Summer 8-13-2018

Forest Disturbance Detection and Aboveground Biomass Modeling Using Moderate-Resolution, Time-Series Satellite Imagery

John B. Kilbride
University of Maine, John.B.Kilbride@gmail.com

Recommended Citation

FOREST DISTURBANCE DETECTION AND ABOVEGROUND BIOMASS

MODELING USING MODERATE-RESOLUTION,

TIME-SERIES SATELLITE IMAGERY

By

John Kilbride

B.A. University of Vermont, 2016

A THESIS

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

(in Forest Resources)

The Graduate School

The University of Maine

August 2018

Advisory Committee:

Shawn R. Fraver, Professor of Forest Resources, Co-Advisor

Daniel J. Hayes, Professor of Forest Resources, Co-Advisor

Aaron Weiskittel, Professor of Forest Resources

Sean P. Healey, United States Forest Service Research Ecologist

FOREST DISTURBANCE DETECTION AND ABOVEGROUND BIOMASS

MODELING USING MODERATE-RESOLUTION,

TIME-SERIES SATELLITE IMAGERY

By John Kilbride

Thesis Advisors: Dr. Shawn R. Fraver and Dr. Daniel J. Hayes

An Abstract of the Thesis Presented
in Partial Fulfillment of the Requirements for the
Degree of Master of Science
(in Forest Resources)

August 2018

Human-induced and natural disturbances are an important feature of forest ecosystems. Disturbances influence forest structure and composition and can impact crucial ecosystem services. However, deriving spatially explicit estimates of past forest disturbance across a large region can prove challenging. Researchers have recognized that remote sensing is an important tool for monitoring forest ecosystems and mapping land use and land cover change. One of the most important sources of remotely sensed imagery is the United States Geologic Survey's Landsat program which has continuously acquired earth observations since 1972. This repository of imagery has the spatial, spectral, and temporal resolution necessary to produce maps of disturbance which are meaningful for the analysis of forested ecosystems.

In this analysis, we utilize the imagery from the Landsat archive to produce maps of forest disturbance from 1985 to 2017 for the New England states and the Canadian Maritime provinces. The change detection maps were developed using stacked generalization, a modeling technique that fuses the outputs of an ensemble of individual change-detection algorithms through the use of a secondary classifier. To better understand the error associated with these

classifications, we quantified the spectral characteristics associated with different harvesting practices. Using two case studies, the 1998 ice storm and the 2016 gypsy moth outbreak in southern New England, we performed experiments to examine how the stacked generalization framework can be utilized to increase the accuracy of disturbance maps following large-scale natural disturbances. The change detection maps developed in this analysis possessed a 98.7% overall accuracy and a 27.5% balance of the errors of omission and commission. Our results indicated that adjusting the probability threshold associated with the secondary classifier in the stacked generalization framework increase the spatial coherence of disturbance patches and better capture the low- to moderate-severity disturbances.

Using the maps of disturbance for the New England states and Maritime Provinces, we derived metrics describing the spectral change magnitude, timing, and percent spectral recovery across the study region. Recent research has found that including metrics of disturbance and recovery processes, derived from the analysis of time-series satellite imagery, can improve the accuracy of AGB models. However, these studies have largely been conducted in regions with relatively homogenous forest composition and structure and disturbance regimes dominated by stand-replacing disturbances. This analysis expands upon the existing literature by exploring how disturbance and recovery metrics can improve the predictions of AGB models in a heterogeneous landscape with a complex land-use history. Gradient boosting models, a sophisticated machine learning technique, were used to produce regional AGB models using spectral, disturbance, and environmental (e.g., topographic, climatological, etc.) metrics. Additionally, we explore how adjusting the rate of mapped disturbance through modifications to the class-inclusion rate associated with the secondary classifier can impact estimates of AGB. We conclude that

landscape heterogeneity, as well as the general lack of stand-replacing disturbances, negatively

impacts the predictive utility of disturbance and recovery metrics for modeling AGB.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE:

**ON DISTURBANCE DETECTION IN MIXED-SPECIES, STRUCTURALLY
COMPLEX, FORESTS USING LANDSAT TIME-SERIES
CHANGE DETECTION**

Human-induced and natural disturbances influence forest structure and composition.
Deriving accurate characterizations of past forest disturbance histories from remotely-sensed
imagery is important for assessing ecosystem services and developing robust tools for carbon
monitoring programs. In this analysis, we develop change detection maps for the forest area of
the New England states and the Canadian Maritime provinces using stacked generalization, a
modeling technique that fuses the outputs of an ensemble of individual change detection models
through the use of a secondary classifier. We also quantify the spectral characteristics associated
with different harvesting intensities, and we examine how the stacked generalization framework
can be utilized to increase the disturbance map accuracy following large-scale natural
disturbances. The classifier developed in this analysis produced a 98.7% overall accuracy and a
27.5% balance of the errors of omission and commission. Our results indicate that adjusting the
probability threshold associated with the secondary classifier can increase the spatial coherence
of disturbance patches and better capture low-severity disturbances associated with partial
disturbances.

1.1. Introduction

The spatial distribution and severity of forest disturbances are critical considerations for
assessing forest productivity, wildlife habitat quality, and ecosystem services, such as carbon
storage (Houghton, 2005; De Groot et al., 2010; Goetz et al., 2009). Satellite imagery provides
the only efficiently means of assessing forest disturbances, in a spatially explicit manner, across

large spatial scales (Kennedy et al., 2014; Hansen et al., 2013; Powell et al., 2010a). Imagery from the United States Geologic Survey's (USGS) Landsat program serves as the principal data source for characterizing historical disturbance because of its moderate spatial resolution and continuous acquisition over the past four decades (Cohen and Goward, 2004; Kennedy et al., 2014). Due to computational limitations, early analyses of forest disturbances were restricted to a small number of images, temporally separated by several years (Coppin and Bauer, 1996; Wilson and Sader, 2002). These techniques have now been supplanted by sophisticated change-detection algorithms that can detect disturbances on an annual or sub-annual time-step using dense time-series imagery (Kennedy et al., 2010; Huang et al., 2010; Zhu and Woodcock, 2014; Hughes et al., 2017). Recent advancements have led to ensemble change detection models using stacked generalization (Healey et al., 2018; Wolpert, 1992). Stack generalization improves the change detection accuracy by synthesizing the outputs of an of ensemble of individual models through the use of a secondary classifier, such random forests (Cohen et al., 2018; Healey et al., 2018; Breiman, 2001).Previous remote-sensing studies have successfully utilized change detection techniques to reconstruct the history of forest disturbances. The resulting spatially extensive maps of forest disturbance have focused on high-severity disturbance events such as clear-cutting, crown fire, and deforestation (Hansen et al., 2013; Hermosilla et al., 2015b). However, disturbance regimes dominated by low- to moderate-severity disturbances such as partial-harvesting, wind storms, ice storms, or insect outbreaks may evade detection for change detection approaches due to low magnitude of change in the spectral values (Cohen et al., 2017). As a result, many of the current assessments of forest disturbance based on change detection may dramatically underestimate forest disturbance rates and spatial extents. Further, investigations that have incorporated disturbance mapping into the biomass modeling or carbon-stock change

2

analyses have been performed in generally homogenous forest types with relatively simple disturbance regimes (Pflugmacher et al., 2014; Frazier et al., 2014). Thus, research that clarifies the precision of change-detection algorithms in structurally diverse forests with mixed-severity disturbance regimes would provide needed insight into the suitability of using change-detection algorithms across a broader range of forest types and disturbance regimes.

Our overarching objective in this study was to produce a focused analysis of forest disturbance across a broad-region that features a structurally complex, mixed-species forest, as well as mixed-severity disturbance regime. The specific objectives were to 1) recreate the history of forest disturbances (1986 – 2017) for the U.S. New England States and the Canadian Maritime provinces, 2) quantify and compare the omission and commission error associated with a range of disturbance severities, and 3) evaluate two large-scale mixed-severity natural disturbance events to assess how adjusting the class inclusion rates of the secondary classifier can improve change detection accuracy. This study used stacked generalization to combine outputs from the LandTrendr change-detection algorithm (Kennedy et al., 2010), run multiple times using spectral bands or indices as the input, to create a spectral-ensemble of modeling outputs (Cohen et al., 2018). Our results will allow assess the performance of change-detection algorithms for estimating past disturbance severity and spatial extent over a broad region with a highly heterogeneous landscape composition.

1.2. Methods

1.2.1 Study area

The study area is ~60,800,000 ha in size and is composed of the U.S. New England states and the Canadian Maritime provinces. Forest species composition transitions along a latitudinal gradient from the hardwood-dominated stands in southern New England to the conifer-

dominated boreal forests in Northern Maine and throughout the Canadian Maritime provinces. The forest structure and composition are heavily influenced by natural regeneration, and mixed-age forest stands are common. Partial harvesting regimes are common in the mixed-species stands in the New England states (Belair and Ducey, 2018; Canham et al., 2013). Clear-cutting is more commonly practiced in the conifer-dominated forest stands found throughout the Canadian Maritime province. Intensively managed, single-species plantations occur in portions of the New Brunswick province (Hennigar et al., 2016). The natural disturbance regime in the region is dominated by low- to moderate-severity, periodic events such as wind storms and insect outbreaks (Fraver et al., 2009; Seymour et al., 2002).

Data from the U.S. Forest Inventory and Analysis program, Nova Scotia permanent sample plot database, and the New Brunswick permanent sample plot database indicate that 117 tree species are present throughout the study area. The Acadian Forest region of the study area in the Maritime Provinces and northern New England feature conifer-dominated or mixed-species stands predominantly composed of balsam fir (*Abies balsamea* (L.) Mill.) red spruce (*Picea rubens* Sarg.), white spruce (*Picea glauca*), white pine (*Pinus strobus* L.), red maple (*Acer rubrum* L.), and American beech (*Fagus grandifolia* Ehrh.) (Hennigar et al., 2016). The hardwood-dominates species in the southern New England states are generally composed of oak (*Quercus rubra* L.), paper birch (*Betula papyrifera* Marsh.), yellow birch (*Betula alleghaniensis* Marsh.), eastern hemlock (*Tsuga canadensis* L.), and sugar maple (*Acer saccharum* Marsh.) (Seymour, 1994; Thompson et al., 2013).

1.2.2. Image processing

USGS Landsat Collections Tier 1 Thematic Mapper (TM), Enhanced Thematic Mapper Plus (ETM+), and Operation Land Imager (OLI) imagery from 1984 – 2017 with less than 85%

cloud cover were selected for analysis. Landsat Collections Tier 1 is considered suitable for time-series analysis with all images possessing a scene-wide radial geometric root mean square error ≤ 12m. Surface reflectance values were calculated using the Landsat Ecosystem Disturbance Adaptive Processing System algorithm and the Landsat 8 Surface Reflectance Code algorithm (Masek et al., 2006; Vermote et al., 2016). The surface reflectance values from the OLI sensor were cross-calibrated with the TM and ETM+ sensors using coefficients from Roy et al. (2016).

Cloud and cloud shadow removal was performed using the FMASK algorithm (Zhu and Woodcock, 2012). The Temporal Dark Outlier Method (TDOM) was used to remove cloud shadows committed by the FMASK algorithm (Housman et al., 2015). The TDOM generates a population of clear-sky observations around each pixel in a Landsat scene. If the value of a pixel is 1.5 standard deviations lower than the pixel-population mean, the pixel is discarded. To account for variation in forest phenology throughout the study region, only pixels within ±32 days of the peak of the growing season were included for further analysis. The peak of the growing season was estimated using information from the Moderate Resolution Imaging Spectroradiometer (MODIS) Land Cover Dynamics dataset (Ganguly et al., 2010).

Prior to temporal segmentation using the LandTrendr algorithm, annual summer-time composited images were produced. Image compositing was performed using the medoid compositing method (Flood, 2013). Qualitative testing indicated the medoid method was more robust to noisy observations than was the pixel-scoring criteria proposed by White et al. (White et al., 2014) and Hermosillia et al. (2015a) and (2015b). The medoid method relies on calculating the Euclidian distance between observations in the spectral space. Consequently, if fewer than three observations were available, the mean spectral value of each band was used. All image

analysis and subsequent change detection analysis were performed using the Google Earth

Engine (GEE) cloud-computing platform (Gorelick et al., 2017).

Using the six spectral bands common to the TM, ETM+, and OLI sensors (i.e., the blue,

green red, near-infrared, and the two shortwave infrared bands), eight spectral indices were

calculated. The three primary components of the Tasseled Cap (TC) transformation, brightness,

greenness, and wetness, were computed using the coefficients from Crist and Cicone (1984).

Tasseled cap angle (Powell et al., 2014) and distance (Duane et al., 2010) were then derived from

these TC components. Three additional vegetation indices, the Normalized Difference

Vegetation Index (NDVI), the Normalized Difference Moisture Index (NDMI), and the

Normalized Burn Ratio (NBR) were also computed (Rouse et al., 1974; Wilson and Sader, 2002;

Key and Benson, 1999). This process yielded a total of 14 spectral bands and indices that were

used to generate an ensemble of LandTrendr models.

1.2.3. TimeSync reference data

An historical record of forest disturbance over a set of reference plots was produced using

the TimeSync image interpretation program (Cohen et al., 2010). This process allows the

interpreter to visualize the time-series of spectral values, extracted from the Landsat archive,

over a single pixel (hereafter referred to as a plot). TimeSync links the plot locations to historical

acquisitions of high-resolution imagery in Google Earth to facilitate the interpretation of the

time-series. The interpreter subdivides the time-series into a sequence of linear segments that

correspond to changes observed in the imagery. At each segment, the interpreter labels the

change process (e.g. no change, wind disturbance, forest harvesting, etc.), and the land use and

land cover at the segment's start and end dates. For this analysis, the image interpreters also

classified each harvest event as either a "partial harvest" or a "clear-cut". To ensure consistent

classifications, all of the plots were evaluated by two interpreters to ensure consistent

classification and to avoid the omission of subtle change-events. Disagreements between two

interpretations were resolved with the aid of a third interpreter.

The reference dataset consisted of 4233 TimeSync plots that contained interpretations

from 1984-2016. Of those plots, 797 were never forested and were excluded from subsequent

analysis. Any segments that were not initially forested, indicated by the classification at the

segments start date, were also removed from the dataset. The final database consisted of 105,339

annual observations that were labeled as either disturbed or not-disturbed based on the change

process indicated by the interpreter. These annual observation serve as the training and validation

data for developing the stacked-generalization ensemble classifier.

1.2.4 Training the base learners

An ensemble of change detection models were developed using the GEE-implementation

of the LandTrendr algorithm (Kennedy et al., 2010; Kennedy et al., 2018). LandTrendr uses a

specified given band or index and partitions the time-series of values for an individual pixel into

linear-segments using a regression-based vertex-identification procedure. Using the first and last

observations within the pixel's time-series as the initial vertices, the subsequent vertex is fit at

the observation with the greatest absolute deviation from the fitted curve. Each of the new

segments are then subset and retained based on the mean square error (MSE) of the fit. The

process is iterated until a stopping criteria is reach. This maximally complex model is then

simplified by iteratively removing the weakest vertices. From the resultant series of models, the

p-value of the F-statistic for the curve fit is used to select a sufficiently complex model to

characterize the pixel's time-series (see Kennedy et al. (2010) for additional details). The

LandTrendr algorithm was applied to all pixels across the study region using each of the shared

TM, ETM+, and OLI spectral bands and indices calculated during image processing. This resulted in each pixel containing the output from the ensemble of fourteen base-learners. Qualitative testing indicated that increasing the sensitivity of the base learners improved the secondary classifier's accuracy at the cost of increasing the false-positive rate of disturbance detection for the individual base-learners. As suggested in Cohen et al. (2018), the segment filtering portion of the LandTrendr algorithm was disabled.

The linear-segments in each of the trained base learners were used to calculate three metrics: 1) the fitted spectral value, 2) the difference in the fitted spectral values between the current year (t) and the previous year (t-1), and 3) a binary label that indicated if the slope of the fitted segment was greater than or equal to zero. This resulted in a total of 42 predictors produced by the base-learners. Prior to their inclusion in the ensemble-models, the conditional importance values (Strobl et al., 2008) of base-learner metrics were assessed using the Party package (Hothorn et al., 2015) in the R programming language (R Core Team, 2017). None of the metrics negatively influenced the secondary classifier performance; thus all metrics selected to develop the multi-spectral ensemble model.

1.2.5 Training the ensemble classifier

The outputs of the individual base-learners were combined using a secondary classifier, here the random forests (RF) algorithm (Breiman, 2001). The modeling dataset consisted of all records in the TimeSync reference database and the corresponding base-learner outputs, for each year of each plot. The dependent variable was the binary "disturbed" or "not disturbed" label, derived from the TimeSync interpretations. Preliminary testing indicated that the RF's out-of-bag (OOB) prediction stabilized after 500 decision trees were included in the RF ensemble. The performance of the RF model was assessed using the RF OOB error. The OOB error reflects the

mean prediction error for a given sample x_i across the all trees in the RF, where x_i was withheld during the bootstrapped aggregation ("bagging") procedure (Breiman, 1996). The error for the secondary classifier was derived from the GEE RF used to produce the wall-to-wall disturbance maps.

The binary disturbance classes in the modeling dataset were highly imbalanced with only ~2.6% of the records in the TimeSync reference dataset labeled as disturbances. Typically, one would either sub-sample the majority class or generate synthetic minority class data using a technique such as synthetic minority over-sampling technique (SMOTE) (Chawla et al., 2002) to balance the training dataset. However, because the sensitivity of the base-learners was increased in order to capture subtle disturbance events (e.g. wind disturbances, partial harvesting), the application of these techniques produced a significant increase in the commission error rate. Training across the entire dataset resulted in a more conservative RF classifier with fewer commission errors. The imbalance in the errors of omission and commission produced by the RF over an imbalanced dataset were accounted for by adjusting the probability threshold that determined class membership. The threshold was set at the point where the errors of omission and errors of commission were balanced. The errors for disturbance mapping were derived from the RF classifier that was used to produce wall-to-wall disturbance maps in Google Earth Engine.

1.2.6 Annual estimates of disturbed forest area

To derive annual estimates of forest disturbance, it was necessary to mask non-forested areas from our analysis. A binary forest mask was developed using the 2011 National Land Cover Dataset (NLCD) and the Agriculture and Agri-Food land cover datasets from 2011-2017 (Fisette et al., 2013; Homer et al., 2015). To account for harvesting and land-use change that may

have influenced the land cover classifications in either dataset, areas of forest gain identified in the Hansen Global Forest Change (2000-2017) were labeled as forests (Hansen et al., 2013). This masking process reduced the chance of falsely attributing forest disturbances to agricultural and wetland ecosystems, which experience strong year-to-year changes in their spectral characteristics.

1.2.7. Modifications to the mapped disturbance rate for capturing subtle disturbance events

Large-scale natural disturbances can vary in their intensity and patch-size across the landscape. A balancing the errors of omission and commission may omit a large portion of the disturbance area that was not impacted by stand-replacing disturbances. Additionally, due to the ensemble classifier being trained on data where the majority of disturbances are not natural in origin (e.g., harvesting, land-use change, etc.), there may be insufficient training data for natural disturbances to be correctly assigned. In this analysis, two case studies selected for analysis: the 1998 ice storm that impacted the Northeastern United States and the 2016 gypsy moth (*Lymantria dispar*) outbreak in southern New England. For both disturbance events, the probability threshold of the secondary classifier was allowed to vary between 0% and -25% in increments of 5% around the balance point of the errors of omission and commission. Decreasing the probability threshold decreases the minimum probability required for membership in the disturbed category and thus increases the mapped area of disturbance. The analysis of the 1998 ice storm focused on forest disturbances that occurred throughout New Hampshire, because of the large portion of the forests which experienced storm-related damage. The analysis of the 2016 gypsy moth outbreak focused on Connecticut, Massachusetts, and Rhode Island.

1.3. Results

1.3.1 Disturbance detection accuracy assessment

The secondary classifier achieved a 98.7% overall accuracy on the samples in TimeSync reference dataset after balancing the errors of omission and commission. A comparison of the TimeSync reference data and the ensemble classifier's predictions is presented in Table 1. Balancing the probability threshold of the secondary classifier produced a 27.2% balance of the errors of omission and commission. Cohen's kappa (k=0.74), a comparison of the classifier's accuracy with the expected accuracy (i.e, randomized predictions), indicated a strong agreement between the change detection results and the reference data.

Table 1.1. Accuracy if the ensemble classifier's prediction, based on accuracy to the TimeSync reference dataset.

		TimeSync reference data			
		Not-disturbed	Disturbed	Row Totals	Errors of Commission
	Not-disturbed	94734	700	95434	0.7%
Ensemble Classification	Disturbed	697	1867	2564	27.2%
	Column Totals	95431	2567	-	-
	Errors of Omission	0.7%	27.3%	-	-

1.3.2. Spectral characteristics of forest harvests

Using the TimeSync interpretations, harvest information was extracted from the disturbance maps. Harvesting events, which generally had lower changes in spectral magnitudes, were more frequently misclassified than disturbances with greater change magnitudes such as clear cutting or land use change (Figure 1). Partial harvesting events were

correctly classified 44.3% of the time. Clear-cuts were correctly classified 77.9% of the time. The general relationship identified between the NBR magnitude and change detection accuracy (i.e., lower change magnitudes produce greater rates of misclassification) was identified in other spectral bands and indices.

1.3.4. Annual disturbance rates by state and province

A summary of the annual forest disturbance rates is presented in Table 2. Across the entire study region, on average 0.93% ± 0.19% of all forested areas were disturbed each year amounting to an annual disturbance rate of 323097 ± 64457.20 ha. Plotting the individual trends in percent area disturbed (Figure 2) indicates large variations in the annual rates of disturbance between different states and provinces. New Brunswick and Maine both possess the greatest rates of annual disturbance both in terms of the percentage of total forest area and in terms of absolute area. Sudden increases against the background disturbance rate during years of broad-scale natural disturbances can be observed in Figure 2. The impact of the 1998 ice storm is prominent in New Hampshire and Maine, and the defoliation caused by the 2016 gypsy moth is observed in Connecticut, Massachusetts, and Rhode Island. Additionally, several other disturbance events are captured, including the impact of wind storm damage from the 2004 nor'easter in Nova Scotia, for example.
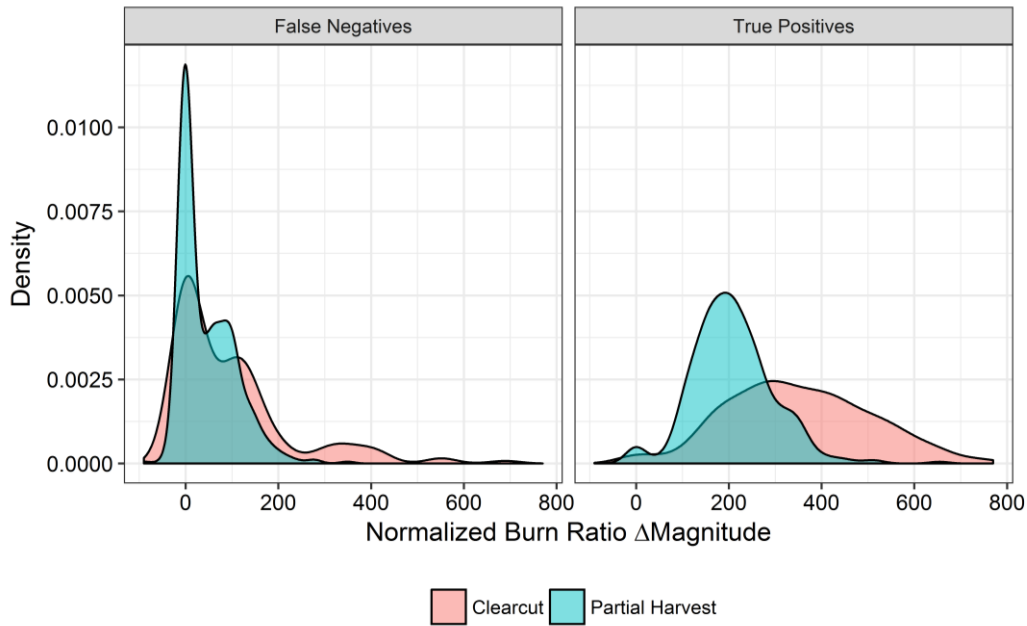
Figure 1.1. Probability density distributions for the Normalized Burn Ratio (NBR) change magnitude of harvests events. The harvests are stratified by classification outcome and by the TimeSync interpreters' classification of the harvest type.

Table 1.2. A summary of the annual disturbance rates, derived from the time-series of disturbance maps (1985-2017), by state and province.

	% of forest area		Total area (ha)	
	Mean	Std. dev.	Mean	Std. dev
Connecticut	0.40%	0.68%	4855	8191
Maine	1.12%	0.20%	102577	18622
Massachusetts	0.40%	0.37%	7317	6609
New Brunswick	1.19%	0.23%	115251	22570
New Hampshire	0.65%	0.27%	17521	7413
Nova Scotia	0.88%	0.29%	61204	20112
Prince Edward Island	0.99%	0.50%	4266	2138
Rhode Island	1.05%	3.63%	2183	7588
Vermont	0.30%	0.09%	7885	2293
Overall	0.93%	0.19%	323097	64457
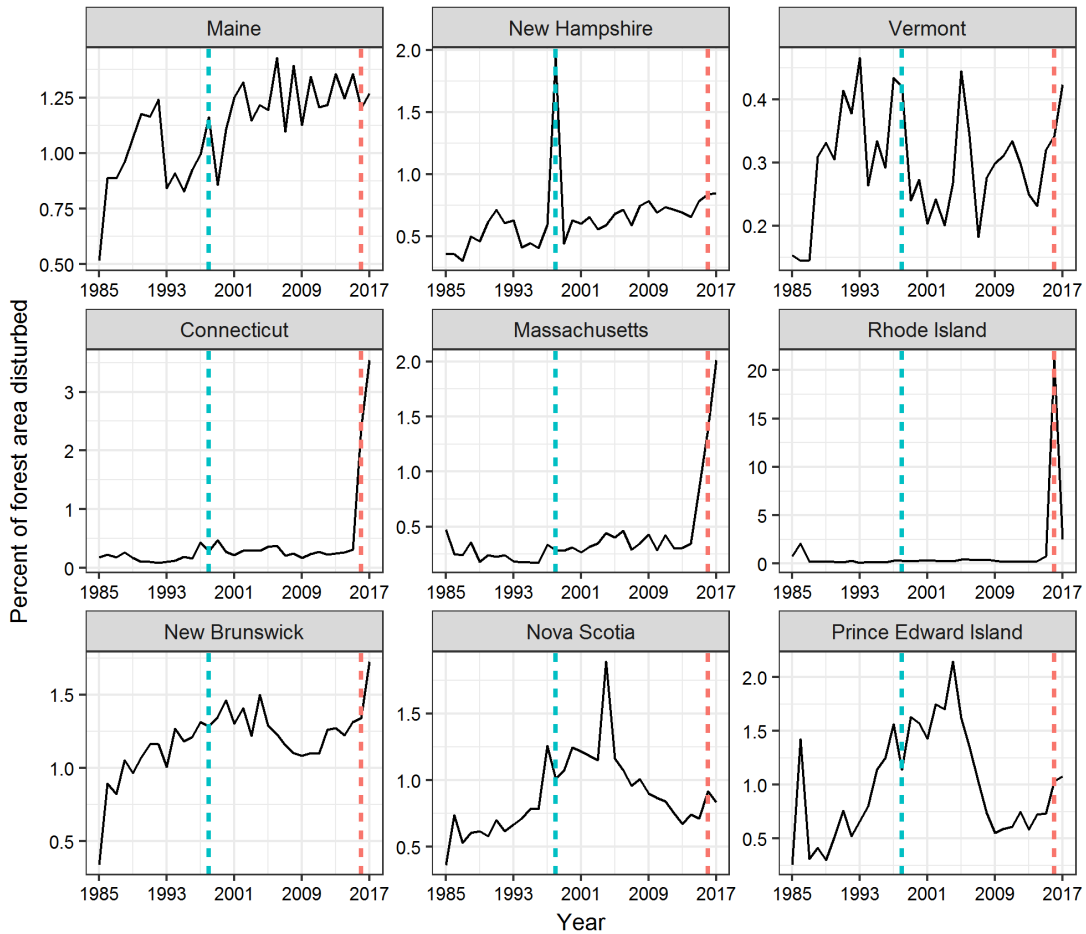
Figure 1.2. Trends in the annual percentage of forest area disturbed by state and province. The two natural disturbances investigated in this paper are denoted with vertical dashed lines: the 1998 ice storm (blue) and the 2016 gypsy moth outbreak (orange). Note that the scale for percent of forest area disturbed on the y-axis varies among panels.

1.3.5. Impact of adjusting the mapped rate of disturbance to capture natural disturbances

1.3.5.1 The 1998 Ice Storm

When mapping disturbances using the balanced probability threshold, the mapped area of disturbance throughout New Hampshire, in 1998, increased by 52632 ha relative to the average annual rate of disturbance over the 1985-2017 time period. Adjusting the probability threshold of the secondary classifier (between 0% and -25%) produced a 28% – 277% (50039 ha - 128337 ha) increase in total mapped disturbance area throughout the state of New Hampshire (Figure 3).

Decreasing the probability threshold improved the spatial cohesion of the disturbed areas (Figure 4). This improvement occurred around the edges of disturbance patches where mixed-pixels were more likely to be classified as "not-disturbed."

1.3.5.2 The 2016 gypsy moth outbreak

Gypsy moth defoliation in 2016 impacted 2.4% (28570 ha) of the forested areas in Connecticut and 1.4% (24,996 ha) of the forested areas in Massachusetts. In Rhode Island, where the outbreak was the most severe, 21.1% (44,030 ha) of the forested areas experienced defoliation. The defoliation in Rhode Island in 2016 was significantly larger than the total area disturbed through the remainder of the time series (28,024 ha). Decreasing the probability threshold (between 0% and -25%) increased the rate of mapped disturbance by 14 – 203% in Connecticut, 19 – 243% in Massachusetts, by 11 – 150% in Rhode Island (Figure 2). While increasing the rate of mapped disturbance improved the spatial cohesion of some disturbance patches, it also increased the number of small, scattered pixel clusters classified as disturbed at higher probability threshold settings (Figure 5).
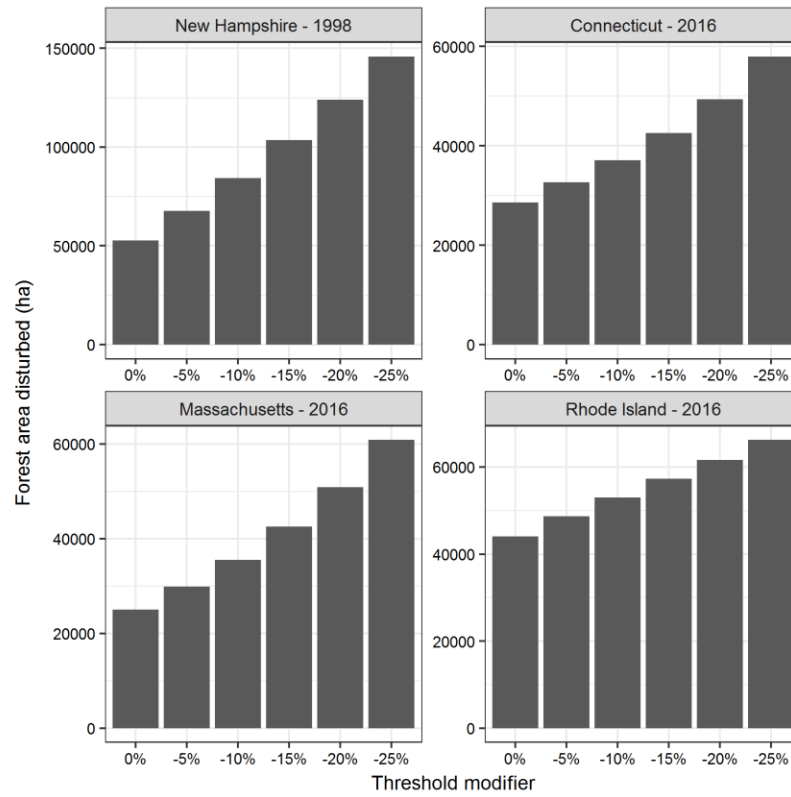
Figure 1.3. Changes in the mapped area of disturbance as the probability threshold of the random forest secondary classifier is adjusted by a modifier, shown for major disturbance years for each state**.**
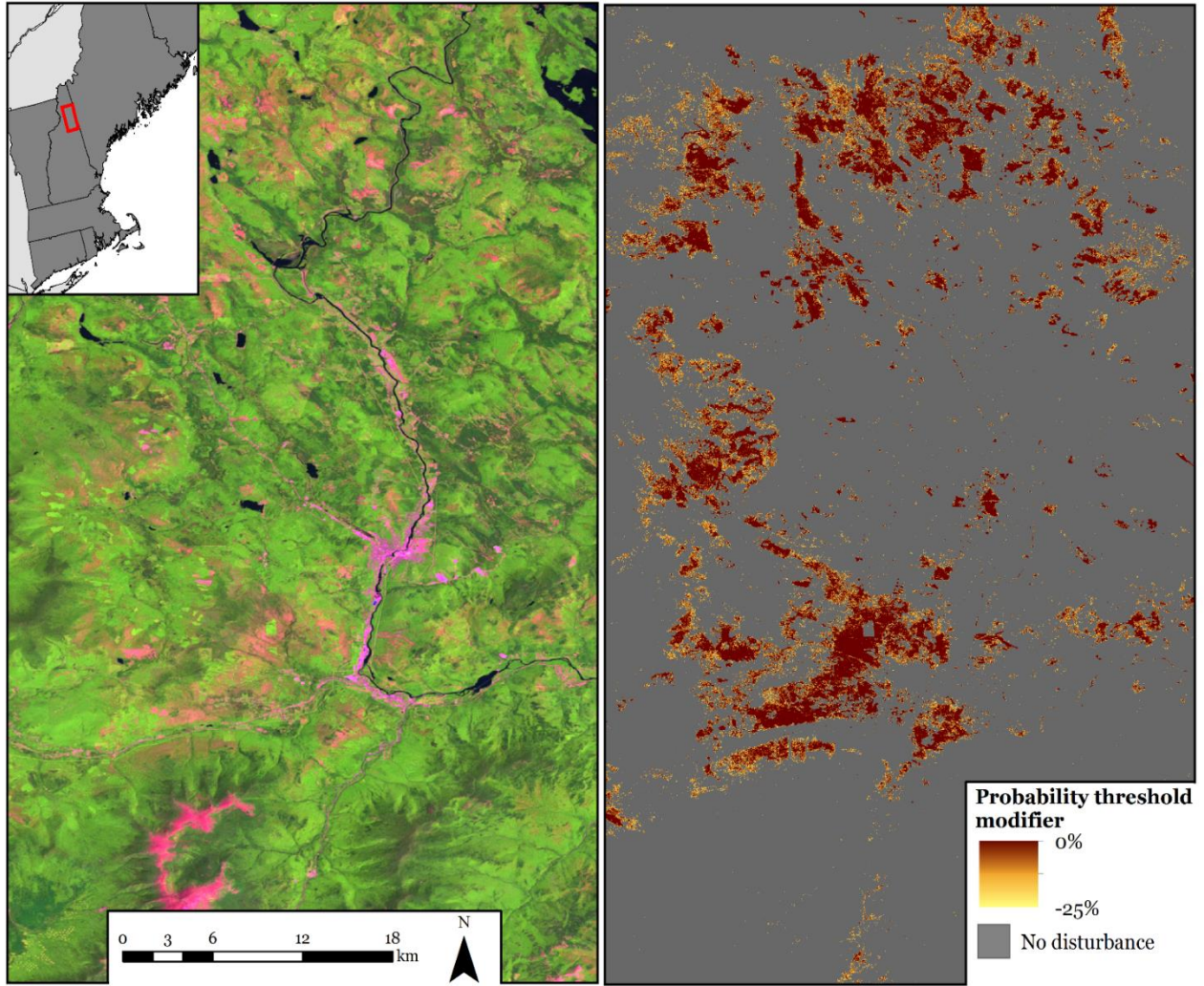
Figure 1.4. A sample of the 1998 Landsat summer-time composited image (left), located in northern New Hampshire, depicting ice storm damage. Change detection output varies as the probability threshold is modified from the balanced threshold (right).
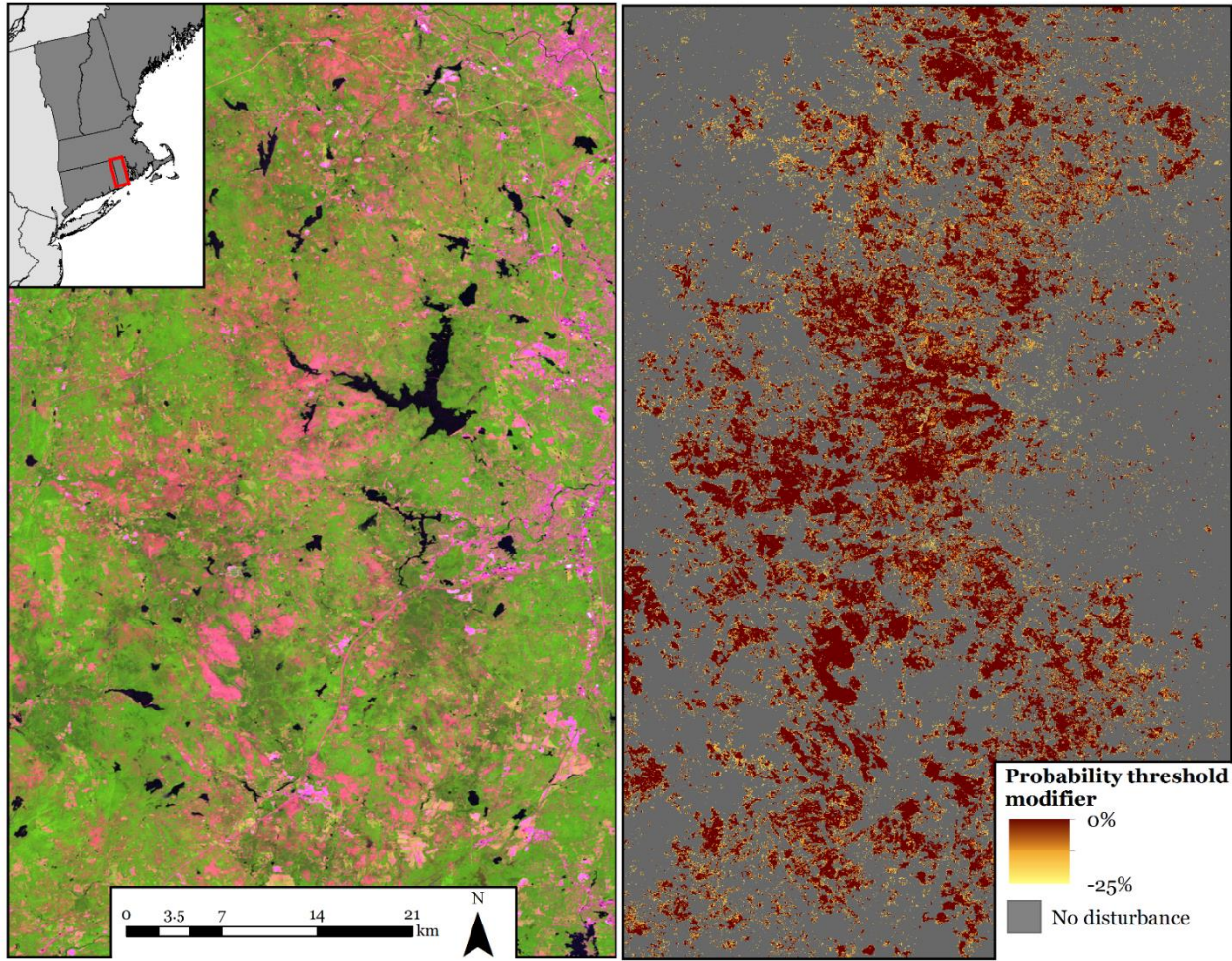
Figure 1.5. A sample of the 2016 Landsat summer-time composited image (left) detailing gypsy moth defoliation in Rhode Island. Change-detection output varies as the probability threshold is modified from the balanced threshold (right).

1.4. Discussion

In this study we have developed maps of forest disturbances from 1985 to 2017 over a large region with a highly heterogeneous landscape and a complex disturbance regime. This allowed us to assess the practicality of the stacked generalization methodology for generating broad-scale regional estimates of disturbance and to assess the spectral characteristics associated with partial-harvests and clearcuts. Additionally, we explored one of the properties associated with the secondary classifier, i.e., the ability to modify the probability threshold that controls class inclusion rates, and evaluated the changes in classification accuracy of large-scale, mixed-intensity disturbance events.

1.4.1. Disturbance mapping

The 98.7% overall accuracy and 27.5% balance of omission and commission errors found in the present study compares favorably to those of previous change detection analyses. Validation of the North American Forest Dynamics CONUS disturbance maps indicated an 84.5% overall accuracy and with an average of 46.4% errors of omission and an average of 46.7% errors of commission (Zhao et al., 2018).Healey et al. (2018) utilized stacked generalization to combine the outputs of an ensemble of change-detection algorithms and achieved a 40% balance of omission and commission errors. Cohen et al. (2018) utilized stacked generalization to combine the outputs from an ensemble of LandTrendr models, run using different spectral indices, and achieved a 29.6% balance of omission and commission error. The improvements observed in Cohen et al. (2018) and our results, relative to those presented in Healey et al., (2018), can be attributed to inclusion of the change in fitted spectral magnitude as a predictor of disturbance, alongside the spectral values and binary disturbance labels produced by the base-learners. The small improvement in accuracy in our study, may be attributed to subtle

improvements in the image pre-processing and optimization of the LandTrendr parameters for application in our study region. Our study demonstrates that across a broad range of forest ecosystems, stacked generalization robustly captures the general trends associated with clear-cut harvesting. However, as discussed in Cohen et al. (2017) and Healey et al. (2018), low-severity natural disturbances and partial harvesting remain challenging to reliably detect using these methods.

1.4.2. Case-studies for modifications of the disturbance threshold

In order to further explore the utility of these methods, we selected two case studies that highlight the challenges of disturbance detection in a region where high-severity, stand-replacing natural disturbances are rather uncommon. These disturbance events were selected for two reason: 1) the disturbances were highly-varied in their intensity across the landscape and 2) the disturbances were extensively documented and provide ample grounds for comparison. The former is important for our analysis because the ability of moderate resolution satellite imagery to capture high-severity disturbances has been well documented (Cohen et al., 2017; Hansen et al., 2013). By selecting disturbances with a broad gradient in their impact across the landscape, we can assess 1) if adjustments made to the probability threshold associated with the secondary classifier can be used improve disturbance classification, and 2) the degree to which these improvements match reference information.

1.4.2.1. Ice storm damage in New Hampshire

The 1998 ice storm damaged 1,375,931 ha of forested area throughout New Hampshire (Irland, 1998). Aerial surveys conducted in 1998 estimated that approximately 161,874 ha of New Hampshire's forests had experienced heavy damage, defined as damage to 50% or more of the canopies on the majorities of trees (Irland, 1998). The mapped area of disturbance produced

by the secondary classifier in the present analysis ranged 52,632 ha to 145,858 ha. The estimate

of disturbed forest produced by the largest modification to the balanced probability threshold (-

25%) was closely aligned with those from the aerial surveys. Given the Landsat sensors'

moderate spatial resolution (30m), estimates of disturbance will emphasize moderate- to high-

severity disturbances. Damage estimates produced by Quebec's Department of Natural

Resources categorized 64% of the 1998 ice storm damage as moderate to slight/trace (Irland,

2000). Assuming a roughly comparable distribution of damage types throughout New

Hampshire, it would not be reasonable to expect change-detection disturbance estimates

comparable to 1,375,931 ha of cumulative damage estimate for that state. This is because

approximately two thirds of the disturbances would correspond to a disturbance patch with a size

or severity not observable by moderate resolution satellite imagery. Adjusting the probability

threshold was a successful mechanism for increases the mapped area of disturbance to more

closely align with estimates of storm damage derived from aerial survey data.

1.4.2.2. Gypsy moth defoliation in Southern New England

The United States Forest Service (USFS) Forest Health Technologies Enterprise Team

(FHTET) produces yearly estimates of defoliation and mortality attributable to forest pests. In

2016, the FHTET estimated 84,906 ha were disturbed in Connecticut, 145,540 ha were disturbed

in Massachusetts, and 99,260 ha were disturbed in Rhode Island. Our tests found that using the

greatest modification (-25%) to the balanced probability threshold, the model classified 57,984

ha of disturbed forests in Connecticut, 60,823 ha of disturbed forests in Massachusetts, and

66,210 ha of disturbed forests in Rhode Island. Our disturbance estimates appear to

underestimate the area estimates by the FHTET, though it is difficult to compare the estimates of

area disturbed from aerial sketch mapping to those the spatially explicit estimates of disturbance

derived from the Landsat imagery. Pasquarella et al. (2017) developed defoliation estimates for the 2016 gypsy moth-outbreak using the Landsat time-series data and the Continuous Change Detection and Classification (CCDC) algorithm. The Landsat CCDC outputs produced disturbed area estimates for Rhode Island between 22,000 ha and 61,800 ha, and was able to accurately capture the areas which had experienced defoliation. The disturbed forest area estimates for Rhode Island in this study show a strong agreement with the 2016 gypsy moth defoliation estimates of Pasquarella et al. (2017). As discussed above, limitations associated with the spatial resolution of the Landsat sensors and data-gaps produced by clouds and atmospheric conditions hamper the ability to detect low-severity disturbances. Modifying the probability threshold was an effective means to probabilistically increase the disturbed-class membership to produce disturbance estimates that better matched the results of other analyses.

1.5. Conclusion

Our analysis has demonstrated that stacked generalization can produce accurate estimates of disturbance in a heterogeneous forest landscape with a complex disturbance regime. Our results corroborate the challenges previously identified surrounding the use of moderate-resolution imagery to capture low-severity disturbances, which typically have small changes in spectral magnitude (Cohen et al., 2018; Cohen et al., 2017). The benefit of utilizing Landsat data is programs long history of data acquisition, allowing for landscape dynamics to be characterized. Fusing the Landsat archive with new satellite imagery with greater spatial or spectral resolutions, such as the European Space Agency's Sentinel-2 A/B platforms, could improve the detection rate of low- to moderate- severity disturbance events.

The results of our two case studies suggest that adjusting the probability threshold associated with the secondary classifier in a stacked generalization change detection framework

can increase the accuracy of disturbance detection for large-scale, natural disturbances with variations in severity. Our results also indicate that the stacked generalization framework has considerable flexibility when used to assess past disturbance in regions characterized by low- to moderate-severity disturbance regimes. Although our training data principally consisted of forest harvesting events, the ensemble classifier mapped gypsy moth defoliation with comparable accuracy to that of a change-detection methodology developed to target a specific disturbance agent. Our results also indicate that disturbance maps produced using the stacked generalization framework could be fine-tuned at specific years or across a specific sub-region to capture large-scale, natural disturbances without inflating the commission errors across the remainder of the time-series.

CHAPTER TWO:

ABOVEGROUND BIOMASS ESTIMATION IN A HETEROGENOUS, MIXED-SPECIES FOREST UTILIZING LANDSAT-DERIVED DISTURBANCE HISTORY

Deriving regional estimates of aboveground biomass (AGB) from remotely sensed data is important for supporting carbon accounting initiatives and monitoring the status of ecosystem services. Recent research has found that including measures of disturbance and recovery processes, derived from the analysis of time-series satellite imagery, can improve the accuracy of spatially modeled AGB estimates. However, these studies have largely been conducted in regions with relatively homogenous forest composition and structure shaped by stand-replacing disturbance regimes. In this analysis, Landsat satellite imagery from 1984-2017 was used to develop disturbance history maps across the forested area of the New England states and the Canadian Maritime provinces. The maps were derived using stacked generalization, a technique that fuses the outputs from an ensemble of change-detection outputs from a time-series segmentation algorithm. This process was used to generate a series of metrics characterizing disturbance severity, timing, and post-disturbance recovery for each 30m pixel in the study region over the 33-year Landsat record. Then, using gradient boosting, regional AGB models were developed using spectral, disturbance, and environmental metrics. Finally, we analyzed how adjusting the probability threshold that controls the class-inclusion rates influenced AGB model performance. This enabled our analysis to test if increasing or decreasing the sensitivity of the secondary classifier could improve the predictive ability of the disturbance and recovery metrics and, consequently, AGB model performance. The AGB models developed in this analysis possessed normalized root mean squared errors between 61.5 and 52.4%. Our analysis indicated that incorporating environmental metrics provided the greatest improvements in model

performance. Including the disturbance metrics also improved AGB model performance. Our results indicate that adjusting the probability threshold could produce a 10.7 million ha increase or 7 million ha decrease in the mapped disturbance area. However, the changes in disturbed area had minimal impact on the AGB model performance. We conclude that landscape heterogeneity, as well as the general lack of stand-replacing disturbances, diminishes the predictive utility of disturbance and recovery metrics for modeling AGB. However, incorporating the disturbance metrics still provided worthwhile increases in model performance.

2.1. Introduction

Considerable recent research has focused on improving the accuracy of spatially explicit aboveground biomass (AGB) estimation (Goetz et al., 2009; Powell et al., 2010b; Herold and Johns, 2007) in order to facilitate carbon accounting programs and manage ecosystem services (Pan et al., 2011; Dixon et al., 1994). One outcome of this research has been the recognition that AGB estimates can be improved by including information on past disturbances (Pflugmacher et al., 2014; Frazier et al., 2014).

The United States Geologic Survey's Landsat program has been acquiring satellite imagery of Earth since 1972, with sensors continually being updated over time as technology improves. The resulting imagery has proven invaluable for a broad range of applications in ecological analysis (Cohen and Goward, 2004). For example, Landsat imagery has been used extensively to model AGB, as it enables large-scale estimates to be derived by modeling the relationship between the imagery and forest inventory measurements (Cohen and Goward, 2004; Lu, 2005; Luther et al., 2006). The amount of forest biomass at a given location is a function of the forest's composition, the accrual of biomass through forest growth and development, and the loss of biomass caused by harvesting and natural disturbances. However, using Landsat imagery

to estimate AGB faces two notable limitations. First, the spectral signal from moderate-resolution satellite imagery saturates as forests approach their late-successional stages (Lu et al., 2012). Second, the value of spectral indices of a disturbed forests can return to pre-disturbance values rather quickly (~1.6 – 4.1 years), meaning that the spectral recovery in imagery will occur long before the disturbed system truly recovers (Pickell et al., 2016).

Incorporating accurate information on forest disturbance into AGB estimates can alleviate the limitations associated with Landsat satellite imagery (Pflugmacher et al., 2014). Following the opening of the formerly privatized Landsat archive in 2009, sophisticated change-detection algorithms (e.g LandTrendr, VCT, VeRDET) have been developed to extract disturbance information from dense time-series of imagery (Kennedy et al., 2010; Hughes, 2014; Huang et al., 2010). Spatially-explicit estimates of disturbance and recovery (DR) processes can be derived from these algorithms to characterize the pre-disturbance state, produce metrics of the disturbance event (eg. duration, severity), and identify spectral trends during recovery (Pflugmacher et al., 2012; Hermosilla et al., 2015a). For example, studies by Pflugmacher et al., (2014) and Frazier et al., (2014) demonstrated that incorporating DR metrics alongside spectral covariates into models of AGB significantly reduced model error and increased explanatory power.

Given that change-detection algorithms have been developed with a specific use-case in mind, they tend to perform better in the specific ecosystems or disturbance regimes for which they were developed (Cohen et al., 2017). To exploit the individual strengths of each change detection algorithm and reduce the errors of omission and commission for disturbance detection, stacked generalization approaches have been developed (Wolpert, 1992). These approaches are based on an ensemble modeling technique where the outputs of individual change-detection

models, referred to as base-learners, are synthesized with the use of a secondary classifier, such as random forests (Breiman, 2001; Healey et al., 2018). Multi-spectral ensemble model approaches, which use a single base-learner algorithm run using different spectral bands or indices, have been shown to improve the accuracy of disturbance detection by making use of the subtle spectral variations associated with various disturbance severities and agents (Cohen et al., 2018). Although ensemble modeling approaches have improved change-detection capabilities, the ability to adjust the rate at which disturbances are mapped, by modifying the class-inclusion rates, has not been explored. Increasing or decreasing the sensitivity of the ensemble model could improve the predictive strength of DR metrics, derived from the disturbance history maps.

In this analysis, we aim to determine the extent to which the contemporary amount and distribution of forest biomass can be determined by characterizing the disturbance history. Our objectives are 1) quantify to what extent Landsat-derived DR metrics improves the prediction forest AGB; 2) determine the number of years of disturbance information required to meaningfully inform biomass estimation; and 3) explore how modifying the probability threshold associated with the secondary classifier alters the predictive strength of the DR metrics in models of AGB. This study expands upon existing research by attempting to develop broad-scale AGB estimate across a structurally complex, mixed-species forested region with a complex disturbance regime and land-use history

2.2. Methods

2.2.1. Study Area

The New England-Canadian Maritime describes the gradient of forest environments that exist across the study region. Forest inventory data compiled throughout the study area (see below) recorded 116 tree species. The forests stand within the Canadian Maritime Provinces and

27

northern Maine, USA, are conifer-dominated and principally composed of balsam fir (*Abies*

balsamea (L.) Mill.) red spruce (*Picea rubens* Sarg.), black spruce (*Picea mariana* Mill.), white

spruce (*Picea glauca*), white pine (*Pinus strobus* L.), red maple (*Acer rubrum* L.), and American

beech (*Fagus grandifolia* Ehrh.) (Hennigar et al., 2016). Central Maine marks a transition from

conifer-dominated stands towards mixed-species stand with greater abundances of American

beech (*Fagus grandifolia* Ehrh.), paper birch (*Betula papyrifera* Marsh.), and yellow birch

(*Betula alleghaniensis* Marsh.) (Seymour, 1994; Thompson et al., 2013). Forests of the Southern

New England states are dominated by broadleaf trees including northern red oak (*Quercus rubra*

L.), eastern hemlock (*Tsuga canadensis* L.) and maple species such as red maple and sugar

maple (*Acer saccharum* Marsh.).

2.2.2. Forest field plot data

Recent forest inventory data (2011-2017) were compiled from national and provincial

inventories (Figure 1) as reference data to develop models of AGB. Trees with a diameter at

breast height (DBH) ≥ 10 cm were selected for analysis. The AGB for each tree was estimated

using version 2 of the component ratio method (CRM2.0) (Radtke et al., 2017). The CRM2.0

was developed for the eastern United States; it reduces the bias and error of the allometric

equations in the original CRM equations (Goodale et al., 2002). The CRM2.0 methodology

requires height measurements for each tree (which was not available) to estimate AGB. To

impute missing tree heights, a height-diameter model was constructed using gradient boosting

models (Friedman, 2001), an ensemble machine learning technique, using the XGBoost package

(Chen et al., 2015). DBH, species, elevation, slope, and climate site index (Jiang et al., 2014)

were used as covariates in the height-diameter model. The plot-level data were used to estimate

biomass densities (Mg ha^{-1}). Table 1 contains a summary of the AGB density estimates for each inventory dataset.

Figure 2.1. The study area and distribution of the forest inventory plots used in this analysis.

Table 2.1. Summary statistics describing the national and provincial permanent sample plot forest inventories included in the analysis.

Dataset	Measurement years	# plot locations	# of plot measurements	AGB mean (Mg ha^{-1})	AGB median (Mg ha^{-1})	AGB SD (Mg ha^{-1})
FIA	2011-2016	5,180	5,787	101.0	91.7	67.4
NB	2017	1,355	1,355	84.7	72.4	79.3
PEI	2011-2017	604	1,311	156.5	150.6	65.7
NS	2011 - 2016	3,039	3,608	69.0	60.0	56.9
Totals	-	10,178	12,061	-	-	-

* Dataset descriptions

FIA - United States Forest Service's Forest Inventory and Analysis phase II plots

NB - New Brunswick provincial sample plots

PEI - Prince Edward Island provincial sample plots

NS - Nova Scotia Department of Natural Resources provincial sample plots

2.2.3. Image processing

Landsat Collection Tier 1 Landsat Thematic Mapper (TM), Enhanced Thematic Mapper plus (ETM+), and Operational Land Imager (OLI) scenes collected from 1984-2017 with less than 85% cloud-cover were collated for image processing. The United States Geologic Survey Earth Resources and Science Center requires all Landsat Collection Tier 1 scenes to possess a radial root mean square error (RMSE) of ≤ 12m to ensure a sufficient spatial accuracy for time-series analysis. Surface reflectance values were calculated using the Landscape Ecosystem Disturbance Adaptive Processing System algorithm (Masek et al., 2006) and the Landsat 8 Surface Reflectance Code (Vermote et al., 2016). Cross-sensor harmonization of the TM/ETM+ and OLI sensors was performed using the coefficients from Roy et al. (2016). Cloud and shadow contaminated pixels were removed using the Fmask algorithm (Zhu and Woodcock, 2012). An additional round of noise removal was performed using the temporal dark-outlier method (Housman et al., 2015) to remove the cloud shadows missed by the Fmask algorithm. To account for phenological variability across the study area, all pixel observation were constrained to a 38-day (+/-) window around the peak-of-the growing season, derived via Moderate Resolution

Imaging Spectroradiometer (MODIS) observations. The pre-processed time-series of images were then composited into annual images using the medoid method (Flood, 2013). Preliminary tests indicated that smoothing the imagery (e.g. applying a 3x3 focal mean filter) to account for plot design and plot geolocation inaccuracies did not improve the subsequent models of AGB. Image processing was conducted in the Google Earth Engine cloud-computing environment (Gorelick et al., 2017).

2.2.4. Spectral predictors

In addition to the six spectral bands common to the TM, ETM+, and OLI sensors, eight additional spectral metrics were computed as covariates for AGB model development. The three primary components of the Tasseled Cap (TC) transformation, i.e., brightness, greenness, and wetness, were computed using the coefficients from Crist and Cicone (1984). Two indices were computed from the TC components: angle (Powell et al., 2014) and distance (Duane et al., 2010). Three additional vegetation indices, the Normalized Difference Vegetation Index (NDVI), the Normalized Difference Moisture Index (NDMI), and the Normalized Burn Ratio (NBR) were computed (Rouse et al., 1974; Wilson and Sader, 2002; Key and Benson, 1999). The formulations for the Tasseled Cap transformations and spectral indices and can be found in the Appendix.

2.2.5. Environmental predictors

Topographic and climatological predictors (collectively referred to as environmental predictors) were developed to help quantify the spatial variation across the landscape. Elevation, slope, and aspect were derived from the Shuttle Radar Topography Mission (SRTM) 1 arc-second DEM. Climatological metrics were produced using the Daymet V3 daily weather and climatological dataset (Thornton et al., 2016). The daily datasets from 1980-2017 were masked

to isolate observations within the growing season using a MODIS-derived phenology dataset. The datasets were then annually aggregated to estimate the annual precipitation, growing degree days, average maximum, mean, and minimum temperature for each 1-km grid cell. The per-pixel average of the yearly metrics were used as predictors. A 1km estimate of climate site index was used to provide a spatial estimate of forest productivity (Jiang et al., 2014).

2.2.6. Disturbance detection with generalized stacking

2.2.6.1. TimeSync reference data

A historical record of forest disturbance was produced using the TimeSync image interpretation program (Cohen et al., 2010). TimeSync allows an interpreter to visualize the time-series of spectral values, derived from the Landsat archive, over a given pixel (hereafter referred to as a plot). TimeSync interfaces with Google Earth to provide access to historical high-resolution imagery to aid the interpreter's decisions. The interpreter subdivides the time-series into a sequence of linear segments that describe the change processes (e.g. no change, forest harvest, vegetation recovery, etc.). The start and end dates for a given segment were used to record information about land use and land cover. To improve the consistency of interpretations across different interpreters, all plots were re-assessed by a single interpreter.

The reference dataset consisted of 4233 TimeSync plots interpreted for change type and timing over the 1984-2016 time-period. Of those plots, 797 were never forested and were thus excluded from subsequent analysis. The remaining plots were again filtered to exclude any segment that were not initially forested as indicated by the classification at the segments start date. This allowed for disturbances that produced a change in forest land use (e.g. road construction) to be incorporated into the subsequent analysis. The final database consisted of 105,339 annual observations that were labeled as either disturbed or not-disturbed based on the

change process indicated by the interpreter. These annual observation serve as the training and validation data for developing the stacked generalization ensemble classifier.

2.2.6.2. Training the base-learners

The initial change detection process was conducted using the GEE-implementation of the LandTrendr temporal segmentation algorithm (Kennedy et al., 2010; Kennedy et al., 2018). LandTrendr outputs serve as the initial models, referred to as "base-learners", whose outputs were then combined by the secondary classifier in the stacked generalization framework (Wolpert, 1992). LandTrendr identifies the temporal trajectory of each pixel's spectral values within the time series through an iterative, regression-based curve-fitting process (see Kennedy et al. (2010) for details). A series of qualitative tests were used to determine the optimum parameterization for the LandTrendr algorithm across the range of forest conditions present within the study area. The testing indicated that increasing the sensitivity of the algorithm, thus increasing the commission rate for individual LandTrendr models, improved the performance of the random forests classifier during the stacking process (as detailed below). Following the suggestion of Cohen et al. (2018), the segment filtering portion of the LandTrendr algorithm was disabled, allowing all segments with a negative slope to be labeled as a disturbance. In total, 14 LandTrendr models were developed, one for each of the spectral predictors used in this analysis.

Each base-learner in the ensemble of LandTrendr models was used to calculate three metrics: the current spectral value as fitted by the LandTrendr algorithm, the difference in the fitted spectral values between the given year and the year prior (i.e., the magnitude of the spectral change), and a binary predictor (disturbed/not disturbed) based upon the slope of the fitted segment. This process resulted in 42 covariates produced by the ensemble of base-learners.

2.2.6.3. Stacking with the ensemble classifier

Combining the outputs of the base-learners was performed using a secondary classifier, here the random forests (RF) algorithm with 500 trees (Breiman, 2001). The modeling dataset consisted of all of the records in the TimeSync reference database and the corresponding base learner outputs, for each year of each plot. The dependent variable was the binary label, indicating "disturbed" or "not-disturbed", assigned by the TimeSync interpreter. The out-of-bag (OOB) error rate was used to assess the RF classification performance. During the construction of each tree in the ensemble, approximately one-third of the original data were withheld in the bootstrapped aggregation ("bagging") procedure (Breiman, 1996). Each sample within the original dataset received an OOB error estimate based on the average of instances in which it was withheld during bagging. For the purpose of the binary classification problem at hand, the OOB error, for a given sample, reflects the predicted probability that a given observation is a member of either class.

The classes in the modeling dataset were highly imbalanced, with only ~2.6% of the records in the TimeSync reference dataset labeled as disturbances. Typically, one would either sub-sample the majority class or generate synthetic minority class data using a technique such as synthetic minority over-sampling technique (Chawla et al., 2002) to balance the training dataset. However, because the sensitivity of the base-learners was increased to capture subtle disturbance events (e.g. low severity wind disturbances, partial harvesting), the application of these techniques produced a significant increase in the commission error rate. Training across the entire dataset generated a more conservative RF classifier with fewer commission errors. The imbalance in the errors of omission and commission, produced by the RF over an imbalanced dataset, were accounted for by adjusting the probability threshold that determined class

membership. The threshold was set so as to balance the errors of omission (i.e., the number of disturbed TimeSync observations incorrectly classified as not disturbed divided by the total number of correctly classified disturbance in the TimeSync dataset) and errors of commission (i.e., the number of non-disturbed TimeSync observations incorrectly classified as a true disturbance divided by the total number of correctly classified disturbance in the TimeSync dataset). The errors reported here were derived from the RF classifier that was used to produce wall-to-wall disturbance maps in Google Earth Engine.

2.2.7. Disturbance and Recovery Metrics

After applying the trained secondary RF classifier across the study region, disturbance and recovery metrics (DR) were computed over the pixels classified as having been disturbed. The DR metrics were calculated using the fitted spectral values produced by the base-learners. Three classes of metrics were calculated for each of the fourteen spectral indices: 1) the magnitude of the greatest severity disturbance (MGSD), 2) the magnitude of the most recent disturbance (MMRD), and 3) the greatest magnitude disturbance percent recovery (PRGMD). The MGMD metrics quantifies the difference between the pre-disturbance and post-disturbance spectral values of a given spectral index, for the disturbance event with the greatest magnitude difference in a given pixel's time-series. The MMRD represents the difference between the pre-disturbance and post-disturbance spectral values for the most recent disturbance event. PRGMD represents the percent recovery of the current year's spectral values to the pre- MGSD spectral values. In addition to these 42 metrics, the number of years since the MGSD and the MMRD were calculated and outputted for each pixel.

2.2.8. AGB model development and evaluation

Models of AGB were developed using the gradient boosting machines with the XGBoost package in R (Chen et al., 2015). Gradient boosting is an ensemble machine learning technique where weak learners, here decision trees, are sequentially grown while optimizing a differentiable loss function (e.g. the mean squared error of the residuals) (Friedman, 2001). The loss function is paired with a regularization term to penalize the complexity of the decision trees and prevent overfitting (Chen et al., 2015; Friedman, 2001). Dropouts, the random removal of decision trees from the boosting ensemble, were incorporated into the boosting process to avoid adding overly-specialized decision trees in later rounds of boosting (Vinayak and Gilad-Bachrach, 2015). Four gradient boosting models were developed using different sets of covariates: Model 1 consisted of only the spectral covariates; Model 2 consisted of the spectral and environmental metrics; Model 3 consisted of the spectral and disturbance metrics; and Model 4 consisted of the spectral, environmental, and disturbance metrics. The structure of these four models allowed us to assess the benefit of adding environmental and disturbance metrics, singly, and in combination, to the standard models based on spectral covariates alone. The permuted variable importance of the predictors in each modeling dataset was assessed using the R package VSURF (Genuer et al., 2015). Variables with low permuted-importance values were removed to improve model parsimony. The gradient boosting models were evaluated using nested cross-validation. The model's hyper parameters were selected using an inner five-fold cross-validation procedure with the outer ten-fold cross-validation procedure providing an estimate of the model's generalized prediction error (Cawley and Talbot, 2010; Pryzant et al., 2017).

The gradient boosting models were evaluated using the outer-fold predictions from the nested cross-fold validation. Model performance was assessed using mean absolute error (MAE), root mean squared error (RMSE), normalized RMSE (nRMSE), bias (computed as the mean difference of the observed and predicted values), and R^2.

2.2.9. Assessing the influence of time-series length on biomass predictions

To assess how time-series length influenced the ability to predict AGB, the DR metrics were iteratively recalculated, with the temporal depth of the satellite image time-series (D) used to calculate the DR metrics increased by one year with each iteration. Two series of RF models were developed using each set of the DR metrics. One series of RF models incorporate the spectral and recalculated DR metrics, the other also incorporated the environmental metrics. The changes in the predictive utility of the DR metrics were evaluated by changes in the nRMSE and R^2 as D was varied. To account for potential changes in the predictive strength of individual DR metrics as the depth of the time-series was increased, all DR metrics were used during the modeling. To serve as a benchmark, when D=0, the DR metrics were removed from the modeling datasets.

2.2.10. Applications of adjusting the secondary classifier's probability threshold

Although mapping disturbance with a balanced error rate may be suitable for general landscape analysis, increasing the rate at which low-severity disturbances are mapped (thus increasing commission error) or decreasing the rate at that they are mapped (thus increasing the omission error) may improve the accuracy of AGB models. The probability threshold of the secondary classifier was allowed to vary by ±25% in increments of 5% around the probability threshold that balanced the errors of omission and the errors of commission (hereafter referred to as the balanced probability threshold). Increasing the threshold increases the number of

observations included in the not-disturbed class, thus decreasing the errors of omission. RF models were developed using the same set of predictors as in Model 4 but with the DR metrics recalculated for each variation of the probability threshold. The RF models with 2500 trees were developed in R using the "ranger" package (Wright and Ziegler, 2015).

Regression-based equivalence tests were used to compare the predictions produced by the RF using DR metrics calculated with modified probability thresholds (collectively referred to as RF_{MOD}) to the predictors of a baseline RF model (RF_{BAL}) where DR metrics were calculated using a balanced probability threshold (Robinson et al., 2005). Equivalence testing inverts the traditional hypothesis test by assuming the dissimilarity of the populations being considered. A rejection of the null hypothesis indicates that the two populations are similar enough as to be considered equivalent (see Robinson and Froese, 2004). Regression-based equivalence tests expand upon the standard two one-sided t-test procedure by assessing the slope and intercept terms of a least-squares regression model fit between the two populations. Equivalence of the intercept term indicates a lack of bias and equivalence of the slope term indicates the degree of proportionality or association between the predictions (Robinson et al., 2005; Hudak et al., 2012). The regression-based equivalence tests were computed using 1000 iterations of non-parametric bootstrap with significance assessed at $\alpha = 0.05$. If the proportion of significant results ($p < 0.05$) during the bootstrap exceeds 0.949, the null hypothesis is rejected. The equivalence interval was set at 5%, this value is quite conservative and ensures that significant test results indicates a high degree of similarity.

2.3. Results

2.3.1. Change detection results.

A 27.5% balance of omission and commission errors was achieved applying the secondary RF classifier in Google Earth Engine. The secondary classifier's accuracy varied between partial harvests and clearcuts (Figure 2). Observations in the TimeSync database labeled as clearcuts were correctly classified 77.9% of the time. Observations labeled as partial harvests were classified correctly 44.3% of the time.
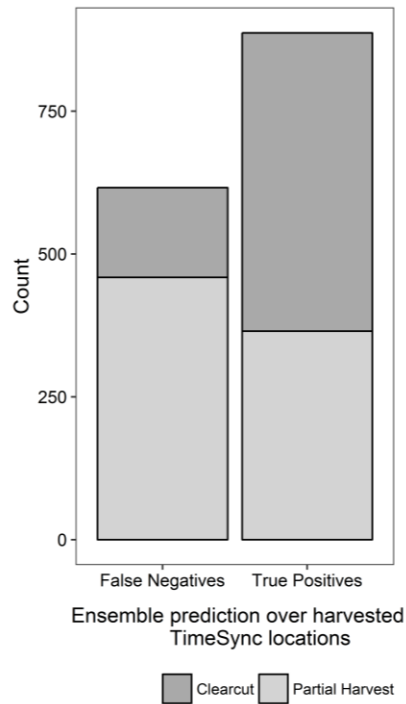
Figure 2.2. The secondary classifier's prediction results over known harvested pixels in the TimeSync database. The true positives and false negatives were stratified by the harvest type.

Table 2.2. Model summary statistics describing the performance gradient boosting models. MAE, RMSE, and bias are expressed in Mg ha^{-1}.

Model	bias	MAE	RMSE	nRMSE	pseudo-R^2
Model 1	1.25	44.15	58.14	61.5%	28.5%
Model 2	1.49	38.49	51.47	54.5%	44.0%
Model 3	0.78	41.06	55.12	58.3%	35.8%
Model 4	1.21	36.38	49.55	52.4%	48.1%

* Dataset descriptions
Model 1 – spectral covariates
Model 2 – spectral and environmental covariates
Model 3 – spectral and disturbance covariates
Model 4 – spectral, environmental, and disturbance covariates

Table 2.3. The ten most important variables in each of the AGB models. Importance was determined using the VSURF permuted importance values.

Rank	Model 1	Model 2	Model 3	Model 4
1	NBR	B2	B2	B2
2	B2	B3	NDMI	NDMI
3	B4	NDMI	NBR	NBR
4	B5	NBR	TCW	GDD
5	B7	GDD	TCA	TCA
6	NDMI	TCW	B5	B3
7	NDVI	B5	B7	Avg. tmin
8	TCG	B7	TCG	TCW
9	TCW	Avg. tmin	B3	CSI
10	TCA	Prcp.	NDVI	Prcp.

Growing Degree Days (Annual GDD); Average minimum temperature (Avg. tmin);
Annual growing season precipitation (Prcp.)

Figure 2.3. Predicted vs. observed aboveground biomass values using the out-of-fold predictions from the gradient boosting models. The dashed line indicates the 1:1 relationship. The relationship between the observed and predicted values is summarized with a least-squares regression curve (solid line).

2.3.2. AGB model performance

A summary of model performance across datasets is provided in Table 2.2. Utilizing spectral metrics (Model 1) alone yielded an R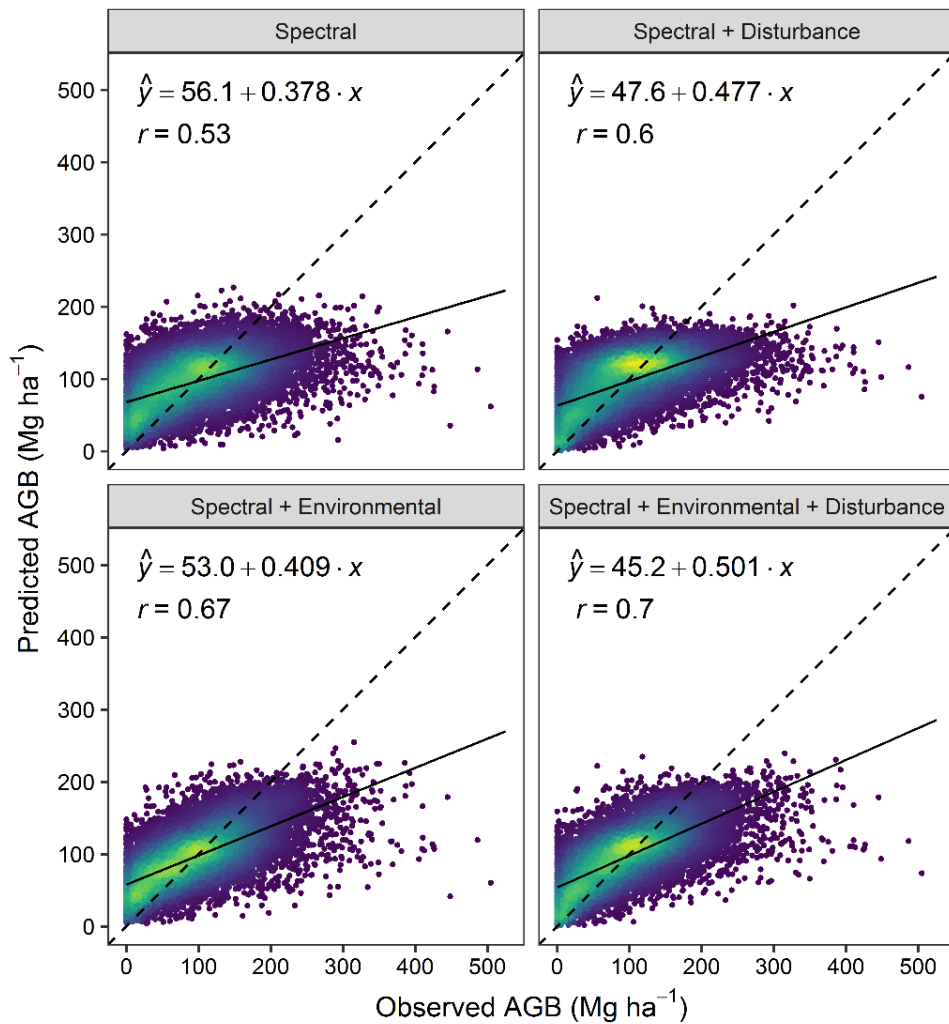MSE = 58.1 Mg ha^{-1} and an R^2 = 0.285. Incorporating environmental covariates (Model 2) improved model performance (RMSE = 51.5 Mg ha^{-1} and R^2 = 0.44) and improved the linear relationship between the predicted AGB values and the observed AGB values (Figure 3). The model developed using the spectral and DR metrics (Model 3), performed better then Model 1, though the gains in performance were not as notable as Model 2 (RMSE = 55.1 Mg ha^{-1} and R^2 = 0.358). The best performing model (Model 4), contained the spectral, environmental, and DR covariates (RMSE = 49.6 Mg ha^{-1} and R^2 = 0.481). This indicates a 1.92 Mg ha^{-1} decrease in RMSE through the inclusion of the DR metrics relative to Model 2.

The ten most important variables for each model, ranked using the permuted importance scores calculated by VSURF, are presented in Table 3. In all four models, green reflectance is ranked as the first or second most important predictor. Although model performance improved by incorporating the DR metrics (Model 3), none of the DR covariates were ranked among the top ten most important variables. The number of growing degree days, average minimum temperature during the growing season, and the annual precipitation were important predictors in both models that included the environmental metrics (Model 2 and 4).

2.3.3. Impact of the time-series length on disturbance and recovery metrics

Incorporating the disturbance metrics initially produced a decrease in model performance, compared to the benchmark models, which excluded the disturbance metrics (Figure 3). The nRMSE and R^2 improved beyond the benchmark models at D=11 when using the spectral and DR metrics and at D=16 when using the spectral, environmental, and DR metrics.

At the full depth of the time-series (D=32), nRMSE decreased by 2.9% and the R^2 increased by 5.4% for the RF model using the spectral and DR metrics. The RF model that included environmental metrics indicated more subtle gains in model performance at the full time-series depth with nRMSE declining by 1.6% and R^2 increasing by 2.7%, although overall model performance was enhanced by the addition of the environmental metrics.

2.3.3. Model performance using adjusted disturbance metrics

Mapping forest area disturbances from 1985-2017 using the balanced probability threshold produced 10.7 million ha of mapped disturbances. Increasing the balanced probability threshold by 25% produced a 7 million ha decrease in total mapped disturbance area, while decreasing the balanced threshold by 25% produced a 10 million ha increase in the mapped disturbance area. However, with the exception of two models, the changes in mapped disturbance area did not influence the RF model's predictions of AGB (Table 2.3). The equivalence testing of the intercept terms indicated that the AGB predictions produced by the RF_{MOD} models were equivalent to the RF_{BAL} predictions and lacked bias. Equivalence testing also indicated that, with the exception of two models, the association of slope terms was equivalent. The RF models that utilized the -20, and -25% modifiers in the construction of the DR metrics had slope terms that were not equivalent to those of the predicted model. Decreasing the balances probability threshold by 20% reduced the RF model's RMSE a 0.86 Mg ha^{-1}, relative to the RF_{BAL} model. Decreasing the balanced probability threshold by 25% further reduced the model error by 0.28 Mg ha^{-1}.
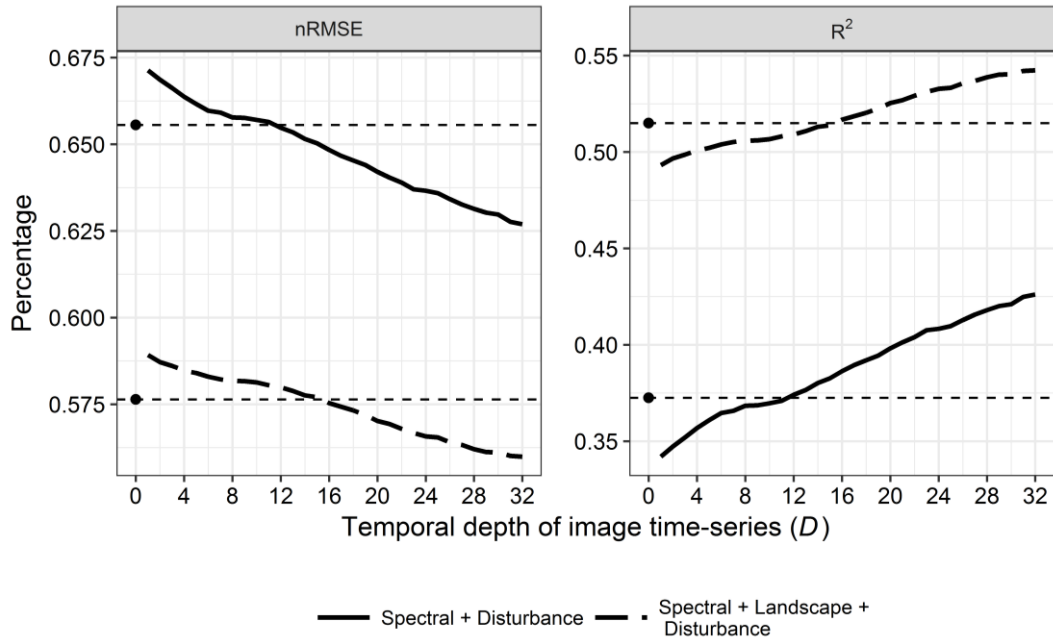
Figure 2.4. Changes in nRMSE and R^2 of random forests models, using two sets of covariates, as the number of images used as the depth of the image time-series (D) is increased. The dots at (*D*=0) indicate the performance of the benchmark models that excluded the DR metrics.

2.4. Discussion

2.4.1 Change detection

Relatively few publications in have utilized stack generalization for the creation of disturbance models. The 27.5% balance of omission and commission errors compares favorably to the results of Healey et al. (2018) whose best secondary classifier possessed a 40% balance of omission and commission errors. The improvements in our results can be attributed to inclusion of the spectral magnitude of the disturbance events alongside the spectral values and binary disturbance labels produced by the base-learners. Cohen et al. (2018) developed a secondary classifier with a 29.6% balance of omission and commission error. Our results indicate that stacked generalization, using an ensemble of LandTrendr models as the base-learners, is an effective technique for capturing forest disturbances. However, capturing low severity disturbance events still poses a challenge

for change-detection, as evidenced by the greater omission rate for partial harvesting compared to clearcutting.

Table 2.4. Results from the regression-based equivalence tests. The AGB predictions of random forests (RF) models developed using disturbance and recovery (DR) metrics generated using the balanced probability thres4hold were compared with predictions derived from models using DR metrics generated using a modified probability thresholds. If the 95% confidence intervals for the slope ($C_{\beta_0}^-$,$C_{\beta_0}^+$) and intercept ($C_{\beta_1}^-$,$C_{\beta_1}^+$) terms are contained within their corresponding equivalence margins ($E_{\beta_0}^-$, $E_{\beta_0}^+$) and (E_1^-, $E_{\beta_1}^+$) the null hypothesis is rejected.

Threshold Modifier	$C_{\beta_0}^-$	$C_{\beta_0}^+$	$E_{\beta_0}^-$	$E_{\beta_0}^+$	H_0: β_0=0	$C_{\beta_1}^-$	$C_{\beta_1}^+$	$E_{\beta_1}^-$	$E_{\beta_1}^+$	H_0: β_1=1
-25 %	95.537	95.937	91.138	100.731	Reject	0.941	0.949	0.950	1.050	Not rejected
-20 %	95.571	95.913	91.109	100.699	Reject	0.947	0.954	0.950	1.050	Not rejected
-15 %	95.562	95.903	91.082	100.670	Reject	0.955	0.961	0.950	1.050	Reject
-10 %	95.587	95.879	91.049	100.633	Reject	0.967	0.973	0.950	1.050	Reject
-5 %	95.634	95.835	90.992	100.571	Reject	0.984	0.988	0.950	1.050	Reject
5 %	95.641	95.824	90.920	100.491	Reject	1.001	1.005	0.950	1.050	Reject
10 %	95.597	95.870	90.871	100.436	Reject	1.003	1.009	0.950	1.050	Reject
15 %	95.556	95.901	90.814	100.374	Reject	1.005	1.012	0.950	1.050	Reject
20 %	95.537	95.945	90.796	100.353	Reject	1.005	1.014	0.950	1.050	Reject
25 %	95.491	95.970	90.804	100.363	Reject	1.008	1.019	0.950	1.050	Reject

2.4.2. AGB model assessment

The AGB models developed in our analysis possessed nRMSEs ranging from 52.4 to 61.5%. These results are similar to, or better than, those previously reported in the literature. Using Landsat spectral metrics, Deo et al. (2017) developed models of AGB with nRMSE between 50.8 - 69.8% in northern Minnesota. Frazier et al. (2014) developed an AGB models, with nRMSE of between 56-71%, using Landsat spectral metrics and LandTrendr-derived DR

metrics in unmanaged conifer stands in British Columbia. Powell et al. (2014) developed AGB

models for individual Landsat scenes using Landsat spectral metrics, climatological metrics, and

topographic metrics throughout the conterminous United States with nRMSEs between 32.5-

86.1%; the AGB models developed using scenes in the northern portion of the Forest Inventory

Analysis's Northeastern region possessed nRMSE between 32.5-60.9%. Given the spatially-

expansive and heterogeneous nature of the forests in our study region, as well as the expansive

validation dataset used to develop the models, our results are comparable to or better than prior,

single-date estimates of AGB using Landsat spectral data.

2.4.3. Influential AGB covariates

The ranking of the most influential predictors for each model provides insight into future

modeling needs. The spectral metrics, as a group, comprised the majority of most influential

predictors in each model. Many of the spectral metrics that ranked among the most important

predictors (e.g. TCW, TCA, NDMI) have previously been identified as being strongly correlated

with AGB and forest canopy density (Frazier et al., 2014; Pflugmacher et al., 2012).

Interestingly, green spectral reflectance (B2) was ranked highly (the most important metric in

Models 2, 3, and 4) in all models. The green reflectance of a forest is strongly influenced by the

presence of softwood species. Green reflectance could, effectively, be serving as a proxy for

proportion of the AGB represented by conifer species. The environmental metrics that appeared

in both Models 2 and 4 (i.e., average minimum temperature, annual precipitation, and growing

degree days) all provide an estimate of site quality. Despite improving model performance, none

of the DR metrics were included in the top ten influence metrics for either models in which they

were included (Model 3 and 4).

2.4.4. Impact of time-series length on AGB prediction

The initial decrease in the performance of the RF models, relative e to the baseline

models (D=0), which excluded the DR metrics, can be explained by the random-feature selection

process used to grow each tree in a RF ensemble (Breiman, 2001). Although RF is normally

quite robust to the inclusion of uncorrelated features, the number of DR metrics calculated in this

study (42) was considerably larger than the number of spectral metrics (14) or environmental

metrics (8). When the full depth of the time-series was considered (D=32), 24.2% of the forest

inventory plots had experienced a disturbance according to the classifier output. At low values of

D, the number of plots that had experienced disturbance became negligible, thus reducing the

metrics' predictive strength and in turn reducing the models' ability to correctly partition the

data. Given the common-place use of RF or similar tree-based machine-learning algorithms to

predict AGB (e.g. Powell et al., 2010a; Powell et al., 2014; Frazier et al., 2014; Pflugmacher et

al., 2012), future studies should utilize robust feature selection in locations where the temporal

depth of the Landsat archive is limited (see Wulder et al., 2016). Increasing the time-series depth

by incorporating Landsat Multispectral Scanner (MSS, available back to 1972) data could

potentially further improve model performance by extending the temporal depth of landscape

disturbance an additional 14 years. Our results did not indicate that the rate of change in nRMSE

and R^2 reached a plateau as was observed in Pflugmacher et al. (2014).

2.4.5. The influence of landscape heterogeneity on the disturbance metrics

Our results indicate smaller increases in model performance when DR metrics are

included in the modeling dataset than those found in previous studies. Our results showed that

including DR metrics alongside spectral metrics improved nRMSE by 3.2% compared to the

model developed using solely spectral metrics. In contrast, Pflugmacher et al. (2012) found a

16% improvement in nRMSE in models developed using Landsat-derived spectral and DR metrics over models developed using single-date Landsat spectral values. Similarly, the best model developed in Frazier et al. (2014) showed the 13% improvement in nRMSE when models incorporated Landsat spectral and DR metrics when compared to a model created using single-date spectral predictors. Pflugmacher et al. (2014) compared models developed relating light detection and ranging (LiDAR) based estimates of AGB to Landsat spectral and DR metrics, reporting an 8.2% decrease in nRMSE when DR metrics were included compared to models utilizing solely single-date spectral metrics.

Notably, previous studies that have explored AGB models developed using DR metrics have been conducted in relatively homogenous landscape (i.e., forests with simpler species composition and structure) and under disturbance regimes dominated by high-severity, stand-replacing disturbances (Frazier et al., 2014; Matasci et al., 2018; Pflugmacher et al., 2014). By comparison, the NEMP region encompasses a much broader range of forest types and structures, and it exhibits a general lack of stand replacing disturbances. For example, the New England states have seen a transition from clear-cutting towards lower-severity, partial harvesting practices (Canham et al., 2013; Belair and Ducey, 2018). In addition, unlike many regions of the US, the natural disturbance regime in this region is dominated by low-severity disturbances that only remove a portion of the forest canopy; stand replacing natural disturbances are quite rare (Lorimer, 1977; Seymour et al., 2002). We suspect that this lack of stand-replacing disturbances weakens the utility of DR metrics for predicting AGB in these systems. Our results also suggest that complex interactions that influence the recovery following disturbance process may not be adequately captured by the spectral magnitude of the disturbance event alone. Including additional complementary datasets such as soils maps, forest composition maps, and additional

information regarding harvesting practices (e.g. reclassification classification of the disturbance map) may improve the predictive strength of the DR metrics.

2.4.6. Using the probability threshold to adjust AGB predictions

As proposed by Healey et al., (2018), our results suggest that variations in the probability threshold used to map disturbances can produce large variations in the extent of the mapped disturbance areas. Here, our tests indicated that adjusting the probability threshold ± 25% on the RF secondary classifier produced a 94% increase or a 66% decrease in the mapped disturbance area. If sufficient information were available, it could be possible to dynamically adjust the probability threshold on a year-to-year basis to capture the changes in annual disturbances produced by documented large-scale natural disturbances (e.g., hurricanes or ice storms) or to account for known changes in public policy (e.g., legislation that places limitations on clear-cutting such as the 1989 Maine Forest Practices Act). However, it is not clear if this would benefit efforts for regional AGB modeling. Regression-based equivalence tests indicated that decreasing the probability threshold by 20% and 25% yield nominal, though significant changes in model predictions. This finding supports our conclusion that, for large heterogeneous landscapes, additional information is necessary to more completely understand the landscape-level relationships between forest composition, forest structure, past disturbance, and AGB density. Our tests indicate that low predictive strength of the DR metrics cannot be attributed to the omission of disturbance events over forest plots. Our tests greatly increased the mapped rate of disturbance across the landscape and only yielded a 0.9 and 1.2% decrease in nRMSE compared to the baseline-model.

2.4.7. Implications for future regional biomass modeling

DR metrics may improve AGB prediction accuracy, relative to the predictions of models developed using single-date Landsat spectral metrics; however, the benefits of the DR metrics are diminished because of region's forest and landscape heterogeneity, as well as a disturbance regime dominated by low-severity events. Our results suggest that future AGB modeling efforts should combine data from various sources to more accurately quantify forest structure across ecological and climatological gradients and to overcome the challenges associated with spectral saturation in high biomass locations (Lu et al., 2012). LiDAR has been demonstrated as an effective tool for quantifying forest structure (Hudak et al., 2012; Lefsky et al., 2002) and has been combined with moderate-resolution spectral imagery and Landsat DR metrics to improve AGB model performance (Zald et al., 2016; Pflugmacher et al., 2014; Deo et al., 2017). Spaceborne- LiDAR acquired by the GLAS sensor carried onboard the ICESat platform has been used to develop AGB estimates for large regions (Boudreau et al., 2008; Mitchard et al., 2012). The Global Ecosystem Dynamics Investigation module and ICESat-2 spaceborne-LiDAR sensors, which are planned to be launched in late-2018, will provide new opportunities to explore how these data can be incorporated into AGB models.

2.5. Conclusion

This analysis developed AGB models for a structurally complex, mixed-species forested region with a complex disturbance regime using Landsat spectral and DR metrics as well as environmental predictors. Our modeling results indicate that the environmental metrics were better predictors of AGB than were the DR metrics, a finding that contrasts with previous results in the literature. Our results suggest that the biomass implications of a past disturbance event cannot be derived solely using the timing, spectral magnitude, and spectral recovery percentage.

Given these metrics were among the most informative in previous analyses, it does not appear that simply calculating different disturbance metrics would overcome the limitations identified in this study. Prior studies had have been performed predominantly in landscapes with relatively homogenous species composition and disturbance regimes dominated high-severity, stand replacing disturbances. We therefore conclude that landscape and disturbance heterogeneity negatively impacts the predictive utility of DR metrics.

Our results suggest that stacked generalization is a robust and flexible change-detection framework. Despite the heterogeneity of the landscape, it was possible to develop a secondary classifier with a 27.5% balance of the omission and commission errors, comparable to previous results in the literature. Additionally, using a secondary classifier allows for the rate of mapped disturbance to be adjusted probabilistically through modifications to the class inclusion rate meaning the technique can be adapted to suit diverse research objectives.

Future research should emphasize incorporating additional sources of data to increase the accuracy of AGB models. In regions with heterogeneous forest composition, significant improvements in model performance will require more direct estimates of forest structural characteristics. Upcoming spaceborne-LiDAR sensors will provide an opportunity to incorporate direct measurements of forest structure into biomass models.

REFERENCES

Belair, E. P., M. J. Ducey, 2018: Patterns in Forest Harvesting in New England and New York: Using FIA Data to Evaluate Silvicultural Outcomes. *J.For.,* **116**, 273-282.

Boudreau, J., R. F. Nelson, H. A. Margolis, A. Beaudoin, L. Guindon, and D. S. Kimes, 2008: Regional aboveground forest biomass using airborne and spaceborne LiDAR in Québec. *Remote Sensing of Environment,* **112**, 3876-3890, doi:https://doi-org.prxy4.ursus.maine.edu/10.1016/j.rse.2008.06.003.

Breiman, L., 1996: Bagging predictors. *Mach.Learning,* **24**, 123-140.

Breiman, L., 2001: Random Forests. *Mach.Learning,* **45**, 5-32, doi:10.1023/A:1010933404324".

Canham, C. D., N. Rogers, and T. Buchholz, 2013: Regional variation in forest harvest regimes in the northeastern United States. *Ecol.Appl.,* **23**, 515-522.

Cawley, G. C., N. L. Talbot, 2010: On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research,* **11**, 2079-2107.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, 2002: SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research,* **16**, 321-357.

Chen, T., T. He, and M. Benesty, 2015: Xgboost: extreme gradient boosting. *R package version 0.4-2,* , 1-4.

Cohen, W. B., S. N. Goward, 2004: Landsat's role in ecological applications of remote sensing. *AIBS Bulletin,* **54**, 535-545.

Cohen, W. B., S. P. Healey, Z. Yang, S. V. Stehman, C. K. Brewer, E. B. Brooks, N. Gorelick, C. Huang, M. J. Hughes, and R. E. Kennedy, 2017: How Similar Are Forest Disturbance Maps Derived from Different Landsat Time Series Algorithms? *Forests,* **8**, 98.

Cohen, W. B., Z. Yang, S. P. Healey, R. E. Kennedy, and N. Gorelick, 2018: A LandTrendr multispectral ensemble for forest disturbance detection. *Remote Sens.Environ.,* **205**, 131-140.

Cohen, W. B., Z. Yang, and R. Kennedy, 2010: Detecting trends in forest disturbance and recovery using yearly Landsat time series: 2. TimeSync — Tools for calibration and validation. *Remote Sens.Environ.,* **114**, 2911-2924, doi:http://dx.doi.org/10.1016/j.rse.2010.07.010.

Coppin, P. R., M. E. Bauer, 1996: Digital change detection in forest ecosystems with remote sensing imagery. *Remote Sens.Rev.,* **13**, 207-234, doi:10.1080/02757259609532305.

Crist, E. P., 1985: A TM Tasseled Cap equivalent transformation for reflectance factor data. *Remote Sens.Environ.,* **17**, 301-306, doi:http://dx.doi.org.prxy4.ursus.maine.edu/10.1016/0034-4257(85)90102-6.

De Groot, R. S., R. Alkemade, L. Braat, L. Hein, and L. Willemen, 2010: Challenges in integrating the concept of ecosystem services and values in landscape planning, management and decision making. *Ecological complexity,* **7**, 260-272.

Deo, R. K., M. B. Russell, G. M. Domke, C. W. Woodall, M. J. Falkowski, and W. B. Cohen, 2017: Using Landsat time-series and LiDAR to inform aboveground forest biomass baselines in northern Minnesota, USA. *Canadian Journal of Remote Sensing,* **43**, 28-47.

Dixon, R. K., S. Brown, R. A. Houghton, A. M. Solomon, M. C. Trexler, and J. Wisniewski, 1994: Carbon pools and flux of global forest ecosystems. *Science,* **263**, 185.

Duane, M. V., W. B. Cohen, J. L. Campbell, T. Hudiburg, D. P. Turner, and D. L. Weyermann, 2010: Implications of Alternative Field-Sampling Designs on Landsat-Based Mapping of Stand Age and Carbon Stocks in Oregon Forests. *For.Sci.,* **56**, 405-416.

E. P. Crist, R. C. Cicone, 1984: A Physically-Based Transformation of Thematic Mapper Data---The TM Tasseled Cap. *IEEE Transactions on Geoscience and Remote Sensing,* **GE-22**, 256-263, doi:10.1109/TGRS.1984.350619.

Fisette, T., P. Rollin, Z. Aly, L. Campbell, B. Daneshfar, P. Filyer, A. Smith, A. Davidson, J. Shang, and I. Jarvis, 2013: AAFC annual crop inventory. *Proc. Agro-Geoinformatics (Agro-Geoinformatics), 2013 Second International Conference on,* IEEE, 270-274.

Flood, N., 2013: Seasonal composite Landsat TM/ETM images using the Medoid (a multi-dimensional median). *Remote Sensing,* **5**, 6481-6500.

Fraver, S., A. S. White, and R. S. Seymour, 2009: Natural disturbance in an old-growth landscape of northern Maine, USA. *J.Ecol.,* **97**, 289-298, doi:10.1111/j.1365-2745.2008.01474.x.

Frazier, R. J., N. C. Coops, M. A. Wulder, and R. Kennedy, 2014: Characterization of aboveground biomass in an unmanaged boreal forest using Landsat temporal segmentation metrics. *ISPRS Journal of Photogrammetry and Remote Sensing,* **92**, 137-146, doi:http://dx.doi.org/10.1016/j.isprsjprs.2014.03.003.

Friedman, J. H., 2001: Greedy function approximation: a gradient boosting machine. *Annals of statistics,* , 1189-1232.

Ganguly, S., M. A. Friedl, B. Tan, X. Zhang, and M. Verma, 2010: Land surface phenology from MODIS: Characterization of the Collection 5 global land cover dynamics product. *Remote Sensing of Environment,* **114**, 1805-1816, doi:https://doi-org.prxy4.ursus.maine.edu/10.1016/j.rse.2010.04.005.

Genuer, R., J. Poggi, and C. Tuleau-Malot, 2015: VSURF: an R package for variable selection using random forests. *The R Journal,* **7**, 19-33.

Goetz, S. J., A. Baccini, N. T. Laporte, T. Johns, W. Walker, J. Kellndorfer, R. A. Houghton, and M. Sun, 2009: Mapping and monitoring carbon stocks with satellite observations: a comparison of methods. *Carbon balance and management,* **4**, 2.

Goodale, C. L., M. J. Apps, R. A. Birdsey, C. B. Field, L. S. Heath, R. A. Houghton, J. C. Jenkins, G. H. Kohlmaier, W. Kurz, and S. Liu, 2002: Forest carbon sinks in the Northern Hemisphere. *Ecol.Appl.,* **12**, 891-899.

Gorelick, N., M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore, 2017: Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sens.Environ.,* **202**, 18-27.

Hansen, M. C., P. V. Potapov, R. Moore, M. Hancher, S. A. Turubanova, A. Tyukavina, D. Thau, S. V. Stehman, S. J. Goetz, T. R. Loveland, A. Kommareddy, A. Egorov, L. Chini, C. O. Justice, and J. R. G. Townshend, 2013: High-Resolution Global Maps of 21st-Century Forest Cover Change. *Science,* **342**, 850-853, doi:10.1126/science.1244693.

Healey, S. P., W. B. Cohen, Z. Yang, C. K. Brewer, E. B. Brooks, N. Gorelick, A. J. Hernandez, C. Huang, M. J. Hughes, and R. E. Kennedy, 2018: Mapping forest change using stacked generalization: An ensemble approach. *Remote Sens.Environ.,* **204**, 717-728.

Hennigar, C., A. Weiskittel, H. L. Allen, and D. A. MacLean, 2016: Development and evaluation of a biomass increment based index for site productivity. *Canadian Journal of Forest Research,* **47**, 400-410.

Hermosilla, T., M. A. Wulder, J. C. White, N. C. Coops, and G. W. Hobart, 2015a: An integrated Landsat time series protocol for change detection and generation of annual gap-free surface reflectance composites. *Remote Sensing of Environment,* **158**, 220-234, doi:http://dx.doi.org.prxy4.ursus.maine.edu/10.1016/j.rse.2014.11.005.

Hermosilla, T., M. A. Wulder, J. C. White, N. C. Coops, and G. W. Hobart, 2015b: Regional detection, characterization, and attribution of annual forest change from 1984 to 2012 using Landsat-derived time-series metrics. *Remote Sensing of Environment,* **170**, 121-132, doi:https://doi-org.prxy4.ursus.maine.edu/10.1016/j.rse.2015.09.004.

Herold, M., T. Johns, 2007: Linking requirements with capabilities for deforestation monitoring in the context of the UNFCCC-REDD process. *Environmental Research Letters,* **2**, 045025.

Homer, C., J. Dewitz, L. Yang, S. Jin, P. Danielson, G. Xian, J. Coulston, N. Herold, J. Wickham, and K. Megown, 2015: Completion of the 2011 National Land Cover Database for the conterminous United States–representing a decade of land cover change information. *Photogrammetric Engineering & Remote Sensing,* **81**, 345-354.

Hothorn, T., K. Hornik, C. Strobl, A. Zeileis, and M. T. Hothorn, 2015: Package 'party'. *Package Reference Manual for Party Version 0.9-998,* **16**, 37.

Houghton, R., 2005: Aboveground forest biomass and the global carbon balance. *Global Change Biol.,* **11**, 945-958.

Housman, I., V. Tanpipat, T. Biswas, A. Clark, P. Stephen, P. Maus, and K. Megown, 2015: Monitoring forest change in southeast Asia: case studies for USAID Lowering Emissions in Asia's Forests. **RSAC-10108-RPT1**.

Huang, C., S. N. Goward, J. G. Masek, N. Thomas, Z. Zhu, and J. E. Vogelmann, 2010: An automated approach for reconstructing recent forest disturbance history using dense Landsat time series stacks. *Remote Sens.Environ.,* **114**, 183-198, doi:http://dx.doi.org/10.1016/j.rse.2009.08.017.

Hudak, A. T., E. K. Strand, L. A. Vierling, J. C. Byrne, J. U. H. Eitel, S. Martinuzzi, and M. J. Falkowski, 2012: Quantifying aboveground forest carbon pools and fluxes from repeat LiDAR surveys. *Remote Sensing of Environment,* **123**, 25-40, doi:https://doi-org.prxy4.ursus.maine.edu/10.1016/j.rse.2012.02.023.

Hughes, M. J., S. D. Kaylor, and D. J. Hayes, 2017: Patch-based forest change detection from Landsat time series. *Forests,* **8**, 166.

Hughes, M. J., 2014: New Remote Sensing Method for Detecting and Quantifying Forest Disturbance and Regeneration in the Eastern United States.

Irland, L. C., 1998: Ice storm 1998 and the forests of the Northeast: A preliminary assessment. *J.For.,* **96**, 32-40.

Irland, L. C., 2000: Ice storms and forest impacts. *Science of The Total Environment,* **262**, 231-242, doi:https://doi-org.prxy4.ursus.maine.edu/10.1016/S0048-9697(00)00525-8.

Jiang, H., P. J. Radtke, A. R. Weiskittel, J. W. Coulston, and P. J. Guertin, 2014: Climate-and soil-based models of site productivity in eastern US tree species. *Canadian Journal of Forest Research,* **45**, 325-342.

Kennedy, R. E., Z. Yang, N. Gorelick, J. Braaten, L. Cavalcante, W. B. Cohen, and S. Healey, 2018: Implementation of the LandTrendr Algorithm on Google Earth Engine. *Remote Sensing,* **10**.

Kennedy, R. E., S. Andréfouët, W. B. Cohen, C. Gómez, P. Griffiths, M. Hais, S. P. Healey, E. H. Helmer, P. Hostert, M. B. Lyons, G. W. Meigs, D. Pflugmacher, S. R. Phinn, S. L. Powell, P. Scarth, S. Sen, T. A. Schroeder, A. Schneider, R. Sonnenschein, J. E. Vogelmann, M. A. Wulder, and Z. Zhu, 2014: Bringing an ecological view of change to Landsat-based remote sensing. *Frontiers in Ecology and the Environment,* **12**, 339-346, doi:10.1890/130066.

Kennedy, R. E., Z. Yang, and W. B. Cohen, 2010: Detecting trends in forest disturbance and recovery using yearly Landsat time series: 1. LandTrendr — Temporal segmentation algorithms. *Remote Sens.Environ.,* **114**, 2897-2910, doi:http://dx.doi.org/10.1016/j.rse.2010.07.008.

Key, C. H., N. C. Benson, 1999: The Normalized Burn Ratio (NBR): A Landsat TM radiometric measure of burn severity. *United States Geological Survey, Northern Rocky Mountain Science Center.(Bozeman, MT), .*

Lefsky, M. A., W. B. Cohen, G. G. Parker, and D. J. Harding, 2002: Lidar Remote Sensing for Ecosystem Studies: Lidar, an emerging remote sensing technology that directly measures the three-dimensional distribution of plant canopies, can accurately estimate vegetation structural attributes and should be of particular interest to forest, landscape, and global ecologists. *Bioscience,* **52**, 19-30, doi:10.1641/0006-3568(2002)052[0019:LRSFES]2.0.CO;2.

Lorimer, C. G., 1977: The presettlement forest and natural disturbance cycle of northeastern Maine. *Ecology,* **58**, 139-148.

Lu, D., Q. Chen, G. Wang, E. Moran, M. Batistella, M. Zhang, G. Vaglio Laurin, and D. Saah, 2012: Aboveground forest biomass estimation with Landsat and LiDAR data and uncertainty analysis of the estimates. *International Journal of Forestry Research,* **2012**.

Lu, D., 2005: Aboveground biomass estimation using Landsat TM data in the Brazilian Amazon. *Int.J.Remote Sens.,* **26**, 2509-2525, doi:10.1080/01431160500142145.

Luther, J. E., R. A. Fournier, D. E. Piercey, L. Guindon, and R. J. Hall, 2006: Biomass mapping using forest type and structure derived from Landsat TM imagery. *International Journal of Applied Earth Observation and Geoinformation,* **8**, 173-187, doi:https://doi.org/10.1016/j.jag.2005.09.002.

Masek, J. G., E. F. Vermote, N. E. Saleous, R. Wolfe, F. G. Hall, K. F. Huemmrich, F. Gao, J. Kutler, and T. Lim, 2006: A Landsat surface reflectance dataset for North America, 1990-2000. *IEEE Geoscience and Remote Sensing Letters,* **3**, 68-72.

Matasci, G., T. Hermosilla, M. A. Wulder, J. C. White, N. C. Coops, G. W. Hobart, and H. S. Zald, 2018: Large-area mapping of Canadian boreal forest cover, height, biomass and other structural attributes using Landsat composites and lidar plots. *Remote Sens.Environ.,* **209**, 90-106.

Mitchard, E. T., S. S. Saatchi, L. White, K. Abernethy, K. J. Jeffery, S. L. Lewis, M. Collins, M. A. Lefsky, M. E. Leal, and I. H. Woodhouse, 2012: Mapping tropical forest biomass with radar and spaceborne LiDAR in Lopé National Park, Gabon: overcoming problems of high biomass and persistent cloud. *Biogeosciences,* **9**, 179-191.

Pan, Y., R. A. Birdsey, J. Fang, R. Houghton, P. E. Kauppi, W. A. Kurz, O. L. Phillips, A. Shvidenko, S. L. Lewis, J. G. Canadell, P. Ciais, R. B. Jackson, S. W. Pacala, A. D. McGuire, S. Piao, A. Rautiainen, S. Sitch, and D. Hayes, 2011: A large and persistent carbon sink in the world's forests. *Science,* **333**, 988-993, doi:10.1126/science.1201609 [doi].

Pasquarella, V. J., B. A. Bradley, and C. E. Woodcock, 2017: Near-real-time monitoring of insect defoliation using Landsat time series. *Forests,* **8**, 275.

Pflugmacher, D., W. B. Cohen, and R. E. Kennedy, 2012: Using Landsat-derived disturbance history (1972–2010) to predict current forest structure. *Remote Sens.Environ.,* **122**, 146-165, doi:http://dx.doi.org/10.1016/j.rse.2011.09.025.

Pflugmacher, D., W. B. Cohen, R. E. Kennedy, and Z. Yang, 2014: Using Landsat-derived disturbance and recovery history and lidar to map forest biomass dynamics. *Remote Sens.Environ.,* **151**, 124-137, doi:http://dx.doi.org/10.1016/j.rse.2013.05.033.

Pickell, P. D., T. Hermosilla, R. J. Frazier, N. C. Coops, and M. A. Wulder, 2016: Forest recovery trends derived from Landsat time series for North American boreal forests. *Int.J.Remote Sens.,* **37**, 138-149, doi:10.1080/2150704X.2015.1126375.

Powell, S. L., W. B. Cohen, R. E. Kennedy, S. P. Healey, and C. Huang, 2014: Observation of trends in biomass loss as a result of disturbance in the conterminous US: 1986–2004. *Ecosystems,* **17**, 142-157.

Powell, S. L., W. B. Cohen, S. P. Healey, R. E. Kennedy, G. G. Moisen, K. B. Pierce, and J. L. Ohmann, 2010a: Quantification of live aboveground forest biomass dynamics with Landsat time-series and field inventory data: A comparison of empirical modeling approaches. *Remote Sens.Environ.,* **114**, 1053-1068, doi:http://dx.doi.org/10.1016/j.rse.2009.12.018.

Powell, S. L., W. B. Cohen, S. P. Healey, R. E. Kennedy, G. G. Moisen, K. B. Pierce, and J. L. Ohmann, 2010b: Quantification of live aboveground forest biomass dynamics with Landsat time-series and field inventory data: A comparison of empirical modeling approaches. *Remote Sens.Environ.,* **114**, 1053-1068, doi:http://dx.doi.org.prxy4.ursus.maine.edu/10.1016/j.rse.2009.12.018.

Pryzant, R., S. Ermon, and D. Lobell, 2017: Monitoring Ethiopian Wheat Fungus with Satellite Imagery and Deep Feature Learning. *Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW),* IEEE, 1524-1532.

R Core Team, 2017: *R: A language and environment for statistical computing.R Foundation for Statistical Computing, Vienna, Austria.2016,* .

Radtke, P., D. Walker, J. Frank, A. Weiskittel, C. DeYoung, D. MacFarlane, G. Domke, C. Woodall, J. Coulston, and J. Westfall, 2017: Improved accuracy of aboveground biomass and carbon estimates for live trees in forests of the eastern United States. *Forestry: An International Journal of Forest Research,* **90**, 32-46.

Robinson, A. P., R. A. Duursma, and J. D. Marshall, 2005: A regression-based equivalence test for model validation: shifting the burden of proof. *Tree Physiol.,* **25**, 903-913.

Robinson, A. P., R. E. Froese, 2004: Model validation using equivalence tests. *Ecological Modelling,* **176**, 349-358, doi:https://doi-org.prxy4.ursus.maine.edu/10.1016/j.ecolmodel.2004.01.013.

Rouse, J., R. Haas, J. Schell, and D. Deering, 1974: Monitoring vegetation systems in the Great Plains with ERTS. *Proc. Third Earth Resources Technology Satellite-1 Symposium. Volume 1: Technical Presentations, section A,* TX, United States, .

Roy, D. P., V. Kovalskyy, H. K. Zhang, E. F. Vermote, L. Yan, S. S. Kumar, and A. Egorov, 2016: Characterization of Landsat-7 to Landsat-8 reflective wavelength and normalized difference vegetation index continuity. *Remote Sens.Environ.,* **185**, 57-70, doi:http://doi.org/10.1016/j.rse.2015.12.024.

Seymour, R. S., 1994: The northeastern region. *Regional silviculture of the United States,* **3**, 31-79.

Seymour, R. S., A. S. White, and P. G. deMaynadier, 2002: Natural disturbance regimes in northeastern North America—evaluating silvicultural systems using natural scales and frequencies. *For.Ecol.Manage.,* **155**, 357-367, doi:http://dx.doi.org/10.1016/S0378-1127(01)00572-2.

Strobl, C., A. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, 2008: Conditional variable importance for random forests. *BMC Bioinformatics,* **9**, 307.

Thompson, J. R., D. N. Carpenter, C. V. Cogbill, and D. R. Foster, 2013: Four Centuries of Change in Northeastern United States Forests. *PLoS ONE,* **8**, 1-15, doi:10.1371/journal.pone.0072540.

Thornton, P., M. Thornton, B. Mayer, N. Wilhelmi, Y. Wei, R. Devarakonda, and R. Cook, 2016: *Daymet: daily surface weather data on a 1-km grid for North America, Version 3.ORNL DAAC, Oak Ridge, Tennessee, USA,* , doi:https://doi.org/10.3334/ORNLDAAC/1328.

Vermote, E., C. Justice, M. Claverie, and B. Franch, 2016: Preliminary analysis of the performance of the Landsat 8/OLI land surface reflectance product. *Remote Sensing of Environment,* **185**, 46-56, doi:https://doi-org.prxy4.ursus.maine.edu/10.1016/j.rse.2016.04.008.

Vinayak, R. K., R. Gilad-Bachrach, 2015: Dart: Dropouts meet multiple additive regression trees. *Proc. Artificial Intelligence and Statistics,* , 489-497.

White, J. C., M. A. Wulder, G. W. Hobart, J. E. Luther, T. Hermosilla, P. Griffiths, N. C. Coops, R. J. Hall, P. Hostert, A. Dyk, and L. Guindon, 2014: Pixel-Based Image Compositing for Large-Area Dense Time Series Applications and Science. *Canadian Journal of Remote Sensing,* **40**, 192-212, doi:10.1080/07038992.2014.945827.

Wilson, E. H., S. A. Sader, 2002: Detection of forest harvest type using multiple dates of Landsat TM imagery. *Remote Sens.Environ.,* **80**, 385-396, doi:http://dx.doi.org/10.1016/S0034-4257(01)00318-2.

Wolpert, D. H., 1992: Stacked generalization. *Neural Networks,* **5**, 241-259.

Wright, M. N., A. Ziegler, 2015: ranger: A fast implementation of random forests for high dimensional data in C and R. *arXiv preprint arXiv:1508.04409,* .

Wulder, M. A., J. C. White, T. R. Loveland, C. E. Woodcock, A. S. Belward, W. B. Cohen, E. A. Fosnight, J. Shaw, J. G. Masek, and D. P. Roy, 2016: The global Landsat archive: Status, consolidation, and direction. *Remote Sensing of Environment,* **185**, 271-283, doi:https://doi-org.prxy4.ursus.maine.edu/10.1016/j.rse.2015.11.032.

Zald, H. S., M. A. Wulder, J. C. White, T. Hilker, T. Hermosilla, G. W. Hobart, and N. C. Coops, 2016: Integrating Landsat pixel composites and change metrics with lidar plots to predictively map forest structure and aboveground biomass in Saskatchewan, Canada. *Remote Sens.Environ.,* **176**, 188-201.

Zhao, F., C. Huang, S. N. Goward, K. Schleeweis, K. Rishmawi, M. A. Lindsey, E. Denning, L. Keddell, W. B. Cohen, and Z. Yang, 2018: Development of Landsat-based annual US forest disturbance history maps (1986–2010) in support of the North American Carbon Program (NACP). *Remote Sens.Environ.,* **209**, 312-326.

Zhu, Z., C. E. Woodcock, 2014: Continuous change detection and classification of land cover using all available Landsat data. *Remote Sens.Environ.,* **144**, 152-171.

Zhu, Z., C. E. Woodcock, 2012: Object-based cloud and cloud shadow detection in Landsat imagery. *Remote Sens.Environ.,* **118**, 83-94, doi:http://dx.doi.org.prxy4.ursus.maine.edu/10.1016/j.rse.2011.10.028.

APPENDICES:

Appendix A:

Table A.1. The formulation and references for the spectral indices and transformations used as spectral covariates.

Spectral Index	Formulation	Reference
Normalized Difference Vegetation Index (NDVI)	(B3 – B4) / (B3 + B4)	Rouse et al. (1974)
Normalized Burn Ratio (NBR)	(B4 – B7) / (B4 + B7)	Key and Benson(1999)
Normalized Difference Moisture Index (NDMI)	(B4 – B5) / (B4 + B5)	Wilson and Sader (2002)
Tasseled Cap Brightness (TCB)	0.2043*B1 + 0.4158*B2 + 0.5524*B3 + 0.5741*B4 + 0.3124*B5 + 0.2303*B7	Crist (1985)
Tasseled Cap Greenness (TCG)	-0.1603*B1 + -0.2819*B2 + -0.4934*B3 + 0.7940*B4 + -0.0002*B5 + -0.1446*B7	Crist (1985)
Tasseled Cap Wetness (TCW)	0.0315*B1 + 0.2021*B2 + 0.3102*B3 + 0.1594*B4 + -0.6806B5 + -0.6109*B7	Crist (1985)
Tasseled Cap Angle (TCA)	$Arctan(TCG / TCB)$	Powell et al. (2010a)
Tasseled Cap Distance (TCD)	$Sqrt(TCG^2 + TCB^2)$	Duane et al. (2010)

BIOGRAPHY OF THE AUTHOR

John Kilbride was born in Portland, Maine, on March 7th, 1994. He was raised in Falmouth, Maine and attended the Falmouth Public Schools. He graduated from the University of Vermont in 2016 with a degree in Geography and minors in Geospatial Technologies and Economics. John is a candidate for the Master of Science degree in Forest Resources from the University of Maine in August 2018.