Journal of Forestry, 2022, 1–13
https://doi.org/10.1093/jofore/fvab075
Research Article - geospatial technologies
Received October 8, 2021; Accepted January 4, 2022
Advance Access publication February 11, 2022

Research Article - geospatial technologies

Ground-Truthing Forest Change Detection Algorithms in Working Forests of the US Northeast

Madeleine L. Desrochers[1], Wayne Tripp[2], Stephen Logan[2], Eddie Bevilacqua[1], Lucas Johnson[3], and Colin M. Beier[1]

[1]Department of Sustainable Resources Management, SUNY College of Environmental Science and Forestry, Syracuse, NY 13210, USA. [2]F&W Forestry, LLC, USA. [3]Graduate Program in Environmental Science, SUNY College of Environmental Science and Forestry, Syracuse, NY 13210, USA.

*Corresponding author: Email: mldesroc@syr.edu

Abstract

The need for reliable landscape-scale monitoring of forest disturbance has grown with increased policy and regulatory attention to promoting the climate benefits of forests. Change detection algorithms based on satellite imagery can address this need but are largely untested for the forest types and disturbance regimes of the US Northeast, including management practices common in northern hardwoods and mixed hardwood-conifer forests. This study ground-truthed the "off-the-shelf" outputs of three satellite-based change detection algorithms using detailed harvest records and maps covering 43,000 ha of working forests in northeastern New York.

Study Implications: Algorithms performed best in detecting clearcuts, but performed much worse and poorly overall in detecting the partial harvest prescriptions (e.g., shelterwoods, thinnings) that were far more common in our ground-truthing data (and for this region). Among the algorithms tested, Landtrendr was consistently superior at both detecting partial harvests and estimating harvest intensity (volume removals), but there still remained substantial room for improvement. Overall, we suggest that these algorithms need further training and tuning to be reliably used for accurate monitoring of harvest-related activities in working forests of the US Northeast.

Keywords: change detection, forest disturbance, remote sensing, working forests, landscape monitoring

Forest lands are becoming increasingly valued for providing natural climate solutions, especially the removal and sequestration of atmospheric greenhouse gases that drive global climate change (Malmsheimer et al. 2008, Fargione et al. 2018). As regulatory and market-based offset programs expand to include more forest landowners across the landscape, there has been a corresponding need for high-resolution, large-scale monitoring capabilities to evaluate both historical and prospective land cover dynamics, land use practices, and their implications for carbon stocks. Such information has become essential for achieving more accurate greenhouse gas accounting, from regional to global scales as well as for carbon offset markets and regulatory programs, to ensure compliance at the individual parcel scale. To meet this need, a growing variety of landscape-scale carbon monitoring frameworks, methods, and tools have emerged to provide monitoring and modeling capabilities based largely on remotely-sensed data, given the impracticalities of field data collection across very large (landscape) scales (Kennedy et al. 2010, Zhu & Woodcock 2014, Healey et al. 2018, Housman et al. 2021).

In addition to proprietary software applications, extensions, and tools for desktop geographic information system (GIS) platforms, there are several freely available forest change detection algorithms or corresponding data products, including Landtrendr (LT-GEE; Kennedy et al. 2010), Continuous Change Detection and Classification (CCDC; Zhu and Woodcock 2014) and the Landscape Change Monitoring System (LCMS; Housman et al. 2021). In general, these algorithms process chronologically through a time-series of satellite images on a pixel-by-pixel basis, evaluate stable or cyclic patterns in pixel values, predict future values based on those patterns, and then identify discontinuities or "breaks" where observed values deviate from predictions. For forest change detection, this process commonly involves tracking one or more measures of "greenness" over time, such as the normalized burn ratio (NBR; USGS 2021a) or normalized difference vegetation index (NDVI; USGS 2021b), identifying when a significant or abrupt change has occurred and then estimating the magnitude of that change (based on deviation from expected values). At present, at least a dozen algorithms of this sort exist to provide large-scale retrospective assessment of forest disturbance rates and patterns (Cohen et al. 2010, Banskota et al. 2014, Zhu 2017). Many of these algorithms were developed because of the availability of standardized data products like Landsat Analysis Ready Data (Cohen and Goward 2004, Dwyer et al. 2018) and open computational platforms like Google Earth Engine. These resources have made the algorithms much more accessible to researchers (Zhu 2017) for a range of applications in environmental science, land planning, and resource management (Powell et al. 2010, Schroeder et al. 2011, Zhu and Lui 2014, Hilsop et al 2019, Yin et al. 2020, Zang and Fan 2020, De Marzo et al. 2021).

However, the validation of the outputs of forest change algorithms (i.e., forest disturbance maps indicating the timing and magnitude of changes) using reliable ground-truth information remains a critical and unmet challenge (Cohen et al. 2010). Historical datasets documenting the spatial and temporal patterns of forest disturbance are rare, expensive to compile and curate (Banskota et al. 2014), and typically not available with the same temporal frequency of Landsat data (Kennedy et al. 2010). For these reasons, most satellite-derived disturbance mapping techniques use aerial imagery and manual design-based validation methods, like TimeSync (Cohen et al. 2010, Thomas et al. 2011). Such expert-based procedures are valid and effective but have inherent limitations, including modeling assumptions and human error, relative to ground-truthing with independent and directly measured (or recorded) data sources. For working forests that are privately owned, the ideal ground-truthing data would be the actual harvest records maintained by the landowner or forester but for many reasons, the availability of this kind of information is (understandably) quite limited.

In this study, by working with management records and digital maps provided by forest landowners, we evaluated the ability of three Landsat-derived forest change algorithms (CCDC, LCMS, and LT-GEE) to detect known harvest disturbances at the correct time(s) and location(s) across a 43,300 ha study landscape during a 30-year period (1990–2019). We also analyzed whether the disturbance magnitude estimates provided by two of the algorithms (CCDC and LT-GEE) explained variation in actual harvest intensity (based on pulpwood volume removals). We focused on northern hardwood forests managed primarily with uneven-aged silviculture (e.g., shelterwoods). By comparing relative measures of performance across algorithms and within similar types of harvest prescriptions, we sought to provide independent ground-truthing using landowner data to represent a type of working forest that, although ubiquitous across the US Northeast, has only played a limited role so far in the training and testing of these algorithms.

Methods

Study Area

We evaluated change detection algorithms across 43,300 ha (107,000 ac) of noncontiguous parcels of forest land in the Adirondack Mountains region of northern New York state (Figure 1) for which we acquired harvest records from two different landowners. Comprising 85% of the study area, the Upper Hudson Woodlands (UHW) consists of 49 parcels that are privately owned and managed as a working forest landscape. The UHW lands have been continously managed for forest products by the current (ATP, the Dutch Pension Fund) and previous (Finch, Pruyn & Co.) landowners since the early 20th century. Harvest records for UHW lands extend back to 2010, which coincides with the timing of land sale and transition of management operations to F&W Forestry LLC for ATP.

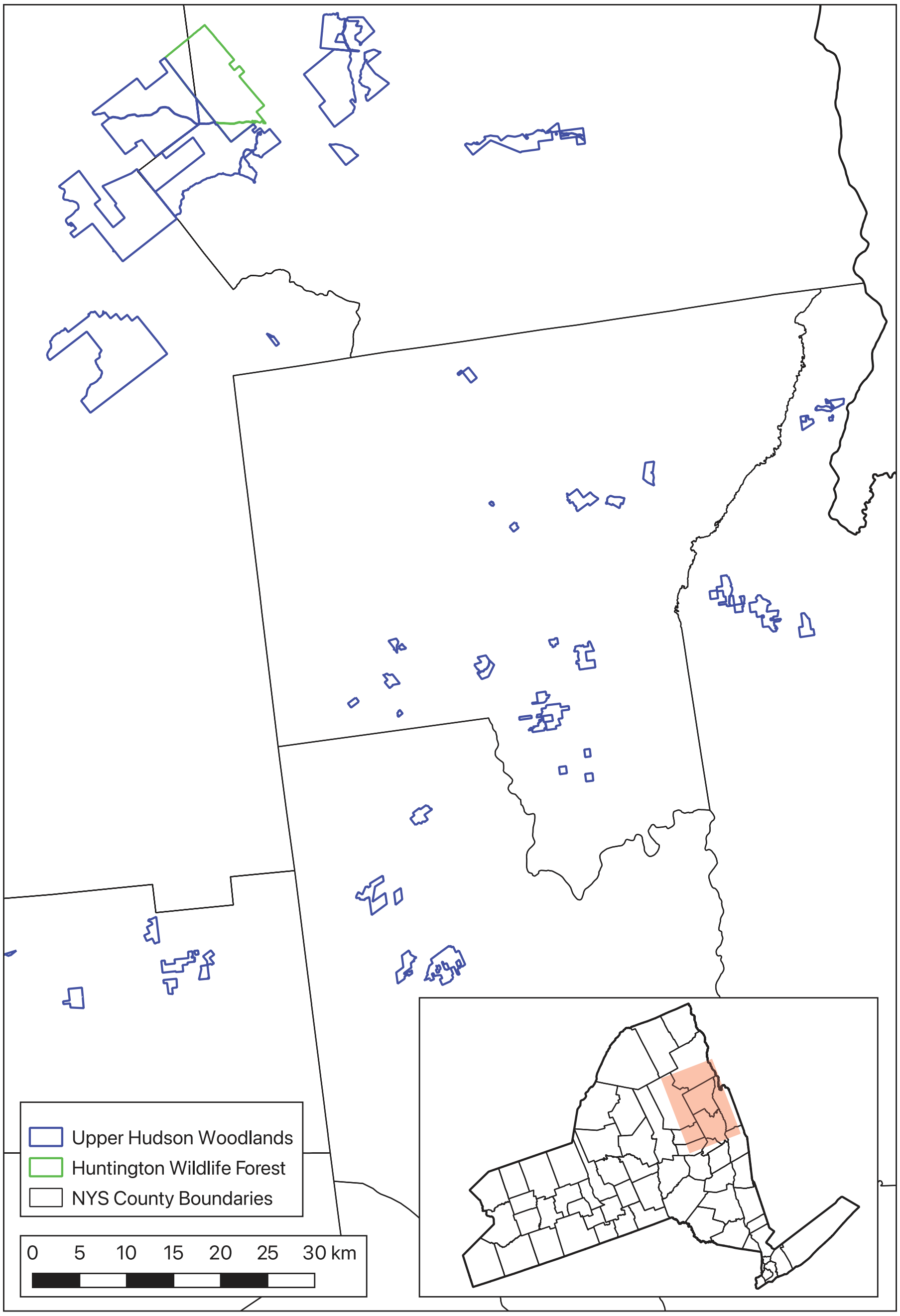

Figure 1. Map of study area including forest parcels in northern New York State (USA). Algorithm performance was evaluated only for areas within parcel boundaries totaling approximately 43,000 ha.

The records included GIS shapefiles of harvest compartments and corresponding tables of the harvest year(s), prescription (harvest type), and volume removals by species and grade. The remaining study area (and single parcel) was Huntington Wildlife Forest (HWF), a 6,000 ha research and education property operated since 1932 by the State University of New York, College of Environmental Science and Forestry, in Newcomb, New York. Forest harvesting at HWF since its acquisition has been conducted for research, teaching, and demonstration purposes. We accessed management records from 1990 to 2019 for HWF that included year, prescription, and volume removal estimates for each harvest operation. Across the two groups of harvests, the average harvest size was 50 ha, with the harvests ranging from 0.25 ha to 226 ha. In total, we evaluated algorithm performance against 229 documented harvest operations, nearly all (94%) of which took place in the last decade (2010–2019). Analysis was conducted on the entire study area; however, to better visualize the algorithm outputs, all map figures (except Figure 1) depict a majority portion of the study area including all of HWF and the northernmost parcels of the UWH.

Preprocessing

We organized all the harvests into five categories: clearcut, shelterwood, thinning, unspecified and other (Table 1). Most harvests were categorized as shelterwood (n = 115) and thinnings (n = 38). The unspecified harvests (n = 55) lacked prescriptions but included harvest year and removal volumes. The other (n = 9) category included miscellaneous prescriptions that were infrequent among the records analyzed. All unspecified harvests were from UHW.

Change Detection Algorithms

The CCDC algorithm is a harmonic regression model that includes elements of intra-annual phenology, gradual interannual change, and abrupt changes (Zhu & Woodcock 2014). Change is assigned when three consecutive deviations from the model are recorded for the same pixel (Zhu & Woodcock 2014). Because the CCDC algorithm uses all available Landsat images, it is able to produce outputs at finer temporal scales than other algorithms—up to every 16 days, the frequency at which Landsat images are produced. To be consistent with the other algorithms, we used just the year of disturbance detected by CCDC.

LT-GEE is a temporal segmentation algorithm that uses both point-to-point and regression-based fitting (Kennedy et al. 2010). The LT-GEE algorithm works on medoid composite values from image stacks from the same time period across multiple years; here, we limited analysis to imagery from July 1 to Aug 31 each year to capture leaf-on season. We evaluated NBR using LT-GEE as it has been found to be most sensitive to forest disturbance events (Kennedy et al. 2010). Using NBR also allowed us to be more consistent with the implementation of the LCMS ensemble.

The LCMS is an ensemble prediction based on outputs of LT-GEE and CCDC that have been implemented using the same Landsat and Sentinel 2 imagery series (Housman et al. 2021). The LCMS outputs including "year of disturbance" are generated by the USDA Forest Service as data products. Disturbance or "loss" magnitude estimates from LCMS have not yet been published for our study area at the time of this study.

To assess algorithm performance, we used two sets of outputs: year of most recent disturbance (Figure 2) and magnitude of most recent disturbance (Figure 3). LT-GEE and CCDC were implemented without calibration to analyze Landsat analysis-ready data in Google Earth Engine. LCMS outputs were acquired from the Forest Service data viewer (https://apps.fs.usda.gov/lcms-viewer/, accessed 18 March 2021). Although "tuning" is recommended for both CCDC and LT-GEE (Landtrendr implemented with Google Earth Engine) applications, we used an "off-the-shelf" application of both algorithms to allow for the most straightforward comparison between them and the LMCS data product. LCMS does not allow end-users to perform tuning, nor has it been specifically tuned for US Northeast forests (N. Pugh, personal communication, November 2020). All algorithm outputs had the same 30 × 30 m grid geometry at an annual time scale and were clipped to the same property boundaries so that areas both known to be harvested and known not be harvested were included in the analysis.

Algorithm Performance Assessment

We evaluated performance of the change detection algorithms by comparing their outputs with detailed forest harvest records and maps provided by landowners using a raster-based GIS overlay analysis. To ground-truth the algorithm outputs against the harvest records, we assigned pixel states based on two binary

Table 1. Summary of forest harvest operations included in this study.

				Upper Hudson Woodlands		Huntington Wildlife Forest	
Category	Harvest types	n	Area (ha)	n	Area (ha)	n	Area (ha)
Clearcut	Clearcuts, permanent clearings	12	119	3	76	9	43
Shelterwood	Shelterwoods	115	6,429	112	6,378	3	51
Thinning	Thinning from above and below	38	1,318	33	1,264	5	54
Other	Salvage, selection systems, OSR, seed tree, strip cuts	9	322	2	149	7	173
Unspecified	N/A	55	3,898	55	3,898	0	0

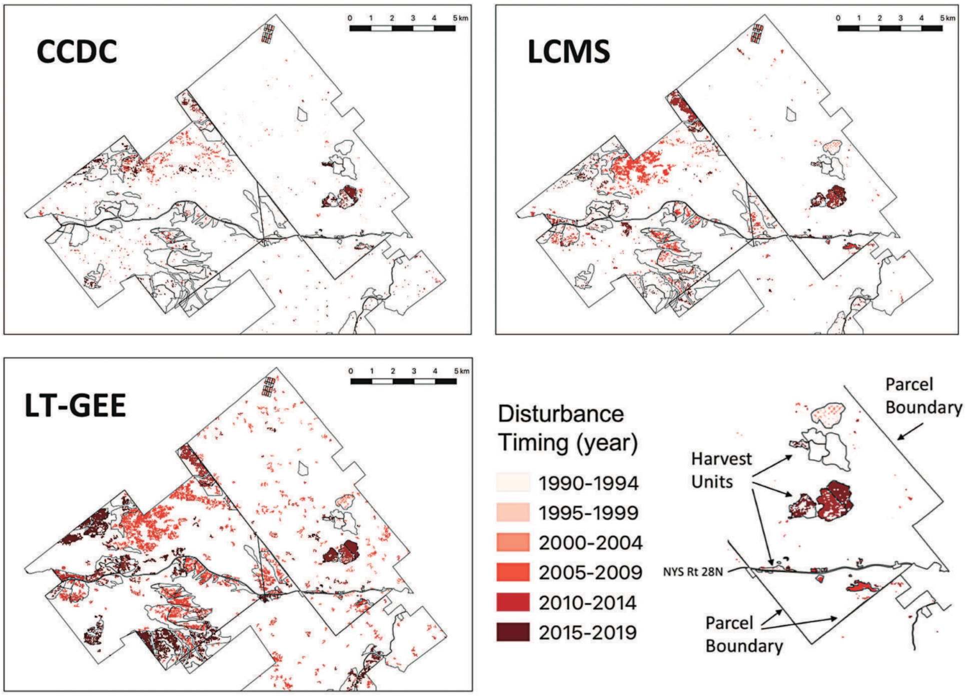

Figure 2. Maps of disturbance timing (year of most recent disturbance) for Continuous Change Detection and Classification (CCDC), Landscape Change Monitoring System (LCMS), and Landtrendr (LT-GEE). Maps show only the northern portion of the study area with the most contiguous tracts of managed forest land for clarity. This analysis was conducted for the entire study area. Inset map next to legend shows examples of parcel boundaries (mostly straight lines of northwest-southeast or northeast-southwest orientation), harvest units (irregular polygons located within property boundaries) and a two-lane highway (NYS Route 28N).

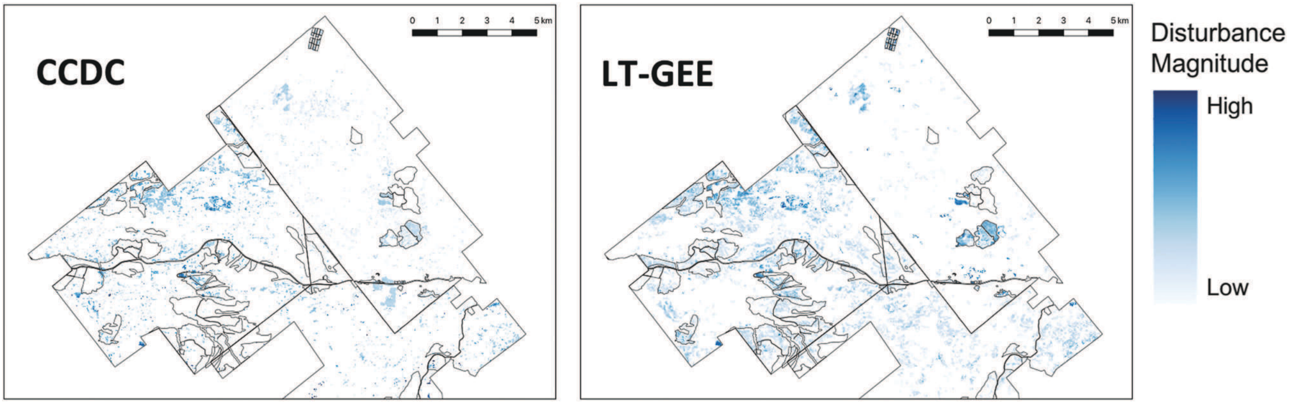

Figure 3. Maps of disturbance magnitude for Continuous Change Detection and Classification (CCDC) and Landtrendr (LT-GEE), showing only a subset of the study area with the most contiguous tracts of managed forest land. Magnitude (loss) outputs were not available for the Landscape Change Monitoring System (LCMS). Note that disturbance magnitudes are unitless indices of deviation from expected spectral values and were not scaled equally between outputs.

conditions, calculated a confusion matrix for each algorithm, and then derived several metrics from these matrices. A detailed description of this process can be found in Figure 4 and is described below.

Raster overlay analysis allowed for pixels to be assigned to one of four groups based on two binary classifications: (1) whether or not a disturbance was detected and (2) whether or not a harvest was recorded at that location. Pixels that met neither condition (no detected disturbance and no recorded harvests) were classified as true negative. Pixels that met both conditions (both a harvest was detected and the year the harvest was detected matched the harvest records) were classified as true positive (TP). Because of the nature of

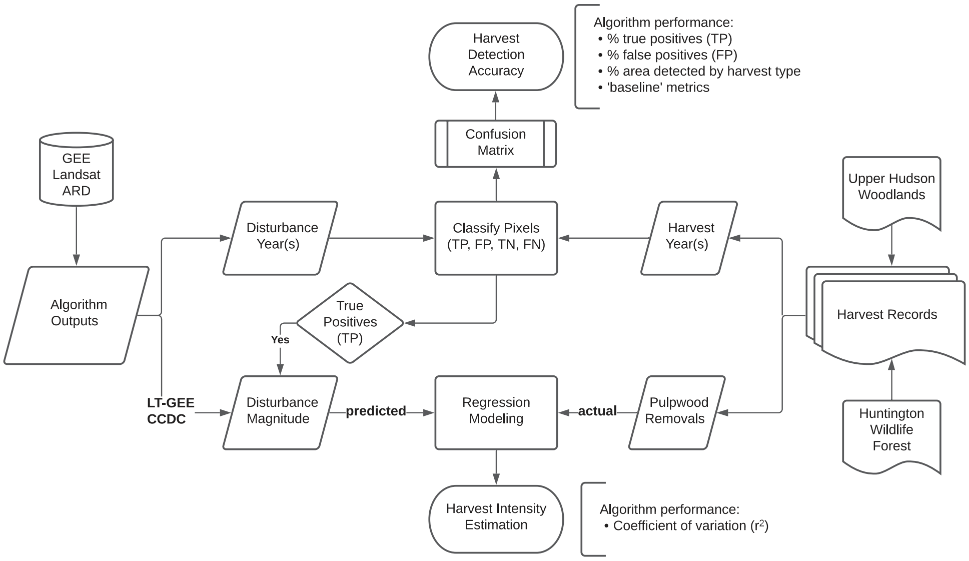

Figure 4. Overview of methods for evaluating performance of forest change detection algorithms (Continuous Change Detection and Classification [CCDC], Landscape Change Monitoring System [LCMS], Landtrendr [LT-GEE]) in detecting known harvest operations.

both Landsat imagery and typical forestry operations in the Northeast US, where much of the harvesting occurs in the winter during leaf off conditions, we recorded any detected disturbances within ± 1 year of the recorded harvest period as a TP. Likewise, for harvests that occurred over multiple years, all harvest years and an additional one-year window on either side of the harvest were considered a satisfactory match.

Pixels that met one but not both conditions fell into one of two categories. False positive (FP) values (commission errors) were those pixels for which a disturbance was detected, but either they were in an area where no harvest occurred, or the detected year of disturbance returned by the algorithm did not match the recorded year(s) of harvest. Finally, false negative (FN) was assigned to pixels where no disturbance was detected for areas where a harvest was recorded (omission errors). Example maps of classified pixels can be found in Figure 5.

We note that areas mapped as FP, although defined as error for our purposes, may represent an accurate detection of disturbance that was simply not represented in harvest operations records. Because these algorithms separate detection from causal attribution, assessing their overall performance would require ground-truthing data on all possible forest disturbances unrelated to harvest operations, which was beyond our current scope. For this reason, our FP results (commission errors) have limited interpretability and have been largely set aside as evidence. For areas mapped as FN, we also note that these algorithms were trained primarily in western North American forests where stand replacing events are the primary disturbance regime, both ecologically (fire) and commercially (clearcut harvesting) (Kennedy et al. 2010; N. Pugh, personal communication, November 2020). Yet most harvest prescriptions used in our study area (and hardwood forests across the broader Northeast US) are less intensive than clearcutting, such as shelterwood, group selection, and thinning. By definition, these practices do not uniformly disrupt or remove the canopy across an entire harvest unit but leave some portion of the canopy intact for a period of time, although some or all of the residual canopy may be harvested years or decades later. Given this context, we expected that most harvest units in our study, which were not clearcuts, would contain nontrivial amounts of FN pixels. Therefore, our assessment of FN pixel results (i.e., where algorithms failed to detect harvest disturbance) was focused on relative performance of the algorithms against the same set of harvest operations.

Confusion matrixes were derived from the pixel classification. The total area represented in each pixel state and the F1 statistic were used to assess the relative accuracy between algorithms. Higher proportions of TP and lower proportions of FN were used as indicators of better algorithm performance. Lower levels of FPs increased the readability of the maps, but because, as previously noted, we could not identify the causal agent of these disturbances, we used this metric in a secondary capacity. The F1 statistic is the harmonic mean of the precision and recall of each algorithm

(Chinchor 1992). We also calculated the total area of detected disturbance across all harvest polygons as a proportion of the study area.

Finally, as a measure of "baseline" or failsafe performance, we calculated the proportion of harvest polygons that did not contain at least one TP pixel for each algorithm. Although in practice, a single pixel would be insufficient to identify a harvest-related disturbance on its own, this allowed us to estimate how often each algorithm failed to detect any disturbance within the boundaries of a known harvest operation.

Once all pixels were classified, the TPs were extracted and zonal statistics were used to determine the number of TPs per polygon. The proportion of TPs for each polygon was averaged by harvest type in the previously described categories. The mean percent harvest area detected was used to compare algorithm performance (e.g., percent harvest area detected) across five different categories of prescriptions: clearcuts, shelterwoods, thinnings, unspecified, and other. ANOVA with a Tukey's HSD was used to test for significant differences in harvest area detected (α = 0.05) between algorithms within each prescription type. We did not compare algorithms across harvest categories due to variable sample sizes and uncertainty from grouping harvests by type/category, instead of working with detailed prescriptions. For instance, because it is possible for a shelterwood to be less intensive than a thinning, we did not expect that formally comparing algorithm performance across such categories would be helpful.

Disturbance Magnitude and Harvest Intensity

The magnitude of disturbance returned by CCDC and LT-GEE algorithms was a unitless value of the Landsat-derived spectral index (e.g., NBR, NDVI), which represented a deviation from the algorithm's expected value for that index for each pixel. We assessed whether algorithm estimates for disturbance magnitude were related to known harvest intensities using only those pixels correctly identified as disturbed (TP) from CCDC and LT-GEE outputs. LCMS disturbance magnitude (or loss) estimates were unavailable at the time of analysis.

Using least-squares regression, we compared LT-GEE and CCDC change magnitude estimates with

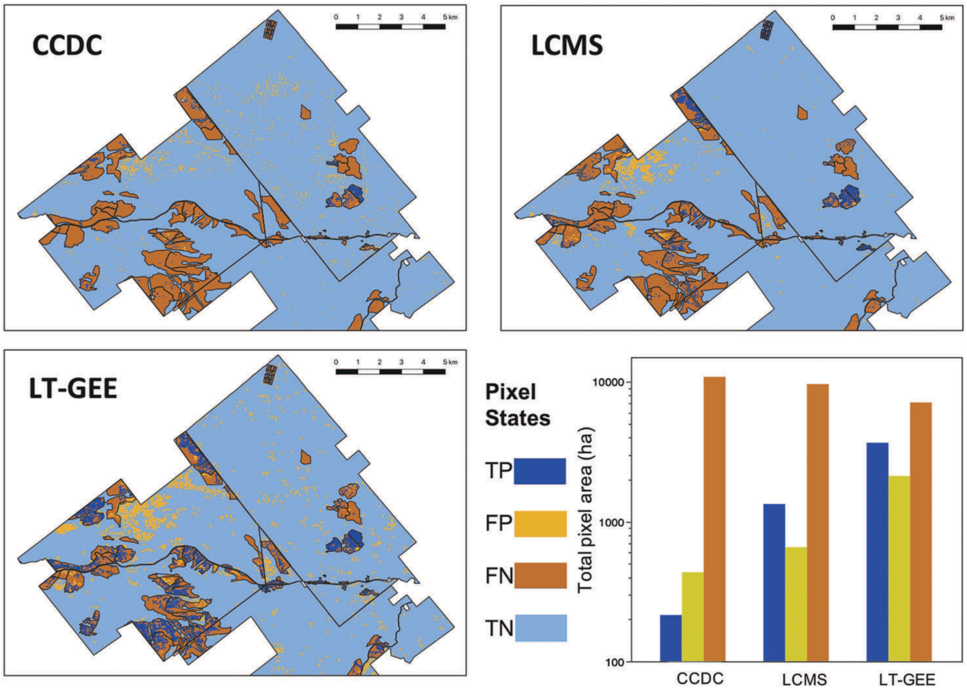

Figure 5. Maps depicting pixel classifications for forest change detection algorithms (Continuous Change Detection and Classification [CCDC], Landscape Change Monitoring System [LCMS], Landtrendr [LT-GEE]) relative to known forest harvest activities. Areas in dark blue represent a timely detection of disturbance within a harvest unit, i.e., true positive (TP) pixels. False positive (FP) pixels are where algorithms detected disturbance outside of known harvest unit boundaries and false negative (FN) are pixels within harvest units where disturbance was not detected at or near the time of harvest. Total areas of TP, FP, and FN pixels tabulated for each algorithm/map are provided for reference (lower right); see Table 2 for same results in percentage format. Maps shows only the northern portion of the study area with the most contiguous tracts of managed forest land for clarity. This analysis was conducted for the entire study area.

recorded pulpwood removal amounts (or depletions) from each UHW harvest unit. The regression model predictor variable was the sum of the pixel-based disturbance magnitudes for each harvest polygon, and the response variable was the total pulpwood removals recorded for the same harvest polygon. Both variables were log-transformed prior to model fitting and the relative performance of CCDC and LT-GEE was assessed based on coefficients of variation and root mean square error (RMSE). All polygons were included in the magnitude analysis, and those polygons that contained no pixels classified as TP had a total magnitude of zero. We focused on pulpwood removals because all of the harvests yielded pulpwood volume, whereas not all yielded sawtimber (overall, pulp removals were about four times sawtimber removals). Yield data from HWF was not included in this analysis due to incomplete records and incompatibility of volume units with UHW records.

Results

Harvest Detection Accuracy

Across 43,300 ha of working forest lands, we found that LT-GEE consistently outperformed LCMS and CCDC in detecting harvest disturbances (Table 2). Focusing on only the areas that were delineated as harvested, LT-GEE detected canopy disturbance for 32.23% of the delineated harvest area and LCMS and CCDC detected 11.69% and 1.87% of the delineated harvested area, respectively. LT-GEE correctly detected disturbance inside harvest units 17 times more often than CCDC (8.70% versus 0.50% of the total study area) and nearly three times more than LCMS (Table 2). Although LT-GEE had the highest rate of TPs, it also had the highest rate of FPs, at 5.46%, compared with 1.52% for LCMS and 1.01% for CCDC. Although more FP pixels can result in "noisy" map outputs, they may actually reflect real disturbances, including natural causes, not represented in harvest records (which we used as the sole basis for ground-truthing). Additionally, LT-GEE had the highest F1 score 0.44, whereas LCMS and CCDC had 0.21 and 0.04, respectively.

Harvest Detection Accuracy by Harvest Type

None of the algorithms consistently detected 50% or greater of recorded harvests, by area, regardless of prescription type. LCMS surpassed the 50% threshold only for clearcut harvests. Overall, LT-GEE performed best and most consistently, as it detected partial harvests better than LCMS and CCDC and had only a slightly lower detection than LCMS for clearcut harvests. LT-GEE was the only algorithm to consistently approach or exceed the 30% harvest area detected benchmark (Figure 6). LT-GEE also performed best with shelterwoods, the most common prescription type, detecting more than twice the area of these harvests compared with LCMS.

Within harvest type groups, there was a significant difference in the performance of all three algorithms for the shelterwood and unspecified category (Figure 6). For thinnings, the performance of LT-GEE was significantly better than LCMS and CCDC, which were not significantly different from each other. For clearcuts, LCMS was significantly better than CCDC, and LT-GEE was intermediate to LCMS and CCDC and not significantly different from either.

Table 2. Harvest detection performance based on a combined confusion matrix (summary of pixel states) that includes results for three algorithms (Continuous Change Detection and Classification [CCDC], Landscape Change Monitoring System [LCMS], and Landtrendr [LT-GEE]) compared against the same harvest records. Values given are percentages of pixels (or area) calculated for each algorithm's outputs separately. Each algorithm was run off the shelf without specific tuning to the study area. True positives indicate pixels where a known harvest occurred and the algorithm detected a disturbance within one year of the harvest period; false negatives indicate where the algorithm failed to detect any timely disturbance within the harvest area; and false positives indicate where disturbance was detected but no known harvest took place.

		Harvest Occurrence	
		Yes	No
Algorithm Detection	Yes	True Positive (TP) CCDC: 0.5% LCMS: 3.1% LT-GEE: 8.7%	False Positive (FP) CCDC: 1.0% LCMS: 1.5% LT-GEE: 5.5%
	No	False Negative (FN) CCDC: 25.2% LCMS: 22.4% LT-GEE: 16.4%	True Negative (TN) CCDC: 73.3% LCMS: 73.0% LT-GEE: 69.4%

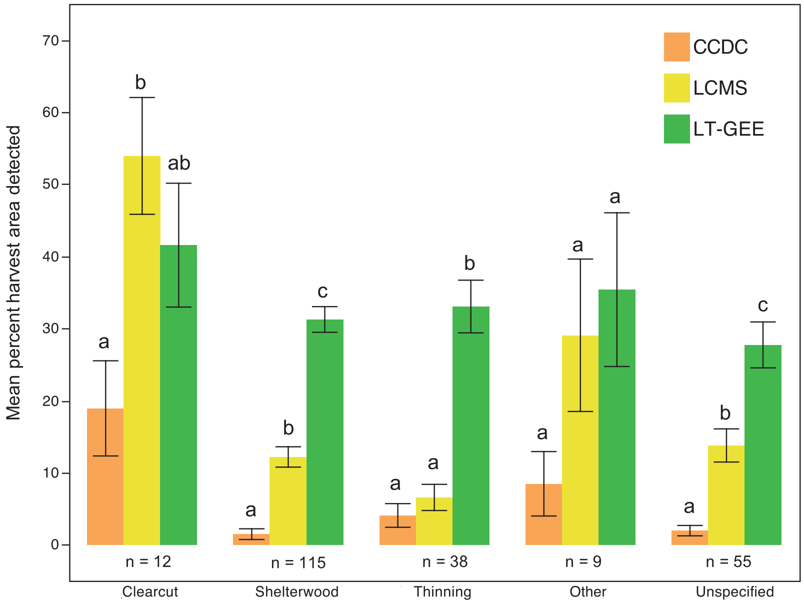

Figure 6. Mean percentage of harvest area detected by harvest type for three Landsat-derived change detection algorithms (Continuous Change Detection and Classification [CCDC], Landscape Change Monitoring System [LCMS], Landtrendr [LT-GEE]). Comparison of means was done between algorithms within each harvest type based on Tukey's HSD. Sample sizes for each harvest type are given below each column cluster, error bars represent ±1 SE, and significant differences among means ($P < 0.05$) are indicated by different lower case letters.

Baseline Detection

CCDC also had the lowest performance of the three algorithms in the baseline metric, which represented a failsafe for harvest detection purposes. CCDC failed to detect any disturbance within polygon boundaries for nearly half (45.8%) of all recorded harvests, compared with less than one-tenth of harvests that were completely undetected by LCMS (7.8%) and LT-GEE (4.8%). When the baseline threshold for detection was raised from one to five pixels (equaling an area just under 0.5 ha), CCDC failed to detect any disturbance for nearly three out of four recorded harvests (74.6%). Raising the detection threshold as above had only nominal effects on baseline detection rates for LCMS and LT-GEE. For monitoring applications, detecting some portion of a harvest disturbance may be preferable to no detection at all, as a higher level of baseline performance can more reliably provide locations to conduct field reconnaissance to assess the disturbance intensity and cause(s). Although the detection of a single pixel is not a reliable indicator of forest harvest or any other disturbance, the relative ability of different algorithms to detect at least a single pixel of change in areas known to have been harvested is an important practical consideration for monitoring purposes.

Magnitude

Based on regression analysis, disturbance magnitudes estimated by LT-GEE (the sum of magnitudes corresponding to all TP pixels for each harvest) were a significant predictor of harvest intensity, using pulpwood depletions as the response variable (Figure 7). As the detected intensity of harvest increased, so did the reported amount of pulpwood harvested. A log-log regression based on LT-GEE explained two-thirds (67.84%) of the variation in the pulpwood removals data, whereas the equivalent CCDC regression explained under one-tenth (9.14%) of the variation in pulpwood depletions. The RMSE for LT-GEE was also much lower than for CCDC, 0.56 compared with 0.80, respectively. CCDC explained much less variation than LT-GEE, in large part due to CCDC's relatively poor baseline detection rate. We identified 105 harvests for which CCDC did not detect a single pixel of change that coincided temporally with recorded harvest operations (see Baseline Detection section). Where no disturbance was detected for a given harvest, the magnitude estimates were defined as null, which resulted in the zero-inflated distribution apparent in the CCDC regression scatterplot (Figure 7).

Discussion

The ability of Landsat-derived change detection algorithms to detect harvests in Adirondack working forests was highly variable, but overall, their performance suggests there is ample room for improvement before they can be used to reliably monitor US Northeast

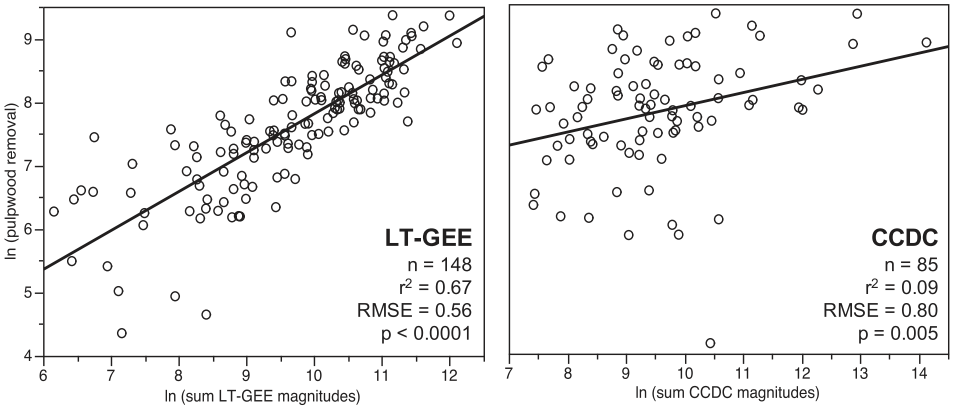

Figure 7. Relationships between disturbance magnitudes (as estimated by Landsat-derived change detection algorithms) and harvest intensity (based on recorded pulpwood removals), based on least squares regression analysis. Pixel magnitudes were summed for each harvest polygon and regressed against pulpwood volumes; all data were log-transformed prior to model fitting. Note the sample size for Continuous Change Detection and Classification (CCDC) (n = 86) is lower than Landtrendr (LT-GEE) (n = 150) due to CCDC's lower baseline detection rates, resulting in more null values that were excluded from analysis.

forest landscapes. All three algorithms performed their best in detecting clearcuts, but two of them (CCDC and LCMS) were much less effective in detecting partial harvest operations such as shelterwoods and thinnings, which are far more common in this region of northern hardwoods (and comprised 94% of the defined prescriptions in our study's harvest records).

By contrast, the LT-GEE algorithm performed equally well across harvest prescriptions and its disturbance magnitude estimates were useful for modeling harvest yields (pulpwood volume removals). These results indicated that LT-GEE has the most off-the-shelf potential for monitoring purposes in the region. However, a tradeoff that came with LT-GEE's superior performance was that of higher sensitivity and output noise. LT-GEE had roughly five times the commission error (FP pixels) as the other algorithms, which resulted in noticeably noisier maps that could make interpretation more difficult in some cases. Yet overall, we found that LT-GEE generated the most reliable and useful outputs, and in a few cases, actually indicated where harvests must have occurred but were missing from our ground-truthing data.

The simultaneous use of multiple algorithms can support the interpretation of, and overall confidence in, results in cases where two or more sets of outputs are in agreement. For any combination of two algorithms, we found agreement on disturbance location about one-fourth (25.61%) of the time. For all three algorithms, we found that agreement decreased by an order of magnitude to roughly 2.5%, consistent with Cohen et al.'s (2017) agreement rate of <5% for any combination of three similar algorithms. Cohen et al. (2017) also found that change detection probability was positively related to the estimated magnitude of the spectral change (i.e., decreased "greenness"), meaning that relatively small spectral changes (<50% reduction in index value) were more likely to go undetected. Because uneven-aged management via partial harvests were by far the most prevalent in our study (94% of defined prescriptions, 70% of all harvests recorded), this lack of sensitivity may explain the relatively low detection rate. Although LT-GEE was able to detect most harvest types at a similar rate and was overall the best performer among the three tested, it only detected on average 30%–45% of each harvest unit (by area).

Study Limitations and Sources of Uncertainty

We also note that the prevalence of partial harvests posed issues for not only the algorithms tested but also for the ground-truthing analysis itself. As discussed, the most common silvicultural prescriptions in northern hardwoods are partial harvests such as shelterwoods and thinnings, which can have residual stocking ranging from 20–80 ft^2 ac^{-1} (4.6–18.4 m^{2}ha^{-1}; Leak et al. 2014). Translating this to a GIS-based ground-truthing analysis, if by definition not all pixels within a harvest unit were disturbed by the harvest operation, then an algorithm that detected only a portion of the harvest area could, in fact, have completely detected the actual harvest operation. However, the available ground-truthing data, which (understandably) did not contain detailed maps of where the canopy was removed or left intact after each harvest operation, precluded any such analysis. Instead, we estimated performance based on the correctly detected area (TP pixels) as a proportion of the total harvest unit area, while recognizing in most cases that the entire unit (polygon) was not actually harvested.

In light of this caveat, our area-based detection results should only be interpreted for algorithm cross-comparison purposes and not as absolute measures of accuracy. For instance, it is possible that LT-GEE's mean detection rate of shelterwood harvests of approximately 30% by area could actually represent a true detection accuracy closer to 50% or 75% or even 90%, depending on the proportion of canopy that was removed in each of these harvests, which was unknown. Moreover, the prevalence of partial harvests and residual canopy coverage required a nuanced interpretation of ground-truthing results, especially FP pixel states (see Algorithm Performance Assessment section). Those pixels classified as FP (commission errors) could in reality represent actual disturbances, such as windthrow or disease-mediated tree mortality, that are not reflected in harvest records and for which historical data is unavailable for ground-truthing. Causal attribution of algorithm-detected disturbances remains a complex challenge that to date has required expert systems to tackle (Cohen et al. 2010); here we side-stepped this issue by assuming that any disturbance detected at the correct place and time was harvest related.

We used the year of most recent disturbance outputs for practical reasons, including the fact that it was the only LCMS output available for cross comparisons. However, this choice could have introduced bias in cases where algorithms detected one or more disturbances prior to the most recent disturbance, and the earlier detections coincided with harvest timing. Such bias would have been unfavorable to algorithm performance estimates; that is, where we erroneously assigned FN instead of TP for an accurate harvest detection. Yet overall, because the vast majority of harvests in this study occurred in the last decade (2010–2019), we suggest that there was a negligible likelihood of a postharvest disturbance unrelated to an earlier harvest operation (e.g., stand reentry for residual canopy tree removal in a shelterwood). To further examine this potential bias, we replaced the CCDC outputs of year of most recent disturbance with year of maximum disturbance as inputs to the confusion matrix and assumed that a harvest would represent the maximum disturbance that occurred within any pixel where multiple disturbances were detected over the study period. We found that multiple (>1) disturbances were detected by CCDC in only 0.03% of the study area and that when we assigned TP to pixels where the year of maximum disturbance suitably matched the harvest timing, the harvest area correctly detected by CCDC increased by 252 ha or 0.5% of the total area of harvest polygons. Overall, we found that pixels with repeat disturbances were rare and had little influence on our results and no meaningful impact on our broader conclusions.

Implications for Further Algorithm Development

For several reasons, our ground-truthing results are best interpreted in a relative context and should not construed as absolute measures of algorithm performance or quality. Our assessment of their off-the-shelf accuracy is not meant to be indicative of their potential performance with more training and tuning, but instead to represent the nature of outputs that a novice user, such as a resource manager or regulatory officer, might generate and attempt to interpret.

In fact, because these algorithms were not developed with the specific goal of detecting harvests in working forests of the eastern US, it was reasonable to expect their performance may be lacking, and that substantial gains in performance could be realized with more focused training and tuning efforts. In this case, CCDC was developed for mixed land cover types of coastal New England and LT-GEE was developed for the dense coniferous forests of the Pacific Northwest (Kennedy et al. 2010, Zhu & Woodcock 2014). Although similar in terms of methodological approach and recent applications in forest change detection, CCDC and LT-GEE were initially developed for the different (but closely related) objectives of land use/land cover change and forest landscape dynamics, respectively. By contrast, LCMS is an ensemble prediction based on LT-GEE and CCDC outputs that used nationwide calibration data, but to date has not been specifically tuned for eastern deciduous forests, including our study area in New York (Housman et al. 2021). Landsat-based change detection algorithms perform best when calibrated for a specific location and set of forest conditions (Cohen et al. 2018).

Looking forward, we recommend the redoubling of efforts to train and tune these algorithms for monitoring US Northeast working forests and offer our enthusiastic and practical support for developers and others seeking to do so. Such investments of time and expertise are needed to ensure that large-scale carbon accounting programs are reliable and that working forest landowners and landscapes are monitored accurately and fairly. We expect that attribution of natural versus harvest-related disturbances will be especially challenging in the eastern US forest landscape, especially at lower harvest intensities, where prescriptions such as group selection are likely to visually mimic natural gap-phase dynamics in unmanaged stands. Here, reliable attribution of the causal factors behind a canopy disturbance could require a more sophisticated analysis that compares patterns of change between working forests and unmanaged reserves. Collaborations between scientists and forest landowners, such as we have developed here, that facilitate sharing of harvest records as ground-truth data will be essential to such efforts.

Conclusion

Change detection algorithms based on satellite imagery are powerful tools for detecting forest change, but their ability to accurately monitor harvest operations and outcomes in working forests of the US Northeast has been little tested. Our ground-truthing analysis based on harvest records from over 43,000 ha of Adirondack (New York) working forests found that off-the-shelf performance of three freely available algorithms was widely variable but generally unsatisfactory especially for detection of partial harvesting practices common in the region. Of the algorithms that we evaluated, LT-GEE performed consistently best in detecting harvesting disturbance across different prescription types and in explaining variability in harvest intensity (based on pulpwood removals). Even so, LT-GEE's average rates of harvest detection were under 50% (by area). The other algorithms performed similarly to LT-GEE for clearcuts but largely failed to detect the much more common types of partial harvests such as shelterwoods and thinnings. Our results suggest that focused training and tuning efforts are needed prior to using these algorithms for monitoring the working forest landscapes of the US Northeast.

Literature Cited

Banskota, A., N. Kayastha, M.J. Falkowski, M.A. Wulder, R.E. Froese, and J.C. White. 2014. Forest monitoring using Landsat time series data: A review. *Can. J. Remote Sens*. 40(5):362–384.

Chinchor, N. 1992. MUC-4 evaluation metrics. *Proceedings of the 4th Conference on Message Understanding*. San Deigo, CA: Association for Computational Linguistics, 22–29.

Cohen, W.B., and S.N. Goward. 2004. Landsat's role in ecological applications of remote sensing. *BioScience* 54(6):535–545.

Cohen, W.B., S.P. Healey, Z. Yang, S.V. Stehman, C.K. Brewer, E.B. Brooks, N. Gorelick, et al. 2017. How similar are forest disturbance maps derived from different Landsat time series algorithms?. *Forests* 8(4):98.

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(12):2911–2924.

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(2):131–140.

De Marzo, T., D. Pflugmacher, M. Baumann, E.F. Lambin, I. Gasparri, and T. Kuemmerle. 2021. Characterizing forest disturbances across the Argentine Dry Chaco based on Landsat time series. *Int. J. Appl. Earth Obs. Geoinf*. 98:102310.

Dwyer, J.L., D.P. Roy, B. Sauer, C.B. Jenkerson, H.K. Zhang, and L. Lymburner. 2018. Analysis ready data: enabling analysis of the Landsat archive. *Remote Sens*. 10(9):1363.

Fargione, J.E., S. Bassett, T. Boucher, S.D. Bridgham, R.T. Conant, S.C. Cook-Patton, P.W. Ellis, et al. 2018. Natural climate solutions for the United States. *Sci. Adv.* 4(11). doi: 10.1126/sciadv.aat1869.

Healey, S.P., W.B. Cohen, Z. Yang, C.K. Brewer, E.B. Brooks, N. Gorelick, A.J. Hernandez, et al. 2018. Mapping forest change using stacked generalization: an ensemble approach. *Remote Sens. Environ*. 204(1):717–728. doi: 10.1016/j.rse.2017.09.029.

Hislop, S., S. Jones, M. Soto-Berelov, A. Skidmore, A. Haywood, and T.H. Nguyen. 2019. A fusion approach to forest disturbance mapping using time series ensemble techniques. *Remote Sens. Environ*. 221:188–197.

Housman, I., L. Campbell, W. Goetz, M. Finco, N. Pugh, and K. Megown. 2021. *U.S. Forest Service landscape change monitoring system methods*. USDA Forest Service. Proj. Rep. GTAC-10225-Brief1, USDA General Technology and Applications Center, Salt Lake City, UT. 27 p. Available online at https://apps.fs.usda.gov/lcms-viewer/tutorials/LCMS_v2020-5_Methods.pdf.

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(12): 2897–2910.

Leak, W.B., M. Yamasaki, and R. Holleran. 2014. *Silvicultural guide for northern hardwoods in the Northeast*. USDA Forest Service, NRS-GTR-132. Northern Research Station, Newtown Square, PA.

Malmsheimer, R.W., P. Heffernan, S. Brink, D. Crandall, F. Deneke, C. Galik, and J. Stewart. 2008. Forest management solutions for mitigating climate change in the United States. *J. For.* 106(3):115–117.

Powell, S.L., W.B. Cohen, S.P. Healey, R.E. Kennedy, G.G. Moisen, K.B. Pierce, and J.L. Ohmann. 2010. 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(5):1053–1068.

Schroeder, T.A., M.A. Wulder, S.P. Healey, and G.G. Moisen. 2011. Mapping wildfire and clearcut harvest disturbances in boreal forests with Landsat time series data. *Remote Sens. Environ.* 115(6):1421–1433.

Thomas, N.E., C. Huang, S.N. Goward, S. Powell, K. Rishmawi, K. Schleeweis, and A. Hinds. 2011. Validation of North American forest disturbance dynamics derived from Landsat time series stacks. *Remote Sens. Environ.* 115(1):19–32.

US Geological Survey. 2021a. Landsat normalized burn ratio. Available online at https://www.usgs.gov/core-science-systems/nli/landsat/landsat-normalized-burn-ratio; Last accessed September 27, 2021.

US Geological Survey. 2021b. NDVI, the foundation for remote sensing phenology. Available online at https://www.usgs.gov/core-science-systems/eros/phenology/science/ndvi-foundation-remote-sensing-phenology?qt-science_center_objects=0#qt-science_center_objects; Last accessed September 27, 2021.

Yin, H., A. Brandão, J. Buchner, D. Helmers, B.G. Iuliano, N.E. Kimambo, K.E. Lewińska, et al. 2020. Monitoring cropland abandonment with Landsat time series. *Remote Sens. Environ.* 246(9):111873.

Zhang, W., and Fan, H. 2020. Application of isolated forest algorithm in deep learning change detection of high resolution remote sensing image. *2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, 753–756. Dalian, China: Institute of Electrical and Electronics Engineers.

Zhu, X., and D. Liu. 2014. Accurate mapping of forest types using dense seasonal Landsat time-series. *ISPRS J. Photogramm. Remote Sens.* 96(10):1–11.

Zhu, Z. 2017. Change detection using Landsat time series: a review of frequencies, preprocessing, algorithms, and applications. *ISPRS J. Photogramm. Remote Sens.* 130(8):370–384.

Zhu, Z. and C.E. Woodcock. 2014. Continuous change detection and classification of land cover using all available Landsat data. *Remote Sens. Environ.* 144(3):152–171.