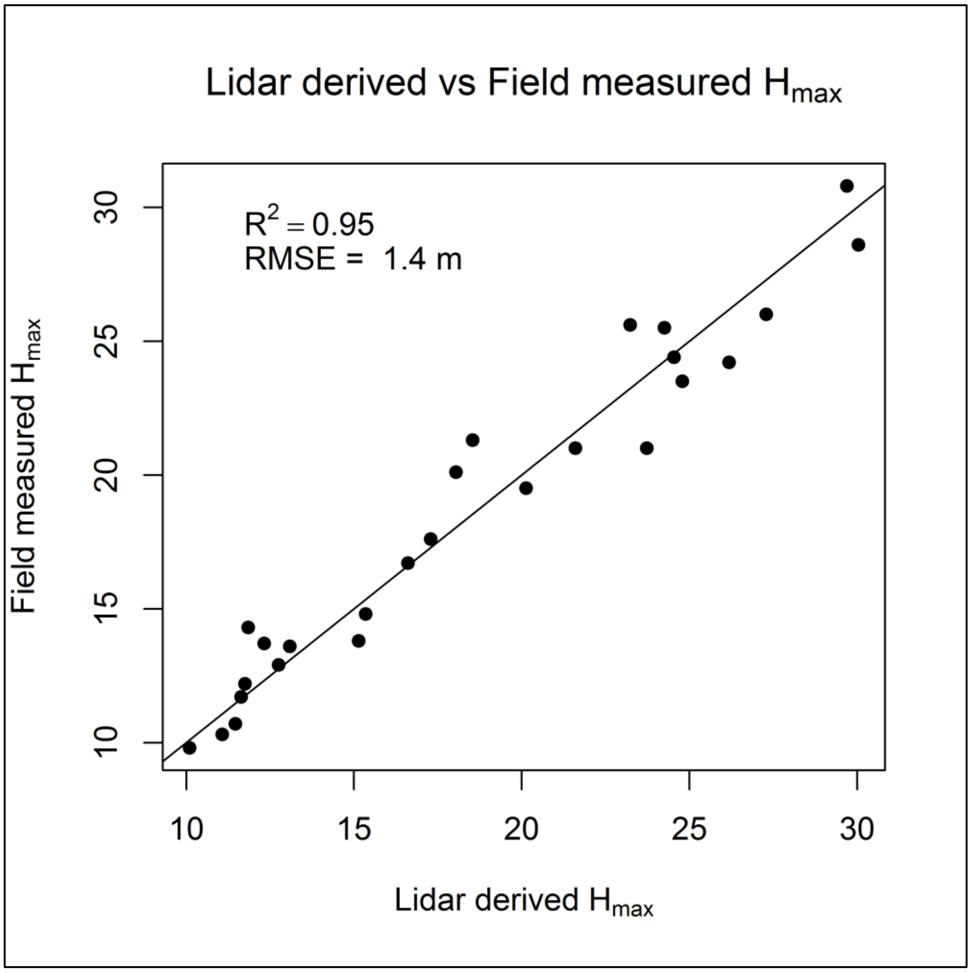


Figure 3.1: The high correlation between lidar derived and field measured maximum canopy height

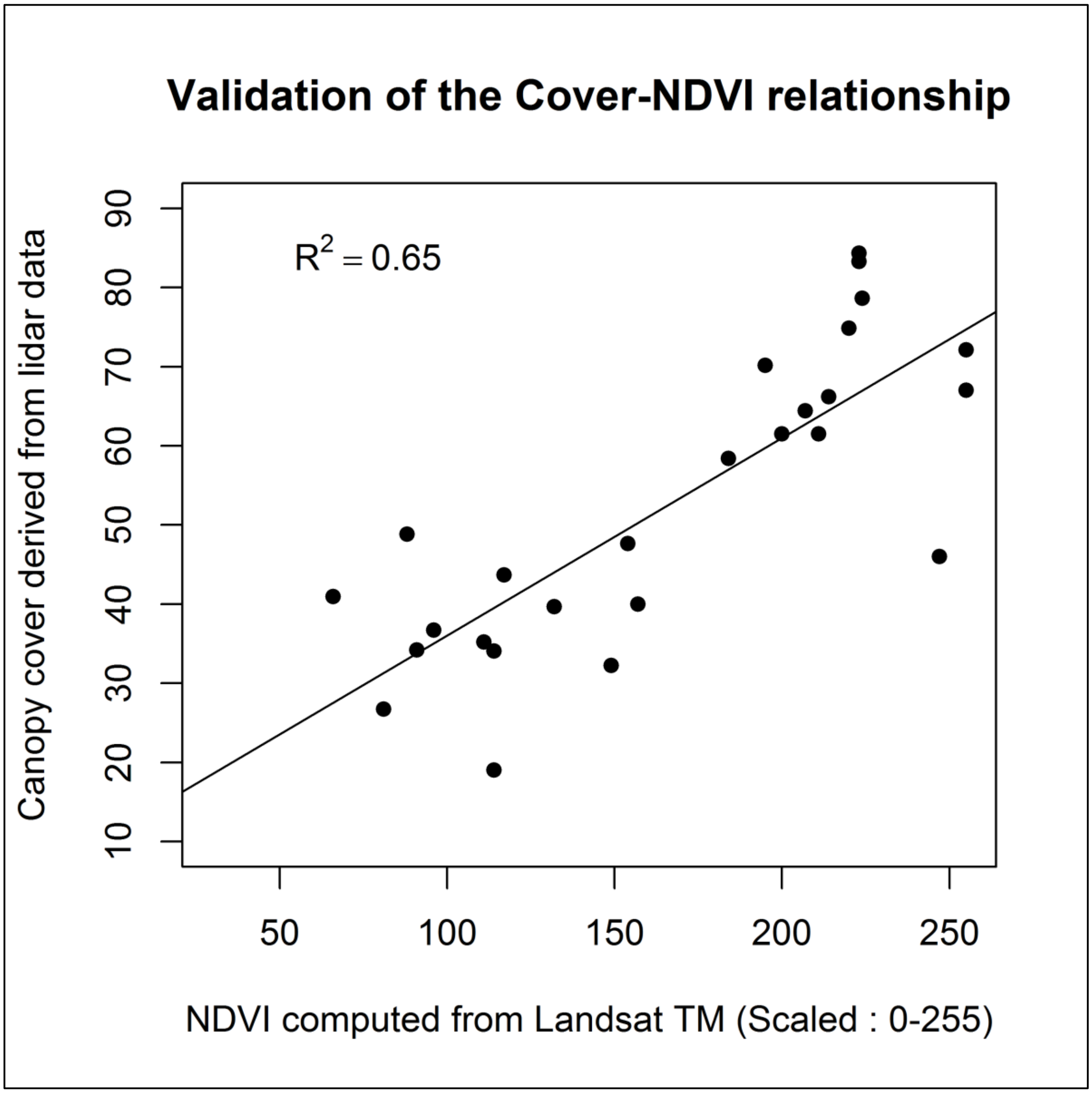


Figure 3.2: Canopy cover-NDVI relationship in the 26 plots enumerated

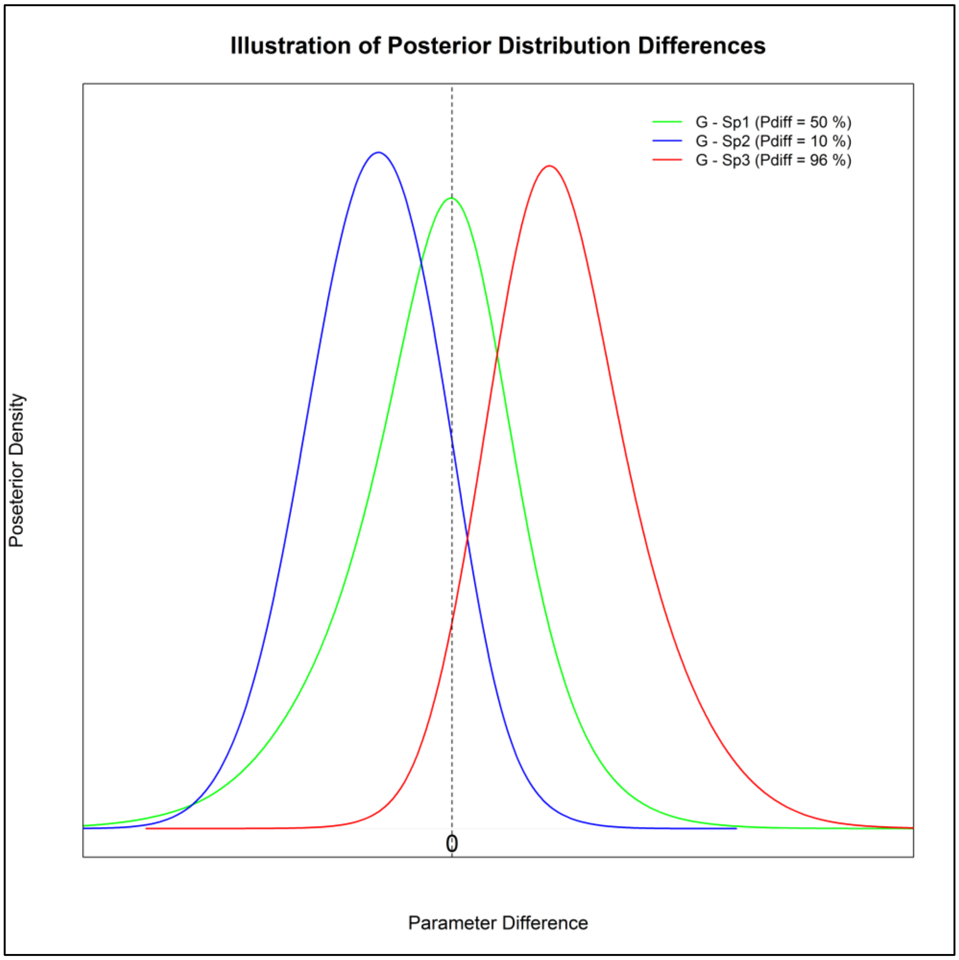


Figure 3.3: Illustration of the posterior distribution differences used to assess model generalizability

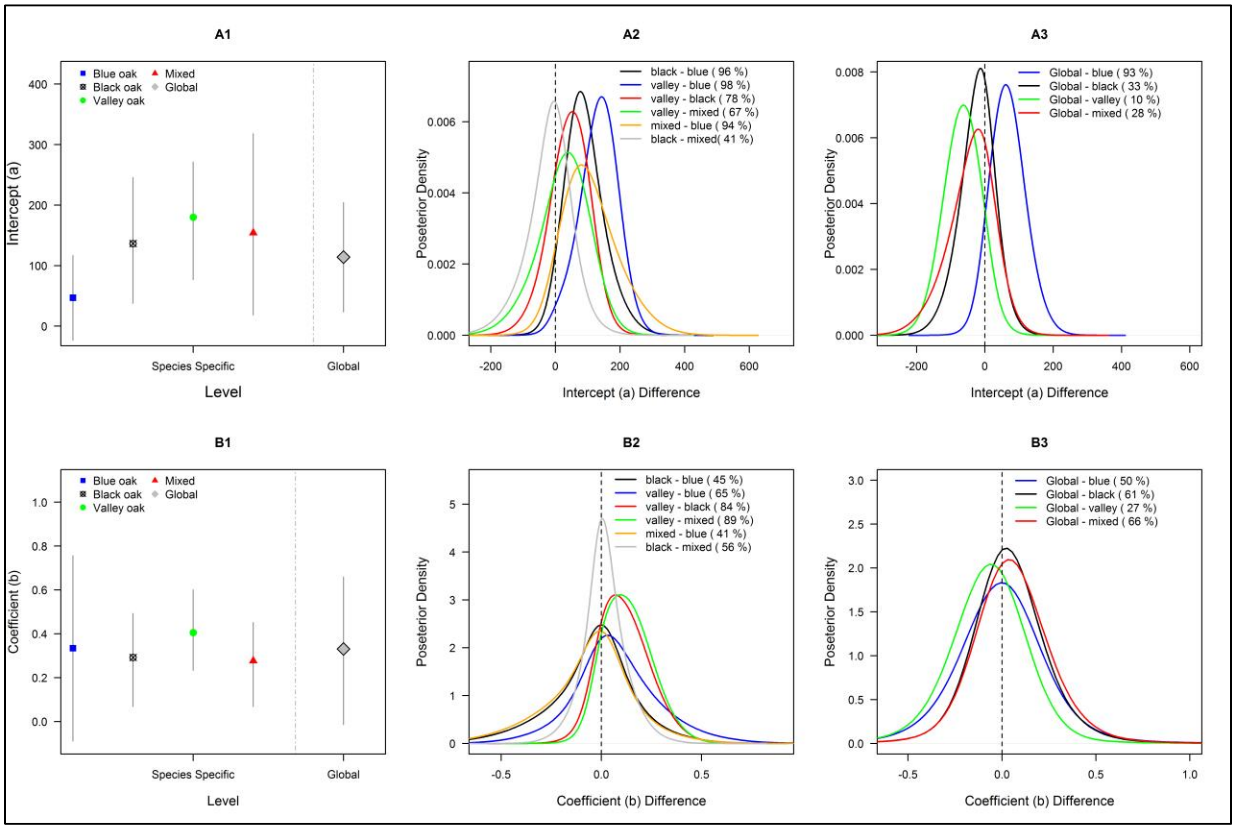


Figure 3.4: Posterior predictive checks for the H_{max} (equation 4) model

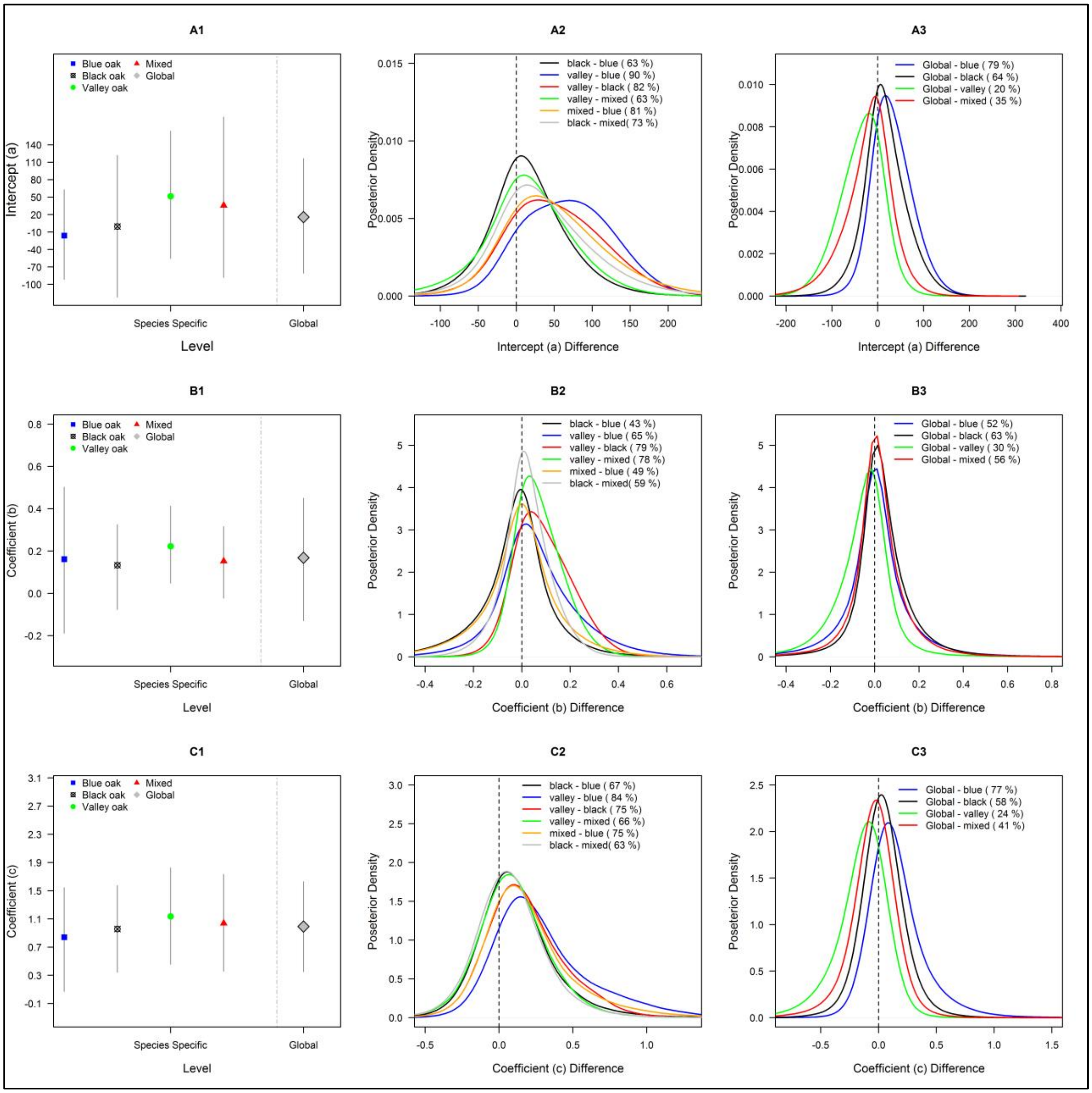


Figure 3.5: Posterior predictive checks for the H_{max} + NDVI (equation 5) model

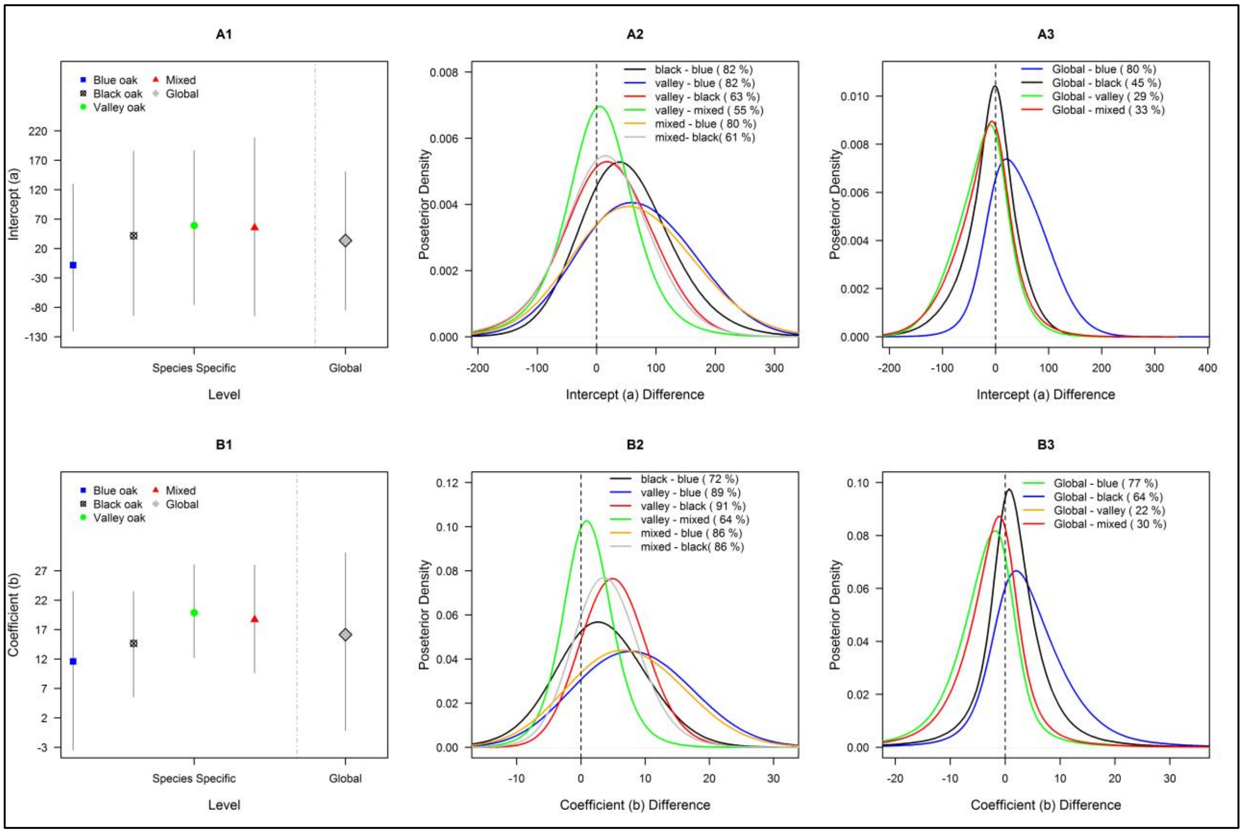


Figure 3.6: Posterior predictive checks for the H_{lorey} (equation 6) model

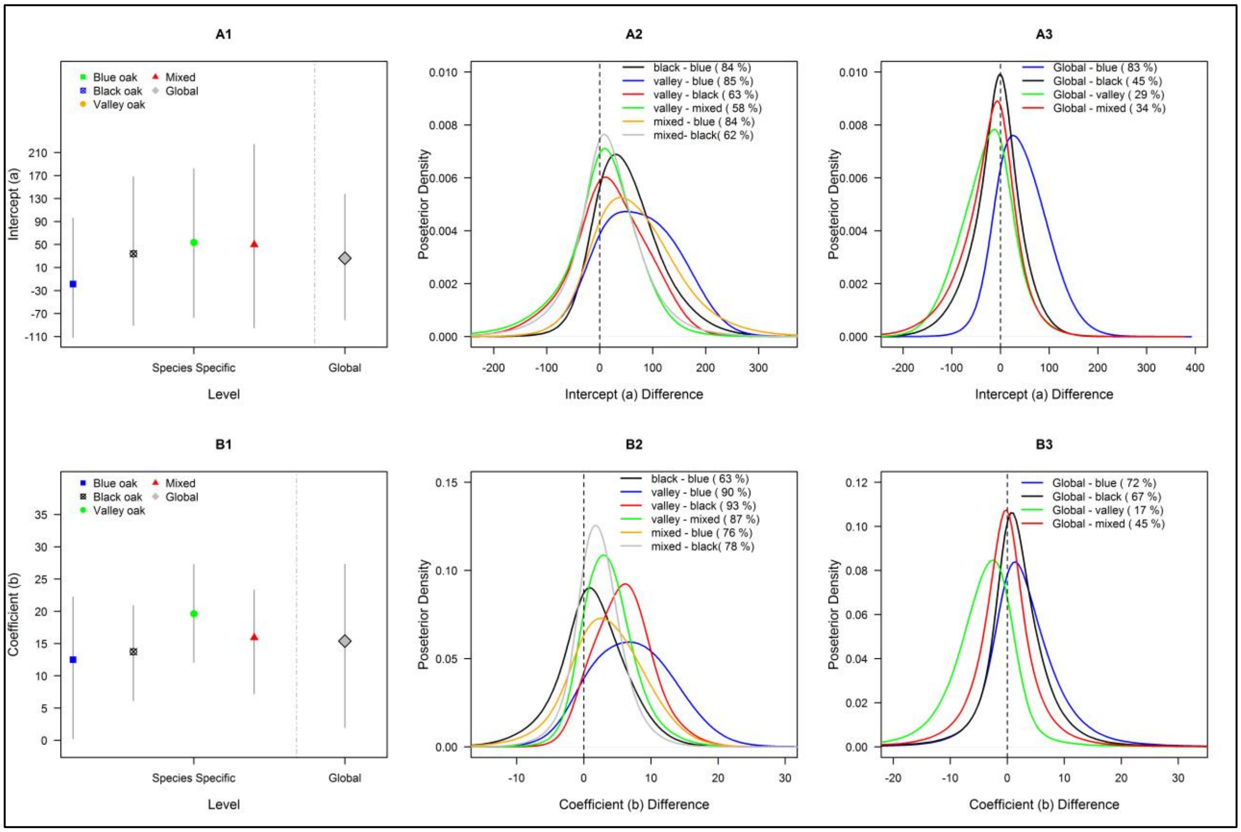


Figure 3.7: Posterior predictive checks for the H_{90} (equation 7) model

CHAPTER 3 REFERENCES

Berger, J. (2006). The Case for Objective Bayesian Analysis. *Bayesian Analysis*, *1*(3).

Bernstein, L. S., Adler-Golden, S. M., Sundberg, R. L., Levine, R. Y., Perkins, T. C., Berk, A., … Hoke, M. (2005). Validation of the QUick Atmospheric Correction (QUAC) algorithm for VNIR-SWIR multi- and hyperspectral imagery. In *SPIE Proceedings, Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XI Vol. 5806* (pp. 668–678). SPIE Digital Library.

Boudreau, J., Nelson, R., Margolis, H., Beaudoin, A., Guindon, L., & Kimes, D. (2008). Regional aboveground forest biomass using airborne and spaceborne LiDAR in Québec. *Remote Sensing of Environment*, *112*(10), 3876–3890.

Brooks, S., Gelman, A., Jones, G., & Meng, X. (2011). *Handbook of Markov Chain Monte Carlo*. Boca Raton, USA: Taylor and Francis Group.

Carlin, B. P., & Thomas, A. L. (2000). *Bayes and Empirical Bayes Methods for Data Analysis* (Second). London, UK: Chapman & Hall.

Chave, J., Andalo, C., Brown, S., Cairns, M. a, Chambers, J. Q., Eamus, D., … Yamakura, T. (2005). Tree allometry and improved estimation of carbon stocks and balance in tropical forests. *Oecologia*, *145*(1), 87–99.

Colgan, M. S., Asner, G. P., Levick, S. R., Martin, R. E., & Chadwick, O. a. (2012). Topo-edaphic controls over woody plant biomass in South African savannas. *Biogeosciences Discussions*, *9*(1), 957–987.

Colgan, M. S., Asner, G. P., & Swemmer, T. (2013). Harvesting tree biomass at the stand level to assess the accuracy of field and airborne biomass estimation in savannas. *Ecological Applications*, *23*(5), 1170–84.

Drake, J. B., Dubayah, R. O., Clark, D. B., Knox, R. G., Blair, J. B., Hofton, M. A., … Prince, S. D. (2002). Estimation of tropical forest structural characteristics using large-footprint lidar. *Remote Sensing of Environment*, *79*(2-3), 305–319.

Drake, J. B., Dubayah, R. O., Knox, R. G., Clark, D. B., & Blair, J. B. (2002). Sensitivity of large-footprint lidar to canopy structure and biomass in a neotropical rainforest. *Remote Sensing of Environment*, *81*(2-3), 378–392.

Duncanson, L. I., Niemann, K. O., & Wulder, M. A. (2010). Estimating forest canopy height and terrain relief from GLAS waveform metrics. *Remote Sensing of Environment*, *114*(1), 138–154. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0034425709002697

Gamon, J. A., Field, C. B., Goulden, M. L., Griffin, K. L., Hartley, E., Joel, G., … Valentini, R. (1995). Relationships between NDVI, canopy structure and photosythesis in three Californian vegetation types. *Ecological Applications, 5*(1), 28–41.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (Third). London, UK: Chapman and Hall.

Gelman, A., & Hill, J. (2009). *Data analysis using regression and multilevel/ hierarchical modeling*. Cambridge, UK: Cambridge University Press.

Ghosh, J. K., Delampady, M., & Samanta, T. (2006). *An Introduction to Bayesian Analysis*. New York, USA: Springer.

Goldstein, M. (2006). Subjective Bayesian Analysis : Principles and Practice Applied subjectivism. *Bayesian Analysis*, *1*(3), 403–420.

Gwenzi, D., & Lefsky, M. A. (2014). Modeling canopy height in a savanna ecosystem using spaceborne lidar waveforms. *Remote Sensing of Environment*, *154*.

Hall, B. (2012). Bayesian Inference. https://datajobs.com/data-science-repo/Bayesian-Inference-[Statisticat].pdf

Heidelberger, P., & Welch, P. D. (1983). Simulation Run Length Control in the Presence of an Initial Transient. *Operations Research, 31*, 1109–1144.

Hobbs, N. T., Andren, H., Persson, J., Aronsson, M., & Chapron, G. (2012). Native predators reduce harvest of reindeer by Sa ´ mi pastoralists. *Ecological Applications*, *22*(5), 1640–1654.

Isenburg, M. (2015). LAStools: award-winning software for rapid LiDAR processing. Retrieved April 2, 2015, from http://www.cs.unc.edu/~isenburg/lastools/

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. New York, USA: Cambridge University Press.

Jenkins, J. C., Chojnacky, D. C., Heath, L. S., & Birdsey, R. A. (2004). *Comprehensive database of biomass equations for North American tree species. General Technical Report NE-319*. USDA Forest Service, Newton Square, PA. USA.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence 2 (12)* (pp. 1137–1143). Morgan Kaufmann, San Mateo, CA.

Lefsky, M. A. (2010). A global forest canopy height map from the Moderate Resolution Imaging Spectroradiometer and the Geoscience Laser Altimeter System. *Geophysical Research Letters*, *37*(15), 1–5.

Lefsky, M. A., Cohen, W. B., Acker, S. A., Parker, G. G., Spies, T. A., & Harding, D. (1999). Lidar Remote Sensing of the Canopy Structure and Biophysical Properties of Douglas-Fir Western Hemlock Forests. *Remote Sensing of Environment*, *70*(3), 339–361.

Lefsky, M. A., Harding, D., Cohen, W. B., Parker, G., & Shugart, H. H. (1999). Surface Lidar Remote Sensing of Basal Area and Biomass in Deciduous Forests of Eastern Maryland, USA. *Remote Sensing of Environment*, *67*(1), 83–98.

Lefsky, M. A., Harding, D. J., Keller, M., Cohen, W. B., Carabajal, C. C., Del Bom Espirito-Santo, F., … de Oliveira Jr, R. (2005). Estimates of forest canopy height and aboveground biomass using ICESat. *Geophysical Research Letters*, *32*(22), 1–4.

Lefsky, M. A., Keller, M., Pang, Y., Camargo, P. B. de, & Hunter, M. O. (2007). Revised method for forest canopy height estimation from Geoscience Laser Altimeter System waveforms. *Journal of Applied Remote Sensing*, *1*.

Lefsky, M., Cohen, W., Parker, G., & David Harding. (2002). Lidar Remote Sensing for Ecosystem Studies. *BioScience*, *52*(1), 19–30.

Litton, C. M., & Kauffman, B. J. (2008). Allometric Models for Predicting Aboveground Biomass in Two Widespread Woody Plants in Hawaii. *Biotropica*, *40*(3), 313–320.

McCarthy, M. A. (2007). *Bayesian Methods for Ecology*. Cambridge, UK: Cambridge University Press.

Means, J. E., Acker, S. A., Harding, D. J., Blair, J. B., Lefsky, M. A., Cohen, W. B., … McKee, W. A. (1999). Use of Large-Footprint Scanning Airborne Lidar To Estimate Forest Stand Characteristics in the Western Cascades of Oregon. *Remote Sensing of Environment*, *67*(3), 298–308.

Mitchard, E. T. a., Saatchi, S. S., White, L. J. T., Abernethy, K. a., Jeffery, K. J., Lewis, S. L., … Meir, P. (2011). Mapping tropical forest biomass with radar and spaceborne LiDAR: overcoming problems of high biomass and persistent cloud. *Biogeosciences Discussions*, *8*(4), 8781–8815.

Mwakalukwa, E. E., Meilby, H., & Treue, T. (2014). Volume and Aboveground Biomass Models for Dry Miombo Woodland in Tanzania. *International Journal of Forestry Research*, *2014*.

Naesset, E. (1997). Determination of mean tree height of forest stands using airborne laser scanner data. *ISPRS Journal of Photogrammetry and Remote Sensing*, *52*(2), 49–56.

Ni-Meister, W., Lee, S., Strahler, A. H., Woodcock, C. E., Schaaf, C., Yao, T., … Blair, J. B. (2010). Assessing general relationships between aboveground biomass and vegetation structure parameters for improved carbon estimate from lidar remote sensing. *Journal of Geophysical Research*, *115*.

Nyström, M., Holmgren, J., & Olsson, H. (2012). Prediction of tree biomass in the forest–tundra ecotone using airborne laser scanning. *Remote Sensing of Environment*, *123*, 271–279.

Pang, Y., Lefsky, M., Andersen, H.-E., Miller, M. E., & Sherrill, K. (2008). Validation of the ICEsat vegetation product using crown-area-weighted mean height derived using crown delineation with discrete return lidar data. *Canadian Journal of Remote Sensing*, *34*(S2), S471–S484.

Pilli, R., Anfodillo, T., & Carrer, M. (2006). Towards a functional and simplified allometry for estimating forest biomass. *Forest Ecology and Management*, *237*(1-3), 583–593.

Plummer, M. (2003). JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leish, & A. Zeileis (Eds.), *Proceedings of the 3rd International Workshop on Distributed Statistical Computing.* Vienna, Austria.

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence Diagnosis and Output Analysis for MCMC. *R News*, *6*, 7–11.

Robert, C. P. (2007). *The Bayesian Choice : From Decision-Theoretic Foundations to Computational Implementation* (Second). New York, USA: Springer Science+Business Media.

Robert, C. P., & Casella, G. (2004). *Monte Carlo Statistical Methods* (Second). New York, USA: Springer.

Robert J. McGaughey. (2010). FUSION/LDV: Software for LIDAR Data Analysis and Visualization. United States Department of Agriculture Forest Service Pacific Northwest Research Station.

Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2005). *WinBUGS User Manual, Version 2.10*. Cambridge, UK: MRC Biostatistics Unit.

Tejon Ranch Conservancy. (2013). Tejon Ranch Conservancy. Retrieved April 2, 2015, from http://www.tejonconservancy.org/

Todd, S. W., & Hoffer, R. M. (1998). Responses of Spectral Indices to Variations in Vegetation Cover and Soil Background. *Photogrammetric Engineering and Remote Sensing*, *64*(9), 915–921.

Tredennick, A. T., Bentley, L. P., & Hanan, N. P. (2013). Allometric convergence in savanna trees and implications for the use of plant scaling models in variable ecosystems. *PloS One*, *8*(3).

VanLaar, A., & Akca, A. (2007). *Forest Mensuration (Managing Forest Ecosystems)* (Second). Dordrecht, The Netherlands: Springer.

Wagenmakers, E.-J., Lee, M. D., Lodewyckx, T., & Iverson, G. J. (2008). Bayesian Versus Frequentist Inference. In H. Hoijtink, I. Klugkist, & P. A. Boelen (Eds.), *Bayesian evaluation of informative hypotheses* (pp. 181–207). New York, USA: Springer.

Waring, R. H., & Running, S. W. (2007). *Forest Ecosystems, Analysis at Multiple Scales* (Third). San Diego, California, US: Elsevier Academic Press.

Wu, J., Aardt, J. A. N. Van, Asner, G. P., Knapp, D., & Erasmus, B. F. N. (2009). LiDAR Waveform-based Woody and Foliar Biomass Estimation in Savanna Environments. In *Silvilaser 2009* (pp. 20–29). Texas, USA.

Xing, Y., de Gier, A., Zhang, J., & Wang, L. (2010). An improved method for estimating forest canopy height using ICESat-GLAS full waveform data over sloping terrain: A case study in Changbai mountains, China. *International Journal of Applied Earth Observation and Geoinformation*, *12*(5), 385–392.

Zapata-Cuartas, M., Sierra, C. a., & Alleman, L. (2012). Probability distribution of allometric coefficients and Bayesian estimation of aboveground tree biomass. *Forest Ecology and Management*, *277*, 173–179.

Zianis, D., & Mencuccini, M. (2004). On simplifying allometric analyses of forest biomass. *Forest Ecology and Management*, *187*(2-3), 311–332.

Zolkos, S. G., Goetz, S. J., & Dubayah, R. (2013). A meta-analysis of terrestrial aboveground biomass estimation using lidar remote sensing. *Remote Sensing of Environment*, *128*, 289–298.

CHAPTER 4: SPATIAL MODELING OF LIDAR-DERIVED WOODY BIOMASS ESTIMATES COLLECTED ALONG TRANSECTS IN A HETEROGENEOUS SAVANNA LANDSCAPE[3]

Synopsis

Transects of lidar waveforms from the Ice Cloud and land Elevation Satellite's Geoscience Laser Altimeter System (ICESat-GLAS) have shown a capability to estimate canopy height and footprint level aboveground biomass even in structurally complex savanna ecosystems. However, for decision making at the landscape level, wall-to-wall maps are preferred since they are more easily integrated with other geospatial data sources. In this work we employed a data fusion approach with deterministic and geostatistical methods to spatially map the variability of aboveground woody biomass across a 11 800 Ha Oak savanna landscape in Santa Clara county, California, USA. We evaluated and compared the utility of inverse distance weighting, cokriging, regression kriging and image segmentation methods to create a wall-to-wall map from footprint level biomass. The 4 methods estimated biomass at independent validation sites with between 39% (inverse distance weighting) and 66% (image segmentation) of variance explained and Root Mean Square Error of 42% and 32% of the mean respectively. Image segmentation results indicated that when more waveforms were available to characterize patch biomass, 78% of variance in biomass was explained (RMSE = 21% of the mean). Overall, the mean pixel biomass predicted by the 4 methods did not differ significantly but the output maps showed

[3] David Gwenzi and Michael Andrew Lefsky, 2015
Submitted to the International Journal of Applied Earth Observation and Geoinformation

marked differences in the estimation precision and ability of each model to mimic the primary variable's trend across the landscape. We conclude that ICESat-GLAS or similar transect-sampling lidar data can be used to create wall-to-wall biomass maps in savannas but the methods work better with a higher sampling intensity and more informative correlated secondary data so as to reproduce the variability of the primary variable across the landscape. We recommend that future satellite lidar missions increase the sampling intensity across track so that biomass observations are made and characterized at the scale at which they vary.

Key words: Savanna; woody biomass; deterministic model; spatial autocorrelation; geostatistics; image segmentation; ICESat-GLAS

1. Introduction

Waveforms from the Ice Cloud and land Elevation Satellite's Geoscience Laser Altimeter System (ICESat-GLAS) have shown a capability to estimate canopy height even in structurally complex savanna ecosystems (Gwenzi & Lefsky, 2014). This canopy height can be used to estimate plot (footprint) level biomass (Lefsky *et al*., 1999; 2005; Gwenzi & Lefsky, 2015). While transect data is sufficient for many scientific studies, wall-to-wall maps are preferred for decision making as they can be applied at all points on a landscape (Turner, 1989; Levin, 1992). Spatial modeling methods are required to derive spatially continuous maps from transect based data such as that from ICESat-GLAS or future missions like the Global Ecosystem Dynamics Investigation (GEDI) lidar (http://eospso.gsfc.nasa.gov/mission-category/55). These kinds of maps can also be used to derive information about how wide scale gradients such as total precipitation and dry season length drive regional biomass distribution, and significantly improve our ability to estimate the carbon flux resulting from land-use change (Chambers *et al*., 2007).

The utility of mapping aboveground biomass by fusing transect based lidar data and other data sources that have a complete horizontal coverage has been tried mostly in forests (Sales *et al*., 2007; Boudreau *et al*., 2008; Nelson *et al*., 2009; Saatchi *et al*., 2011; Mitchard *et al*., 2012) . The main approaches used have been 1) deterministic methods, 2) geostatistical methods and 3) image segmentation followed by regression analysis using proxies of remote sensing and other environmental covariates in distribution models. In the latter two approaches lidar data is used to capture the 3 dimensional structure of vegetation at plot level and other auxiliary data sources are used to provide a complete two dimensional coverage. The wall-to-wall output not only gives the mean or total quantities but also shows the variability of the biomass across the landscape under study. Commonly used auxiliary data include radar backscatter, spectral indices derived from optical passive remote sensing imagery and topography indices derived from digital elevation models (DEM).

1.1. Deterministic methods

This approach relies explicitly on the first “law” of geography (Tobler, 1970) which states that “everything is related to everything else, but near things are more related than distant things." Interpolation/extrapolation of field measured values to larger areas is done by employing methods such as moving average, tessellation and inverse distance weighting (IDW). All of these methods are “deterministic” in the sense that there is no statistical estimation of parameters used in the method. Surfaces are created from measured points using mathematical functions, based on either the extent of similarity or the degree of smoothing in the data. Deterministic methods are relatively fast since they use algorithms that require fewer assumptions and input parameters. The problem with these aforementioned methods is that they lack the explicit spatial information of the distribution of the concerned variable which can be addressed by the use of geostatistics.

1.2. Geostatistical methods

This group of methods involve decomposing the unknown value (*z*) at any location into a mean component (*m*) and a residual component (*s*). The variability of the mean component determines the method to use in the subsequent modeling (Goovaerts, 1997). If the mean is assumed to be constant (an unusual case), a stationarity based method known as ordinary kriging (OK) is the best to use. When the mean is spatially variable (the more common case), then it is modeled by expressing it as a function of auxiliary variables using non-stationary methods that belong to a group called hybrid geostatistical methods (McBratney *et al*., 2000). The auxiliary variables could be any that explain the distribution of the response variable of interest for example remote sensing derived vegetation spectral indices, soil type, forest type etc. for aboveground biomass.

These hybrid geostatistical methods are classified into two main groups depending on the properties of the input data (Hengl *et al*., 2003). In the first (cokriging), estimates are made using the spatial correlation of the primary variable with itself, spatial correlation of a secondary wall-to-wall variable with itself, and a “cross-correlogram” that describes the cross correlation between the primary and secondary variable. The second group is referred to variously as kriging with unknown mean, kriging with a trend model or kriging with external drift but at least 3 different approaches are recognized: 1) universal kriging (UK) where the trend is modeled as a function of coordinates, 2) kriging with external drift (KED) where the drift is defined externally by auxiliary variables instead of the coordinates and 3) regression kriging (RK) where the drift and residuals are fitted separately and then summed.

1.3. Image segmentation

This approach involves delineating patches within an image of the study area based on its spectral and/or textural qualities. After image segmentation, each of the resulting polygons is attributed with measures of centrality or spread (such as mean, range and standard deviation) of the auxiliary variable layers. These polygon statistics are then used as independent variables in modeling the biomass (Mitchard *et al*., 2012). For small scale studies and where financial resources are not limiting, high spectral and or spatial resolution data has been used to derive the covariates (Cho *et al*., 2011; Cho *et al*., 2012; Naidoo *et al*., 2012). This approach would be suitable for savannas also because high spectral resolution and high spatial resolution images will capture the irregularity of the vegetation that results from marked differences in biotic and abiotic factors such as topography, rainfall, herbivory and human impacts. Hyperspectral imagery is good in species discrimination while high spatial resolution imagery can clearly differentiate the tree and grass segments of the landscape at finer scales. However, for large extents, use of such data is at present prohibitively expensive to acquire and methods based on moderate to low resolution data that are freely available are more appropriate. At regional to global level other ecologically related layers such as ecoregion classifications can be used in attributing the polygons (M A. Lefsky, 2010; Mitchard et al., 2012)

Although generally more accurate, the geostatistical and image segmentation approaches are more sophisticated and the necessity of employing such complicated processes depends on the purpose of the modeling. The aim of this work was to use ICESat-GLAS footprint level biomass in combination with wall-to-wall remotely sensed data and topographically derived variables to estimate the woody biomass values at unsampled points and consequently its spatial distribution across the landscape. The secondary objective was to compare the accuracy of the 3 earlier mentioned groups of approaches for this kind of work to come up with recommendations for large extent and global mapping efforts. Our hypothesis was that these different approaches would give slightly different results but the intensive process of the geostatistics approach would give a more insightful solution that provides a better understanding of the ecosystem compared to the other two approaches. Where not specified, the word biomass in this paper refers to aboveground woody biomass as estimated in Gwenzi & Lefsky (2015).

2. Materials and methods

2.1. Study Site

We used data from the same site and ICESat-GLAS footprints reported in our previous work (Gwenzi & Lefsky, 2014). These are the Santa Clara Oak Savannas located in California, USA (Figure 4.1). For details of the vegetation and terrain structure, we refer readers to the above mentioned paper.

2.2. Data

2.2.1. Raster data

A Landsat 5 TM image collected on 9 June, 2006 was acquired from the USGS distribution site (http://earthexplorer.usgs.gov/). We chose a scene from the summer since it is the time when trees green up and the year 2006 was chosen to match approximately with the time the ICESat-GLAS waveforms and other lidar data used in Gwenzi & Lefsky (2014) were collected (2003-2006). Atmospheric correction was done on the image using the Quick Atmospheric Correction (QUAC) module of the Environment for Visualizing Images (ENVI) software (Bernstein *et al.*, 2005). For radar, we used the Phased Array type L- band Synthetic Aperture Radar (PALSAR) data collected by the Advanced Land Observing Satellite (ALOS). Two Level 4.1 scenes from June 2007 were obtained from the Japanese Space System's integrated ASTER/PALSAR distribution site (https://ims.ersdac.jspacesystems.or.jp) and mosaicked. We could not find a 2006 scene for our study site so we assumed that the vegetation changes within the 4 years range (2003-2007) of our data acquisition times would not significantly influence the results. Both the horizontal transmit, horizontal receive (HH) and horizontal transmit, vertical receive (HV) polarization data in sigma nought units were used for this work. The spatial resolution of remote sensing analyses should be equal to or at least not finer than the size of the field plots used to calibrate the remote sensing (Naesset, 2002; Gonzalez *et al.*, 2010). Since most of our GLAS waveforms were from laser 3 observation period with a nominal footprint diameter of 55 m, we did our analyses at the 55 m pixel size and resampled all the remote sensing data accordingly. A Shuttle Radar Topography Mission DEM was obtained from the above mentioned USGS site. From the DEM we obtained 3 topography related variables i.e. absolute elevation in meters, slope in percent and aspect in degrees.

2.2.2. Footprint level biomass

We calculated footprint level biomass using the model developed for canopy height-biomass modeling in our previous work (Gwenzi & Lefsky, 2015) in a nearby oak savanna site, the Tejon ranch conservancy in Kern County, California. For each footprint, biomass was calculated in Mg/Ha units as a function of maximum canopy height and NDVI using the following model:

$$Biomass = 0.29\,H_{\max}^{2} + 1.02NDVI - 41.11 \qquad (1)$$

where $H_{\max}$ is the footprint's maximum canopy height computed from lidar data and NDVI is a measure of greenness scaled from 0 to 255 computed from Landsat TM image's Red and Near Infrared bands' reflectance values. For all the mapping methods, the data set was split into training (2/3) validation (1/3) datasets.

2.3. Spatial autocorrelation analysis

The spatial autocorrelation statistical formulas described in this section are presented in Reich (2008). Geographic coordinates of the GLAS waveforms were used to determine the inverse distance between points in computing a spatial weight matrix that was used in equation 2 below to calculate Moran's I. The weights were calculated in distance lags of 500 m and then rescaled from 0 to 1 such that when points *i* and *j* are closest to each other they had the highest weight (1), which decreases to near zero when they are furthest from each other but still within 500 m, and eventually zero if they are more than 500 m apart. The Moran's I values were then plotted against distance to give a correlogram that shows how the spatial autocorrelation of biomass changes with distance. Moran's I was calculated using equation:

$$I(d) = \frac{n\sum_i \sum_j w_{ij}(d) z_i z_j}{W(d)\sum_i z_i^2} \tag{2}$$

where *d* is the distance class, *w_{ij}(d)* is the weight at distance d, *W(d)* is the sum of all the weights at distance *d* , $z_i = (y_i - \bar{y})$ and $z_j = (y_j - \bar{y})$. The information from the correlogram was then used to decide on default lag distances to use in computing semi-variograms (section 2.5.2)

2.4. Mapping Methods

2.4.1. Deterministic Approach

The inverse distance weighting (IDW) method was chosen to represent this group. IDW interpolates by using the average of weighted observational data points within a given search radius based on the inverse distance from the estimation point (Dirks *et al.*, 1998). The distance is raised to some power based on minimum error and in this work we used a power of 2 as recommended by literature (Kruizinga & Yperlaan, 1978; Dirks *et al.*, 1998). Equation 3 shows the IDW model. The number of nearest neighbors and search radius used in IDW were based on the results of the spatial correlation analysis explained in section 2.3 above.

$$P_{est} = \frac{\sum_{i=1}^{n} P_i / d_i}{\sum_{i=1}^{n} 1/d_i} \tag{3}$$

where P_{est} is the estimation for the unsampled point, d_i is the distance of that point to a sampled point *i* within the search radius and whose value is P_i, *n* is the number of points in the search radius

2.5. Geostatistical approach

2.5.1. Cokriging

In cokriging, estimates are made using the spatial correlation of the primary variable (biomass) with itself, the spatial correlation of the secondary wall-to-wall variable (radar backscatter in this case) with itself, and a cross-correlogram that describes the cross-correlation between the primary and secondary variable. The HV layer was selected since it was the one with the highest correlation to the biomass point data (r = 0.72). The spatial dependence is calculated by treating Moran's I as a special case of the cross-correlation statistic:

$$I_{YZ} = \frac{1}{W}\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(z_j - \bar{z})}{\sqrt{Var(y)Var(z)}} \tag{4}$$

where w_{ij} is a scalar that quantifies the degree of spatial association or proximity between locations *i* and *j* , *W* is the sum of all n^2 values of w_{ij}, *Var(y)* is the sample variance of y_i, and *Var(z)* is the sample variance of z_i.

The cross-correlation statistic simultaneously tests the following hypotheses: a) Is variable *y* spatially correlated? b) Is variable *z* spatially correlated? c) Are variables *y* and *z* spatially cross-correlated? If $y_i = z_i$ the cross-correlation statistic is equivalent to Moran's I. Table 4.1a shows the linear relationship between biomass and backscatter HV and Figure 4.2 shows the cross-correlogram computed by the above formula.

2.5.2. Regression Kriging

Regression kriging involves fitting a linear model to the data to identify the variables that significantly explain the large scale trend (trend surface) of the dependent variable and then separately modeling the resulting residuals. This method requires that the predictors must be: 1) available at every sample point and 2) linearly related to the response variable (Hengl *et al*., 2003). After fitting the large scale trend model, the residuals are analyzed for spatial autocorrelation. If the residuals are not spatially autocorrelated then it means the variability in the mean component of the data is explained by large scale trends only. If they are spatially autocorrelated then it means there is a local spatial dependency that is not explained by the large scale trend. If the residuals are approximately normally distributed they can then be spatially modeled by OK. The residuals kriging layer added to the trend surface will give a distribution of the response variable of interest across the concerned landscape.

An Ordinary Least Squares (OLS) model was fit to relate the footprint biomass to 5 variables: radar backscatter HV polarization (HV), radar backscatter HH polarization (HH), elevation, terrain aspect and slope (results shown in table 4.1b). The residuals were checked for conformity to kriging requirements (normal distribution, presence of spatial autocorrelation and stationarity) by means of calculating Morani's I and examining their histogram and *qq* plots. As Figure 4.3 shows, the residuals satisfied the ordinary kriging requirements.

The OLS model was then used to develop the trend surface and the spatial structure of the residuals was modeled using OK. To accomplish this, semi-variograms were computed to determine the kriging parameters to use. We computed the semi-variance ($\gamma(h)$) from distance 0 to 10 000 m using equation 5 and then divided the distance range into 20 equal lags to show the semi-variogram as a plot of the median semi-variance of each class (y-axis) against the center of each distance class (x-axis).

$$\gamma(h) = \frac{1}{2n(h)} \sum_{(ij|h_{ij}=h)} (z_{i-} z_j)^{2} \tag{5}$$

where *h* is the distance separating sample locations *i* and *j*, *z_i* is the variable of interest at location *i* and *n(h)* is the number of data pairs separated by distance *h*.

We checked for anisotropy (the directional tendency of spatial autocorrelation) by calculating and plotting semi-variograms of the residuals in 3 directions (0, 45 and 90 degrees) and compared these with the omnidirectional semi-variogram (Figure 4.4). The range and sill values for all four semi-variograms were approximately the same, suggesting there was no anisotropy in the residuals. This means the trend surface model had captured the large scale variability so isotropic models could be used in kriging the residuals. In plotting the semi-variograms, we tried the Gaussian, spherical, and exponential models (equations 6-8) and then selected one that fit the data best. Some studies use the AIC values of the models to determine the best one to use but in our case the AIC values were not very different so we instead used a cross validation approach to determine the Mean Squared Error of Prediction (MSEP) associated with each model and selected the best model as the one that had the lowest MSEP values. An advantage of this cross validation method is that its output can also be used to determine the optimum number of nearest neighbors to use in the interpolation. This means that during kriging, the search range will not be fixed but varies according to the distribution of the nearest neighbors, a more reasonable approach for sample data that is not uniformly stratified. The exponential model was the best in terms of this criterion (Table 4.2). After modeling the spatial structure of the residuals, the output raster was added to the trend surface to obtain the final estimation map.

Spherical model: $$\gamma(h) = c_0 + c_1\left[\frac{3h}{2a} - 0.5\left(\frac{h}{a}\right)^3\right] \tag{6}$$

Gaussian model: $$\gamma(h) = c_0 + c_1\left[1 - \exp\left(\frac{h^2}{a^2}\right)\right] \tag{7}$$

Exponential model: $$\gamma(h) = c_0 + c_1\left[1 - \exp\left(\frac{-h}{a}\right)\right] \tag{8}$$

where for all models, the parameter a is the range, h is the distance lag and c_0+c_1 is the sill of the semi-variogram. When $h > a$, $\gamma(h) = c_0 + c_1$.

2.6. Image Segmentation Approach

Image segmentation involves partitioning image pixels into multiple patches (objects/segments) that have similar characteristics such as color, intensity or texture. Definiens Developer (Definiens, 2007) software was chosen for image segmentation due to its record of success for natural resources applications (van Aardt, Wynne, & Scrivani, 2008). Definiens builds objects in a bottom-up procedure that starts from seed pixels which are then merged into polygons, which are further merged until user supplied spectral and spatial heterogeneity criteria is met (Benz *et al.*, 2004; Definiens, 2007). We used two of the software's main algorithms in a hierarchical manner. We first segmented the image using the contrast split algorithm and then did a multiresolution segmentation on the result. The contrast split algorithm segments an image or image objects into dark and bright regions. This gave a first cut distinction between high tree density, moderate tree density, grass and bare areas. The multiresolution algorithm is an optimization procedure which identifies single image objects and merges them with their neighbors, based on relative homogeneity criteria. The segmentation is controlled by the main heterogeneity criteria of color and shape and a scale parameter, which defines the maximum allowable heterogeneity of the objects. This further split the objects created in stage one (contrast split) using an equal weight for color and shape and a scale factor of 20, a combination that optimized the quality of the segmentation as determined by visual inspection.

Patches that were coincident with one or more ICESat-GLAS footprints were randomly split into training (two thirds) and test (the other third) datasets. Within each patch we calculated the average biomass as well as statistics (mean and standard deviation) describing the independent variables (HV, HH, elevation, slope and aspect). The training data set was used to develop models whose results are shown in Table 4.1c. The first was an OLS model generated to relate average patch biomass to the statistics of the independent variables. Estimates made between adjacent patches are expected to have a positive covariance (autocorrelation). To investigate this we created a binary spatial weight matrix where a weight of 1 was assigned to adjacent patches and 0 for non-adjacent patches and calculated Moran's I for the residuals of the OLS model. Since the residuals were significantly ($p<0.05$) spatially autocorrelated, we accounted for spatial autocorrelation by also doing the estimation using a second model, which is a spatial autoregressive error (SAR) model (Anselin &Bera, 1998; Reich, 2008). The model used was:

$$Y = X\beta + (I - \lambda W)^{-1}\varepsilon \tag{9}$$

where Y is the response variable, X is a vector of the independent variables, β is a vector of the regression coefficients, I is an identity matrix, λ is spatial autoregressive coefficient, W is a spatial weight matrix and ε is an independent error term.

The SAR model is defined by adding a spatial structure term to the OLS model's residuals thus partitioning the error term into a spatial structure residual and a random residual. The cause of the residual autocorrelation is typically assumed to arise from the exclusion of an unobserved endogenous spatially structured covariate, that were it measured would explain the spatial autocorrelation in the residuals. This model therefore controls for the effect of correlated errors arising from an inherently spatial process or spatial autocorrelation in the measurement errors of the variables in the model (Anselin & Bera, 1998). We then used the OLS and the SAR models separately to estimate the average biomass for every other patch in the whole landscape.

In Figure 4.7 we present the results for both the SAR and OLS models but we used only the SAR model for every other subsequent analysis (including the production of the final estimation map) since it had better results. We also investigated the effect of lidar sampling intensity on the results of the image segmentation method. Firstly we compared the validation residual for each patch to the number of footprints in that patch and found out that patches with fewer footprints tended to have higher residuals. We then developed the estimation model separately for 3 main groups of patches according to the number of coincident footprints i.e. group one with only one footprint, group two with 2 or more footprints and group three with 3 or more footprints. There were 223 patches overall and 82 of these had at least 2 footprints so we split them into 55 for training and 27 for validation. However, there were only 34 patches with 3 or more footprints thus splitting them into training and validation sets would result in a very low sample size so we only used a 10 fold cross validation approach for this group (Kohavi, 1995). In addition to using an independent validation set, we also did a 10 fold cross validation for the other 2 groups for comparison purposes.

2.7. Validation of the output maps

The validation data set mentioned in the sections above were used to determine how at a point with known value, the estimation from interpolation by each method departed from the observed value. We extracted the predicted biomass values from the output layer of each method and correlated them with the observed values. The R^2 and RMSE statistics were then used to determine the level of accuracy. The validation was therefore at the footprint level for IDW, cokriging and RK and at both the footprint and patch level for the image segmentation approach.

2.8. Final biomass distribution maps

To minimize edge effects we did the analysis on raster layers at an extent (Figure 4.1) greater than the study area boundary and then extracted the focal study site results from the output (Wickham *et al*., 2008). Due to the use of many remote sensing independent variables, impossible results such as negative biomass values are expected from some few pixels. To avoid these unreasonable results, we truncated every negative biomass value to 0. The study area also has a few non-vegetation areas (mainly roads and buildings). It would have been ideal to mask these out but they are relatively very small and doing so would have introduced many more edges and hence increase edge related errors. If necessary, analysis can be done on the full extent like we did and then mask out these small patchy non-vegetation areas in the very end. To determine the total study site biomass predicted by each method, we calculated each pixel’s absolute biomass value by multiplying the Mg/Ha value by 0.3025 (the pixel area in Ha) and then summed up all pixel biomass values. To compare these total biomass values per method we divided the processing extent into 32 regular blocks (Figure 4.1) and used them as replicates. We then calculated the total biomass of each block as predicted by each method and used a one way analysis of variance (ANOVA) to compare the means.

3. Results

3.1. Spatial Autocorrelation

Overall there was a significant ($p < 0.001$) positive spatial autocorrelation in the footprint biomass data. The spatial autocorrelation started off very high at small distance lags and decreased to zero around 1500 m, after which it fluctuated randomly and at very low positive and negative values (Figure 4.5).

3.2. Biomass estimation and the resulting maps

The study site estimation maps produced by each method are shown in Figure 4.6. One way analysis of variance showed that there was no significant difference ($p > 0.05$) among the total biomass values predicted by the 4 methods at the $\alpha= 0.05$ level. The distribution of biomass (as determined using the 32 comparison blocks) was similar among these methods (Table 4.3). Figure 4.7 shows the validation plots and statistics for each method. Overall, the IDW and cokriging methods were poor at reproducing the structure (distribution) of the primary variable across the landscape. IDW was worse as it gave an over-smoothed distribution while concentrating the gradient only on areas close to the observation points. These two methods also tended to heavily underestimate the primary variable on the high biomass plots. Regression kriging and the image segmentation methods gave estimation surfaces that better visually resembled the distribution of the primary variable. This suggests that regression kriging and image segmentation approaches are better than IDW and cokriging since they had a higher accuracy (as measured by R^2 and RMSE values) and a better ability to reproduce the spatial variability of the primary variable.

For the image segmentation approach, the estimation accuracy was dependent on the number of footprints coincident with each patch as shown by Figure 4.8. Estimation error was lower for patches that had a higher number of coincident footprints as evidenced by the residuals of each patch being significantly negatively correlated to the number of footprints ($p = 0.04$, $r= 0.25$, Figure 4.8b).

4. Discussion

The spatial autocorrelation range of ~ 1500 m provides a reference point to guide spatial sampling in this landscape. Any transect sampling approach will be improved when there are enough transects so that an adequate number of points fall in the search radius of an interpolation method. GLAS sampling is high along track, with footprints separated by ~ 172 m but the spacing between ascending and descending tracks in a single orbit is in km, varying with latitude (Abdalati *et al*., 2010). Spatial sampling in heterogeneous ecosystems like savannas would be ideal if it is sufficiently dense so that trends can be characterized on the scales at which they vary. The wide across track separation of GLAS footprints has resulted in large scale studies that resort to averaging of observations to the land cover type/ patch level (Boudreau *et al*., 2008; Lefsky, 2010; Nelson *et al*., 2009; Mitchard *et al*., 2012). This results in some smaller/local scale trends going unnoticed as they are masked by the larger trends captured by the scale at which observations are made. In spite of our good results, better sampling densities are recommended for future spaceborne platforms. In this context we conclude that NASA’s GEDI mission will be successful for landscapes of this type. GEDI will sample all the land between 50° North and South latitudes using 3 High Output Maximum Efficiency Resonator (HOMER) lasers, whose beams will be divided into 14 parallel tracks of 25 m contiguous footprints on the ground (Stysley *et al*., 2015). With a track separation of 500 m, one swath will cover about 6.5 km, a configuration dense enough to fulfill the spatial autocorrelation conditions we identified in this work.

The non-significant difference in the total landscape biomass predicted by the 4 methods suggests that the pros and cons for each method must be based on other factors. IDW requires fewer assumptions and input parameters and the algorithm is fairly quick hence when interpolation is just needed for a quick rough picture on which to base secondary objectives, it would be adequate. When there are no predictor variables that significantly explain the global trend, the distribution can be treated as heavily dependent on distance only and hence IDW will be most appropriate. Likewise when one has a very high density of sample points and the global (landscape) mean is of primary importance, IDW would be acceptable. In this case high accurate (densely sampled) areas can compensate for low accuracy areas while areas of overestimation can cancel out some areas of underestimation. The failure of the IDW method to portray the actual pattern of landscape biomass makes it less ideal when the local mean is preferred more to the global mean. The “bull’s-eye” effect around data points and oversmoothing of unsampled areas make IDW inappropriate for mapping biomass as a determinant of flow based ecosystem services such as water flow, habitat availability and associated ecological phenomenon like herbivory and fire.

Cokriging and RK exploit the correlation of the primary variable to at least one covariate hence their estimation maps are better at mimicking the primary variable’s trend in the landscape. The portrayal of the landscape is important for management purposes. With the results, not only can we explain the landscape in terms of the mapped process, but also in terms of what influences it, which is important for making decisions about issues like selective logging or restorative management since priority areas can be singled out. Visual inspection of the biomass maps indicate that the RK estimation map better represents the landscape pattern of biomass than does the cokriging map. This is mainly because the cokriging analysis used only one covariate, i.e. radar backscatter HV while RK used 4 (radar backscatter HV and HH, slope and aspect).

It is important to note that the furthest western part of the study area is outside the spatial autocorrelation range of 1500 m from the nearest observation point so the RK results in that part are only based on the trend surface model and not the residuals’ kriging surface. This does not affect our cross validation results because they are based on observations from the adequately sampled area. Slope and aspect are bio-geographically very important variables in rugged terrain since vegetation distribution and growth depend on sun angle and elevation (Day & Monk, 1974; O’loughlin, 1981). Aboveground biomass in savannas has been found to be highly sensitive to topographic factors such as absolute or relative elevation, especially when considered at a landscape level (Colgan *et al*., 2012). Cokriging and RK are also better for understanding purposes as they offer relatively more information than IDW. The semi variogram or cross correlogram analysis required for obtaining mapping parameters provide interpretive values beyond the methods’ role in interpolation and explain well the nature, intensity and extent of the spatial distribution of the data, which deterministic methods like IDW cannot. These detailed trend analyses are important in savannas since the vegetation is highly irregular in canopy and crown shape, height and other structural dimensions, showing inter and even intra-species spectral variability due to natural differences in topography, rainfall, herbivory and human impacts (Naidoo *et al*., 2012).

The demonstration of the accuracy of the image segmentation approach relies on the presence of sufficient footprints to adequately characterize each patch. The use of the average value as the sampling unit has a problem of smoothing variability but this may be acceptable when larger scale trends are of particular interest. At regional to global scales, even other functional characteristics based data sources such as land cover maps and ecoregion types can be used as in related forest based studies (Boudreau *et al*., 2008; Nelson *et al*., 2009; Lefsky, 2010). In a savanna study Colgan *et al*. (2012) also pointed out the importance of edaphic factors at smaller scales. The key point is that the observation unit to use will depend on the scale and purpose of modelling.

At very small scales, other high resolution data can be used to further segment the image to finer resolutions than we did in this work. Discrete return lidar data is commonly used either in identifying patches, or providing more covariates to use in training the mapping models (Antonarakis, Richards, & Brasington, 2008; Arroyo, Johansen, Armston, & Phinn, 2010; van Aardt et al., 2008). At moderate to large scale work like ours, other studies like Mitchard *et al*. (2012) have used the Radar Forest Degradation Index (RFDI) to classify the study site into different patches based on vegetation type. We tried this index but it was not significantly ($P>0.05$) correlated to biomass, probably because our landscape was not as diverse as theirs which ranged from very dense tropical forests to very open savannas. The reduced estimation error associated with patches that had a higher number of footprints suggests that the relatively high spatial density of observations for missions such as GEDI will be a successful strategy for reducing uncertainty in biomass estimates.

The differences in the distribution of biomass obtained by the 4 methods are mostly a result of the models' different capabilities to utilize secondary variables. IDW simply uses the sampled points' values and the Euclidean distance between them thus estimates at points farther away from the sample data become blocky. Moreover, if only non-zero biomass plots are used in mapping, the method cannot capture and isolate zero biomass areas like water bodies. On the other hand the secondary variables (e.g radar backscatter) used in a method like RK can isolate such areas. If the patch delineation method is accurate enough, the image segmentation method can also isolate these areas. This was demonstrated in this work where the IDW and cokriging estimation maps could not show the small water body in the southwestern quadrant of the study area that can be clearly seen in Figure 4.1. Regression kriging and image segmentation methods managed to capture and isolate this water body as seen in Figure 4.6. The regression kriging's output map looks fuzzier than that of the image segmentation method most probably due to the different grain/resolution in the data used by the two approaches. RK's estimation was done at the pixel level (finer resolution) while image segmentation was at the patch level. As a result, rare or small objects are lost as grain becomes coarser, especially if they are dispersed (Turner *et al*., 1989; Levin, 1992; Qi & Wu, 1996).

The validation plots (Figure 4.7) clearly show that IDW and cokriging underestimate the higher biomass areas. Holding other factors constant, this would mean that the total landscape biomass predicted by these two methods would be much lower than that of the other two methods. However, this was not the case in our results, most probably because the former methods compensated for this by overpredicting those zero and low biomass areas as explained above, resulting in the uncertainties canceling out. We cannot tell conclusively why the image segmentation method predicted the highest (although not significantly) biomass values among all the other methods but our speculation is that it may have resulted from the edge effects associated with using numerous objects (patches) as polygons. It is more likely that the image segmentation approach generally overestimated the biomass in some areas since our sample plots' mean biomass is much lower (112 Mg/Ha) and literature has shown that the average biomass in this type of ecosystem is less than 120 Mg/Ha (Battles *et al.*, 2008). The higher correlation between predicted and observed biomass and lower RSME associated with the image segmentation method at the patch level is partly due to the fact that variability was reduced by means of using the average patch value as the smallest unit. However, the image segmentation approach still gave the best results even when validation was done at the footprint level (figure 4.7).

5. Conclusions

Our work has demonstrated the use of ICESat-GLAS to create wall-to-wall biomass maps in savannas, and the results are likely to be applicable to other waveform lidar remote sensing missions. Biomass estimates would likely be much improved if the sampling intensity is increased so that biomass measurements are optimally sampled and landscape patches are better characterized. The mapping methods work better if they are more informed by correlated secondary variables so as to reproduce the variability of the primary variable across the landscape. Building on these results, increased sampling intensity and improved methods with future missions will increase our confidence in change detection studies related to landscape level biomass and the associated ecological functionalities.

6. Tables and figures

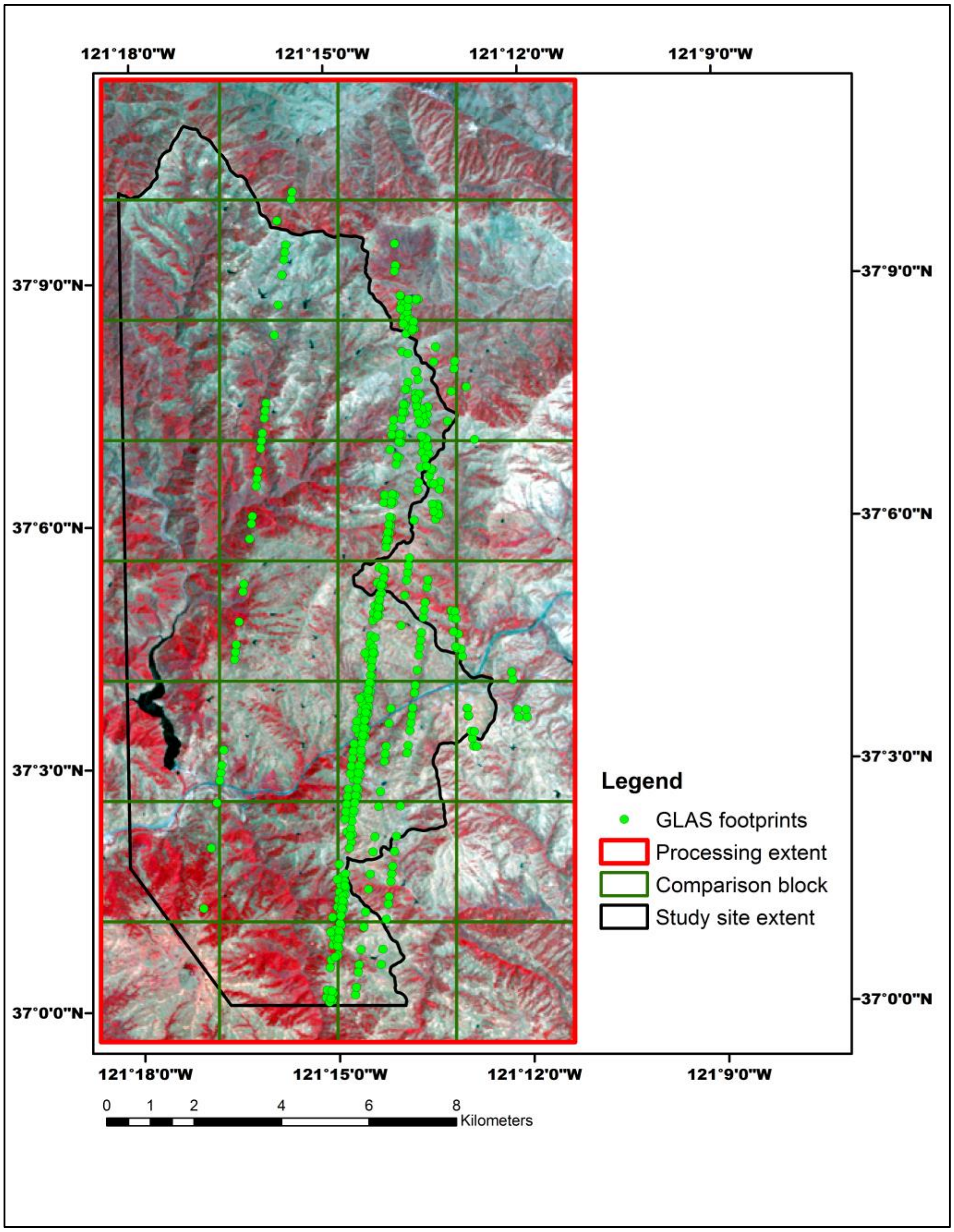


Figure 4.1: Map showing Landsat TM false color (432) image, study area extent and location of ICESat-GLAS footprints used as plots in this work.

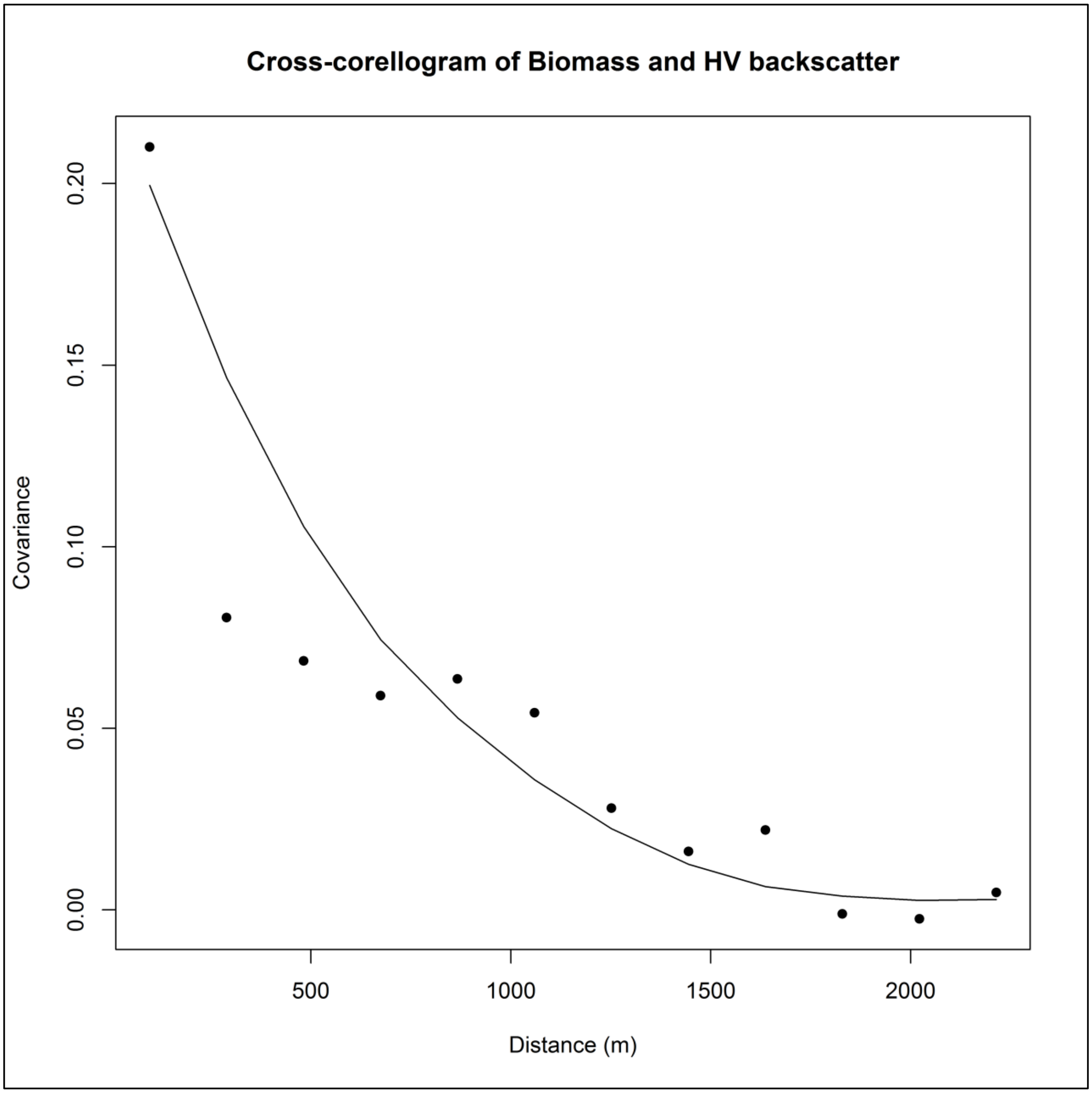


Figure 4.2: Cokriging cross-correlogram

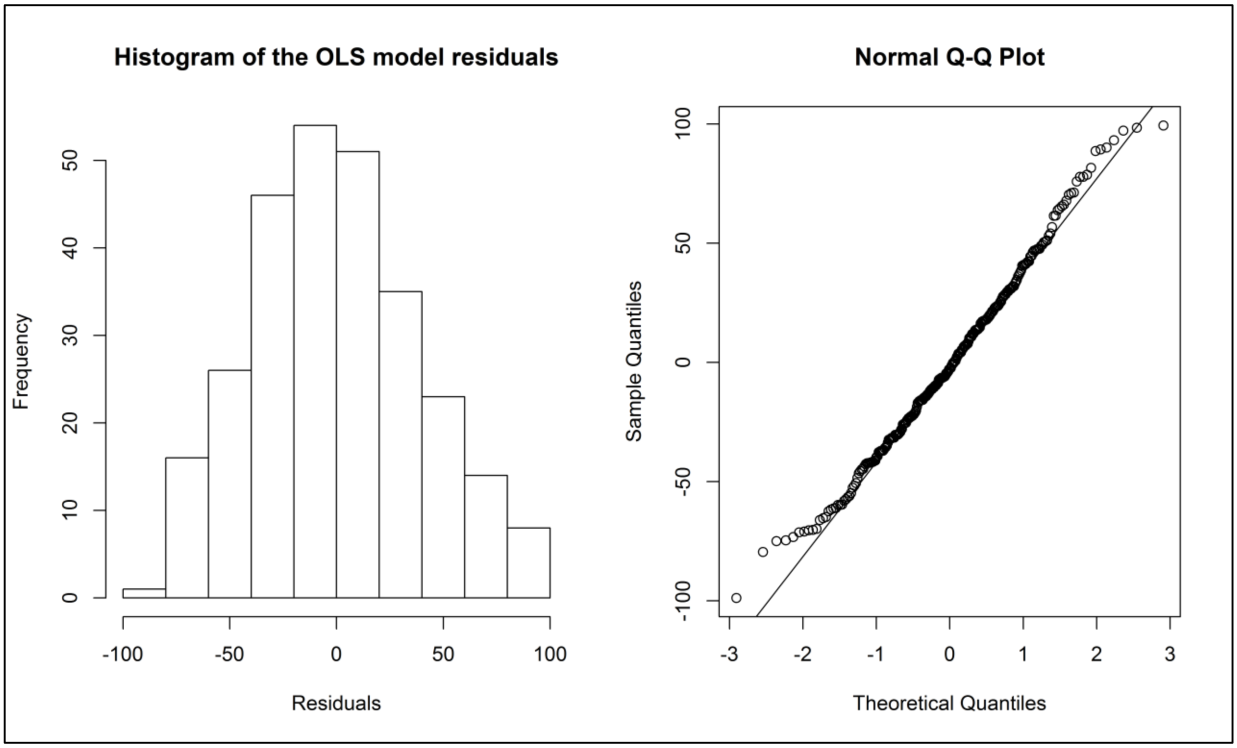


Figure 4.3: Diagnostic plots of the trend surface OLS model's residuals. Morani's I = 0.04, p value = 0.000961

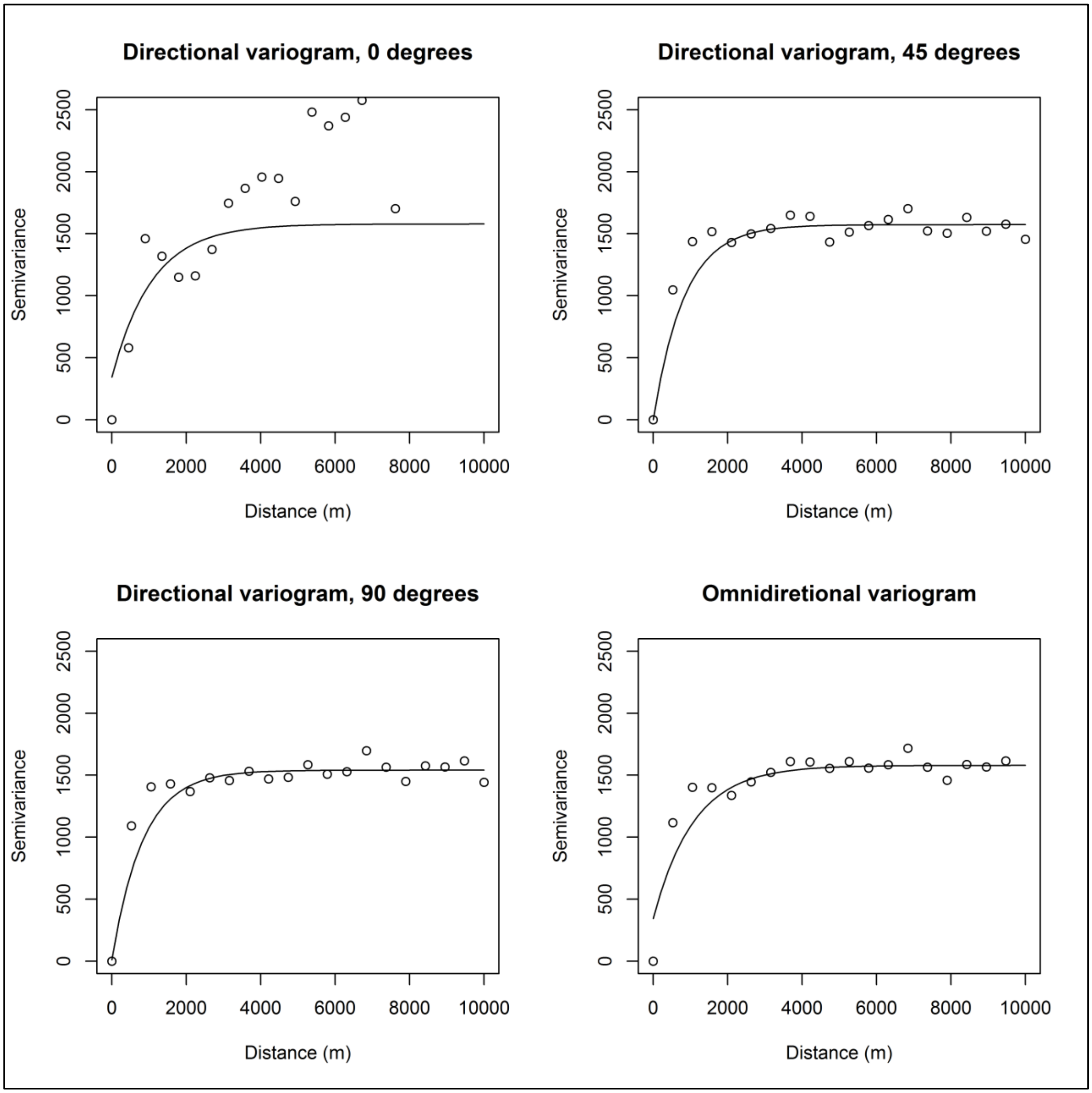


Figure 4.4: Semi-variograms for the OLS model residuals with an exponential model (equation 8) curve fit

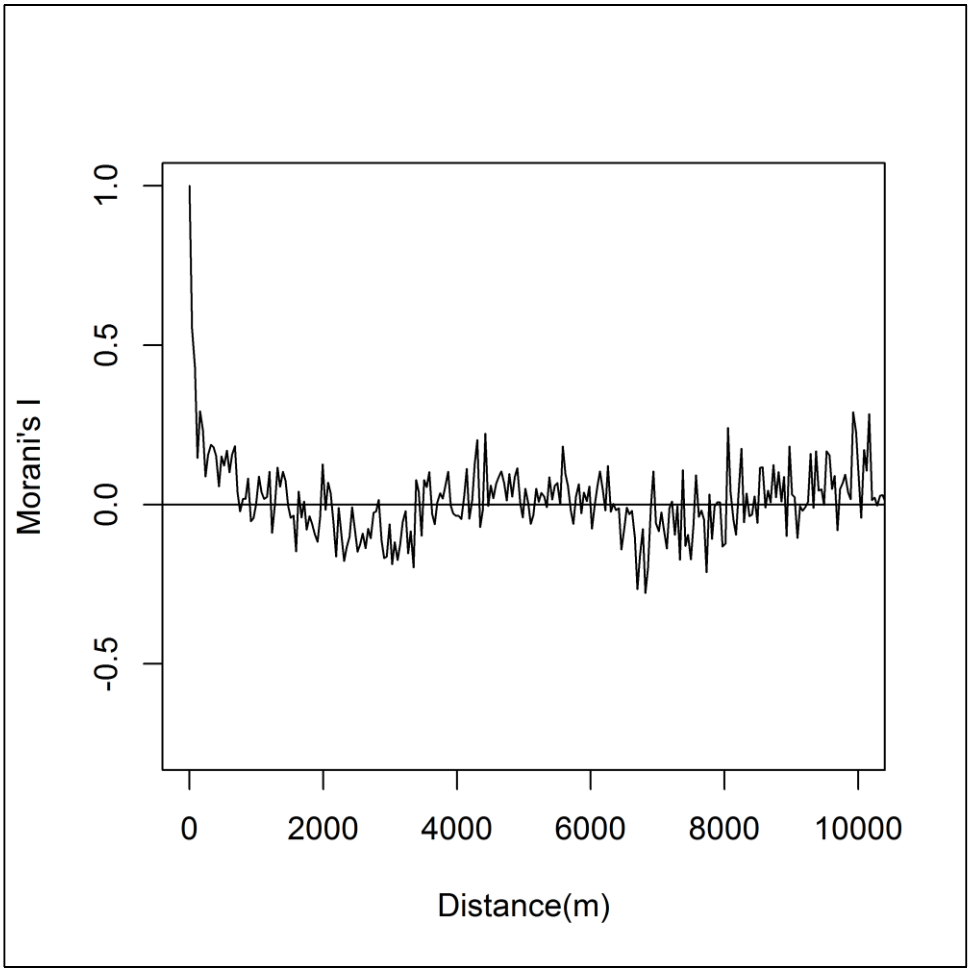


Figure 4.5: Correlogram showing how the spatial autocorrelation of biomass varies with distance

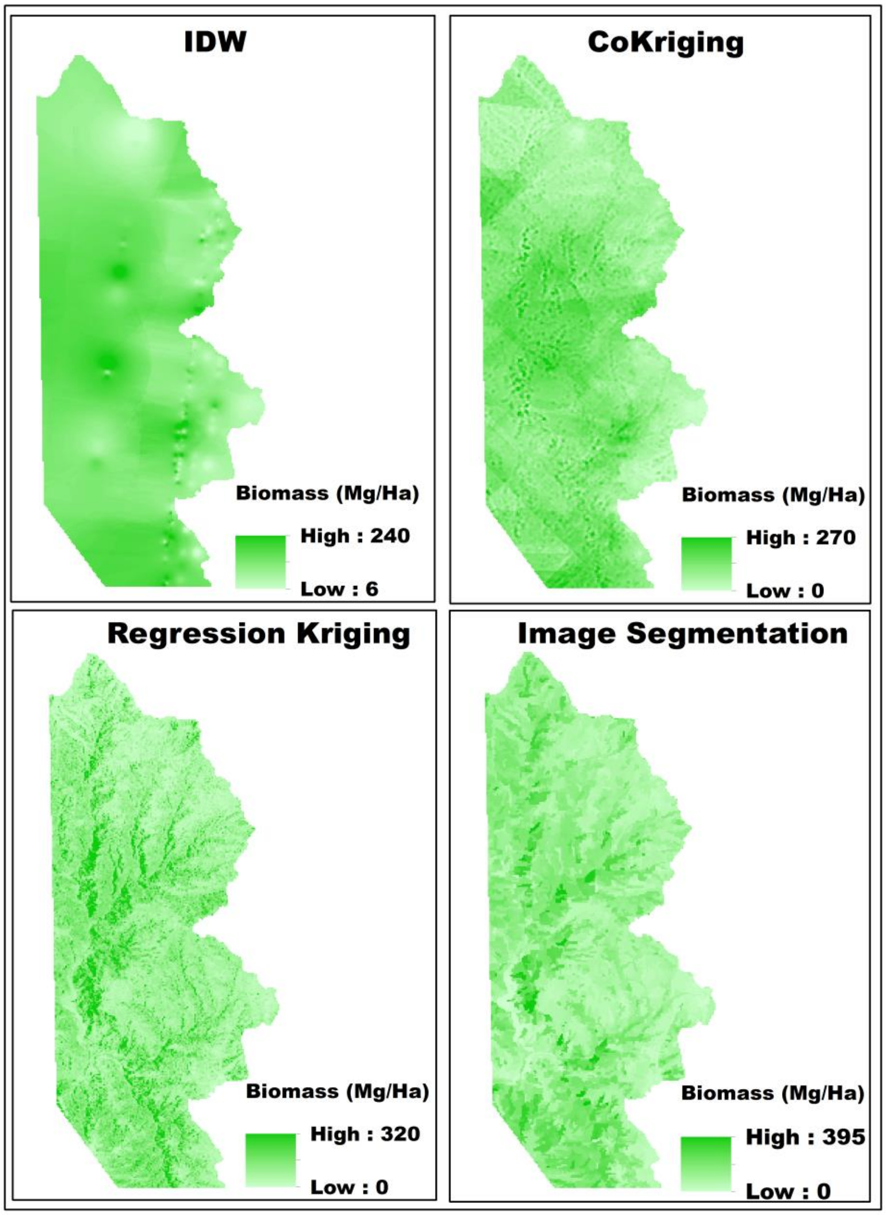


Figure 4.6: Estimation maps of the 4 methods used in interpolation

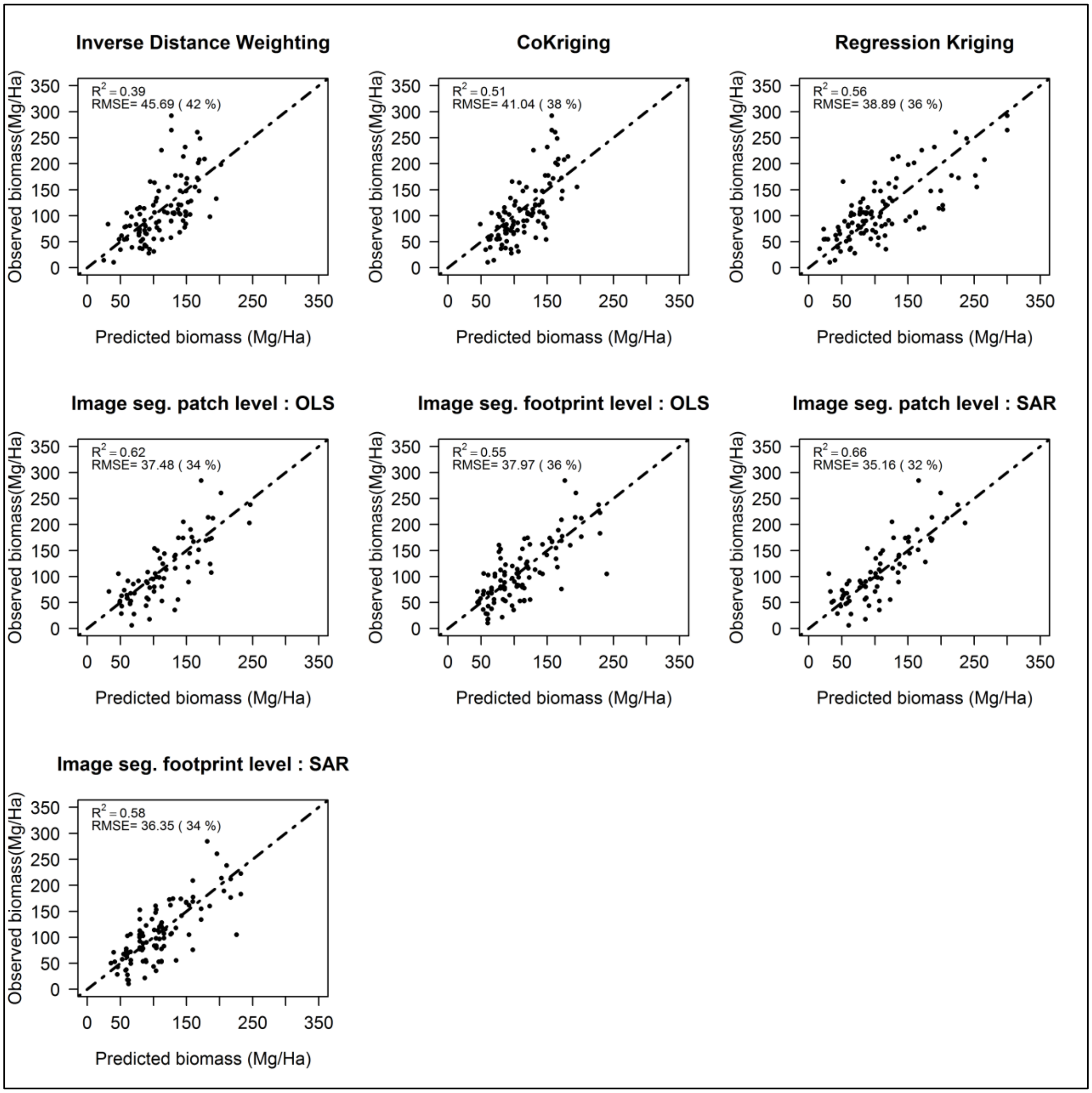


Figure 4.7: Validation plots and statistics for the 4 methods used in interpolation

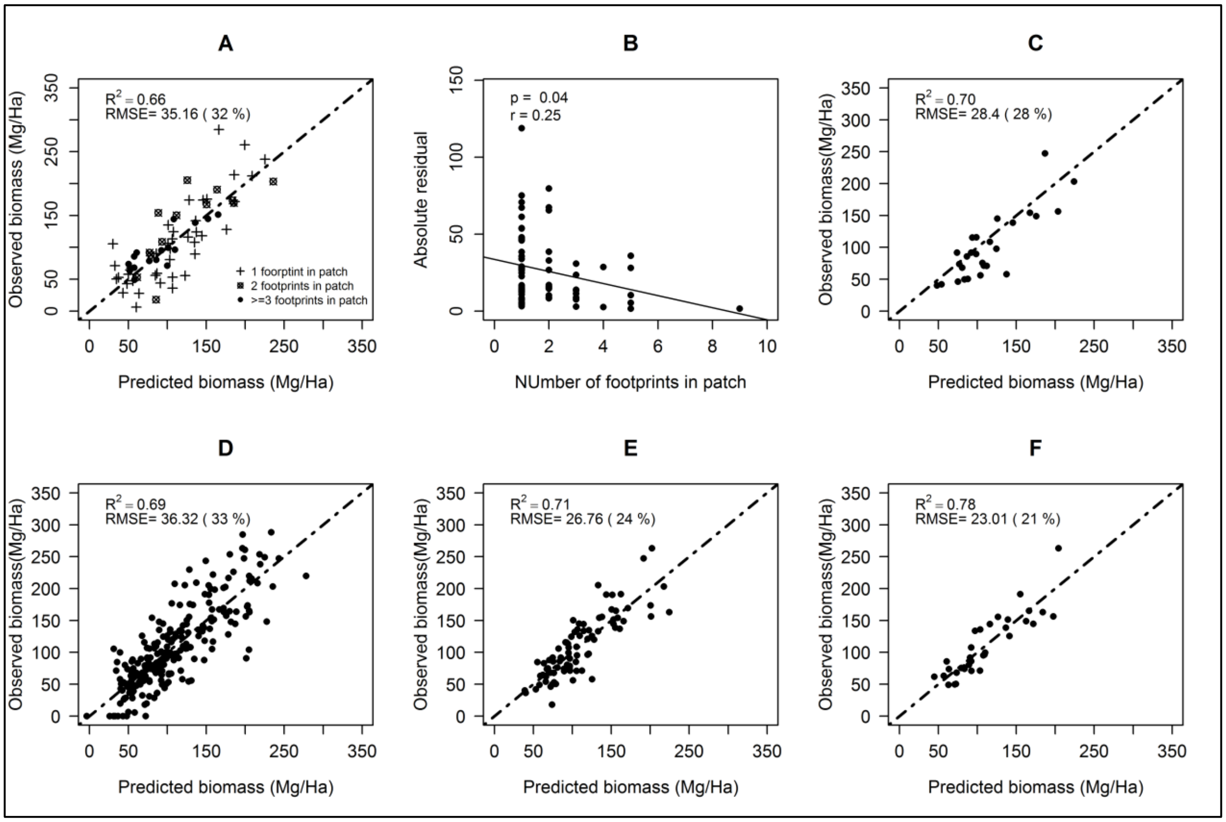


Figure 4.8: Relationship between estimation error and number of coincident footprints. A: independent validation using all patches, B: correlation of absolute residual and number of footprints in patch, C: independent validation using only patches with 2 or more footprints, D: a 10 fold cross validation using all patches, E: a 10 fold cross validation using only patches with 2 or more footprints, F: a 10 fold cross validation using only patches with 3 or more footprints.

Table 4.1: Intermediate regression model results

Method	**Model type**	**Model parameters**			**Model**
		Variable	**Coefficient**	***p* value**	**R^2 value**
a. Cokriging	Linear model of footprint level biomass as a function of the highest correlated auxiliary data layer	Intercept	56.23	<0.0001	0.52
		HV	5866.83	<0.0001	
b. Regression Kriging	OLS for trend surface	Intercept	53.81	<0.0001	0.57
		HV	4620.84	<0.0001	
		Aspect	-0.11	<0.0001	
		HH	211.88	0.0011	
		Slope	0.89	0.0076	
c. Image segmentation	SAR for patch level biomass	Intercept	64.43	<0.0001	0.78
		Mean HV	6252.95	<0.0001	
		Mean slope	3.64	<0.0001	
		HH std	-142.49	0.0001	
		Mean aspect	-0.10	0.0062	
		HV std	-2542.21	0.0404	
	OLS for patch level biomass	Intercept	18.81	0.0424	0.69
		Mean HV	9318.10	<0.0001	
		Mean HH	-503.62	<0.0001	
		Mean slope	5.03	<0.0001	
		Mean aspect	-0.09	0.0317	

Table 4.2: Semi-variogram statistics for the models fit to the regression kriging trend surface residuals

Model	**Nugget**	**Range**	**Sill**	**AIC**	**MSEP range**
Exponential	0.05	1203.56	1573.98	250.55	1799.55 – 2027.20
Gaussian	0.86	1104.65	1539.33	249.39	2750.65 – 3088.31
Spherical	0.21	1325.47	1538.87	249.55	1981.68 – 2309.33

Table 4.3: Study site biomass estimations by the 4 methods

Estimation Method	**Study site mean pixel biomass (Mg/Ha)**	**Total Study site biomass (Tg)**
Inverse distance weighting	119	1.40
Cokriging	117	1.39
Regression kriging	118	1.39
Image segmentation	121	1.43

CHAPTER 4 REFERENCES

Abdalati, W., Zwally, H. J., Bindschadler, R., Csatho, B., Farrell, S. L., Fricker, H. A., … Webb, C. (2010). The ICESat-2 Laser Altimetry Mission. *Proceedings of the IEEE, 98*(5), 735–751.

Anselin, L., & Bera, A. K. (1998). Spatial dependence in linear regression models with an introduction to spatial econometrics. In A. Ullah & G. D. E. A (Eds.), *Handbook of Applied Economic Statistics* (pp. 237–289). New York, USA: Marcel Dekker.

Antonarakis, a. S., Richards, K. S., & Brasington, J. (2008). Object-based land cover classification using airborne LiDAR. *Remote Sensing of Environment, 112*(6), 2988–2998.

Arroyo, L. A., Johansen, K., Armston, J., & Phinn, S. (2010). Integration of LiDAR and QuickBird imagery for mapping riparian biophysical parameters and land cover types in Australian tropical savannas. *Forest Ecology and Management, 259*(3), 598–606.

Battles, J. J., Jackson, R. D., Shlisky, A., & Bartolome, J. W. (2008). Net Primary Production and Biomass Distribution in the Blue Oak Savanna 1. In A. Merenlender, D. McCreary, & K. L. Purcell (Eds.), *Proceedings of the Sixth Symposium on Oak Woodlands : Today's Challenges, Tomorrow's Opprtunities* (pp. 511–524). USDA Forest Service, General Technical Report PSW-GTR-217.

Benz, U. C., Hofmann, P., Willhauck, G., Lingenfelder, I., & Heynen, M. (2004). Multi-resolution, object-oriented fuzzy analysis of remote sensing data for GIS-ready information. *ISPRS Journal of Photogrammetry and Remote Sensing*, *58*(3-4), 239–258.

Bernstein, L. S., Adler-Golden, S. M., Sundberg, R. L., Levine, R. Y., Perkins, T. C., Berk, A., … Hoke, M. (2005). Validation of the QUick Atmospheric Correction (QUAC) algorithm for VNIR-SWIR multi- and hyperspectral imagery. In *SPIE Proceedings, Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XI Vol. 5806* (pp. 668–678). SPIE Digital Library.

Boudreau, J., Nelson, R., Margolis, H., Beaudoin, A., Guindon, L., & Kimes, D. (2008). Regional aboveground forest biomass using airborne and spaceborne LiDAR in Québec. *Remote Sensing of Environment*, *112*(10), 3876–3890.

Chambers, J. Q., Asner, G. P., Morton, D. C., Anderson, L. O., Saatchi, S. S., Espírito-Santo, F. D. B., … Souza, C. (2007). Regional ecosystem structure and function: ecological insights from remote sensing of tropical forests. *Trends in Ecology & Evolution*, *22*(8), 414–23.

Cho, M. A., Mathieu, R., Asner, G. P., Naidoo, L., van Aardt, J., Ramoelo, A., … Erasmus, B. (2012). Mapping tree species composition in South African savannas using an integrated airborne spectral and LiDAR system. *Remote Sensing of Environment*, *125*, 214–226.

Cho, M. A., Naidoo, L., Mathieu, R., & Asner, G. P. (2011). Mapping savanna tree species using Carnegie Airborne Observatory hyperspectral data resampled to WorldView-2 multispectral configuration. In *34th International Symposium on Remote Sensing of Environment*. Sydney, Australia.

Colgan, M. S., Asner, G. P., Levick, S. R., Martin, R. E., & Chadwick, O. a. (2012). Topo-edaphic controls over woody plant biomass in South African savannas. *Biogeosciences Discussions*, *9*(1), 957–987.

Day, F. P., & Monk, C. D. (1974). Vegetation Patterns on a Southern Appalachian Watershed. *Ecology*, *55*(5), 1064.

Definiens. (2007). *Definiens Developer 7—User Guide*. Munich, Germany: Definiens AG.

Dirks, K. N., Hay, J. E., Stow, C. D., & Harris, D. (1998). High-resolution studies of rainfall on Norfolk Island Part II : Interpolation of rainfall data. *Journal of Hydrology*, *208*, 187–193.

Gonzalez, P., Asner, G. P., Battles, J. J., Lefsky, M. a., Waring, K. M., & Palace, M. (2010). Forest carbon densities and uncertainties from Lidar, QuickBird, and field measurements in California. *Remote Sensing of Environment*, *114*(7), 1561–1575.

Goovaerts, P. (1997). *Geostatistics for Natural Resources Evaluation*. Oxford, UK: Oxford University Press.

Gwenzi, D., & Lefsky, M. A. (2014). Modeling canopy height in a savanna ecosystem using spaceborne lidar waveforms. *Remote Sensing of Environment*, *154*.

Gwenzi, D., & Lefsky, M. A. (2015). Plot level aboveground woody biomass modeling using canopy height and auxiliary remote sensing data in a heterogeneous savanna. *Journal of Applied Remote Sensing*, (Submitted).

Hengl, T., Heuvelink, G. B. M., & Stein, A. (2003). Comparison of kriging with external drift and regression-kriging. Technical Note, International Institute for Geo-information Science and Earth Observation (ITC), Enschede, http://www.itc.nl/library/Academic output.

Kruizinga, S., & Yperlaan, G. J. (1978). Spatial interpolation of daily totals of rainfall. *Journal of Hydrology*, *36*, 65–73.

Lefsky, M. A. (2010). A global forest canopy height map from the Moderate Resolution Imaging Spectroradiometer and the Geoscience Laser Altimeter System. *Geophysical Research Letters*, *37*(15), 1–5.

Lefsky, M. A., Harding, D., Cohen, W. B., Parker, G., & Shugart, H. H. (1999). Surface Lidar Remote Sensing of Basal Area and Biomass in Deciduous Forests of Eastern Maryland, USA. *Remote Sensing of Environment*, *67*(1), 83–98.

Lefsky, M. A., Harding, D. J., Keller, M., Cohen, W. B., Carabajal, C. C., Del Bom Espirito-Santo, F., … de Oliveira Jr, R. (2005). Estimates of forest canopy height and aboveground biomass using ICESat. *Geophysical Research Letters*, *32*(22), 1–4.

Levin, S. A. (1992). The problem of pattern and scale in Ecology. *Ecology*, *73*(6), 1943–1967.

McBratney, A. B., Odeh, I. O. a., Bishop, T. F. a., Dunbar, M. S., & Shatar, T. M. (2000). An overview of pedometric techniques for use in soil survey. *Geoderma*, *97*(3-4), 293–327. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0016706100000434

Mitchard, E. T. a., Saatchi, S. S., White, L. J. T., Abernethy, K. a., Jeffery, K. J., Lewis, S. L., … Meir, P. (2012). Mapping tropical forest biomass with radar and spaceborne LiDAR in Lopé National Park, Gabon: overcoming problems of high biomass and persistent cloud. *Biogeosciences*, *9*(1), 179–191.

Naesset, E. (2002). Predicting forest stand characteristics with airborne scanning laser using a practical two-stage procedure and field data. *Remote Sensing of Environment*, *80*(1), 88–99. http://doi.org/10.1016/S0034-4257(01)00290-5

Naidoo, L., Cho, M. a., Mathieu, R., & Asner, G. (2012). Classification of savanna tree species, in the Greater Kruger National Park region, by integrating hyperspectral and LiDAR data in a Random Forest data mining environment. *ISPRS Journal of Photogrammetry and Remote Sensing*, *69*, 167–179.

Nelson, R., Ranson, K. J., Sun, G., Kimes, D. S., Kharuk, V., & Montesano, P. (2009). Estimating Siberian timber volume using MODIS and ICESat/GLAS. *Remote Sensing of Environment*, *113*(3), 691–701.

O'loughlin, E. M. (1981). Saturation regions in catchments and their relations to soil and topographic properties. *Journal of Hydrology*, *53*, 229–246.

Qi, Y., & Wu, J. (1996). Effects of changing spatial resolution on the results of landscape pattern analysis using spatial autocorrelation indices. *Landscape Ecology*, *11*(1), 39–49.

Reich, R. M. (2008). *Spatial Statistical Modeling of Ecosystem Resources and the Environemt, Short Course Manual*. Fort Collins, CO USA: Colorado State University.

Saatchi, S. S., Harris, N. L., Brown, S., Lefsky, M., Mitchard, E. T. a, Salas, W., … Morel, A. (2011). Benchmark map of forest carbon stocks in tropical regions across three continents. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(24), 9899–904.

Sales, M. H., Souza Jr., C. M., Kyriakidis, P. C., Roberts, D. a., & Vidal, E. (2007). Improving spatial distribution estimation of forest biomass with geostatistics: A case study for Rondônia, Brazil. *Ecological Modelling*, *205*(1-2), 221–230.

Stysley, P. R., Coyle, D. B., Kay, R. B., Frederickson, R., Poulios, D., Cory, K., & Clarke, G. (2015). Long term performance of the High Output Maximum Efficiency Resonator (HOMER) laser for NASA׳s Global Ecosystem Dynamics Investigation (GEDI) lidar. *Optics & Laser Technology*, *68*, 67–72. http://doi.org/10.1016/j.optlastec.2014.11.001

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, *46*(2), 234–240.

Turner, M. G., Neill, R. V. O., Gardner, R. H., & Milne, B. T. (1989). Effects of changing spatial scale on the analysis of landscape pattern. *Landscape Ecology*, *3*(4), 153–162.

Van Aardt, J. A. N., Wynne, R. H., & Scrivani, J. a. (2008). Lidar-based Mapping of Forest Volume and Biomass by Taxonomic Group Using Structurally Homogenous Segments. *Photogrammetric Engineering & Remote Sensing*, *74*(8), 1033–1044.

Wickham, J. D., Riitters, K. H., Wade, T. G., & Homer, C. (2008). Temporal change in fragmentation of continental US forests. *Landscape Ecology*, 891–898.

CHAPTER 5: PROSPECTS OF PHOTON COUNTING LIDAR FOR SAVANNA ECOSYSTEM STRUCTURAL STUDIES[4]

Synopsis

The next planned spaceborne lidar mission is the Ice, Cloud and land Elevation Satellite 2 (ICESat-2), which will use the Advanced Topographic Laser Altimeter System (ATLAS) sensor, a photon counting technique. To pre-validate the capability of this mission for studying three dimensional vegetation structure in savannas, we assessed the potential of the measurement approach to estimate canopy height in a savanna landscape. We used data from the Multiple Altimeter Beam Experimental Lidar (MABEL), an airborne photon counting lidar sensor developed by NASA's Goddard Space Flight Center. ATLAS-like data was generated using the MATLAS simulator, which adjusts MABEL data's detected number of signal and noise photons to that expected from the ATLAS instrument. Transects flown over the Tejon ranch conservancy in Kern County, California, USA were used for this work. For each transect we chose to use data from the near infrared channel that had the highest number of photons. We segmented each transect into 50 m, 25 m and 14 m long blocks and aggregated the photons in each block into a histogram based on their elevation values. We then used an automated algorithm to identify cut off points where the cumulative density of photons from the highest elevation indicates the presence of the canopy top and likewise where such cumulative density from the lowest elevation indicates the presence of the mean terrain elevation. MABEL derived height metrics

[4] David Gwenzi, Michael Andrew Lefsky, David Harding, Vijay Suchdeo, 2015
In preparation

were moderately correlated to DRL derived height metrics (r^2 and RMSE values ranging from 0.60 to 0.73 and 2.9 m to 4.4 m respectively) but MATLAS simulation resulted in more modest correlations with DRL indices (r^2 ranging from 0.4 – 0.64 and RMSE from 3.6 m to 5.2 m). Simulations also indicated that the number of signal photons will be substantially lower and this reduced canopy height estimation precision especially in areas of low density vegetation cover. On the basis of the simulated data, there is reason to believe that the ability of ICESat-2 to estimate height will be similar or worse than the original ICESat mission.

Key Words: Photon counting lidar, ICESat-2, MABEL, MATLAS, Savanna, Canopy height

1. Introduction

Lidar remote sensing provides a means to directly estimate the three dimensional biophysical parameters of vegetation using the physical interactions of an emitted laser pulse with the vegetation structure being illuminated. In ecology, one widely demonstrated application of lidar has been the estimation of canopy height which is in turn related to aboveground woody biomass, an important quantity in monitoring the dynamics and storage of carbon in vegetation. Due to their small spatial coverage and high acquisition costs, small footprint discrete return lidar (DRL) systems are ideally useful for small extents. For the opposite reasons, spaceborne large footprint waveform lidar systems have been the primary source of data for studies of larger extents (Hall et al., 2011). The Geoscience Laser Altimeter System (GLAS) aboard the Ice, Cloud and land Elevation Satellite (ICESat), was a spaceborne lidar sensor that provided waveform data and demonstrated a capability to estimate canopy height in various ecosystems (Lefsky *et al*., 2007; Duncanson *et al*., 2010; Lefsky, 2010; Xing *et al*., 2010; Simard *et al*., 2011;

Gwenzi & Lefsky, 2014). ICESat was decommissioned in 2010 and the earliest planned future mission is its successor, ICESat-2, which will use the Advanced Topography Laser Altimeter System (ATLAS).

Unlike GLAS, that used a full waveform recording technique, ATLAS will use a single photon counting technique. A single photon counting lidar (SPL) system fires thousands of laser pulses per second and records the travel time of individual photons that are reflected back to the sensor. The photons' time of arrival and the instrument's Global Positioning System (GPS) and Inertial Measurement Unit (IMU) orientation are used to calculate the distance the light travelled and hence the elevation of the surface below. The high level of sensitivity of a SPL at low energy expenditure promises extended laser lifetimes and makes it possible to fly at higher altitudes, thus providing larger coverage. The plan for ATLAS is to use a single pulse at 532 nm wavelength which will be split into 6 transmit beams arranged in 3 pairs. The configuration will give a distance of 3.3 km between each pair with a 90 m separation between the members of each pair. Using a 10 kHz repetition rate at an altitude of ~500 km will produce overlapping footprints of nominally 14 m diameter at 70 cm intervals along track (Abdalati et al., 2010). The location of the ICESat-2 footprint will be known but the origin of the recorded photons within the footprint will be unknown (Rosette et al., 2011). The primary objective of ICESat-2 will be the quantification of ice sheets and sea ice but as with ICESat-GLAS, vegetation height retrieval for biomass assessment is a science objective, although not a mission requirement.

The Multiple Altimeter Beam Experimental Lidar (MABEL) is an airborne simulator of ATLAS that was developed by NASA's Goddard Space Flight Center to pre-validate the ICESat-2 mission. MABEL flights were carried out on NASA's ER-2, a high altitude aircraft (http://www.nasa.gov/centers/armstrong/aircraft/ER-2/index.html). The sensor uses laser pulses in the red (532 nm, obtained by a frequency doubler) and near infrared (1064 nm) wavelengths at a variable repetition rate of 5-25 kHz. Typically, it uses a 10 kHz repetition rate and laser pulse length of 2 ns. At the platform's nominal speed of 200 ms^{-1}, a pulse is will be emitted every 4 cm along the track (McGill *et al.*, 2013). At the ER-2's operational altitude of 20 km, the laser illuminates a spot (footprint) of ~2 m in diameter, within the telescope's field of view of ~4 m. The output of the MABEL laser at the two wavelengths is split into 8 near infrared (1064 nm) and 16 red (532 nm) beams which can be off-nadir pointed at ±3°. With this configuration, a flight altitude of 20 km results in a swath width of up to 2.10 km. The details of MABEL configuration are given in (McGill *et al*., 2013).

The MABEL instrument was flown aboard the ER-2 on several missions above various earth surfaces between the years 2010 and 2014 at different times of the day. The variation of conditions under which it was flown provides different levels of solar background and other atmospheric conditions necessary to test signal detection algorithms for different surfaces, including vegetation. The aggregation of the time tagged photons along the ground track allows for vertical profiles to be created, on which vegetation and terrain elevations can be computed. MABEL was not intended to be an exact duplicate of ATLAS but was meant to provide the measurement concept and data for algorithm development with the flexibility to explore science and engineering trade spaces (McGill *et al*., 2013). This paper reports work that used MABEL data from one selected channel, and simulated ATLAS data, to investigate the prospects of photon counting lidar in retrieving 3-D vegetation structural attributes in a savanna landscape.

We hope to provide a base for any other photon counting lidar remote sensing work that aims at calculating canopy height, biomass and consequently carbon storage/dynamics in such ecosystems.

2. Methods and materials

2.1 Study Site

This research was conducted in the oak savannas of Tejon Ranch Conservancy (figure 5.1). The 2008 Tejon Ranch Conservation and Land Use Agreement between Tejon Ranch Company and a group of conservation organizations resulted in the creation of this 72 000 Ha conservancy. The conservancy was created to protect the ranch and implement science based stewardship, thus preserving, enhancing and restoring the native biodiversity and ecosystem values of the Tejon Ranch and Tehachapi Range for the benefit of California's future generations (Tejon Ranch Conservancy, 2013). These oak savannas comprise mainly of Blue oaks (*Quercus douglasii*), Black oaks (*Quercus kelloggii*) and Valley oaks (*Quercus lobata*). Other species found in this ecosystem are Canyon live oak (*Quercus chrysolepis*), Interior live oak (*Quercus wislizeni*), the California Buckeye (*Aesculus californica*) and a few conifers. Blue oak woodlands are dominant at the lower elevations (between 500 and 1 000 m), Black oak woodlands are dominant in higher elevation areas (> 1 200 m) while Valley oak woodlands are found on both lower (400- 600 m) and higher (1400- 1800 m) elevations. Grass dominates the understories of Blue and Valley oaks while shrubs are found in combination with grass in the understory of Black oaks.

2.2 MABEL altimetry

We used MABEL data collected during February 2012 day time flights using a NASA ER-2 aircraft flying at an altitude of 20 km. At that altitude the laser footprint diameter is 2 m, substantially smaller than the 14 m footprint planned for ATLAS. At the nominal aircraft speed of 200 m/sec and laser repetition rate of 5,000 pulses per second the spacing between footprints was 4 cm as compared to the ATLAS 70 cm spacing. Because of the high altitude, and the small diameter of the receiver telescope, the probability of detecting (PD) a photon per laser fire reflected from the surface was very low and highly variable between different wavelengths and beam. Although ATLAS's laser will be at 532 nm, in this study a 1064 nm channel was used because for this MABEL campaign damage to the green optical fibres caused unacceptably low signal density. Using a different aerial single photon airborne lidar, Harding *et al*., (2011) showed that the vertical distribution of reflected photons from forest canopies does not substantially differ between 532 and 1064 nm. Therefore, we think the use of 1064 nm MABEL data to simulate the 532 nm ATLAS results is justified. Channel 49 was selected for these analyses since on average it had the highest number of detected photons for the data available (3.6 photons per meter).

Data from 2014 MABEL flights in the ER-2 indicated that the geolocation precision of MABEL was 30 m RMSE, but this figure was derived after significant engineering improvements had been made to MABEL and errors for our 2012 data are likely to be larger. To minimise geolocation error we extracted photons that were classified as terrain (see section 2.3 for details) from the MABEL data and co-registered them to a Digital Elevation Model (DEM) derived from DRL data. The DRL data was collected in July 2012 by a commercial lidar vendor at an average density of 1 point per m^2 and was validated with field data as explained in Gwenzi & Lefsky (2015). Co-registration was performed by comparing root mean square errors for the difference between the terrain elevations for the MABEL and DRL datasets. The MABEL data was shifted in the x and y directions to create a surface of RMSE error as a function of the distance shifted and the shift that resulted in the lowest error was added to the original MABEL coordinates. We also adjusted for elevation errors by adding the median difference between the MABEL and DRL derived terrain elevations to the MABEL elevation points. While this would be unacceptable for a study of absolute elevation, we are only concerned with the relative height of the vegetation and the result of the elevation shift was to allow for visual comparison of the two datasets.

2.3 MATLAS simulations

MATLAS simulation is the process of generating ATLAS-like data by adjusting the number of MABEL's detected number of signal and noise photons to that predicted using ATLAS instrument model design cases. The design cases developed by the ICESat-2 science team provide values for the physical parameters used to model instrument performance so they can be compared to the precision and accuracy needed to meet the mission science requirements. Using the design cases, the MATLAS simulator transforms MABEL data in five ways. 1) MATLAS simplifies the trajectory of the airborne MABEL data (which varies with platform roll, pitch and yaw) to simulate the less variable ATLAS trajectory. 2) The MABEL spatial resolution is degraded to match the larger ATLAS footprints. 3) Photons in the MABEL data are classified into signal and noise classes. 4) MABEL signal photons are subsampled to match the expected ATLAS signal photon density. 5) If simulated background noise levels exceed those observed in the MABEL source data, (i.e. due to varying solar elevation angles) noise levels are adjusted, while retaining the observed spatial variability of solar background noise caused by changing surface reflectance along the flight line.

Two design cases were evaluated for this study, a day time design case and a night time design case. This allow for an investigation of the different contributions of background noise and instrument noise. For forest targets the physical properties described by the design cases include SEA, atmospheric transmission, nominal canopy and terrain bi-directional reflectance values, reflectance multiplier factors for the laser retro-reflectance (i.e. for “hot spot” parallel illumination and view angles), canopy height, leaf area index and terrain slope and roughness. For this study, both the strong and weak beams that ATLAS will use were simulated, but only the strong beams were evaluated for their ability to estimate canopy height. Due to simplifying assumptions in the DC input parameters and modelling method the expected levels are only meant to be approximations for the different forest cover types, not rigorous predictions.

2.3.1 Trajectory simplification and decrease of spatial resolution

As with all airborne data, the MABEL beam tracks on the ground form complex sinusoidal patterns due to changes in aircraft roll, pitch and yaw at various frequencies and amplitudes. To produce the MATLAS simulations the following steps are applied. To produce an ATLAS-like straight track more amenable to simulation, all photons in a 60-second (approximately12 km) MABEL file are projected perpendicularly to a straight-line track defined by a best fit to all the photon latitude and longitude locations. The data is then divided into 14 m long segments.

2.3.2 Classification of MABEL photons

To classify photons in the MABEL data, the lowest terrain elevation within each segment is identified and a section below that elevation is used to identify the mean and standard deviation of total observed background noise, which includes both solar and instrument noise. The segment is divided into 2 m vertical cells and an expected number of noise photons is identified for that cell using average noise statistics and a Poisson distribution. If more than that number of photons is present in a cell, that excess amount is removed by randomly sampling the photons present in the cell. The remaining photons are classified as noise.

This simple method of classification is purposefully meant to not take into account the spatial structure of the signal photon population, as might be done in a more sophisticated surface tracking methodology. One of the purposes for developing MATLAS is to provide realistic ATLAS-like data that retains the statistical properties of the MABEL source data, for the development of algorithms to be used in ATLAS processing. Using a surfacing tracking algorithm in the MATLAS classification could impose a hidden selection bias potentially altering the characteristics of the observed photon point cloud in a way that erroneously influences the development of ATLAS processing algorithms.

2.3.3 Modelling ATLAS signal photons

The number of MABEL signal photos exceeds the number of expected ATLAS photons, so that the photons to be used in the MATLAS simulation can be selected by subsampling the MABEL data. The subsampling factor to select the expected number of ATLAS signal photons from the MABEL signal population is derived from the ratio of ATLAS to MABEL signal photons. The total number of expected ATLAS signal photons is computed as the number of 14 m ATLAS footprints along the track multiplied by the predicted number of signal photons per laser pulse. The number of footprints is equal to the track length divided by the 0.7 m footprint spacing (7 km/sec spacecraft velocity and 10 kHz laser fire rate). The subsampling factor is equal to the ratio divided by 20 because each MABEL signal photon can be observed 20 times due to the overlapping 14 m footprints. For each ATLAS footprint signal photons are randomly selected at the subsampling rate from the MABEL photons encompassed by the footprint. In this way the variability of signal photon density in the segments is preserved while the total number of expected ATLAS photons is matched for the entire track. The elevation of the selected photons are shifted by a random value between ±10 cm so that if a photon is selected more than once each occurrence will have a unique elevation. The latitude and longitude of the selected photons are assigned to the center of the ATLAS footprint, as will be the case for ATLAS data products.

2.3.4 Modelling ATLAS noise photons

MATLAS noise cannot be produced by subsampling because an increased noise rate (as compared to the MABEL source data) is required when simulating a higher SEA than the angle at the time of data collection. Instead solar noise is generated at the predicted rate. In the first step, instrument noise is removed from the noise population. The instrument noise rate is constant and determined for each of the beams from MABEL data collected at night. Using that rate and a random distribution the occurrence of instrument noise photons is determined for each 14 m x 2 m cell. When using MABEL data acquired during the daytime as the simulation source, the number of instrument noise photons is subtracted from the number of noise photons present in the cell leaving the number of solar noise photons. The total number of solar photons in each

14 m segment defines its along track spatial variability introduced by the variation fractional components of sunlit and shaded canopy and terrain surfaces viewed in the instrument pointing direction, and their reflectance.

The solar noise rate predicted for the DC defines the total number of photons that should be observed along the length of the ATLAS track. As in the signal photon case, a scaling factor for the entire track is defined as the total number of predicted solar noise photons vs MABEL photons classified as solar noise. The noise photons are assigned the latitude and longitude of the footprint center at an elevation randomly distributed throughout the 14 m wide column. For a DC simulation with a SEA larger than when the MABEL data was collected, the number of solar noise photons will increase. Conversely, it will decrease for a lower DC SEA. For night time simulations with the sun below the horizon the number of solar noise photons is set to zero. The pattern of noise variability is only meant to be a representative occurrence for the DC land cover type, not a rigorous treatment of what ATLAS will observe for that specific location and SEA. In the MATLAS product each photon is identified as being signal, solar noise or instrument noise.

2.4 Terrain and canopy top identification

Within the transect, the derivation of canopy heights was evaluated within segments of varying resolution (14 m, 25 m, 50 m), herein referred to as blocks. We chose the 14 m block size to represent the planned ICESat-2's footprint size. The 25 m block size was chosen to compare with the footprint size of a successful medium resolution airborne lidar sensor, the Laser Vegetation Imaging Sensor (LVIS) (Blair, *et al*., 1999) and a recently approved mission, the Global Ecosystem Dynamics Investigation (GEDI) (Stysley *et al*., 2015) lidar. We evaluated the 50 m block size to compare with ICESat-GLAS and determine the effect of analysis at a much coarser resolution. For each block length, photons were aggregated into a histogram at 0.5 m vertical resolution. The raw data for each channel has a large quantity of noise photons collected below and above the signal photons. We considered those photons whose elevation was within the 2.5 σ of the mean elevation to be potential signal photons. The histogram for each block was used to derive two height metrics for that corresponding block: H_{max} defined as the maximum canopy height minus mean terrain elevation and H_{90} defined as the 90th percentile canopy height minus mean terrain elevation without a cut off threshold.

As a preliminary analysis, we selected blocks for which the vertical distribution of the photons showed clear breaks, (i.e. a likely canopy top and terrain elevation). These breaks were then used to define average values for the percentage of photons associated with the canopy top and mean terrain elevation. Percentages were calculated relative to the highest elevation for canopy top and relative to the lowest elevation for mean terrain elevation. From the histograms, we found out that on average for all the block sizes, the break for canopy top corresponded to the elevation of the lowest bin among the top 2.5 % photons. The mean terrain elevation corresponded to the elevation of the most numerous bin among the bottom 20 %, 25 % and 27.5 % of photons for the 14 m, 25 m and 50 m block sizes respectively. On the basis of these results we then implemented an algorithm (figure 5.2) in R (R Core Team, 2014) that used these percentages to identify the canopy top and mean terrain elevation for each block at each of the three resolutions.

2.5 Analysis

In this work, we present results from the transect that had the highest variability in terms of vegetation cover and terrain relief (Transect 5 in figure 5.1). Height metrics calculated from the MABEL and MATLAS data were validated by comparing them with the equivalent metrics derived from the DRL data. From the DRL data we created a Digital Terrain Model and maximum and 90th percentile canopy height Digital Surface Models at 2 m resolution using LAStools (Isenburg, 2015). To extract validation height metrics, the transect ground track was divided into blocks coincident with those used in MABEL/MATLAS data analysis, with a cross-track buffer of 2 m to represent the approximate footprint diameter of MABEL. As with the MABEL/MATLAS data height extraction algorithm, H_{max} and H_{90} for each block were obtained by subtracting the mean terrain elevation from the maximum and 90th percentile canopy elevations respectively. We used r^2 and RMSE statistics to determine the deviation of the MABEL derived metrics from the DRL derived metrics. We removed all the points for which the residuals showed clearly unrealistic values, presumably due to processing error. The maximum allowable residual was calculated as the sum of the mean tree height (11 m) observed during field work (Gwenzi & Lefsky, 2015) and the average vertical extent of terrain, obtained using the block size and the study area's mean terrain slope (22 degrees). Figures comparing MABEL/MATLAS and DRL derived height metrics show the residual points, which are plotted in grey.

Data co-registration, height calculation, vegetation prolife generation and validation procedures were done for MATLAS simulated day and night data in the same manner as those for MABEL.

We computed signal and noise photons statistics for channel 49 of transect 5 to show the expected differences in data quality between MABEL and ATLAS.

3. Results

3.1 MABEL

Profiles of transect 5 at a densely vegetated and open canopy areas at the three different block sizes are presented as Figures 5.3 and 5.4. Profiles at the 25 m and 14 m block sizes clearly had a better representation of the vegetation than those at 50 m, primarily due to the smaller contribution of terrain variability for the smaller block sizes. Smaller block sizes are also better at matching the spatial scale of the canopy. The MABEL derived height metrics were moderately correlated to DRL derived metrics with only slight differences in the quality of estimates for the H_{max} and H_{90} metrics (Figure 5.5). The correlation coefficient between the MABEL and DRL metrics was highest at the 25 m block size and lowest at the 14 m block size implying that the spatial resolution of analysis has a significant influence on the results. However, the meaningful MABEL profiles at 14 m and the DRL data’s strong agreement with field conditions (Gwenzi & Lefsky, 2015) suggests that the low agreement between MABEL and DRL height metrics at this block size is mainly due to the spatial resolution of analysis getting finer than the geolocation error of the MABEL data.

3.2 MATLAS

The MATLAS simulation process showed that the quantity of signal photons will be substantially reduced (relative to the MABEL data) for both day and night time acquisitions

(Table 5.1). These are average values computed for channel 49 of transect 5 at 50 m block sizes. Of the two MATLAS design cases, night data has a higher ratio of signal to noise photons, but the accuracy of the height metrics derived from the night data was just slightly better (slightly lower RMSE) than those from day data implying that the most important factor for the algorithm we implemented is the density of photons per block. MABEL does not only have a higher ratio of signal to noise photons but also has a high total photon density, which is about 6 and 7 times greater than that for the MATLAS day and night data respectively. Solar background contributes much of the noise in the MATLAS day data which reduces the ratio of signal to noise photons to a value below 1 compared to as high as 10.26 for MATLAS night and 76.41 for MABEL data. Because of the reduced photon density, the vegetation profiles (Figure 5.6-5.7; 5.9-5.10) from MATLAS data are less representative compared to those of MABEL and the correlations between the MATLAS derived height metrics and DRL metrics are poorer as shown in table 5.2 and figures 5.8 and 5.11.

4. Discussion

Our algorithm provided encouraging height derivation results but were not impressive when compared to other high spatial resolution lidar techniques. We believe that better validation results could have been obtained with more accurately longitude-latitude geolocated photons. Since our cut-off values for canopy top and mean ground elevation derivation were empirically generated, we expect them to vary as a function of vegetation and terrain factors like canopy structure, mean stem density and slope variability. As such, other variables may need to be considered for a large scale application. In this paper we present a first cut method suitable for evaluation, not a full blown algorithm.

Our results were likely affected by the differences in vegetation phenology over the different seasons in which the MABEL and DRL data were collected. MABEL data was collected in February, a season during which the majority of tree species in the study site are leaf off and grass growth is at peak. DRL data was collected in July, a season within opposite phenology- leaf on trees and dry grass. The consequences of these differences are twofold: MABEL penetrated the canopy more under leaf off conditions and hence had more terrain returns in densely vegetated deciduous tree areas where DRL may not have penetrated well under leaf on conditions. On the other hand, MABEL under leaf off conditions had a higher probability of missing the canopy tips where DRL had higher chances of capturing these tips as surface area is greatly increased by the leaf on conditions. These effects would be compounded because the lower pulse energy of MABEL is less likely to record energy from the very highest elevations and more likely to penetrate to the terrain. Moreover different herbaceous vegetation grows on the open areas over these different seasons in the study area, which can also contribute to the differences in the terrain elevations in the sparse sections of the transect obtained from each data set.

Profiles drawn at the 50 m block size do not adequately represent the vegetation along the transect because of the high heterogeneity of both relief and vegetation within this area. The variability of these two factors is finer than the aggregation length of 50 m. The less dramatic change in the visual appearance of the vegetation profiles from 25 to 14 m block sizes and the reduced accuracy from 25 m to 14 m block size suggests that for MABEL/MATLAS data, 25 m is the best aggregation length to use. This concurs with waveform lidar studies where 25 m footprint data from LVIS (Blair *et al*., 1999; Drake *et al*., 2002; Anderson *et al*., 2006) provided better results compared to GLAS data with footprint sizes greater than 50 m in diameter.

The spatial resolution of MATLAS data is limited due to the large, overlapping footprints that remove the specifics of the location of each received photon. It is possible that photons returned from later pulses have come from behind (relative to along-track travel) the photons received earlier. In single beam simulations, as used in this study, only the along-track component of reduced resolution is introduced. The cross-track resolution remains equal to the 2 m MABEL footprint diameter. This does not does give a true estimation of the sampling capability of the ATLAS sensor across track. Better MATLAS simulation results could have been obtained with a MABEL cross-track resolution that matches the 14 m ATLAS resolution. The MATLAS simulator includes the creation of a composite product that combines several adjacent MABEL beams to enlarge the simulated footprint size in the cross-track direction. However, composite products were not used in this study because a method is not available yet that addresses the complexly intersecting nature of the MABEL sinusoidal ground tracks which introduce inconsistent simulated footprint widths along track.

The MATLAS simulation results suggest that the actual data from ICESat-2's ATLAS sensor will give poorer results than MABEL. The higher background noise levels from a spaceborne platform and combination of lower sampling rates and larger footprints will reduce ICESat-2 data's reliability. Having few photons in areas that already have low vegetation cover makes it difficult to characterize the vertical distribution of the vegetation at small aggregation lengths.

Increasing the aggregation block size will also increase the variability in both relief and tree height within each block.

Full waveform lidar has already proven to provide good results for vegetation and the best block size of 25 m identified in this work suggests that medium resolution future missions like GEDI (http://science.nasa.gov/missions/gedi/) with a canopy height measurement RMSE target of 1 m will be a better option than ICESat-2 for biomass studies. GEDI will measure the biophysical attributes of vegetation areas between 50° North and South latitudes using 3 High Output Maximum Efficiency Resonator (HOMER) lasers, whose beams will be divided into 14 parallel tracks of 25 m contiguous footprints on the ground (Stysley et al., 2015)

5. Conclusions

For such a structurally complex savanna system, the results we obtained from MABEL data are encouraging but it appears unlikely that ICESat-2 will provide the kind of data required for a reliable mapping of the biophysical properties of savanna vegetation. However, this does not rule out the use of single photon lidar as a technique, provided a sensor with a higher number of observed photons is used. Current resources may be better off concentrated on preparing for the GEDI mission, scheduled to launch by the end of this decade. ICESat-2 will however still be useful for those latitudes not covered by GEDI or as a necessary bridge with future missions to ensure continuity of monitoring canopy height, biomass and carbon stocks using spaceborne lidar.

6. Tables and figures

Table 5.1: Summaries of the MABEL and MATLAS design cases data. The dashes are a result of no data because at the moment, MABEL noise data is not sub-classified into instrument and background noise and for MATLAS night data background noise is assumed to be zero

Data variable	Data source		
	MABEL	MATLAS day	MATLAS night
Average number of total photons per 50 m block	710	124	97
Average number of signal photons per 50 m block	701	89	89
Average number of noise photons per 50 m block	9	124	9
Contribution of instrument noise to total noise	-	7 %	100 %
Contribution of background noise to total noise	-	93 %	0 %
Ratio of signal to total noise photons	76.41	0.72	10.26
Ratio of signal to instrument noise photons	-	10.6	10.26
Ratio of signal to background noise photons	-	0.78	-

Table 5.2: MABEL and MATLAS height metrics validation results

Height Metric	Data source	Block size and validation results					
		50		25		14	
		R^2 (%)	RMSE (m)	R^2 (%)	RMSE (m)	R^2 (%)	RMSE (m)
H_{max}	MABEL	74	4.1	75	3.2	64	3.0
	MATLAS night	71	4.4	69	3.6	62	3.1
	MATLAS day	73	4.6	70	3.5	58	3.3
H_{90}	MABEL	71	4.2	72	3.0	60	2.9
	MATLAS night	67	4.4	69	3.1	56	3.0
	MATLAS day	64	4.6	64	3.3	50	3.3

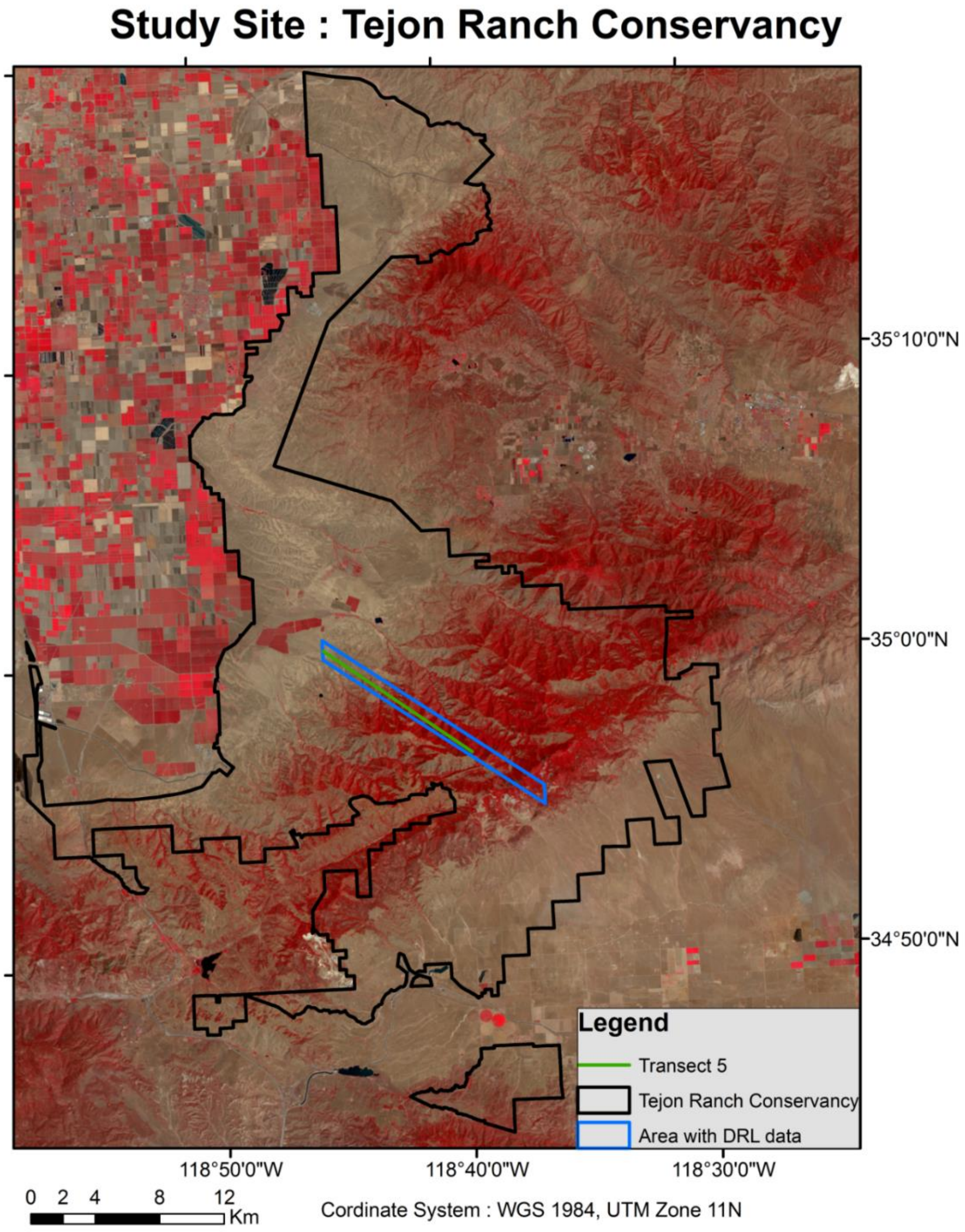
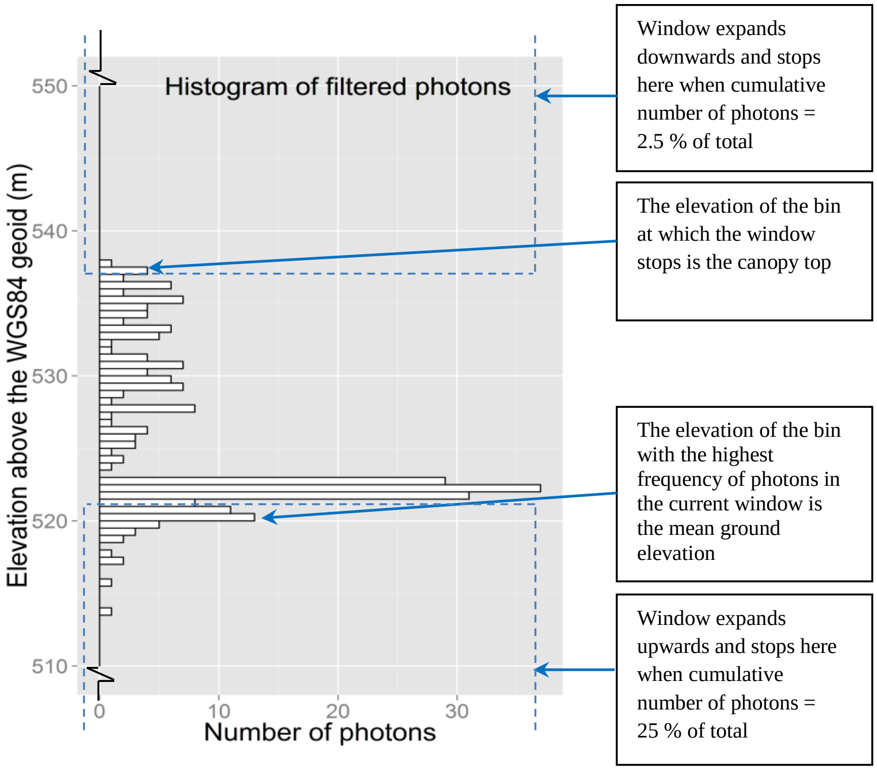


Figure 5.1. Map showing Landsat TM false color (432) image of study area and location of DRL data extent and transect 5.



Window expands downwards and stops here when cumulative number of photons = 2.5 % of total

The elevation of the bin at which the window stops is the canopy top

The elevation of the bin with the highest frequency of photons in the current window is the mean ground elevation

Window expands upwards and stops here when cumulative number of photons = 25 % of total

Figure 5.2. Diagrammatic representation of the height calculation algorithm (an example of a 25 m block from transect 5)

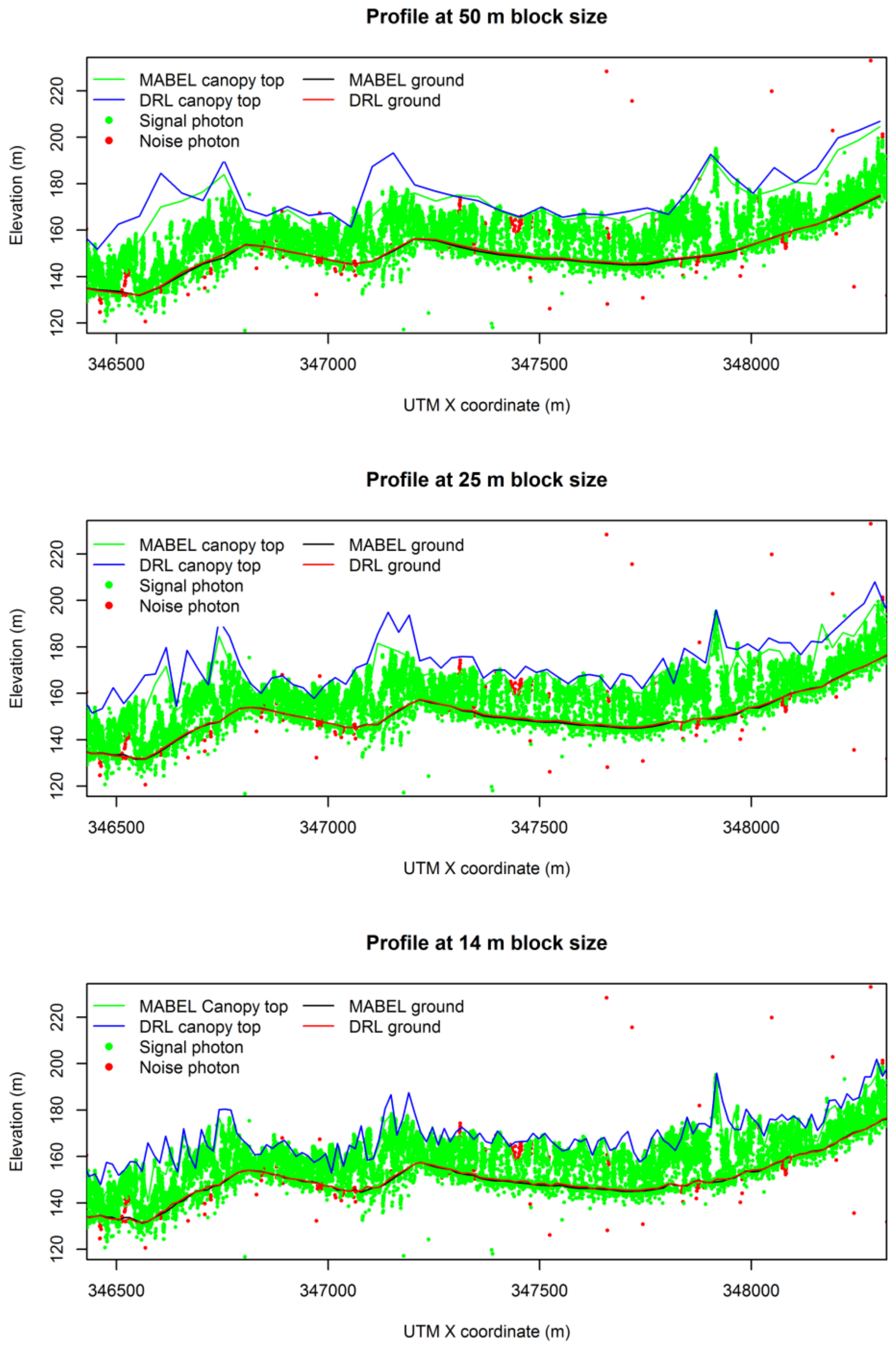


Figure 5.3: Vegetation profiles crated from MABEL data and their comparison to DRL profiles on a typical dense vegetation area along transect 5. Terrain elevation has been reduced by a factor of 10 to enhance vertical variability

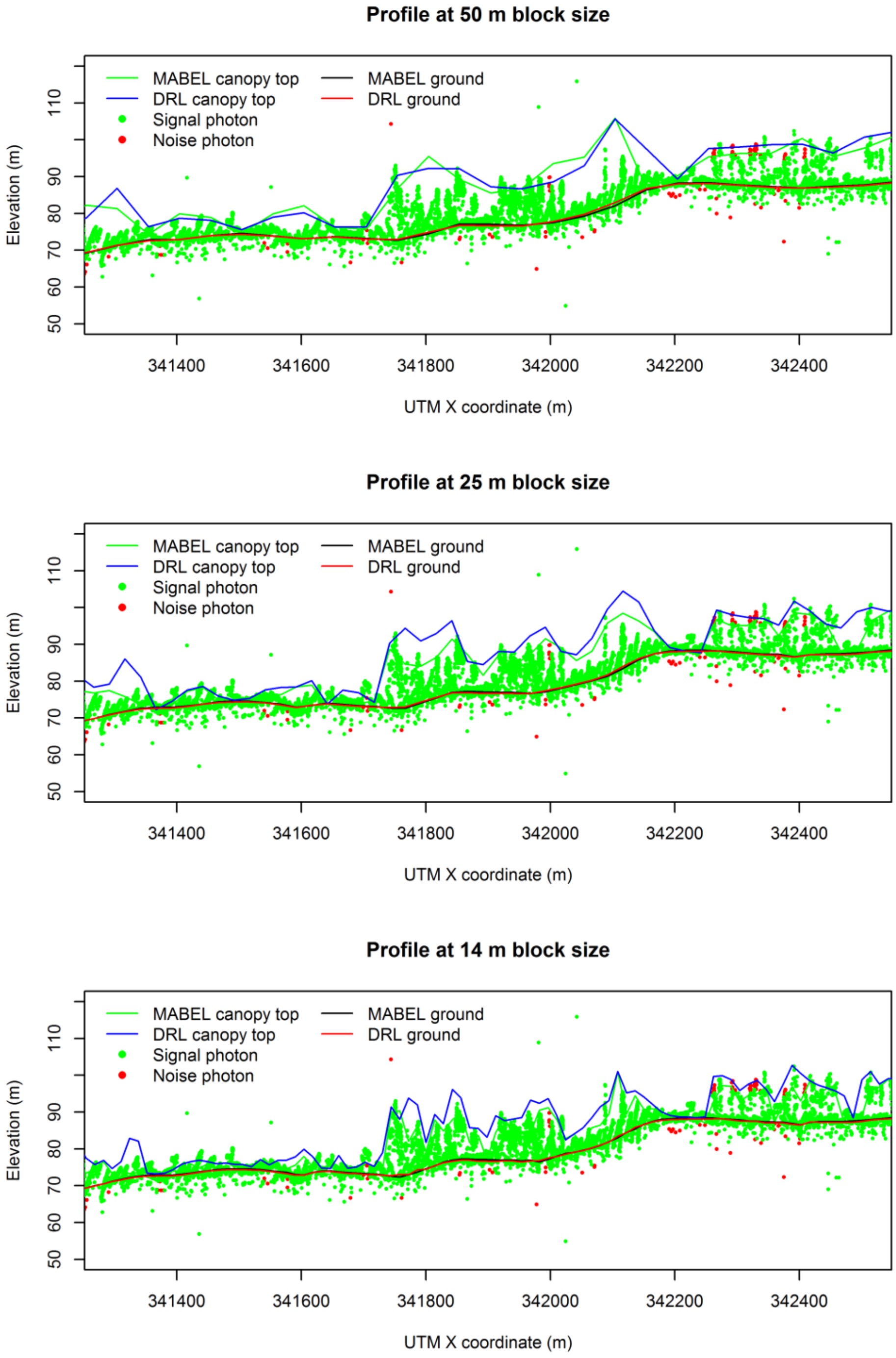


Figure 5.4: Vegetation profiles crated from MABEL data and their comparison to DRL profiles on a typical dense vegetation area along transect 5. Terrain elevation has been reduced by a factor of 10 to enhance vertical variability

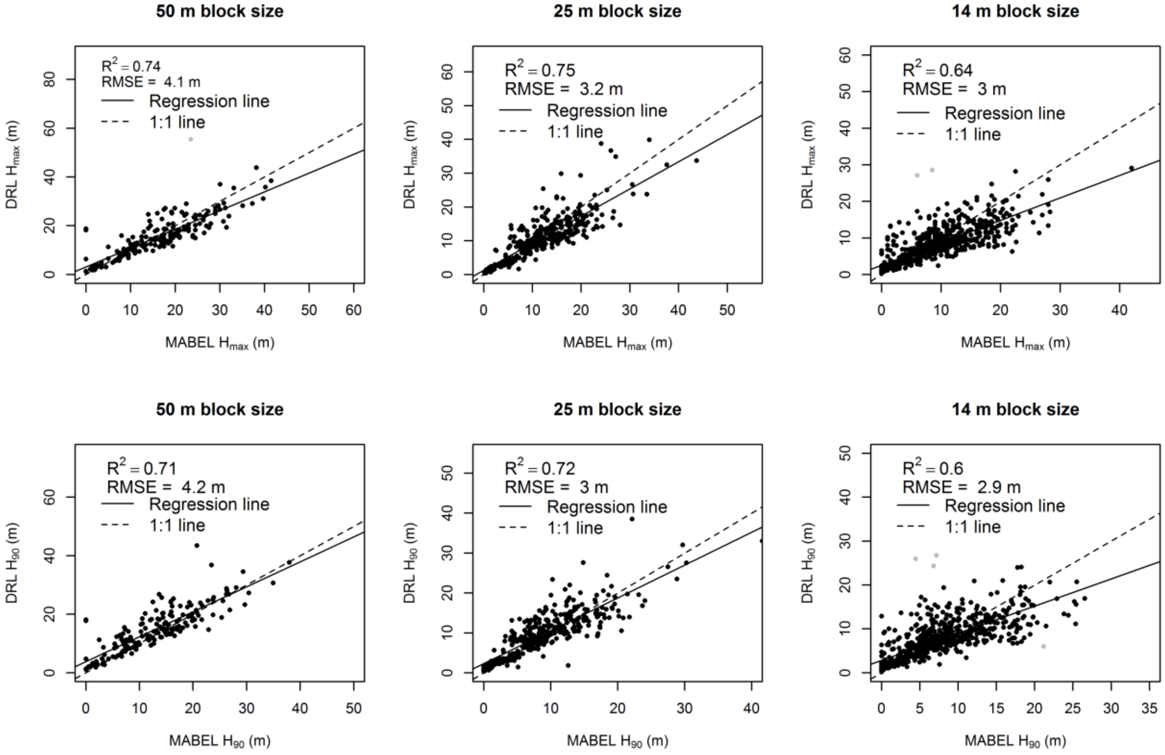
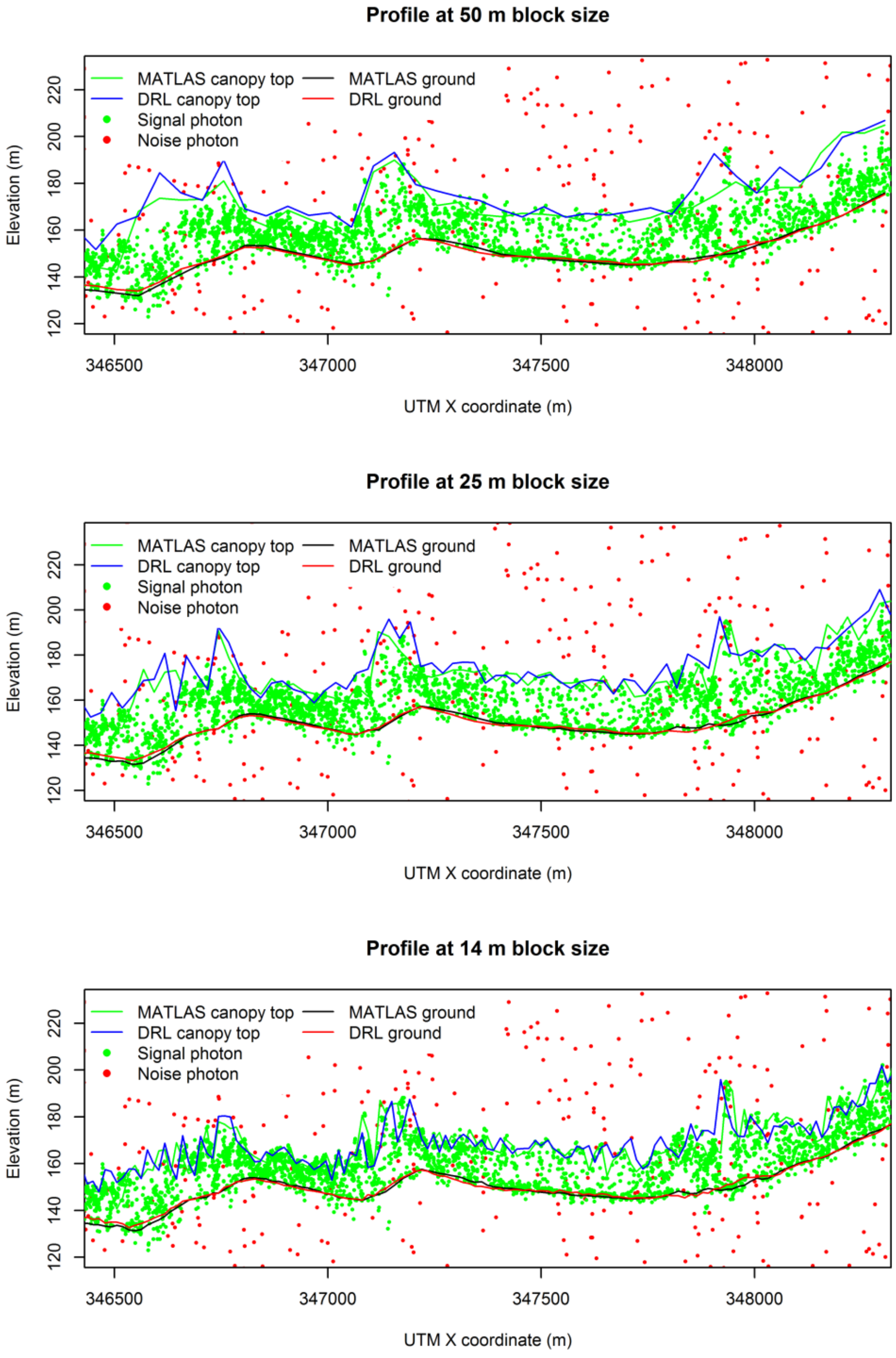


Figure 5.5: MABEL canopy height metrics validation. The gray points were considered outliers and excluded from analysis as explained in the methods section.

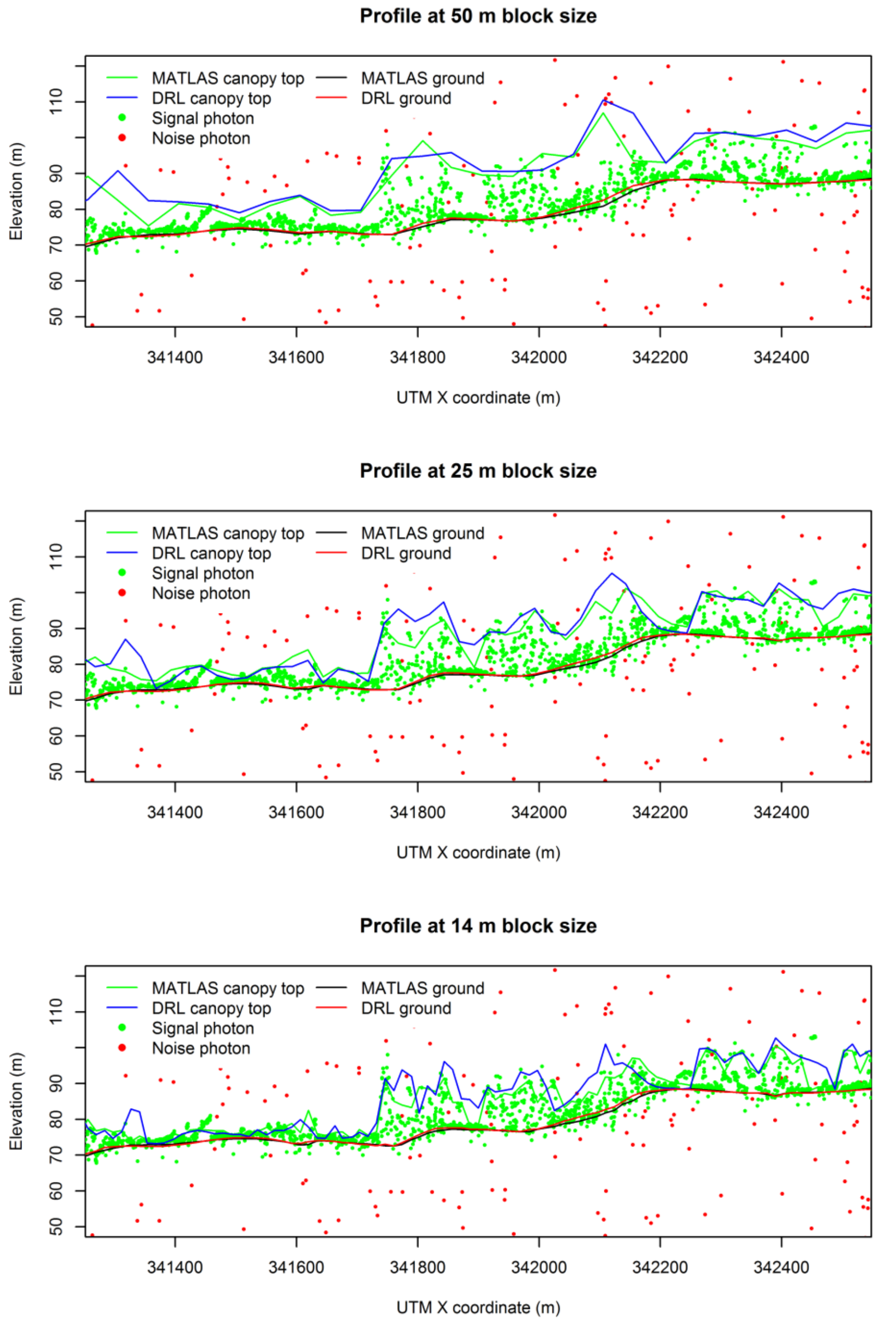


Profile at 50 m block size

Profile at 25 m block size

Profile at 14 m block size

Figure 5.6: Vegetation profiles crated from MATLAS day data and their comparison to DRL profiles on a typical dense vegetation area along transect 5. Terrain elevation has been reduced by a factor of 10 to enhance vertical variability


Profile at 50 m block size

Profile at 25 m block size

Profile at 14 m block size

Figure 5.7: Vegetation profiles crated from MATLAS day data and their comparison to DRL profiles on a typical open canopy area along transect 5. Terrain elevation has been reduced by a factor of 10 to enhance vertical variability

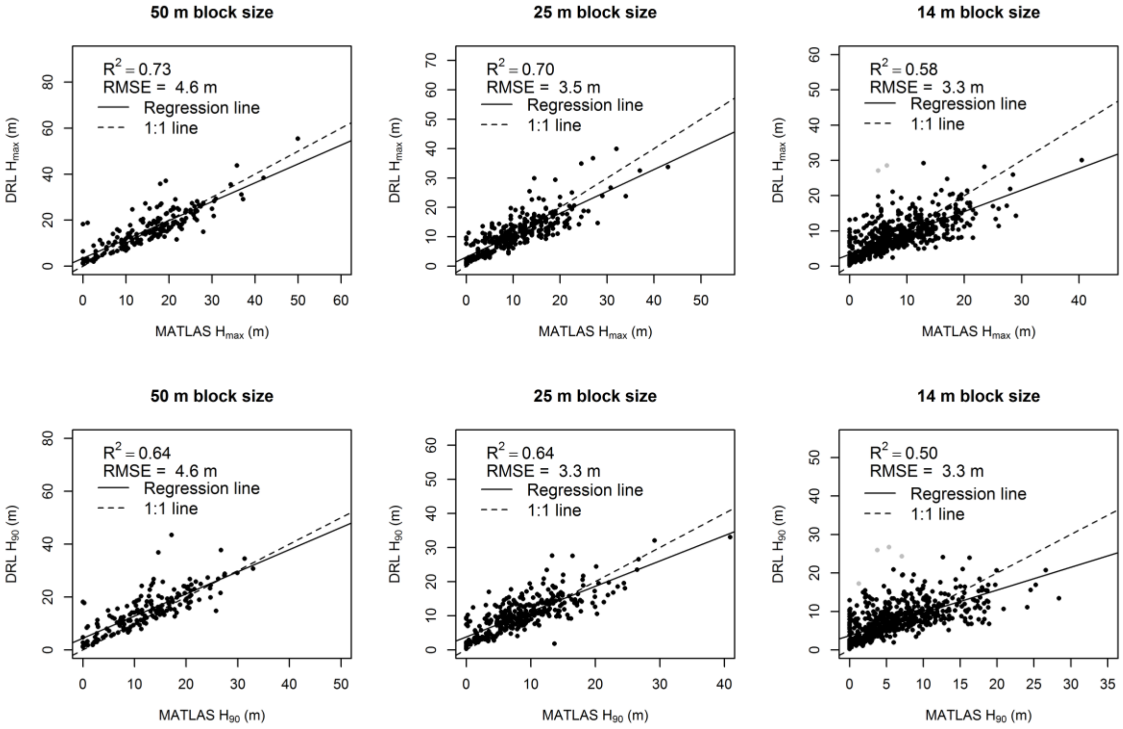
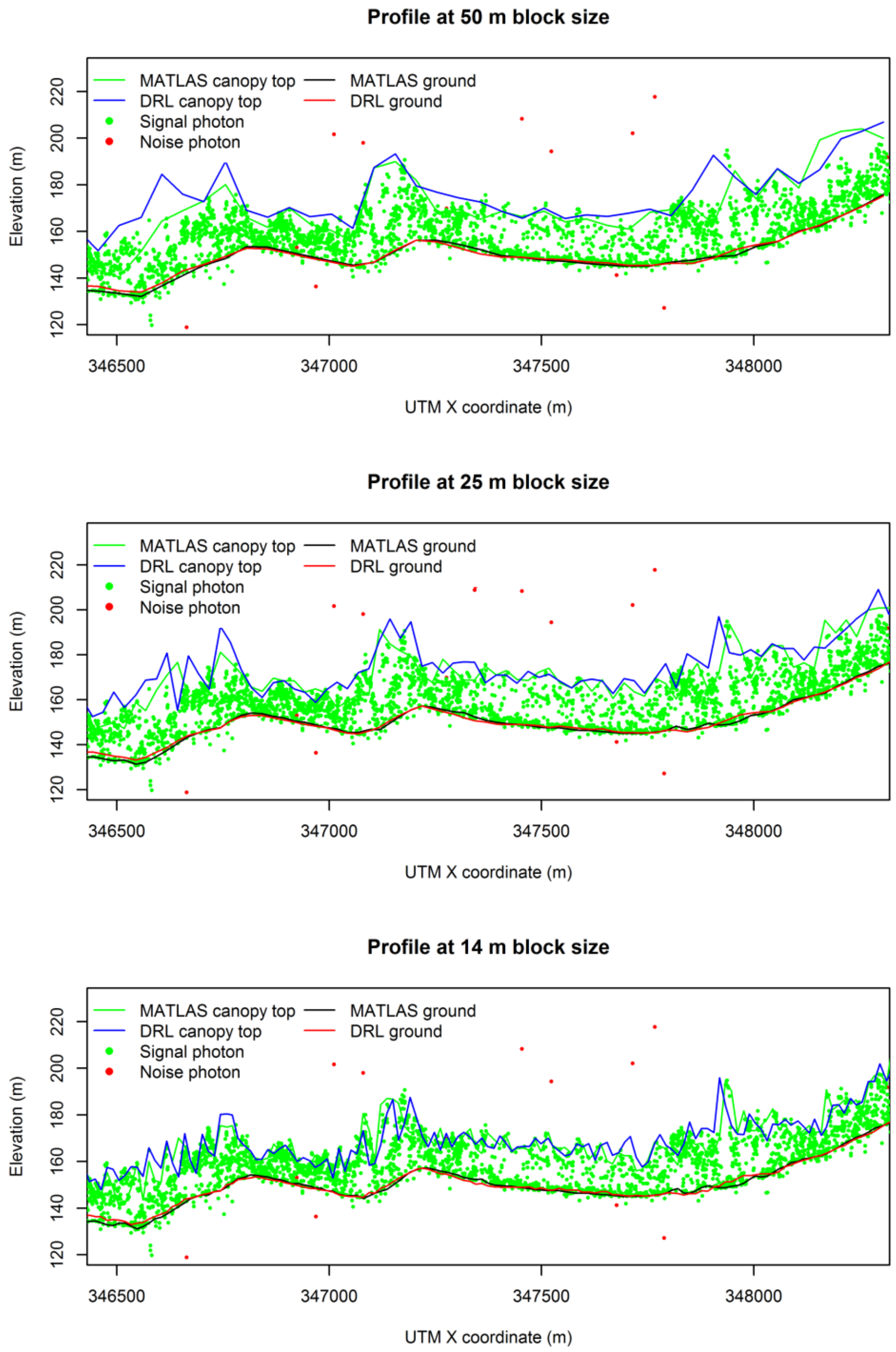


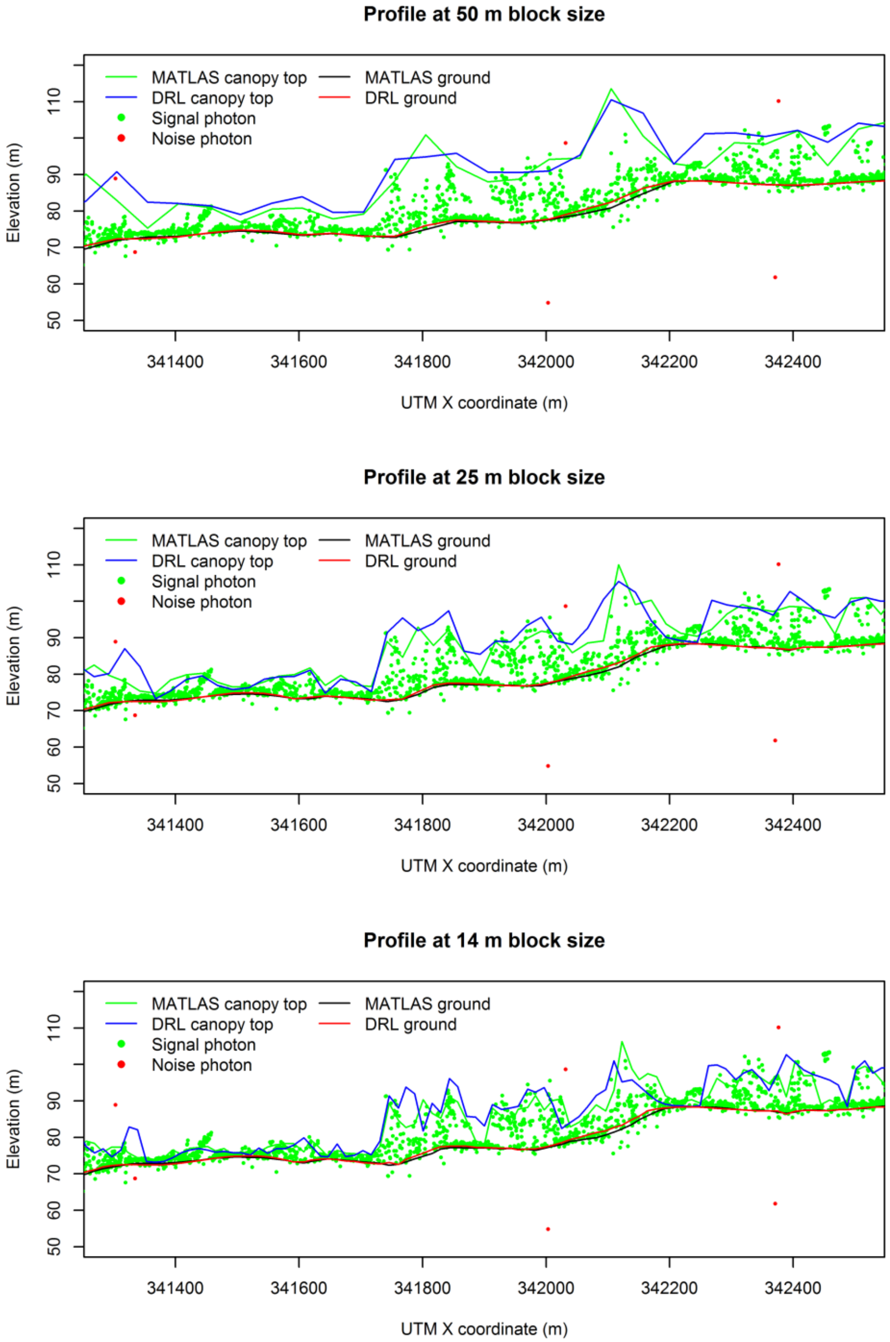
Figure 5.8: MATLAS day data height metrics validation. The gray points were considered outliers and excluded from analysis as explained in the methods section.


Profile at 50 m block size

Profile at 25 m block size

Profile at 14 m block size

Figure 5.9: Vegetation profiles crated from MATLAS night data and their comparison to DRL profiles on a typical dense vegetation area along transect 5. Terrain elevation has been reduced by a factor of 10 to enhance vertical variability



Profile at 50 m block size

Profile at 25 m block size

Profile at 14 m block size

Figure 5.10: Vegetation profiles crated from MATLAS night data and their comparison to DRL profiles on a typical open canopy area along transect 5. Terrain elevation has been reduced by a factor of 10 to enhance vertical variability

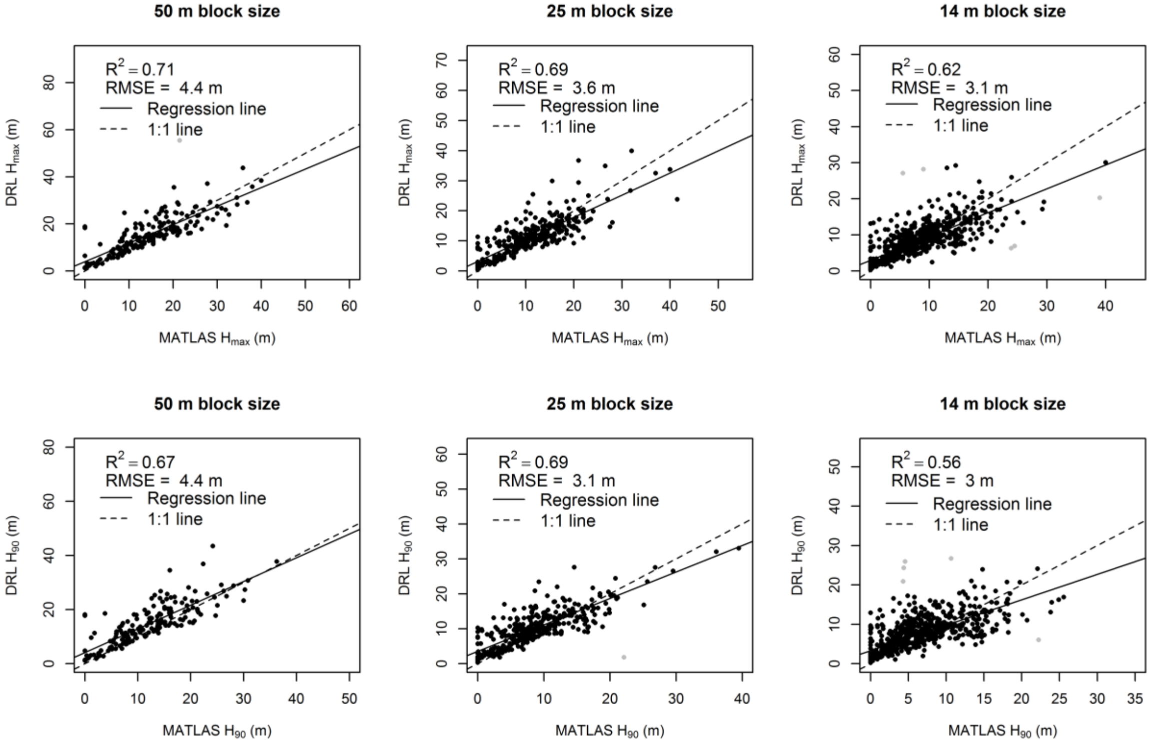


Figure 5.11: MATLAS night data height metrics validation. The gray points were considered outliers and excluded from analysis as explained in the methods section.

CHAPTER 5 REFERENCES

Abdalati, W., Zwally, H. J., Bindschadler, R., Csatho, B., Farrell, S. L., Fricker, H. A., … Webb, C. (2010). The ICESat-2 Laser Altimetry Mission. *Proceedings of the IEEE, 98*(5), 735–751.

Anderson, J., Martin, M. E., Smith, M.-L., Dubayah, R. O., Hofton, M. A., Hyde, P., … Knox, R. G. (2006). The use of waveform lidar to measure northern temperate mixed conifer and deciduous forest structure in New Hampshire. *Remote Sensing of Environment*, *105*(3), 248–261 (14).

Blair, J. B., Rabine, D. L., & Hofton, M. a. (1999). The Laser Vegetation Imaging Sensor: a medium-altitude, digitisation-only, airborne laser altimeter for mapping vegetation and topography. *ISPRS Journal of Photogrammetry and Remote Sensing*, *54*(2-3), 115–122.

Drake, J. B., Dubayah, R. O., Knox, R. G., Clark, D. B., & Blair, J. B. (2002). Sensitivity of large-footprint lidar to canopy structure and biomass in a neotropical rainforest. *Remote Sensing of Environment*, *81*(2-3), 378–392.

Duncanson, L. I., Niemann, K. O., & Wulder, M. A. (2010). Estimating forest canopy height and terrain relief from GLAS waveform metrics. *Remote Sensing of Environment*, *114*(1), 138–154.

Gwenzi, D., & Lefsky, M. A. (2014). Modeling canopy height in a savanna ecosystem using spaceborne lidar waveforms. *Remote Sensing of Environment*.

Gwenzi, D., & Lefsky, M. A. (2015). Plot level aboveground woody biomass modeling using canopy height and auxiliary remote sensing data in a heterogeneous savanna. *Journal of Applied Remote Sensing*, (Submitted).

Hall, F. G., Bergen, K., Blair, J. B., Dubayah, R., Houghton, R., Hurtt, G., … Wickland, D. (2011). Characterizing 3D vegetation structure from space: Mission requirements. *Remote Sensing of Environment*, *115*(11), 2753–2775.

Harding, D. J., Dabney, P. W., Valett, S., He, X., Xu, J., & Ferreira, V. (2011). Polarimetric, two-color, photon-counting laser altimeter measurements of forest canopy structure. In *International Symposium on Lidar and Radar Mapping 2011: Technologies and Applications*.

Isenburg, M. (2015). LAStools: award-winning software for rapid LiDAR processing. Retrieved April 2, 2015, from http://www.cs.unc.edu/~isenburg/lastools/

Lefsky, M. A. (2010). A global forest canopy height map from the Moderate Resolution Imaging Spectroradiometer and the Geoscience Laser Altimeter System. *Geophysical Research Letters*, *37*(15), 1–5.

Lefsky, M. A., Keller, M., Pang, Y., Camargo, P. B. de, & Hunter, M. O. (2007). Revised method for forest canopy height estimation from Geoscience Laser Altimeter System waveforms. *Journal of Applied Remote Sensing*, *1*.

McGill, M., Markus, T., Scott, V. S., & Neumann, T. (2013). The Multiple Altimeter Beam Experimental Lidar (MABEL): An Airborne Simulator for the ICESat-2 Mission. *Journal of Atmospheric and Oceanic Technology*, *30*(2), 345–352.

R Core Team. (2014). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. Vienna, Austria.

Rosette, J., Field, C., Nelson, R., Cook, B., DeCola, P., & Degnan, J. (2011). Single-Photon LiDAR for Vegetation Analysis.

Simard, M., Pinto, N., Fisher, J. B., & Baccini, A. (2011). Mapping forest canopy height globally with spaceborne lidar. *Journal of Geophysical Research*, *116*(G04021), 1–12.

Stysley, P. R., Coyle, D. B., Kay, R. B., Frederickson, R., Poulios, D., Cory, K., & Clarke, G. (2015). Long term performance of the High Output Maximum Efficiency Resonator (HOMER) laser for NASA′s Global Ecosystem Dynamics Investigation (GEDI) lidar. *Optics & Laser Technology*, *68*, 67–72.

Tejon Ranch Conservancy. (2013). Tejon Ranch Conservancy. Retrieved April 2, 2015, from http://www.tejonconservancy.org/

Xing, Y., de Gier, A., Zhang, J., & Wang, L. (2010). An improved method for estimating forest canopy height using ICESat-GLAS full waveform data over sloping terrain: A case study in Changbai mountains, China. *International Journal of Applied Earth Observation and Geoinformation*, *12*(5), 385–392.