# Improving Salt Marsh Digital Elevation Model Accuracy with Full-Waveform Lidar and Nonparametric Predictive Modeling

**Jeffrey N. Rogers[1,2], Christopher E. Parrish[3,4], Larry G. Ward[1,4,6], David M. Burdick[5,6]**

**11/20/17**

**1 - Department of Earth Sciences, University of New Hampshire, Durham, NH 03824**
**2 - Center for Coastal Studies, Provincetown, MA 02657**
**3 - School of Civil and Construction Engineering, Oregon State University, Corvallis, OR 97331**
**4 - Center for Coastal and Ocean Mapping/Joint Hydrographic Center, School of Marine Science and Ocean Engineering, University of New Hampshire, Durham, NH 03824**
**5 - Department of Natural Resources and the Environment, University of New Hampshire, Durham, NH 03824**
**6 - Jackson Estuarine Laboratory, School of Marine Science and Ocean Engineering, University of New Hampshire, Durham, NH 03824**

## ABSTRACT

Salt marsh vegetation tends to increase vertical uncertainty in light detection and ranging (lidar) derived elevation data, often causing the data to become ineffective for analysis of topographic features governing tidal inundation or vegetation zonation. Previous attempts at improving lidar data collected in salt marsh environments range from simply computing and subtracting the global elevation bias to more complex methods such as computing vegetation-specific, constant correction factors. The vegetation specific corrections can be used along with an existing habitat map to apply separate corrections to different areas within a study site. It is hypothesized here that correcting salt marsh lidar data by applying location-specific, point-by-point corrections, which are computed from lidar waveform-derived features, tidal-datum based elevation, distance from shoreline and other lidar digital elevation model based variables, using nonparametric regression will produce better results. The methods were developed and tested using full-waveform lidar and ground truth for three marshes in Cape Cod, Massachusetts, U.S.A. Five different model algorithms for nonparametric regression were evaluated, with TreeNet's stochastic gradient boosting algorithm consistently producing better regression and classification results. Additionally, models were constructed to predict the vegetative zone (high marsh and low marsh). The predictive modeling methods used in this study estimated ground elevation with a mean bias of 0.00 m and a standard deviation of 0.07 m (0.07 m root mean square error). These methods appear very promising for correction of salt marsh lidar data and, importantly, do not require an existing habitat map, biomass measurements, or image based remote sensing data such as multi/hyperspectral imagery.

Index words: *Spartina alterniflora*, Random Forests, TreeNet Stochastic Gradient Boosting, regression trees, CART, DEM correction

## 1. Introduction

Salt marshes are saline wetlands dominated by grasses and other plants adapted to periodic flooding, usually as a result of tidal forcing (Mitsch and Gosselink, 2000). Salt marshes are found throughout middle to high latitudes and exhibit characteristic patterns of vegetation zonation that are often based on an elevation gradient (Morris et al., 2005; Zedler et al., 1999). Salt marshes provide valuable ecosystem functions, such as critical wildlife and biodiversity support, water quality improvement, and coastal storm protection (Costanza et al., 1997; Mitsch and Gosselink, 2000). Geomorphically, salt marshes are often separated from adjacent tidal flats by a ramp or abrupt change in elevation caused by increased sedimentation, peat development and decreased erosion due to vegetation (Crooks et al., 2002; Fagherazzi et al., 2006). These low-lying landforms are poised systems, balancing accretion and storage with erosion and oxidation of sediments in response to tidal flooding (Roman and Burdick, 2012) and, therefore, are sensitive to increases in water levels resulting from sea-level rise (SLR). In general, very small variations in elevation, which affect inundation, available sediment, nutrients and salinity, determine whether salt marsh species thrive, survive or fail (Morris et al., 2002). Therefore, SLR is a major cause of concern for coastal scientists and managers.

Accurately determining salt marsh elevation is fundamental to understanding almost every aspect of marsh system science and management, including response to SLR and storm surge inundation, in terms of adaptation and resiliency. However, obtaining high-resolution, high-accuracy digital elevation models (DEMs) of salt marshes can be difficult, costly, and time consuming using traditional data collection methods. The importance of lidar (light detection and ranging) for conducting rapid surveys of salt marshes has been recognized (Brock and Sallenger, 2001), and the technology is often proposed as a substitute for field-based data sets collected by either differential leveling or RTK GNSS (Real-Time Kinematic Global Navigation Satellite System) surveys (Montane and Torres, 2006; Schmid et al., 2011). Further, the ability to map major plant communities using remote sensing, which always appears out of reach with each new technological breakthrough, would be of great value to salt marsh ecologists and managers.

An inherent problem with the use of lidar in salt marshes is the vegetation typically increases the vertical uncertainty. That uncertainty can be quantified empirically as the root mean square error (RMSE), obtained by comparison against RTK GNSS, as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(Z_i - Z_{i,c})^2} \qquad (1)$$

where $Z_i$ is the i$^{th}$ lidar-derived elevation and $Z_{i,c}$ is the corresponding ground control elevation. The RMSE can also be decomposed into the bias, $\mu$ (the mean difference between what the lidar determines to be bare earth elevation and ground control) and standard deviation of elevation differences about the mean, $\sigma$. For large sample sizes, $N$, the following relation is expected to hold (Stewart et al., 2009):

$$RMSE^2 \approx \mu^2 + \sigma^2 \qquad (2)$$

For lidar to serve as a viable technology in salt marsh research and planning, the observed uncertainty in elevation needs to be less (preferably much less) than the elevation ranges of ecological importance (Sadro et al., 2007). For instance, if the uncertainty due to vegetative impacts on the determination of elevation from the lidar signal is greater than the elevation range determining species dominance and habitat, then lidar is not useful for restoration planning, hydrologic modeling, and SLR studies. These uncertainties can be seasonally driven depending on marsh location or region, since many marsh systems cycle between senescence and peak growth conditions. Quantifying uncertainties of salt marsh lidar data and applying corrections to produce accurate DEMs has, to date, been only partially resolved. In general, uncorrected lidar datasets from salt marshes lack sufficient accuracy for use in the tasks mentioned above (Hladik and Alber, 2012; Rosso et al., 2006; Schmid et al., 2011). However, research to determine the extent to which lidar penetrates the salt marsh canopy and methods to correct for vegetation-induced elevation uncertainty have begun to achieve results (Buffington et al., 2016; Gopfert and Heipke, 2006; Hladik and Alber, 2012; Hladik et al., 2013; Medeiros et al., 2015; Populus et al., 2001; Rogers et al., 2016; Rosso et al., 2006; Schmid et al., 2011).

## 1.1 Previous Research

Prior attempts at developing correction techniques for vegetation-induced lidar uncertainty have involved: 1) subtracting off a global (i.e., computed for the entire data set) elevation bias; 2) filtering/interpolation/classification methods (Schmid et al., 2011); 3) reduction based on canopy height, density , or above ground biomass coverage (Medeiros et al., 2015; Wang et al., 2009); 4) subtraction of species-specific bias based on vegetation cover maps (Hladik and Alber, 2012; Hladik et al., 2013); and 5) use of Normalized Difference Vegetation Index (NDVI) (Buffington et al., 2016). Due to the spatial variation in elevation uncertainty across a marsh (Parrish et al., 2014), subtracting a global bias tends to overcorrect the elevation error in some places and under correct in others. Filtering and interpolation correction methods are greatly hindered by the dearth of true ground returns from the low, dense growing salt marsh vegetation and the potential inaccuracies introduced by uncertainty in the separation of ground and vegetation returns (Rogers et al., 2016; Sadro et al., 2007; Schmid et al., 2011; Wang et al., 2009). While relationships between vegetation canopy height, percent coverage and lidar uncertainty have been observed (Gopfert and Heipke, 2006; Populus et al., 2001; Schmid et al., 2011), these methods often fail to produce the desired level of elevation correction in a salt marsh.

Advancements in salt marsh DEM correction methods have been made by conducting species-specific elevation correction (Hladik and Alber, 2012; Hladik et al., 2013; McClure et al., 2016; Sadro et al., 2007). Since the error is primarily attributable to vegetation and tends to be species-dependent, this method vastly improved DEM accuracy by focusing the appropriate amount of correction where it is needed. Unfortunately, a requirement of vegetation-based correction techniques is *a priori* knowledge of species distribution. From past project experience, existing vegetation maps are typically unavailable, too outdated, too coarse, or too inaccurate for many project sites. If a project requires collecting this information, it would also necessitate additional fieldwork or multi/hyperspectral sensor data that adds to cost, time and introduced errors. However, even if vegetation data were available and accurate, salt marsh species often present ranges of elevation uncertainty that fall in a continuous distribution rather than a constant (Rogers et al., 2016). Lidar uncertainty in salt marsh environments is influenced by vegetation height, stem density, biomass, and species growth habit (Buffington et al., 2016; Hladik and Alber, 2012; Rogers et al., 2015; Schmid et al., 2011). These vegetation characteristics vary over the marsh surface as a function of edaphic conditions (nutrients, salinity, sulfide concentrations, lower redox potential) and time of year, as well as other factors (Bertness and Ellison, 1987; Byrd and Kelly, 2006; Mendelssohn et al., 1981; Mitsch and Gosselink, 2000). For example, medium-form *Spartina alterniflora* has a height range of 50 - 100 cm, and one would expect the observed lidar uncertainty to have a range as well. It seems unlikely that each vegetation species/ecophene region would require a constant DEM correction factor across its entire extent (Hladik and Alber, 2012; Hladik et al., 2013).

## 1.2 Full-waveform and Nonparametric Modeling Approach

An alternate method to the problem of salt marsh lidar elevation correction involves the use of full-waveform lidar systems. Full-waveform equipment records a time series of backscattered energy with a digitizer and a high-capacity storage device. The amplitude of the laser return is dependent on the power of the transmitted pulse, the range, the surface-intercepted fraction of the pulse, the surface reflectance, the incidence angle, and the fraction of the pulse returned toward the sensor (Lefsky et al., 2002). As a result only a small fraction of the transmitted energy from the initial pulse returns to the sensor from the ground target (Wagner et al., 2008). Ground targets, such as vegetation, soil and other objects tend to have a rough surface at the near infrared (NIR) wavelengths commonly used in topographic lidar and generally scatter lidar energy diffusely, at least as a first-order approximation. Water is often observed as a data void since most of the energy is absorbed or undergoes specular reflection in a direction away from the sensor, although some strong, specular returns from near-nadir beams (i.e., directly below the aircraft) are often observed.

Full-waveform digitizing systems reveal the vertical distribution of the targets for the nadir beams and resolve surfaces closer together in the range direction than discrete-return lidar (DRL) systems (Anderson et al., 2008; Drake et al., 2002; Lefsky et al., 2002; Parrish et al., 2011). Data processing techniques for full-waveform lidar usually involve computationally-complex decomposition or deconvolution (Jutzi and Stilla, 2006) of the returned backscatter into relevant peaks to generate denser point clouds then would be available from DRL (Mallet and Bretar, 2009; Wagner et al., 2008). Studies utilizing simple, feature-based waveform metrics have started to demonstrate utility in the waveform data beyond these resource intensive approaches (Adams et al., 2012; Muss et al., 2013; Parrish et al., 2014; Rogers et al., 2015, 2016).

In a previous study by the authors, it was observed that distributions of vegetation height display unique, species-based characteristics (**Figure 1**) (Rogers et al., 2016). While this relationship appeared to be particularly true with *S. alterniflora* and *Salicornia spp.*, *S. patens* and *D. spicata* maintained very similar growth characteristics and range of elevation dominance. In New England salt marshes, a known association between elevation and vegetation height exists, such that as marsh elevation decreases the vegetation height increases (**Figure 2**). It has also been determined that individual marsh species exhibit varying ranges of elevation uncertainty unique to their growth and form (Hladik et al., 2013; Rogers et al., 2016; Schmid et al., 2011). Therefore, the ability to discriminate between species using these and other observable characteristics and relationships might play a role in determining a lidar elevation correction strategy. Furthermore, a relationship between metrics derived from lidar waveform features (in particular waveform width's association with elevation uncertainty and vegetation height) (Parrish et al., 2014; Rogers et al., 2015, 2016), suggested that a non-parametric modeling approach might lead to a successful correction technique.

Problems with numerous independent variables and complex, possibly nonlinear relationships lend themselves to the use of machine learning and nonparametric modeling techniques. Unlike typical statistical analysis of dependent and independent variables that utilize single or multiple regression techniques to make predictions of variable outcome, nonparametric modeling does not necessitate any hypothesis concerning variable distribution as a prerequisite to analysis (Bourennane et al., 2014). Nonlinear approaches are often required in environmental modeling problems due to the complex and often concealed relationships between predictor variables (Tayyebi and Pijanowski, 2014).

The research presented here investigates the following: 1) the potential correction of vegetation-induced elevation error using full-waveform lidar feature-based metrics such as waveform width and amplitude, as well as salt marsh surface characteristics such as slope and rugosity derived from the DRL, as inputs into a battery of nonparametric modeling algorithms; 2) the use of nonparametric modeling and DRL-derived salt marsh surface characteristics (i.e. no full-waveform inputs included) to reduce vegetation-induced error; and 3) creation of a vegetative zone maps using the same modeling parameters and a training set of known vegetation species locations. The ultimate goal of this work is to enable generation of models that can correct salt marsh lidar-derived DEMs to a level suitable for ecological and SLR applications. Also, it may be possible to derive vegetative classification maps from lidar data (with limited ground truth efforts) that could assist researchers with locating habitat, research planning, or vegetation modeling.

## 2. Methods

The study sites are comprised of three individual, mesotidal salt marshes located on protected shorelines of Cape Cod, Massachusetts (Moors marsh: ~2.0 km$^2$; Pamet River marsh: ~2.0 km$^2$; and Great Island Middle marsh; ~0.3 km$^2$) (**Figure 3**). The study area is characterized by semidiurnal tides with a mean range of ~2.83 m (NOAA, 2013) . The marshes were selected based on the following criteria: 1) they are physically close to one another, but hydrologically separate, 2) they contain large stands of the major marsh species present in northeastern United

States, and 3) they are easily accessible enabling collection of field data within a narrow time window.

### 2.1 Vegetative Community

Low marsh environments dominated by *Spartina alterniflora* (smooth cordgrass) are most commonly found in the studied marshes. However, small topographic highs (typically isolated) and small borders of high marsh located in the landward portions of the marshes are dominated by *S. patens* (salt marsh hay), *D. spicata* (spike grass) and *Salicornia spp.* (glasswort) (Portnoy et al., 2003) **(Figure 4)**. Salt marsh vegetation demonstrates zonation driven by small elevation changes and edaphic conditions (Bertness and Ellison, 1987). Varying plant morphologies and growth habits have evolved by each vegetation species to adapt to the harsh conditions found in tidal marshes. The vegetation occurs as homogeneous, near monoculture stands for the three major species and one genus (*Spartina alterniflora, Spartina patens, Distichlis spicata*, and *Salicornia spp.).* Within each vegetative community there is variability in growth habit and height. For example, *Spartina alterniflora* at these sites has three distinct variations or ecophenes caused by edaphic factors, often reported as short form (0-50 cm; SF), medium form (50-100 cm; MF), and tall form (>100 cm; TF) (Anderson and Treshow, 1980; Hladik and Alber, 2012; Ornes and Kaplan, 1989; Pennings and Bertness, 2001; Reimold et al., 1973; Wiegert and Freeman, 1990). Tall-form *S. alterniflora* ranges up to 2 m in height and is typically found at lower elevations and along estuarine creeks. In contrast, SF *S. alterniflora* is commonly found in high marsh depressions with higher salinity, greater sulfide concentrations and/or lower redox potential (Mitsch and Gosselink, 2000).

## 2.2 Lidar Data Collection

Approximately 37 $km^2$ of lidar data was collected by The National Center for Airborne Laser Mapping (NCALM) on July 20th, 2010 centered on the daily predicted low tide (± 90 minutes) during peak biomass. An Optech Gemini Airborne Laser Terrain Mapper (ALTM) and an Optech 8-bit IWD intelligent waveform digitizer (serial number 08DIG017) were mounted in a twin-engine Cessna 337 Skymaster. Data were collected at a pulse repetition rate of 70 kHz and a flight speed of 60 m/s and altitude of 600 m (**Table 1**). DRL was collected concurrently using the Optech hardware-based constant fraction discriminator and time interval meter. Waveform data were sampled at 1 ns intervals and delivered in Optech's NDF (digitizer file) binary format with an IDX index file and CSD (corrected sensor data) file.

The salt marshes studied were comprised of low-growing marsh vegetation, "bare earth" and water features and did not include trees, buildings, or other structures such that the dataset was almost entirely composed of single return pulses (Rogers et al., 2015). It is important to note that the lidar system used in this study had a long transmit pulse width of ~12 ns [full-width at half-maximum (FWHM)] at 70 kHz PRF, corresponding to ~1.8 m of range. Salt marsh vegetation with heights significantly less than the range-equivalent transmit pulse width typically show return waveforms that contain just a single peak (Parrish et al., 2014; Rogers et al., 2015).

Elevations delivered in NAVD88 were converted to a local tidal datum, mean high water (MHW), using NOAA's Vertical Datum Transformation (VDatum) version 3.2 (Yang et al., 2013) for consistency with NOAA shoreline definitions. Conversions performed in VDatum do introduce some additional uncertainty in vertical coordinates (Cooper et al., 2013). For the Gulf of Maine VDatum region, the NAVD88-MHW transformation uncertainty is reported by NOAA to be 11 cm (1-$\sigma$) (NOAA, 2017b). However, due to the small spatial extents of our project sites, the NAVD88-MHW separation is very nearly constant throughout the sites, and, therefore, the vertical datum transformation uncertainty can be treated as systematic uncertainty (i.e., removable, through a simple global bias correction), in contrast to the more complex, spatially-varying elevation uncertainties considered elsewhere in this paper.

### 2.3 Field Data Collection

To characterize the salt marsh environment, ~2,800 ground control points (GCPs) were established in various zones including tidal sandflats, low marsh, and high marsh. GCPs were collected along quasi-randomly oriented and spaced transect lines using 30-sec RTK GNSS occupation times. In this case, quantity of training and test data for the model generation outweighed the need for a small number of truly random sample stations. Hard surfaces such as roads and parking lots in close proximity to the marshes were also surveyed to analyze for the overall lidar dataset accuracy (Rogers et al., 2016). Marsh surface elevations and hard target GCPs were collected with a Trimble NetR5 base station network with cellular-based correction and a Trimble R8 Model 3 RTK GNSS rover. Due to the conditions found in salt marsh environments, special care was needed when using the rover to ensure vertical accuracy (Torres and Styles, 2007). A GNSS survey rod was modified with a 12 cm diameter flat base to keep the rod from depressing into the unconsolidated mud and peat. Transects were taken through the marsh to record ground elevations, with an average point spacing of 5-7 m. The GNSS equipment provided an RMSE of < 1 cm in the horizontal and 2 cm in the vertical (based on comparisons against geodetic control within the survey site), with elevations referenced to NAVD88 using GEOID09 (the latest NGS geoid model available at the time). At each location surface conditions were recorded such as the presence of sand, mud or dominant vegetation species and canopy height for later use in the model.

## 2.4 Model Predictor Variables

A custom process was developed with ArcGIS10, QCoherent LP360 and MATLAB to extract lidar waveforms from the provided data files and compute waveform shape-related metrics. This research leverages previous work on waveform shape metrics by the authors (Parrish et al., 2014; Rogers et al., 2015, 2016). The effects of variable lidar incidence angle on the waveform metrics, were tested in a previous study (Parrish et al., 2014) and found that, with the low flying height (600 m), narrow beam divergence (0.25 mrad), relatively flat terrain of salt marshes, and relatively small scan angles (±21°) used in the Cape Cod data acquisition, the effects are negligible. This finding is consistent with that of Bretar et al. (2009). Each lidar point within a subset of the studied marshes had a number of waveform features calculated, including lidar echo width, mean, area under the curve (AUC), skewness, and peak amplitude (**Table 2**). Each of the feature metrics was exported as an individual ASCII file and gridded in ArcGIS Spatial Analyst using an inverse distance weighting (IDW) with a 1 m cell size.

The DRL dataset used to produce predictor variables for the model included uncorrected lidar elevations and other surface measures such as rugosity and slope **(Table 2)**. Lidar point clouds in the LAS file format were preprocessed using QPS Fledermaus PFM 3D v7.43. Lidar data evaluation and cleaning were performed using the PFM 3D point cloud editor to remove artifacts as well as erroneous or non-natural points that could influence the gridding results. Elevations were converted to MHW in VDatum v. 3.2 and gridded using an IDW interpolation method with a cell size of 1 m and a search radius of 1 (Rogers et al., 2016). IDW uses a weighted average of the *n* nearest elevation points, where the weights are inversely proportional to distance from the cell being analyzed (Ries, 1993). Comparison studies between interpolation methods suggest that results between various methods (inverse distance weighting; ordinary kriging; universal kriging; multiquadratic radial basis function; and regularized spline with tension) are not appreciably different if the sampling density is high, but under low sampling density kriging techniques are preferable (Chaplot et al., 2006). Since this study acquired lidar with approximately 5 pts/$m^2$ point density (return density), the IDW interpolation method was chosen as the preferred method due to its fast processing speed. In addition, using a search radius of 1 restricted the final cell elevation to be based only on returns from that cell. Landscape metrics derived from the lidar DEM such as surface slope (the rate of change in value from each cell to its neighbors (Burrough and McDonell, 1998)), and three measures of curvature (fourth-order polynomials of a surface on a cell-by-cell basis (curvature, profile curvature and planimetric curvature (Zevenbergen and Thorne, 1987)), were calculated with ArcGIS v10. Rugosity, which is a measure of surface roughness (Sappington et al., 2007), was calculated using Benthic Terrain Modeler for ArcGIS10 (NOAA, 2017a) using the gridded DRL elevation.

It was critical that the elevation data used in this research be referenced to a local tidal datum such as MHW as opposed to NAVD88 orthometric heights or NAD83 ellipsoid heights because salt marsh vegetation speciation is tidally driven. A relationship has been established between tidal datum elevations (i.e. Mean High Water [MHW]) and the frequency of salt marsh species occurrence (Lefor et al., 1987; Mckee and Patrick, 1988; Morris et al., 2005). Therefore a tidal datum is the best possible method to analyze difference in topographic height and speciation that will assist with model pattern recognition. Another reason the MHW datum was chosen was to be consistent with the NOAA Continually Updated Shoreline Product (CUSP) (NOAA, 2016).

Distance from the shoreline was the only model input variable not taken directly from the lidar metrics but was a derivative product from the lidar. The distance from the shoreline has a direct influence on inundation frequency and edaphic conditions and, therefore, vegetation speciation (Andrew and Ustin, 2009; Griffin et al., 2011; Hladik and Alber, 2014; Sanderson et al., 2001). The -1.0 m MHW shoreline (i.e., the -1.0 m elevation contour, relative to MHW) was extracted from the lidar following procedures used by NOAA NGS (Graham et al., 2003; White et al., 2011). For this study, the -1.0 m MHW contour line closely followed the lowest most extent of vegetation. Also referenced was a 2009 3-band (RGB) MassGIS high resolution (0.3 m pixel) orthophoto captured one year prior to the lidar survey. The final shoreline was an interpretation of the extracted shoreline and the orthophotography and in this case represents -1.0 m or the lowest extent of vegetation. This orthophoto was also used in the photo interpretation of marsh vegetation zones.

Full-waveform lidar was collected for the entire geographic area covering the selected salt marshes. However, to enable the processing and multiple model runs to execute in a reasonable amount of time, priority subregions (red boxes in **Figure 3**) were identified and used in the analysis. Therefore, the model training dataset included only field collected RTK GNSS data for "true" ground elevation in MHW data that were bound by the extracted subset of full-waveform data (n = 785 out of total collected n > 2800). The data file also included the dominant vegetation species found at each location and was intersected with the multiple predictor grid layers calculated above. Using the "extract multivalues to point" utility in ArcGIS10, all XY locations were attributed with the corresponding waveform or surface values found in **Table 2**.

Distance to shoreline was calculated in meters for each point using a “multiple minimum distance” script in ArcGIS and the positive direction was defined to be shoreward of this line. This same subset of ground control data was also used for the DRL model runs for consistency and comparability between the various models.

## 2.5 Models and Model Construction

The complex and often nonlinear relationships between predictors can be extracted using nonparametric, computer-based, predictive modeling with the 13 predictor variables available in this study (**Table 2**), without any prior assumptions as to the distribution of the variables. All models used in this study were created using Salford Predictive Modeler version 7.0, a commercially available software by Salford Systems (www.salford-systems.com). A battery of five nonparametric model runs were conducted including Stochastic Gradient Boosting of Trees [TreeNet], Multivariate Adaptive Regression Splines [MARS], Generalized Path Seeker Model [GPSM], Random Forest [RF], Classification and Regression Trees [CART]) (**Table 3**). One parametric model was also used (Stepwise Least Squares Regression). The support vector machine (SVM) algorithm was not selected as a testing model in favor of using Stochastic Gradient Boosting (TreeNet), which has been found to match or exceed SVM’s prediction performance (Zhang et al., 2017). The 785 GCPs of available data from the three hydrologically separate study marshes were combined into one database and then partitioned into “learn” (n = 560 [71%]) and “test” (n = 225 [29%]) datasets. The modeling software randomly selects records from the provided dataset based on the user preference of the required test partition size. The commonly-referenced standard is an 80/20 split of test to learn records. However, in this analysis a slightly more robust training sample size of nearly 30% (70/30) was established to ensure model accuracy on the independent dataset. The test data are held back from the model development process making them completely independent of the model learn data and are used solely for model validation. Up to 20 model runs with randomly selected learn and test sets were conducted in order to verify robustness of the results and to ensure that the model results were not dependent on a randomly selected best case scenario from the learn dataset. Models were then evaluated for their performance using three criteria: 1) a high regression coefficient of determination ($R^2$) with the independent test dataset; 2) similar regression coefficients between learn and test datasets; and 3) the closeness of fit of the final regression equation line to a perfect 1:1. Therefore, a perfect model would produce an $R^2$ value equal to 1 and an equation of y = x. An algorithmic-level description of the different models is available in the references listed in the last column of **Table 3**. In the implementation of each of the following models, the algorithm rules were selected to maximize the accuracy and then tested on the independent test dataset.

The predictor parameter variables existed for every pixel (1 $m^2$) of the marsh surface and the final developed models were then "scored" against the complete marsh-wide grid of 525,941 pixels. To accomplish this, the parameter (P) grids were exported to ASCII format from ArcGIS and a single spreadsheet (X, Y, Z, P1-Pn) was created. The table was imported into the Salford Predictive Modeler software and using the best performing correction models, the entire table was scored for each geographic coordinate with all the required predictor variables. Finally, the X, Y and model corrected Z values were exported and gridded in ArcGIS to create a new DEM.

## 2.6 Vegetation Classification

A similar method was also employed to create vegetation classification maps for the entire marsh based solely on the waveform shape-based metrics and not spectrum. Firstly a 2009 3-band

(RGB) MassGIS high resolution (0.3 m pixel) orthophoto captured one year prior to the lidar survey was used in conjunction with the field collected locations (n = 785) of ground conditions and species dominance to create a GIS map of marsh vegetation zones. The marsh zonation field map consisted of three classes: bare ground (GR), high marsh vegetation [*S. patens*, *Salicornia spp.*, *D. spicata*, short-form *S. alterniflora*] (HM), and low marsh vegetation [tall-form and medium-form *S. alterniflora*] (LM). Next, the field data spreadsheet from the elevation correction model runs with X, Y, Z, P1-Pn and species dominance was updated to include the new parameter of marsh zone (GR, LM, HM). The table was imported into the Salford Predictive Modeler software and classification models were created and confusion matrices were constructed for the best performing models. Lastly, the previously discussed method for creating a grid of the entire marsh based on the model results was used to develop vegetation zonation classification maps. These maps were then visually analyzed for consistency with the aerial/field interpretation maps.

## 3. Results

Predictive modeling runs were conducted with a variety of parameters from both full-waveform and DRL sources in addition to DRL only analyses. Since the set of predictors sufficient to provide discriminatory power and high predictive model accuracy for correcting vegetation induced uncertainty in salt marshes was unknown, a number of commonly used lidar derivative products were added to the full-waveform metrics for analysis. These parameters were added and removed from successive modeling runs to test results. In addition, multiple nonparametric algorithms were utilized to find the best performing model and ensure validity of all the model results to minimize the possibility of overfitting. Final development of the best performing model was also conducted up to 20 times using randomly selected different learning and test datasets from the available data to verify model consistency. Field and lidar data from three different marshes were used to create the model, which also limits the possibility of the model becoming overly fit to a single location.

### 3.1 Elevation Correction Models Using Full-waveform Metrics and DRL Predictors

The results of five different regression-based nonparametric models and one parametric model are presented in **Table 4**. The dataset used for these model runs included all available waveform metrics as well as those predictors derived from the DRL elevation data from the same flight (**Table 2**). The resulting models produced "test" sample regression coefficients ranging from $R^2$ = 0.919 to 0.963 with regression line slopes from 0.897 to 0.982 and y intercepts near 0. The top two most successful models were TreeNet and MARS with test sample $R^2$ values of 0.96 and slopes within 4% of 1:1. Since the learn and test sample results were very close in $R^2$ values, the model was scored (i.e. run) against all of the available data with ground truth RTK GNSS elevations (Learn + Test samples, n = 785) and plotted with the original uncorrected lidar data to visualize the improvement. The TreeNet algorithm produced better results than MARS on the independent test sample with a tighter linear clustering for the scored dataset of all available data with an $R^2$ of 0.982 compared to an uncorrected lidar $R^2$ of 0.797 **(Figure 5a**). The MARS model results appear to be a little more scattered than the TreeNet model with additional negative residuals (**Figure 5b**).

Predictor variable importance is a significant tool in evaluating model results. The Salford Systems modeling software assigns the most important variable a score of 100 and all other variables are rescaled relative to the most important variable. This importance score measures the performance of the variable as a primary or surrogate splitter for each individual tree evaluated and not its value in relation to other trees. As more and more trees are used in the model construction, more predictor variables have an opportunity to influence decision trees. Since relative variable importance within decision trees in any given model and across models can greatly differ, variable importance can't be an absolute value or percentage.

An apparent trend exists in the variable importance among the various nonparametric models (**Table 5**). The obvious and most influential variable when calculating corrected elevation is uncorrected lidar elevation. The second most important variable in 4 of 5 nonparametric models was waveform width. The CART model defined distance from shoreline as the second and waveform width as the third most important variables. The predictive power of waveform width is consistent with previous findings by the authors in relation to observed lidar uncertainty and vegetation characteristics such as height (Parrish et al., 2014; Rogers et al., 2015, 2016). However, the third most important variable was not consistent across models. In two of five cases (TreeNet and Random Forest) the third most important variable was distance from shoreline, but in the MARS and Generalized Path Seeker models, surface curvature and waveform amplitude, respectively, were the third most important variable.

Error caused by the salt marsh vegetation on lidar returns was evident in the uncorrected dataset by comparing the vegetated field RTK GNSS measurements used in this study (n = 694, 91 GCPs were bare ground) with lidar derived elevations from the NCALM dataset (Rogers et al., 2016). Uncorrected lidar measurements exhibited a positive bias, $\mu$, of 0.24 m over the "all vegetation" ground control data (**Table 6**) and a standard deviation, $\sigma$, of 0.23 m (0.33 m RMSE). Separated by species type, most of the overall vegetation error can be attributed to just *S. alterniflora* with an observed bias of 0.35 m and standard deviation of 0.22 m (0.41 m RMSE).

The other species surveyed (*S. patens*, *D. spicata*, and *Salicornia spp.*), had a bias of between 0.05 to 0.06 m with standard deviations ranging from 0.05 to 0.08 m (0.07 - 0.10 m RMSE).

The output corrected elevations from the TreeNet and MARS models were both evaluated in a similar manner to the uncorrected lidar and the TreeNet model and exhibited an overall vegetation bias, $\mu$, of 0.00 and standard deviation, $\sigma$, of 0.07 m (0.07 m RMSE) compared to the ground control data. Also, the biases of *S. alterniflora* and the other species were reduced substantially after correction with the TreeNet model (0.01 to 0.02 m; **Table 6**). The MARS model correction produced similar results, but with a slightly larger standard deviation ($\mu$ = 0.00 m; $\sigma$ = 0.10 m), and less reduction in bias for the shorter species compared with the TreeNet model results (0.10 m RMSE).

The reason it is possible for the correction technique discussed here to reduce both the mean and standard deviation (typically associated with systematic and random uncertainty components) is that the corrections are performed on a point-by-point basis. This type of correction reduces systematic errors due to vegetation cover at each particular spot location, in contrast to methods that operate on the entire data set and can only account for a global bias, while the final model accuracy assessment is performed on the entire dataset. The frequency distribution of uncorrected residuals demonstrated a range of lidar error unique to each species surveyed (**Figure 6a**). Three of the four target species had similar residual distributions, but *S. alterniflora* was offset and had a long, asymmetric tail. A histogram of the TreeNet corrected residuals illustrates a tight grouping around 0 m with only *S. alterniflora* exhibiting small shoulders on either side (**Figure 6b**).

## 3.2 Elevation Correction Models Using Discrete-Return Lidar Predictors

Using the same algorithms as implemented with the full-waveform dataset, new model runs were conducted with only predictor variables derived from the DRL elevation data such as rugosity and slope (**Table 2**). These models, without the use of the waveform feature-based metrics, produced test sample regression coefficients ranging from 0.828 to 0.911 and regression line slopes from 0.799 to 0.913 with intercepts slightly below 0 (**Table 4**). TreeNet and Random Forest (RF) created the two most successful models with test sample $R^2$ values of approximately 0.91 and slopes within 9% and 14% of a 1:1 line, respectively compared to an uncorrected lidar $R^2$ of 0.797. The TreeNet algorithm **(Figure 5c**) had slightly more scatter on the scored dataset of all available data than the RF algorithm (**Figure 5d**). However, the TreeNet model results had a significantly better slope line and y intercept than RF. The RF results had residuals that suggested a more pronounced overestimation of bare ground (sandflats) and an underestimation of high marsh vegetation. Both models with only DRL data contained significantly more scatter and underestimation than models developed using all of the waveform predictors. Variable importance for the DRL-based models also showed uncorrected lidar elevation was most influential, with the second most important variable typically being rugosity (**Table 7**). Model variation in variable importance was illustrated in the CART model, which considered distance from shoreline as the second most important variable and rugosity the third.

The top two DRL-based models, TreeNet and Random Forest, were also evaluated on their ability to remove overall lidar bias as well as species bias (**Table 6**). The TreeNet corrected data exhibited an overall vegetation bias, $\mu$, of -0.01 m and standard deviation, $\sigma$, of 0.14 m (0.14 m RMSE), but species bias contributions varied widely (-0.05 to 0.10 m; **Table 6**). The Random Forest model correction produced a similar results with $\mu$ = -0.01 and $\sigma$ = 0.11 m (0.11 m RMSE). However, the shorter vegetation species had a tendency to be underestimated, producing negative bias of between -0.07 and -0.08 m. A TreeNet residuals histogram exhibits a symmetric grouping around 0 m with *S. alterniflora* with moderate shoulders on either side (**Figure 6c**).

### 3.3 Vegetation Classification Models

Dominant species or ground type had been collected as part of the field data for the 785 RTK GNSS locations across the three marshes that overlapped the extracted waveform data footprints. A model developed to separately classify the three major species and one genus (*S. alterniflora, S. patens, D. spicata*, and *Salicornia spp.*) did not produce useful results due to similarities in growth characteristics and waveform response that created considerable class confusion. Therefore a simplified approach was attempted, relying on zonation to classify vegetation. The zonation model employed only three classes: bare ground (GR), high marsh vegetation [*S. patens*, *Salicornia spp.*, *D. spicata*, short-form *S. alterniflora*] (HM), and low marsh vegetation [tall-form and medium-form *S. alterniflora*] (LM). Three model algorithms were evaluated and their prediction success, the ability to discriminate between the three classes, is presented in a confusion matrix (**Table 8**). The TreeNet model produced the highest success rate with an overall classification accuracy of 92% in the independent test dataset with the lowest success in the GR class. Random Forest and CART models also performed well. Variable importance of each of the three zonation models was evaluated (**Table 9**) and, as with the waveform based elevation correction models found in **Table 4**, the three most important predictors were waveform width, uncorrected lidar elevation, and distance from shoreline.

The models were scored against the complete lidar dataset for Moors marsh (525,941 grid cells) with all 13 predictor variables to create classified grids of vegetation. As a reference and for comparison, a vegetation zonation map was created using traditional aerial photo interpretation and ground-truth data (**Figure 7a**). The field map displays a system dominated by low marsh with a large central channel and several scattered high marsh regions, which are presumably topographic highs. Comparisons between maps generated by the various classification models produced similar predictions, with some performing better in high marsh and others better at discriminating between low marsh and unvegetated tidal flats (**Table 8**). The best performing model, TreeNet, produced the most accurate classification map (**Figure 7b**). Data gaps are typically water features such as salt ponds that are shown as white. The resultant grid distinctly displays the two vegetative regions. The model had some difficulty in interpreting bare ground just inside the shoreline contour and confused it with high marsh vegetation, possibly due to the dense macroalgae that was present. There were also several high marsh areas identified by the model that were not interpreted as high marsh (SF *S. alterniflora*) from either the field or aerial survey. A subsequent site visit to the marsh confirmed that these were indeed areas that should be classified as high marsh that were missed from the original aerial photo interpretation used to prepare the field zonation map.

# 4. Discussion

## 4.1 Nonparametric DEM Correction

Although a few case studies have been conducted using predictive modeling in salt marshes to determine habitat, vegetative species, and edaphic conditions (Andrew and Ustin, 2009; Griffin et al., 2011; Hladik and Alber, 2014; Sanderson et al., 2001; Sellars and Jolls, 2007), to our knowledge there are no other salt marsh studies that use predictive modeling to develop a DEM correction technique nor any that employ full-waveform lidar metrics. The predictive modeling developed here provides a viable alternative to previous methods of DEM corrections. By applying nonparametric modeling on a location-specific, point-by-point basis, our methods reduced not only the global bias, but also the standard deviation of elevation residuals when an empirical accuracy assessment for the entire data set was performed. The models developed using both full-waveform and DRL surface predictors were successful at adapting to each pixel's varying predictors, eliminating a majority of the vegetation-induced bias. The models accomplished this without *a priori* knowledge of vegetation species location and using only a single remote sensing platform. Although many of the algorithms evaluated in this study provided good results, the TreeNet algorithm consistently outperformed the others. The final model achieved an exceptional $R^2$ of 0.96 on the test dataset and 0.98 on the combined learn and test datasets, which dropped the overall bias from the uncorrected 0.24 to 0.00 m, the standard deviation, $\sigma$, from 0.23 to 0.07 m, and RMSE from 0.33 to 0.07 m. This reduction was achieved for lidar data collected at peak vegetative conditions.

The TreeNet algorithm, otherwise known as stochastic gradient boosting, was consistently the best performing algorithm used in this study. It is capable of consistently generating extremely accurate models for both regression and classification. To accomplish this, TreeNet generates thousands of small decision trees (< 6 terminal nodes), from a random sample of the data that sequentially eliminate residuals and converge on a highly accurate model (Derrig and Francis, 2008; Friedman, 2002). TreeNet has the ability to handle contaminated or missing data that can be very challenging for other data mining methods, such as neural networks, by rejecting training data points that are too much at variance with the existing model. The response variable mean square error or average negative log likelihood is successively lowered through applying numerous trees until an optimal model is achieved.

The strong results observed in this study might suggest that the model may be overfitting the data. While this is a valid consideration, it should be noted that the model algorithms used in this study, in particular TreeNet, are designed to be highly resistant to overfitting. TreeNet resists overfitting since very small trees are used instead of one large tree and therefor the models produce substantially higher accuracies (Friedman, 2002). TreeNet uses several regularization techniques to minimize overfitting such as a gradual build up the model through successive gradient boosting iterations (trees). Variables are introduced one at a time, but are only permitted to adjust the model outcome by very small coefficients (Friedman, 2002). Increasing the number of trees reduces the error on the learn dataset and the software determines the optimal tree that minimizes overfitting and error. In addition, another method of overfitting regularization employed by TreeNet consists of the subsample size, which is a constant fraction of the size of the training set. A small subsample size introduces randomness into the algorithm by forcing the regression trees to be fit to reduced datasets at each boosting iteration (Friedman, 2002). Another method of ensuring validity of the models (i.e. absence of overfitting), would be comparison of the results of multiple nonparametric algorithms. The results from the various algorithms used in this study based on very different mathematical formulas and concepts produced a cluster of similar results giving further indication that the data were not overfit. Also, data from three regional marshes were used to create the model limiting the possibility that the model results are site specific. Additionally, final model accuracy and overfitting was assessed by performing the model creation multiple times using randomly selected different learning and test datasets from the available data to verify model consistency.

The set of predictors for correcting uncertainty in salt marshes chosen here appears to be sufficient to provide discriminatory power and high predictive model accuracy. In some of the models this list could be paired back and still achieve similar results. In addition to uncorrected lidar elevation, waveform width appears to be the variable with the strongest predictive power, although several other predictors such as distance from shoreline, rugosity and waveform amplitude also played key roles in some models. Previous research has suggested a relationship between waveform width, vegetation height and lidar uncertainty (Parrish et al., 2014; Rogers et al., 2015, 2016). This relationship can be attributed to the convolution of the laser pulse with an extended target (i.e., taller vegetation results in greater spreading of the return pulse) (Rogers et al., 2015). Distance from shoreline also played a key role in the developed models. As distance increased from the shoreline (i.e., the lowest elevational extent of vegetation), vegetation height tended to decrease as well. However, this relationship may not always be the case in all marsh environments. Although variations in rugosity (surface roughness) were slight across much of the uncorrected DEM surface, there were perceptible differences between vegetation species, presumably representative of growth habits, which were used in the correction process. For example, *S. alterniflora* stands appeared to have greater rugosity than high marsh species. The predictive power of waveform amplitude was likely due to increased planimetric obscuration (i.e. vegetation coverage) with plant height, as well as the near infrared wavelength of the laser, which is preferentially reflected by healthy vegetation (Rogers et al., 2015). Not surprisingly, waveform amplitude and waveform standard deviation (a collinear variable with waveform width used in this study) were found to account for nearly 75% of the variability in vegetation height (Rogers et al., 2015).

The uncorrected lidar DEM for Moors Marsh displays highly variable elevations with undulating clusters of vegetation growth (**Figure 8a**). However, the scored results from the TreeNet full-waveform model for the same geographic area produced a vastly improved DEM, suggesting that the model performs extremely well at removing vegetation-induced uncertainty (**Figure 8b**). All high elevation clustering visible in the uncorrected DEM was removed and the underlying smooth topographic surface was revealed. Topographic highs hidden in the original DRL dataset are now plainly visible after model correction. Species-based correction methods have been found to create step like patterns in marsh DEMs when transitioning from one species polygon to another and step removal required additional smoothing algorithms that would increase DEM inaccuracy (Hladik et al., 2013). This was particularly true within the ecophenes of *S. alterniflora* (Hladik et al., 2013). A map depicting the difference between the uncorrected lidar and the full-waveform corrected DEMs confirms the extent of vegetation-based uncertainty reduction (**Figure 8d**). Although the overall DEM bias is clearly improved with species-based correction methods (Hladik et al., 2013), nonparametric modeling with full-waveform predictors improves error removal, while compensating for changing vegetation conditions on a pixel by pixel basis, resulting in more accurate DEMs.

The availability of lidar waveform data to the user community is still relatively limited. Therefore, since most researchers may not have access to or the ability to process raw waveform data at present, elevation correction of the raw salt marsh lidar DEM using only DRL data sources (i.e. no waveform model predictors) would be a valuable alternative to full-waveform based correction even if it were slightly less accurate. However, there is one waveform-based parameter that is regularly supplied with DRL systems that is helpful and can improve DRL corrections. In addition to recording return pulse time to correspond with elevation, most topographic lidar systems record the intensity, or the waveform amplitude (typically scaled to an arbitrary range of 0-255), of the return pulse. Lidar intensity typically represents the peak amplitude of the return pulse, and is a function of the reflectivity of the surface at the laser wavelength (as well as range, incidence angle, and other variables). Since waveform amplitude was found to correlate well with some salt marsh biophysical parameters (Rogers et al., 2015) and was a moderate contributor in the full-waveform model, intensity was included in the DRL based models. The lidar intensity value provided with the NCALM data delivery was uncalibrated, but since the data were collected for all three marshes with the same sensor and in one continuous flight, intensity values by ground feature type from site to site are not expected to vary significantly.

As anticipated, the DRL-based model did not produce corrected DEMs of similar quality to models created using full-waveform feature based metrics. Nevertheless, the use of the DRL data predictors and intensity did greatly improve the resulting DEM over the uncorrected lidar with an $R^2$ = 0.93 with a slope within 9% of a 1:1 line and brought the RMSE down from 0.33 to 0.14 m. These results were comparable to several other advanced correction methods and as with the waveform-based methods, the DLR-based nonparametric approach does not require *a priori* species information or other remote sensing data inputs such as multi/ hyperspectral imagery (Buffington et al., 2016; Hladik et al., 2013; Medeiros et al., 2015). The use of this type of model may be acceptable in circumstances where partial correction is better than correction accomplished by some other means or no correction at all. This is particularly the case when data acquisition does not specify recording full-waveform returns or when processing historical DRL datasets. Scored results for the full geographic area produced an improved DEM (**Figure 8c**) over the uncorrected lidar dataset (**Figure 8a**). Differences between the uncorrected and the DRL corrected DEMs suggest that the model performs reasonably well at removing vegetation-induced uncertainty (**Figure 8e**). However, a comparison of the waveform model difference map (**Figure 8d**) and the DRL model difference map (**Figure 8e**) reveals that the DRL model under-corrected elevations in areas of tallest vegetation and over-corrected in areas with the shortest vegetation (**Figure 8f**). This is particularly prevalent in areas that could be identified as SF *S. alterniflora* dominant.

### 4.2 Vegetation Classification

Salt marsh vegetation mapping is traditionally performed using field based data, aerial interpretation or classification from spectral signatures found in multi/hyperspectral imagery to show patterns in time and space as plants respond to changes in important drivers like hydrology, sea level, and sediment supply (**Figure 7**) (Kirwan et al., 2011; Konisky, 2012). A logical extension of the uncertainty correction modeling was to test its ability to map vegetation based on the strong relationships between waveform-based metrics and vegetation biophysical parameters (Rogers et al., 2015). This predictive modeling based classification method relied solely on lidar data based parameters and did not use the spectral properties typically used in vegetation classification. However, due to the similarities in biophysical characteristics between some of the vegetation found at this and other northeastern salt marshes, producing an individual species based map from lidar metrics proved difficult, as other researchers have found when classifying vegetation based on spectral characteristics (Fernandez-Nunez et al., 2017; Hladik and Alber, 2014; Hladik et al., 2013; McClure et al., 2016; Medeiros et al., 2015).

Salt marsh ecologists often refer to the vegetative zonation within the marsh system as high marsh (HM) and low marsh (LM) and these designations represent both the species present, as well as frequency of inundation, which are integrally related. High marsh vegetation species in northeastern United States typically include *S. patens*, *D. spicata*, *Salicornia spp.* and often short-form *S. alterniflora*, while the low marsh is comprised primarily of medium and tall-form

*S. alterniflora.* Using a combination of predictor variables including waveform width, rugosity, and distance from shoreline, several useful models were created that were based on our 785 sample locations with the best model having an overall classification accuracy of 92%. A three zone model (high marsh, low marsh, and bare ground) was produced using this model and interpolated for each pixel across the entire marsh.

In some cases, the model appeared to have some difficulty in interpreting bare ground just inside the shoreline contour and confused it with high marsh vegetation. It has been reported that classification of multi/hyperspectral imagery of *S. alterniflora* also has difficulty in this zone due to spectral confusion with mixed pixels that include mud: "the *Spartina* problem" (Hladik et al., 2013). However, the cause in this case is likely in part be due to the presence of large mats of macroalgae on rocks (*Ascophyllum nodosum var. scorpioides* and *Fucus vesiculosus var. spiralis*). Macroalgae was not evaluated in this study, but are commonly is found in the intertidal zone and might produce a similar biologic induced waveform response to that of high marsh vegetation based on some of its biophysical characteristics such as its short height. Further testing is needed to corroborate this observation.

The vegetation maps created in this study have been derived solely from lidar data and without the use of any spectra derived from aerial photography or multi/hyperspectral imagery. There is little if any spectral difference between the three ecophenes of *S. alterniflora* (Artigas and Yang, 2005; Schmidt and Skidmore, 2003), and using traditional remote sensing classification methods often results in considerable confusion among the classes. Overall classification accuracies from other studies using spectral signatures or hybrid approaches of lidar and hyperspectral imagery ranged from 59% to >90% (Hladik et al., 2013; Rosso et al., 2006; Wang et al., 2007). That the nonparametric modeling of the full-waveform metrics could achieve similar or better classification results without the use of spectra is significant. The classification based on lidar modeling appears a viable alternative to differentiate salt marsh vegetation into identifiable regions or classes.

The results of this study are consistent with the CART nonparametric vegetation classification models conducted by Hladik and Alber (2014) that do not use full-waveform metrics as predictor variables. Vegetation zonation mapping is commonly used by salt marsh scientists to investigate marsh habitat and monitor changes in the marsh over time due to tidal restrictions, restored flow after a restoration project, storm assessment, or the potential impacts or monitoring of SLR. In future studies, salt marsh mapping using full-waveform lidar and nonparametric, predictive modeling could be automated and provide standardized results with minimal human input or interpretation, which may allow for rapid, unbiased assessments of vegetation zones. Although more research is needed to assess its full capabilities, this new vegetation classification method may also prove to be more efficient and/or more accurate than some of the traditional methods currently being employed. Another possible future research direction for research could be to add spectral values from the various bands of multi/hyperspectral imagery as predictor variables to the waveform model to produce potentially highly accurate vegetation classification maps. The combination of precise elevation (+/- 2 cm) and vegetation maps from full-waveform LIDAR [or: 'remote sensing'] would allow ecologists and managers to track salt marsh responses to restoration and other management actions as well as observe and predict responses to rising sea levels, such as plant community migration and loss by drowning.

## 5. Conclusions

The utility of salt marsh DEMs based on lidar is weakened by vegetation-induced uncertainty, which continues to challenge researchers and coastal managers who desire to use high resolution lidar datasets for regional or site-specific analysis. Without a satisfactory correction method, lidar-based DEM models are often unsuitable for restoration planning, hydrologic modeling, storm impact analysis, SLR adaptability studies or other applications where fine topographic details are necessary. The main conclusions drawn from this research are: 1) nonparametric predictive modeling techniques, coupled with full-waveform shape-based metrics, provide a powerful tool to reduce elevation uncertainty due to salt marsh vegetation, even during peak vegetation growth conditions. The highest performing model produced an $R^2$ of 0.98, a slope within 4% of a 1:1 line, reduced bias, $\mu$, from 0.24 m to 0.00 m, and standard deviation, $\sigma$, from 0.23 to 0.07 m (0.33 to 0.07 m RMSE); 2) in addition to DRL uncorrected lidar elevation, waveform width was determined to be the most significant predictor variable in nearly all models that used waveform feature-based metrics; 3) moderately successful models can be built from predictors based solely on DRL sources (with intensity), which may provide adequate correction when full-waveform lidar is not available. The best models resulted in an $R^2$ of 0.92, slopes within 9% of 1:1, reduced bias to -0.01 m, and standard deviation to 0.14 m (0.14 m RMSE); and 4) accurate salt marsh zone classification maps (overall classification accuracy >90%) can be created using only a lidar data source and without multi/hyperspectral imagery.

The coupling of nonparametric modeling tools and GIS has become standard practice in many different environmental fields such as land use, geomorphology, soil science, and wildlife habitat (Bourennane et al., 2014; Gutierrez et al., 2009; Meissner et al., 2014; Tayyebi and Pijanowski, 2014; Timm and McGarigal, 2012). Full-waveform lidar combined with predictive modeling tools appears to deliver highly accurate salt marsh elevation models and vegetation maps by reducing vegetation-induced lidar uncertainty. The developed model was able to reduce both systematic and random error as computed for the entire data set by applying location-specific, point-by-point corrections obtained via the nonparametric regression methods. The explanation for the ability to reduce both $\mu$ and $\sigma$ is that some of what is computed as the "random error" of the full dataset is, in fact, due to vegetation-induced systematic error that exists at the individual point level, and these model corrections are applied on a point-by-point basis. Corrected elevation surfaces will be tremendously useful to support coastal research and management objectives, while also minimizing the amount of expensive, time-consuming field work. The ability to properly correct salt marsh DEMs should allow the creation of better inundation models such as SLAMM (Sea- Level Affecting Marshes Model) (Chu-Agor et al., 2011) and the detailed assessment of the impacts of sea level rise on marsh health and resilience. Corrected DEMs should also help to plan and monitor the results of salt marsh restoration projects. The five nonparametric models created in this study employed different algorithms to reduce elevation uncertainty, yet provided a relatively narrow range of results. The use of multiple algorithms producing similar results provides further validation of a successful outcome despite the complex variable relationships and interactions.

It is important to note that, since the data for this study were acquired in 2010, a number of important advancements have been made in airborne lidar technology. These include Geiger-mode (Abdullah, 2016) and single-photon lidar (Stoker et al., 2016), short pulse width systems (Wright et al., 2016), and multi-wavelength lidar systems (Morsy et al., 2016). The Geiger-mode and single-photon technologies offer the potential for higher data densities and higher flight altitudes, but cannot provide waveforms (although some authors have developed techniques to aggregate returns to create something akin to a waveform). For the short-pulse width systems, it is presently unclear to what extent the methods developed in this study will work (or whether they are even necessary, since the ability to resolve multiple returns—including the ground return —in dense marsh vegetation may improve). Multi-wavelength (e.g., 1550, 1064, and 532 nm), waveform-resolving lidar systems appear well suited for extending the work presented here, but future research is needed to investigate this. Additional topics recommended for future work include: a) assessing whether models created in this study can be successfully scored against full-waveform data from other northeastern salt marshes without substantially modifying the developed model; b) extending this type of analysis to marshes in different regions of the country with differing vegetation species; c) analyzing full-waveform data taken from marsh systems in winter (senescent conditions) to determine if this technique is adaptable to data collected at different times of the year and perhaps lowering RMSE further.

**Acknowledgements**

We would like to acknowledge the following organizations and people for their contributions and funding of this research. National Center for Airborne Laser Mapping (NCALM) for providing a seed data grant that flew the 2010 lidar coverage for Cape field areas. Michael Sartori and Juan Fernandez of NCALM provided information related to the waveform acquisition. Funding for Larry G. Ward was provided by UNH/NOAA Joint Hydrographic Center Award NA10NOS4000073. Mark Adams from the Cape Cod National Seashore and Dr. Mark Borrelli from the Provincetown Center for Coastal Studies provided RTK GNSS equipment for field data collection and software for data processing. We would also like to express gratitude to Dr. Mary Martin, Dr. Susan Adamowicz, and Dr. Graham Giese for their input and thoughtful reviews that helped improve this manuscript. This research was conducted

as part of the requirements for completion of a doctoral degree at the University of New Hampshire.

## References


Abdullah, Q.A., 2016. A star is born: The state of new lidar technologies. Photogrammetric Engineering and Remote Sensing 82, 307-312.

Adams, T., Beets, P., Parrish, C., 2012. Extracting More Data from LiDAR in Forested Areas by Analyzing Waveform Shape. Remote Sensing 4, 682-702.

Anderson, C.M., Treshow, M., 1980. A Review of Environmental and Genetic Factors That Affect Height in *Spartina alterniflora* Loisel (Salt-Marsh Cord Grass). Estuaries 3, 168-176.

Anderson, J.E., Plourde, L.C., Martin, M.E., Braswell, B.H., Smith, M.L., Dubayah, R.O., Hofton, M.A., Blair, J.B., 2008. Integrating waveform lidar with hyperspectral imagery for inventory of a northern temperate forest. Remote Sensing of Environment 112, 1856-1870.

Andrew, M.E., Ustin, S.L., 2009. Habitat suitability modelling of an invasive plant with advanced remote sensing data. Diversity and Distributions 15, 627-640.

Artigas, F.J., Yang, J.S., 2005. Hyperspectral remote sensing of marsh species and plant vigour gradient in the New Jersey Meadowlands. International Journal of Remote Sensing 26, 5209-5220.

Bertness, M.D., Ellison, A.M., 1987. Determinants of Pattern in a New England Salt Marsh Plant Community. Ecological Monographs 57, 129-147.

Bourennane, H., Couturier, A., Pasquier, C., Chartin, C., Hinschberger, F., Macaire, J.-J., Salvador-Blanes, S., 2014. Comparative performance of classification algorithms for the development of models of spatial distribution of landscape structures. Geoderma 219–220, 136-144.

Breiman, L., 2001. Random forests. Machine Learning 45, 5-32.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth, Belmont, CA.

Bretar, F., Chauve, A., Bailly, J.S., Mallet, C., Jacome, A., 2009. Terrain surfaces and 3-D landcover classification from small footprint full-waveform lidar data: application to badlands. Hydrology and Earth System Sciences 13, 1531-1544.

Brock, J., Sallenger, A., 2001. Airborne topographic LIDAR mapping for coastal and resource management. U.S. Geological Survey Open File Report, 01-46, p. 4.

Buffington, K.J., Dugger, B.D., Thorne, K.M., Takekawa, J.Y., 2016. Statistical correction of lidar-derived digital elevation models with multispectral airborne imagery in tidal marshes. Remote Sensing of Environment 186, 616-625.

Burrough, P.A., McDonell, R.A., 1998. Principles of Geographical Information Systems. Oxford University Press, New York.

Byrd, K.B., Kelly, M., 2006. Salt marsh vegetation response to edaphic and topographic changes from upland sedimentation in a pacific estuary. Wetlands 26, 813-829.

Chaplot, V., Darboux, F., Bourennane, H., Leguedois, S., Silvera, N., Phachomphon, K., 2006. Accuracy of interpolation techniques for the derivation of digital elevation models in relation to landform types and data density. Geomorphology 77, 126-141.

Chu-Agor, M.L., Munoz-Carpena, R., Kiker, G., Emanuelsson, A., Linkov, I., 2011. Exploring vulnerability of coastal habitats to sea level rise through global sensitivity and uncertainty analyses. Environmental Modelling & Software 26, 593-604.

Cooper, H.M., Fletcher, C.H., Chen, Q., Barbee, M.M., 2013. Sea-level rise vulnerability mapping for adaptation decisions using LiDAR DEMs. Progress in Physical Geography 37, 745-766.

Costanza, R., D'Arge, R., DeGroot, R., Farber, S., Grasso, M., 1997. The value of the world's ecosystem services and natural capital. Nature 387, 253-280.

Crooks, S., Schutten, J., Sheern, G.D., Pye, K., Davy, A.J., 2002. Drainage and elevation as factors in the restoration of salt marsh in Britain. Restoration Ecology 10, 591-602.

Derrig, R., Francis, L., 2008. Distinguishing the forest from the TREES: a comparison of tree-based data mining methods. Variance 2, 184-208.

Drake, J.B., Dubayah, R.O., Clark, D.B., Knox, R.G., Blair, J.B., Hofton, M.A., Chazdon, R.L., Weishampel, J.F., Prince, S.D., 2002. Estimation of tropical forest structural characteristics using large-footprint lidar. Remote Sensing of Environment 79, 305-319.

Fagherazzi, S., Carniello, L., D'Alpaos, L., Defina, A., 2006. Critical bifurcation of shallow microtidal landforms in tidal flats and salt marshes. Proceedings of the National Academy of Sciences of the United States of America 103, 8337-8341.

Fernandez-Nunez, M., Burningham, H., Ojeda Zujar, J., 2017. Improving accuracy of LiDAR-derived digital terrain models for saltmarsh management. Journal of Coastal Conservation (Springer Science & Business Media B.V.) 21, 209-222.

Friedman, J.H., 1991. Multivariate Adaptive Regression Splines. Annals of Statistics 19, 1-67.
Friedman, J.H., 2002. Stochastic gradient boosting. Computational Statistics & Data Analysis 38, 367-378.

Friedman, J.H., 2012. Fast sparse regression and classification. International Journal of Forecasting 28, 722-738.

Gopfert, J., Heipke, C., 2006. Assessment of Lidar DTM in coastal vegetated areas. International Archives of Photogrammetry, Remote Sensing, and Spatial Information Sciences 36, 79-85.
Graham, D., Sault, M., Bailey, C.J., 2003. National Ocean Service shoreline - Past, present, and future. Journal of Coastal Research, 14-32.

Griffin, P.J., Theodose, T., Dionne, M., 2011. Landscape Patterns of Forb Pannes Across a Northern New England Salt Marsh. Wetlands 31, 25-33.

Gutierrez, A.G., Schnabel, S., Contador, J.F.L., 2009. Using and comparing two nonparametric methods (CART and MARS) to model the potential distribution of gullies. Ecological Modelling 220, 3630-3637.

Hladik, C., Alber, M., 2012. Accuracy assessment and correction of a LIDAR-derived salt marsh digital elevation model. Remote Sensing of Environment 121, 224-235.

Hladik, C., Alber, M., 2014. Classification of salt marsh vegetation using edaphic and remote sensing-derived variables. Estuarine Coastal and Shelf Science 141, 47-57.

Hladik, C., Schalles, J., Alber, M., 2013. Salt marsh elevation and habitat mapping using hyperspectral and LIDAR data. Remote Sensing of Environment 139, 318-330.

Jutzi, B., Stilla, U., 2006. Range determination with waveform recording laser systems using a Wiener Filter. ISPRS Journal of Photogrammetry and Remote Sensing 61, 95-107.

Kirwan, M.L., Murray, A.B., Donnelly, J.P., Corbett, D.R., 2011. Rapid wetland expansion during European settlement and its implication for marsh survival under modern sediment delivery rates. Geology 39, 507-510.

Konisky, R., 2012. Role of Simulation Models in Understanding the Salt Marsh Restoration Process, in: T., R.C., Burdick, D. (Eds.), Tidal Marsh Restoration: A Synthesis of Science and Practice. Island Press, Washington, pp. 253-276.

Lefor, M.W., Kennard, W.C., Civco, D.L., 1987. Relationship of salt marsh plant distributions to tidal levels in Connecticut, USA. Environmental Management 11, 61-68.

Lefsky, M.A., Cohen, W.B., Parker, G.G., Harding, D.J., 2002. Lidar remote sensing for ecosystem studies. BioScience 52, 19-30.

Mallet, C., Bretar, F., 2009. Full-waveform topographic lidar: State-of-the-art. ISPRS Journal of Photogrammetry and Remote Sensing 64, 1-16.

McClure, A., Liu, X., Hines, E., Ferner, M.C., 2016. Evaluation of Error Reduction Techniques on a LIDAR-Derived Salt Marsh Digital Elevation Model. Journal of Coastal Research 32, 424-433.

Mckee, K., Patrick, J.W.H., 1988. The Relationship of Smooth Cordgrass (*Spartina alterniflora*) to Tidal Datums: A Review. Estuaries 11, 143-151.

Medeiros, S., Hagen, S., Weishampel, J., Angelo, J., 2015. Adjusting Lidar-Derived Digital Terrain Models in Coastal Marshes Based on Estimated Aboveground Biomass Density. Remote Sensing 7, 3507-3525.

Meissner, K., Fiorentino, D., Schnurr, S., Arbizu, P.M., Huettmann, F., Holst, S., Brix, S., Svavarsson, J., 2014. Distribution of benthic marine invertebrates at northern latitudes - An evaluation applying multi-algorithm species distribution models. Journal of Sea Research 85, 241-254.

Mendelssohn, I.A., McKee, K.L., Patrick, W.H., 1981. Oxygen Deficiency in Spartina alterniflora Roots - Metabolic Adaptation to Anoxia. Science 214, 439-441.

Mitsch, W.J., Gosselink, J.G., 2000. Wetlands, third ed. John Wiley & Sons, New Jersey.

Montane, J.M., Torres, R., 2006. Accuracy assessment of lidar saltmarsh topographic data using RTK GPS. Photogrammetric Engineering and Remote Sensing 72, 961-967.

Morris, J.T., Porter, D., Neet, M., Noble, P., Schmidt, L., Lapine, L.A., Jensen, J.R., 2005. Integrating Lidar elevation data, multispectral imagery and neural network modeling for marsh characterization. International Journal of Remote Sensing 26, 5221-5234.

Morris, J.T., Sundareshwar, P.V., Nietch, C.T., Kjerfve, B., Cahoon, D.R., 2002. Responses of coastal wetlands to rising sea level. Ecology 83, 2869-2877.

Morsy, S., Shaker, A., El-Rabbany, A., LaRocque, P.E., 2016. AIRBORNE Multispectral Lidar Data for Land-Cover Classification and Land/Water Mapping Using Different Spectral Indexes. ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences 3.

Muss, J.D., Aguilar-Amuchastegui, N., Mladenoff, D.J., Henebry, G.M., 2013. Analysis of Waveform Lidar Data Using Shape-Based Metrics. IEEE Geoscience and Remote Sensing Letters 10, 106-110.

NOAA, 2013. Tides and Currents. NOAA Center for Operational Oceanographic Products and Services, http://tidesandcurrents.noaa.gov/station_info.shtml?stn=8446121.

NOAA, 2016. NOAA Continually Updated Shoreline Product (CUSP). National Oceanic and Atmospheric Administration, https://shoreline.noaa.gov/data/datasheets/cusp.html.

NOAA, 2017a. Benthic Terrain Modeler. NOAA Office for Coastal Management, https://coast.noaa.gov/digitalcoast/tools/btm.

NOAA, 2017b. Estimation of Vertical Uncertainties in VDatum. NOAA National Ocean Service, https://vdatum.noaa.gov/docs/est_uncertainties.html.

Ornes, W.H., Kaplan, D.I., 1989. Macronutrient status of tall and short forms of Spartina alterniflora in a South Carolina salt marsh Marine Ecology-Progress Series 55, 63-72.

Parrish, C.E., Jeong, I., Nowak, R.D., Smith, R.B., 2011. Empirical Comparison of Full-Waveform Lidar Algorithms: Range Extraction and Discrimination Performance. Photogrammetric Engineering and Remote Sensing 77, 825-838.

Parrish, C.E., Rogers, J.N., Calder, B.R., 2014. Assessment of Waveform Features for Lidar Uncertainty Modeling in a Coastal Salt Marsh Environment. IEEE Geoscience and Remote Sensing Letters 11, 569-573.

Pennings, S.C., Bertness, M.D., 2001. Salt Marsh Communities, in: Bertness, M.D., Gaines, S.D., Hay, M.E. (Eds.), Marine Community Ecology. Sinauer Associates Inc., Sunderland, MA, pp. 289-316.

Populus, J., Barreau, G., Faxilleau, J., Kerdreux, M., L'Yavanc, J., 2001. Assessment of the Lidar topographic technique over a coastal area, Proceedings of CoastGIS'01: 4th International Symposium on GIS and Computer Mapping for Coastal Zone Management, Halifax, Nova Scotia, p. 11.

Portnoy, J.W., Roman, C.T., Smith, S.M., Gwilliam, E., 2003. Estuarine habitat restoration at Cape Cod National Seashore: the Hatches Harbor prototype. Park Science 22, 8.

Reimold, R.J., Gallagher, J.l., Thompson, D.E., 1973. Remote Sensing of Tidal Marsh. Photogrammetric Engineering and Remote Sensing 39, 477-488.

Ries, L., 1993. Areas of influence for IDW-interpolation with isotropic evnironmental data. Catena 20, 199-205.

Rogers, J.N., Parrish, C.E., Ward, L.G., Burdick, D.M., 2015. Evaluation of field-measured vertical obscuration and full waveform lidar to assess salt marsh vegetation biophysical parameters. Remote Sensing of Environment 156, 264-275.

Rogers, J.N., Parrish, C.E., Ward, L.G., Burdick, D.M., 2016. Assessment of Elevation Uncertainty in Salt Marsh Environments Using Discrete-Return and Full-Waveform Lidar. Journal of Coastal Research Special Issue: 76, 107-122.

Roman, C.T., Burdick, D., 2012. Tidal Marsh Restoration: A Synthesis of Science and Practice. Island Press, Washington.

Rosso, P.H., Ustin, S.L., Hastings, A., 2006. Use of lidar to study changes associated with *Spartina* invasion in San Francisco Bay marshes. Remote Sensing of Environment 100, 295-306.

Sadro, S., Gastil-Buhl, M., Melack, J., 2007. Characterizing patterns of plant distribution in a southern California salt marsh using remotely sensed topographic and hyperspectral data and local tidal fluctuations. Remote Sensing of Environment 110, 226-239.

Sanderson, E.W., Foin, T.C., Ustin, S.L., 2001. A simple empirical model of salt marsh plant spatial distributions with respect to a tidal channel network. Ecological Modelling 139, 293-307.

Sappington, J.M., Longshore, K.M., Thompson, D.B., 2007. Quantifying landscape ruggedness for animal habitat analysis: A case study using bighorn sheep in the Mojave Desert. Journal of Wildlife Management 71, 1419-1426.

Schmid, K.A., Hadley, B.C., Wijekoon, N., 2011. Vertical Accuracy and Use of Topographic LIDAR Data in Coastal Marshes. Journal of Coastal Research 27, 116-132.

Schmidt, K.S., Skidmore, A.K., 2003. Spectral discrimination of vegetation types in a coastal wetland. Remote Sensing of Environment 85, 92-108.

Sellars, J.D., Jolls, C.L., 2007. Habitat modeling for Amaranthus pumilus: An application of light detection and ranging (LIDAR) data. Journal of Coastal Research 23, 1193-1202.

Stewart, J.P., Hu, J., Kayen, R.E., Lembo Jr, A.J., Collins, B.D., Davis, C.A., O'Rourke, T.D., 2009. Use of airborne and terrestrial lidar to detect ground displacement hazards to water systems. Journal of Surveying Engineering 135, 113-124.

Stoker, J.M., Abdullah, Q.A., Nayegandhi, A., Winehouse, J., 2016. Evaluation of single photon and Geiger mode Lidar for the 3D Elevation Program. Remote Sensing 8, 767.

Tayyebi, A., Pijanowski, B.C., 2014. Modeling multiple land use changes using ANN, CART and MARS: Comparing tradeoffs in goodness of fit and explanatory power of data mining tools. International Journal of Applied Earth Observation and Geoinformation 28, 102-116.

Timm, B.C., McGarigal, K., 2012. Fine-scale remotely-sensed cover mapping of coastal dune and salt marsh ecosystems at Cape Cod National Seashore using Random Forests. Remote Sensing of Environment 127, 106-117.

Torres, R., Styles, R., 2007. Effects of topographic structure on salt marsh currents. Journal of Geophysical Research-Earth Surface 112, 14.

Wagner, W., Hollaus, M., Briese, C., Ducic, V., 2008. 3D vegetation mapping using small-footprint full-waveform airborne laser scanners. International Journal of Remote Sensing 29, 1433-1452.

Wang, C., Menenti, M., Stoll, M.-P., Belluco, E., Marani, M., 2007. Mapping mixed vegetation communities in salt marshes using airborne spectral data. Remote Sensing of Environment 107, 559-570.

Wang, C., Menenti, M., Stoll, M.P., Feola, A., Belluco, E., Marani, M., 2009. Separation of Ground and Low Vegetation Signatures in LiDAR Measurements of Salt-Marsh Environments. IEEE Transactions on Geoscience and Remote Sensing 47, 2014-2023.

White, S.A., Parrish, C.E., Calder, B.R., Pe'eri, S., Rzhanov, Y., 2011. LIDAR-Derived National Shoreline: Empirical and Stochastic Uncertainty Analyses. Journal of Coastal Research, 62-74.

Wiegert, R.G., Freeman, B.J., 1990. Tidal salt marshes of the southeast Atlantic Coast : a community profile. U.S. Dept. of the Interior, Fish and Wildlife Service, Washington, D.C.

Wright, C.W., Kranenburg, C., Battista, T.A., Parrish, C., 2016. Depth Calibration and Validation of the Experimental Advanced Airborne Research LiDAR, EAARL-B. Journal of Coastal Research SI 76, 4-17.

Yang, Z., Myers, E., Jeong, I., White, S., 2013. VDatum for the Gulf of Maine: Tidal Datums and Topography of the Sea Surface. NOAA Technical Memorandum NOS CS 31.

Zedler, J.B., Callaway, J.C., Desmond, J.S., Vivian-Smith, G., Williams, G.D., Sullivan, G., 1999. Californian salt-marsh vegetation: An improved model of spatial pattern. Ecosystems 2, 19-35.

Zevenbergen, L.W., Thorne, C.R., 1987. Quantitative Analysis of Land Surface Topography. Earth Surface Processes and Landforms 12, 47-56.

Zhang, C.S., Liu, C.C., Zhang, X.L., Almpanidis, G., 2017. An up-to-date comparison of state-of-the-art classification algorithms. Expert Systems with Applications 82, 128-150.

# Figure Captions and Tables

**Figure 1 (a) Histogram of vegetation height for each of the surveyed species. (b) Frequency of occurrence by elevation range (MHW) for each vegetation species (n = 2,899). (SPAL - *Spartina alterniflora*, DISP - *Distichlis spicata*, SPPA - *Spartina patens*, SASP - *Salicornia spp.*)**

**Figure 2: Scatterplot of vegetation height and terrain elevation (MHW) at each RTK GNSS location (n = 2,899). Open circles are *Spartina alterniflora* and closed circles are all other species (*Spartina patens*, *Distichilis spicata*, and *Salicornia spp.*).**

**Figure 3: Site locus map. Insets are 1) Moors marsh, 2) Pamet marsh, and 3) Great Island – middle marsh. RTK GNSS points are color coded by dominant vegetation species/ground type. Red boxes are the extent of Full-waveform data used in the model creation analysis.**

**Figure 4: Pamet Marsh – Vegetation showing (left to right) *Spartina alterniflora*, *Salicornia spp.*, and *Spartina patens* zonation along a man-made dike.**

**Table 1: Flight parameters of NCALM July 20th, 2010.**

| Flight Parameter | Value |
|---|---|
| Flying Speed (m/sec) | 60 |
| Altitude (m) | 600 |
| Swath Overlap (%) | 50 |
| Laser Beam Divergence (mrad) | 0.25 |
| Pulse Rate Frequency (kHz) | 70 |
| Transmit Pulse Width (ns) | 12 |
| Scan Rate (kHz) | 40 |
| Scan Angle (degrees) | ± 21 |
| Point Return Density (pts/$m^2$) | 5 |
| Laser Footprint Diameter (m) | 0.15 |

**Table 2: Waveform metrics and surface characteristics available to the model predictor variables.**

| Waveform and Surface Metrics | | | |
|---|---|---|---|
| **Source** | **Symbol** | **Metric Name** | **Description** |
| **Full-waveform** | $A$ | Waveform amplitude | Maximum of received echo (i.e., peak value) |
| | $AUC$ | Area under curve | Trapezoidal numerical integration of echo |
| | $\mu_{\omega}$ | Waveform mean | A measure of the “center” of the return pulse |
| | $g_1$ | Waveform skewness | A measure of the asymmetry of the return pulse; positive for our waveforms, which are right skewed |
| | $w$ | Waveform width | Width (FWHM) of return pulse |
| **Derived from Discrete Lidar** | $\gamma$ | Curve | The curvature of a surface is the fourth-order polynomial calculated on a cell-by-cell basis. |
| | $\gamma_{pl}$ | Curve Plan | This is the curvature of the surface in the direction perpendicular to slope |
| | $\gamma_{pr}$ | Curve Profile | This is the curvature of the surface in the direction of slope |
| | $d$ | Distance | Distance (m) from the -1 mean high water (MHW) contour line (or lowest extent of vegetation). Positive values for shoreward and negative values for seaward distances. |
| | $Z$ | Elevation | Lidar elevation as derived from the discrete-return dataset using a 1 x 1 m cell size and inverse distance weighting interpolation method. |
| | $i$ | Intensity | Lidar intensity is the magnitude, of the return pulse. It represents the reflectivity of the surface at the laser wavelength scaled between 0-255. |
| | R | Rugosity | Measure of terrain variation of grid cells within a neighborhood in three-dimensions. Output raster values range from 0 (no terrain variation) to 1 (complete terrain variation). |
| | $m$ | Slope | Slope is the maximum rate of change in value from each cell to its neighbors calculated as a percent. |

**Table 3: Regression and classification models used with their descriptions, benefits and detriments.**

| Model | Description | Pros | Cons | References |
|---|---|---|---|---|
| Classification and Regression Trees (CART) | Creates classification trees using binary recursive partitioning to predict the group association based on one or more predictor variables. | Ability to handle missing data; Can often reveal important data relationships that sometimes remain concealed using other analytical methods | Regression based models are limited in the output response to data clustering based on the terminal node assignment | (Breiman et al., 1984) |
| Multivariate Adaptive Regression Splines (MARS) | Approximates functions by capturing essential nonlinearities and interactions but still produces results in a form similar to a traditional regression | Predicts continuous numeric outcome; Uncovers important data patterns; Produces output equations similar to those used in traditional regression approaches. | Not capable of categorical classifications | (Friedman, 1991) |
| TreeNet - Stochastic Gradient Boosting | Generates thousands of small decision trees, less than 6 terminal nodes, from a random sample of the data that sequentially eliminate residuals and converge on a highly accurate model | Highly resistant to over fitting of the data since very small trees are used instead of one large tree and the models produce substantially higher accuracies | Does not produce equation style regression output; lacks interpretable decision trees as are found with CART | (Friedman, 2002) |
| Random Forests | Random Forests is an ensemble of many CART trees that are not influenced by each other | Ability to spot outliers/anomalies; Discovering data patterns; Identifying important predictors; Predict future outcomes. | Produces somewhat more accurate classification models than regression | (Breiman, 2001) |
| Generalized Path Seeker Model (GPSM) | A forward stepping model that builds linear regressions that are additive with predictors and cannot discover on its own nonlinear relationships or interactions without the help of an analyst. | Well suited to using more predictor columns than observation records; Can handle highly correlated predictors (colinearity); Finds a compact model with good performance | Does not handle missing values and will enforce row deletions to compensate for missing predictor values. | (Friedman, 2012) |

**Table 4: Model results from full-waveform and discrete-return lidar based models. The "learn" sample was used to build the model while the "test" sample is independent and used for confirming model results. The scored data column is the results of the model on the combined learn and test samples. The regression line equation for the scored model is displayed to give an indication of how close to a 1:1 relationship the model created. A perfect model would have an $R^2$ value of 1 and an equation of y = x. Models results are sorted in order by performance (best to worst), which is determined using three criteria: a high independent "test" sample $R^2$ result, similarity of $R^2$ results between the "learn" and "test" results, and closeness of fit of the final regression equation line to a 1:1 correlation. * (The learn sample $R^2$ for Random Forest [RF] models, otherwise known as "OOB" [out-of-bag], is always 1 and therefore not reported.)**

| Type | Models | Learn (n = 560) | Test (n = 225) | Scored (n = 785) | Equation |
|---|---|---|---|---|---|
| Waveform | TreeNET | 0.990 | 0.963 | 0.982 | y = 0.9748x - 0.0103 |
| | MARS | 0.967 | 0.960 | 0.964 | y = 0.9642x - 0.0169 |
| | GPSM | 0.934 | 0.948 | 0.938 | y = 0.9329x - 0.0327 |
| | Regression | 0.934 | 0.947 | 0.938 | y = 0.9326x - 0.0327 |
| | RF | * | 0.959 | 0.984 | y = 0.8971x - 0.0488 |
| | CART | 0.939 | 0.919 | 0.934 | y = 0.9964x - 0.0009 |
| Discrete | TreeNET | 0.934 | 0.910 | 0.926 | y = 0.9126x - 0.0388 |
| | RF | * | 0.911 | 0.959 | y = 0.8652x - 0.0649 |
| | MARS | 0.857 | 0.872 | 0.862 | y = 0.8567x - 0.0720 |
| | CART | 0.917 | 0.880 | 0.905 | y = 0.9139x - 0.0407 |
| | GPSM | 0.817 | 0.832 | 0.827 | y = 0.7992x - 0.0990 |
| | Regression | 0.820 | 0.828 | 0.823 | y = 0.8201x - 0.0872 |

**Figure 5: a) Plot of RTK GNSS elevations to raw lidar elevation (closed circles) and the same lidar points corrected with the TreeNet model (open circles) using full-waveform and discrete-return lidar data. b) Plot of RTK GNSS elevations to raw lidar elevation (closed circles) and the same lidar points corrected with the MARS model (open circles) using full-waveform and discrete-return lidar data. c) Plot of RTK GNSS elevations to raw lidar elevation (closed circles) and the same lidar points corrected with the TreeNET model (open circles) using only discrete-return lidar data sources. d) Plot of RTK GNSS elevations to raw lidar elevation (closed circles) and the same lidar points corrected with the Random Forest model (open circles) using only discrete-return lidar data sources. All elevations are in local mean high water (MHW) tidal datum.**

**Table 5: Variable importance is presented for each of the models that use full-waveform and discrete-return lidar data predictors. The most important variable is given a score of 100 and all other variables importance are rescaled relative to the most important variable. The top 3 important variable from each model run are highlighted in bold.**

| Symbol | Predictor Variable | TreeNet | MARS | GPSM | RF | CART |
|---|---|---|---|---|---|---|
| $A$ | Amplitude | 9.05 | 3.16 | **8.54** | 0.24 | 2.71 |
| $AUC$ | Area under curve | 7.71 | - | 1.8 | 0.21 | 1.96 |
| $\mu_\omega$ | Waveform mean | 9.85 | 3.02 | - | 0.07 | 10.7 |
| $g_1$ | Waveform skewness | 7.77 | 4.15 | 2.19 | 0.07 | 4.58 |
| $w$ | Width | **52.11** | **42.62** | **39.16** | **24.11** | **30.26** |
| $Z$ | Elevation | **100** | **100** | **100** | **100** | **100** |
| $\gamma$ | Curve | 6.58 | **7** | 4.91 | 0.05 | 6.13 |
| $\gamma_{pl}$ | Curve Plan | 7.55 | - | - | 0.02 | 2.5 |
| $\gamma_{pr}$ | Curve Profile | 7.32 | - | 6.21 | 0.08 | 3.28 |
| $d$ | Distance | **16.77** | 2.86 | 1.27 | **3.22** | **65.95** |
| $R$ | Rugosity | 8.49 | 5.21 | - | - | 14.14 |
| $m$ | Slope | 7.92 | 3.83 | 4.08 | - | 4.89 |

**Table 6: Residuals by species for uncorrected lidar and top two models for both full-waveform and discrete-return lidar model results**

| Model | Species | N | Mean | Min | Max | SD | RMSE |
|---|---|---|---|---|---|---|---|
| **Uncorrected Lidar** | All Vegetation | 694 | 0.24 | -0.20 | 1.11 | 0.23 | 0.33 |
| | *S. alterniflora* | 446 | 0.35 | -0.20 | 1.11 | 0.22 | 0.41 |
| | *S. patens* | 123 | 0.06 | -0.17 | 0.18 | 0.05 | 0.08 |
| | *Distichlis spicata* | 39 | 0.05 | -0.07 | 0.11 | 0.05 | 0.07 |
| | *Salicornia spp.* | 86 | 0.06 | -0.12 | 0.32 | 0.08 | 0.10 |
| | | | | | | | |
| **TreeNET Waveform** | All Vegetation | 694 | 0.00 | -0.43 | 0.29 | 0.07 | 0.07 |
| | *S. alterniflora* | 446 | -0.01 | -0.43 | 0.27 | 0.08 | 0.08 |
| | *S. patens* | 123 | 0.01 | -0.08 | 0.16 | 0.04 | 0.04 |
| | *Distichlis spicata* | 39 | 0.02 | -0.05 | 0.14 | 0.04 | 0.04 |
| | *Salicornia spp.* | 86 | 0.02 | -0.10 | 0.29 | 0.06 | 0.06 |
| | | | | | | | |
| **MARS Waveform** | All Vegetation | 694 | 0.00 | -0.42 | 0.49 | 0.10 | 0.10 |
| | *S. alterniflora* | 446 | -0.02 | -0.42 | 0.43 | 0.11 | 0.11 |
| | *S. patens* | 123 | 0.01 | -0.10 | 0.26 | 0.07 | 0.07 |
| | *Distichlis spicata* | 39 | 0.03 | -0.06 | 0.16 | 0.06 | 0.06 |
| | *Salicornia spp.* | 86 | 0.05 | -0.11 | 0.49 | 0.09 | 0.10 |
| | | | | | | | |
| **TreeNET Discrete** | All Vegetation | 694 | -0.01 | -0.72 | 0.57 | 0.14 | 0.14 |
| | *S. alterniflora* | 446 | -0.05 | -0.72 | 0.37 | 0.14 | 0.15 |
| | *S. patens* | 123 | 0.04 | -0.14 | 0.48 | 0.10 | 0.11 |
| | *Distichlis spicata* | 39 | 0.04 | -0.11 | 0.57 | 0.11 | 0.12 |
| | *Salicornia spp.* | 86 | 0.10 | -0.07 | 0.37 | 0.09 | 0.13 |
| | | | | | | | |
| **Random Forest Discrete** | All Vegetation | 694 | -0.01 | -0.60 | 0.56 | 0.11 | 0.11 |
| | *S. alterniflora* | 446 | 0.03 | -0.22 | 0.56 | 0.11 | 0.11 |
| | *S. patens* | 123 | -0.07 | -0.47 | 0.04 | 0.08 | 0.11 |
| | *Distichlis spicata* | 39 | -0.07 | -0.60 | 0.04 | 0.10 | 0.12 |
| | *Salicornia spp.* | 86 | -0.08 | -0.33 | 0.08 | 0.07 | 0.11 |

**Figure 6: a) Frequency of occurrence for uncorrected lidar residuals (lidar – RTK GNSS = ΔZ) by vegetation species (n = 694) across all three marsh sites. b) Frequency of occurrence for residuals as corrected by the TreeNet model using full-waveform and discrete-return lidar predictors (n = 694). c) Frequency of occurrence for TreeNet model residuals for discrete-return lidar predictors (n = 694). The red lines in each graph represent the combined total of all *S. alterniflora* ecophenes residuals.**

**Table 7: Variable importance is presented for each of the models that use only the discrete-return lidar data predictors. The most important variable is given a score of 100 and all other variables importance are rescaled relative to the most important variable. The top 3 important variable from each model run are highlighted in bold. " - " represents not found significant or used by the model.**

| Symbol | Predictor Variable | TreeNet | MARS | GPSM | Random Forest | CART |
|---|---|---|---|---|---|---|
| $\gamma$ | Curve | 14.21 | - | 1.23 | 0.45 | 0.78 |
| $\gamma_{pl}$ | Curve Plan | 12.29 | - | - | - | 3.18 |
| $\gamma_{pr}$ | Curve Profile | 17.9 | **13.78** | 2.186 | 0.72 | 7.51 |
| $d$ | Distance | 20.79 | - | 1.86 | **0.73** | **65.17** |
| $Z$ | Elevation | **100** | **100** | **100** | **100** | **100** |
| $\iota$ | Intensity | **23.87** | - | - | **2.44** | 14.68 |
| $R$ | Rugosity | **24.72** | **14.92** | **55.33** | 2.2 | **21.96** |
| $m$ | Slope | 14.68 | - | **6.28** | 0.2 | 14.74 |

**Table 8: Confusion matrices for the three classification models created to identify vegetation zonation. The three zones are bare ground (GR), high marsh vegetation [*S. patens*, *Salicornia spp.*, *D. spicata*, and short-form *S. alterniflora*] (HM), and low marsh vegetation [tall-form and medium-form *S. alterniflora*] (LM). The shaded diagonal (grey) contains the cases of agreement between the model and learn or test datasets.**

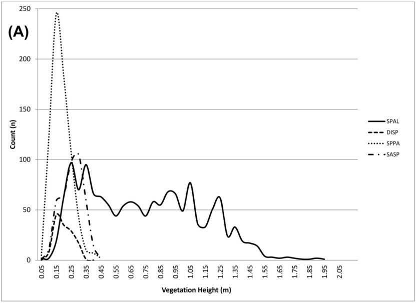
| Model | Class | Test Dataset | | | | | Learn Dataset | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | N | Correct | GR | HM | LM | N | Correct | GR | HM | LM |
| TreeNet | GR | 58 | 94.8% | 55 | 2 | 1 | 33 | 81.8% | 27 | 2 | 1 |
| | HM | 179 | 98.9% | 0 | 177 | 2 | 69 | 92.8% | 0 | 64 | 5 |
| | LM | 230 | 99.6% | 0 | 1 | 229 | 89 | 95.5% | 1 | 3 | 85 |
| | Total | 467 | 98.7% | 55 | 180 | 232 | 191 | 92.1% | 28 | 72 | 91 |
| Random Forest | GR | 58 | 89.7% | 52 | 5 | 1 | 33 | 97.0% | 32 | 1 | 0 |
| | HM | 265 | 77.4% | 9 | 205 | 51 | 110 | 83.6% | 4 | 92 | 14 |
| | LM | 237 | 82.3% | 15 | 27 | 195 | 82 | 86.6% | 6 | 5 | 71 |
| | Total | 560 | 80.7% | 76 | 237 | 247 | 225 | 86.7% | 42 | 98 | 85 |
| CART | GR | 58 | 91.4% | 53 | 4 | 1 | 33 | 87.9% | 29 | 4 | 0 |
| | HM | 265 | 83.4% | 1 | 221 | 43 | 110 | 86.4% | 0 | 95 | 15 |
| | LM | 237 | 81.9% | 12 | 31 | 194 | 82 | 78.0% | 5 | 13 | 64 |
| | Total | 560 | 83.6% | 66 | 256 | 238 | 225 | 83.6% | 34 | 112 | 79 |

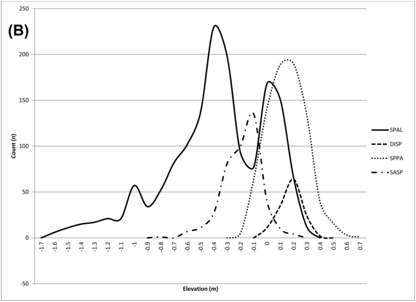
**Table 9: Variable importance is presented for each of the zonation models using all available predictors. The most important variable is given a score of 100 and all other variables importance are reported are rescaled relative to the most important variable. The top three important variables from each model run are highlighted in bold.**

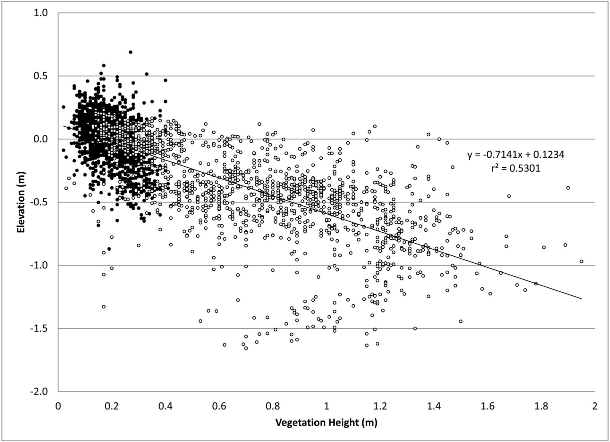
| Symbol | Predictor Variable | TreeNet | Random Forest | CART |
|---|---|---|---|---|
| $A$ | Waveform Amplitude | 35.66 | 12.4885 | 7.4541 |
| $AUC$ | Area under curve | 26.15 | 10.5998 | 10.2763 |
| $m_w$ | Waveform mean | 29.18 | 10.3625 | 19.2795 |
| $g_1$ | Waveform skewness | 21.36 | 8.1181 | 15.1362 |
| $w$ | Waveform Width | **100** | **100** | **92.5945** |
| $Z$ | Elevation | **68.63** | **98.8685** | **100** |
| $\gamma$ | Curve | 15.64 | 4.79 | 7.1263 |
| $\gamma_{pl}$ | Curve Plan | 21.33 | 3.8031 | 0.9588 |
| $\gamma_{pr}$ | Curve Profile | 24.25 | 7.3549 | 1.8469 |
| $d$ | Distance | **60.37** | **68.5338** | **77.8668** |
| $i$ | Intensity | 51.62 | 25.7267 | 38.6824 |
| $R$ | Rugosity | 32.92 | 22.667 | 41.7609 |
| $m$ | Slope | 29.32 | 9.5455 | 8.0322 |

**Figure 7: a) Map of Moors marsh vegetative zones developed from field collected data and interpretation from a 2009 high resolution aerial photograph. Salt ponds are not identified on this map. b) Map of marsh vegetative zones derived from the TreeNet model using all available predictors. Salt ponds and other water features are visible as data voids (white). Red ovals represent areas of high marsh vegetation (SF *Spartina alterniflora*) not interpreted using standard techniques, but detected by the full-waveform nonparametric model. Yellow circles are "bare ground" that have been misclassified as high marsh possibly due to the presence of macroalgae.**

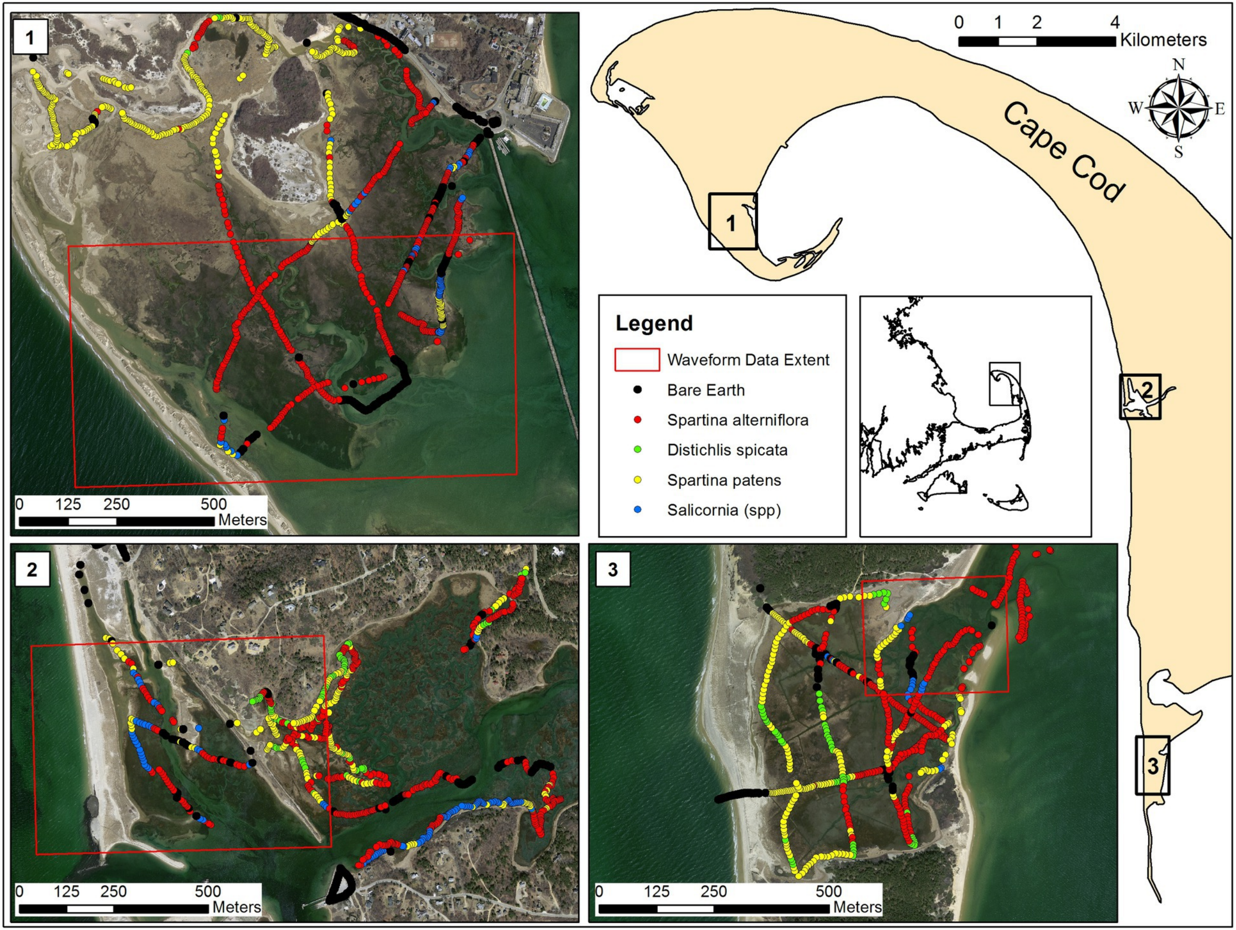
**Figure 8: a) Uncorrected lidar DEM of last (single) returns using an Inverse Distance Weighting algorithm with a radius of 1 cell. b) Full-waveform corrected DEM using the developed TreeNet model. Notice the visible topography that was hidden in the uncorrected DEM by vegetation-induced bias. c) Corrected DEM using discrete-return lidar derived predictor TreeNet model. Results are an improvement over the uncorrected DEM, but still contain significant vegetation-induced bias as compared to the full-waveform corrected DEM. d) Difference map between the uncorrected lidar DEM and the Waveform TreeNet model corrected DEM. e) Difference map between the uncorrected lidar DEM and the discrete-return lidar corrected DEM using the developed TreeNet model. f) Difference map between the full-waveform corrected difference map and the discrete-return lidar corrected difference map. These differences (m) are the "improvement" of the full-waveform model over the discrete-return lidar model at removing the vegetative induced bias. Elevations in panels a, b, and c are in meters and referenced to local MHW datum. Differences in panels d, e, f are attributed to model "removed" vegetation-induced bias and are measured in meters.**

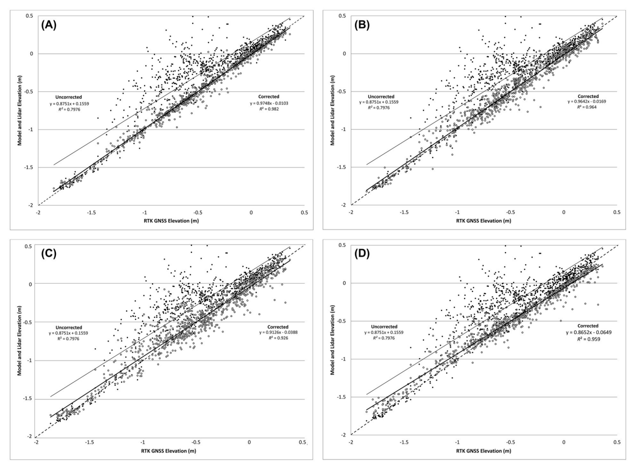

(A)
Count (n)
Vegetation Height (m)
SPAL
DISP
SPPA
SASP
0
50
100
150
200
250
0.05
0.15
0.25
0.35
0.45
0.55
0.65
0.75
0.85
0.95
1.05
1.15
1.25
1.35
1.45
1.55
1.65
1.75
1.85
1.95
2.05


(B)
Count (n)
Elevation (m)
SPAL
DISP
SPPA
SASP
-50
0
50
100
150
200
250
-1.7
-1.6
-1.5
-1.4
-1.3
-1.2
-1.1
-1
-0.9
-0.8
-0.7
-0.6
-0.5
-0.4
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7

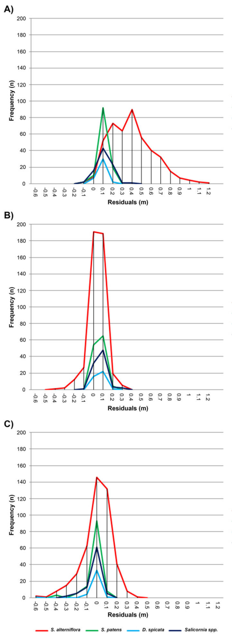


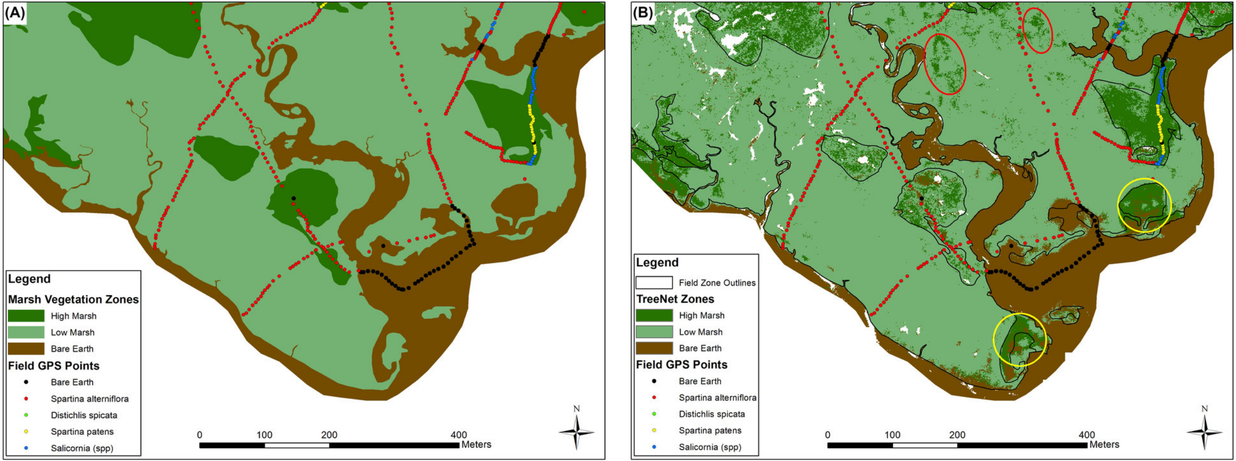
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
Elevation (m)
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
Vegetation Height (m)
y = -0.7141x + 0.1234
r² = 0.5301

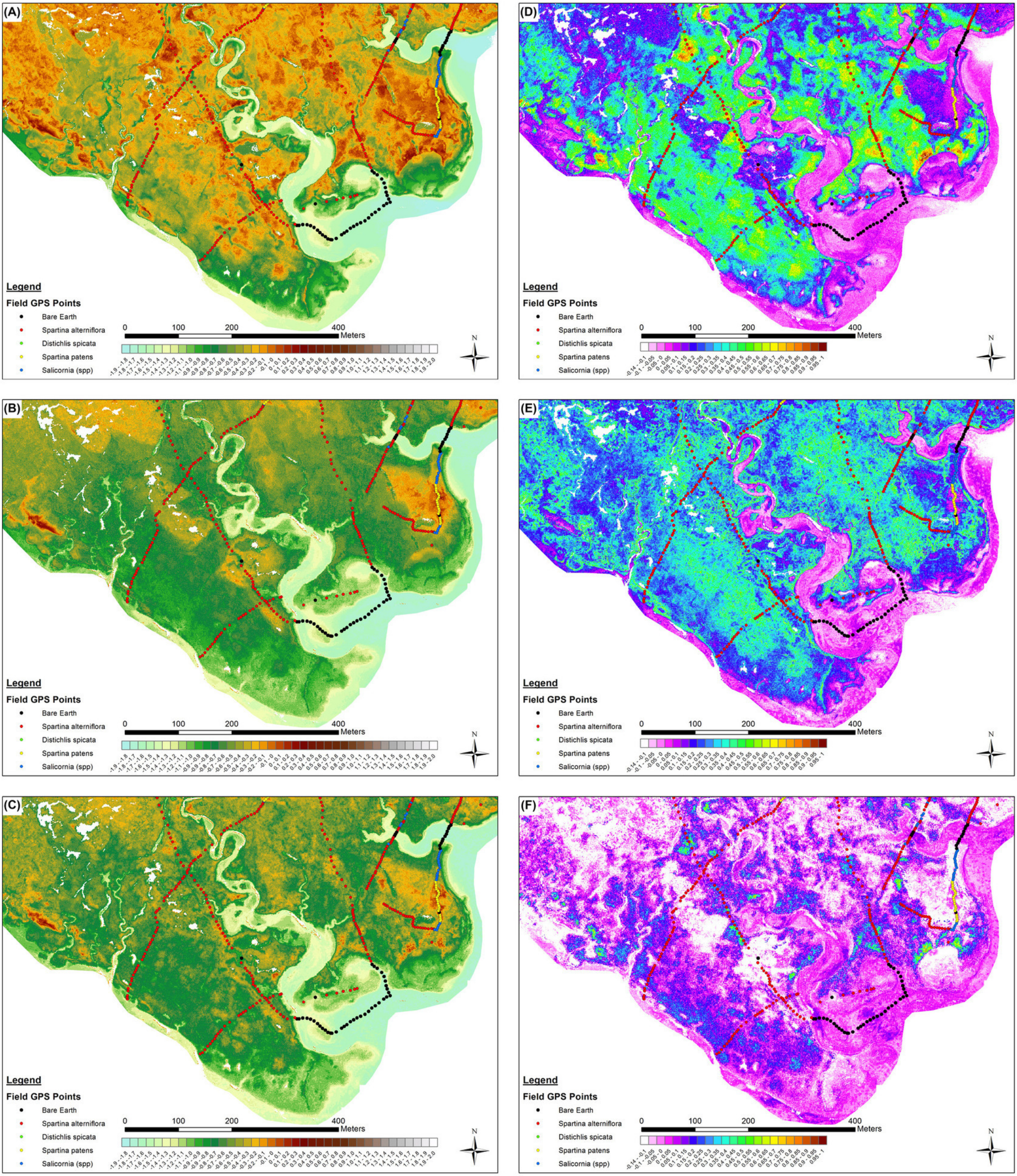

1
2
3
0 125 250 500 Meters
0 1 2 4 Kilometers
N
S
E
W
Cape Cod
Legend
Waveform Data Extent
Bare Earth
Spartina alterniflora
Distichlis spicata
Spartina patens
Salicornia (spp)



(A)
Model and Lidar Elevation (m)
RTK GNSS Elevation (m)
Uncorrected
y = 0.8751x + 0.1559
R² = 0.7976
Corrected
y = 0.9748x - 0.0103
R² = 0.982
(B)
Uncorrected
y = 0.8751x + 0.1559
R² = 0.7976
Corrected
y = 0.9642x - 0.0169
R² = 0.964
(C)
Uncorrected
y = 0.8751x + 0.1559
R² = 0.7976
Corrected
y = 0.9126x - 0.0388
R² = 0.926
(D)
Uncorrected
y = 0.8751x + 0.1559
R² = 0.7976
Corrected
y = 0.8652x - 0.0649
R² = 0.959


A)
B)
C)
Frequency (n)
Residuals (m)
0
20
40
60
80
100
120
140
160
180
200
-0.6
-0.5
-0.4
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1.1
1.2
*S. alterniflora*
*S. patens*
*D. spicata*
*Salicornia spp.*


(A)
Legend
Marsh Vegetation Zones
High Marsh
Low Marsh
Bare Earth
Field GPS Points
Bare Earth
Spartina alterniflora
Distichlis spicata
Spartina patens
Salicornia (spp)
0
100
200
400
Meters
N
(B)
Legend
Field Zone Outlines
TreeNet Zones
High Marsh
Low Marsh
Bare Earth
Field GPS Points
Bare Earth
Spartina alterniflora
Distichlis spicata
Spartina patens
Salicornia (spp)
0
100
200
400
Meters
N


(A)
Legend
Field GPS Points
Bare Earth
Spartina alterniflora
Distichlis spicata
Spartina patens
Salicornia (spp)
0 100 200 400
Meters
N
(D)
Legend
Field GPS Points
Bare Earth
Spartina alterniflora
Distichlis spicata
Spartina patens
Salicornia (spp)
0 100 200 400
Meters
N
(B)
Legend
Field GPS Points
Bare Earth
Spartina alterniflora
Distichlis spicata
Spartina patens
Salicornia (spp)
0 100 200 400
Meters
N
(E)
Legend
Field GPS Points
Bare Earth
Spartina alterniflora
Distichlis spicata
Spartina patens
Salicornia (spp)
0 100 200 400
Meters
N
(C)
Legend
Field GPS Points
Bare Earth
Spartina alterniflora
Distichlis spicata
Spartina patens
Salicornia (spp)
0 100 200 400
Meters
N
(F)
Legend
Field GPS Points
Bare Earth
Spartina alterniflora
Distichlis spicata
Spartina patens
Salicornia (spp)
0 100 200 400
Meters
N