# BIOMASS DENSITY BASED ADJUSTMENT OF LIDAR-DERIVED DIGITAL ELEVATION MODELS: A MACHINE LEARNING APPROACH

by

KHALID ABDELWAHAB

B.S. University of Central Florida, 2016

A thesis submitted in partial fulfillment of the requirements
for the degree of Master of Science
in the Department of Civil, Environmental and Construction Engineering
in the College of Engineering & Computer Science
at the University of Central Florida
Orlando, Florida

Summer Term

2019

# ABSTRACT

Salt marshes are valued for providing protective and non-protective ecosystem services. Accurate digital elevation models (DEMs) in salt marshes are crucial for modeling storm surges and determining the initial DEM elevations for modelling marsh evolution. Due to high biomass density, lidar DEMs in coastal wetlands are seldom reliable. In an aim to reduce lidar-derived DEM error, several multilinear regression and random forest models were developed and tested to estimate biomass density in the salt marshes near Saint Marks Lighthouse in Crawfordville, Florida. Between summer of 2017 and spring of 2018, two field trips were conducted to acquire true elevation and biomass density measures. Lidar point cloud data were combined with vegetation monitoring imagery acquired from Sentinel-2 and Landsat Thematic Mapper (LTM) satellites, and 64 field biomass density samples were used as target variables for developing the models. Biomass density classes were assigned to each biomass sample using a quartile approach. Moreover, 346 in-situ elevation measures were used to calculate the lidar DEM errors. The best model was then used to estimate biomass densities at all 346 locations. Finally, an adjusted DEM was produced by deducting the quartile-based adjustment values from the original lidar DEM. A random forest regression model achieved the highest pseudo $R^2$ value of 0.94 for predicting biomass density in $g/m^2$. The adjusted DEM based on the estimated biomass densities reduced the root mean squared error of the original DEM from 0.38 m to 0.18 m while decreasing the raw mean error from 0.33 m to 0.14 m, improving both measures by 54% and 58%, respectively.

# ACKNOWLEDGMENT

I would first like to express my gratitude to my supervisor, Dr. Stephen C. Medeiros, for supporting me throughout my graduate journey. His expertise was invaluable in the production of this work. His work ethic and mentorship style will have a lasting impact on my professional career and personal life.

Secondly, I would like to thank my colleagues from Louisiana State University and the University of South Carolina for their help in the collection and processing of the data used in this work.

Finally, I would like to thank my parents for patiently supporting me throughout my academic journey, my wife for always comforting and encouraging me and my brothers for always having my back.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# CHAPTER I: INTRODUCTION

Salt marshes provide protective and non-protective ecosystem services. The first section of this chapter aims to provide the reader with background information about these coastal systems, explain how these systems are evolving under current climatic conditions, provide an economic valuation based on the literature, show how these systems are monitored by coastal engineers using Digital Elevation Models (DEMs) and explain the need for more accurate DEMs in salt marshes. The second section is aimed at demonstrating the connection between reduced DEM accuracy and biomass density in these densely vegetated areas while showcasing two notable publications that address the issue. Finally, the third section introduces machine learning and describes relevant literature.

## 1.1 Background

### 1.1.1 Salt Marshes as Ecosystems

Salt Marshes are complex ecosystems in which aquatic, marine and terrestrial organisms coexist (Pomeroy & Wiegert, 2012). These ecosystems are found around the globe in areas of middle to high latitudes and are estimated to cover an area of 0.36 million square kilometers (Mitsch, Gosselink, Zhang, & Anderson, 2009). In their book, Pomeroy and Wiegert describe salt marshes as plastic coastal features that form under the protection of barriers which suppress high wave energy. Furthermore, they clarify that once a salt marsh is fully developed, it becomes resilient to oceanic exposure (Pomeroy & Wiegert, 2012). From a hydrologic viewpoint, these systems are crucial for maintaining a stable shoreline, that is efficient in dissipating storm surges and absorbing

floodwaters (E. B. Barbier et al., 2011). Naturally, coastal systems have the ability to counterbalance the effect of normal sea level rise (Fagherazzi et al., 2012). For the most part of the last two millennia, sea level rose at a rate of one millimeter per year, and coastal wetlands maintained a state of equilibrium (James T Morris, Sundareshwar, Nietch, Kjerfve, & Cahoon, 2002). However, anthropogenic activities that contribute to the accelerated sea level rise reduce the resiliency of coastal wetlands (Donnelly & Bertness, 2001).

### 1.1.2 Hydrogeomorphology of Salt Marshes

In coastal regions, salt marshes are regarded as natural defense systems. Therefore, it is vital to understand their hydrogeomorphology. The performance of models that simulate storm events, or normal tide cycles, primarily depends on the accuracy of digital elevation models (S. Medeiros, Hagen, Weishampel, & Angelo, 2015). Various factors affect the evolution of salt marsh topography. On one hand, sea level is rising; on the other, marsh platforms are changing in elevation. Salt marshes can achieve a constant relative sea level (RSL) when they are in a state of stable equilibrium (James T Morris et al., 2002).

In coastal systems, changes in RSL arise from changes in mean sea level (MSL) and marsh platform (also known as marsh table) elevation. When sea levels rise at a rate higher than that of the marsh platform, marine transgression is said to occur; on the other hand, landward migration takes place when the rates of accretion are higher than the rise of MSL (Priest, 2011). Marsh table elevation increase can be driven by biogenic accretion or sediment deposition while its decrease can be due to SLR and land subsidence (James T Morris et al., 2002; Morton, Bernier, & Barras,

2006). This interplay between physical and biological processes has prompted researchers to develop models that integrate biological and physical feedbacks.

The increase in the rate of relative sea level rise (RSLR), due to anthropogenic effects, is a threat to salt marshes (James T Morris et al., 2002; Priest, 2011). In fact, it is estimated that 20-45% of the total area of coastal wetlands will be lost by the year 2100; that number increases to 70% when the contribution of anthropogenic activities is considered (Craft et al., 2009; Nicholls, Hoozemans, & Marchand, 1999). MSL is expected to rise by 0.5 to 1.6 meters by 2100 (Jevrejeva, Moore, & Grinsted, 2010; Rahmstorf, 2007). Moreover, land subsidence caused by the loss of upstream sediment and resource extraction (e.g. water, oil and natural gas) has resulted in the conversion of thousands of square kilometers of marsh land to open water in the Gulf of Mexico in Northwestern Florida, Mississippi River Delta and coastal Texas (Coplin & Galloway, 1999; Kesel & Sciences, 1988; S. Medeiros et al., 2015). Although the elevation range in salt marsh systems is relatively small (about 2 meters) (Mckee & Patrick, 1988), even slight changes in ground elevation can have enormous impacts on the overall health of these systems (Hladik & Alber, 2012).

Vertical accretion can be due to organic or inorganic matter. Vertical accretion caused by organic matter is linked to the productivity of the marshes; low marshes' organic matter contributes to vertical accretion as much as inorganic sediments while high marshes' organic matter contributes twice as much as inorganic sediments (Bricker-Urso, Nixon, Cochran, Hirschberg, & Hunt, 1989). High marsh areas are those areas farther away from creeks and shorelines. Due to their relatively high platform elevation, less sediment is deposited in those areas by waves; therefore, biogenic

accretion is the dominant driver of marsh table elevation (Friedrichs & Perry, 2001). On the other hand, in low marsh areas, sediment deposition is the main driver of vertical accretion. Within the low marsh zone, 70-80% of deposition takes place in flocculated form (i.e. minute particles form into larger clumps before settling); larger sediment particles (larger than 20 μm) tend to settle as individual particles (Christiansen, Wiberg, & Milligan, 2000). As wave energy dissipates into the high marsh zone, where biogenic accretion is the main driver, inorganic particles tend to settle without flocculation (Christiansen et al., 2000). Furthermore, organic and inorganic sediment deposition rates can be altered by changes in the concentrations of these particles within the proximity of the marsh area (Friedrichs & Perry, 2001). Changes in sediment concentration can be attributed to changes in factors like tidal velocity and off-shore erosion (Christiansen et al., 2000; French & Spencer, 1993; Reed, Spencer, Murray, French, & Leonard, 1999). Also, human-driven change in sediment concentration have been reported in deltaic wetlands due to being trapped by upstream dams (Kesel & Sciences, 1988).

While understanding the mechanisms of marsh platform evolution is vital, the intent of this paper is to produce more accurate depictions of the DEMs in coastal wetlands, which are often inaccessible, using remote sensing techniques. The need for accurate mapping of these habitats is essential for conservatists to make progressive decisions, and for emergency management officials to increase preparedness. Particularly, DEMs used in marsh elevation models such as Marsh Equilibrium Model (MEM), Hydro-MEM, Sea Level Affecting Marsh Model (SLAMM) and Wetland Accretion Rate Model of Ecosystem Resilience (WARMER) (Alizad et al., 2016; M. Swanson et al., 2014; James T Morris et al., 2002; Park, Lee, Mausel, & Howe, 1991) need to

represent initial in-situ conditions (current state) in order to produce sound predictions about the evolution of coastal wetlands. As stated, salt marsh ecosystems provide protective and non-protective services. Next, a synthesis of literature on the economic value of salt marshes is included.

### 1.1.3 Economic Value of Salt Marshes

On an annual basis, coastal wetlands consisting primarily of salt marshes (about two thirds of total area) are estimated to save USD 23.2 billion in storm damage repair costs (Costanza et al., 2008). This figure was calculated using a model developed by Costanza and colleagues that explained 60% of the variation in relative damages using wetland area as a predictor along with wind speed (Costanza et al., 2008). This is achieved by the effectiveness of salt marshes in attenuating floodwaters and dissipating wave energy (E. B. Barbier et al., 2011). There is a lack of studies addressing the economic value of erosion control services provided by coastal wetlands (E. B. Barbier et al., 2011). However, it has been established that salt marshes are effective in dampening wave and current energy; in one study, the height of waves traveling 7 meters inland was reduced by 60% where salt marshes are present as opposed to a 33% reduction in marsh-free mud zones (Morgan, Burdick, Short, & Coasts, 2009).

Aside from flood-damage prevention and erosion control, salt marsh contributions to the seafood industry are indispensable; they provide breeding habitats and nurseries that enhance near-shore fisheries (E. B. J. E. p. Barbier, 2007). Also, one quarter of blue crab- and two thirds of shrimp production in the Gulf Mexico can be attributed to salt marshes (Zimmerman, Minello, & Rozas,

2002). In Florida, recreational fishing on the East and West coast marsh areas are valued at USD 7,452 per acre (Bell, 1997). Additionally, GBP 32.80 per person from the communities surrounding the Severn Estuary Wetlands in the UK was the evaluated economic benefit attained by the sustainable management of these systems (Birol & Cox, 2007). Moreover, water purification services provided by the marshes in southern Louisiana, USA have been estimated to save between USD 785 to USD 15,000 per acre in water treatment cost (Breaux, Farber, & Day, 1995). This is achieved by suspended particle deposition as well as nutrient and pollutant uptake (E. B. Barbier et al., 2011).

### 1.1.4 Importance of Salt Marsh DEMs

The presence of salt marshes in coastal regions increases the resiliency of shorelines against oceanic exposure (Costanza et al., 2008). The protective capability of these systems is measured using hydraulic models that simulate SLR scenarios as well as storm events (Hladik & Alber, 2012; S. Medeiros et al., 2015). Increased DEM accuracy in salt marshes will prevent these models from underestimating the effect of modeled scenarios. This is true since models using higher than actual ground elevations will underestimate inundation depth and frequency (S. Medeiros et al., 2015; Shastry & Durand, 2019). The topography of coastal zones, expressed by DEMs, is a primary parameter for detecting vulnerability against inundation; in low-lying areas such as salt marshes, elevation is the most important factor (Gesch, 2009).

### 1.1.5 Remote Sensing Data in Salt Marshes

There are several factors that affect the accuracy of remotely sensed data in salt marshes. These factors affect both estimates of elevation and biomass density. Errors in remotely sensed data can arise from temporal and spatial variations in acquisition. In addition, overall biomass density, cloud cover and water level can affect the quality of remotely sensed data in salt marshes.

Dense vegetation cover in salt marshes hinders the performance of many remote sensing techniques. The inability of the laser used in lidar to penetrate salt marsh vegetation results in a high bias in elevation estimates; this can be detrimental for models that rely on lidar-derived elevation data for the prediction of tidal, storm surge and SLR effects (Gesch, 2009; S. Medeiros et al., 2015; James T. Morris et al., 2005). Moreover, remote sensing data used to monitor vegetation in salt marshes suffer from what is known as the saturation problem. This issue arises when a certain species-specific threshold value of canopy closure and leaf area density is reached. Under these conditions, color-infrared satellite imagery fails to highlight variations in biomass density (Lu, 2006; Waring et al., 1995). As discussed in the following section, distinguishing biomass density across the marsh table is essential for adjusting lidar-derived DEMs.

Cloud cover can also have a detrimental effect on the quality of satellite imagery. This is especially true of the Gulf Coast in Florida due to the high storm frequency. Cloud distortion is inevitable, however, since this study is not concerned with creating a time series using remotely sensed data, images were selected during cloud-free periods. Nonetheless, various studies have been published in an aim to detect and improve the quality of cloud-distorted satellite imagery (Choi &

Bindschadler, 2004; Tahsin, Medeiros, Hooshyar, & Singh, 2017; Tseng, Tseng, & Chien, 2008; Zhu & Woodcock, 2012).

Another issue that can affect the quality of remote sensing in coastal areas is water level. When monitoring salt marsh vegetation, it is crucial to understand the spectral influence of water on hyperspectral imagery. The frequent inundation of salt marshes can influence the spectral reflectance of the marsh grasses (Kearney, Stutzer, Turpie, & Stevenson, 2009). Also, Kearney and colleagues demonstrated that under flood conditions, leaf area index (LAI) and normalized difference vegetation index (NDVI) can be misrepresented (Kearney et al., 2009). Since salt marshes are frequently inundated, it is likely that NDVI, on its own, is not suitable for monitoring biomass.

### 1.1.6 Field Sampling in Salt Marshes

Since marshlands span large areas, remote sensing is the most efficient way to monitor changes in these habitats. Collecting in-situ data from salt marsh habitats has proven to be an inefficient way to monitor changes in ecosystem characteristics (Jensen, Olson, Schill, Porter, & Morris, 2002). Aside from being non-invasive, remote sensing is more suitable than field measurements in large areas. While being labor-intensive, it is crucial to conduct field surveys that can capture in-situ conditions in order to validate the performance of remotely sensed elevation and biomass data.

True elevation can be measured using Global Positioning Systems (GPS) with Real Time Kinematics (RTK) capabilities. These systems can provide elevation measures that are well within the vertical resolution of remote sensing measures, and can be used to estimate the error in lidar-derived DEMs (Hladik & Alber, 2012; S. Medeiros et al., 2015; Takasu & Yasuda, 2008).

Biomass density samples are collected and processed so they can be utilized for improving the performance of remotely sensed vegetation monitoring techniques; accurate description of biomass is vital for models that incorporate biological aspects in estimating salt marsh evolution (James T Morris et al., 2002). Also, marsh plant productivity varies on a seasonal basis. Therefore, when coupling field and remote sensing data, the effect of temporal variation should be considered. If not addressed, major modelling uncertainties are likely be present (S. Medeiros et al., 2015).

This paper will focus on improving the accuracy of DEMs based on the correlation of lidar error with the density of marsh vegetation; The correlation discussed in previous works (Hladik & Alber, 2012; S. Medeiros et al., 2015) is outlined in the following section.

## 1.2 Biomass-Based Enhancement of Salt Marsh DEMs

### 1.2.1 DEM and Biomass Cover

To reduce the inaccuracy caused by biomass density, Hladik and Alber conducted a species-specific accuracy assessment and DEM correction study (Hladik & Alber, 2012). Their study utilized lidar-derived DEMs coupled with true elevations acquired during a field survey using Global Positioning Systems (GPS) with Real Time Kinematics (RTK) capabilities. The RTK

elevations were subtracted from the lidar-derived DEM to assess its accuracy and determine the high bias. Ten cover classes were considered (ranging from mud flats to tall Spartina Alterniflora), and their corresponding median DEM errors were calculated. Finally, lidar-derived DEM was adjusted within each class by subtracting the median error of the corresponding class. Although they found that the overestimation in elevation is greatest in areas with tall vegetation, the study concluded that canopy height cannot fully explain the error in lidar-derived elevation data. They suggested that other factors such as leaf orientation and biomass and stem density be investigated (Hladik & Alber, 2012).

In 2015, Medeiros et al. carried out a study aimed at increasing DEMs accuracy using estimated biomass densities (S. Medeiros et al., 2015). Conforming to the conclusions of the study outlined above, this study focused on biomass density as the main driver of lidar-derived DEM error (i.e. not species-specific). To estimated biomass densities, a series of ordinary least squares multi-linear regression models were assessed. In these models, field biomass densities (expressed in $g/m^2$) were used as dependent variables. The independent variables were acquired from lidar surveys, the red and near-infrared bands (bands 2 and 3) of the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER), and interferometric synthetic aperture radar (IfSAR). Moreover, ASTER data were used to calculate vegetation indices such as the normalized difference vegetation index (NDVI) and the simple-ratio vegetation index (VI). IfSAR and lidar data were utilized for estimating canopy height for use in the regression. Biomass densities were categorized from high to low, and two biomass density classification systems were introduced while retaining the elevation errors within each class. The best biomass density model achieved an adjusted r-squared

of 0.82 using canopy height, ASTER and IfSAR data. The biomass density was then estimated for the entire area and used as the lone predictor of DEM error. A control scenario was compared to all DEM-adjustment scenarios and a two-class quartile adjustment scenario was adopted. The adjusted DEM showed an improvement of 38% as measured by the RMS error in elevation (S. Medeiros et al., 2015).

These studies outline the necessity of developing rigorous and spatially variable techniques that can yield more accurate DEMs in salt marshes and that unadjusted elevation models are not reliable for coastal management and research applications.

## 1.3 The Potential of Machine Learning for Estimating Biomass Density and Platform Elevation

### 1.3.1 Background

Multiple regression is the most common approach mitigating errors in remotely sensed data (Lu, 2006). In this approach, multiple sources of remotely sensed data are used in different combinations to reduce error in estimates of biomass density and ground elevation. Medeiros et al. developed a series of ordinary least square (OLS) regression models by combining remotely sensed and in-situ data (S. Medeiros et al., 2015). While these models have resulted in notable improvements to DEMs, they come with the inherent assumptions of linearity and normal distribution. Therefore, non-parametric machine learning algorithms can potentially explain more of the variability between in-situ measures and remote sensing data.

Machine learning is a branch of artificial intelligence (AI) (Hsieh, 2009). In the book *Machine Learning*, machine learning is described as follows: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E" (Mitchell, 1997). Another author defines machine learning as "programming computers to optimize a performance criterion using example data or past experience" (Alpaydin, 2014). There are different types of machine learning algorithms such as supervised and unsupervised learning. In supervised learning algorithms, input-output examples are used to train a model (Ayodele, 2010). This approach is adopted in this study due to its applicability.

Since the 1990s, environmental scientists have utilized machine learning techniques to help them understand the complex nature of environmental processes (Hsieh, 2009). In water resource engineering, machine learning has been utilized in both hydraulics (Bhattacharya, Price, & Solomatine, 2007; Rasouli, Hsieh, & Cannon, 2012; Roushangar, Akhgar, Salmasi, & Shiri, 2014) and hydrology (Ahmad, Kalra, & Stephen, 2010; Hong & Computation, 2008; Xingjian et al., 2015). In the field of remote sensing, plenty of research studies have been conducted for vegetation mapping but were mainly focused on delineating vegetation covers (Baker, Lawrence, Montagne, & Patten, 2006; Ghedira, Bernier, & Ouarda, 2000; Pal & Mather, 2003; Szantoi et al., 2015). While these studies address the use of supervised learning algorithms for classification purposes, this paper uses similar techniques to estimate above ground biomass density in order to adjust lidar-derived DEMs.

There are many types of machine learning algorithms that have been utilized in the estimation of biomass (Ali, Greifeneder, Stamenkovic, Neumann, & Notarnicola, 2015; Breidenbach, Næsset, Lien, Gobakken, & Solberg, 2010; Cutler, Boyd, Foody, & Vetrivel, 2012; Gleason & Im, 2012; Jachowski et al., 2013; Mutanga, Adam, & Cho, 2012). Furthermore, machine learning algorithms have attracted more researchers in the field of remote sensing due to their ability to process large datasets without assumptions of linearity (Ali et al., 2015). Gleason and Im used linear mixed-effects (LME) regression, random forest (RF) regression and support vector regression to estimate forest biomass (Gleason & Im, 2012). For the same purpose, Cutler and colleagues used artificial neural networks to estimate forest biomass in sites in Brazil, Malaysia and Thailand (Cutler et al., 2012). Similarly, Jachowski et al. tested several machine learning algorithms including support vector machines (SVM) for the estimation of mangrove biomass in Thailand.

While the aforementioned studies investigate biomass estimation in forests, the scope of this paper was to estimate biomass densities in salt marshes using multi-linear regression, artificial neural networks (ANN) and random forest regression. The use of multi-linear regression analyses was discussed in section 1.2. Here, the performance of ANN and RF regression for biomass estimation is discussed.

### 1.3.2 Case Studies

#### *1.3.2.1 Artificial Neural Networks (ANN)*

The performance of ANNs in estimating forest biomass was investigated by Cutler and colleagues in forests spanning latitudes between 2° 28′ S and 19° 31′ N to determine whether they can produce

generalizable models (Cutler et al., 2012). They used SAR backscatter alongside reflectance of the 6 non-thermal bands of LTM. Moreover, 144 field plots from all three locations were used as field samples in the study; due to missing values in the remote sensing data, only 94 plots were used in the analysis. Unlike linear regression, ANNs make no assumptions about the distribution of the data and are, therefore, generalizable even when used with noisy data (Bishop, 1995).

Many machine-learning algorithms, including ANNs, are specified by hyperparameters. Not to be confused with model parameters, hyperparameters are set before the beginning of the training process and are independent of the learning experience. In the case of ANNs, hyperparameters consist of, among other things, the number of hidden layers, the number of nodes (neurons) within each layer, the learning rate and momentum and the overall architecture of the network. In this study, the investigators did not tweak the learning rate or momentum of the optimizer as this was beyond the scope of the study. Instead, they used a software package that allowed the testing of many networks with differing numbers of hidden layers and neurons.

Once data was preprocessed and prepared for use as input, the networks were trained under four different scenarios. Under the first scenario, the networks were trained using one site at a time. The samples from each site were divided into training and testing sets. The second approach was to train the networks also by site, but to validate the performance of the best model using the samples from the remaining two sites. The third approach was to combine the samples from all the sites (i.e. not site-specific) and to split them into training and testing subsamples. Those three scenarios were first tested with one feature only (SAR backscatter), and later combined with all

the other features. The authors stated that the expected negative effects of collinearity were reduced by feeding the network different combinations of input variables. Again, the limited scope of the analysis did not include the selection of optimum feature-combination. Finally, the fourth approach was to repeat the third scenario with the inclusion of LTM data.

Under the first scenario, weak correlation was found when using SAR backscatter as the lone predictor of biomass. SAR backscatter suffers from the saturation issue discussed in section 1.1, which explains, according to the authors, its poor performance (Imhoff, 1995; Lucas et al., 2006). A network trained under the fourth scenario, which included both SAR and LTM data, yielded the strongest correlation to biomass (Cutler et al., 2012).

#### *1.3.2.2 Random Forest (RF)*

Simply put, a random forest algorithm is a collection of connected decision trees (Breiman, 2001). Random forests train by randomly sampling a number of subsamples (with replacement) equal to the number of trees in the forest (an important hyperparameter that should be optimized). Then, trees grow by splitting (branching out) whenever a reduction in out-of-bag (OOB) error that exceeds a certain threshold is attainable. The error associated with splitting is calculated at each step, and the tree continues to grow until error improvement is minimized or maximum tree depth (another hyperparameter) is reached (Gleason & Im, 2012).

Mutanga et al. utilized this algorithm for the estimation of biomass densities in a subtropical estuarine wetland (Mutanga et al., 2012). To estimate biomass, 82 samples were collected and weighed on-site. Also, WorldView-2 imagery data were acquired, processed, then transformed into canopy reflectance. Canopy reflectance values coinciding with the locations of the 82 field samples were extracted for use as predictors in the model. Later, the data were split into training and testing subsamples using a 30% test size. All possible pairs of the 8 spectral bands were used to calculate a total of 56 NDVI values at each sampling location. The model was then trained to estimate biomass density using 57 samples and validated using 25 samples. The authors use the "Random Forest" package within R environment software for implementing the algorithm. The number of trees (n_estimators) as well as the number of predictors used for splitting a node (min_sample_split) were optimized by holding other parameters constant and trying out forests with number of trees ranging from 500 to 9,500 at 1,000 intervals and number of minimum features to use for splitting a node with integers ranging from 1 to 25.

Random forests can calculate an importance factor for each of the predictors by calculating the mean decrease in accuracy as determined by the out-of-bag samples. This is somewhat analogous to Pearson's correlation coefficient in that large importance factors are attributed to the most critical predictors for the regression. However, the difference here is that the importance score is only relevant to the model in question and does not directly describe physical correlations (Archer & Kimes, 2008).

## 1.4 Summary

This synthesis highlights the value of salt marshes and the importance of accurately depicting their topography. Also, the relationship between digital elevation model (DEM) error and biomass density was underscored. In particular, two relevant works were summarized (Hladik & Alber, 2012; S. Medeiros et al., 2015). Moreover, two promising machine learning algorithms (Artificial Neural Networks and Random Forest) were introduced with examples from literature (Cutler et al., 2012; Gleason & Im, 2012). The following chapters report the development of biomass density estimation models as well as the adjustment of lidar-derived DEMs based on the estimated vegetation densities.

# CHAPTER II: METHODS AND JUSTIFICATIONS

This chapter describes the methods used in developing the biomass density estimation models, collecting field and remote sensing data and the adjustment of a lidar-derived Digital Elevation Model (DEM) over the study area. The workflow is summarized in the first section of this chapter. In the next section, the process of collecting, combining and preprocessing field and remote sensing data is described. Then, a classification technique of field biomass densities is reported (S. Medeiros et al., 2015). Furthermore, the development of the estimation models is detailed. Finally, a description of the technique adopted for adjusting lidar-derived DEM is discussed.

## 2.1 Workflow

The objective of this study was to increase the accuracy of DEMs in salt marshes using biomass densities. The study was carried out in the following fashion:

1. Field samples consisting of biomass densities and elevations were collected.
2. Remote sensing data were appended to the dataset containing the biomass density and elevation measures.
3. Using the in-situ biomass densities, three classes (high, medium and low) were defined for the purpose of classifying the model-estimated biomass densities.
4. A series of multilinear regression and random forest (RF) models were developed to estimate biomass densities.

5. The best performing models from the previous step were selected for the estimation of biomass densities across the entire study area.
6. Lidar-derived DEM over the study area was adjusted based on the estimated biomass densities.

## 2.2 Data Collection and Processing

The dataset used in the analysis consists of field and remote sensing data. In this section, the study area is described. Moreover, the collection of ground elevation data, both in the field and remotely, is discussed. Also, the field biomass sampling technique is detailed followed by a description of the remotely sensed data used for representing the vegetation cover. Finally, the preprocessing of the dataset is detailed.

### 2.2.1 Study Area

This study was conducted at the salt marshes of the St. Marks National Wildlife Refuge located in Wakulla County in the Florida Panhandle (See figure 1). The salt marshes in this area, consisting primarily of juncus, are ideal for the study since they are located in an undisturbed coastal zone (Subrahmanyam & Drake, 1975).

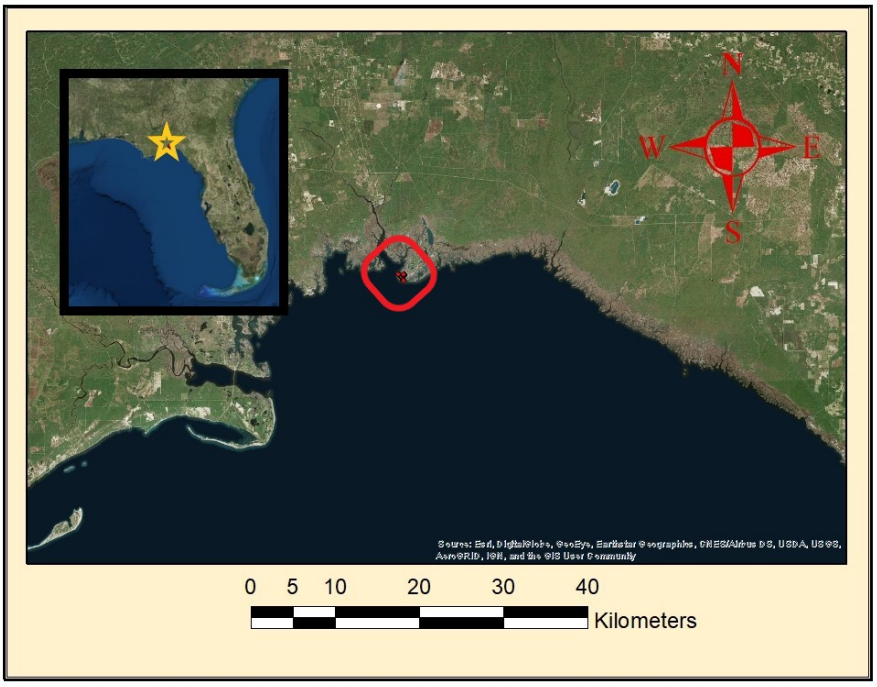


*Figure 1. Location of the Saint Marks National Wildlife Refuge in the Florida Panhandle*

### 2.2.2 Data Collection

#### *2.2.2.1 Elevation*

The DEM that was adjusted in this study is provided to the public by the Florida Division of Emergency Management (FDEM). The development of this DEM is described in the final report of the Specific Purpose LiDAR Survey (Dewberry, 2009). Raw lidar return data are classified into unclassified, ground, noise, water and overlap as per the requirements of FDEM; only unclassified and ground returns were retained for the analysis. Nominal pulse spacing and point cloud density values are reported at 1.25 m and 4.26 points per $m^2$, respectively. Lidar return

values were used as predictors in the developed biomass density models while the final DEM product as provided by FDEM was used as the basis for adjustment.

In the field, 377 spot elevation data points were collected along transects spanning the entire elevation profile of the salt marsh (see figure 2). Two field surveys were conducted in the summer of 2017 and spring of 2018. The spacing between survey points was set in accordance to the resolution of the available lidar data. Elevations were acquired using RTK-GPS in order to calculate the error in lidar-derived DEM before and after the adjustment. The lidar returns used to construct the DEM are reported at 18.6 cm vertical accuracy at the 95%-confidence (Dewberry, 2009). The error from return values adds uncertainty to the constructed DEM which has an inherent error due to the presence of marsh vegetation. The accuracy of the DEM is tested and improved herein.

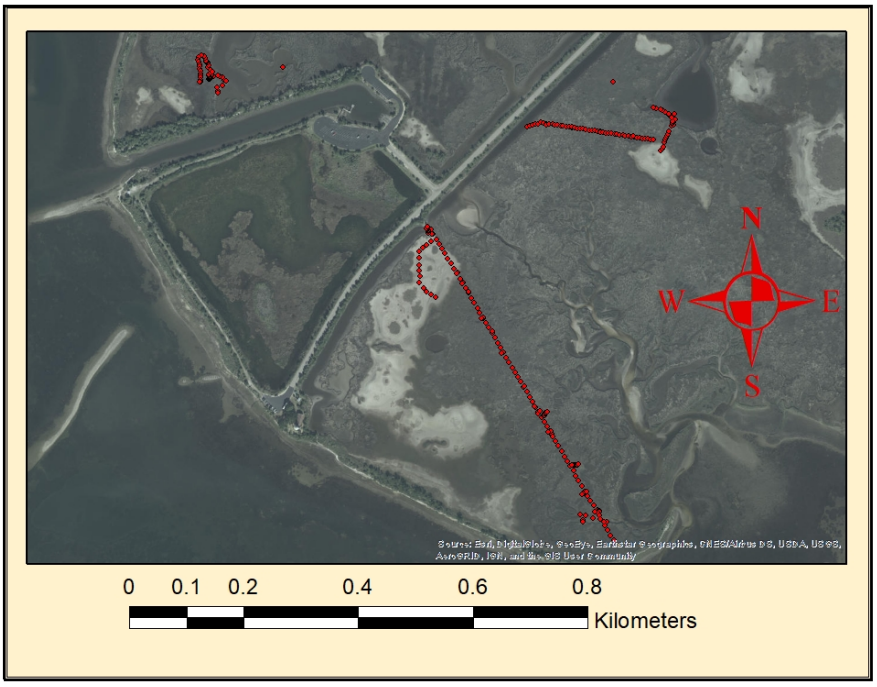


*Figure 2. RTK survey conducted in the study area during two field visits (Summer 2017 and Spring 2018)*

### *2.2.2.2 Biomass Density*

Biomass densities were collected in the field at 107 locations associated with the elevation survey points. These samples were used in the development of the model as the target variables. These samples were collected throughout the elevation range of the marsh in order to determine the distribution of the biomass growth. To reflect natural conditions, zero biomass samples (37/107) were collected at high and low elevations. The lack of vegetation within these elevations can be attributed to either relatively high or low frequency and duration of inundation. Increased inundation, or lack thereof, is considered a stressor for marsh plants. Due to the high salinity of ocean water, increased frequency and duration of inundation events cause the salinity

of the marsh habitat to soar; increased salinity is equivalent to water deficit (Jiang, Luo, Chen, Li, & Science, 2009). The procedure for collecting biomass samples is describe next.

At 107 out of 377 locations of the RTK-GPS survey, biomass samples were acquired. Locations where no biomass was present (total of 37) were simply labeled as zero biomass (ZBM). For locations with observed biomass, a 0.25 m × 0.25 m square made of PVC was randomly tossed in the vicinity of the corresponding elevation measure. Vegetation whose stems are enclosed by the PVC square was harvested and collected in plastic bags that were labeled to correspond to the RTK survey. Three samples per location were harvested to capture variance. These samples were then transported to the lab for further processing. To preserve the samples, they were transported in large containers with dry ice. In the lab, sediment was washed off the samples and they were left to air-dry. The samples were then placed in an oven at 105° C and dried to a constant (dry) weight. The dry weight of each sample was divided by the area of the square used for harvesting the samples in order to determine the biomass density in grams per square meter. For each sampling location, the weighted average of the samples was reported and appended to the corresponding XYZ location.

In order to estimate biomass densities across the study area, lidar returns accompanied by vegetation monitoring satellite data were used. The satellite scenes chosen for the analysis were from the same months of the field trips (July 2017 and March 2018) and carefully chosen on cloud-free days. The lidar intensity data described in the previous section were also used for biomass density prediction as well as the estimation of canopy height and canopy cover ratio

(two additional features to be used for biomass density prediction). The canopy height was calculated simply by subtracting the digital surface model (DSM), as extracted from lidar first return, from the bare earth elevation DEM. The canopy cover ratio takes on values between 0 and 1, where 0 implies zero biomass while 1 indicates very dense vegetation cover.

The reflectance values of the blue, green, red and near-infrared bands were acquired from two satellites: Landsat 8 and Sentinel-2 (ESA, 2017; USGS, 2017). Landsat 8 is part of the Landsat Program, a joint effort between the National Aeronautics and Space Administration (NASA) and the United States Geological Survey (USGS) (Lauer, Morain, Salomonson, & Sensing, 1997). Sentinel-2 was launched in a joint effort by the European Commission (EC) and the European Space Agency (ESA) (Drusch et al., 2012). The reflectance data are reported in $W/m^2/\mu m$ with a resolution of 10 m and 30 m for Sentinel 2 and Landsat, respectively (figures 3 and 4).

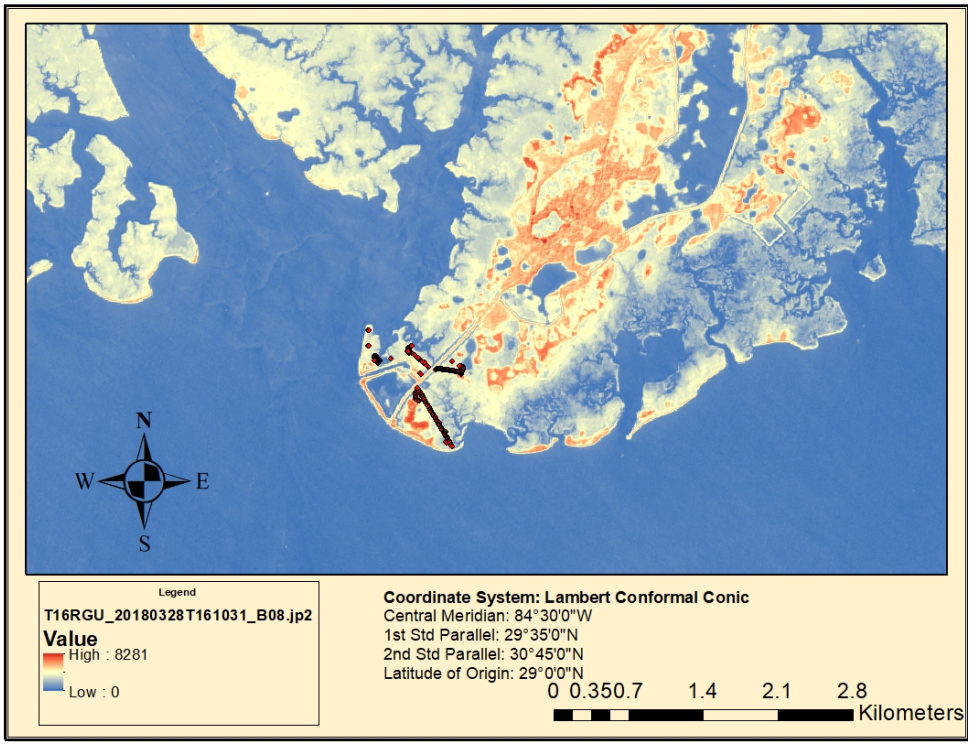


*Figure 3. Sentinel 2 scene Band 8 (NIR) from March-28-2019*

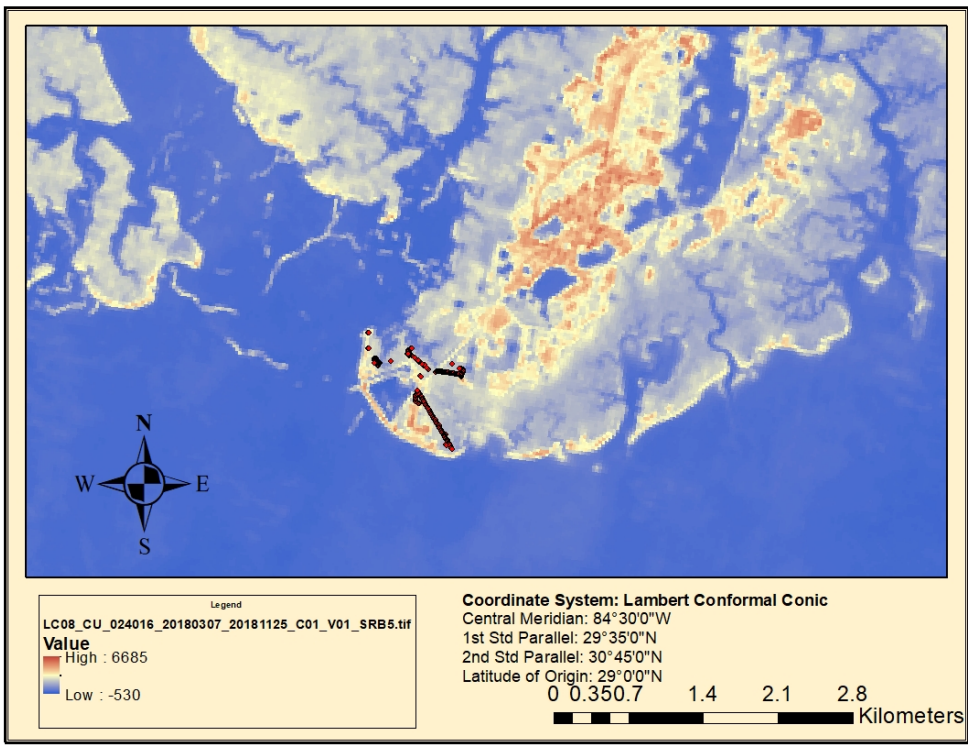


*Figure 4. Landsat 8 Band 5 (NIR) scene from March-07-2018*

### 2.2.3 Data Preprocessing

#### *2.2.3.1 Spatial Combination and Indices Calculations*

The data described above were imported to ArcMap (ESRI, 2011). All the data were referenced using State Plane Florida North NAD83 2011. First, the RTK survey data were displayed on a base map. Then, satellite images were imported and visualized on the same map, and the values of the relevant bands closest to RTK measures were extracted. The same process was repeated using the lidar-derived DEM which was used as the basis for adjustment. Moreover, lidar return values were extracted for every point of the RTK survey in the same manner. The values

extracted from Landsat 8 and Sentinel 2 were then used to create columns with vegetation indices. The indices used in the analysis are the Normalized Difference Vegetation Index (NDVI), the Vegetation Index (VI), the Global Environment Monitoring Index (GEMI) and the Modified Soil-Adjusted Vegetation Index (MSAVI) (see equations 1 through 5 (Eastwood, Yates, Thomson, & Fuller, 1997; Pinty & Verstraete, 1992)).

$$NDVI = \frac{NIR-Red}{NIR+Red} \quad (1)$$

$$VI = \frac{NIR}{Red} \quad (2)$$

$$GEMI = \gamma(1-0.25\gamma) - \frac{Red-0.125}{1-Red} \quad (3)$$

$$where\ \gamma = \frac{2(NIR^2-Red^2)+1.5\ NIR+0.5Red}{NIR+Red+0.5} \quad (4)$$

$$MSAVI = \frac{(2NIR+1)-\sqrt{(2NIR+1)^2-8(NIR-Red)}}{2} \quad (5)$$

Moreover, the elevations from the RTK survey were subtracted from the lidar-derived elevations to obtain a column with DEM errors. This was later used to calculate the quartile errors within each biomass density class and to assess the accuracy of the original DEM before and after the adjustment.

### *2.2.3.2 Preparing Data for Analysis*

After the relevant data and indices were spatially combined, the rows containing missing data were dropped out of the dataset. Further model-specific preprocessing is discussed in the Model Development and Selection section. The data used in the analysis are summarized in table 1. Next, the classification of biomass densities is discussed.

*Table 1. Summary of data used in the analysis*

| Data Set | Resolution | Acquisition Date | Use in Analysis |
|---|---|---|---|
| Lidar DEM | 1.52 m | Summer 2007 | Basis for adjustment |
| Lidar Returns | 4 pts/$m^2$ | Summer 2007 | Biomass Regression |
| Sentinel 2 | 10 m | Summer 2017, Spring 2018 | Biomass Regression |
| Landsat 8 | 30 m | Summer 2017, Spring 2018 | Biomass Regression |
| GPS-RTK Elevation | n/a | Summer 2017, Spring 2018 | Assessing DEM Accuracy |
| Biomass Density | n/a | Summer 2017, Spring 2018 | Biomass Regression |

## 2.3 Classification of Biomass

Biomass density classes were defined for the purpose of classifying the predicted biomass densities before adjusting the lidar-derived DEM. Zero biomass samples were omitted for this part. The biomass densities were then ranked in a descending order ranging from 3330.24 g/$m^2$

to 17.92 g/m$^2$. The 66$^{th}$ and 33$^{rd}$ percentiles were then calculated to delineate the classes into high, medium and low. Lidar-derived DEM error was calculated for each class using a quartile approach; the 75$^{th}$ percentile value of the DEM error within the high-density class, the 50$^{th}$ percentile (or median) within the medium class, and the 25$^{th}$ percentile within the low-density class were retained. The methods used for developing the models and the selection of the best model are discussed in the following section.

## 2.4 Model Development and Selection

A series of multilinear regression and random forest (RF) models for estimating above ground biomass density were tested as part of the analysis. In this section, the development of the models is discussed. Also, the selection criterium of the best performing model is described.

### 2.4.1 Multilinear Regression

Several models were developed using different combinations of predictor (independent) variables. The regression models were created using the Scikit Learn library for Python (Pedregosa et al., 2011). The different combinations of predictors were selected based on an analysis of correlation. Combinations of features with significant correlation to the target variable (biomass density) were selected. Due to the relatively limited dataset (64 biomass trainable samples), a bootstrapping technique was implemented. The algorithms were trained on all the samples except for one (n-1) which was preserved for validation. The process is then repeated n-times and the validation score is averaged and reported. Before fitting the model, variance inflation factors (VIF) were calculated for each set of features. This was done to prevent

redundancy in features within a model. Features with high VIF values within a combination were not used for fitting a model. The results from these analyses are summarized in the following chapter.

### 2.4.2 Random Forest Regression (RF)

The random forest regressor from the Scikit Learn library was used for fitting the RF models (Pedregosa et al., 2011). As discussed in chapter one, each machine learning algorithm is governed by a set of hyperparameters. For the RF models tested in this paper, a hyperparameter optimization technique (RandomizedSearchCV) was adopted. This optimizer fits the model with a series of hyperparameter combinations that are defined by the user. For each fit, the algorithm uses a combination of defined hyperparameters chosen randomly while retaining the scores of the fit. Then, the hyperparameters associated with the best performing fit are reported. The hyperparameters for a random forest include the number of trees, the maximum tree depth, the number of features to consider for a split, the minimum number of samples to split a node and the minimum number of samples to comprise a node. Other hyperparameters were left at their default values as the optimization of those was beyond the scope of this paper. All the features were used to train the RF models because variance inflation does not occur as in the case of multilinear regression. One of the perks of using a random forest (RF) regressor is that it can rank the input variables according to their importance in predicting the target variable. This becomes more valuable for analyses involving larger datasets; one can exclude those features ranking low in importance to save computation cost and time. The results of the RF models are reported in the next chapter.

### 2.4.3 Model Selection

The target of these models is to estimate biomass densities in order to use those values for adjusting the lidar-derived DEM of the study area. The performance of these models was compared using the Root Mean Square Error (RMSE). The adjusted $R^2$ values for the regression analysis are also reported. The best performing models from the multilinear regression models as well as from the RF models were used to classify biomass densities as discussed earlier. After adjusting the DEM, the model resulting in the highest improvement is recommended.

## 2.5 Adjusting the DEM

The biomass densities at the RTK survey points are estimated using the best performing models. Then, the points are classified according to the threshold biomass density values discussed in section 2.3 into high, medium and low. Finally, the adjustment values of the corresponding biomass density class are subtracted from the lidar-derived DEM to create an adjusted DEM. The performance of adjusted DEMs is then compared with the original DEM using the RMSE as well as the raw mean errors. In the next chapter, the results of the analyses described above are summarized and discussed.

# CHAPTER III: RESULTS AND DISCUSSION

This chapter begins with a summary of the results from the DEM accuracy assessment. Next, the results from the biomass density model development are reported. The adjusted DEM is then reported, and its accuracy is compared with the original DEM's RMSE and raw mean error. Finally, the results are interpreted and discussed in the last section.

## 3.1 Results

### 3.1.1 DEM Accuracy Assessment and Control Scenarios

The accuracy of the lidar-derived DEM was assessed as follows:

1. The lidar elevation values coinciding with the RTK survey points were extracted.
2. The RTK elevations were subtracted from the lidar elevations to calculate the high bias.
3. RMSE and raw mean error were calculated.

DEM assessment results are summarized below in table 2. An adjustment model should only be considered if it outperforms adjusting the DEM by subtracting the random error values or, simply, median DEM error value. Consequently, random values within the range of DEM error were generated for every point in the RTK survey and were subtracted from the lidar DEM elevations. The median DEM error value was also subtracted from the lidar DEM elevations. The RMSE and raw mean errors for the control scenarios are included in table 2 for comparison.

*Table 2. Results from DEM accuracy assessment*

| Measure | Original DEM | Randomly Adjusted DEM | Median-Adjusted DEM |
|---|---|---|---|
| RMSE (m) | 0.38 | 0.52 | 0.34 |
| Improvement | n/a | -34% | 12% |
| Raw Mean Error (m) | 0.33 | 0.41 | 0.34 |
| Improvement | n/a | -24% | -3% |

### 3.1.2 Biomass Density Estimation Model

#### *3.1.2.1 Correlation Analysis*

Before developing the regression models, the correlation between the features and the target variable (field biomass density) was investigated. The Pearson correlation coefficient was calculated for each input variable and is displayed in figure 5 below.

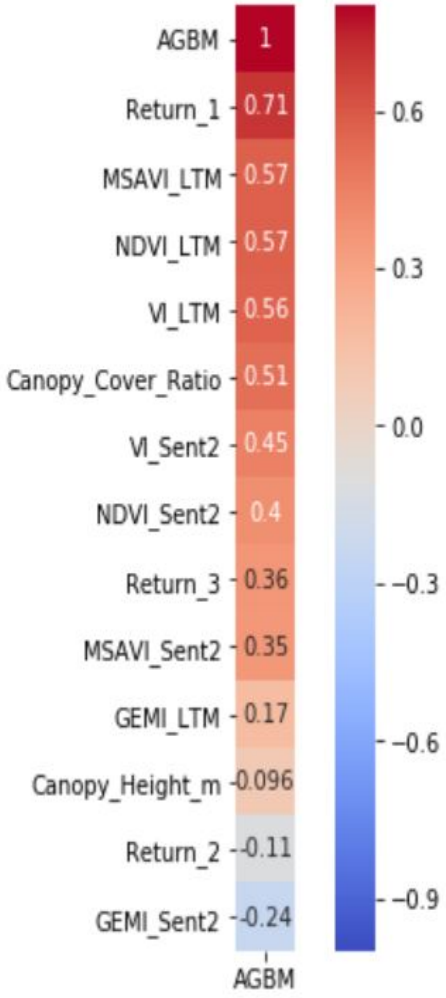


*Figure 5. A heat map displaying Pearson correlation coefficient values of the predictor variables against above ground biomass (ABGM)*

#### *3.1.2.2 Multilinear Regression*

A series of multilinear regression models were developed using different combinations of features. Each selected combination went through a variance inflation factor (VIF) check to determine the extent multicollinearity within predictors. All combinations used in the multilinear regression models had VIF scores lower than 4. Also, AIC scores were reported for model comparison. The table below summarizes the results of the multilinear regression models (table 3).

*Table 3. Summary of the Multilinear Regression Models*

| Model Name | Features | VIF Scores | Coefficients | Intercept | Adj. $R^2$ | AIC |
|---|---|---|---|---|---|---|
| Regression 1 | Return_1 | 1.15 | 12.9 | -12.92 | 0.57 | 276.3 |
| | VI_Sent2 | 1.15 | 5.02 | | | |
| Regression 2 | Return_1 | 1.56 | 9.93 | -15.22 | 0.62 | 269.4 |
| | VI_Sent2 | 1.16 | 5.67 | | | |
| | Canopy_Cover_Ratio | 1.37 | 5.82 | | | |
| Regression 3 | Return_1 | 1.67 | 7.06 | 40.90 | 0.79 | 233.2 |
| | Canopy_Cover_Ratio | 1.70 | 0.85 | | | |
| | GEMI_Sent2 | 3.75 | -26.96 | | | |
| | MSAVI_Sent2 | 3.49 | 32.39 | | | |
| Regression 4 | Return_1 | 1.75 | 7.65 | 38.44 | 0.80 | 231.8 |
| | Canopy_Height_m | 1.17 | -0.48 | | | |
| | Canopy_Cover_Ratio | 1.80 | 1.55 | | | |
| | MSAVI_Sent2 | 3.55 | 31.49 | | | |
| | GEMI_Sent2 | 3.85 | -25.85 | | | |

#### *3.1.2.3 Random Forest (RF) Regression*

A series of RF models were also tested. First, hyperparameter optimization was carried out using the Randomized Search Cross Validation (RandomizedSearchCV) from the Scikit Learn library (Pedregosa et al., 2011). Further tuning by trial and error was conducted for reasons included in the discussion below. The result of the best performing RF model (figure 6) was reported as RMSE = 183.93 g/m$^2$ biomass density. The pseudo $R^2$ was calculated as follows:

$$R^2 = 1 - \frac{MSE}{Var(y)} \quad (6)$$

Where MSE is the mean squared error and Var(y) is the variance within the biomass samples used in training the model. This model achieved and pseudo $R^2$ of 0.94.

### 3.1.3 DEM Adjustment

The biomass density classes that were defined along with the corresponding quartile adjustment values are reported in table 4.

*Table 4. Biomass Classes and their Corresponding Adjustment Values*

| BM Class | Quartile Adjustment (m) | BM Threshold (g/m$^2$) |
|---|---|---|
| High | 0.44, 75$^{th}$ Percentile | 1111.68 |
| Medium | 0.38, Median | 720.64 |
| Low | 0.19, 25$^{th}$ Percentile | 17.92 |

Predicted biomass density values below the low-class threshold were treated as zero biomass and, thus, no adjustments were made to the lidar DEM there. After predicting biomass densities across the study area, the original DEM was adjusted, and the performance of the adjustment scenarios is show here (table 5).

*Table 5. Lidar-derived DEM Adjustment Summary*

| DEM | RMSE (m) | Improvement (%) | Raw Mean Error (m) | Improvement (%) | Standard Deviation (m) | p-Value |
|---|---|---|---|---|---|---|
| Original | 0.38 | N/A | 0.33 | N/A | 0.2 | N/A |
| Median-Adjusted | 0.34 | 12% | 0.34 | -3% | 0 | 1 |
| Randomly Adjusted | 0.52 | -34% | 0.41 | -24% | 0.32 | <0.0001 |
| Regression Adjusted | 0.19 | 51% | 0.15 | 56% | 0.12 | <0.0001 |
| Random Forest Adjusted | 0.18 | 54% | 0.14 | 58% | 0.11 | <0.0001 |

In the next section, the results summarized herein are discussed.

## 3.2 Discussion

### 3.2.1 Digital Elevation Model Accuracy

The accuracy of the lidar-derived DEM was assessed by subtracting the RTK (true) elevations from the original DEM. Table 2 reports the accuracy in terms of RMSE and raw mean error. The poor performance of the lidar survey is evident when comparing the RMSE and raw error (0.38 m and 0.33 m, respectively) to the range and standard deviation over the study area (0.73 m and 0.14 m). The magnitude of error is significant considering the relatively low relief within a salt marsh platform.

Median DEM error was also calculated in order to create the first control scenario. The value was subsequently subtracted from the original DEM to create an "improved" version. The deduction of median error (0.34 m) from the lidar DEM improved the performance by 12% in terms of RMSE, however, the mean raw error witnessed a slight decrease of 3% as indicated in table 2.

Finally, another control scenario was added for comparison using error values randomly generated from within the range of lidar DEM error. Nonetheless, this resulted in reduced DEM accuracy with reported RMSE and mean errors of 0.52 m and 0.41 m, respectively (table 2).

### 3.2.2 Biomass Density Models

Two biomass density estimation models are outlined in this paper: a multilinear regression model and a random forest (RF) model.

#### *3.2.2.1 Multilinear Regression*

The Pearson correlation heatmap (figure 4) was examined to find potential candidate features for the linear regression model. The highest correlation with above ground BM density was associated with lidar intensity return 1. This is the first laser pulse reflected to the lidar sensor. When dense vegetation is present, the probability of ricocheting off canopy is increased; hence, the high correlation with the first laser pulse. This supports the findings of previous works (Hladik & Alber, 2012; S. Medeiros et al., 2015). Moreover, the lack of correlation between biomass density and canopy height was particularly interesting. It is expected that canopy height correlates to the magnitude of DEM error as laser pulses that fail to penetrate to the marsh platform are likely to bounce off canopies. Nonetheless, the density is the actual source of error. In fact, lidar is known to perform well in forests with spars, yet tall, vegetation (Akay, Oğuz, Karas, Aruga, & assessment, 2009). The canopy height estimates are found by subtracting a digital surface model (DSM) from a DEM (or subtracting the corresponding lidar returns if point cloud data are available); to use canopy height is analogous to using a faulty DEM to improve itself. As a result, another canopy related measure was tested, namely canopy cover ratio. This ratio is similar to the leaf area index (LAI), but its performance is not affected by the shape of the leaves; LAI underestimates canopy cover when branches and stems are prevalent (Bréda, 2003). Furthermore, the dimensionless canopy cover ratio used herein is derived from lidar point cloud

data using ground and non-ground returns. Basically, it is a measure of how much ground is visible from lidar perspective. To calculate canopy cover ratio, simply, divided non-ground lidar return count by the total count.

The other indices (VI, NDVI, MSAVI and GEMI) have been used by researchers interested in quantifying biomass (Kearney et al., 2009; S. Medeiros et al., 2015). Eastwood and colleagues indicated that MSAVI and GEMI are most suitable for monitoring salt marsh vegetation (Eastwood et al., 1997). These indices were calculated using reflectance data acquired from Sentinel-2 and Landsat 8. The variation in the correlation with biomass density between indices from different satellites can be attributed to differences in spatial resolution.

The suitability of different combinations of remote sensing data was tested using the VIF score. The regression models summary (table 3) lists those factors. If VIF scores are not verified, models can be misinterpreted. For example, the addition of extra predictors may enhance the performance of a model at the expense of generalization ability.

Because field sampling in marsh areas is labor-intensive, in-situ samples are usually limited. The relatively small feature-to-sample ratio prompted the use of a bootstrapping technique known as leave-on-out. In this approach, the model is fitted using all the samples except for one which is retained for validation. This iterative process is a common approach when dealing with smaller datasets. Splitting the data into training and testing samples (e.g. 20-to-80% ratio) was

investigated. Since only 64 trainable field density samples were available, splitting the data without a bootstrapping approach yielded inferior results. Using 20% to 30% of the sample for testing the model is appealing. With a relatively small dataset, however, some important features may become absent from the training process. This approach was adopted in a previous study with a similar site setting (S. C. Medeiros, Hagen, & Weishampel, 2015).

The best performing model is summarized in table 6 below. This model was adopted to classify biomass densities across the RTK survey in order to determine DEM adjustment values.

*Table 6. Top-performing regression model*

| **Model Name** | **Features** | **VIF Scores** | **Coefficients** | **Intercept** | **Adjusted $R^2$** |
|---|---|---|---|---|---|
| Regression 4 | Return_1 | 1.75 | 7.65 | 38.4391736 | 0.80 |
| | Canopy_Height_m | 1.17 | -0.48 | | |
| | Canopy_Cover_Ratio | 1.80 | 1.55 | | |
| | MSAVI_Sent2 | 3.55 | 31.49 | | |
| | GEMI_Sent2 | 3.85 | -25.85 | | |

#### *3.2.2.2 Random Forest Regression Model*

One of the advantages of using machine learning algorithms, particularly RF, is that collinearity is not a concern. The generalization ability of a prediction model is the objective. When fitting a model, the goal is to learn the patterns behind the variation of a target variable. Overfitting occurs when a model learns the noise within a dataset. Since this noise is particular to that dataset (e.g. due to errors in the sampling technique), the generalization ability is diminished. On the other hand, under fitting occurs when not enough training is conducted due to either a small sample that does not describe the variation of the target variable, or due to not optimizing the hyperparameters of the algorithm. Random forest algorithms are well-known for their ability to prevent overfitting (Gleason & Im, 2012; Mutanga et al., 2012). Once hyperparameters are carefully selected, the algorithm creates training sets randomly picked from the range of predictors (with replacement). While these sets may not contain all the input data, they are equal in size to the original set. This takes place at the tree level; if some important features are not used in training the tree, other trees in the forest will capture them since all features have an equal chance to be selected. The performance in predicting the target variable for each tree is retained in order to calculate the overall performance of the random forest. Using many trees reduces the chance of model overfitting.

While the performance of the random forest is far superior to that of the linear regression model (94% vs 80% $R^2$), they performed similarly in terms of enhancing the performance of the DEM within the study area (54% vs 51% improvement in RMSE for RF and linear model, respectively). Nonetheless, in terms of estimating zero biomass (ZBM) points within the survey

the random forest model dominates again. The mean biomass densities for model-predicted biomass at points defined as ZBM within the survey are approximately 69 $g/m^2$ and 1,215 $g/m^2$ for RF and linear regression models, respectively. As discussed in the introduction, identifying zero biomass locations is of utmost importance for the parameterization of marsh evolution models.

### 3.2.3 Adjusting the DEM

The performance of the adjusted DEM was assessed by comparing the new DEMs to the original one in terms of RMSE and raw mean error. The results are summarized in table 5 in the results section. While the results in terms of improving the DEM are not widespread when comparing the linear regression and RF models, it is important to note that the RF model's biomass values are more representative of in-situ conditions. This is an indication that the RF model is capable of being generalized and testing it using different datasets is encouraged.

The findings discussed in this chapter mark an improvement to a previous species independent biomass-based DEM adjustment technique (S. Medeiros et al., 2015). The model developed in this paper indicates that the RF model is able to reliably classify zero biomass (ZBM) locations which can be crucial to the development of marsh models such as the MEM, Hydro-MEM, SLAMM and WARMER (Alizad et al., 2016; M. Swanson et al., 2014; James T Morris et al., 2002; Park et al., 1991). The next chapter summarizes the findings reported and discussed above. Moreover, recommendations are provided for consideration in future investigations.

# CHAPTER IV: CONCLUSIONS AND FUTURE WORK

## 4.1 Conclusions

Field and remote sensing data were used to develop biomass density estimation models. The best performing model was selected to classify biomass density using only remote sensing data. Adjustments were applied to the original DEM and a reduction of 54% and 58% to the RMSE and mean raw error, respectively, was achieved. Mean raw error dropped from 33 cm to 14 cm.

Developed models included a series of multilinear and random forest regressions. The best models from both analyses performed similarly in terms of reducing lidar-derived DEM error. However, the random forest regression (pseudo $R^2 = 0.94$) was superior to the multilinear regression ($R^2 = 0.80$) in estimating above ground biomass density. Moreover, the multilinear regression model was proved unable to identify zero biomass locations. On the other hand, the RF model was able to classify all zero biomass samples as either ZBM or low density for the entire population except for one sample.

## 4.2 Future Work

The models developed in this study can be improved by increasing the number of trainable samples. This can be attainable given standardized biomass density sampling techniques. Expandable datasets can be of interdisciplinary interest. Additionally, the random forest biomass density model can be enhanced by using an exhaustive hyperparameter optimization algorithm. However, this is computationally expensive since all combinations of potential hyperparameters are used for evaluating the model. Finally, the ability of the RF model to identify zero biomass

locations should be further investigated. Such a model can be used to delineate vegetated from non-vegetated areas which is crucial for modeling the evolution of coastal wetlands.

# REFERENCES


Ahmad, S., Kalra, A., & Stephen, H. J. A. i. W. R. (2010). Estimating soil moisture using remote sensing data: A machine learning approach. *33*(1), 69-80.

Akay, A. E., Oğuz, H., Karas, I. R., Aruga, K. J. E. m., & assessment. (2009). Using LiDAR technology in forestry activities. *151*(1-4), 117-125.

Ali, I., Greifeneder, F., Stamenkovic, J., Neumann, M., & Notarnicola, C. (2015). Review of Machine Learning Approaches for Biomass and Soil Moisture Retrievals from Remote Sensing Data. *7*(12), 16398-16421.

Alizad, K., Hagen, S. C., Morris, J. T., Bacopoulos, P., Bilskie, M. V., Weishampel, J. F., & Medeiros, S. C. (2016). A coupled, two-dimensional hydrodynamic-marsh model with biological feedback. *Ecological Modelling, 327*, 29-43. doi:https://doi.org/10.1016/j.ecolmodel.2016.01.013

Alpaydin, E. (2014). *Introduction to machine learning*: MIT press.

Archer, K. J., & Kimes, R. V. (2008). Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis, 52*(4), 2249-2260. doi:https://doi.org/10.1016/j.csda.2007.08.015

Ayodele, T. O. (2010). Types of machine learning algorithms. In *New advances in machine learning*: IntechOpen.

Baker, C., Lawrence, R., Montagne, C., & Patten, D. J. W. (2006). Mapping wetlands and riparian areas using Landsat ETM+ imagery and decision-tree-based models. *26*(2), 465.

Barbier, E. B., Hacker, S. D., Kennedy, C., Koch, E. W., Stier, A. C., & Silliman, B. R. (2011). The value of estuarine and coastal ecosystem services. *81*(2), 169-193. doi:10.1890/10-1510.1

Barbier, E. B. J. E. p. (2007). Valuing ecosystem services as productive inputs. *22*(49), 178-229.

Bell, F. W. J. E. E. (1997). The economic valuation of saltwater marsh supporting marine recreational fishing in the southeastern United States. *21*(3), 243-254.

Bhattacharya, B., Price, R., & Solomatine, D. J. J. o. H. E. (2007). Machine learning approach to modeling sediment transport. *133*(4), 440-450.

Birol, E., & Cox, V. (2007). Using choice experiments to design wetland management programmes: The case of Severn Estuary Wetland, UK. *Journal of Environmental Planning and Management, 50*(3), 363-380. doi:10.1080/09640560701261661

Bishop, C. M. (1995). *Neural networks for pattern recognition*: Oxford university press.

Breaux, A., Farber, S., & Day, J. J. J. o. e. m. (1995). Using natural coastal wetlands systems for wastewater treatment: an economic benefit analysis. *44*(3), 285-291.

Bréda, N. J. J. J. o. e. b. (2003). Ground-based measurements of leaf area index: a review of methods, instruments and current controversies. *54*(392), 2403-2417.

Breidenbach, J., Næsset, E., Lien, V., Gobakken, T., & Solberg, S. (2010). Prediction of species specific forest inventory attributes using a nonparametric semi-individual tree crown approach based on fused airborne laser scanning and multispectral data. *Remote Sensing of Environment, 114*(4), 911-924. doi:https://doi.org/10.1016/j.rse.2009.12.004

Breiman, L. J. M. l. (2001). Random forests. *45*(1), 5-32.

Bricker-Urso, S., Nixon, S., Cochran, J., Hirschberg, D., & Hunt, C. J. E. (1989). Accretion rates and sediment accumulation in Rhode Island salt marshes. *12*(4), 300-317.

Choi, H., & Bindschadler, R. J. R. S. o. E. (2004). Cloud detection in Landsat imagery of ice sheets using shadow matching technique and automatic normalized difference snow index threshold value decision. *91*(2), 237-242.

Christiansen, T., Wiberg, P. L., & Milligan, T. G. (2000). Flow and Sediment Transport on a Tidal Salt Marsh Surface. *Estuarine, Coastal and Shelf Science, 50*(3), 315-331. doi:https://doi.org/10.1006/ecss.2000.0548

Coplin, L. S., & Galloway, D. (1999). *Houston-Galveston, Texas: Managing coastal subsidence*.

Costanza, R., Pérez-Maqueo, O., Martinez, M. L., Sutton, P., Anderson, S. J., & Mulder, K. J. A. A. J. o. t. H. E. (2008). The value of coastal wetlands for hurricane protection. *37*(4), 241-249.

Craft, C., Clough, J., Ehman, J., Joye, S., Park, R., Pennings, S., . . . Environment, t. (2009). Forecasting the effects of accelerated sea-level rise on tidal marsh ecosystem services. *7*(2), 73-78.

Cutler, M. E. J., Boyd, D. S., Foody, G. M., & Vetrivel, A. (2012). Estimating tropical forest biomass with a combination of SAR image texture and Landsat TM data: An assessment of predictions between regions. *ISPRS Journal of Photogrammetry and Remote Sensing, 70*, 66-77. doi:https://doi.org/10.1016/j.isprsjprs.2012.03.011

Dewberry. (2009). *Final Report Of Specific Purpose Lidar Survey—Lidar, Breaklines and Contours for Franklin County Florida*. Retrieved from Tallahassee, FL, USA:

Donnelly, J. P., & Bertness, M. D. J. P. o. t. N. A. o. S. (2001). Rapid shoreward encroachment of salt marsh cordgrass in response to accelerated sea-level rise. *98*(25), 14218-14223.

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., . . . Bargellini, P. (2012). Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. *Remote Sensing of Environment, 120*, 25-36. doi:https://doi.org/10.1016/j.rse.2011.11.026

Eastwood, J., Yates, M., Thomson, A., & Fuller, R. J. I. J. o. R. S. (1997). The reliability of vegetation indices for monitoring saltmarsh vegetation cover. *18*(18), 3901-3907.

ESA. (2017). *Sentinel-2B* [INS-NOBS].

ESRI. (2011). ArcGIS Desktop: Release 10. Redlands, CA: Environmental Systems Research Institute.

Fagherazzi, S., Kirwan, M. L., Mudd, S. M., Guntenspergen, G. R., Temmerman, S., D'Alpaos, A., . . . Craft, C. J. R. o. G. (2012). Numerical models of salt marsh evolution: Ecological, geomorphic, and climatic factors. *50*(1).

French, J. R., & Spencer, T. (1993). Dynamics of sedimentation in a tide-dominated backbarrier salt marsh, Norfolk, UK. *Marine Geology, 110*(3), 315-331. doi:https://doi.org/10.1016/0025-3227(93)90091-9

Friedrichs, C. T., & Perry, J. E. J. J. o. C. R. (2001). Tidal salt marsh morphodynamics: a synthesis. 7-37.

Gesch, D. B. J. J. o. C. R. (2009). Analysis of lidar elevation data for improved identification and delineation of lands vulnerable to sea-level rise. 49-58.

Ghedira, H., Bernier, M., & Ouarda, T. (2000). *Application of neural networks for wetland classification in RADARSAT SAR imagery.* Paper presented at the IGARSS 2000. IEEE 2000 International Geoscience and Remote Sensing Symposium. Taking the Pulse of the Planet: The Role of Remote Sensing in Managing the Environment. Proceedings (Cat. No. 00CH37120).

Gleason, C. J., & Im, J. (2012). Forest biomass estimation from airborne LiDAR data using machine learning approaches. *Remote Sensing of Environment, 125*, 80-91. doi:https://doi.org/10.1016/j.rse.2012.07.006

Hladik, C., & Alber, M. J. R. S. o. E. (2012). Accuracy assessment and correction of a LIDAR-derived salt marsh digital elevation model. *121*, 224-235.

Hong, W.-C. J. A. M., & Computation. (2008). Rainfall forecasting by technological machine learning models. *200*(1), 41-57.

Hsieh, W. W. (2009). *Machine learning methods in the environmental sciences: Neural networks and kernels*: Cambridge university press.

Imhoff, M. (1995). Radar backscatter and biomass saturation-Ramifications for global biomass inventory IEEE Transactions on Geoscience and Remote Sensing. 33 (2): 511–518. doi: 10.1109/36.377953 View Article PubMed. In: NCBI.

Jachowski, N. R. A., Quak, M. S. Y., Friess, D. A., Duangnamon, D., Webb, E. L., & Ziegler, A. D. (2013). Mangrove biomass estimation in Southwest Thailand using machine learning. *Applied Geography, 45*, 311-321. doi:https://doi.org/10.1016/j.apgeog.2013.09.024

Jensen, J. R., Olson, G., Schill, S. R., Porter, D. E., & Morris, J. J. G. I. (2002). Remote sensing of biomass, leaf-area-index, and chlorophyll a and b content in the ACE Basin National Estuarine Research Reserve using sub-meter digital camera imagery. *17*(3), 27-36.

Jevrejeva, S., Moore, J., & Grinsted, A. J. G. r. l. (2010). How will sea level respond to changes in natural and anthropogenic forcings by 2100? *, 37*(7).

Jiang, L.-F., Luo, Y.-Q., Chen, J.-K., Li, B. J. E., Coastal, & Science, S. (2009). Ecophysiological characteristics of invasive Spartina alterniflora and native species in salt marshes of Yangtze River estuary, China. *81*(1), 74-82.

Kearney, M. S., Stutzer, D., Turpie, K., & Stevenson, J. C. J. J. o. C. R. (2009). The effects of tidal inundation on the reflectance characteristics of coastal marsh vegetation. 1177-1186.

Kesel, R. H. J. E. G., & Sciences, W. (1988). The decline in the suspended load of the lower Mississippi River and its influence on adjacent wetlands. *11*(3), 271-281.

Lauer, D. T., Morain, S. A., Salomonson, V. V. J. P. E., & Sensing, R. (1997). The Landsat program: Its origins, evolution, and impacts. *63*(7), 831-838.

Lu, D. J. I. j. o. r. s. (2006). The potential and challenge of remote sensing-based biomass estimation. *27*(7), 1297-1328.

Lucas, R. M., Cronin, N., Moghaddam, M., Lee, A., Armston, J., Bunting, P., & Witte, C. J. R. S. o. E. (2006). Integration of radar and Landsat-derived foliage projected cover for woody regrowth mapping, Queensland, Australia. *100*(3), 388-406.

M. Swanson, K., Drexler, J., H. Schoellhamer, D., Thorne, K., L. Casazza, M., T. Overton, C., . . . Takekawa, J. (2014). *Wetland Accretion Rate Model of Ecosystem Resilience (WARMER) and Its Application to Habitat Sustainability for Endangered Species in the San Francisco Estuary* (Vol. 37).

Mckee, K. L., & Patrick, W. J. E. (1988). The relationship of smooth cordgrass (Spartina alterniflora) to tidal datums: a review. *11*(3), 143-151.

Medeiros, S., Hagen, S., Weishampel, J., & Angelo, J. J. R. S. (2015). Adjusting lidar-derived digital terrain models in coastal marshes based on estimated aboveground biomass density. *7*(4), 3507-3525.

Medeiros, S. C., Hagen, S. C., & Weishampel, J. F. (2015). A Random Forest Model Based on Lidar and Field Measurements for Parameterizing Surface Roughness in Coastal Modeling. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 8*(4), 1582-1590. doi:10.1109/JSTARS.2015.2419817

Mitchell, T. M. (1997). *Machine Learning*: McGraw-Hill.

Mitsch, W. J., Gosselink, J. G., Zhang, L., & Anderson, C. J. (2009). *Wetland ecosystems*: John Wiley & Sons.

Morgan, P. A., Burdick, D. M., Short, F. T. J. E., & Coasts. (2009). The functions and values of fringing salt marshes in northern New England, USA. *32*(3), 483-495.

Morris, J. T., Porter, D., Neet, M., Noble, P. A., Schmidt, L., Lapine, L. A., & Jensen, J. R. (2005). Integrating LIDAR elevation data, multi-spectral imagery and neural network modelling for marsh characterization. *International Journal of Remote Sensing, 26*(23), 5221-5234. doi:10.1080/01431160500219018

Morris, J. T., Sundareshwar, P., Nietch, C. T., Kjerfve, B., & Cahoon, D. R. J. E. (2002). Responses of coastal wetlands to rising sea level. *83*(10), 2869-2877.

Morton, R. A., Bernier, J. C., & Barras, J. A. J. E. G. (2006). Evidence of regional subsidence and associated interior wetland loss induced by hydrocarbon production, Gulf Coast region, USA. *50*(2), 261.

Mutanga, O., Adam, E., & Cho, M. A. (2012). High density biomass estimation for wetland vegetation using WorldView-2 imagery and random forest regression algorithm. *International Journal of Applied Earth Observation and Geoinformation, 18*, 399-406. doi:https://doi.org/10.1016/j.jag.2012.03.012

Nicholls, R. J., Hoozemans, F. M., & Marchand, M. J. G. E. C. (1999). Increasing flood risk and wetland losses due to global sea-level rise: regional and global analyses. *9*, S69-S87.

Pal, M., & Mather, P. M. J. R. s. o. e. (2003). An assessment of the effectiveness of decision tree methods for land cover classification. *86*(4), 554-565.

Park, R. A., Lee, J. K., Mausel, P., & Howe, R. J. W. R. R. (1991). Using remote sensing for modeling the impacts of sea level rise. *3*(2), 0-2.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Dubourg, V. J. J. o. m. l. r. (2011). Scikit-learn: Machine learning in Python. *12*(Oct), 2825-2830.

Pinty, B., & Verstraete, M. J. V. (1992). GEMI: a non-linear index to monitor global vegetation from satellites. *101*(1), 15-20.

Pomeroy, L. R., & Wiegert, R. G. (2012). *The ecology of a salt marsh* (Vol. 38): Springer Science & Business Media.

Priest, B. (2011). *Effects of elevation and nutrient availability on the primary production of Spartina alterniflora and the stability of southeastern coastal salt marshes relative to sea level rise.* University of South Carolina,

Rahmstorf, S. J. S. (2007). A semi-empirical approach to projecting future sea-level rise. *315*(5810), 368-370.

Rasouli, K., Hsieh, W. W., & Cannon, A. J. J. J. o. H. (2012). Daily streamflow forecasting by machine learning methods with weather and climate inputs. *414*, 284-293.

Reed, D. J., Spencer, T., Murray, A. L., French, J. R., & Leonard, L. J. J. o. C. C. (1999). Marsh surface sediment deposition and the role of tidal creeks: Implications for created and managed coastal marshes. *5*(1), 81-90.

Roushangar, K., Akhgar, S., Salmasi, F., & Shiri, J. J. J. o. H. (2014). Modeling energy dissipation over stepped spillways using machine learning approaches. *508*, 254-265.

Shastry, A., & Durand, M. (2019). Utilizing Flood Inundation Observations to Obtain Floodplain Topography in Data-Scarce Regions. *6*(243). doi:10.3389/feart.2018.00243

Subrahmanyam, C., & Drake, S. H. J. B. o. M. S. (1975). Studies on the animal communities in two north Florida salt marshes Part I. Fish communities. *25*(4), 445-465.

Szantoi, Z., Escobedo, F. J., Abd-Elrahman, A., Pearlstine, L., Dewitt, B., Smith, S. J. E. m., & assessment. (2015). Classifying spatially heterogeneous wetland communities using machine learning algorithms and spectral and textural features. *187*(5), 262.

Tahsin, S., Medeiros, S., Hooshyar, M., & Singh, A. J. R. S. (2017). Optical cloud pixel recovery via machine learning. *9*(6), 527.

Takasu, T., & Yasuda, A. (2008). *Evaluation of RTK-GPS performance with low-cost single-frequency GPS receivers.* Paper presented at the Proceedings of international symposium on GPS/GNSS.

Tseng, D.-C., Tseng, H.-T., & Chien, C.-L. (2008). Automatic cloud removal from multi-temporal SPOT images. *Journal of Applied Mathematics & Computation, 205*(2), 584-600.

USGS. (2017). *Landsat 8 Thematic Mapper* [UINT16].

Waring, R. H., Way, J., Hunt, E. R., Morrissey, L., Ranson, K. J., Weishampel, J. F., . . . Franklin, S. E. J. B. (1995). Imaging radar for ecosystem studies. *45*(10), 715-723.

Xingjian, S., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., & Woo, W.-c. (2015). *Convolutional LSTM network: A machine learning approach for precipitation nowcasting.* Paper presented at the Advances in neural information processing systems.

Zhu, Z., & Woodcock, C. E. J. R. s. o. e. (2012). Object-based cloud and cloud shadow detection in Landsat imagery. *118*, 83-94.

Zimmerman, R. J., Minello, T. J., & Rozas, L. P. (2002). Salt marsh linkages to productivity of penaeid shrimps and blue crabs in the northern Gulf of Mexico. In *Concepts and controversies in tidal marsh ecology* (pp. 293-314): Springer.