Int J Appl Earth Obs Geoinformation 86 (2020) 101985

Contents lists available at ScienceDirect

Int J Appl Earth Obs Geoinformation

journal homepage: www.elsevier.com/locate/jag

Estimating tree canopy cover using harmonic regression coefficients derived from multitemporal Landsat data

Jill M. Derwin[a,*], Valerie A. Thomas[a], Randolph H. Wynne[a], John W. Coulston[b], Greg C. Liknes[c], Stacie Bender[d], Christine E. Blinn[a], Evan B. Brooks[a], Bonnie Ruefenacht[d], Robert Benton[d], Mark V. Finco[d], Kevin Megown[d]

[a] Virginia Tech, Forest Resources and Environmental Conservation, Blacksburg, VA 24061 USA
[b] United States Forest Service, Southern Research Station, Blacksburg, VA 24061 USA
[c] United States Forest Service, Northern Research Station, St Paul, MN 55108 USA
[d] United States Forest Service, Geospatial Technology and Applications Center, Salt Lake City, UT 84138 USA

ARTICLE INFO

Keywords:
Tree Canopy Cover
Landsat time series
Harmonic regression
Image compositing
Random forest regression models

ABSTRACT

The goal of this study was to evaluate whether harmonic regression coefficients derived using all available cloud-free observations in a given Landsat pixel for a three-year period can be used to estimate tree canopy cover (TCC), and whether models developed using harmonic regression coefficients as predictor variables are better than models developed using median composite predictor variables, the previous operational standard for the National Land Cover Database (NLCD). The two study areas in the conterminous USA were as follows: West (Oregon), bounded by Landsat Worldwide Reference System 2 (WRS-2) paths/rows 43/30, 44/30, and 45/30; and South (Georgia/South Carolina), bounded by WRS-2 paths/rows 16/37, 17/37, and 18/37. Plot-specific tree canopy cover (the response variable) was collected by experienced interpreters using a dot grid overlaid on 1 m spatial resolution National Agricultural Imagery Program (NAIP) images at two different times per region, circa 2010 and circa 2014. Random forest model comparisons (using 500 independent model runs for each comparison) revealed the following (1) harmonic regression coefficients (one harmonic) are better predictors for every time/region of TCC than median composite focal means and standard deviations (across times/regions, mean increase in pseudo R^2 of 6.7% and mean decrease in RMSE of 1.7% TCC) and (2) harmonic regression coefficients (one harmonic, from NDVI, SWIR1, and SWIR2), when added to the full suite of median composite and terrain variables used for the NLCD 2011 product, improve the quality of TCC models for every time/region (mean increase in pseudo R^2 of 3.6% and mean decrease in RMSE of 1.0% TCC). The harmonic regression NDVI constant was always one of the top four most important predictors across times/regions, and is more correlated with TCC than the NDVI median composite focal mean. Eigen analysis revealed that there is little to no additional information in the full suite of predictor variables (47 bands) when compared to the harmonic regression coefficients alone (using NDVI, SWIR1, and SWIR2; 9 bands), a finding echoed by both model fit statistics and the resulting maps. We conclude that harmonic regression coefficients derived from Landsat (or, by extension, other comparable earth resource satellite data) can be used to map TCC, either alone or in combination with other TCC-related variables.

1. Introduction

Tree canopy cover (TCC) is the proportion of the forest floor covered by the vertical projection of tree crowns (Jennings, 1999). TCC is an essential measure of forest health and productivity and is used in applications like climate change mitigation, forest management, and pest and disease monitoring. TCC influences wildlife habitat (North et al., 2017), forb yield (Muoghalu and Isichei, 1991), proportion of C3 vs. C4 grasses (Peterson et al., 2007), urban property values (Pandit et al., 2014), and rainfall partitioning (Owens et al., 2006), among others.

At the national scale, TCC is commonly studied annually. Numerous authors have shown the benefits of describing multitemporal

* Corresponding author.
E-mail addresses: jmd59@vt.edu (J.M. Derwin), thomasv@vt.edu (V.A. Thomas), wynne@vt.edu (R.H. Wynne), jcoulston@fs.fed.us (J.W. Coulston), staciebender@fs.fed.us (S. Bender), cblinn@vt.edu (C.E. Blinn), evbrooks@vt.edu (E.B. Brooks), bruefenacht@fs.fed.us (B. Ruefenacht), robertbenton@fs.fed.us (R. Benton), mfinco@fs.fed.us (M.V. Finco), kamegown@fs.fed.us (K. Megown).

https://doi.org/10.1016/j.jag.2019.101985
Received 22 February 2019; Received in revised form 16 August 2019; Accepted 1 October 2019
Available online 22 November 2019

Table 1
Phenology metrics used to map tree canopy cover using satellite data. Only the sensors and bands shown to be important in the resulting tree canopy cover models are listed. Abbreviations: AVHRR, Advanced Very High Resolution Radiometer; MODIS, MOderate-resolution Imaging Spectrometer; OLI, Operational Land Imager, SPOT VGT, Satellite Pour l'Observation de la Terre VEGETATION; TM, Thematic Mapper; NIR, near infrared; SWIR, shortwave infrared; NDVI, normalized difference vegetation index; NDMI, normalized difference moisture index; EVI, enhanced vegetation index; FAPAR, fraction of absorbed photosynthetically active radiation; TC B,G,W, tasseled cap brightness, greenness, and wetness; std, standard deviation; SOS, start of season; DSINT, dry season integral (annual integral minus the integral from start of season to end of season); RDERIV, right derivative (rate of decrease after growing season maximum).

Study	Sensors(s)	Bands/Indices	Curve Fitting	Phenology Metrics
Hansen and DeFries (2004)	AVHRR (monthly composite ordinal by calendar year)	red NIR NDVI BT	No	2nd 6th mean (2nd:6th) mean (2nd:4th) |6th–2nd|
Gessner et al. (2013)	MODIS	blue red SWIR (2.13 μm) NDVI EVI	Fourier	minimum* maximum mean* median amplitude
Karlson et al. (2015)	Landsat OLI	NDVI	No	minimum maximum mean median product* *std*
Brandt et al. (2016)	MODIS SPOT VGT	FAPAR	TIMESAT	DSINT* SOS* RDERIV
Ruefenacht (2016)	Landsat TM	all optical NDVI NDMI TC B,G,W	No	median composite*

*Denotes a particularly important variable.

reflectance data using phenology metrics to estimate TCC, as opposed to selecting one or only a few images throughout the growing season. DeFries et al. (1995) proposed the use of NDVI temporal dynamics to estimate (vegetation) continuous fields of canopy cover. Hansen and DeFries (2004) produced annual global TCC estimates at 8 km spatial resolution using per-pixel phenological metrics derived from Advanced Very High Resolution (AVHRR) monthly composites. Training data were developed using much higher spatial resolution imagery. Phenology metrics utilized were selected monthly composite values coupled with key descriptive statistics for the red, near infrared (NIR), normalized difference vegetation index, and brightness temperature bands.

Phenology metrics have been used as predictor variables in other, more recent, studies in which satellite imagery was used to map tree canopy cover (Table 1). Gessner et al. (2013) mapped the fraction of woody canopy cover in southern Africa using phenology metrics (Table 1) derived from Fourier-smoothed (via harmonic regression, Fig. 1) annual time series of MODIS imagery with blue, red, SWIR (2.13 μm), NDVI, and enhanced vegetation index (EVI) bands. Their resulting random forest models had an RMSE of 8.1% cover in the Kalahari and 3.1% cover in central and eastern Namibia. The three most important variables were the SWIR3 minimum in the dry season, the annual NDVI mean, and the rainy season NDVI mean. Karlson et al. (2015) mapped tree canopy cover at a 30 m pixel resolution in Burkina Faso using Landsat Operational Land Imager (OLI)-derived spectral, texture, and phenology variables. All three variable types (including one phenology metric) were necessary in the resulting random forests model that required only five variables (R^2 = 77%, RMSE = 8.9% cover) derived using backwards feature elimination to find the smallest possible model. Brandt et al. (2016) compared phenology metrics derived using TIMESAT (Jönsson and Eklundh, 2004) smoothed data from the Moderate Resolution Imaging Spectroradiometer (MODIS) and Satellite pour l'Observation de la Terre (SPOT) VEGETATION (VGT) for their utility in mapping percent woody cover at a 1 km scale across the Sahel. The most important phenology metrics, regardless of sensor, were the dry season integral and the starting point of the growing season, defined as 20% of the amplitude. The resulting multiple linear regression models had good predictive ability (R^2, MODIS = 73% and R^2, VGT = 70%) and excellent RMSEs (MODIS, 3.0% cover and VGT, 3.2% cover), especially given the challenging nature of the arid landscape.

Ruefenacht (2016) mapped tree canopy cover using 3 × 3 pixel window focal means of the growing season median composite value for each of the six multispectral bands from the Landsat Thematic Mapper (TM), plus derived indices NDVI, tasseled cap brightness, greenness, and wetness, and the normalized difference moisture index (NDMI). The growing season was defined as the period in which NDVI was within 0.1 of the growing season maximum. Tree canopy cover was mapped in Multi-Resolution Land Characteristics Consortium Mapping Zones 16, 23, 48, 54, and 59. Model quality varied across zones, with R^2s ranging from 50% to 86% and RMSEs ranging from 12.8% cover to 15.3% cover. Median composites were used operationally to map tree canopy cover as an element of the National Land Cover Database (NLCD) 2011 suite of nationwide geospatial data layers (Ruefenacht et al., 2015).

De Beurs and Henebry (2010) categorized methods by which land surface phenology can be characterized as follows: (1) thresholds, (2) derivatives, (3) smoothing functions, and (4) fitted models. Given that standardized markers of start and end of season are first order objectives of these techniques, thresholds and derivatives are less well suited to develop predictor variables not related to growing season timing or length. De Beurs and Henebry (2010) further note that complicated fitted models such as Gaussian local functions result in parameter coefficients with no clear ecological meaning. As such, simpler smoothing functions such as the autoregressive moving average or Fourier analysis can result in parsimonious models with physically meaningful coefficients with only slight loss to model fit. Given (among other issues) the subjective, heuristic nature of selecting a suitable lag time for autoregressive moving average models of land surface phenology, they play an increasingly less prominent role.

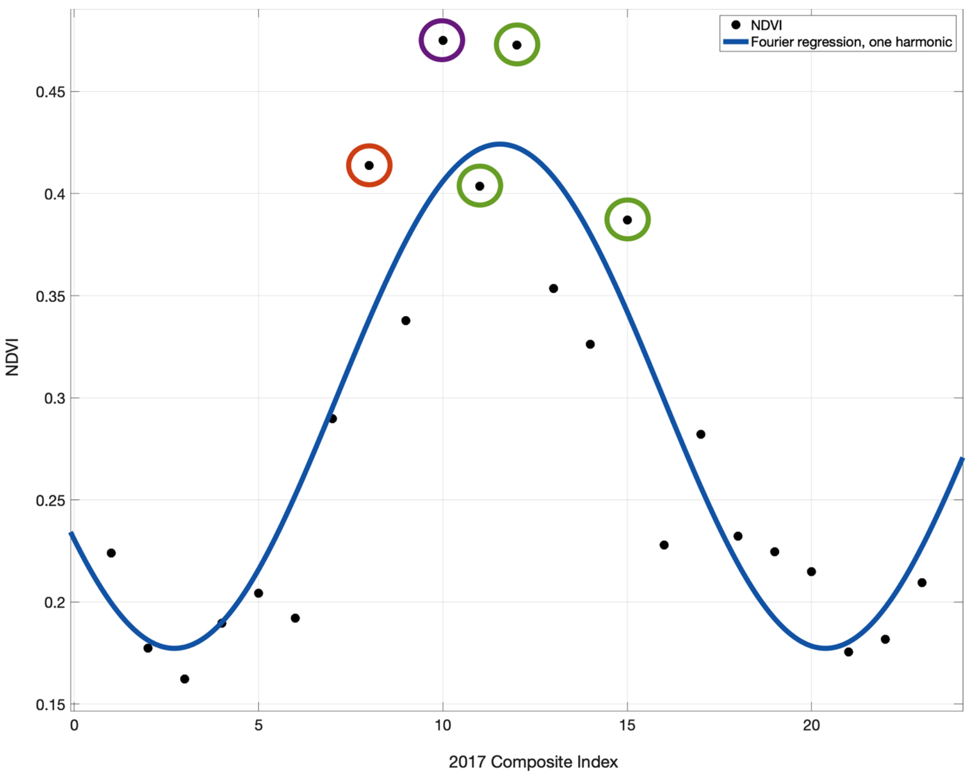

Fig. 1. Median composite and harmonic regression. This figure illustrates the median composite method using NDVI data from MODIS 250 m 16-day composites from 2017 (MYD13Q1 collection 6, coordinates 37.4378°, −80.2727°). Landsat median composite images are calculated using the median reflectance value from up to 15 images from three years preceding the target year. In this example, the maximum MODIS NDVI value is shown, circled, in violet. The Landsat median composite selects observations within 0.1 GLOVIS maximum annual NDVI value (circled in green or red for this example, using annual maximum MODIS NDVI) are used to calculate the median composite, circled in red. (Note that for non-composited Landsat data the earliest and latest value within 0.1 NDVI of the maximum serve to define the range of data used, and the pixel-specific effects of cloud cover result in a median being computed with different numbers of values from pixel to pixel.) The figure also demonstrates the harmonic regression method with MODIS NDVI data. Harmonic regression of surface reflectance observations uses all available observations (black dots) and fits a harmonic curve (blue line) following Eq. (1). Both methods exclude observations identified as cloud cover. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Fourier (harmonic) analysis of annual spectral band reflectances or, more commonly, vegetation indices derived therefrom, has long been used to characterize land surface phenology (Moody and Johnson, 2001). Fourier analysis is sensitive to systematic signal changes but relatively insensitive to (nonsystematic) noise (Moody and Johnson, 2001; De Beurs and Henebry, 2010). Brooks et al. (2012) define Fourier series as “superimposed sequences, over an interval of time, of a constant with sines and cosines of increasing integer multiples of the original frequency based on the time interval” (see Eq. (1) and Fig. 1). The constant is the mean of the series, and the sine and cosine pairings at specified frequencies are called the harmonics of the series (Brooks et al., 2012) (Eq. (1)).

$$f(t) = a_0 + \underbrace{(a_1 \cos t + b_1 \sin t)}_{\text{first harmonic}} + \underbrace{(a_2 \cos 2t + b_2 \sin 2t)}_{\text{second harmonic}} + \underbrace{(a_3 \cos 3t + b_3 \sin 3t)}_{\text{third harmonic}} + \cdots \quad (1)$$

where t is the “time from some initial epoch” (Shureman, 1940), a_0 is the constant, and for each harmonic n, the coefficients a_n and b_n are the amplitudes of the cosine and sine waves, respectively (Smith, 1997). Using harmonic regression on Landsat time-series data to produce a simplified curve overcomes both noise and gaps caused by clouds, snow, or instrument problems, which is a major advantage to the approach (Brooks et al., 2012; Wilson et al., 2018).

For land surface phenology applications, the period is one year, defined in the same units as t. The first harmonic is also known as the fundamental frequency, and is the frequency that the time domain repeats itself (Smith, 1997). As Moody and Johnson (2001) noted, for vegetation indices the first harmonic is particularly associated with the overall productivity of the region, with the constant (a_0) the mean. Regional and continental applications, particularly across biomes, often use only the first harmonic even though the second (and subsequent) harmonics can improve fit. This is due to the second harmonic being variable across years and ecosystems due to its sensitivity to secondary vegetation and climate anomalies (Moody and Johnson, 2001). Fig. 1 demonstrates harmonic regression of surface reflectance observations and compares it to the median composite method, previously described.

Advantages of harmonic regression include (1) the orthogonality of the sine and cosine functions (Weisstein, 2010) and (2) the ability to store the coefficients as rasters (Brooks et al., 2012) (Fig. 2), and (3) the proven utility of the coefficients themselves in a wide variety of mapping applications, both for categorical (Brooks et al., 2016; Wilson et al., 2018) and continuous (Immerzeel et al., 2005; Wilson et al., 2018) variables. However, no prior study (Table 1) has used harmonic regression coefficients derived from moderate resolution resource satellite imagery as predictor variables to estimate tree canopy cover.

The goal of this study was to evaluate whether harmonic regression coefficients derived using all available cloud-free observations in a given Landsat pixel for a three-year period (corresponding with the timing of imagery used for training data collection) can be used to estimate tree canopy cover, and whether using harmonic regression coefficients as predictor variables is an improvement over calculating Landsat-derived predictor variables using the median composite, the current operational standard for NLCD. We explored two research questions in service of this goal, as follows: (1) Are harmonic regression coefficients better predictors of TCC than median composite variables for the same vegetation index and wavelength bands? (2) Can harmonic regression coefficients, when added to a full suite of median composite and terrain variables improve the quality of TCC models (simulating the operational framework for the original 2011 NLCD TCC product vs. that used for the 2016 NLCD TCC product)? In order to answer the first question, we compared models using harmonic regression coefficients for NDVI, SWIR1 and SWIR2, all good predictors of TCC (Table 1; Asner and Lobell, 1996; Lobell et al., 2001) to the median composite values for those same three bands. To answer the second question we compared models using (a) the full suite of median composite and terrain variables used in the original 2011 TCC product, and (b) all predictor variables used in (a) *plus* the harmonic regression coefficients from NDVI, SWIR1, and SWIR2.

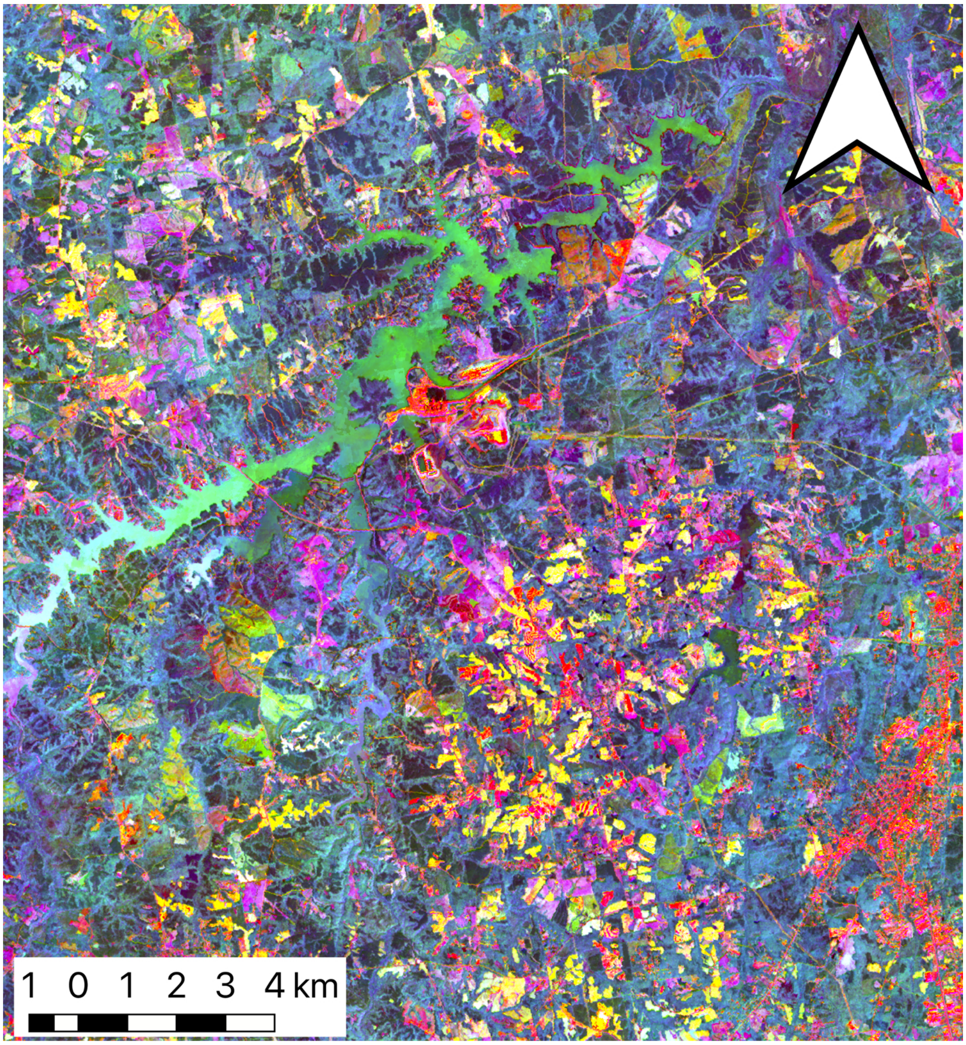

Fig. 2. Harmonic regression coefficients (one harmonic) for the Landsat 8 OLI panchromatic band (15 m spatial resolution) using all cloud-free observations for each pixel from 2014–2016. a_0 shown in red, a_1 shown in green, and b_1 shown in blue. Image boundary coordinates (x:y pairs, Conus Albers, EPSG:5070) upper left 1,485,996 m, 1,634,473 m; lower right 1,505,392 m, 1,613,539 m. Note the quality of land cover and land use discrimination that is possible even using harmonic regression coefficients from only one wavelength band. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2. Data and methods

2.1. Study area

There were two study areas (Fig. 3). The first was in Oregon (USA), bounded by Landsat Worldwide Reference System 2 (WRS-2) paths/rows 43/30, 44/30, and 45/30, and the second was in Georgia and South Carolina (USA) in WRS-2 paths/rows 16/37, 17/37, and 18/37. There are a variety of land cover types represented across the study areas, though the western study area has large tracts with reduced TCC, such as shrublands or pasture. The southern study area has a more diverse, dynamic landscape, characterized by managed forests, interspersed with suburban development.

2.2. Training (response) data

Training data were collected using a photo interpretation (PI) approach that superimposes an evenly spaced 90 m × 90 m dot grid, approximately the area covered by an FIA field plot, over 1 m resolution NAIP images for a sample of photo interpretation points. These data were collected by experienced interpreters as part of a joint effort between the United States Forest Service's (USFS) Forest Inventory and Analysis (FIA) program and the USFS's Geospatial Technology and Applications Center (GTAC) (Toney et al., 2009; Goeking et al., 2012). NAIP images are acquired after green up, when leaves are on trees, so the PI process assumed those conditions. Image acquisitions occurred during summer months of 2009 and 2010 for Oregon and Georgia respectively, and spring of 2011 for South Carolina for Time 1. For Time 2, acquisition occurred during the summer months of 2012 and 2013 for Oregon and Georgia respectively, and during the fall of 2013 for South Carolina (Table 2).

At every photo interpretation point, trained interpreters indicated whether or not each of 105 dots fell on tree canopy, and plot-level percent TCC was calculated by dividing the number of canopy observations by the total number of dots observed. 2626 interpreted points were used, 1360 in the southern study area and 1266 in the western study area (Table 3).

2.3. Training (response) data characteristics

For the south, both time periods have relatively consistent bimodal distributions (Fig. 4, upper half and Fig. 5, left) of observed TCC that are left-skewed. For time 1, the mean is 66.8% and the median is 86.2%. For time 2 the mean is 65.8% and the median is 83.5%. In the west, both time periods also have relatively consistent distributions, but they are not bimodal (Fig. 4, lower half and Fig. 5, right). They are skewed to the right rather than to the left. For time 1, the median is 0.9% and the mean is 18.2%. For time 2 the mean is 18.6% and median is 0%. All sets of observations (both regions/dates) have a minimum of 0% and a maximum of 100%.

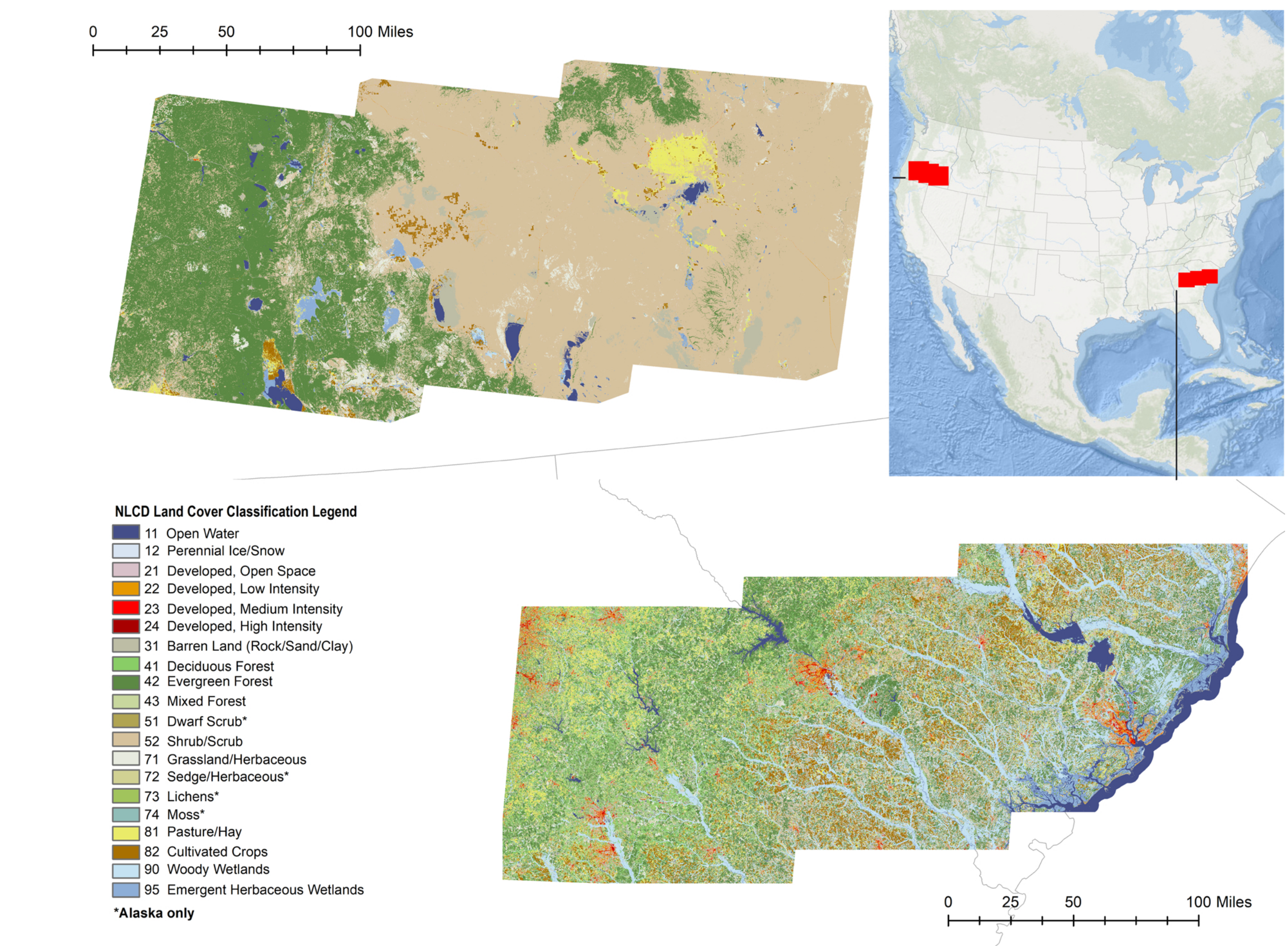

Fig. 3. Western(top) and Southern(bottom) study areas with 2011 National Land Cover Database classification.

Table 2
US National Agricultural Imagery Program (NAIP) image acquisition dates. Note that NAIP aerial images are collected with three visible bands and one near infrared band on clear days to ensure that images have less than 10 percent cloud cover per quarter quad tile. National coverage has been obtained every 3 years since 2009 (USDA Farm Service Agency, 2018).

Year	State	Date Range
2009	Oregon	June 1 to August 8
2012		June 10 to August 4
2010	Georgia	July 3 to October 6
2013		June 28 to October 26
2011	South Carolina	April 19 to May 26
2013		August 25 to October 30

Table 3
Number of photo interpretation (response data) samples per region and time period.

	Time 1	Time 2
South	1360	1360
West	1266	1266
Total	2626	2626

2.4. Development of predictor variables

2.4.1. Predictor variables for median composite models

Following Ruefenacht (2016), we developed a suite of imagery corresponding to Time 1 (2009 West, 2010–2011 South) and Time 2 (2012 West, 2013 South) of the response data by region. Landsat median composite images are a composite of observations from up to 15 dates from the three years leading up to the target year (Fig. 1). These observations were obtained during the annual periods when Global Visualization Viewer (GLOVIS) NDVI curves for forest classes (forest, evergreen forest, mixed forest, and woody wetlands) were within 10% of their annual maximum. Only images with less than 70% cloud cover were used, and clouds and cloud shadows were masked out of the remaining images using Fmask (Zhu and Woodcock, 2012). The median composite predictor variables (Table 7) include the means and standard deviations (calculated using a 90 m × 90 m focal window; Ruefenacht, 2016) of NDVI, NDMI, reflectances of Landsat 7 bands 1–5 and 7, reflectances of bands 2–7 and 9 for Landsat 8, and tasseled cap bands 1–6. In addition to median composite imagery, terrain data from a 30 m digital elevation model were derived. These included the means and standard deviations of elevation, slope, aspect, sin(aspect), and cos (aspect).

2.4.2. Harmonic regression coefficients from Landsat time series

Harmonic regression predictor variables (one harmonic, Table 7) were derived from multitemporal stacks of all available Landsat 5 and 7 images, acquired from USGS, for 3-year periods (with some pixel-specific exceptions as noted below). For Time 1, the training period started in 2008 for the western scenes and 2009 for the southern scenes. For Time 2, the period was from 2012 to 2014. This amounted to 65 dates per path/row at a minimum. To ensure a good harmonic curve fit, an R^2 threshold of 0.9 was applied for the Time 1 data using the EWMACD algorithm with the parameters listed in Table 4(Brooks et al., 2016; EWMACD v.1.8.7, Brooks, 2019). If the curve fit was less than the

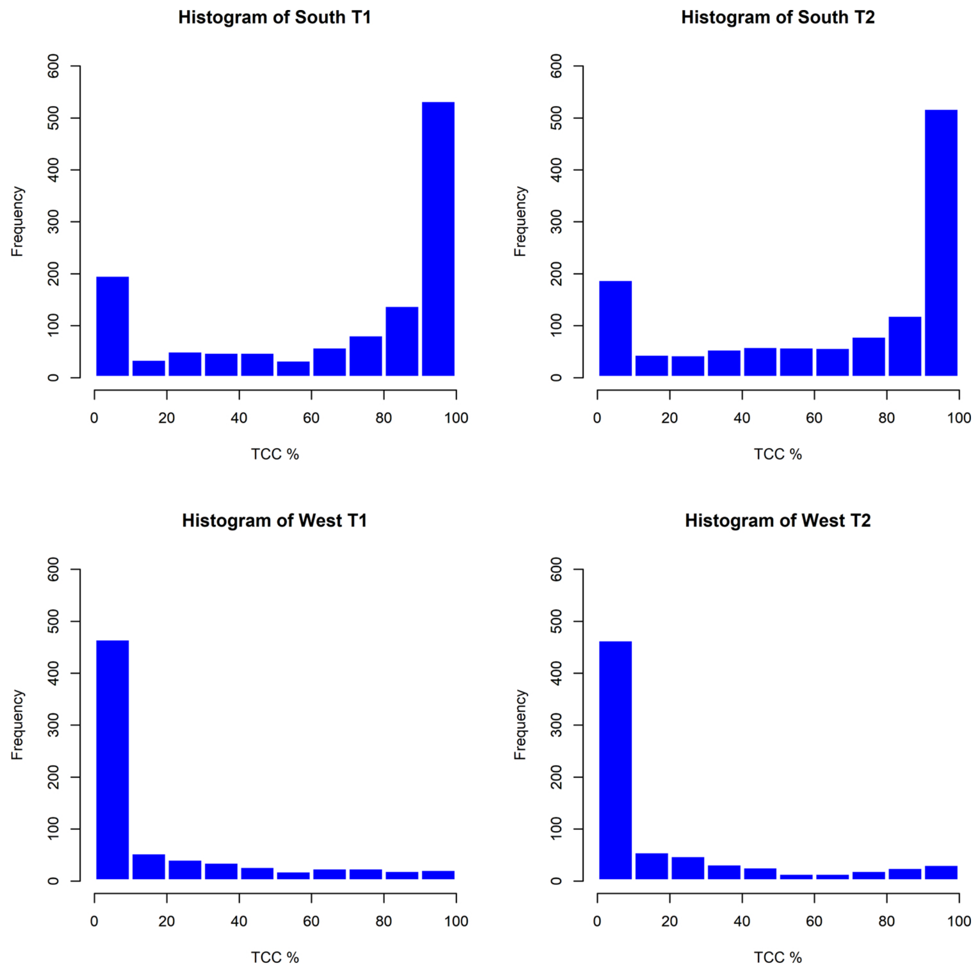

Fig. 4. Histograms of photo interpreted TCC observations for T1 (time 1) and T2 (time 2) in the South and West.

threshold, the temporal window of training observations was expanded to include more scenes from earlier dates, allowing time 1's training period to potentially extend to 2012. Therefore, the total number of input scenes (dates) varies by pixel, with the maximum number of scenes for any pixel being 192. Pre-processing for these data included the masking of clouds, cloud shadows, snow, and water from all images using cfmask provided in the LEDAPS level-1 product.

We compiled these stacks for the short wave infrared bands (SWIR 1 and SWIR 2), and NDVI (Rouse et al., 1974). For each data type, all acquisitions were stacked into multitemporal raster stacks, with each pixel representing a vector of reflectance values chronologically ordered through time. To determine the harmonic regression coefficients for the NDVI and SWIR data, a single harmonic curve (found to be superior to a two harmonic curve in preliminary analyses) was fit to each pixel in the multitemporal stack to derive the constant, sine, and cosine coefficients that describe the temporal behavior for that pixel (Brooks et al., 2012, Fig. 1, Eq. (1)).

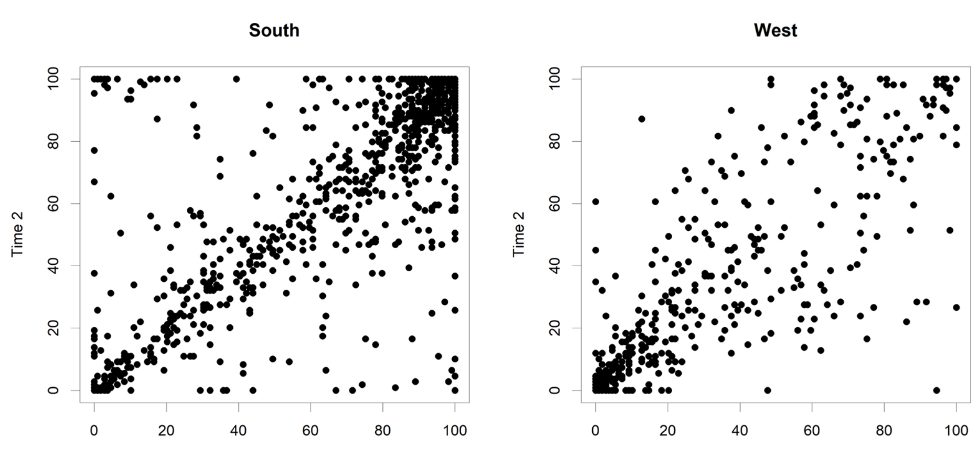

Fig. 5. Scatter plots comparing Time 1 and Time 2 photo interpreted TCC observations for the South and West.

Table 4
Parameters used in the EWMACD algorithm.

EWMACD.v.1.8.5 parameters	Inputs
trainingPeriod	'static'
numberHarmonicsSine	1
numberHarmonicsCosine	1
xBarLimit1	1.5
xBarLimit2	20
lambda	0.3
lambdaSigs	3
rounding	TRUE
persistence	10
reverseOrder	FALSE
trainingFitMinimumQuality	0.9
parallelFramework	'snow'
summaryMethod	'mean'
outputType	'coefficients only'
trainingPeriod	'static'
lowthresh	0
minTrainingLength	15
maxTrainingLength	30

Table 5
Parameters used in the randomForest algorithm.

randomForest parameters	Inputs
ntree	500
corr.bias	True
replace	True
mtry	4
nodesize	1

Table 6
Additional R libraries used in computing.

Additional libraries	Usage
raster (Hijmans, 2018)	Raster analysis
spatial.tools (Greenberg, 2018)	Spatial data analysis
foreach (Microsoft Corporation and Weston, 2017b)	Advanced for-loops
doSNOW (Microsoft Corporation and Weston, 2017a)	Parallelization with foreach

2.5. Predictive modeling with random forests

Given that the predictor variables are derived from Landsat reflectances and indices, there is multicollinearity in the predictor variables (not shown). This is somewhat mitigated among the harmonic regression coefficients, but there is still strong correlation between coefficients with the same function (i.e., the cosines for SWIR 1 and SWIR 2 are highly correlated, as are the sines). This multicollinearity causes few, if any, deleterious effects on random forest regression prediction.

To address the study objectives, four groups of variables (Table 7) were incorporated into random forest models (Breiman, 2001) using R's randomForest package (version 3.4.2; Liaw and Wiener, 2002), using the parameters listed in Table 5. Additionally, R packages listed in Table 6 were used for computing. The first two models included only variables derived from the NDVI and SWIR bands, to allow for a direct comparison between the median composite and harmonic regression approaches. The second study objective was addressed by the third and fourth models. The third model includes the same variables that were used to develop the 2011 NLCD Tree Canopy Cover product (Coulston et al., 2012; Ruefenacht, 2016). Finally, a hybrid of the median composite and harmonic regression approaches is investigated. Separate sets of models were trained to predict tree canopy cover for two time periods in 2010 and 2013 (using predictor variables and response data from appropriate date ranges) and the two study areas (South and West), discussed above.

2.5.1. Iterative modeling framework

Each model was run in R (R Core Team, 2018) with an iterative bootstrapping approach to include all available in-bag training data while also using an out-of-bag sample for model evaluation. Values from each of the input variables were extracted to plot locations and saved to a master table. For in-bag training samples, plots were randomly selected from this table with replacement, allowing for repeat plots in each bootstrapped sample. In-bag training plots were then used to estimate a random forest model based on the extracted values at each sample location. This process was repeated up to 500 times generating 500 forests of 500 decision trees. The same in-bag and out-of-bag plots were used for corresponding iterations of the four models that used different groups of predictor variables (i.e., Full model, Median Composite and Terrain, Harmonic Coefficient, and Median Composite models).

The number of points in the in-bag ample was equal to the number of extracted plots, and because of the sampling replacement, approximately 37% (Coulston et al., 2016) of each bootstrapped sample was left out. These out-of-bag samples were not used in model fitting but were used for model testing, comparing predicted and observed values for each of the 500 models (using their associated out-of-bag samples), and aggregating metrics across all models.

2.6. Model comparison

We compared models using several different measures. Variable importance for the full suite of model variables was used to assess the relative importance of included predictors, while mean correlation and error for the 500 model runs was compared for the four model types. Inherent dimensionality of predictor variables was used to determine the relative amount of information available in each model type. Models were mapped spatially in the southern study area to ensure that output maps were visually consistent. Finally, prediction variance was obtained by land cover type by calculating the mean variance of the 500 model predictions at each sample point and then obtaining the mean of those variances by land cover class.

2.6.1. Variable importance and correlations

Model strength was compared on the basis of mean RMSE (Eq. (2)) and mean pseudo R^2 (Eq. (3)) for the bootstrapped models, calculated for each model using out-of-bag samples as follows:

$$\text{RMSE} = \sqrt{\frac{\Sigma_{i=1}^{n}(\hat{x}_i - x_i)^2}{n}} \tag{2}$$

$$\text{Pseudo } R^2 = 1 - \frac{\Sigma_{i=1}^{n}(x_i - \hat{x}_i)^2}{\Sigma_{i=1}^{n}(x_i - \bar{x})^2} \tag{3}$$

where x_i is the ith observation from the out-of-bag population with n total observations, $\hat{x}_i$ is the ith prediction, and $\bar{x}$ is the mean of the out-of-bag observations.

Both metrics were calculated by comparing observed photo interpretations to predicted values obtained for each random forest model using the 'predict' function in R's randomForest package (version 3.4.2; Liaw and Wiener, 2002).

High levels of correlation among predictor variables can reduce those variables' relative importance. Thus correlation matrices were generated and assessed in order to better inform variable importance analysis. Correlation matrices and variable importance measures were aggregated for each of the model types to determine the relative contribution of different groups of variables in the model. Once correlation was considered, mean decrease in node impurity was obtained for all variables in each combination of model type, time period, and region. This metric was obtained using the R randomForest package (Liaw and

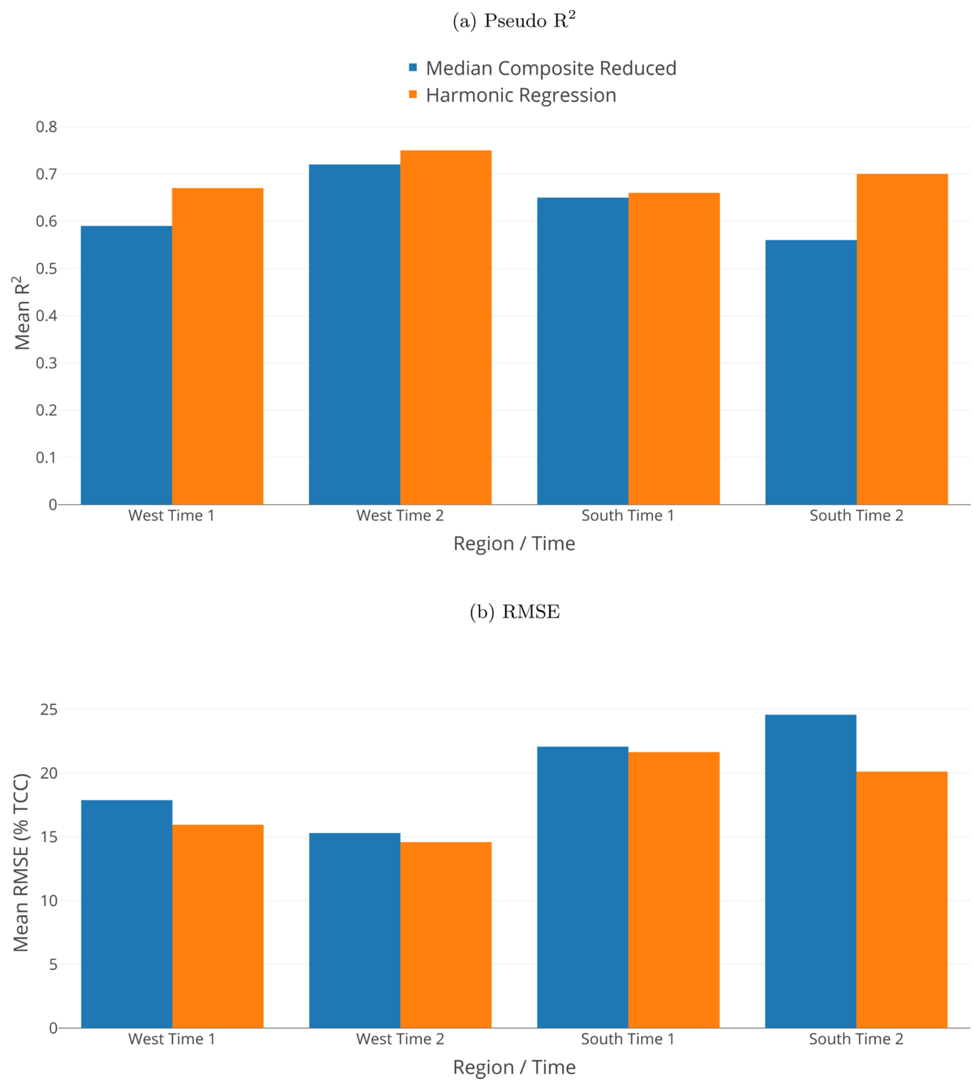

Fig. 6. Comparison of models using harmonic regression coefficients for NDVI, SWIR1 and SWIR2 to the median composite values for those same three bands.

Wiener, 2002). Mean decrease in node impurity identifies the amount that node variance (an indication of impurity) decreases at each split and averages that decrease by predictor variable.

The ten variables whose removal would contribute to the largest decrease in node impurity were ranked, with the largest decrease given a point value of 10 (for most important), and the 10th largest decrease given a 1. All other variables were given a ranking of 0. For each of the four model classes, these rankings were added for all 500 model iterations. Thus, a variable that was considered to be most important for all 500 iterations would receive a value of 5000, while variables that were not considered to be in the top ten most important for any iterations, would be given a point value of 0. This method reflects a similar rank score technique used in Schroeder et al. (2017). Due to the presence of correlated variables in our suite of predictor variables, these rank interpretations were evaluated primarily by comparing the relative contribution of relevant groups of predictor variables instead of the individual variables themselves.

2.6.2. Inherent dimensionality

The inherent dimensionality of each of the four predictor variable sets was assessed using principal components analysis (PCA). PCA utilizes a linear transformation which rotates the input images so that the original axes are orthogonal to one another with decreasing variance for each successive component (Eklundh and Singh, 1993). This methodology is commonly used for many remote sensing applications, including data reduction and feature extraction (Ren et al., 2014; Eklundh and Singh, 1993). PCA assumes (1) a high signal-to-noise ratio, and, (2) for some analytical applications, that input variables have a Gaussian distribution. Landsat data, the only satellite imagery used in this study, has a high signal-to-noise ratio (Barsi and Markham, 2013), and use of PCA for descriptive, exploratory applications does not require strict adherence to distribution assumptions (Jolliffe et al., 2016).

In PCA, eigen values resulting from the transformation can be used to determine the proportion of total variance explained in each resulting component. Typically a large portion of the variance from the original dataset can be found in the first principal component, with decreasing contributions in each successive component. The final components usually have relatively little variance. Therefore, using only the sequential number of bands which collectively represent a large proportion of the original data (usually 95–99%), data can be reduced. Inherent dimensionality is defined in this study as the number of principal components necessary to explain 99% of the cumulative variance. This method is used by Schlamm et al. (2008), as well as Lungu et al. (2017)). We utilize the inherent dimensionality to quantify and compare the variance explained by different groups of predictor variables as a proxy for the amount of unique information they provide.

2.6.3. Mapping TCC across the landscape

For each of the four model ensembles (per date/region), maps were made using the model whose pseudo R^2 was the closest to the mean

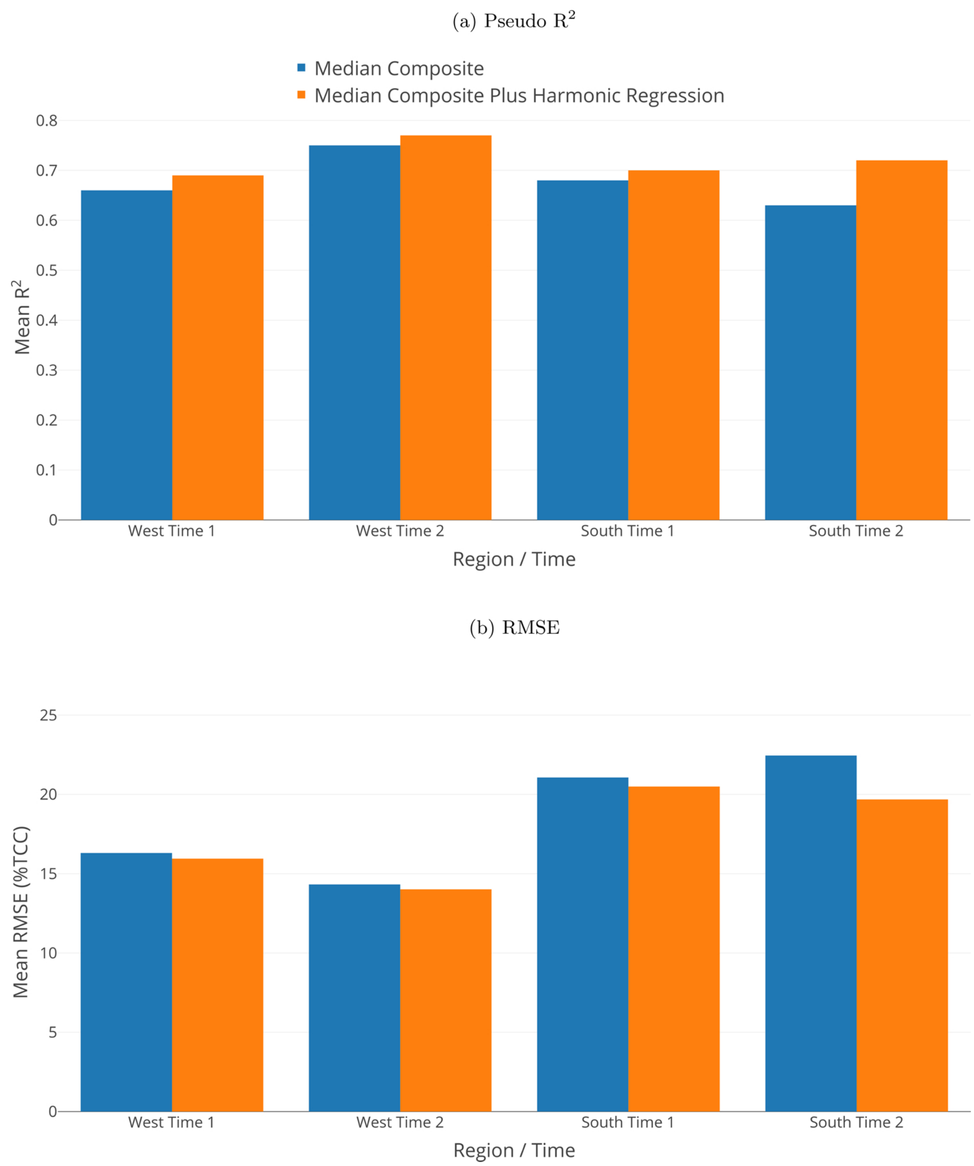

Fig. 7. Comparison of models using the full suite of median composite and terrain variables used in the original 2011 TCC product to models using the full suite of median composite and terrain variables used in the 2011 TCC product *plus* the harmonic regression coefficients from NDVI, SWIR1, and SWIR2.

pseudo R^2 from 500 iterations. We did not use the model with the highest pseudo R^2 to avoid selecting an over-fit model. This model was applied on a pixel-by-pixel basis using a stack of all model inputs. We then conducted both qualitative visual inspections and quantitative assessments. Simple image differencing was used to highlight spatial differences in models. These difference maps were reclassified to identify areas with greater than |30%| difference in tree canopy cover to identify patterns or land features where the models disagreed.

2.6.4. *Variance of predictions across land cover types and TCC values*

For each plot location used in the training dataset the 500 out-of-bag predictions were used to calculate the plot variance in predicted TCC. Plot variance for predictions can give a sense for how consistently a value of TCC is predicted at a location. Land cover classes were obtained from NLCD for each plot location from the training dataset. Plot variance metrics were averaged for plots in each NLCD land class. These values were plotted in a bar chart to compare the variance associated with each land cover class and to observe patterns relating these metrics to land cover.

3. Results

Mean pseudo R^2 was stable, across all four models, at 200 iterations and beyond. As such, the mean pseudo R^2 and mean RMSE at 500 iterations was compared among the resulting 500 random forest models. Using the same input bands (NDVI, SWIR1, and SWIR2; objective 1), harmonic regression coefficients slightly outperformed the median composite values (Fig. 6). Minimum and maximum pseudo R^2 values for the 500 random forest models (not shown) range from a low of 0.55 to a high of 0.82 for the harmonic regression model and from a low of 0.44 to a high of 0.80 for the median composite model. A similar slight improvement with harmonic regression was found for RMSE, with a reduced mean RMSE in all instances (Fig. 6).

To address our second research question (comparing the models using the full suite of median composite and terrain variables used in the original 2011 TCC product to models using all those variables *plus* the harmonic regression coefficients from NDVI, SWIR1, and SWIR2) we again compared the mean pseudo R^2 and mean RMSE at 500 iterations. The full (hybrid) model slightly outperforms the other models in all instances (Fig. 7). The full (hybrid) model also has the

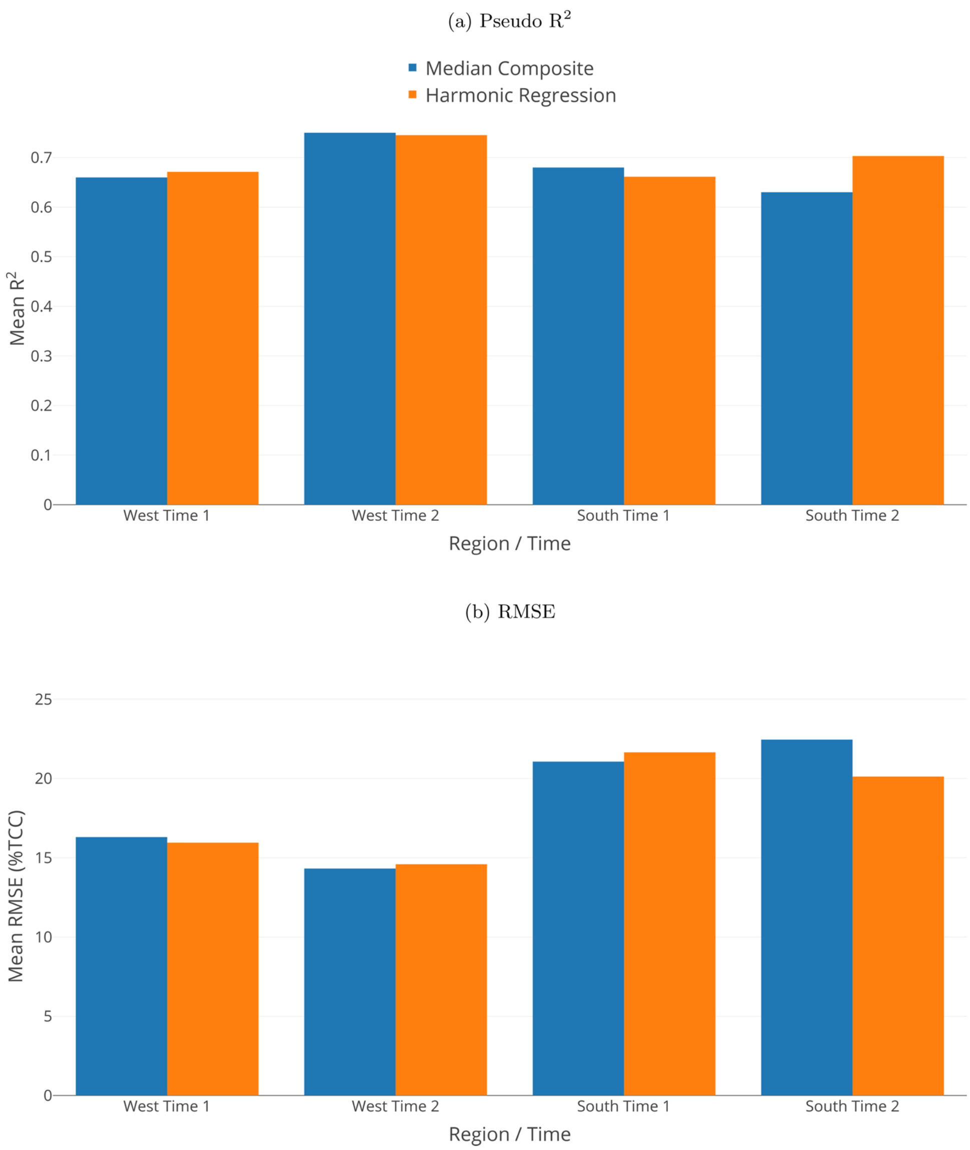

Fig. 8. Comparison of models using the full suite of median composite and terrain variables used in the original 2011 TCC product to models using the harmonic regression coefficients for NDVI, SWIR1, and SWIR2.

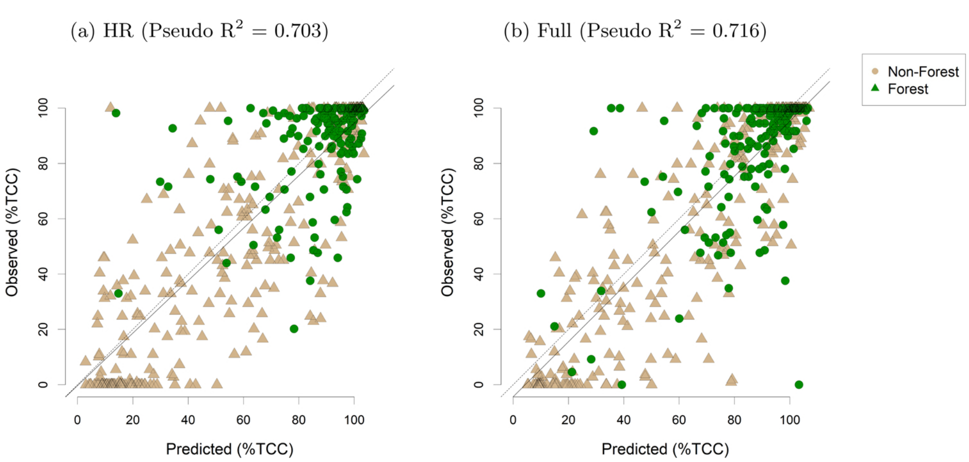

Fig. 9. Observed vs. predicted TCC for mean (with respect to pseudo R^2 out of 500) Time 2 models for the southern study area for the two models using harmonic regression coefficients as part of the predictor variable suite, colored by 2011 NLCD land cover at sample locations. Brown triangles are non-forest; green circles are forest. The dotted line represents the 1:1 relationship between x and y axes ($y = x$) and the solid line represents the linear relationship between observed and predicted TCC. The equation for this linear relationship is $y = 0.9489 * x - 0.2913$ for the HR model and $y = 0.996 * x - 4.301$ for the full model. Variable to axis assignment follows current best practice guidelines (Piñeiro et al., 2008). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

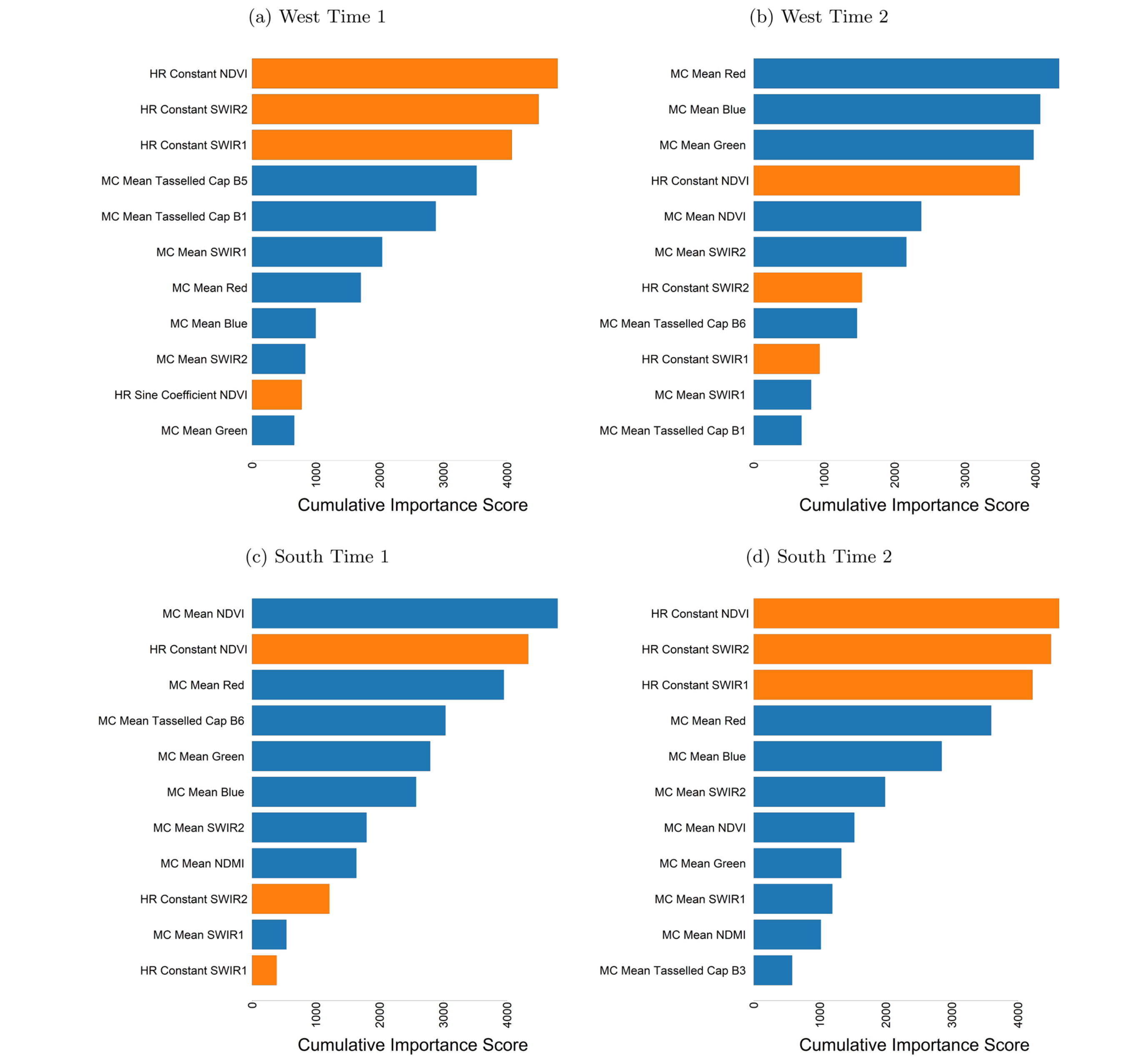

Fig. 10. Top 11 variable importances from full random forests model for each region/time. Median composite means are shown in blue, harmonic regression coefficients in orange. Variable importances were calculated as follows: (1) For each iteration, rank the ten variables whose removal would contribute to the largest increase in node impurity, largest = 10 and 10th (smallest) = 1. (2) Assign all other variables zero. (3) Repeat for 500 iterations (independent model runs), sum to obtain cumulative importance score for each variable. A variable that was considered to be most important for all 500 iterations would receive a value of 5000, while variables that were not in the top ten most important for any iterations, would be given a point value of 0. Abbreviations: HR, harmonic regression; MC, median composite; NDVI, normalized difference vegetation index; SWIR, shortwave infrared. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

highest min/max pseudo R^2, in all instances (not shown), ranging from a low of 0.58 to a high of 0.83.

A comparison of models using the full suite of median composite and terrain variables used in the original 2011 TCC product (38 variables) to the harmonic regression coefficients from NDVI, SWIR1, and SWIR2 (9 variables) revealed mixed model performance (Fig. 8). The 38 median composite and terrain variables outperformed the 9 harmonic regression coefficients in two regions/times (West Time 2 and South Time 1), with the harmonic coefficients leading to an improved model in the other two (West Time 1 and South Time 2).

Observed vs. predicted scatterplots show similar trends. An example for the HR NDVI and SWIR South Time 2 model with the maximum R^2 (0.78), and differentiated by forest/nonforest (according to NLCD land cover class), is shown as Fig. 9. There is more unexplained variance in the south, which is consistent across all models (not shown). It is also evident that the observed data contains more 0% and 100% values than are predicted by the model. In the observed data in the south, most of the forested land cover has high tree canopy cover, while the lower tree

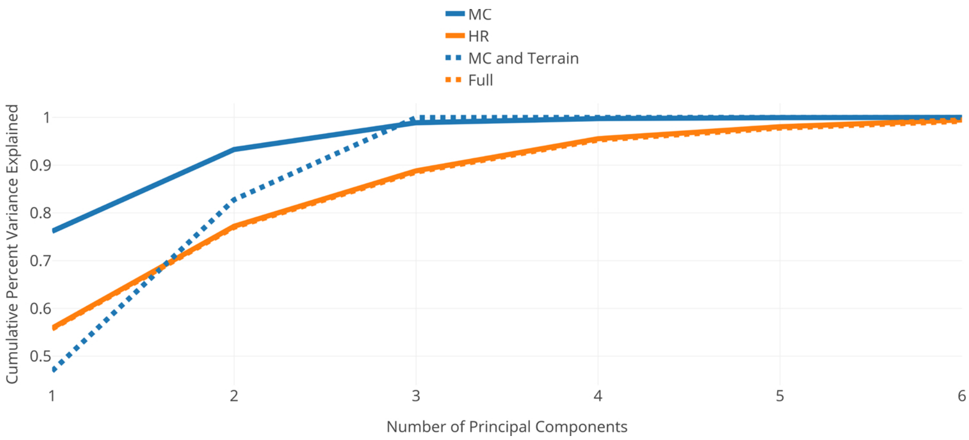

Fig. 11. Eigen analysis of predictor variable sets (South Time 2 shown). The cumulative percent variance explained for a given principal component is calculated as the sum of the eigen values up to and including that principal component divided by the sum of all eigen values for that predictor variable set. Only six principal components are shown for clarity, but the number of principal components for a given predictor variable set is the same as the number of predictor variables in that set. Inherent dimensionality as used here is the number of principal components it takes to explain nearly all the variance, defined as 99%. Note the following: (1) harmonic regression coefficients, alone, or in combination with other predictor variables, have higher inherent dimensionality (and thus information content), and (2) the quantity of available information provided by the harmonic regression coefficients is not substantially increased when looking at the full suite of predictor variables (47 bands), though this is not an indication of the utility of the information provided in predicting TCC.

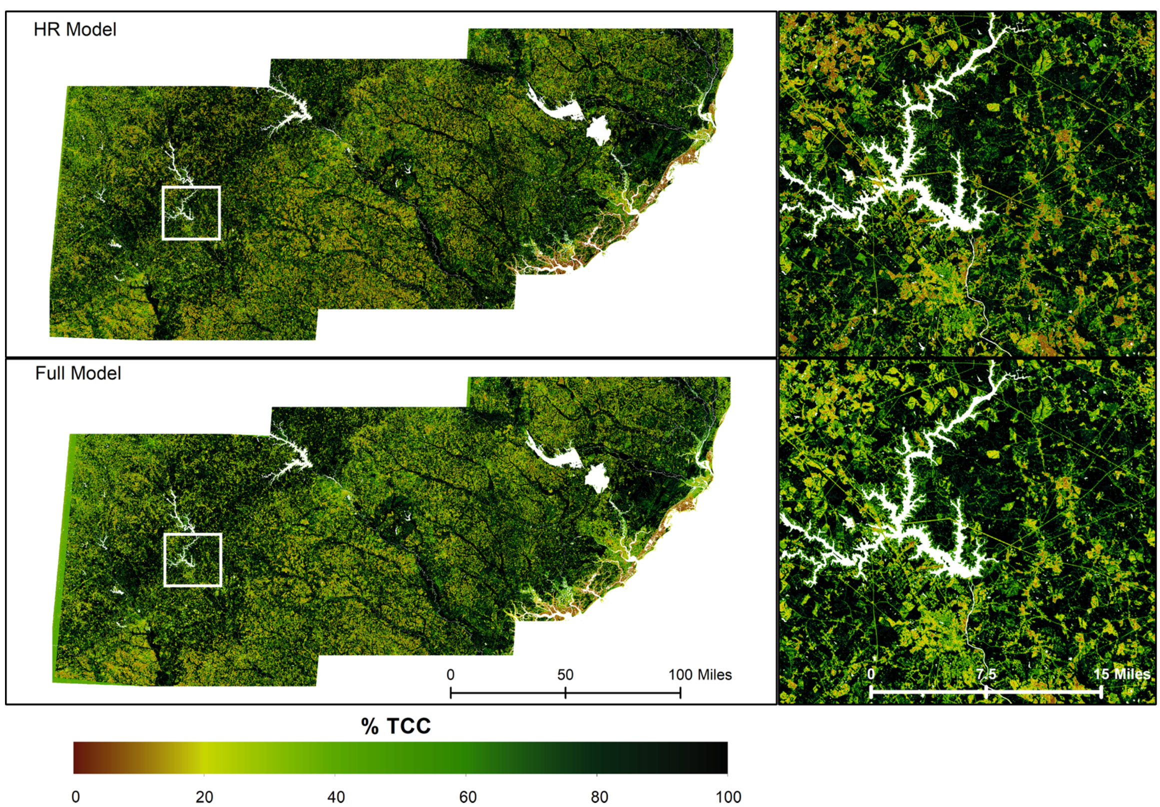

Fig. 12. TCC maps (Time 2) for the southern study area for the two models (*HR* and *Full*) using harmonic regression coefficients as part of the predictor variable suite.

canopy cover values are dominated by nonforest, which includes pasture, grassland, scrub/shrub, etc. There is a more uniform distribution of canopy cover for the samples in the west (excluding zeroes; Fig. 4), which at least partially explains the stronger model performance there.

3.1. Variable importance and correlations

Variable importances for the full (hybrid) model are shown in (Fig. 10). The harmonic regression NDVI constant is always one of the top four most important predictors. In all cases, all three harmonic regression constants and all three median composite mean visible bands are within the top 11 most important variables. The correlations between our top predictor variables and tree canopy cover are shown for the top three variables from each of the models (Table 8). Note that the three predictor variables most correlated with TCC are the harmonic regression constants, and that NDVI is the most important of the bands/ratios. Table 9 shows the correlations between the two NDVI predictor variables and tree canopy cover for each region/time. Note that the harmonic regression constant is always as well or better correlated with tree canopy cover than the median composite value.

3.2. Inherent dimensionality

The principal components analysis of our different sets of input parameters revealed that the models that include harmonic coefficient variables require more principal components to reach 99% cumulative variance than models without these variables (Fig. 11). The full suite of variables (*Full*) had the greatest inherent dimensionality, followed by the harmonic coefficients alone (*HR*), then the median composite and terrain variables (*MC and Terrain*), and finally the median composite

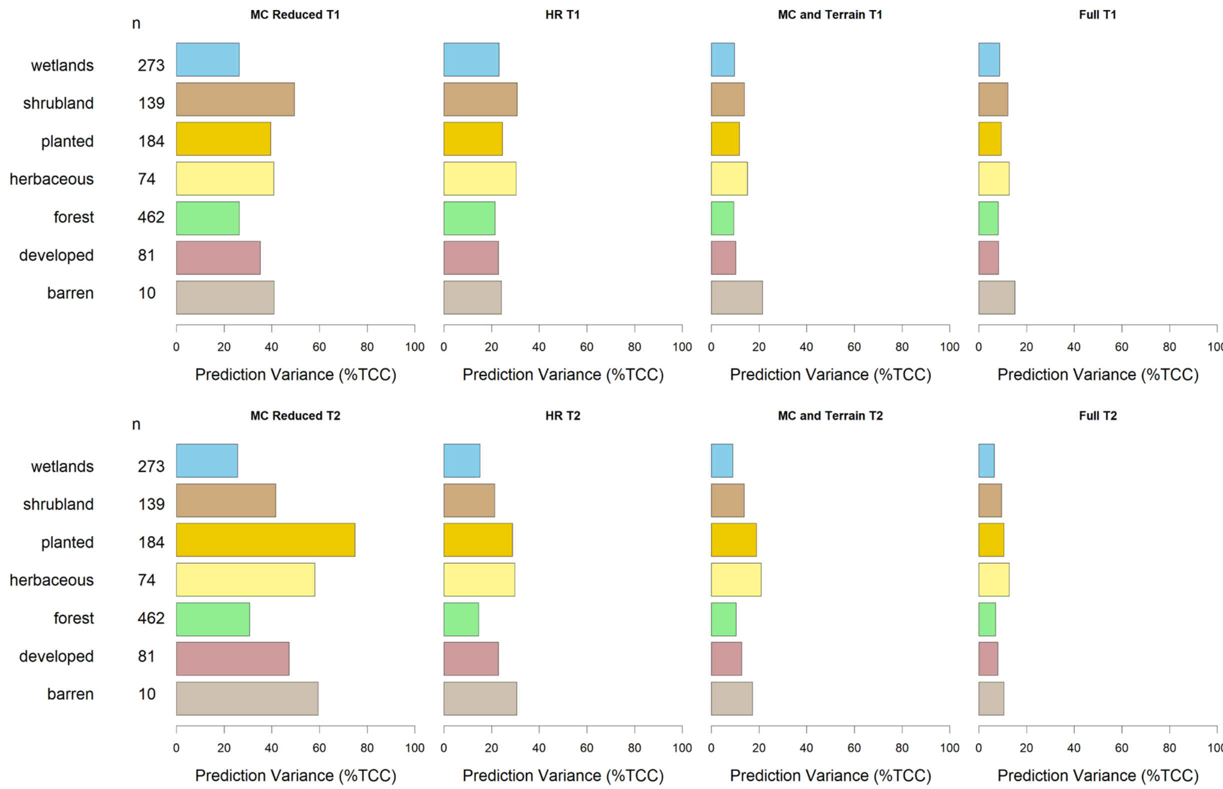

Fig. 13. Mean variance by NLCD land cover class (water excluded) for South, all models.

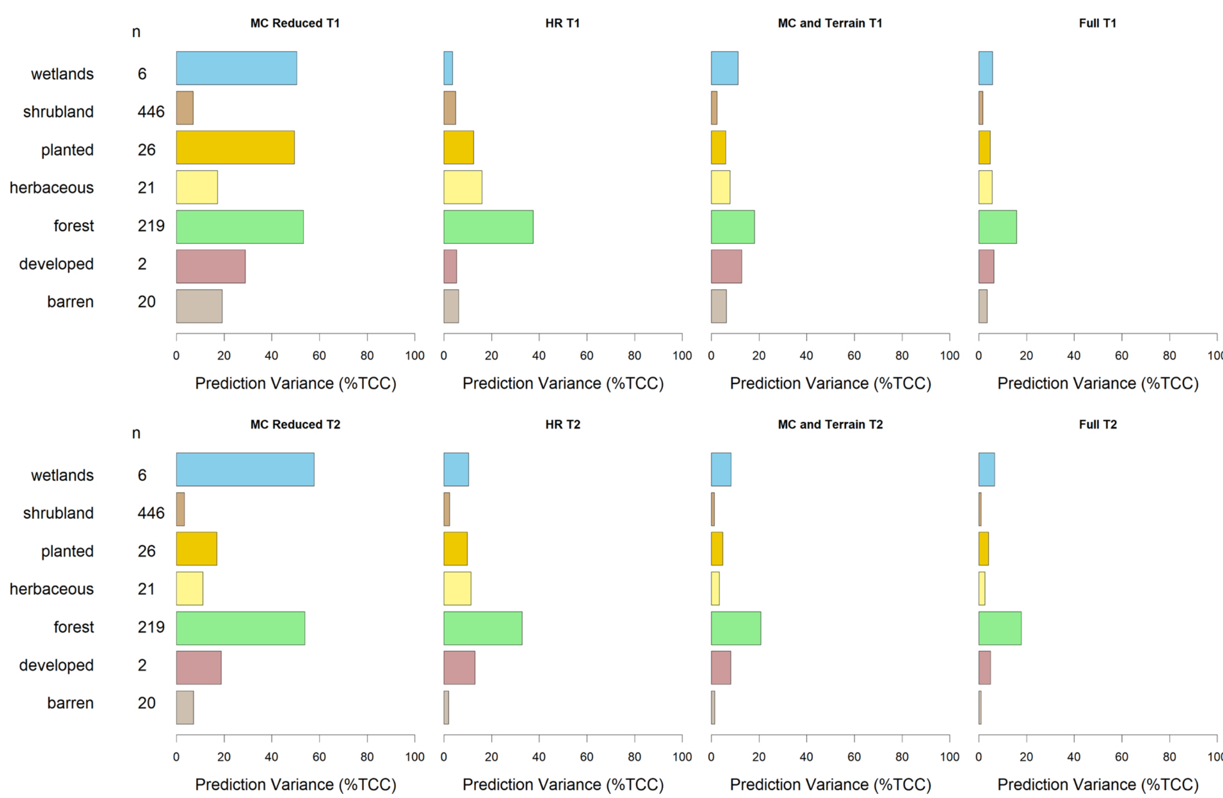

Fig. 14. Mean variance by NLCD land cover class (water excluded) for West, all models.

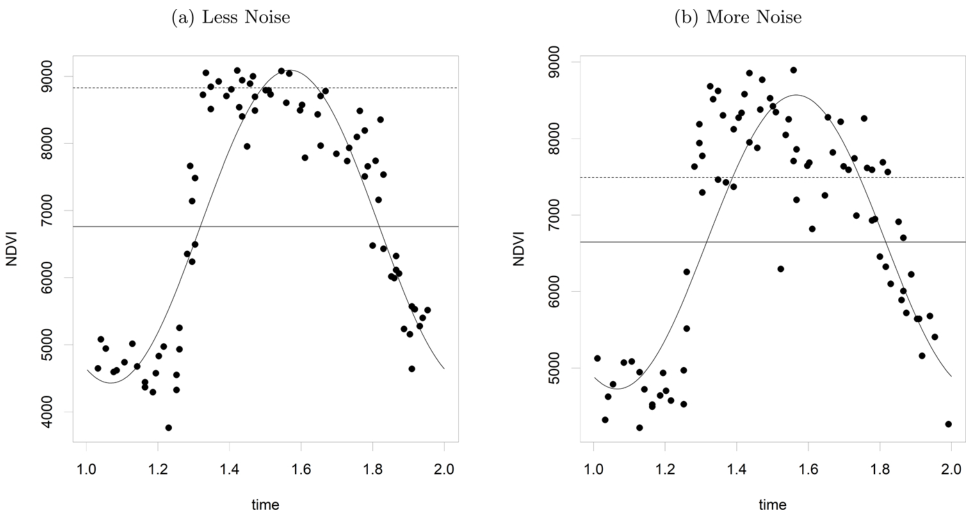

Fig. 15. Observations from two similar South Time 1 plots with fitted harmonic regression line (solid harmonic fit), harmonic regression constant (solid horizontal line), and median composite value (dotted horizontal line) shown. Note that the median composite value is properly located near the peak of the growing season maximum when little noise is present (a, left), but is lower with noisy observations (b, right).

Table 7
Model-specific predictor variables. Abbreviations: MC, median composite; HR, harmonic regression; NDVI, normalized difference vegetation index; NIR, near infrared; SWIR, shortwave infrared; NDMI, normalized difference moisture index; TC, tasseled cap; DEM, digital elevation model.

Model	Number of predictor variables	Variable class	Variables
MC reduced	6	MC[a]	NDVI SWIR1 SWIR2
HR	9	HR[b]	NDVI SWIR1 SWIR2
MC and terrain	38	MC[a]	Blue Green Red NIR SWIR1 SWIR2 NDVI NDMI TC (all)
		Terrain[a]	DEM Slope Aspect sin, cos aspect
Full	47	MC[a]	Blue Green Red NIR SWIR1 SWIR2 NDVI NDMI TC (all)
		Terrain[a]	DEM slope aspect sin, cos aspect
		HR[b]	NDVI SWIR1 SWIR2

[a] *Focal $\bar{x}$ and *std* in 3 × 3 window.
[b] a_0, a_1, b_1 (Eq. (1)), the constant, cosine amplitude, and sine amplitude of the Fourier fit.

NDVI and SWIR variables (*MC Reduced*). The fact that the harmonic regression coefficients for three bands alone convey more information than the 38 variables in the *MC and Terrain* model is striking. However, increased information in a set of predictor variables does not necessarily mean that the target variable will be better estimated.

3.3. Mapping TCC across the landscape

Maps generated from the harmonic regression and full models are generally consistent across the landscape (Fig. 12). Careful examination of these images revealed some anomalies due to cloud cover or striping from (SLC-off) Landsat 7. This causes a reduction in the number of input observations that inform the original harmonic regression fit. Areas with very few observations due to annually or seasonally persistent cloud cover could result in similar anomalies, though these were not observed.

3.4. Variance of predictions across land cover types and TCC values

Tree canopy cover models show some variability in plot-level variance when aggregated by land cover class (Figs. 13 and 14). Mean variance in predicted value by land cover reveals the greatest inconsistency in prediction across land cover classes in the *MC Reduced* model with the most consistent predictions overall in the *Full* model. In the South, variance is lower in the forest and wetland categories. Higher variance can be observed in the other land cover types, without consistent pattern across model types and time periods. In the West, there is higher variance in forest plots, and in some cases (*MC Reduced* T1, *MC and Terrain* T1, *MC Reduced* T2) in the wetland plots as well. There is also lower variance in shrubland plots, and in some cases barren plots. Across land cover classes, models using harmonic regression coefficients outperform those in which the median composite values are used for the SWIR bands and NDVI (*HR* vs. *MC Reduced*). Adding other bands/indices along with the terrain variables resulted in additional decreases in variance across land cover classes (*MC and Terrain*, *Full*).

4. Discussion

All predictor sets performed reasonably well for a given region. However, in the direct comparison (objective 1; *MC Reduced* vs. *HR*) using the same index (NDVI) and wavelength bands (SWIR1 and SWIR2), the models using harmonic regression coefficients consistently outperformed those using median composite means and standard deviations with respect to both R^2 and RMSE across study areas and time periods (Fig. 7). This is partially explained by the consistently high variable importance metrics from the harmonic regression coefficients, particularly the constants, despite the high correlation among predictors. The improved utility of harmonic regression coefficients with respect to median composite values is also supported by the higher

Table 8
Correlation matrix of seven-most important predictor variables and tree canopy cover from the South, time 2 model. The three most important variables from the full random forests model for each region/time (Fig. 10) are shown; there are only seven since the same predictor variables could be among the three most important across regions/times. Abbreviations: TCC, tree canopy cover; MC, median composite; B, blue; G, green; R, red; a_0, harmonic regression constant; NDVI, normalized difference vegetation index; SWIR, shortwave infrared.

	MC B	MC G	MC R	MC NDVI	a_0 NDVI	a_0 SWIR1	a_0 SWIR2
TCC	−0.64	−0.64	−0.66	0.61	0.74	−0.68	−0.70
MC B		0.96	0.96	−0.69	−0.74	0.67	0.76
MC G			0.96	−0.57	−0.71	0.73	0.79
MC R				−0.68	−0.74	0.73	0.80
MC NDVI					0.78	−0.32	−0.44
a_0 NDVI						−0.66	−0.75
a_0 SWIR1							0.96

Table 9
Correlations between TCC in each region/time and the two NDVI predictor variables. Abbreviations: TCC, tree canopy cover; MC, median composite; a_0, harmonic regression constant; NDVI, normalized difference vegetation index.

	MC NDVI	a_0 NDVI
West Time 1	0.77	0.82
West Time 2	0.83	0.85
South Time 1	0.74	0.74
South Time 2	0.61	0.74

inherent dimensionality of the harmonic regression predictors (Fig. 11) and the high correlations of the harmonic regression constants with TCC (Table 8; note that the constants have the highest correlation coefficients of the seven most important predictor variables).

While there are multiple possible explanations for why the harmonic regression constant outperforms the median composite value for a given band, the two most likely are (1) the decreased sensitivity of the harmonic regression constant to systematic noise (Moody and Johnson, 2001; De Beurs and Henebry, 2010) and (2) the use of another sensor (MODIS, via GLOVIS) to define the peak NDVI interval within which the median value is computed. Noise tends to lower the median composite value (Fig. 15), leading to inconsistencies in the median composite value for the same TCC. With respect to the second possible explanation, the increased temporal resolution of MODIS makes it a sensible choice for choosing the interval within which the median value is computed, but MODIS-derived NDVI, while similar to NDVI computed from the Landsat sensors, is slightly different (Steven et al., 2003).

The coefficients derived from harmonic regression of all-available cloud-free moderate resolution earth resource satellite data have been used in a variety of earth science applications, including image classification (Liang, 2001), stratified estimation of dynamic forest biophysical parameters (Brooks et al., 2016), estimation of static forest biophysical parameters besides TCC (Wilson et al., 2018), and, with this study, TCC. However, despite these successes using harmonic regression, there are alternative approaches to curve fitting that have been developed or tested using the same or similar data sets that might prove to be superior to harmonic regression for this and related applications. These include the double hyperbolic tangent model (Vrieling et al., 2018) applied to Sentinel 2 observations and a Bayesian hierarchical modeling framework (Senf et al., 2017) applied to Landsat observations, among others. While Atkinson et al. (2012) found harmonic regression the best of four tested approaches (Fourier analysis, asymmetric Gaussian model, double logistic model, and the Whittaker filter) for estimating phenological parameters reliably using Medium Resolution Imaging Spectrometer (MERIS) composite data, their study did not use moderate resolution earth resource satellite data nor did it test the more modern approaches just noted. As such, use of alternative curve fitting approaches as a precursor to TCC estimation using moderate resolution earth resource satellite data is a potentially fruitful area of further inquiry.

Mapping yielded visually consistent results across models. However, due to the inclusion of Landsat 7 SLC-off images in our time-series, subtle striping can be observed in models with harmonic regression variables in regions with fewer unmasked pixel observations. This finding is consistent with other studies (Wilson et al., 2018), and may be a reason to exclude Landsat 7 observations when feasible.

Mean variance in predicted values yielded inconsistent results across regions (Figs. 13 and 14). In the South, the two land cover types with more samples (wetlands and forest areas) showed lower variance, as expected since there was more training data available for those types. In the West, shrublands were the most prevalent land cover type and had consistently low variance values as well. However, contrary to these other examples, the land cover type with most consistently high variance, at least in the West, was forest, which is the second most prevalent land cover type in terms of sample points. There are several possible reasons for the high mean prediction variance in NLCD forest in the West, including the following: (1) It is common to confuse shrublands (which, if there are no trees present, should have a TCC value of 0) with forest areas, both spectrally and during photo interpretation. This could lead to spectrally similar training data having drastically different observed TCC values, thus limiting the predictive ability of the model. (2) The majority of the forest types in the west are coniferous which may not provide as much contrast in reflectance between tree and non-tree pixels and may not show as much variation over time resulting in more influence from noise. In all cases (land cover, region, and time), models with large numbers of predictor variables (*MC and Terrain*, 38; *Full*, 47; Table 7) reduced variance. Additionally, this observation may also indicate that the *MC and Terrain* variables, while possessing less information overall, may have a greater explanatory power than the harmonic regression coefficients in predicting TCC.

Given recent advances in high performance computing platforms that host the Landsat archives (i.e., Google Earth Engine, Gorelick et al., 2017; NASA Earth Exchange, Nemani, 2011), the selection of an algorithm that reduces input data and pre-processing streams (i.e., image downloads, cloud masking, and development of predictor variables) is not a major factor of consideration. Instead, the focus can be more on striving for analysis techniques that result in more parsimonious predictive models, with fewer variables but higher explanatory power. This concept represents the biggest advantage of harmonic regression over the median composite approach. The inherent dimensionality shown in Fig. 11 highlights the increased information contained in the models that contain harmonic regression predictors. This makes sense, given the large number of input bands/observation dates over which the Fourier curve is fit (ranging from 60–192 for the 3-year training period). This is illustrated with a simple correlation between TCC and the important predictor variables (Table 8), which indicates that the harmonic regression constants are more strongly correlated to TCC than any median composite mean or standard deviation. Moreover, though not an objective of this study, a comparison of Figs. 6 and 7 reveals instances where *HR* (9 predictor variables) is equivalent to or outperforms *MC and Terrain* (38 predictor variables).

The variable importance results (Fig. 10 and Table 8) highlight the importance of the visible bands to the prediction of TCC, although the correlation matrix (Table 8) confirms that these bands are highly correlated. Our analysis did not include a visible band in the harmonic regression predictor variables, so direct comparison is not possible. However, the results suggest that harmonic regression coefficients from a visible band (or panchromatic band as shown in Fig. 2) could be valuable to TCC prediction. While this study found the median composite values derived from the tasseled cap bands to be less important than those derived using the visible bands, harmonic regression coefficients derived from tasseled cap Landsat time series stacks (especially, and atypically, TC band 6) could also be useful.

Finally, we would like to reiterate that there are instances where striping from Landsat 7 SLC-off images has produced visually inconsistent predictions from harmonic regression coefficients. We have seen that these inconsistencies relate to differences in the number of observations available to the harmonic regression algorithm. Even absent striping, cloud-prone areas or areas with cloud cover over a consistent portion of the season may not enable sufficient observational representation of annual phenology, thus reducing the value of harmonic regression coefficients fit to a poorly articulated curve. Median composite methodologies could also be susceptible to these issues, especially if observations are lost during the peak of the growing season from striping or clouds.

5. Conclusions

We conclude that harmonic regression coefficients are better predictors of TCC than median composite variables for the same vegetation index and wavelength bands. Correlations support this finding, with improved R^2 and RMSE values for the harmonic regression coefficient models when compared to the median composite reduced models, as well as increased inherent dimensionality in harmonic regression coefficient variables. Further, plot-level prediction variance is lower for the harmonic regression coefficient models than it is for the median composite reduced models for almost every land cover type, regardless of region or time period.

We further conclude that harmonic regression coefficients can be added to a full suite of median composite and terrain variable to improve the quality of TCC models. Variable importance analysis indicates that harmonic regression variables are uniquely instrumental to model success, with the harmonic regression constants appearing consistently in the top 11 predictor variables. Correlations further support this finding, with improved R^2 and RMSE values for the full model when compared to the median composite and terrain model, and overall. Finally, plot-level prediction variance is lowest for the full models than any other model for almost every land cover type, regardless of region or time period. Overall, we have found that harmonic regression coefficients can be used to estimate TCC, both on their own, and when combined with a suite of median composite and terrain variables. The 9 harmonic coefficient variables alone successfully predict TCC without sacrificing predictive power when compared to models drawing from greater numbers of variables (*MC and Terrain*, n = 38; *Full*, n = 47; Table 7).

Acknowledgements

This study was funded by the 'TCC 2021 Research and Development' grant, the 'NLCD Percent Tree Canopy Cover Layer - Evaluation of Data Sources to Predict Canopy Cover', and the 'NLCD Percent Tree Canopy Cover Layer - Evaluation of the 2011 Product and Alternative Approaches to Detect Change' grant from the USDA Forest Service. It was also supported in part by the Virginia Agricultural Experiment Station and the McIntire-Stennis Program of NIFA, USDA, "Detecting and Forecasting the Consequences of Subtle and Gross Disturbance on Forest Carbon Cycling").

References

Asner, G.P., Lobell, D.B., 1996. A biogeophysical approach for automated SWIR unmixing of soils and vegetation. Remote Sens. Environ. 74, 99–112.

Atkinson, P.M., Jeganathan, C., Dash, J., Atzberger, C., 2012. Inter-comparison of four models for smoothing satellite sensor time-series data to estimate vegetation phenology. Remote Sens. Environ. 123, 400–417.

Barsi, J.A., Markham, B.L., 2013. Early radiometric performance assessment of the Landsat-8 Operational Land Imager (OLI). Proc. SPIE 8866:8866 – 8866 – 12.

Brandt, M., Hiernaux, P., Tagesson, T., Verger, A., Rasmussen, K., Diouf, A.A., Mbow, C., Mougin, E., Fensholt, R., 2016. Woody plant cover estimation in drylands from earth observation based seasonal metrics. Remote Sens. Environ. 172, 28–38.

Brooks, E.B., 2019. Exponentially Weighted Moving Average Change Detection – Script and Sample Data. Blacksburg: Virginia Tech Department of Forest Resources and Environmental Conservation. Virginia Tech Department of Forest Resources and Environmental Conservation, Blacksburg, doi:10.7294/W4WD3XHK.

Brooks, E.B., Coulston, J.W., Wynne, R.H., Thomas, V.A., 2016. Improving the precision of dynamic forest parameter estimates using Landsat. Remote Sens. Environ. 179, 162–169. https://doi.org/10.1016/j.rse.2016.03.017.

Brooks, E.B., Thomas, V.A., Wynne, R.H., Coulston, J.W., 2012. Fitting the multitemporal curve: a Fourier series approach to the missing data problem in remote sensing analysis. IEEE Trans. Geosci. Remote Sens. 50 (9), 3340–3353. https://doi.org/10.1109/TGRS.2012.2183137.

Coulston, J.W., Blinn, C.E., Thomas, V.A., Wynne, R.H., 2016. Approximating prediction uncertainty for random forest regression models. Photogramm. Eng. Remote Sens. 82 (3). https://doi.org/10.14358/PERS.82.3.189.

Coulston, J.W., Moisen, G.G., Wilson, B.T., Finco, M.V., Cohen, W.B., Brewer, C.K., 2012. Modeling percent tree canopy cover – a pilot study. Photogramm. Eng. Remote Sens. 78ID – 22 (7), 715–727. https://doi.org/10.14358/PERS.78.7.715.

De Beurs, K.M., Henebry, G.M., 2010. Spatio-temporal statistical methods for modelling land surface phenology. Phenological Research: Methods for Environmental and Climate Change Analysis. Springer, Dordrecht. https://doi.org/10.1007/978-90-481-3335-2.

DeFries, R.S., Field, C.B., Fung, I., Justice, C.O., Los, S., Matson, P.A., Matthews, E., Mooney, H.A., Potter, C.S., Prentice, K., Sellers, P.J., Townshend, J.R.G., Tucker, C.J., Ustin, S.L., Vitousek, P.M., 1995. Mapping the land surface for global atmosphere-biosphere models: toward continuous distributions of vegetation's functional properties. J. Geophys. Res. 100 (D10), 20. https://doi.org/10.1029/95JD01536.

Eklundh, L., Singh, A., 1993. A comparative analysis of standardised and unstandardised principal components analysis in remote sensing. Int. J. Remote Sens. 14 (7), 1359–1370. https://doi.org/10.1080/01431169308953962.

Gessner, U., Machwitz, M., Conrad, C., Dech, S., 2013. Estimating the fractional cover of growth forms and bare surface in savannas. A multi-resolution approach based on regression tree ensembles. Remote Sens. Environ. 129, 90–102. https://doi.org/10.1016/j.rse.2012.10.026.

Goeking, S.A., Liknes, G.C., Lindblom, E., Chase, J., Jacobs, D.M., Benton, R., 2012. A GIS-Based Tool for Estimating Tree Canopy Cover on Fixed-Radius Plots Using High-Resolution Aerial Imagery. In: Morin, Randall S., Liknes, Greg C. (Eds.), Moving from status to trends: Forest Inventory and Analysis (FIA) symposium 2012; 2012 December 4-6. U.S. Department of Agriculture, Forest Service, Northern Research Station, Baltimore, MD, pp. 237–241 Newtown Square, PA Gen. Tech. Rep. NRS-P-105.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google Earth Engine: Planetary-Scale Geospatial Analysis for Everyone. Remote Sens. Environ. 202, 18–27. https://doi.org/10.1016/j.rse.2017.06.031.

J.A. Greenberg. spatial.tools: R Functions for Working with Spatial Data 2018; R package version 1.6.0 https://CRAN.R-project.org/package=spatial.tools.

Hansen, M.C., DeFries, R.S., 2004. Detecting long-term global forest change using continuous fields of tree-cover maps from 8-km Advanced Very High Resolution Radiometer (AVHRR) data for the years 1982–99. Ecosystems 7 (7), 695–716. https://doi.org/10.1007/s10021-004-0243-3.

R.J. Hijmans. Raster: Geographic Data Analysis and Modeling 2018; R package version 2.8-19 https://CRAN.R-project.org/package=raster.

Immerzeel, W., Quiroz, R., De Jong, S., 2005. Understanding precipitation patterns and land use interaction in Tibet using harmonic analysis of SPOT VGT-S10 NDVI time series. Int. J. Remote Sens. 26 (11), 2281–2296. https://doi.org/10.1080/01431160512331326611.

Jennings, S., 1999. Assessing forest canopies and understorey illumination: canopy closure, canopy cover and other measures. Forestry 72 (1), 59–74. https://doi.org/10.1093/forestry/72.1.59.

Jolliffe, I.T., Cadima, J., Cadima, J., 2016. Principal component analysis: a review and recent developments. Philos. Trans. R. Soc. 374, 16. https://doi.org/10.1098/rsta.2015.0202.

Jönsson, P., Eklundh, L., 2004. TIMESAT – a program for analyzing time-series of satellite sensor data. Comput. Geosci. 30 (8), 833–845. https://doi.org/10.1016/j.cageo.2004.05.006.

Karlson, M., Ostwald, M., Reese, H., Sanou, J., Tankoano, B., Mattsson, E., 2015. Mapping tree canopy cover and aboveground biomass in Sudano-Sahelian woodlands using landsat 8 and random forest. Remote Sens. 7 (8), 10017–10041. https://doi.org/10.3390/rs70810017.

Liang, S., 2001. Land-cover classification methods for multi-year AVHRR data. Int. J.

Remote Sens. 22 (8), 1479–1493. https://doi.org/10.1080/01431160120833.
Liaw, A., Wiener, M., 2002. Classification and regression by randomforest. R News 2 (3), 18–22. https://www.r-project.org/doc/Rnews/Rnews_2002-3.pdf.
Lobell, D.B., Asner, G.P., Law, B.E., Treuhaft, R.N., 2001. Subpixel canopy cover estimation of coniferous forests in Oregon using SWIR imaging spectrometry. J. Geophys. Res. Atmos. 106 (D6), 5151–5160. https://doi.org/10.1029/2000JD900739.
Lungu, C., Ersali, S., Szefler, B., Pîrvan-Moldovan, A., Basak, S., Diudea, M.V., 2017. Dimensionality of big data sets explored by Cluj descriptors. Studia Universitatis Babes-Bolyai, Chemia 62 (3). https://doi.org/10.24193/subbchem.2017.3.16.
Microsoft Corporation, Weston, S., 2017a. doSNOW: Foreach Parallel Adaptor for the ‘snow’ Package. R package version 1.0.16.. https://CRAN.R-project.org/package=doSNOW.
Microsoft Corporation, Weston, S., 2017b. Foreach: Provides Foreach Looping Construct for R. R package version 1.4.4. https://CRAN.R-project.org/package=foreach.
Moody, A., Johnson, D.M., 2001. Land-surface phenologies from AVHRR using the discrete Fourier transform. Remote Sens. Environ. 75, 305–323. https://doi.org/10.1016/S0034-4257(00)00175-9.
Muoghalu, J.I., Isichei, A.O., 1991. Effect of tree canopy cover on the yield, crude protein and fibre content of forb species in Nigerian Guinea savanna. Vegetatio 95 (2), 167–175. https://doi.org/10.1007/BF00045215.
Nemani, R., 2011. NASA earth exchange: next generation earth science collaborative. In: International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences – ISPRS Archives, Vol. XXXVIII-8/W20. Bhopal. ISPRS.
North, M.P., Kane, J.T., Kane, V.R., Asner, G.P., Berigan, W., Churchill, D.J., Conway, S., Gutiérrez, R.J., Jeronimo, S., Keane, J., Koltunov, A., Mark, T., Moskal, M., Munton, T., Peery, Z., Ramirez, C., Sollmann, R., White, A.M., Whitmore, S., 2017. Cover of tall trees best predicts California spotted owl habitat. Forest Ecol. Manag. 405 (August), 166–178. https://doi.org/10.1016/j.foreco.2017.09.019.
Owens, M.K., Lyons, R.K., Alejandro, C.L., 2006. Rainfall partitioning within semiarid juniper communities: effects of event size and canopy cover. Hydrol. Process. 20 (15), 3179–3189. https://doi.org/10.1002/hyp.6326.
Pandit, R., Polyakov, M., Sadler, R., 2014. Valuing public and private urban tree canopy cover. Aust. J. Agric. Resour. Econ. 58 (3), 453–470. https://doi.org/10.1111/1467-8489.12037.
Peterson, D.W., Reich, P.B., Wrage, K.J., 2007. Plant functional group responses to fire frequency and tree canopy cover gradients in oak savannas and woodlands. J. Veg. Sci. 18, 3–12. https://doi.org/10.1111/j.1654-1103.2007.tb02510.x.
Piñeiro, G., Perelman, S., Guerschman, J.P., Paruelo, J.M., 2008. How to evaluate models: Observed vs. predicted or predicted vs. observed? Ecol. Model. 216 (3–4), 316–322. https://doi.org/10.1016/j.ecolmodel.2008.05.006. arXiv: arXiv:1011.1669v3.
R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
Ren, J., Zabalza, J., Marshall, S., Zheng, J., 2014. Effective feature extraction and data reduction in remote sensing using hyperspectral imaging [applications corner]. IEEE Signal Process. Mag. 31 (4), 149–154. https://doi.org/10.1109/MSP.2014.2312071.
Rouse, J.W., J. Haas, R., Deering, D., Schell, J., Harlan, J., 1974. Monitoring the Vernal Advancement and Retrogradation (Green Wave Effect) of Natural Vegetation. Technical Report. NASA/GSFC Type III Final Report, Greenbelt, MD, USA.
Ruefenacht, B., 2016. Comparison of Three Landsat TM Compositing Methods: A Case Study Using Modeled Tree Canopy Cover 82. Photogramm. Eng. Remote Sens., pp. 199–211. https://doi.org/10.14358/PERS.82.3.199.
Ruefenacht, B., Benton, R., Johnson, V., Biswas, T., Baker, C., Finco, M., Megown, K., Coulston, J., Winterberger, K., Riley, M., 2015. Forest Service Contributions to the National Land Cover Database (NLCD): Tree Canopy Cover Production. In: Stanton, S.M., Christensen, G.A. (Eds.), Pushing boundaries: new directions in inventory techniques and applications: Forest Inventory and Analysis (FIA) symposium 2015. 2015 December 8–10. U.S. Department of Agriculture, Forest Service, Pacific Northwest Research Station, Portland, Oregon, pp. 241–243 Gen. Tech. Rep. PNW-GTR-931.
Schlamm, A.A., Messinger, D.W., Basener, W.F., 2008. Geometric estimation of the inherent dimensionality of a single material cluster in multi- and hyperspectral imagery. Proc. SPIE, Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XIV, 69661G 69661G, 394–403. https://doi.org/10.1117/12.776903.
Schroeder, T.A., Schleeweis, K.G., Moisen, G.G., Toney, C., Cohen, W.B., Freeman, E.A., Yang, Z., Huang, C., 2017. Testing a Landsat-based approach for mapping disturbance causality in U.S. forests. Remote Sens. Environ. 195, 230–243. https://doi.org/10.1016/j.rse.2017.03.033.
Senf, C., Pflugmacher, D., Heurich, M., Krueger, T., 2017. A Bayesian hierarchical model for estimating spatial and temporal variation in vegetation phenology from Landsat time series. Remote Sens. Environ. 194, 155–160. https://doi.org/10.1016/j.rse.2017.03.020.
Shureman, P., 1940. Department of Commerce Coast and Geodetic Survey Manual of Harmonic Analysis, Special Publication No. 98. Technical Report. U.S. Department of Commerce Coast and Geodetic Survey, Washington.
Smith, S.W., 1997. Continuous signal processing. In: Smith, S.W. (Ed.), Digital Signal Processing: A Practical Guide for Engineers and Scientists. Elsevier, pp. 243–260. https://doi.org/10.1016/B978-0-7506-7444-7/50050-9.
Steven, M.D., Malthus, T.J., Baret, F., Xu, H., Chopping, M.J., 2003. Intercalibration of vegetation indices from different sensor systems. Remote Sens. Environ. 88 (4), 412–422. https://doi.org/10.1016/j.rse.2003.08.010.
Toney, C., Shaw, J.D., Nelson, M.D., 2009. A stem-map model for predicting tree canopy cover of forest inventory and analysis (FIA) plots. In: USDA Forest Service Proceedings. January 2009. pp. 1–19.
USDA Farm Service Agency, 2018. NAIP Coverage 2002–2018. US Department of Agriculture Farm Production and Conservation- Business Center, Salt Lake CIty, UT. https://www.fsa.usda.gov/Assets/USDA-FSA-Public/usdafiles/APFO/status-maps/pdfs/NAIP_Coverage_2018.pdf.
Vrieling, A., Meroni, M., Darvishzadeh, R., Skidmore, A.K., Wang, T., Zurita-Milla, R., Oosterbeek, K., O’Connor, B., Paganini, M., 2018. Vegetation phenology from Sentinel-2 and field cameras for a Dutch barrier island. Remote Sens. Environ. 215 (July 2017), 517–529. https://doi.org/10.1016/j.rse.2018.03.014.
Weisstein, E.W., 2010. Fourier Series. Wolfram Research, Inc. (accessed 4 October 2018). http://mathworld.wolfram.com/FourierSeries.html.
Wilson, B.T., Knight, J.F., McRoberts, R.E., 2018. Harmonic regression of Landsat time series for modeling attributes from national forest inventory data. ISPRS J. Photogramm. Remote Sens. 137, 29–46. https://doi.org/10.1016/j.isprsjprs.2018.01.006.
Zhu, Z., Woodcock, C.E., 2012. Object-based cloud and cloud shadow detection in Landsat imagery. Remote Sens. Environ. 118, 83–94. https://doi.org/10.1016/j.rse.2011.10.028. eprint:9808047v2 (arXiv:cond-mat).