Remote Sensing of Environment xxx (xxxx) xxxx

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

Quality control and assessment of interpreter consistency of annual land cover reference data in an operational national monitoring program

Bruce W. Pengra[a,*], Stephen V. Stehman[b], Josephine A. Horton[c], Daryn J. Dockter[a], Todd A. Schroeder[d], Zhiqiang Yang[e], Warren B. Cohen[e,f], Sean P. Healey[g], Thomas R. Loveland[h]

[a] Stinger Ghaffarian Technologies, contractor to the U.S. Geological Survey, Earth Resources Observation and Science (EROS) Center, Sioux Falls, SD 57198, USA
[b] College of Environmental Science and Forestry, State University of New York, Syracuse, NY 13210, USA
[c] Innovate! Inc., contractor to the U.S. Geological Survey EROS Center, Sioux Falls, SD 57198, USA
[d] U.S. Forest Service Southern Research Station, Knoxville, TN 37919, USA
[e] Department of Forest Ecosystems and Society, Oregon State University, Corvallis, OR 97331, USA
[f] U.S. Forest Service, Pacific Northwest Research Station, 3200 SW Jefferson Way, Corvallis, OR 97331, USA
[g] U.S. Forest Service, Rocky Mountain Research Station, 507 25th Street, Ogden, UT 84401, USA
[h] U.S. Geological Survey EROS Center, Sioux Falls, SD 57198, USA

ARTICLE INFO

Keywords:
LCMAP
Validation
Land cover
Land use
Disturbance
Time series
Landsat
TimeSync

ABSTRACT

The U.S. Geological Survey Land Change Monitoring, Assessment and Projection (USGS LCMAP) initiative is working toward a comprehensive capability to characterize land cover and land cover change using dense Landsat time series data. A suite of products including annual land cover maps and annual land cover change maps will be produced using the Landsat 4-8 data record. LCMAP products will initially be created for the conterminous United States (CONUS) and then extended to include Alaska and Hawaii. A critical component of LCMAP is the collection of reference data using the TimeSync tool, a web-based interface for manually interpreting and recording land cover from Landsat data supplemented with fine resolution imagery and other ancillary data. These reference data will be used for area estimation and validation of the LCMAP annual land cover products. Nearly 12,000 LCMAP reference sample pixels have been interpreted and a simple random subsample of these pixels has been interpreted independently by a second analyst (hereafter referred to as "duplicate interpretations"). The annual land cover reference class labels for the 1984–2016 monitoring period obtained from these duplicate interpretations are used to address the following questions: 1) How consistent are the reference class labels among interpreters overall and per class? 2) Does consistency vary by geographic region? 3) Does consistency vary as interpreters gain experience over time? 4) Does interpreter consistency change with improving availability and quality of imagery from 1984 to 2016? Overall agreement between interpreters was 88%. Class-specific agreement ranged from 46% for Disturbed to 94% for Water, with more prevalent classes (Tree Cover, Grass/Shrub and Cropland) generally having greater agreement than rare classes (Developed, Barren and Wetland). Agreement between interpreters remained approximately the same over the 12-month period during which these interpretations were completed. Increasing availability of Landsat and Google Earth fine resolution data over the 1984 to 2016 monitoring period coincided with increased interpreter consistency for the post-2000 data record. The reference data interpretation and quality assurance protocols implemented for LCMAP demonstrate the technical and practical feasibility of using the Landsat archive and intensive human interpretation to produce national, annual reference land cover data over a 30-year period. Protocols to estimate and enhance interpreter consistency are critical elements to document and ensure the quality of these reference data.

1. Introduction

Understanding land cover and land cover change is central to managing the Earth's natural capital such as soils, forests, water resources, biodiversity and climate (Foley et al., 2005). Global and large-area land cover and land change maps have important and even foundational uses in a variety of research, management and policy applications. Land cover and land change data are crucial inputs for

* Corresponding author.
E-mail address: bpengra@contractor.usgs.gov (B.W. Pengra).

https://doi.org/10.1016/j.rse.2019.111261
Received 25 September 2018; Received in revised form 22 May 2019; Accepted 11 June 2019

climate models (Prestele et al., 2017), forest management and research (Hansen et al., 2013) and the study of biodiversity and habitat loss (Hoekstra et al., 2004). The expanding scope and importance of land cover and land change science drive a need for more accurate land cover and land change data to support these uses (Turner 2nd et al., 2007).

In response to those needs the U.S. Geological Survey (USGS) Land Change Monitoring, Assessment and Projection (LCMAP) initiative will provide a suite of annual map products including land cover and land cover change maps, as well as newly developed change products aimed at predicting a range of land surface disturbances. Initial output will cover the conterminous United States (CONUS) and eventually include Hawaii and Alaska. These outputs are produced from dense time series Landsat data using the USGS Analysis Ready Data (ARD). The USGS ARD are "geo-registered, top of atmosphere and atmospherically corrected products defined in a common equal area projection, accompanied by spatially explicit quality assessment information, and appropriate metadata" (Dwyer et al., 2018).

Assessing the accuracy of land cover and land change maps is widely recognized as an integral part of land cover and land change studies (Olofsson et al., 2014). The overall accuracy and the thematic, spatial and temporal variations in accuracy of land cover data can have a major impact on the products, predictions and conclusions made using that data (Foody, 2015; Olofsson et al., 2013). Nevertheless, the time and resources required to collect reference data for a large-area probability sample present a major challenge, often leading to limited and inadequate validation efforts (Foody, 2010). The resources required to collect reference data via field visits for a large, probability sample from an area the size of CONUS would be prohibitive. In addition, many land cover and land change products, including those being created by the LCMAP initiative, require historical reference data often covering some or all of the Landsat data record through time. Adequate historical field data with the required consistency are not available to provide the reference data required for large-area maps such as those produced by LCMAP. Consequently, it has become common practice to collect reference data using remotely sensed data (Olofsson et al., 2014). By using data of higher quality (e.g., finer resolution) and/or by using more accurate classification methods (e.g., expert interpretation), it is assumed possible to obtain data of higher quality than the map data (Olofsson et al., 2014).

Most large-area map accuracy assessments employ multiple interpreters to obtain the reference class labels (e.g., Scepan et al., 1999; Zhu et al., 2000; Powell et al., 2004; Clark et al., 2012; Wickham et al., 2017). Human interpreters do not always arrive at the same reference classification even when given the same sources of information for determining the reference label, and if interpreters disagree, the possibility of error in the reference classification exists. As stated by Congalton (1991), "It is obvious that in order to adequately assess the accuracy of the remotely sensed classification, accurate ground, or reference data must be collected." The longstanding recognition of the importance of accurate reference data was further highlighted by Foody's (2010, 2013) quantification of the impact of reference data error on accuracy and area estimates. Foody (2013) provided an example for a binary classification in which reference classification error resulted in an overestimate of class abundance (area) by nearly a factor of six even though the accuracy of the reference data was over 90%. Whereas Foody (2013) focused on the impact of reference data error, McRoberts et al. (2018) quantified the impact of variability in the reference class labels on the standard error of area estimates and found that failing to account for interpreter variability could produce underestimates of the standard error by a factor of 2.3.

Good practice guidelines for accuracy assessment (Olofsson et al., 2014) recommend assessing the uncertainty of reference data. To estimate the variability of the LCMAP reference land cover data and also to ensure that the reference data are of high quality, an approach for quality assurance and quality control (QA/QC) was developed and implemented based on obtaining duplicate interpretations from independent analysts for a subset of the full reference sample. Pairwise comparison of these duplicate interpretations provided the basis for estimating agreement among interpreters and for calibration of interpreters through individual and group feedback. We use the duplicate reference annual land cover interpretations for 1984 through 2016 obtained from the sample data to evaluate the following questions. 1) What is the agreement among interpreters overall and by land cover class? 2) Does interpreter agreement differ across four large geographic regions? 3) Does agreement change over the time period during which the interpretations were obtained (i.e., does agreement change as interpreters gain experience)? 4) Does agreement vary as the quality and density of Landsat and fine resolution imagery have improved over the time period monitored, 1984–2016?

In this article, we estimate agreement among multiple interpreters for a large-area (national) land cover monitoring program spanning a long time series (over 30 years). The assessment of interpreter consistency in the LCMAP response design protocol is embedded within an ongoing land cover monitoring activity and consequently the results of this assessment estimate interpreter agreement in a realistic operational setting. Another novel aspect of the LCMAP response design protocol is that a portion of the data collected to estimate agreement and provide information for interpreter feedback is obtained from a probability subsample of the full sample. This aspect of the protocol provides a rigorous basis for estimating interpreter agreement within a design-based inference framework.

2. Use of interpreters in land-cover studies

The response design protocols for reference class labeling in large-area map accuracy assessments vary widely in terms of how interpreters are used. In some applications a single interpreter is employed (Bicheron et al., 2008; Hermosilla et al., 2015, 2018; Sexton et al., 2013) and hence no information is available regarding interpreter variability. More commonly, multiple interpreters are required because of the large sample size of reference classifications. For example, Feng et al. (2016) used 12 analysts to interpret nearly 28,000 sample points to obtain reference data for three epochs of global forest cover change. Typically interpreters undergo common training to establish a baseline of consistency. For example, the interpreters may initially collectively work on a common set of pixels and discuss these cases to establish consistency (Wickham et al., 2017; Tsendbazar et al., 2018) before proceeding to work independently to obtain the interpretations for the reference sample. Further, over the life of the project interpreter consistency may be enhanced by group review of selected cases, often focusing on challenging examples, in an ongoing process of interpreter training and calibration (e.g., Sleeter et al., 2013, Sec. 2.5).

Operationally, if multiple interpreters examine the same sample pixels it will be necessary to resolve disagreements to produce the final reference class labels used in the assessment. For example, the mode reference class may be used (McRoberts et al., 2018), interpreters may be reconvened to decide by consensus a final label (Powell et al., 2004; Zhu et al., 2000), or an expert interpreter may be called upon to provide the final label (Clark et al., 2012). In one of the first large-area accuracy assessments conducted (Scepan et al., 1999), three experts independently interpreted each sample pixel. For the map label to be considered correct, two of the three interpreters had to have assigned a reference label that matched the map label; otherwise, the pixel was deemed incorrectly labeled. However, in these cases for which the map label was incorrect, if the interpreters disagreed on their reference labels the pixel was removed from the assessment (this occurred for approximately 50% of the pixels labeled as incorrect). Scepan et al. (1999) did not report a quantitative evaluation of agreement among the multiple interpreters. The National Land Cover Database (NLCD) of the United States provides additional examples of large-area map accuracy assessments in which multiple interpreters were used (Wickham et al.,

2013, 2017). Although the NLCD response design protocol included provisions for interpreter training and in-progress interpreter feedback to enhance consistency, the NLCD did not carry out a study to quantitatively evaluate interpreter agreement. Gong et al. (2013) employed four interpreters in a first round and three interpreters in a second round of reference data collection for over 36,000 sample units, but no information regarding interpreter consistency was reported.

For some applications in which multiple interpreters were used, the percentage of cases for which interpreters disagreed has been reported. For example, Clark et al. (2012) reported 10% disagreement between duplicate interpretations and Zhu et al. (2000) reported 30% disagreement between duplicates. However, it is very rare that a quantitative analysis of interpreter consistency is reported. Powell et al. (2004) is a notable exception as they examined agreement among five interpreters providing reference labels for a sample of 790, 30 m × 30 m pixels selected from an area covering approximately 54,000 km^2 in Rondônia, Brazil. The legend consisted of five classes at a single point in time, with average agreement between pairs of interpreters reported as 86% overall, 49% for second-growth forest, 81% for pasture, and 92% for primary forest, and no report for Urban/bare soil and Water (Powell et al., 2004, Table 3). Mann and Rothley (2006) employed three interpreters working with a five-class legend and an area covering 2.7 km × 4.4 km to examine how estimates of accuracy varied over the different interpreters. However, they did not report agreement among the three interpreters.

Although issues associated with multiple interpreters and their consistency have been present throughout the history of land cover monitoring, few studies have estimated agreement among interpreters. In those cases where agreement has been estimated, the area mapped has been relatively small and the reference data represent only a single date. In our study, we estimate pairwise interpreter agreement for a probability subsample of nearly 3000 pixels representative of CONUS and a time series of over 30 years. Our results documenting interpreter consistency therefore inform the ongoing development of methods for collecting reference data for large-area, land cover monitoring efforts targeting a long time series of annual observations.

3. Methods

3.1. Overview of LCMAP

A condensed overview of the LCMAP initiative is provided to set the context for the reference data protocol (i.e., response design) that is the focus of this article. LCMAP annual land cover data are being produced using the Continuous Change Detection and Classification (CCDC) algorithm developed by Zhu and Woodcock (2014). The land cover legend includes the classes Developed, Cropland, Tree Cover, Grass/Shrub, Wetland, Water, Snow/Ice, Disturbed and Barren. Annual maps of land cover and land cover change spanning 1985–2017 will be produced for CONUS and four large reporting regions created by aggregating Omernik ecoregions (Fig. 1) (Omernik and Griffith, 2014). Parallel to the LCMAP mapping effort, reference data are being collected for a simple random sample of pixels from CONUS. The reference data will be used to estimate accuracy of the LCMAP products and to produce estimates of land cover composition and change. The reference interpretations are obtained independently of the map classification.

The reference sample consists of single, Landsat-resolution grid-pixels (30 m × 30 m). The sample frame was defined by the full CONUS extent of the National Land Cover Database, which shares the same grid system as LCMAP (Homer et al., 2012). Simple random sampling was used to select the initial sample of 25,000 pixels because it is easy to implement, simple to analyze and amenable to future augmentation to increase the sample size from targeted classes or regions within CONUS. A primary motivation for implementing simple random sampling was that reference data collection needed to begin long before the LCMAP products would be available. Although stratified sampling is often justified to increase the sample size from rare classes (Olofsson et al., 2014), it was not a viable option given the unavailability of maps to construct the strata. The LCMAP sampling strategy includes provision to augment the initial simple random sample to targeted geographic areas or rare classes once the LCMAP products are available to construct strata (Overton and Stehman, 1996). Sample pixels are processed (interpreted) in a random order, so the 11,900 sample pixels analyzed here constitute a simple random sample from CONUS.

3.2. TimeSync reference data collection

The TimeSync tool (Cohen et al., 2010) provides efficient interpreter access to Landsat data. TimeSync accommodates the specific map projection parameters that had been adopted by the LCMAP initiative and was customized to collect a broad range of attributes that could be translated to the LCMAP land cover classes. A system of data collection and data quality assurance was developed through a collaboration between USGS and the U.S. Forest Service (USFS) Landscape Change Monitoring System (LCMS) initiative. The class definitions and rules for collection of the full set of attributes are defined in the Joint Response Design (JRD) (see Supplementary materials).

The Landsat input data are obtained from the Google Earth Engine collection of Landsat 5, 7 and 8 and converted to image files. Data are reprojected to the LCMAP grid system in Albers Equal Area Conic, World Geodetic System 1984 (WGS84). TimeSync uses two basic forms of Landsat display: 1) annual image chips and 2) pixel values graphed through time. Image chips consist of 255 × 255-pixel single-date Landsat subsets from 1984 to 2016 growing season images, which are displayed in sequence (Fig. 2). Analysts can access all usable images for each year allowing them to replace images impacted by clouds, cloud shadow or other data quality issues. Image chips can be displayed in three band combinations.

TimeSync also graphs the spectral values of all cloud free Landsat observations for each sample pixel (Fig. 3). Interpreters can select from Landsat bands and several indices for this “trajectory” display. The display can be toggled to show the values for only the display image date or the values for all Landsat observations in the data collection not identified as cloud or cloud shadow by the Fmask algorithm (from CFmask acquired with pre-Collection 1 data and the Landsat QA band for more recently acquired Collection 1 data) (Zhu and Woodcock, 2012).

3.3. Interpretation protocol

Collection of this reference dataset is a collaboration between the LCMS project of the USFS and the LCMAP initiative of the USGS. To accommodate the different data needs of the two projects, a unique implementation of TimeSync was developed to collect attributes serving the needs of both projects. Because of this the resulting dataset includes more attributes than either individual project requires. These attributes are combined in different ways by the USGS and USFS to produce the output reference data needed by the respective LCMAP and LCMS projects. The interpretation protocols followed to collect these attributes are defined in the JRD produced by USGS and USFS personnel.

3.3.1. Subsampling of pixels and assignment for duplicated interpretations

The subsample of pixels selected for duplicated interpretations provided data used to estimate interpreter consistency and for ongoing interpreter training and calibration. Pixels were processed in sets ranging from 600 to 1400 pixels, with Set 1 being the first 600 pixels in the randomized sample list of all 25,000 sample pixels and subsequent sets created by continuing sequentially through that randomized list (Table 1). Thus, the sets reflect the temporal order in which the pixels were interpreted. Twelve sets represent the 11,900 pixels that had been interpreted at the time of this writing. For Sets 1–12, a minimum of

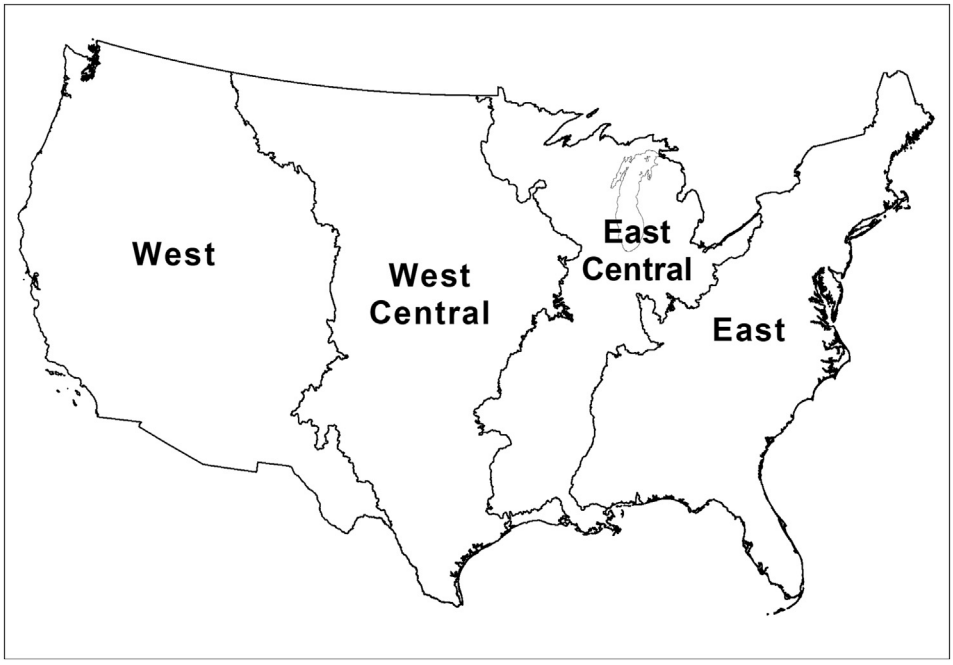

Fig. 1. LCMAP reporting regions from aggregated Omernik ecoregions.

10% of the reference sample pixels were selected for duplicate interpretation via simple random sampling. Only these randomly selected duplicated pixels were used to estimate interpreter agreement, thus ensuring that these estimates were produced from a probability sample. Other pixels selected for a second interpretation were purposively chosen based on QA/QC goals such as targeting geographic regions or classes that were more difficult to interpret consistently. For some sets, fewer purposively selected duplicate interpretations were needed to address the targeted QA/QC concerns, and so in those sets the sample size for the simple random subset was increased. Duplicated sample pixels (random and purposive) that had interpreter disagreement were used to identify issues with specific classes and interpreters, to flag pixels for review and editing, and to provide feedback to interpreters.

The team of interpreters varied in size from 5 to 11 over the course of the study. Interpreter experience ranged from multiple years of working with thematic land cover to mostly classroom experience working with forestry mapping (Table S1). Each interpreter was randomly assigned 100–200 pixels per set with approximately 3 weeks allocated to complete interpretations as the interpreter's schedule allowed (based on an expected 20 h per week per interpreter). Every pixel was randomly assigned to an interpreter for the initial interpretation, with approximately 60% of the pixels also assigned to another independent, randomly assigned interpreter to conduct a duplicate interpretation for QA/QC (Table 1).

3.3.2. *Visual interpretation and classification*

Interpreters were provided with training regarding workflows, response design, class definitions, image interpretation and use of ancillary data. Detailed guidelines and class definitions were available for reference in the JRD. Feedback and ongoing training were also provided to interpreters via e-mail and group teleconferences throughout the collection process. Interpreter questions about any aspect of the process, including individual sample pixels, were answered by the QA/QC review team and shared with all interpreters at the end of each set.

Landsat spectral data displayed in TimeSync were the primary information used for interpretations. Interpreters were also expected to consult the fine resolution aerial imagery in Google Earth for each pixel. These data were often supplemented with older aerial imagery available through EarthExplorer (USGS, 2018) from the USGS National Aerial Photography Program (NAPP) and the National High-Altitude Photography (NHAP) program. Data such as Monitoring Trends in Burn Severity (MTBS) fire polygons (MTBS, 2018) and National Wetland Inventory polygons (USFWS, 2018) were available as well. Interpreters were to use these data to support interpretation of Landsat and fine resolution imagery, but the greatest weight of evidence was to be given to the Landsat data.

Interpreters recorded attributes in three general categories: 1) land use, 2) land cover, and 3) land change process (Fig. 4). Land use and land cover attributes were recorded by the interpreters at vertices (see red circles Fig. 3). For each change in cover, a new vertex is needed.

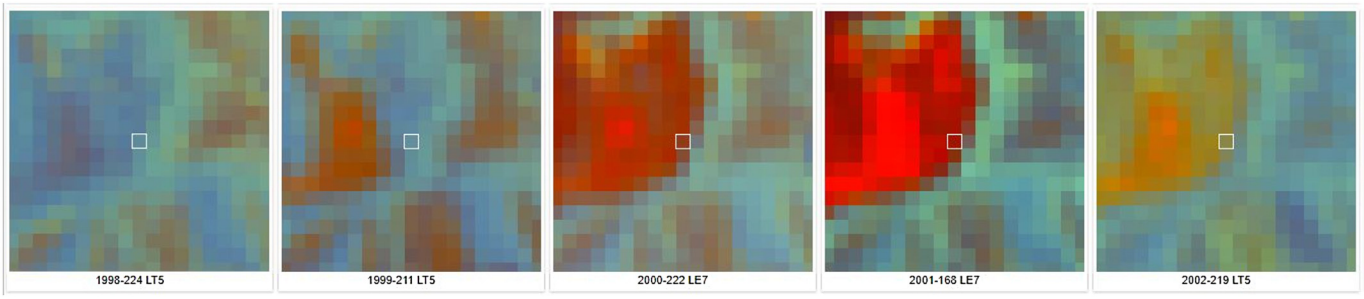

Fig. 2. Five years of Landsat image chips from the TimeSync 1984–2016 time series sequence for sample pixel 11,039, showing a forest harvest over the sample pixel outline (white squares).

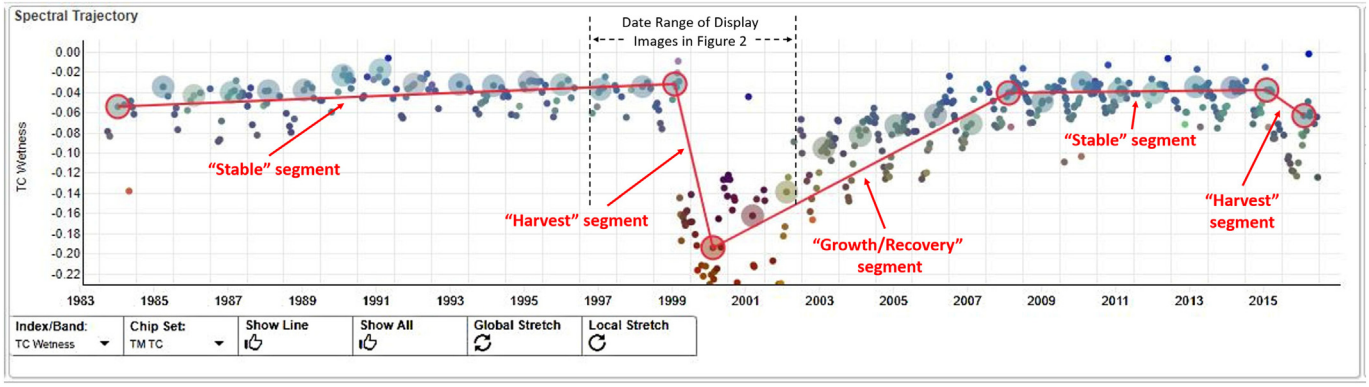

Fig. 3. TimeSync trajectory display gives interpreters a summary view of values for all cloud free Landsat observations. Land use and land cover class labels are recorded at vertices and extrapolated backward to the next vertex. "Segments" between vertices record change processes related to disturbance such as Fire and Harvest (forest), and non-disturbance processes such as Stable and Growth/Recovery.

The attribute at each vertex should describe the use and cover for the year it is recorded and for the years going back in time to the last year before the next vertex. If there is more than one cover or use between these vertices, then another vertex needs to be added to record that different attribute. These attributes are converted into annual labels by extrapolating the recorded values at each vertex back until another labeled vertex is reached. Primary and secondary land use and land cover labels were assigned based on the proportional composition of the single 30 m × 30 m sample pixel as interpreted from the fine resolution imagery. Additional characteristics such as wetland status and the presence of mining or specific types of agriculture were recorded for some classes by marking checkboxes.

Change processes (Fig. 4) are recorded in TimeSync as 'segments' extending between the vertex representing the year before the change can be seen in the data, forward in time to the last vertex representing a year where change from the prior year is evident. Segments range in length from the entire time series to a single year. Change processes impacting any portion of the pixel were recorded. The full set of labels assigned for land use, land cover and change process categories was not designed to produce data that would be used directly as three discrete legends of land characteristics. The cover, use and change characteristics were defined in such a way as to provide the necessary information to enable a crosswalk to the classes of the LCMAP or LCMS legends. This was done with a scripted set of rules for the LCMAP data used in this analysis.

3.4. Quality assurance and quality control

The QA/QC process was designed to meet three objectives: 1) estimate consistency of the reference data among interpreters and over the time span of data collection; 2) use agreement diagnostics and interpretation reviews as a basis for ongoing feedback and calibration of interpreters; and 3) identify errors or inconsistencies in interpretations and correct problems for the final version of these data. Most of the QA/QC process relied on comparing interpretations at sample pixels that had been independently interpreted by random pairs of analysts. All pixels that were assigned for a second interpretation were compared for agreement based on the LCMAP crosswalked land cover classes (Section 3.3.2). Upon completion of each set, a contingency table was created showing overall and per-class agreement between the duplicate interpretations. A second type of contingency table was constructed for each individual interpreter comparing their interpretations to the corresponding duplicate interpretations (Table 2). For QA/QC purposes these interpreter-specific evaluations included both randomly and purposively selected duplicate pixels. However, when estimating interpreter agreement, only those pixels selected for the probability subsample were used.

These individual and overall tables were provided to interpreters at the completion of each set to inform them of which classes they were frequently interpreting differently from the group. When individual interpreters showed repeated disagreement in a category of comparison this was considered an indication that their interpretations were deviating from the group and potentially inconsistent with the JRD. For example, interpreter 137 (Table 2) has labeled well less than half as many observations as Wetland compared to the mix of other interpreters who have interpreted the same duplicate pixels (98 versus 269). This was considered a strong indication that this interpreter was deviating from the JRD for the Wetland class. Interpreters were notified of specific interpretation mistakes discovered during the review process and provided with the QA/QC reviewer's recommended edits. In addition, the QA/QC team used the overall confusion matrix assembled from all duplicate interpretations to identify examples where additional general interpretation guidance was needed for the group as a whole. Teleconferences were held periodically to provide guidance regarding common interpretation issues, to supply further training and to allow for interpreter discussion and knowledge sharing.

4. Results

4.1. Interpreter agreement

A subsample of 2952 out of the 11,900 sample pixels completed to date was randomly selected for duplicate interpretation by a second interpreter. This randomly selected subset was the basis for estimating consistency among all interpretations in the full dataset. Agreement was assessed using data that had been crosswalked to the LCMAP land cover

Table 1
Allocation of sample size and duplicate interpretations per set (sets represent time order of interpretations).

	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8	Set 9	Set 10	Set 11	Set 12	Total
Number of sample pixels	600	600	800	1400	1400	1100	1000	900	1000	1200	1000	900	11,900
Random duplicated interpretations	209	220	400	140	313	270	233	197	233	307	233	197	2952
Purposive duplicated interpretations	191	179	0	660	487	430	367	303	367	493	367	303	4147
Total interpretations	1000	999	1200	2200	2200	1800	1600	1400	1600	2000	1600	1400	18,999

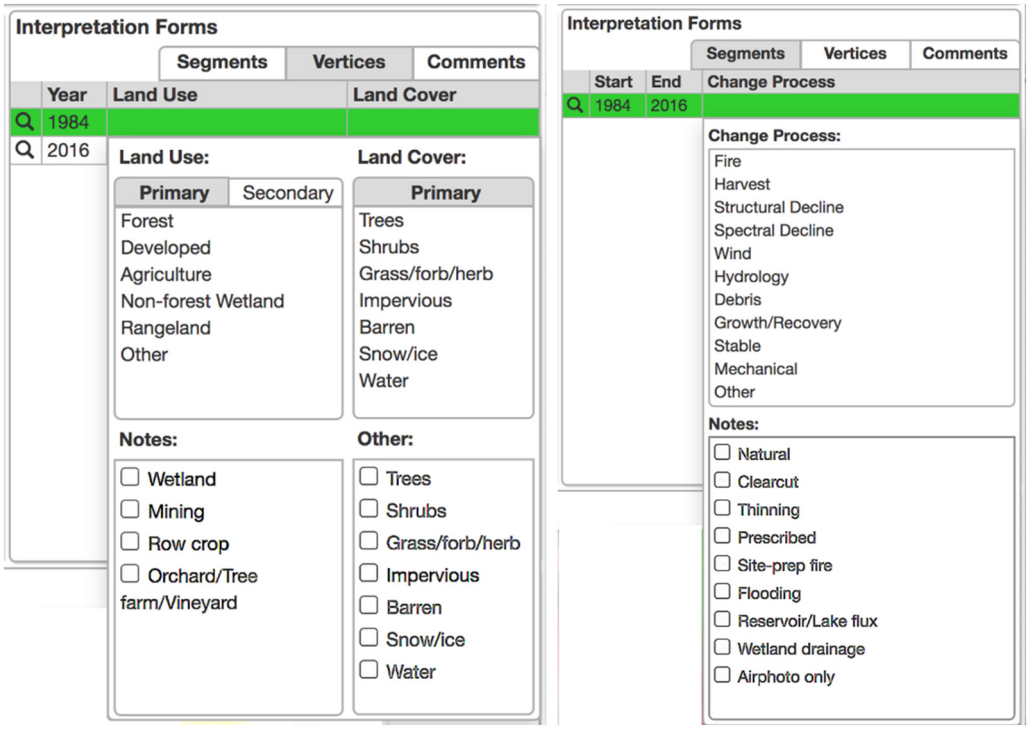

Fig. 4. Land use and land cover attribute labels (left) and change attribute labels (right) available to interpreters in TimeSync.

classes. At each sample pixel, class labels for each year (1984–2016) were compared (Table 3). Because the rows and columns of Table 3 represent random pairs of interpreters, per-class agreement (Table 4) is computed as the average of the row and column agreement.

Overall agreement estimated from the randomly selected subsample of pixels was 88% (Table 3). The four most prevalent classes (Water, Tree Cover, Grass/Shrub and Cropland) had CONUS-level agreement ranging from 89% to 94% (Table 4). The Disturbed (46%), Barren (56%) and Wetland (74%) classes had less agreement, which was expected because these classes have historically been challenging to interpret accurately and consistently. Overall agreement was similar across regions varying from 86% to 89%, but class-specific agreement was much more variable by region. Grass/Shrub was especially variable ranging from 63% in the East Central to 91% in the West Central and the West (Tables 4, S2a–S2d). Tree Cover agreement was 80% in the West Central where tree cover is rarer and often fragmented compared to 92% in the East where tree cover more frequently occurs in large homogeneous patches. The Disturbed class agreement also varied widely with 35% agreement in the West Central compared to 55% agreement in the West (Table 4).

Table 2
Contingency table comparing individual Interpreter 137 to all other interpreters (random and purposively selected pixels from Set 10) for purposes of interpreter evaluation and feedback.

		Other Interpreters - Set 10									
		Water	Developed	Disturbed	Barren	Tree cover	Grass/shrub	Cropland	Wetland	Total	Agreement %
Interpreter 137 - Set 10	Water	151		1					13	165	92
	Developed		262	5						267	98
	Disturbed		11	14		12	3	2	1	43	33
	Barren						1			1	0
	Tree cover			28		1706	8		97	1839	93
	Grass/shrub		68	17	65	156	2204	96	94	2700	82
	Cropland			2			32	328		362	91
	Wetland			1		33			64	98	65
	Total	151	341	68	65	1907	2248	426	269	5475	
										4729	Agree
	Agreement %	100	77	21	0	89	98	77	24	86.4	**Overall agreement %**

Table 3
Overall and per-class agreement between interpreters for the random subsample of pixels (CONUS summary).

Initial interpretations		Duplicate interpretations									
		Water	Developed	Disturbed	Barren	Tree cover	Grass/shrub	Cropland	Wetland	Total	Agreement %
	Water	4917	58	39	3	33	68		68	5186	95
	Developed	76	4105	60	6	299	225	64	18	4853	85
	Disturbed	6	65	422	1	193	101	38	72	898	47
	Barren	3	1	13	608	29	419	1		1074	57
	Tree cover		303	179		24335	1693	79	587	27176	90
	Grass/shrub	66	420	105	461	1444	31156	1381	106	35139	89
	Cropland		175	58	13	115	1152	17120	62	18695	92
	Wetland	165		54		632	337	71	3152	4411	71
	Total	**5233**	**5127**	**930**	**1092**	**27080**	**35151**	**18754**	**4065**	**97432**	
										85815	Agree
	Agreement %	94	80	45	56	90	89	91	78	**88.1%**	**Overall agreement**

Table 4
Interpreter agreement (%) by LCMAP class and region (Fig. 1) estimated from the randomly selected subsample. Values shown are the average of the row and column agreement for data organized as in Table 3.

Agreement (%)					
	East	East Central	West Central	West	CONUS
Water	98	95	87	91	94
Developed	83	87	75	82	82
Disturbed	49	38	35	55	46
Barren	96		0	56	56
Tree cover	92	89	80	90	90
Grass/shrub	68	63	91	91	89
Cropland	83	93	93	91	91
Wetland	77	76	65	72	74
Overall	86	87	89	89	88

Interpreter confusion most often occurred between classes that are known to frequently be difficult to distinguish or ambiguous (Table 5). For example, 79% of the disagreement involving Cropland was with Grass/Shrub, and 32% of the disagreement involving Grass/Shrub was with Cropland (Table 5). For Wetland, the most common disagreement was with Tree Cover (56%), and for Developed the two most common classes of disagreement were Grass/Shrub (36%) and Tree Cover (34%). Barren class disagreement was almost exclusively with Grass/Shrub and found in the arid and semi-arid areas of the West reporting region.

4.2. Agreement over temporal sets of interpretation

The QA/QC process provided feedback to interpreters with the intention of improving interpretation agreement over time. As described in Section 3.3.1, the randomized list of sample pixels was divided into sets of pixels for interpretation, so Set 1 through Set 12 represent the ordering in time that interpretations were completed. Overall agreement between interpreters varied from 86% (Set 1) to 91% (Set 8) but did not show a strong trend over time based on the 12 sets completed (Table 6). Per-class agreement was also generally consistent as most classes showed minimal or no trend in agreement over the 12 sets (Table 6). Some of the observed fluctuation in agreement is attributable to small sample size per set, particularly for the rare classes. Much of the feedback provided to interpreters focused on consistent recording of the change processes that crosswalk to the LCMAP Disturbed class. The positive trend in agreement for Disturbed (Table 6) indicates this feedback was beneficial. For the correlations (r) reported in Table 6, the *p*-values for testing if the correlation is 0 versus an alternative hypothesis that the correlation is not equal to 0 are greater than 0.15 for all classes except Disturbed ($p = 0.07$) and Barren ($p = 0.04$).

4.3. Temporal variation in agreement through the data time series

To evaluate whether generally increasing availability of data (Landsat, Google Earth imagery and ancillary data) through the

Table 5
Disagreement (%) distribution by class expressed as the percent of cases of disagreement in which one interpreter assigned the column-heading class and the other interpreter assigned the row-heading class (randomly selected duplicate pixels).

Class	Water	Developed	Disturbed	Barren	Tree cover	Grass/shrub	Cropland	Wetland
Water	*	8	5	1	1	2	0	11
Developed	23	*	13	1	11	8	7	1
Disturbed	8	7	*	1	7	3	3	6
Barren	1	0	1	*	1	11	0	0
Tree cover	6	34	38	3	*	39	6	56
Grass/shrub	23	36	21	93	56	*	79	20
Cropland	0	14	10	1	3	32	*	6
Wetland	40	1	13	0	22	6	4	*
Total	100	100	100	100	100	100	100	100

Table 6
Agreement (%) between interpreters for each LCMAP class per processing set for the subsample of randomly selected pixels (Set 1 and Set 12 are the first and last sets of pixels processed). The correlation (r) between agreement and set number is also provided.

	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8	Set 9	Set 10	Set 11	Set 12	r
Water	97	88	94	95	97	89	94	97	90	96	95	99	0.27
Developed	89	83	85	77	88	82	76	80	74	83	85	78	−0.41
Disturbed	34	43	35	49	50	41	57	57	51	49	44	51	0.55
Barren	66	75	43	66	56	55	0	0	0	0	75	0	−0.60
Tree cover	87	92	89	90	88	91	91	91	89	90	90	87	−0.05
Grass/shrub	85	89	89	86	90	87	90	94	88	90	86	86	0.14
Cropland	93	86	93	90	89	93	92	91	93	93	91	92	0.37
Wetland	62	85	77	76	53	73	80	79	80	64	74	85	0.27
Overall	85	88	88	87	88	88	89	91	88	89	88	87	0.34

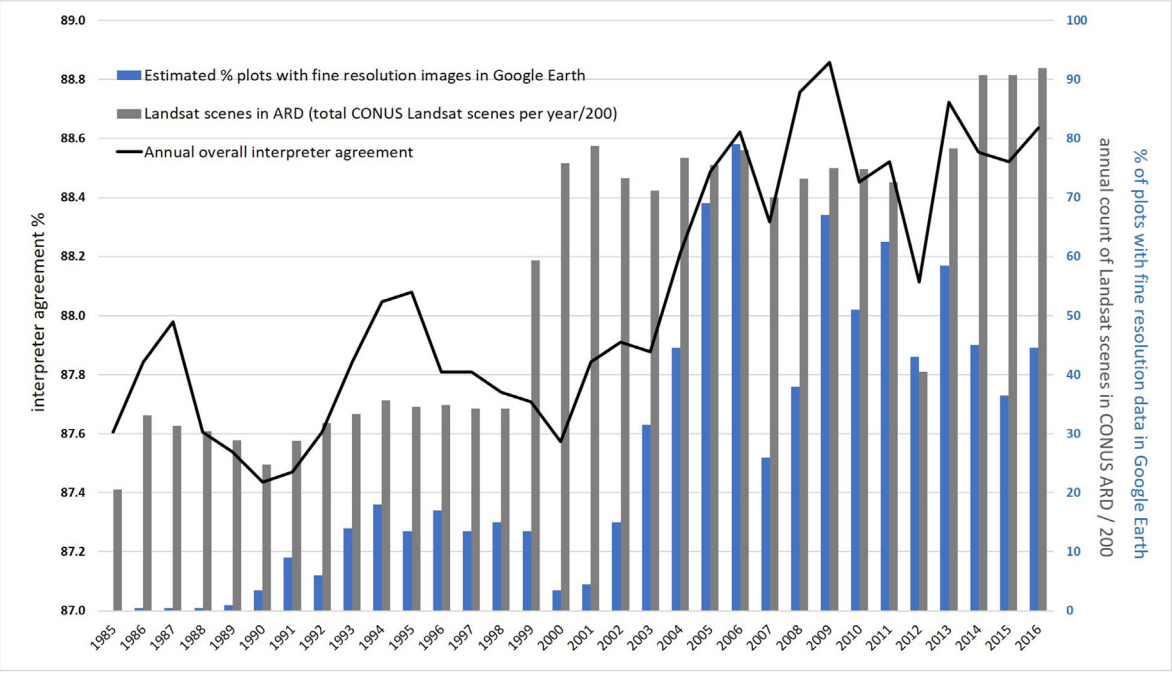

Fig. 5. Overall interpreter agreement (%) and Landsat and Google Earth fine resolution imagery availability 1985–2016.

33 years of the time series would increase interpreter agreement we graphed overall interpreter agreement per data acquisition year with Landsat and Google Earth data density. Fine resolution image availability was determined for each year of the time series at 200 randomly selected pixels in Google Earth and used as an estimate of fine resolution image coverage. The relative density of all available Landsat scenes per year for CONUS was computed as the number of scenes represented in the ARD data divided by 200 (for scale). Although an increase in overall agreement was detectable from 2001 to 2006, coinciding with increased availability of Landsat data and an increase in aerial photo coverage in Google Earth for the same period (Fig. 5), the improvement was only about 1%. Interpreter agreement for Barren, Water, Developed and Wetland classes was essentially flat. Increasing agreement was seen in Cropland, Grass/Shrub, Tree Cover and Disturbed classes (Table S3).

5. Discussion

The response design protocol developed for the LCMAP and LCMS projects has been demonstrated to be a viable operational methodology for obtaining reference data for a national, long-term land cover monitoring program. The information obtained from the probability subsample of duplicate interpretations provides quantitative estimates of agreement as well as critical information to further train interpreters to enhance consistency and accuracy of the reference class labels. While variability in reference class labeling has long been recognized, our study quantifies the level of consistency achievable by well-trained interpreters provided with the same resources of imagery and class definitions. Overall agreement between pairs of analysts interpreting randomly selected pixels was 88%. Interpreter agreement varied by class (46% to 94%), with lower agreement observed for Disturbed (46%), Barren (56%) and Wetland (74%) as these rarer classes are very challenging to identify consistently. However, the LCMAP QA/QC process includes review by senior interpreters and correction of obvious interpretation and data entry errors. Therefore, the final reviewed and edited reference data are expected to be more consistent than the initial baseline level of agreement estimated from the independent pairs of interpreters. Nevertheless, agreement results obtained in our study highlight the critical importance of training interpreters and implementing quality control procedures to monitor interpreters and provide feedback to improve consistency.

When multiple interpreters are used but the reference class labels disagree, a decision must be made on how the reference classification will be recorded. A common strategy is to review all cases in which interpretations disagree and provide a single ‘final’ interpretation that resolves the disagreement (e.g., by consensus). For a large dataset such as that of LCMAP, comprehensive expert or consensus review of all interpretations would be impractical and compete for resources that are needed to interpret a larger reference sample. In addition, because the

reference labels represent a long time series, any effort to resolve disagreements must be done in a manner that does not create illogical sequences of land change in the data. Currently LCMAP uses the information from the duplicated interpretations to prioritize pixels for detailed review by three senior interpreters. Duplicated interpretations that fail any of several tests of agreement are reviewed by senior interpreters and the senior interpreter chooses the better of the duplicated interpretations or makes other edits to determine the final interpretation. All sample pixels undergo evaluation to identify obvious interpretation and data entry errors (e.g., illogical sequences of land cover) and such errors are corrected within TimeSync.

The consistency of agreement across the 12 sets of interpretations completed to date (Section 4.2 and Table 6) is an encouraging testament to the success of interpreter training and monitoring. However, there are cautionary lessons to be learned as well. The complexity of creating and executing an interpretation protocol in TimeSync to accommodate multiple applications such as LCMAP and LCMS, compounded by the difficulty of training and maintaining consistency across a changing cast of interpreters was very challenging. The necessity of meeting the data requirements for two different projects with different class definitions and different priorities has very likely had at least some negative impact on data quality. Based on our experience with this work, the tradeoffs between the potential efficiency of jointly collecting reference data for multiple applications versus the much simpler interpretation protocol possible where data are collected for a single set of class definitions and data requirements need to be carefully considered.

6. Conclusions

A lack of consistent, high-quality reference data is a challenge faced by many land resource scientists and managers when seeking to evaluate land cover and map products. While interpreter agreement has long been recognized as a potential issue, most studies of agreement have been limited to relatively small spatial and temporal extents. Typically, the response designs for land cover monitoring studies covering large spatial extents (e.g., national or continental) over a long time series (e.g., 30 years) have not included a rigorous probability sampling protocol for quantifying interpreter agreement and using this agreement information in ongoing training of interpreters to enhance the quality of the reference data. The response design developed for LCMAP includes these features and has been successfully implemented in an operational, national land cover monitoring framework.

The primary conclusions based on the results from obtaining duplicate interpretations of nearly 3000 randomly selected reference sample pixels are the following: 1) a well-trained and calibrated team of interpreters can achieve overall agreement among pairs of interpreters of 88% with more prevalent classes such as Tree Cover, Grass/Shrub, Water and Cropland having greater agreement (89% to 94%) than rare classes such as Disturbed, Barren and Wetland (46% to 74%); 2) overall and per-class interpreter agreement varied regionally; 3) the interpreter training and QA/QC protocols maintained consistent agreement over the time period that pixels for the 12 sets were interpreted with some improvement in consistency of the problematic Disturbed class; and 4) overall interpreter agreement increased, albeit by only slightly more than 1%, for years after 2000, when more and better quality imagery became available. Because most large-area reference datasets are obtained by teams of interpreters, the results of this study provide conclusive evidence for including protocols in the response design to estimate and monitor interpreter consistency, and that providing feedback to interpreters can improve accuracy and consistency of the reference data. Efforts to improve and estimate reference data quality will continue to be an ongoing initiative in LCMAP.

Declaration of Competing Interest

No potential conflict of interest was reported by the authors.

Acknowledgements

This work was supported with funding from the U.S. Geological Survey's Land Resources Mission Area, in part under USGS Contract G15PC00012. This research was also funded by the U.S. Forest Service (IRDB-LCMSNWFP) Landscape Change Monitoring System, the U.S. Forest Service Information Resources Direction Board and the Northwest Forest Plan. We thank Alexander Hernandez of Utah State University (USU) for TimeSync technical assistance and supporting the USU student interpreters. Thanks also to all the other interpreters who worked on this data collection: Roger Auch, Janis Taylor, Steve Kambly, Alexandra Hernandez, Ellie Leydsman, Mikala Solos, Tommy Thompson, Dalton Newbold and Helen Zhong. Thanks to the U.S. Forest Service Geospatial Technology and Applications Center and Mark Finco, Jennifer Lecker and Chris Gay for their help. We also thank Christopher Barber at EROS Data Center and three anonymous reviewers for their detailed and helpful reviews of this manuscript. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.rse.2019.111261.

References

Bicheron, P., Defourny, P., Brockmann, C., Schouten, L., Vancutsem, C., Huc, M., ... Arino, O., 2008. GLOBCOVER: Products Description and Validation Report.

Clark, M.L., Aide, T.M., Riner, G., 2012. Land change for all municipalities in Latin America and the Caribbean assessed from 250-m MODIS imagery (2001 – 2010). Remote Sens. Environ. 126, 84–103. https://doi.org/10.1016/j.rse.2012.08.013.

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 (12), 2911–2924. https://doi.org/10.1016/j.rse.2010.07.010.

Congalton, R.G., 1991. A review of assessing the accuracy of classifications of remotely sensed data. Remote Sens. Environ. 37, 35–46. https://doi.org/10.1016/0034-4257(91)90048-B.

Dwyer, J., Roy, D., Sauer, B., Jenkerson, C., Zhang, H., Lymburner, L., 2018. Analysis ready data: enabling analysis of the Landsat archive. Remote Sens. (9), 10. https://doi.org/10.3390/rs10091363.

Feng, M., Sexton, J.O., Huang, C., Anand, A., Channan, S., Song, X.-P., Townshend, J.R., 2016. Earth science data records of global forest cover and change: assessment of accuracy in 1990, 2000, and 2005 epochs. Remote Sens. Environ. 184, 73–85. https://doi.org/10.1016/j.rse.2016.06.012.

Foley, J.A., Defries, R., Asner, G.P., Barford, C., Bonan, G., Carpenter, S.R., ... Snyder, P.K., 2005. Global consequences of land use. Science 309 (5734), 570–574. https://doi.org/10.1126/science.1111772.

Foody, G.M., 2010. Assessing the accuracy of land cover change with imperfect ground reference data. Remote Sens. Environ. 114 (10), 2271–2285. https://doi.org/10.1016/j.rse.2010.05.003.

Foody, G.M., 2013. Ground reference data error and the mis-estimation of the area of land cover change as a function of its abundance. Remote Sens. Lett. 4, 783–792. https://doi.org/10.1080/2150704X.2013.798808.

Foody, G.M., 2015. Valuing map validation: the need for rigorous land cover map accuracy assessment in economic valuations of ecosystem services. Ecol. Econ. 111, 23–28. https://doi.org/10.1016/j.ecolecon.2015.01.003.

Gong, P., Wang, J., Yu, L., Zhao, Y., Zhao, Y., et al., 2013. Finer resolution observation and monitoring of global land cover: first mapping results with Landsat TM and ETM + data. Int. J. Remote Sens. 34, 2607–2654. https://doi.org/10.1080/01431161.2012.748992.

Hansen, M.C., Potapov, P.V., Moore, R., Hancher, M., Turubanova, S.A., Tyukavina, A., ... Townshend, J.R., 2013. High-resolution global maps of 21st-century forest cover change. Science 342 (6160), 850–853. https://doi.org/10.1126/science.1244693.

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

Hermosilla, T., Wulder, M.A., White, J.C., Coops, N.C., Hobart, G.W., 2018. Disturbance-informed annual land cover classification maps of Canada's forested ecosystems for a 29-year Landsat time series. Can. J. Remote. Sens. 44 (1), 67–87. https://doi.org/10.

1080/07038992.2018.1437719.
Hoekstra, J.M., Boucher, T.M., Ricketts, T.H., Roberts, C., 2004. Confronting a biome crisis: global disparities of habitat loss and protection. Ecol. Lett. 8 (1), 23–29. https://doi.org/10.1111/j.1461-0248.2004.00686.x.
Homer, C.H., Fry, J.A., Barnes, C.A., 2012. The national land cover database. In: U.S. Geological Survey Fact Sheet 2012-3020.
Mann, S., Rothley, K.D., 2006. Sensitivity of Landsat/IKONOS accuracy comparison to errors in photointerpreted reference data and variations in test point sets. Int. J. Remote Sens. 27 (22), 5027–5036. https://doi.org/10.1080/01431160600784291.
McRoberts, R.E., Stehman, S.V., Liknes, G.C., Næsset, E., Sannier, C., Walters, B.F., 2018. The effects of imperfect reference data on remote sensing-assisted estimators of land cover class proportions. ISPRS J. Photogramm. Remote Sens. 142, 292–300. https://doi.org/10.1016/j.isprsjprs.2018.06.002.
Monitoring Trends in Burn Severity (MTBS), 2018. Monitoring Trends in Burn Severity (MTBS). Retrieved from. https://www.mtbs.gov/.
Olofsson, P., Foody, G.M., Stehman, S.V., Woodcock, C.E., 2013. Making better use of accuracy data in land change studies: estimating accuracy and area and quantifying uncertainty using stratified estimation. Remote Sens. Environ. 129, 122–131. https://doi.org/10.1016/j.rse.2012.10.031.
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. https://doi.org/10.1016/j.rse.2014.02.015.
Omernik, J.M., Griffith, G.E., 2014. Ecoregions of the conterminous United States: evolution of a hierarchical spatial framework. Environ. Manag. 54 (6), 1249–1266. https://doi.org/10.1007/s00267-014-0364-1.
Overton, W.S., Stehman, S.V., 1996. Desirable design characteristics for long-term monitoring of ecological variables. Environ. Ecol. Stat. 3 (4), 349–361. https://doi.org/10.1007/bf00539371.
Powell, R.L., Matzke, N., de Souza, C., Clark, M., Numata, I., Hess, L.L., ... Roberts, D.A., 2004. Sources of error in accuracy assessment of thematic land-cover maps in the Brazilian Amazon. Remote Sens. Environ. 90 (2), 221–234. https://doi.org/10.1016/j.rse.2003.12.007.
Prestele, R., Arneth, A., Bondeau, A., de Noblet-Ducoudre, N., Pugh, T.A.M., Sitch, S., ... Verburg, P.H., 2017. Current challenges of implementing anthropogenic land-use and land-cover change in models contributing to climate change assessments. Earth Syst. Dyn. 8 (2), 369–386. https://doi.org/10.5194/esd-8-369-2017.
Scepan, J., Menz, G., Hansen, M.C., 1999. The DISCover validation image interpretation process. Photogramm. Eng. Remote. Sens. 65 (9), 1075–1081.
Sexton, J.O., Urban, D.L., Donohue, M.J., Song, C., 2013. Long-term land cover dynamics by multi-temporal classification across the Landsat-5 record. Remote Sens. Environ. 128, 246–258. https://doi.org/10.1016/j.rse.2012.10.010.
Sleeter, B.M., Sohl, T.L., Loveland, T.R., Auch, R.F., Acevedo, W., Drummond, M.A., ... Stehman, S.V., 2013. Land-cover change in the conterminous United States from 1973 to 2000. Glob. Environ. Chang. 23 (4), 733–748. https://doi.org/10.1016/j.gloenvcha.2013.03.006.
Tsendbazar, N.E., Herold, M., de Bruin, S., Lesiv, M., Fritz, S., Van De Kerchove, R., ... Pekel, J.F., 2018. Developing and applying a multi-purpose land cover validation dataset for Africa. Remote Sens. Environ. 219, 298–309. https://doi.org/10.1016/j.rse.2018.10.025.
Turner 2nd, B.L., Lambin, E.F., Reenberg, A., 2007. The emergence of land change science for global environmental change and sustainability. Proc. Natl. Acad. Sci. U. S. A. 104 (52), 20666–20671. https://doi.org/10.1073/pnas.0704119104.
U.S. Fish and Wildlife Service, 2018. National Wetlands Inventory. Retrieved from. https://www.fws.gov/wetlands/.
USGS, 2018. EarthExplorer. Retrieved from. https://earthexplorer.usgs.gov/.
Wickham, J.D., Stehman, S.V., Gass, L., Dewitz, J., Fry, J.A., Wade, T.G., 2013. Accuracy assessment of NLCD 2006 land cover and impervious surface. Remote Sens. Environ. 130, 294–304. https://doi.org/10.1016/j.rse.2012.12.001.
Wickham, J.D., Stehman, S.V., Gass, L., Dewitz, J.A., Sorenson, D.G., Granneman, B.J., ... Baer, L.A., 2017. Thematic accuracy assessment of the 2011 National Land Cover Database (NLCD). Remote Sens. Environ. 191, 328–341. https://doi.org/10.1016/j.rse.2016.12.026.
Zhu, Z., Woodcock, C.E., 2012. Object-based cloud and cloud shadow detection in Landsat imagery. Remote Sens. Environ. 118, 83–94. https://doi.org/10.1016/j.rse.2011.10.028.
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. https://doi.org/10.1016/j.rse.2014.01.011.
Zhu, Z., Yang, L., Stehman, S.V., Czaplewski, R.L., 2000. Accuracy assessment for the U. S. Geological Survey regional land-cover mapping. Photogramm. Eng. Remote. Sens. (66), 1425–1435 (doi:10.1.1.476.517).