Remote Sensing of Environment 204 (2018) 717–728

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

Mapping forest change using stacked generalization: An ensemble approach

Sean P. Healey[a,*], Warren B. Cohen[b], Zhiqiang Yang[c], C. Kenneth Brewer[d], Evan B. Brooks[e], Noel Gorelick[f], Alexander J. Hernandez[g], Chengquan Huang[h], M. Joseph Hughes[c], Robert E. Kennedy[c], Thomas R. Loveland[i], Gretchen G. Moisen[a], Todd A. Schroeder[j], Stephen V. Stehman[k], James E. Vogelmann[i], Curtis E. Woodcock[l], Limin Yang[n,o], Zhe Zhu[m]

[a] US Forest Service Rocky Mountain Research Station, United States
[b] US Forest Service Pacific Northwest Research Station, United States
[c] Oregon State University, United States
[d] US Forest Service Washington Office, United States
[e] Virginia Polytechnic Institute and State University, United States
[f] Google Corporation, United States
[g] Utah State University, United States
[h] University of Maryland, United States
[i] USGS Earth Resources Observations and Science Center, United States
[j] US Forest Service Southern Research Station, United States
[k] State University of New York College of Environmental Science and Forestry, United States
[l] Boston University, United States
[m] Texas Tech University, United States
[n] Stinger Ghaffarian Technologies (SGT, Inc.), United States
[o] U.S. Geological Survey Earth Resources Observation and Science Center, United States

ARTICLE INFO

Keywords:
Multiple classifier systems
Landsat
Change detection
Ensemble

ABSTRACT

The ever-increasing volume and accessibility of remote sensing data has spawned many alternative approaches for mapping important environmental features and processes. For example, there are several viable but highly varied strategies for using time series of Landsat imagery to detect changes in forest cover. Performance among algorithms varies across complex natural systems, and it is reasonable to ask if aggregating the strengths of an ensemble of classifiers might result in increased overall accuracy. Relatively simple rules have been used in the past to aggregate classifications among remotely sensed maps (e.g. using majority predictions), and in other fields, empirical models have been used to create situationally specific algorithm weights. The latter process, called "stacked generalization" (or "stacking"), typically uses a parametric model for the fusion of algorithm outputs. We tested the performance of several leading forest disturbance detection algorithms against ensembles of the outputs of those same algorithms based upon stacking using both parametric and Random Forests-based fusion rules. Stacking using a Random Forests model cut omission and commission error rates in half in many cases in relation to individual change detection algorithms, and cut error rates by one quarter compared to more conventional parametric stacking. Stacking also offers two auxiliary benefits: alignment of outputs to the precise definitions built into a particular set of empirical calibration data; and, outputs which may be adjusted such that map class totals match independent estimates of change in each year. In general, ensemble predictions improve when new inputs are added that are both informative and uncorrelated with existing ensemble components. As increased use of cloud-based computing makes ensemble mapping methods more accessible, the most useful new algorithms may be those that specialize in providing spectral, temporal, or thematic information not already available through members of existing ensembles.

[*] Corresponding author.
E-mail address: seanhealey@fs.fed.us (S.P. Healey).

http://dx.doi.org/10.1016/j.rse.2017.09.029
Received 9 March 2017; Received in revised form 16 September 2017; Accepted 22 September 2017
Available online 29 September 2017
0034-4257/ Published by Elsevier Inc.

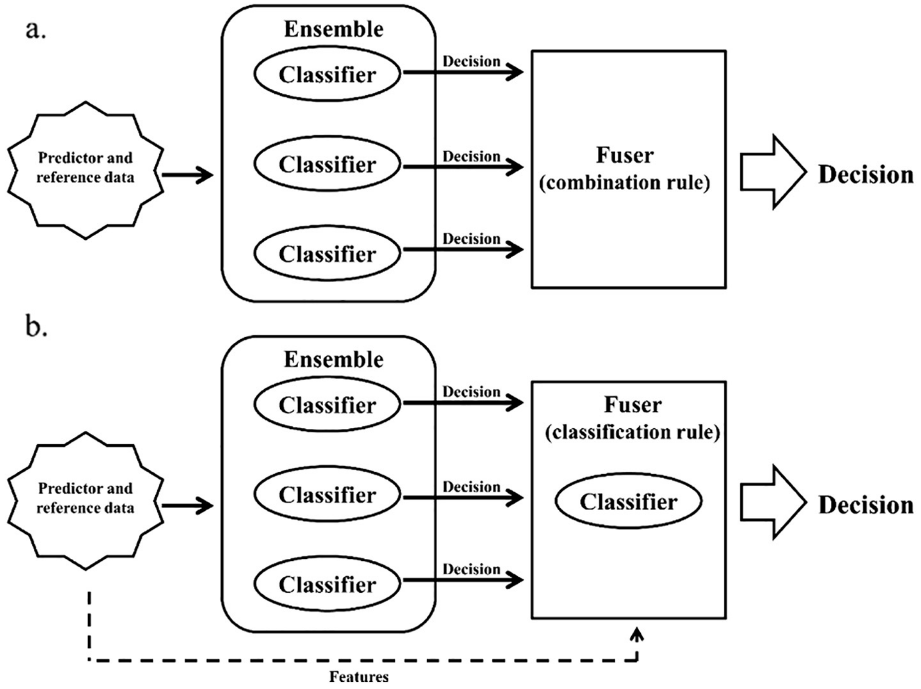

Fig. 1. Topologies of commonly used MCS. Fuser functions may either use a combination rule, such as voting or averaging, which requires only classifier outputs (panel a), or they may call upon features from a learning dataset to facilitate weighting of outputs on the basis of model performance (panel b). This second approach has been termed "stacking."
Figure adapted from Woźniak et al. (2014).

1. Introduction

1.1. The challenge of mapping subtle forest cover loss

Land cover change due to both human and natural disturbance processes has a profound effect on how ecosystems function, affecting biogeochemical (Chambers et al., 2007; Kurz et al., 2009) and hydrological cycles (Seilheimer et al., 2013), habitat conditions (Spies et al., 2010), and availability of social and economic human benefits (González-Olabarria and Pukkala, 2011). Characterization of land cover change has therefore emerged as a discipline with a central bearing on many fields of study (Turner et al., 2007). The Landsat platform has been a primary source of land change information, capable of detecting important vegetative and disturbance patterns because of the sensor's long history and appropriate temporal, spatial, and spectral properties (Cohen and Goward, 2004). The Sentinel and SPOT platforms have also proven useful for this task (e.g., (Antropov et al., 2016; Li et al., 2016; Verhegghen et al., 2016)). The free release of all images in the Landsat archive (in 2008) has led to the development of many new algorithms capable of using temporally dense observations to increase the breadth, accuracy, and precision of land cover change characteristics that can be mapped (Wulder et al., 2012).

However, like most remote sensing problems, there are many factors that can increase the complexity of detecting forest change, particularly beyond the relatively straightforward stand-replacing disturbances targeted in earlier efforts (e.g. Healey et al., 2008). Cohen et al.'s (2016) national survey of forest disturbance processes found that low-magnitude forest decline was the most common cause of disturbance, particularly in the Western US. Likewise, US Forest Service inventory data indicates that partial harvests are more commonly practiced than clearcuts across the country (Smith et al., 2009), and the inter-agency Monitoring Trends in Burn Severity project (Schwind et al., 2010) found that only 36% of the area burned by 13,400 large fires in the US between 1985 and 2010 had moderate or greater severity (Finco et al., 2012). For any given low-magnitude disturbance, subtle removals of forest canopy may increase spectral reflectance in both the visible and mid-infrared wavelengths if removal of vegetation reveals brighter soils, but reflectance may actually decrease if canopy removal increases the contribution of shadowing to the spectral signal or if charring occurs (Schroeder et al., 2011). Consistency of spectral response across space and time may also be compromised by phenology, atmosphere, topography, soil type, forest type, and forest structure.

There are several change detection algorithms which target lower-magnitude change (e.g. (DeVries et al., 2015; Healey et al., 2006; Meigs et al., 2015)) in very specific scenarios, but it is an open question if Landsat or other remote sensing platforms can be used across complex landscapes to detect the full range of disturbance magnitudes and types without also introducing detrimental levels of false-positive (i.e., commission) error. It should be noted that while the term "change detection" is used here for the process of mapping forest disturbance, that process is very much subject to error and actually represents a prediction of change more than a definitive discovery. The more accurate "change prediction" is not used here both because of convention and to distinguish the current monitoring task from work involved with assessment of future events (e.g. (Seidl et al., 2014).

1.2. Multiple classifier systems

This paper presents a test of the idea that an ensemble of change detection algorithms can be used together to obtain forest disturbance maps of greater accuracy and sensitivity than maps from any single automated algorithm. Wolpert and Macready (1997) demonstrated, in their "No Free Lunch" theorems, that if an algorithm performs well in one class of problems, it necessarily "pays" for that accuracy with degraded performance on a set of all remaining problems. If different algorithms have different specialties, particularly if those specialties are diverse, combination of those algorithms in Multiple Classifier Systems (MCS) should improve global performance (Oza and Tumer, 2008). We use the term "classifier" to refer to any generalizing algorithm or model that produces a hypothesis about an object using a set of learning data. A variety of tools have been used as classifiers across disciplines, from logistic regression to nearest neighbor imputation and support vector machine methods (e.g. (Sáez et al., 2013); (Kavzoglu et al., 2014)), and this paper focuses on a variety of algorithms that make use of time series analysis with Landsat imagery.

Analytical approaches based on MCS now play an important role in tasks ranging from detecting computer security risks to diagnosing disease (Woźniak et al., 2014). This paper focuses on a class of MCS

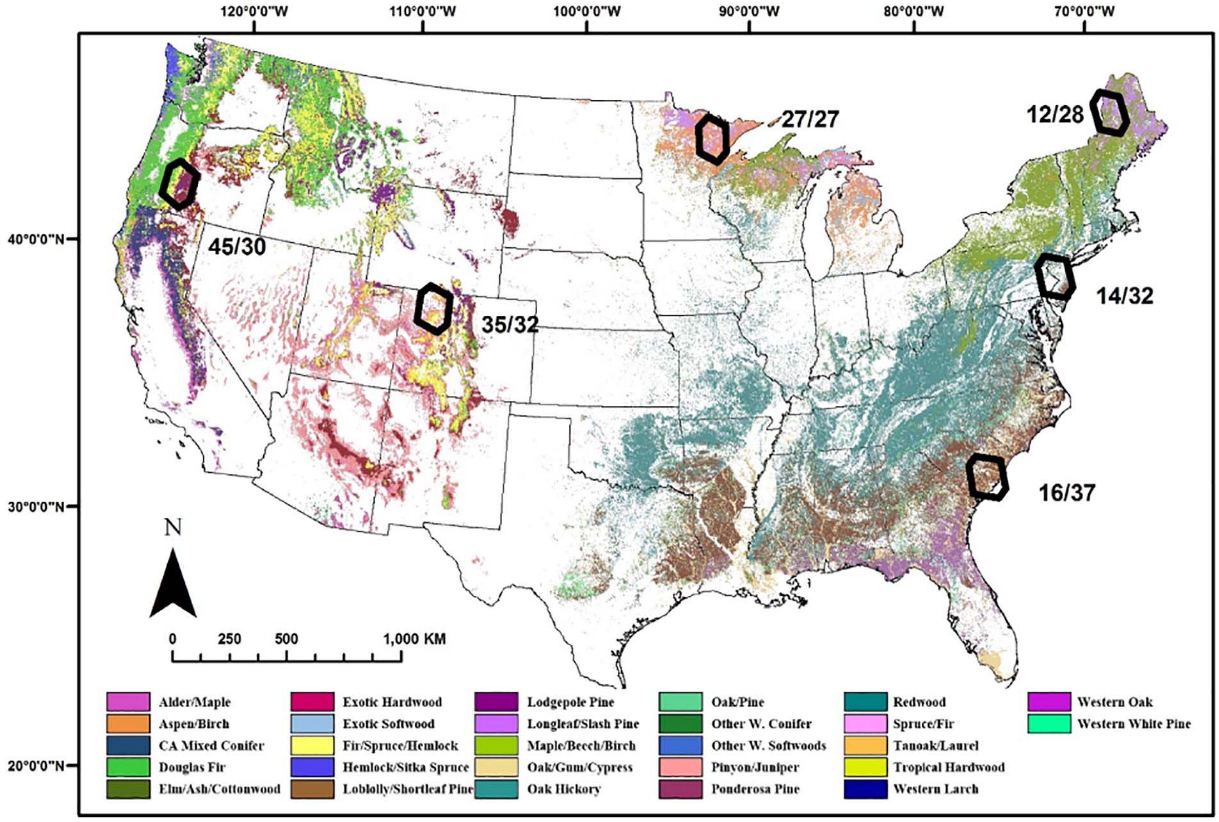

Fig. 2. Study areas across the United States, labeled by World Reference System Path/Row. Forest type groups (Ruefenacht et al., 2008) are also shown, with mapped colors described in the legend.

which applies an ensemble of classifiers to a problem simultaneously and then uses a fusion rule to employ a meta-classification process.

Fig. 1 illustrates two types of fusion rules: one which uses a simple combination rubric such as an average or majority (a), and one which uses a secondary model to re-weight the classifiers according to their performance against similar cases in the reference data (b). Random Forests (RF; (Breiman, 2001)) is a prominent example of an MCS which uses a combination rule. RF creates an ensemble of similar classifiers by training decision tree-based models with random partitions of the input data, a process called "bagging" (Breiman, 1996a). An RF prediction is typically a simple function of how the trees vote (e.g. either the average or majority of predictions). RF is widely used with remotely sensed data to map ecological variables (e.g., (Cutler et al., 2007; Li et al., 2017; Powell et al., 2010; Prasad et al., 2006)), including land cover change (Ahmed et al., 2017; Grinand et al., 2013).

Simple combination-rule fusing processes may also be used with heterogeneous ensembles – those that involve fundamentally different classifiers, in contrast to the trees used by RF which differ only by the particular subset of input data used for calibration. Engler et al. (2013) predicted tree species using remotely sensed data and eight separate probabilistic models. An ensemble prediction was derived by taking a weighted average of the eight models' predicted probability of membership for each species class. Kandel et al. (2015) took a similar averaging approach to combine and map relative occurrence indices for wildlife. Foody et al. (2007) took the majority vote of five binary classifiers to map the occurrence of fenland.

1.3. Ensemble-rule stacking

Wolpert (1992) is credited with the suggestion of a secondary model (as in Fig. 1.b) as an alternative to simple combination rules. This approach, termed "stacked generalization," or "stacking," works by predicting the original classifiers' areas of poor performance with respect to independent or bootstrapped reference data. Detected systematic errors are used in the second-level classifier to improve prediction accuracy. In other words, whereas the outputs of a typical RF model and the ensembles mentioned above emphasize model agreement (using voting or mean predictions), stacking can differentially weight dissenting models within an ensemble based on performance against empirical reference data.

While most stacking applications to date have used a form of regression as the secondary classifier (e.g. (Breiman, 1996b); (Xing et al., 2016)), we hypothesize that using an ensemble classifier such as RF to fuse individual maps will improve performance not only over the original classifiers but also over regression-mediated stacking. Schroeder et al. (2017) found an RF-fused ensemble to be effective in the categorization of the cause of mapped disturbances. In the case of forest change detection, using RF with a stacking approach requires assembly of an ensemble of independent mapping algorithms, the outputs of which can be used as the inputs to a secondary RF model calibrated with independent reference data. Unlike the parametric methods proposed by Wolpert (1992), use of an ensemble method for stacking would not explicitly identify classifier weights in the form of regression parameters. However, an ensemble-rule stacking process may address overfitting of the training data, which is known to be problematic when individual classifiers are used for stacking (Breiman, 1996b; Reid and Grudic, 2009).

We report a test of this approach in the context of forest disturbance detection using both Random Forests and logistic regression models to build the secondary model that accomplishes the stacking function. This tested approach represented an expansion of what is typically termed "stacking," both because of the use of Random Forests in the secondary model, and because some of the ensembles tested contained inputs other than the results of independent learners (some contained unclassified imagery or topographic inputs, for example). Specifically, this test was conducted at six diverse sites across the United States, with error rates of individual learners compared to those of both regression- and ensemble-rule stacking processes. An ensemble of leading change detection algorithms was used in this test of the stacking process, with the intent of representing the variety of available approaches instead of cataloging the strengths of each approach. Results of this study are relevant both for forest change detection and other complex classification problems where compelling alternative learning algorithms exist.

2. Methods

2.1. Study areas and reference data

The study area was the non-overlapping area of six World Reference System (WRS-2) Landsat frames (labeled by path/row): 45/30; 35/32; 27/27; 16/37; 14/32; and 12/28 (Fig. 2). These scenes encompass a broad range of forest types, and the covered forests are affected by a variety of disturbance agents across a range of magnitudes. The scenes cover areas known to have experienced industrial logging (particularly scenes 12/28 and 16/37), insect epidemics (35/32), fires (45/30) and wind events (27/27 and 16/37).

All Landsat Thematic Mapper (TM) and Enhanced Thematic Mapper (ETM +) imagery for the period of 1984–2012 that was available in the United States Geological Survey Earth Resources and Science Center (USGS EROS) archive (as of April, 2013) was retrieved and uniformly pre-processed. Pre-processing included conversion to surface reflectance (Masek et al., 2006), automated generation of cloud and cloud shadow masks (Zhu and Woodcock, 2012), and terrain correction provided with Landsat Level-1T products (Loveland and Dwyer, 2012). This standardized set of all available imagery was provided to developers of each of the algorithms described in the next section.

A reference dataset characterizing historical forest disturbance across the study area was acquired using a tool called TimeSync (Cohen et al., 2010). This tool facilitates visual interpretation of both historical Landsat imagery and historical finer-resolution aerial images served by Google Earth. Among the data recorded were the start and end dates of forest disturbances occurring at each reference point from 1986 to 2011, as well as the type of disturbance (including “harvest,” “fire,” “forest decline,” and “wind”). Areas were identified as forests in the reference data development process if fine-resolution imagery indicated the presence of, or potential for developing, at least 10% tree cover, as well as the absence of other land uses. Any reduction of forest canopy that was perceptible through inspection of either the Landsat or fine-resolution time series was registered as a disturbance. All observations were based upon visual inspection of both all available Landsat imagery and fine-resolution aerial imagery. Two interpreters were used for each sample point, and a third was used as an arbiter if there was a disagreement. While definitions of what constitutes a disturbance vary among projects, as discussed earlier, this arbitration process for every TimeSync observation was assumed to reduce measurement error with respect to the definition used in this project to negligible levels.

Reference points were chosen using a simple random sample; 300 30-meter pixels per scene (1800 total). Of these locations, 1303 were determined through TimeSync analysis to be in forested conditions. At each forested location, observed disturbances were recorded across 27 years (1985–2011), with long-duration disturbances such as insect outbreaks counting as disturbances in every year they were observed to be active. This sample was acquired for three purposes. First, it formed the basis for statistical estimation of landscape-wide disturbance rates over time (Olofsson et al., 2014). Second, the sample was used to evaluate the performance of disturbance maps at the pixel level, as described below. Lastly, the reference points provided a basis for training and evaluating empirical models of disturbance through stacking, also described below.

Since the TimeSync reference data collection process required extraction and staging of all available Landsat data, as well as compilation of fine-resolution aerial imagery, there was incentive to gather as much information as possible at each location. Consequently, TimeSync observations at each 1-year interval for every selected location were used to evaluate model and algorithm performance in predicting historical disturbances. Some degree of autocorrelation in classifier performance may be expected across time at the same location. Forest conditions generally remain relatively stable, and performance in one date is likely to be similar to performance in the next, particularly when no disturbance occurs. In this study, there were 26 1-year change/no change observations at each of the 1303 forested locations. This collection of classifier performance observations can be considered as a temporal cluster sample across randomly allocated locations (i.e., each location is a cluster and the observations for each year are considered secondary sampling units). Under such a design, one would take the mean of cluster (location) means as a measure of overall performance. Because an equal number of observations was made at the location of each temporal cluster (one observation for each 1-year period), each observation (year within location) was treated as having resulted from an equal probability of selection, allowing for an unweighted analysis to combine observations over locations. Comparisons here of model and algorithm performance were based upon both a false-positive detection error rate (“errors of commission”) and a false-negative detection error rate (“errors of omission”). The false positive rate was predicted as the number of mapped disturbances not corresponding to a detected disturbance in the reference data divided by the number of all mapped disturbances. The false-negative error rate was the number of reference disturbances missed in the map divided by the total number of disturbances observed in the reference dataset.

2.2. Base learners

Eight automated change detection algorithms, described briefly here and more thoroughly in the cited references, were applied to the standardized Landsat dataset described in the previous section. Each algorithm used a subset of imagery consistent with published requirements and methods (cited below). Specifically, Continuous Change Detection and Classification (CCDC) and Exponentially Weighted Moving Average Change Detection (EWMACD) used all available L1T imagery, and the others targeted composites of growing season imagery. Individual algorithms were termed “base learners” (or “BLs) here because they produce the basic classifications used to build various forms of MCS. The reader is cautioned that the term “BL” does not imply a machine learning component in any of the algorithms. BL algorithms included:

1. CCDC – Continuous Change Detection and Classification (Zhu and Woodcock, 2014) – Time series functions of all clear (cloud- and shadow-free) pixels are fit using ordinary least squares. These models are intended to capture seasonality, trend and breaks. Predictions are made in a forward fashion from preceding acquisitions, and when these predictions differ significantly from actual reflectance values in three consecutive acquisitions, a land cover change is inferred. In addition to producing forest change maps, CCDC can predict cloud-free “synthetic” images (Zhu et al., 2015) for any date. Annual synthetic outputs, instead of surface reflectance, were used as inputs for MIICA and ITRA (described below). VCT and LandTrendr (also below) were each run with CCDC synthetic imagery and, independently, with surface reflectance imagery.
2. VCT – Vegetation Change Tracker (Huang et al., 2010; Thomas et al., 2011) – Cloud-free annual composites are converted to a multi-band z-score metric of similarity to local undisturbed forest conditions. Abrupt disturbances are predicted when this metric shifts (and stays) away from forested conditions.
3. LandTrendr – (Kennedy et al., 2010; Kennedy et al., 2012) – Cloud-free annual composites are used to create single-band time series for

each pixel, and these time series are then segmented into discrete periods of growth, disturbance, or recovery. Changes may be abrupt or gradual.

4. EWMACD – Exponentially Weighted Moving Average Change Detection (Brooks et al., 2014) – EWMACD analyzes residuals between observed pixel values and predictions generated by fitting a harmonic regression model (Brooks et al., 2012) to a fixed set of designated training imagery. A user-tuned collection of statistical process control tools, specifically tuned in this study to detect low-magnitude (e.g., thinning) and/or long-term changes, are then applied to the residual time series to identify changed pixels at each point in the time series.
5. MIICA – Multi-index Integrated Change Analysis (Jin et al., 2013) – Thresholds related to spectral change magnitude and direction are developed empirically across bi-temporal (one-year interval) periods. These thresholds use four different indices (differenced normalized burn ratio, differenced normalized difference vegetation index, change vector, and relative change vector maximum), and identify biomass increase, decrease, or no change.
6. VerDET – Vegetation Regeneration and Disturbance Estimates through Time (Hughes et al., 2017) – Annual cloud-free composites are segmented using a piecewise linear function. For each pixel, the slopes of each temporal segment are interpreted as "disturbed," "stable," or "regenerating," allowing identification of fast or slow disturbances of different magnitudes.
7. ITRA – Image Trends from Regression Analysis (Vogelmann et al., 2012) – annual cloud-free composites are sub-divided into three periods, and a linear regression model is fit to each period. The emphasis is on capturing longer-term trends in forests, shrublands, and other ecosystems.
8. Shapes-NBR – (Meyer, 2013; Moisen et al., 2016) – for each pixel, a semi-parametric additive regression algorithm provides a smoothed trajectory constrained to behave in an ecologically sensible manner, assuming one of six possible "shapes", or patterns through time. From those fitted trajectories, parameters are generated that summarize the temporal pattern, including: year(s) of inflection; magnitude of change; and pre- and post-inflection rates of growth or recovery. Originally intended only as a predictor of forest change (Schroeder et al., 2017) instead of as a stand-alone algorithm, the shapes algorithm was applied to a composite series of NBR images, and disturbances were inferred through basic rules applied to inflection points.

Each of these algorithms was adapted to produce year-by-year binary ("disturbed" or "not disturbed") maps of forest disturbance, which were compared to similarly simplified TimeSync observations. Each year of mapped longer-term disturbance events was labeled "disturbed." For every year a pixel was classified as "disturbed" by a particular algorithm, the variables listed in Table 1 were also saved as potential inputs to ensemble models. Because time series algorithms frequently exhibit errors in the first and last time periods, only change maps from 1985 to 2011 (matching the 27 years of TimeSync observations) were used in this process, ignoring outputs for 1984–85 and 2011–12. The year 2012 was the endpoint of the time series because that was the most recent complete year of Landsat acquisition at the initiation of the project. Also, a composite map (called the "Union" map) was created in which a pixel was labeled as "disturbed" in a given 1-year period if any of the eight algorithms labeled it so.

2.3. Stacking of base learners and other spatial predictors

The stacking process was pursued here one year at a time, with date-neutral models predicting binary "disturbed/non-disturbed" status based upon varying ensemble inputs. Both logistic regression and RF models were tested and trained with the TimeSync reference dataset described above. Inputs for the logistic model were determined through a stepwise model selection process based upon Akaike's Information Criterion, operating upon a list of potential inputs that included the BL outputs listed in Table 1, plus a forest type group map (Ruefenacht et al., 2008; see Fig. 2), the most basic topographic variables (slope, elevation, and aspect), and Landsat imagery, composited as described in Section 2.4, for both the beginning and ending year of the year-to-year period. Logistic regression modeling was carried out using the glm function of the R statistical programming platform (R Core Team, 2015).

Stacking with a RF fusion rule was also carried out in R, using the randomForest package (Liaw and Wiener, 2002). Five different sets of inputs were tested to better understand contributions of particular input classes: **1) Landsat only** – a single composite Landsat image (created using methods described in Section 2.4) from both the "before" and "after" years of the 2-date period; 2) BLs alone – the binary change maps from each change detection algorithm and ancillary outputs listed in Table 1 for all pixels labeled as "disturbed"; 3) BLs + Landsat – combination of set #s 1 and 2; 4) BLs + Topography + Forest Type – combination of #2 and the pixel's slope, elevation, and cos (aspect), plus, the value from the Ruefenacht (2008) forest type group map (see

Table 1
Base learner output variables.

Base learner	Output	Variable type
CCDC	Magnitude of spectral change	Continuous
EWMACD	Magnitude of spectral change	Continuous
ITRA	Magnitude of spectral change	Categorical (high, medium, low)
LandTrendr-surface Reflectance	Magnitude of spectral change	Continuous
	Duration of declining segment	Continuous
LandTrendr-synthetic	Magnitude of spectral change	Continuous
	Duration of declining segment	Continuous
MIICA	Magnitude of spectral change	Categorical (biomass increase, decrease, no change)
VCT-surface reflectance	NBR change magnitude	Continuous
	NDVI change magnitude	Continuous
	"udist" magnitude	Continuous
VCT-synthetic	NBR change magnitude	Continuous
	NDVI change magnitude	Continuous
	"udist" magnitude	Continuous
Shapes	Shape Type	Categorical (flat, decreasing, increasing, jump, growth-to-decline, decline-to-growth)
	Absolute spectral change magnitude	Continuous
	Relative spectral change magnitude	Continuous
	Duration of declining segment	Continuous
	Prior rate of change	Continuous
	Posterior rate of change	Continuous

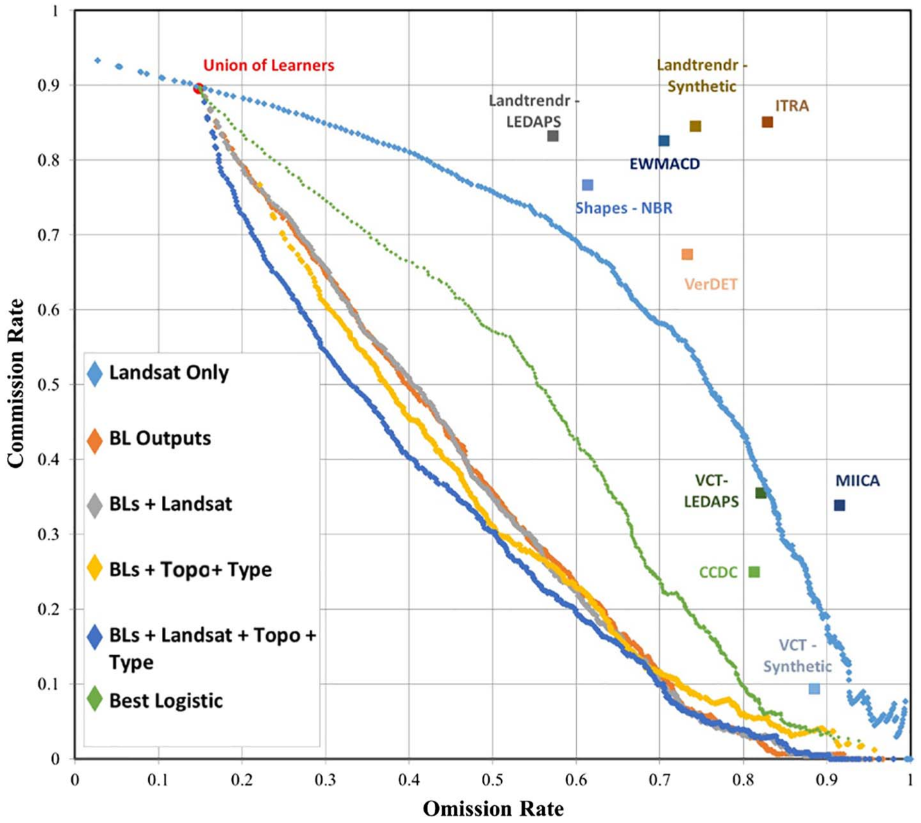

Fig. 3. Performance of each of the tested BLs, RF model combinations and the best logistic model in terms of lowest resulting omission and commission error rates against the complete TimeSync reference dataset. Model outputs are represented as a series of points because both RF and logistic models allow flexible class inclusion rules. Points closest to the origin (0,0) have the lowest rates of omission and commission error.

Fig. 2); 5) BLs + Landsat + Topography + Forest Type – combination of the inputs for model #s 1 and 4. Initial testing showed that results were stable at and above 500 decision trees per model, so each of these combinations was run with 500 trees.

Two measures were taken to enhance the modeling process for stacking using both RF and logistic regression. First, the model-building dataset was restricted to only those 1-year observations where a disturbance was either recorded in TimeSync or predicted by at least one BL (or both). Initial testing suggested that skewing the model-building dataset in this way toward potentially confusing cases improved model performance. Effects of this model-building step were not formally evaluated, however. Secondly, in recognition of the improbability of using stacking to predict disturbances detected by zero BLs, all such pixels (that is, pixels outside of the union map described in Section 2.2) were automatically labeled as "not disturbed," regardless of what the model would have determined for those cases.

Error assessment explicitly accounted for both of these steps in assuring that all algorithms and models were assessed at each 1-year interval at all 1303 forested TimeSync sample points. Specifically, all of the reference sample points automatically labeled "undisturbed" (in both the RF and logistic models) because they fell outside of the union of BL disturbance outputs were simply compared to their TimeSync label to determine classification success or failure. These cases were either counted as true negatives (not in error) or false negatives (i.e., errors of omission). For the remaining sample units (that is, those used for model calibration by virtue of falling in the union of mapped disturbed pixels), evaluation of disturbance predictions used either: 10-fold cross-validation for the logistic model; or (for RF models) comparison of TimeSync interpretations against predictions created from decision trees for which the individual data points had not randomly been selected to calibrate (also called "out-of-bag" predictions). To summarize, despite the fact that not all TimeSync observations were used to train the ensemble models, every member of the TimeSync sample was used to evaluate error in each BL and ensemble model output.

2.4. Temporal sliding of inputs to harmonize change maps and reference data

Because the above stacking process is applied on a year-by-year basis, disagreement among BLs about exactly when a disturbance occurs can be a source of noise in the stacking model. There is variability among the image selection processes used by the change detection algorithms described above: some use all cloud-free pixels, and some use a single date annually. For single-date algorithms, if a forest fire occurs on 01 July of a particular year, it will be detected in that year (assuming it is detected at all) if the representative image was acquired in July or August. If a June image is selected, the fire would be mapped in the following year. Thus, disagreement among algorithms in year of detection might be a function of image date selection rather than differing classification criteria. Image date discrepancies may also affect agreement with TimeSync reference data, potentially introducing further noise and degrading performance of the stacking model.

An approach that is here called "sliding" was used to minimize effects of temporal mismatch issues. The year of every disturbance detected by either a BL or in the reference data was allowed to "slide" one year forward or backward depending on which interval showed the larger increase in shortwave infrared reflectance (Landsat band 5) for that pixel. Reflectance differences were determined through a surface reflectance composite developed from the cloud-free pixel closest in time to a scene-wide target date. This composite target date was the center of the period in the annual scene-level NDVI curve (derived by collapsing MODIS NDVI, 2000–2015) when NDVI rose above its 60th percentile threshold (generally in the summer). Target Julian dates were: 205 (Path 12 Row 28); 201 (P 14 R 32, P 35 R 32, and P 16 R 37); 209 (P 27 R 27); and 237 (P 45 R 30).

It should be emphasized that disturbances were not "double-counted": sliding simply adjusts the year of predicted disturbance in a

Table 2
Model performance when a decision tree agreement threshold for classification of disturbance is used that balances omission and commission error in the validation dataset.

Model	Error rate (omission = commission)
Landsat only	0.65
BLs alone	0.44
BLs + Landsat	0.44
BLs + Topography + Forest Type	0.43
BLs + Landsat + Topography + Forest Type	0.40
Logistic	0.54

way that increases likelihood of alignment among mapping approaches. Sliding was not practiced for long-term disturbances stretching across more than four years. Ensemble-rule models were also run and subjected to the above error assessment without sliding to verify that the process improved prediction accuracy.

3. Results

Fig. 3 shows the accuracy of several types of classifiers against our reference sample of TimeSync observations at the six study sites. These classifiers included: the binary (“disturbed”/”not disturbed”) output of each of the BLs, represented as individual points; the stacked ensemble map using a logistic regression fusion rule; and stacked ensembles using RF decision rules for each of the five sets of inputs described above (Section 2.3). Ensemble model results are displayed in Fig. 3 as a linear series of points representing performance at progressively more restrictive disturbance classification thresholds (i.e., requiring increasing unanimity among decision trees for RF models and increasing probability values for the logistic regression-rule model).

Table 2 summarizes the model results given in Fig. 3 by focusing on performance at the decision tree threshold (i.e., the proportion of decision trees required to label a pixel “disturbed”) that balanced the omission and commission error rates in the validation dataset. In other words, Table 2 expresses performance along the 1:1 line in Fig. 3.

Rates of omission and commission were much greater for each of the BLs against TimeSync's broad definition of disturbance than they were against the data in their published validations (cited earlier). The average annual disturbance rate across the TimeSync sample was 6.2%. Some algorithms skewed conservative with smaller rates of commission and higher omission (VCT, MIICA, CCDC), while others were the opposite (particularly LandTendr). Omission and commission error rates were most balanced for VerDET and ITRA. Smoothing the input time series with “synthetic” imagery produced by CCDC reduced commission error and increased omission error for VCT, while it only increased omission error for LandTrendr. MIICA, too, used CCDC-processed synthetic imagery, but was not run with surface reflectance for comparison.

The “Landsat only” RF model, which like the other ensemble models benefited from being trained with TimeSync data (and a definition of disturbance directly corresponding to that used in the validation dataset), produced marginally smaller error rates than most of the BLs. The two stacking approaches (RF- and logistic regression-rule fusion) using the BL outputs as independent variables resulted in much more accurate models than the Landsat-only model (Table 2). The RF-rule classifier produced lower omission and commission rates than the more conventional logistic regression-rule approach. In general, the RF-rule classifiers gained slight performance increases as additional inputs were added.

The RF models each performed slightly better with respect to the reference data when temporal “sliding” of BL- and TimeSync-detected forest changes was practiced prior to the stacking procedure. Assessed at the point where omission and commission error rates were balanced (see Fig. 3 and Table 2, both of which included the effects of sliding), not sliding reduced the performance of the best model from 40% to 42%.

While the results of this paper rely primarily upon the designed error assessment shown in Fig. 3, it was important to verify that the model predictions could be translated into actual maps. Fig. 4 illustrates mapped results for 2008 and 2009, using the model with the smallest assessed balance point between omission and commission error Fig. 3: BLs + Landsat + Topography + Forest Type) in Scene 45/30 (Oregon). Different colors represent different proportions of decision trees in the ensemble “voting” for a disturbance prediction. Using decision tree agreement as a metric of Random Forests' certainty in relation to the data used to train the model, the areas of greater certainty of disturbance (red and orange classes) clustered into patches associated with forest harvest and a large forest fire (2009). Some of these areas were also predicted as “disturbed” in either preceding or subsequent years, although usually by a minority of decision trees (light gray in Fig. 4).

Fig. 5 shows an aggregated 2005–2011 disturbance map created from outputs such as those in Fig. 4. The year of any pixel labeled as “disturbed” by more than either 50% (upper map) or 75% (lower map) of decision trees was labeled with the year of the detected event. Using the probability structure created by the stacking process in this way allows relatively large changes in the area mapped as “disturbed” (specified in the center of Fig. 5) without major changes to the map's spatial patterns. Additionally, the impact of such changes on accuracy can be directly inferred from the series of model omission/commission points shown in Fig. 3: more conservative thresholds result in reduction of commission error, but also cause an increase in omission error.

4. Discussion

Implications of this experiment are discussed here in terms of: 1) how the results suggest a different paradigm in both forest change detection in particular and resource mapping in general (Section 4.1); 2) how stacking offers a simple means for aligning maps with independent estimates based on an underlying probability structure (Section 4.2), and; 3) how burgeoning cloud computing platforms may accommodate both the sharing of code and large computing resources needed to realize the accuracy benefits illustrated in this paper (Section 4.3).

4.1. Stacking and change detection

The definition of “disturbance” is not absolute. The events that kill, injure, or remove trees exist on temporal, spatial, and magnitude gradients (Masek et al., 2015), and the point on those gradients where a disturbance becomes meaningful varies depending upon the type of disturbance and the particular application, as well as disciplinary perspectives. The relativity of what constitutes a “meaningful” disturbance, and the prevalence mentioned earlier of subtle and complex forest changes, create important potential for “generalization error,” which is the error that results when an algorithm is applied outside of the realm for which it was optimized. While each of the BLs has performed well in published validation exercises, apparent error when measured against our highly inclusive disturbance reference database was quite large, with omission and/or commission error rates often above 60%.

It is instructive that an RF model built with only a composite “before” and “after” Landsat image performed with slightly reduced error rates against the TimeSync reference data than most of the BLs. This suggested the role of generalization error: the RF model was trained with data perfectly aligned with the validation data, while algorithms accessing much more imagery were calibrated independently and were therefore searching for slightly different disturbance profiles. This in itself may be an advantage of stacking – outside of any improvements in the acuity of the change detection process, the secondary model aligns the map classes with the definitions built into a specific application's reference data. It will rarely be possible for a general user to “re-tune”

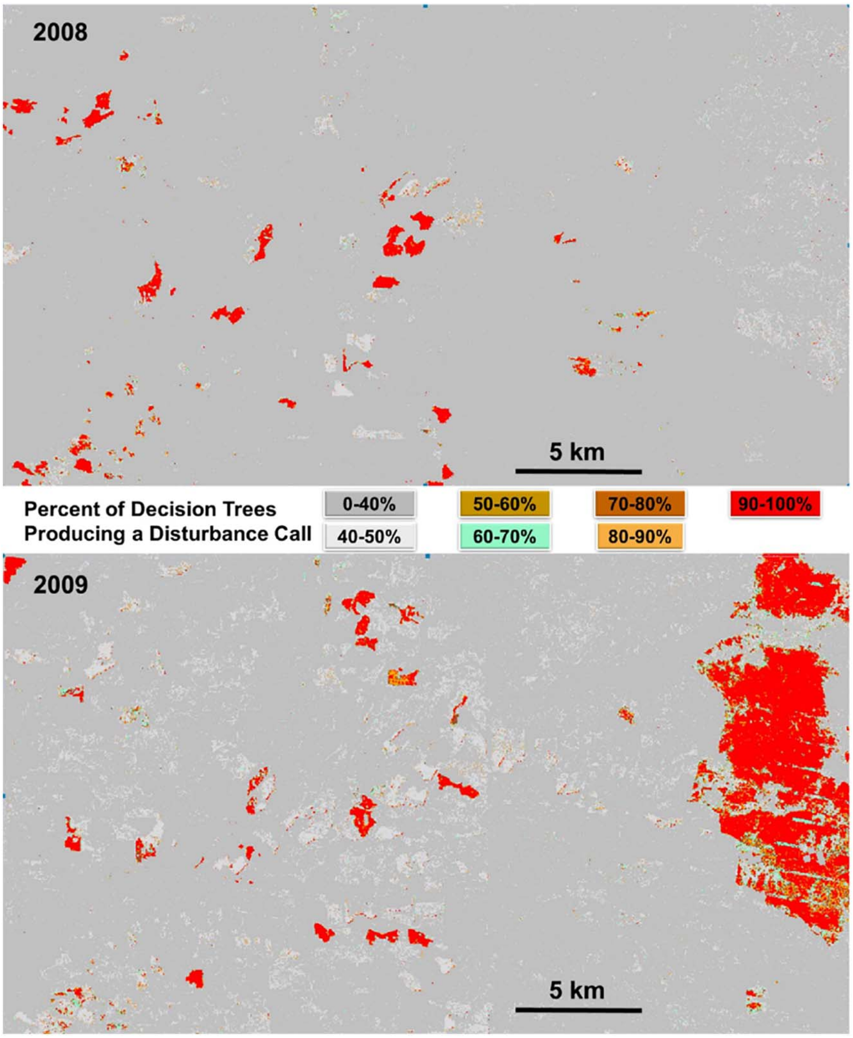

Fig. 4. Predicted disturbance in an area in Oregon (Path/Row 45/30) mapped in successive years with the BLs + Landsat + Topography + Forest Type ensemble model. Different colors, as indicated in the legend, were assigned to different ranges of decision tree agreement for the RF model used in the stacking process.

the search profile of an individual BL, despite the increasing trend toward open-source code. With stacking, a user with an empirical reference dataset should be able to access publically available change classifiers in a way that is harmonized with their own needs.

Beyond alignment advantages, our results suggest that stacking of multiple BLs also increases change detection acuity. Adding BL outputs greatly improved classifier performance above using image pairs alone, with selected supplementary inputs (forest type map, topography) generating further improvements (Table 2). The apparent value of supplementary inputs in change classification was also seen by Schroeder et al. (2017) and Kennedy et al. (2015), who used Landsat time series with RF to predict disturbance type (fire, harvest, etc).

It was initially hypothesized that an ensemble of classifiers with diverse specialties, if integrated with an effective "fusion" rule (see Fig. 1), could be used to improve change classification. Recent research has demonstrated that Landsat-based change detection algorithms do indeed produce diverse products, often varying not only in spatial pattern but in proportion of the landscape labeled as "disturbed" (Cohen et al., 2017).The current study focused on estimating the performance value of stacking, leaving investigation of the specific specialties of different BLs to future work. Results of such an investigation could provide a basis for narrowing the ensemble if computing costs were restrictive. Furthermore, understanding how existing algorithms and ensembles perform on different classes of problems may enable targeting of new ensemble members toward areas of aggregate poor performance.

The tested BLs vary in their use of time series context as well as the spectra they consider, resulting in different patterns of omission and commission error (Fig. 3). While the focus of this study did not require controlling for the effect of different temporal and spectral parameters among the algorithms, it is nevertheless possible to observe the effects of different temporal smoothing options for VCT and LandTrendr. Each was run upon an annual time series of both surface reflectance and synthetic imagery produced through CCDC (Zhu et al., 2015). Synthetic CCDC images allow an atmosphere-free and phenologically controlled look at the landscape, although the signal associated with disturbances missed by CCDC is minimized along with atmospheric noise.

As might be expected, VCT showed a corresponding reduction in commission error and an increase in omission error with synthetic imagery. LandTrendr's output had more omission error with synthetic imagery, but not the expected decreased commission error. More research is needed into the value and effect of synthetic imagery when used with complex time series analyses. Since input bands clearly have an impact on algorithm performance, it would also be of interest to explore alternative ensemble configurations. Under the theory that ensembles add value when member classifiers are informative in non-overlapping domains, it might be possible to create a diverse and effective ensemble for use in stacking simply by applying a series of different threshold and noise filtering parameters when running a single algorithm across a series of bands or transformations.

Traditional parametric stacking, using logistic regression, generated more errors than ensembles using a secondary (RF) ensemble process as the fusion rule. This result is consistent with the reduction of overfitting often attributed to ensemble learners such as RF, but gains in model

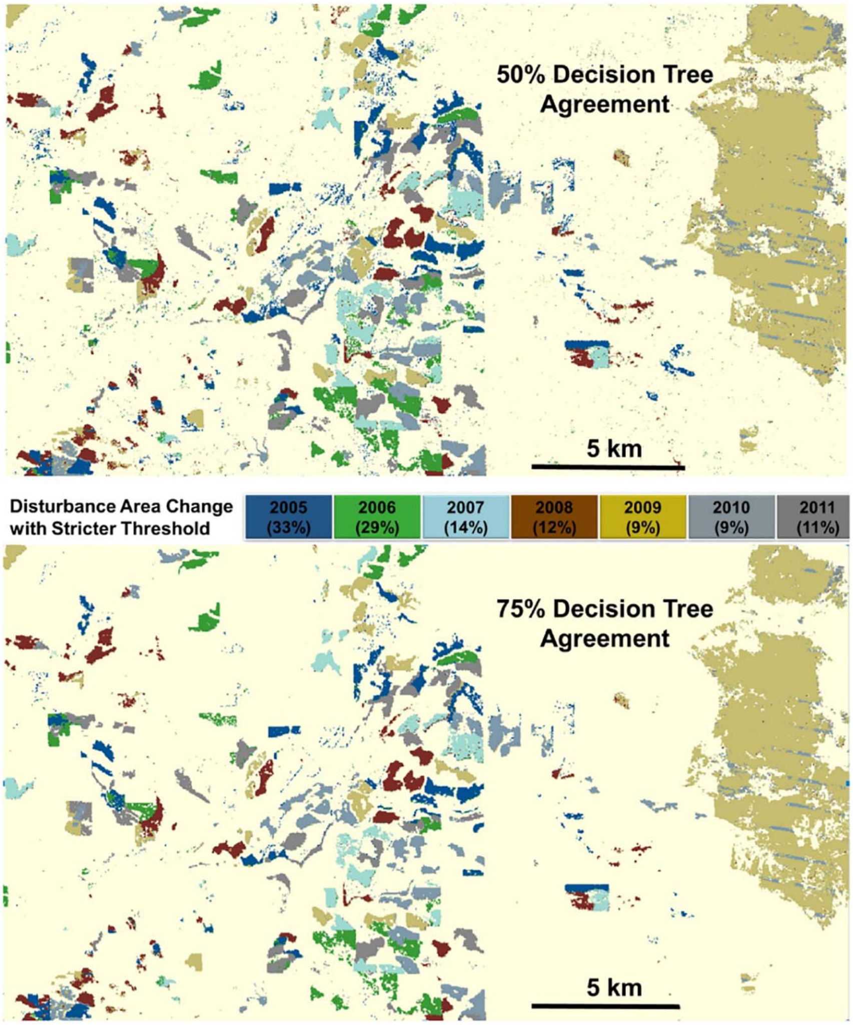
50% Decision Tree Agreement
5 km
Disturbance Area Change with Stricter Threshold
2005 (33%)
2006 (29%)
2007 (14%)
2008 (12%)
2009 (9%)
2010 (9%)
2011 (11%)
75% Decision Tree Agreement
5 km

Fig. 5. An example of modifying the amount of change displayed in a map by varying the RF agreement threshold associated with the model used to implement the stacking process. Mapped disturbance in each year is color-coded, and the percent area reduction for each year under the larger agreement threshold is displayed in the legend by year.

performance must be weighed against the inability of "black box" machine learning algorithms to explicitly parameterize contributions and interactions among independent variables. In applications focusing on these exploratory questions rather than map accuracy, parametric fusion rules may indeed be attractive.

The benefits of "sliding" – the procedure proposed here for increasing temporal alignment of disturbance dates among BLs and reference data – are more ambiguous. When this laborious process was implemented with the lowest-error RF-based ensemble model, the average classification improvement was only on the order of a 1.85% reduction in omission and commission. This level of improvement may not be practically beneficial, and in any case should be considered against substantial additional computing costs.

As mentioned earlier, disturbances resulting in partial canopy removal and subtle spectral changes are common in many ecosystems and represent the majority of error and disagreement among change detection algorithms (Cohen et al., 2017). The distribution of BL performance in Fig. 3 suggests that change detection algorithms and models can successfully detect these changes (i.e., achieve low levels of omission), but doing so requires relaxation of thresholds for the "disturbed" class, which results in additional commission error. The 30-m spatial grain of the Landsat platform may be a limiting factor; localized events affecting only a small portion of the ground field of view may result in minor spectral changes that are easily confused with noise. It is noteworthy, however, that stacking with RF, and with logistic regression to lesser extent, seems to improve the terms of this tradeoff. Omission error is not eliminated, but its reduction comes with a smaller cost in commission error (Fig. 3).

The potential uses of stacking extend well beyond the Landsat platform and beyond the task of change detection. There is nothing Landsat-specific about the tested process, and in fact the topographic independent variables contributing to the best-performing (i.e., lowest-error) model in this paper were not Landsat-based. Stacking could be performed upon change detection maps predicted from a dense time series composed of harmonized Sentinel-2 and Landsat data, or one might map changes using each sensor in parallel and then integrate the outputs through stacking. The focus of this paper was forest change detection, but stacking could be applied to classification problems involving detection of any surface feature, and similar performance benefits above individual BL algorithms should be apparent whenever the ensemble contains inputs which are both informative and non-overlapping.

4.2. *Stacking as an instrument of matching map totals to inventory totals*

Map totals or pixel counts are frequently used to estimate the area of a particular cover class or a population parameter such as mean biomass. However, many of these applications address uncertainty in an ad-hoc manner that does not correctly reflect precision at the population level (Stahl et al., 2016). Good practice guidelines suggest that inferences about populations be based upon probability samples of high-quality reference data (Olofsson et al., 2014), and that maps may be useful as auxiliary variables in the reduction of sample-based uncertainty (Stehman, 2009). It is only under specific sets of assumptions e.g. (McRoberts et al., 2014) that remotely sensed maps may themselves form the basis for the estimates.

This raises the question of how to reconcile authoritative population-level estimates predicted from designed samples with the spatial context provided by remotely sensed maps. When maps are combined as spatial inputs to ecosystem models, an approach called "PDF weaving" can be used with Monte Carlo error simulations to ensure agreement between maps values and design-based estimates (Healey et al., 2014). Specifically, constraints built into systems of linear equations are used to construct probability density functions (PDFs) that control how map values are varied in the Monte Carlo process, such that: 1) the probability of error being simulated for any individual map pixel is determined by pixel/plot validation metrics, and 2) when all pixels are added together for a given error simulation, they equal a value drawn randomly from a distribution defined by the standard error of the authoritative population estimate (Healey et al., 2014). This approach is not useful, however, when a single visualization is required.

One possible approach is to change the number of map units labeled as a particular class (in the current context, "disturbed" or "not disturbed") so that the area associated with that class matches the independent estimate (or upper/lower confidence bounds). Stacking with either logistic regression or RF fusion rules produces class association values within a [0,1] interval that can at least loosely be interpreted as ordinal probabilities. As illustrated in Fig. 5, if more of a particular class is needed to ensure map agreement with independent estimates, more pixels can be added by dropping the probability threshold used for classification (Freeman and Moisen, 2008). The threshold needed for any targeted area of disturbance (or other type of mapped class) should be evident from the cumulative distribution function of decision tree agreement proportions across the map. Fig. 3 indicates that changing such thresholds offers predictable tradeoffs in the balance of false positive and false negative error rates, but Fig. 5 suggests that altering thresholds for the purpose of synchronizing mapped and sample-based estimates does not fundamentally alter spatial patterns apparent in the map.

Geospatial filters (implementing neighborhood functions, or minimum map units, for example) could alternatively be used to increase or decrease the number of pixels labeled as "disturbed," but such a process would, computationally, be less direct than simply choosing a decision tree agreement threshold from a cumulative distribution function of voting scores. More importantly, addition or elimination of "disturbed" pixels should ideally be targeted by some underlying measure of likelihood; this is achieved with the RF stacking scores, but not through simple geometric filters.

4.3. New methods and new strategies

While this investigation demonstrated performance benefits of aggregating multiple classifiers in an ensemble-based learning system, there are practical barriers to broader implementation:

1) Access to code for multiple algorithms – complex image-processing algorithms have traditionally been, if not proprietary, difficult to share because of dependencies on local system properties;
2) Image acquisition and pre-processing costs – a few of the algorithms described here leverage every available clear pixel-level observation in the Landsat archive, representing massive system demands;
3) Computation volume – running multiple algorithms and then implementing a secondary fusion model clearly represents a higher computation load than a single classifier.

As cloud computing becomes more common in remote sensing, these barriers to stacking are becoming less important. Cloud-based platforms typically offer access to massive remote computing resources, making computationally intensive approaches like stacking more feasible. The sharing of ideas enabled through such platforms may be just as important. The authors of the algorithms included in this study have committed to making their processes easily accessible through Google Earth Engine (GEE). GEE enables adroit and standardized access to the entire USGS EROS Landsat archive (including surface reflectance products), with an application programming interface (API) allowing modular (and shareable) approaches to data processing. Easy sharing of code and standardized system properties on platforms such as GEE may both fertilize innovation and provide access to the algorithms needed to build an ensemble.

Any shift toward ensemble methods may foster new attitudes in algorithm development. To the extent that ensembles benefit from the addition of informative BLs that are uncorrelated with existing members, it may be productive to develop algorithms targeting elevated performance in narrow, under-represented problems rather than performance across a range of problems already addressed by other processes. Wolpert and Macready's (1997) "No Free Lunch" theorems suggest that the range of a classifier's optimization is limited. With the potential of stacking to weight ensemble member responses according to their empirically determined niche, an algorithm that does only one thing well can be a critical advance if that one thing is a unique contribution.

5. Conclusions

Comparisons against a reference dataset in six areas across the US demonstrated that the performance of multiple forest change detection algorithms can be improved through their inclusion in a learning ensemble. Any algorithm will suffer from generalization error when applied to a problem outside of the conditions for which it was optimized. In light of the highly variable and complex spectral response of different types of forest disturbances across diverse ecosystems, the potential for generalization error in this field is high. One advantage of stacked generalization, or "stacking" (Wolpert, 1992), is simply alignment of classifier outputs with the definitions and parameters implicit in the training data collected for the application.

For the reference data collected for this study, which included subtle changes often ignored by automated change detection processes, stacking produced results superior to those of any BL. Conservative BLs minimized mapping of false changes (with commission error rates in the 20–30% range), but did so at the cost of omission error rates around 80–90%. More inclusive BLs somewhat reduced omission error (60–75%), but they typically included false positive errors at rates above 75%. Stacking greatly reduced commission error while keeping omission error relatively small; stacking with a conventional parametric model balanced omission/commission error at 54%, while stacking with RF models reduced error rates to approximately 40%. The addition to the stacking ensemble of non-BL inputs, such as topography and raw imagery, improved omission and commission error rates by approximately 4%. Because different BLs use different imagery and may map detected land cover changes in different years, maximizing cross-algorithm agreement by "sliding" of detected changes forward or backward by a single year may be useful. However, our tests showed only marginal accuracy improvements, which may in some applications not be worth the additional processing.

It is likely that advantages of stacking observed here will be robust beyond forest change detection. To the extent that detection of complex or subtle surface phenomena require specialized signal processing, ensembles may facilitate combination of different algorithms' specialties in the context of a single, more general application. Stacking through comparison with empirical data provides an effective approach to weighting alternative learning algorithms. As cloud computing enables greater access to different algorithms and allows more efficient processing of large datasets, the advantages of ensemble methods are likely to become more relevant.

Acknowledgements

The authors wish to thank several individuals for their contributions to this study: Kevin Megown; Jeff Eidenshink; Brian Schwind; Karen Schleeweis; and Mathew Gregory. Institutional support was also provided by the Forest Service Information Resources Direction Board through the Landscape Change Monitoring System, the USGS Earth Resources Observations and Science Center, Google Earth Engine, the Oregon Department of Forestry, and the NASA Carbon Monitoring System (NASA reference number: NNH13AW62I).

References

Ahmed, O.S., Wulder, M.A., White, J.C., Hermosilla, T., Coops, N.C., Franklin, S.E., 2017. Classification of annual non-stand replacing boreal forest change in Canada using Landsat time series: a case study in northern Ontario. Remote Sens. Lett. 8, 29–37.

Antropov, O., Rauste, Y., Väänänen, A., Mutanen, T., Häme, T., 2016. Mapping forest disturbance using long time series of Sentinel-1 data: case studies over boreal and tropical forests. In: 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 3906–3909.

Breiman, L., 1996a. Bagging predictors. Mach. Learn. 24, 123–140.

Breiman, L., 1996b. Stacked regressions. Mach. Learn. 24, 49–64.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

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.

Brooks, E.B., Wynne, R.H., Thomas, V.A., Blinn, C.E., Coulston, J.W., 2014. On-the-fly massively multitemporal change detection using statistical quality control charts and Landsat data. IEEE Trans. Geosci. Remote Sens. 52, 3316–3332.

Chambers, J.Q., Fisher, J.I., Zeng, H., Chapman, E.L., Baker, D.B., Hurtt, G.C., 2007. Hurricane Katrina's Carbon Footprint on U.S. Gulf Coast Forests. Science 318, 1107.

Cohen, W.B., Goward, S.N., 2004. Landsat's role in ecological applications of remote sensing. Bioscience 54, 535–545.

Cohen, W.B., Yang, Z., Kennedy, R., 2010. Detecting trends in forest disturbance and recovery using yearly Landsat time series: 2. TimeSync—tools for calibration and validation. Remote Sens. Environ. 114, 2911–2924.

Cohen, W.B., Yang, Z., Stehman, S.V., Schroeder, T.A., Bell, D.M., Masek, J.G., Huang, C., Meigs, G.W., 2016. Forest disturbance across the conterminous United States from 1985–2012: the emerging dominance of forest decline. For. Ecol. Manag. 360, 242–252.

Cohen, W., Healey, S., Yang, Z., Stehman, S., Brewer, C., Brooks, E., Gorelick, N., Huang, C., Hughes, M., Kennedy, R., Loveland, T., Moisen, G., Schroeder, T., Vogelmann, J., Woodcock, C., Yang, L., Zhu, Z., 2017. How similar are forest disturbance maps derived from different Landsat time series algorithms? Forests 8, 98.

Cutler, D.R., Edwards, T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., Lawler, J.J., 2007. Random forests for classification in ecology. Ecology 88, 2783–2792.

DeVries, B., Verbesselt, J., Kooistra, L., Herold, M., 2015. Robust monitoring of small-scale forest disturbances in a tropical montane forest using Landsat time series. Remote Sens. Environ. 161, 107–121.

Engler, R., Waser, L.T., Zimmermann, N.E., Schaub, M., Berdos, S., Ginzler, C., Psomas, A., 2013. Combining ensemble modeling and remote sensing for mapping individual tree species at high spatial resolution. For. Ecol. Manag. 310, 64–73.

Finco, M.V., Quayle, B., Zhang, Y., Lecker, J., Megown, K.A., Brewer, C.K., 2012. Monitoring Trends in Burn Severity (MTBS): monitoring wildfire activity for the past quarter century using Landsat data. In: Morin, R.S., Liknes, G.C. (Eds.), Forest Inventory and Analysis (FIA) Science Symposium. USDA Forest Service, Northern Research Station, Baltimore, MD, pp. 222–228.

Foody, G.M., Boyd, D.S., Sanchez-Hernandez, C., 2007. Mapping a specific class with an ensemble of classifiers. Int. J. Remote Sens. 28, 1733–1746.

Freeman, E.A., Moisen, G.G., 2008. A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. Ecol. Model. 217, 48–58.

González-Olabarria, J.-R., Pukkala, T., 2011. Integrating fire risk considerations in landscape-level forest planning. For. Ecol. Manag. 261, 278–287.

Grinand, C., Rakotomalala, F., Gond, V., Vaudry, R., Bernoux, M., Vieilledent, G., 2013. Estimating deforestation in tropical humid and dry forests in Madagascar from 2000 to 2010 using multi-date Landsat satellite images and the random forests classifier. Remote Sens. Environ. 139, 68–80.

Healey, S.P., Yang, Z.Q., Cohen, W.B., Pierce, D.J., 2006. Application of two regression-based methods to estimate the effects of partial harvest on forest structure using Landsat data. Remote Sens. Environ. 101, 115–126.

Healey, S.P., Cohen, W.B., Spies, T.A., Moeur, M., Pflugmacher, D., Whitley, M.G., Lefsky, M., 2008. The relative impact of harvest and fire upon Landscape-level dynamics of older forests: lessons from the Northwest Forest Plan. Ecosystems 11, 1106–1119.

Healey, S.P., Urbanski, S.P., Patterson, P.L., Garrard, C., 2014. A framework for simulating map error in ecosystem models. Remote Sens. Environ. 150, 207–217.

Huang, C., Goward, S.N., Masek, J.G., Thomas, N., Zhu, Z., Vogelmann, J.E., 2010. An automated approach for reconstructing recent forest disturbance history using dense Landsat time series stacks. Remote Sens. Environ. 114, 183–198.

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

Jin, S., Yang, L., Danielson, P., Homer, C., Fry, J., Xian, G., 2013. A comprehensive change detection method for updating the National Land Cover Database to circa 2011. Remote Sens. Environ. 132, 159–175.

Kandel, K., Huettmann, F., Suwal, M.K., Ram Regmi, G., Nijman, V., Nekaris, K.A.I., Lama, S.T., Thapa, A., Sharma, H.P., Subedi, T.R., 2015. Rapid multi-nation distribution assessment of a charismatic conservation species using open access ensemble model GIS predictions: red panda (*Ailurus fulgens*) in the Hindu-Kush Himalaya region. Biol. Conserv. 181, 150–161.

Kavzoglu, T., Sahin, E.K., Colkesen, I., 2014. Landslide susceptibility mapping using GIS-based multi-criteria decision analysis, support vector machines, and logistic regression. Landslides 11, 425–439.

Kennedy, R.E., Yang, Z., Cohen, W.B., 2010. Detecting trends in forest disturbance and recovery using yearly Landsat time series: 1. LandTrendr—temporal segmentation algorithms. Remote Sens. Environ. 114, 2897–2910.

Kennedy, R.E., Yang, Z., Cohen, W.B., Pfaff, E., Braaten, J., Nelson, P., 2012. Spatial and temporal patterns of forest disturbance and regrowth within the area of the Northwest Forest Plan. Remote Sens. Environ. 122, 117–133.

Kennedy, R.E., Yang, Z., Braaten, J., Copass, C., Antonova, N., Jordan, C., Nelson, P., 2015. Attribution of disturbance change agent from Landsat time-series in support of habitat monitoring in the Puget Sound region, USA. Remote Sens. Environ. 166, 271–285.

Kurz, W.A., Dymond, C.C., White, T.M., Stinson, G., Shaw, C.H., Rampley, G.J., Smyth, C., Simpson, B.N., Neilson, E.T., Trofymow, J.A., Metsaranta, J., Apps, M.J., 2009. CBM-CFS3: a model of carbon-dynamics in forestry and land-use change implementing IPCC standards. Ecol. Model. 220, 480–504.

Li, W., Ciais, P., MacBean, N., Peng, S., Defourny, P., Bontemps, S., 2016. Major forest changes and land cover transitions based on plant functional types derived from the ESA CCI Land Cover product. Int. J. Appl. Earth Obs. Geoinf. 47, 30–39.

Li, Z.-W., Xin, X.-P., Tang, H., Yang, F., Chen, B.-R., Zhang, B.-H., 2017. Estimating grassland LAI using the Random Forests approach and Landsat imagery in the meadow steppe of Hulunber, China. J. Integr. Agric. 16, 286–297.

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R News 2, 18–22.

Loveland, T.R., Dwyer, J.L., 2012. Landsat: building a strong future. Remote Sens. Environ. 122, 22–29.

Masek, J.G., Hayes, D.J., Hughes, M.J., Healey, S.P., Turner, D.P., 2015. The role of remote sensing in process-scaling studies of managed forest ecosystems. For. Ecol. Manag. 355, 109–123.

Masek, J.G., Vermote, E.F., Saleous, N.E., Wolfe, R., Hall, F.G., Huemmrich, K.F., Gao, F., Kutler, J., Lim, T.K., 2006. A Landsat surface reflectance dataset for North America, 1990–2000. IEEE Geosci. Remote Sens. Lett. 3, 68–72.

McRoberts, R.E., Næsset, E., Gobakken, T., 2014. Estimation for inaccessible and non-sampled forest areas using model-based inference and remotely sensed auxiliary information. Remote Sens. Environ. 154, 226–233.

Meigs, G.W., Kennedy, R.E., Gray, A.N., Gregory, M.J., 2015. Spatiotemporal dynamics of recent mountain pine beetle and western spruce budworm outbreaks across the Pacific Northwest Region, USA. For. Ecol. Manag. 339, 71–86.

Meyer, M.C., 2013. Semi-parametric additive constrained regression. J. Nonparametric Stat. 25, 715.

Moisen, G.G., Meyer, M.C., Schroeder, T.A., Liao, X., Schleeweis, K.G., Freeman, E.A., Toney, J.C., 2016. Shape selection in Landsat time series: a tool for monitoring forest dynamics. Glob. Chang. Biol. 22, 3518–3528.

Olofsson, P., Foody, G.M., Herold, M., Stehman, S.V., Woodcock, C.E., Wulder, M.A., 2014. Good practices for estimating area and assessing accuracy of land change. Remote Sens. Environ. 148, 42–57.

Oza, N.C., Tumer, K., 2008. Classifier ensembles: select real-world applications. Information Fusion 9, 4–20.

Powell, S.L., Cohen, W.B., Healey, S.P., Kennedy, R.E., Moisen, G.G., Pierce, K.B., Ohmann, J.L., 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, 1053–1068.

R Core Team, 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Prasad, A., Iverson, L., Liaw, A., 2006. Newer classification and regression tree techniques: bagging and Random Forests for ecological prediction. Ecosystems 9, 181–199.

Reid, S., Grudic, G., 2009. Regularized linear models in stacked generalization. In: Benediktsson, J., Kittler, J., Roli, F. (Eds.), Multiple Classifier Systems. Springer, Berlin Heidelberg, pp. 112–121.

Ruefenacht, B., Finco, M.V., Nelson, M.D., Czaplewski, R., Helmer, E.H., Blackard, J.A., Holden, G.R., Lister, A.J., Salajanu, D., Weyermann, D., Winterberger, K., 2008. Conterminous U.S. and Alaska forest type mapping using forest inventory and analysis data. Photogramm. Eng. Remote. Sens. 74, 1379–1388.

Sáez, J.A., Galar, M., Luengo, J., Herrera, F., 2013. Tackling the problem of classification with noisy data using Multiple Classifier Systems: analysis of the performance and robustness. Inf. Sci. 247, 1–20.

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

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.

Schwind, B., Brewer, K., Quayle, B., Eidenshink, J.C., 2010. Establishing a nationwide baseline of historical burn-severity data to support monitoring of trends in wildfire effects and national fire policies. In: Pye, J.M., Rauscher, H., Sands, Y., Lee, D.C., Beatty, J.S. (Eds.), Advances in Threat Assessment and Their Application to Forest and Rangeland Management. USDA Forest Service, Pacific Northwest Research Station, Portland, OR, pp. 381–396.

Seidl, R., Schelhaas, M.-J., Rammer, W., Verkerk, P.J., 2014. Increasing forest disturbances in Europe and their impact on carbon storage. Nat. Clim. Chang. 4, 806–810.

Seilheimer, T.S., Zimmerman, P.L., Stueve, K.M., Perry, C.H., 2013. Landscape-scale modeling of water quality in Lake Superior and Lake Michigan watersheds: how useful are forest-based indicators? J. Great Lakes Res. 39, 211–223.

Smith, W.B., Miles, P.D., Perry, C.H., Pugh, S.A., 2009. Forest Resources of the United States, 2007. F.S. U.S. Department of Agriculture, Washington, DC, pp. 336.

Spies, T.A., Miller, J.D., Buchanan, J.B., Lehmkuhl, J.F., Franklin, J.F., Healey, S.P., Hessburg, P.F., Safford, H.D., Cohen, W.B., Kennedy, R.S.H., Knapp, E.E., Agee, J.K., Moeur, M., 2010. Underestimating risks to the northern spotted owl in fire-prone forests: response to Hanson et al. Conserv. Biol. 24, 330–333.

Stahl, G., Saarela, S., Schnell, S., Holm, S., Breidenbach, J., Healey, S., Patterson, P., Magnussen, S., Naesset, E., McRoberts, R., Gregoire, T., 2016. Use of models in large-area forest surveys: comparing model-assisted, model-based and hybrid estimation. For. Ecosystems 3, 5.

Stehman, S.V., 2009. Model-assisted estimation as a unifying framework for estimating the area of land cover and land-cover change from remote sensing. Remote Sens. Environ. 113, 2455–2462.

Thomas, N.E., Huang, C., Goward, S.N., Powell, S., Rishmawi, K., Schleeweis, K., Hinds, A., 2011. Validation of North American Forest Disturbance dynamics derived from Landsat time series stacks. Remote Sens. Environ. 115, 19–32.

Turner, B.L., Lambin, E.F., Reenberg, A., 2007. The emergence of land change science for global environmental change and sustainability. Proc. Natl. Acad. Sci. 104, 20666–20671.

Verhegghen, A., Eva, H., Ceccherini, G., Achard, F., Gond, V., Gourlet-Fleury, S., Cerutti, P., 2016. The potential of sentinel satellites for burnt area mapping and monitoring in the Congo Basin forests. Remote Sens. 8, 986.

Vogelmann, J.E., Xian, G., Homer, C., Tolk, B., 2012. Monitoring gradual ecosystem change using Landsat time series analyses: case studies in selected forest and rangeland ecosystems. Remote Sens. Environ. 122, 92–105.

Wolpert, D.H., 1992. Stacked generalization. Neural Netw. 5, 241–259.

Wolpert, D.H., Macready, W.G., 1997. No free lunch theorems for optimization. IEEE Trans. Evol. Comput. 1, 67–82.

Woźniak, M., Graña, M., Corchado, E., 2014. A survey of multiple classifier systems as hybrid systems. Information 16, 3–17.

Wulder, M.A., Masek, J.G., Cohen, W.B., Loveland, T.R., Woodcock, C.E., 2012. Opening the archive: How free data has enabled the science and monitoring promise of Landsat. Remote Sens. Environ. 122, 2–10.

Xing, W., Chen, X., Stein, J., Marcinkowski, M., 2016. Temporal predication of dropouts in MOOCs: reaching the low hanging fruit through stacking generalization. Comput. Hum. Behav. 58, 119–129.

Zhu, Z., Woodcock, C.E., 2012. Object-based cloud and cloud shadow detection in Landsat imagery. Remote Sens. Environ. 118, 83–94.

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

Zhu, Z., Woodcock, C.E., Holden, C., Yang, Z., 2015. Generating synthetic Landsat images based on all available Landsat data: predicting Landsat surface reflectance at any given time. Remote Sens. Environ. 162, 67–83.