Remote Sensing of Environment 257 (2021) 112357

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

Thematic accuracy assessment of the NLCD 2016 land cover for the conterminous United States

James Wickham [a,*], Stephen V. Stehman [b], Daniel G. Sorenson [c], Leila Gass [d], Jon A. Dewitz [e]

[a] Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, NC 27111, USA
[b] College of Environmental Science and Forestry, State University of New York, Syracuse, NY 13210, USA
[c] U.S. Geological Survey, 909 1st Ave., Seattle, WA 98104, USA
[d] U.S. Geological Survey, 520 N. Park Ave., Tucson, AZ 85179, USA
[e] Earth Resources and Observation Science (EROS) Center, U.S. Geological Survey, 47914 252nd St., Sioux Falls, SD 57198, USA

ARTICLE INFO

Editor: Dr Marie Weiss

Keywords:
Accuracy assessment
Cognitive psychology
Forest disturbance
Land cover
Land cover change
NLCD
Reference data quality
Stratified sampling
Urbanization

ABSTRACT

The National Land Cover Database (NLCD) is an operational land cover monitoring program providing updated land cover and related information for the United States at five-year intervals. NLCD2016 extends temporal coverage to 15 years (2001–2016). We collected land cover reference data for the 2011 and 2016 nominal dates to report land cover accuracy for the NLCD2016 database 2011 and 2016 land cover components at Level II and Level I and for Level I 2011–2016 land cover change using two definitions of agreement. For both the 2011 and 2016 land cover components, single-date Level II overall accuracies (OA) were 72% (standard error of ±0.9%) when agreement was defined as match between the map label and primary reference label only and 86% (± 0.7%) when agreement also included the alternate reference label. The corresponding level I OA for both dates were 79% (± 0.9%) and 91% (± 1.0%). The 2011–2016 user's and producer's accuracies (UA and PA) were ~ 75% for forest loss and PA for water loss, grassland loss, and grass gain were > 70% when agreement included a match between the map label and either the primary or alternate reference label. Depending on agreement definition and level of the classification hierarchy, OA for the 2011 land cover component of the NLCD2016 database was about 4% to 7% higher than OA for the 2011 land cover component of the NLCD2011 database, suggesting that the changes in mapping methodologies initiated for production of the NLCD2016 database have led to improved product quality. Additionally, we used the reference dataset collected for assessment of the NLCD2011 database to assess the 2001 and 2006 land cover components of the NLCD2016 database. OA for the 2001 and 2006 land cover components was 1% - 5% lower than OA for the 2011 and 2016 land cover components of the NLCD2016 database. Higher OA for 2011 and 2016 land cover components of the NLCD2016 database relative to OA for its 2001 and 2006 components may be attributable to differences in reference data quality.

1. Introduction

An overarching objective of the NLCD mapping project is to be a long-term land cover monitoring program for the United States (Yang et al., 2018). The release of NLCD2011 (Homer et al., 2015) was the first realization of that objective because it included land cover and impervious cover for three points in time (2001, 2006, and 2011), providing Landsat-based, digitally produced, continental-scale data suitable for examining trends for several land cover themes (e.g., forest, urban, and impervious cover). The release of NLCD 2016 (https://www.mrlc.gov) extends the NLCD land cover time series from 10 years to 15 years (2001 — 2016) (Homer et al., 2020).

NLCD is an operational program. Operational programs tend to have stricter standards than, for example, experimental programs because they are defined by exogenous factors such as administrative requirements, production deadlines, user community needs, and product standards (Hansen and Loveland, 2012). To meet user community needs and with an eye toward improving product quality, NLCD has always invested substantially in production-oriented research to support its long-term monitoring objective (e.g., Homer et al., 2004; Jin et al., 2013). Methodological improvement and refinement are standard operating procedures in the years immediately following an NLCD

* Corresponding author at: 109 TW Alexander Dr. MD: 343-05, Research Triangle Park, NC 27711, USA.
E-mail addresses: wickham.james@epa.gov (J. Wickham), dsorenson@usgs.gov (D.G. Sorenson), lgass@usgs.gov (L. Gass), dewitz@usgs.gov (J.A. Dewitz).

https://doi.org/10.1016/j.rse.2021.112357
Received 28 September 2020; Received in revised form 22 January 2021; Accepted 12 February 2021
Available online 20 February 2021
0034-4257/Published by Elsevier Inc.

release, and accuracy assessments for each NLCD release suggest that methodological improvements have led to enhanced product quality (Table 1). NLCD accuracy assessments follow release of NLCD products and the results of these assessments are incorporated in the planning and production of subsequent releases.

NLCD2016 mapping methods were a substantial expansion of those used to produce prior NLCD databases. NLCD2016 was based on multi-temporal modeling of spectral and geographic data (Homer et al., 2020; Jin et al., 2019; Yang et al., 2018), whereas previous NLCD products were based primarily on spectral classification and identification of change (Jin et al., 2013). An in-depth description of the NLCD2016 procedures is found in Yang et al. (2018). Briefly, NLCD production used automated procedures to select suitable Landsat images at 2–3-year intervals between 2001 and 2016, model land cover and land-cover change with pixel- and object-based methods, and then post-process the classification results for spatial and temporal coherence of land cover labels. Accuracy assessment, like classification methods research, has been a consistent aspect of NLCD production. As noted in Table 1, statistically rigorous accuracy assessments (Stehman, 2001; Stehman et al., 2003), have accompanied each NLCD release (Homer et al., 2007, 2015; Fry et al., 2011; Wickham et al., 2010, 2013, 2017).

Our overall objective is to report on the thematic accuracy of NLCD2016 land cover. Specifically, we focused on land cover accuracy at Level II and Level I of the NLCD classification hierarchy and Level I land cover change. Because each new release of the NLCD database includes new versions of previously produced land cover data, we use the semantical construction "NLCDxxxx yyyy," where xxxx refers to the year identifying the database and "yyyy" refers to the nominal year of the land cover component. For example, NLCD2016 2011 would refer to the land cover component for the nominal year 2011 in the NLCD2016 database, whereas NLCD2011 2011 would refer to the land cover component for the nominal year 2011 in the NLCD2011 database. The NLCD 2016 database now includes seven dates of land cover (2001, 2003, 2006, 2008, 2011, 2013, and 2016). One reason for the inclusion of the "interstitial" dates (2003, 2008, and 2013) was to identify land cover trajectories more precisely, which in turn were used to inform the classification of the individual dates (Yang et al., 2018).

2. Methods

The accuracy assessment methodology for NLCD2016 was nearly identical to that used for NLCD2011 (Wickham et al., 2017). The main differences were: 1) the NLCD2016 accuracy assessment did not include an east-west geographic stratification, and; 2) the number of sample pixels was reduced from 8000 for the NLCD2011 accuracy assessment to 4629. The smaller sample size, due to limited resources for reference data collection, affects precision of the accuracy estimators (i.e., larger standard errors) but does not change the feature that the estimators are unbiased (Stehman, 2001). For the sample data collected in association with the NLCD2016 assessment, reference labels were obtained for the 2011 and 2016 components of the NLCD2016 database. To estimate accuracy of the 2001 and 2006 land cover components of the NLCD2016 database, we used the reference sample data collected for the accuracy assessment of the NLCD2011 database (Wickham et al., 2017). Use of these previously collected reference data allowed us to report accuracies for the four main years of the NLCD2016 database land cover (2001, 2006, 2011, 2016).

Table 1
Conterminous United States overall accuracies (OA; %) and (standard errors) by NLCD product year. Agreement is based on a match between the map and primary or alternate reference labels. NLCD accuracy assessments are reported in Wickham et al. (2010, 2013, 2017). Accuracy trends for land cover components (Year of Land Cover) are reported column-wise.

NLCD Product	Year of Land Cover			
Year	2001	2006	2011	2016
Level II				
2001[a]	78.8 (2.1)			
2006	79.0 (0.8)	78.0 (0.8)		
2011	83.2 (0.5)	82.8 (0.5)	82.0 (0.5)	
2016	83.7 (0.5)	83.6 (0.5)	86.8 (0.7)	86.4 (0.6)
Level I				
2001[a]	80.4 (1.9)			
2006	85.0 (0.4)	84.0 (0.7)		
2011	89.3 (0.4)	89.0 (0.4)	88.0 (0.4)	
2016	89.2 (0.5)	89.2 (0.5)	90.5 (0.6)	90.6 (0.6)

[a] OA and (SE) based on an unweighted average of regional values.

2.1. Sampling design

The reference sample data collected for the accuracy assessment of the NLCD2016 database were obtained from a stratified random sample of pixels with the strata defined using the NLCD2016 database. The 21 strata consisted of 15 no-change strata, derived from 16 NLCD Level II land cover classes (Table 2), and 6 change strata created from the NLCD2016 2011 — 2016 change map (Table 3). As was true of previous NLCD accuracy assessments, choosing stratified sampling and defining the strata were motivated by the objective of precise estimation of user's accuracy for key classes, specifically Level II land cover and the five land cover change themes identified as change strata. Stratified sampling assured a sufficient sample size to precisely estimate user's accuracy for these targeted classes even if the classes were rare. We did not include the NLCD class perennial ice/snow as a no change stratum because it comprised less than 0.0006% of the population (i.e., map area). The six land cover change strata created from the NLCD2016 2011–2016 change map were urban gain, forest loss, forest gain, agriculture loss,

Table 2
National Land Cover Database (NLCD) land cover legend for Level II and Level I of the classification hierarchy and class (codes). See https://www.mrlc.gov/data/legends/national-land-cover-database-2016-nlcd2016-legend for a complete description of NLCD classes.

Class (code)	Description
Level II	
Water (11)	Open water with generally <25% vegetation or soil cover
Ice/snow (12)[a]	>25% permanent ice or snow
Urban, open space (21)	Dominated by vegetation; impervious cover (IC) ≤ 20%
Urban, low intensity (22)	Vegetation, 20% < IC ≤ 49%
Urban, med. Intensity (23)	Vegetation, 50% < IC ≤ 79%
Urban, high intensity (24)	Vegetation, IC ≥ 80%
Barren (31)	Vegetation <15% cover (e.g., bedrock, desert pavement)
Deciduous forest (41)	>20% trees; >75% of trees shed foliage seasonally
Evergreen forest (42)	>20% trees; >75% of trees maintain foliage year-round
Mixed forest (43)	>20% trees; deciduous and evergreen each <75% cover
Shrubland (52)	>20% woody vegetation <5 m
Grassland (71)	>80% herbaceous cover that is not managed (e.g., tilling)
Pasture (81)	>20% herbaceous cover for livestock, seed, or hay
Cropland (82)	>20% herbaceous or woody cover cultivated for human consumption or use
Woody wetlands (90)	>20% woody cover on saturated soil
Herbaceous wetlands (95)	>80% herbaceous cover on saturated soil
Level I	
Water (10)	Classes 11 and 12
Urban (20)	Classes 21–24
Barren (30)	Class 31
Forest (40)	Classes 41–43
Shrubland (50)	Class 52
Grassland (70)	Class 71
Agriculture (80)	Classes 81 and 82
Wetland (90)	Classes 90 and 95

[a] Not included in accuracy assessment of the NLCD2016 database.

Table 3
Accuracy assessment strata for the sampling design of the NLCD2016 reference database. All change strata are resolved at Level I of the classification hierarchy (see Table 2). The population comprises an area of 7,792,667 km^2 (8,658,519,089 pixels).

Stratum	# pixels	Description
1	144,260,856	Water, no change
2	255,518,661	Open urban, no change
3	130,760,739	Low intensity urban, no change
4	59,945,943	Medium intensity urban, no change
5	21,576,459	High intensity urban, no change
6	88,893,116	Barren, no change
7	825,184,784	Deciduous forest, no change
8	979,538,237	Evergreen forest, no change
9	316,640,402	Mixed forest, no change
10	1,864,314,143	Shrubland, no change
11	1,159,802,223	Grassland, no change
12	559,677,646	Pasture, no change
13	1,429,403,249	Cropland, no change
14	380,971,449	Woody wetlands, no change
15	117,993,354	Herbaceous wetlands, no change
16	146,679,429	Catch-all stratum; all 2011—2016 change pixels not in strata 17—21
17	6,772,002	Urban gain; not urban (2011) to urban (2016)
18	73,557,194	Forest loss; forest (2011) to not forest (2016) except forest to urban
19	70,038,668	Forest gain; not forest (2011) to forest (2016)
20	4,483,066	Agriculture (Ag) loss; Ag (2011) to not Ag (2016) except Ag to forest or urban
21	22,289,822	Ag gain; not Ag (2011) to Ag (2016), except for forest to Ag

agriculture gain, and all other ("catch-all") change. Because strata must be mutually exclusive sets of pixels and loss from one category is gain in another, we created change strata in the sequence listed as follows (Table 3):

1) Urban gain: not labeled as urban in 2011 and labeled as urban in 2016;
2) Forest loss: labeled as forest in 2011 and not labeled as forest in 2016, except for forest to urban (already included in the urban gain stratum);
3) Forest gain: not labeled as forest in 2011 and labeled as forest in 2016;
4) Agriculture loss: labeled as agriculture in NLCD2016 2011 and not labeled as agriculture in 2016, except for agriculture to forest (already included in the forest gain stratum) and agriculture to urban (already included in the urban gain stratum);
5) Agriculture gain; not labeled as agriculture in NLCD2016 2011 and labeled as agriculture in 2016, except for forest to agriculture (already included in the forest loss stratum);
6) Catch-all: all 2011–2016 change pixels not in the 5 strata defined above.

Sample pixels selected from the catch-all stratum were dominated by shrubland and grassland changes because of their geographically extensive coverage in the western United States. The six change strata were consistent with those used in the accuracy assessment of the NLCD2011 database (Wickham et al., 2017).

The population for reference data collection was based on the dual map labels resultant from combining NLCD2016 2016 and NLCD2016 2011 in a GIS and classifying each dual-label combination into one of the 21 strata (i.e., pixel counts in Table 3). To sample individual pixels from each stratum for reference data collection, we created strata-specific maps, assigned a random number (integer with no duplicate numbers) to each pixel in the stratum, and then selected the pixels with random numbers 1 through n where n equaled the number of sample pixels pre-assigned for each stratum. The population used for accuracy assessment was slightly different from the NLCD2016 land cover data itself. In addition to exclusion of perennial snow/ice, the NLCD2016 accuracy assessment (as well as all previous NLCD accuracy assessments) did not include coastal water or the Great Lakes because their inclusion would result in nearly all water sample pixels being located in these large water features.

2.2. Response design

Reference data collection was accomplished by two interpreters with experience in the production of all previous NLCD products (NLCD2001, NLCD2006, and NLCD2011) and all previous NLCD accuracy assessments. The rules (response design protocols) established to guide reference label assignment included: 1) reference label assignment without knowledge of the map classification (blind interpretation); 2) use of Google earth™ online imagery as the source for reference label interpretation, satisfying the criterion that the reference medium be of higher quality than the mapping medium (Olofsson et al., 2014; Stehman and Foody, 2019); 3) use of a pixel as the spatial support unit (Stehman and Wickham, 2011), and; 4) assignment of primary and alternate reference labels to promote consistency in reference label assignment (Mann and Rothley, 2006). Reference label consistency across interpreters was further supported by weekly, web-enabled conference calls and by requiring the two interpreters to initiate data collection by jointly determining the reference class labels for 50 sample pixels to establish consistency. These meetings were used to collaboratively evaluate reference labels for a selection of sample pixels and included an additional person familiar with the NLCD project who consulted on and contributed to reference label assignment for the sample pixels discussed on the conference calls. Further information on response design protocols and reference data is provided in appendix A (Supplementary data).

Assignment of the most appropriate reference labels to the *ground condition* of the sample pixel (response design element 3) is the main objective of reference label assignment (Stehman and Wickham, 2011). To support reference label assignment (response design protocol 1), the interpreters were provided Keyhole Markup language Zipped (KMZ) files of the sample pixel locations for display in Google Earth™ (response design protocol 2). The KMZ sample pixel locations were represented as a point (center of pixel) and a 3-x-3 pixel window surrounding the point. The 3-x-3 pixel window was included to facilitate the use of landscape context for reference label assignment (Stehman and Czaplewski, 1998). For example, the presence of a road in the 3-x-3 pixel window, can be used to inform whether an alternate label of urban would be appropriate for a sample pixel located in a forested area.

Assignment of primary and alternate reference labels (response design protocol 4) has been implemented for all NLCD accuracy assessments (Stehman et al., 2003; Wickham et al., 2010, 2013, 2017). In addition to the inherent edginess in land cover maps (e.g., road in the 3-x-3 pixel window of a forested sample pixel), distinction between shrubland and forest, pasture and grassland, and many other land cover classes can be ambiguous. Primary and alternate labels have been used to account for the inherent ambiguity, geolocation error, and mixed pixels (Appendix A). Their use can be considered a special case of the linguistic scale, fuzzy membership analysis (Stehman et al., 2003, p. 513) reported by Gopal and Woodcock (1994). The use rate for alternate labels was 39% at level I and 62% at Level II of the classification hierarchy, which is consistent with the use rates reported for accuracy assessment of the NLCD2011 database (Wickham et al., 2017). Alternate label use rates are expected to be higher for Level II than Level I because of the greater similarity among Level II classes.

2.3. Analysis

General estimation theory of probability sampling was the foundation of the analysis component (cf. Särndal et al., 1992). The known inclusion probabilities from the stratified (by map class) random design

(Stehman, 2001; Stehman and Czaplewski, 1998) were incorporated into the sample-based estimates using the indicator variable formulation of the estimators (Stehman, 2014). Overall accuracy (OA) was estimated as:

$$\hat{\mathscr{O}} = \left(\frac{1}{N}\right)\sum_{h=1}^{H} N_h \hat{p}_h \tag{1}$$

where $\hat{p}_h$ is the sample proportion of correctly classified pixels in stratum h, N is the total number of sample pixels, N_h is the population size of stratum h, and the summation is over all H strata (H = 21). User's accuracy (UA) and producer's accuracy (PA) were estimated as a ratio $R = Y/X$, where Y is the population total of y_u, where

$$y_u = \left\{ \begin{array}{c} 1 \; if \; pixel \; u \; satisfies \; condition \; A \\ 0 \; if \; pixel \; u \; does \; not \; satisfy \; condition \; A \end{array} \right\} \tag{2}$$

and X is the population total of x_u, where

$$x_u = \left\{ \begin{array}{c} 1 \; if \; pixel \; u \; satisfies \; condition \; B \\ 0 \; if \; pixel \; u \; does \; not \; satify \; condition \; B \end{array} \right\} \tag{3}$$

To estimate user's and producer's accuracy of class K (K can be any land cover class or land-cover change class), condition A would be that the map class and reference class are both K (agreement). For user's accuracy, condition B would be that the map is class K, and for producer's accuracy, condition B would be that the reference class is K. The combined ratio estimator (Cochran, 1977, Section 6.11) for UA and PA is then

$$\hat{R} = \frac{\hat{Y}}{\hat{X}} = \frac{\sum_{h=1}^{H} N_h \bar{y}_h}{\sum_{h=1}^{H} N_h \bar{x}_h} \tag{4}$$

where $\bar{x}_h$ is the sample mean of x_u in stratum h (i.e., Table 4) and $\bar{y}_h$ is the sample mean of y_u in stratum h. The estimated variance of the combined ratio estimator is.

$$\hat{V}\left(\hat{R}\right) = \left(\frac{1}{\hat{X}^2}\right)\left[\sum_{h=1}^{H} N_h^2 \left(1 - n_h/N_h\right)\left(s_{yh}^2 + \hat{R}^2 \, s_{xh}^2 - 2\hat{R} s_{xyh}\right) \Big/ n_h\right] \tag{5}$$

where n_h is the sample size in stratum h, s_{yh}^2 and s_{xh}^2 are the sample variances for y_u and x_u for stratum h and s_{xyh} is the sample covariance for y_u and x_u for stratum h, and

$$\hat{X} = \sum\nolimits_{h=1}^{H} N_h \bar{x}_h \tag{6}$$

is the estimated population total of the x_u. The values for y_u and for $\bar{y}_h$ and s_{yh}^2 equal zero (0) for a stratum in which no sample pixels satisfy condition A (the condition defining the numerator of $\hat{R}$), and similarly the values of x_u, $\bar{x}_h$, and x_{yh}^2 equal zero (0) for a stratum in which no pixels satisfy condition B (the condition defining the denominator of $\hat{R}$). Estimates were computed using version 9.3 of SAS (Statistical Analysis Software, SAS, Inc., Cary, North Carolina, USA).

We report accuracy estimates for agreement based on the map label matching the primary reference label only and for the map label matching either the primary or alternate label. Agreement for change accuracy estimation includes two of the four possible combinations of reference labels: 1) 2011 primary and 2016 primary, and 2) 2011 alternate and 2016 alternate (see Appendix A for examples). The other two possible combinations (2011 primary and 2016 alternate; and 2011 alternate with 2016 primary) were not considered. All four possible combinations were considered in previous NLCD assessments (Wickham et al., 2013, 2017). As noted previously, 39% of the sample pixels had an alternate reference label at Level I, the level of the classification hierarchy used to assess land cover change.

Olofsson et al. (2014) recommended estimating area of land cover

Table 4
Agreement between map and reference labels for NLCD2016 2016 for the continental United States at Level II of the classification hierarchy. Class codes are in Table 2. Cell entries are percentages of map area (see Table 3) based on a match between the map label and the primary reference label only. Map Bias (MB) is the map area estimate (Total [row]) minus the reference classification area estimate (Total [column]). TSE is the standard error of the area estimate based on the reference classification. OA, UA, PA and their (standard errors) are the overall, user's and producer's accuracies and OAa, UAa and PAa are the corresponding versions of the estimates when agreement included the alternate reference label. OA = 72.1% (± 0.9%). OAa = 86.9% (± 0.6%).

	Reference																			
Map	11	21	22	23	24	31	41	42	43	52	71	81	82	90	95	Total	UA	UAa	MB	n
11	1.5870	0.0088				0.0003	0.0266				0.0088	0.0091	0.0097	0.0178	0.0620	1.7300	92 (2)	99 (1)	0.05	221
21	0.0119	2.1389	0.3708	0.0606	0.0240		0.1083	0.0951	0.0124	0.0606	0.0008	0.0606	0.0364			2.9802	72 (3)	88 (2)	−2.11	359
22		0.5132	0.6585	0.2629	0.0121		0.0309	0.0061	0.0003	0.0063	0.0063	0.0124	0.0185	0.0003	0.0003	1.5280	43 (3)	74 (3)	−0.02	316
23		0.0638	0.1575	0.3398	0.1360	0.0005	0.0130				0.0049		0.0046			0.7201	47 (4)	77 (3)	0.00	244
24	0.0003	0.0228	0.0217	0.0481	0.1671	0.0023					0.0010					0.2642	63 (3)	75 (3)	−0.08	283
31	0.0003	0.0435		0.0108		0.6007	0.0108	0.0216			0.1513	0.0114	0.0117		0.0003	1.0355	58 (5)	77 (4)	0.09	106
41		0.2043	0.0389				8.0140	0.2206	0.6841	0.3242	0.0098	0.2010		0.0389		9.7358	82 (2)	90 (2)	−2.17	308
42		0.2407	0.0033			0.0462	0.2043	9.9763	0.4878	0.5998	0.1912			0.0527		11.8023	85 (2)	93 (2)	−1.43	395
43	0.0256	0.1088	0.0256				0.9402	0.5368	1.9629	0.0865	0.0832	0.0098		0.0033		3.7826	52 (4)	75 (4)	0.19	180
52		0.5546	0.1090			0.1638	0.3703	1.6054	0.0236	15.4062	3.9059	0.3586	0.2218	0.0545	0.0003	22.7740	68 (2)	87 (2)	3.73	679
71		0.0982				0.0802	0.4517	0.2412	0.0149	1.9944	9.3697	1.8073	0.3481	0.0003	0.0182	14.4242	65 (3)	91 (2)	0.28	496
81		0.4702	0.1053			0.0521	0.5278	0.0521	0.0010	0.0552	0.1574	4.1855	0.7897	0.0010	0.1043	6.5016	64 (4)	77 (4)	−1.20	154
82	0.0020	0.5210	0.0573			0.0020	0.1187	0.0583		0.1218	0.1299	0.8612	14.8772		0.0031	16.7525	89 (2)	93 (1)	0.35	516
90	0.0115	0.0564					1.0436	0.4231	0.3949	0.1583	0.0282	0.0285		2.0201	0.3051	4.4696	45 (4)	61 (4)	2.13	170
95	0.0426	0.0493					0.0459		0.0098	0.0608	0.0908	0.1608	0.0809	0.1491	0.8094	1.4994	54 (4)	69 (4)	0.20	202
Total	1.6811	5.0945	1.5479	0.7222	0.3392	0.9482	11.9059	13.2366	3.5917	19.0479	14.1393	7.7062	16.3984	2.3380	1.3028	100.0000				
TSE	0.0490	0.3437	0.1530	0.0450	0.0303	0.1459	0.4230	0.4307	0.2810	0.6451	0.6362	0.5071	0.4048	0.1979	0.1311					
PA	94 (2)	42 (3)	43 (4)	47 (4)	49 (4)	63 (9)	67 (2)	75 (2)	55 (4)	81 (2)	66 (2)	54 (3)	91 (2)	86 (3)	62 (6)					
PAa	95 (2)	61 (4)	70 (6)	84 (4)	78 (6)	77 (9)	79 (2)	85 (2)	78 (4)	95 (1)	91 (1)	76 (3)	94 (1)	96 (2)	80 (6)					
n	208	512	255	219	215	75	436	508	164	566	476	287	470	104	134					4629

and land cover change from the reference classification rather than the map classification because of the fundamental principle of accuracy assessment that reference data quality is superior to map quality. Accordingly, we estimated map bias as another class-specific accuracy measure in addition to user's and producer's accuracies, where map bias is defined as the difference between the area of the class determined from the map labels and the area of the class determined from the reference labels (i.e., map area minus reference area). Map bias represents the bias incurred by employing "pixel counting" to calculate percent area of the land cover classes relative to the percent area estimated using the reference classification as the best assessment of ground condition (Olofsson et al., 2014).

The results of the analysis protocol (Section 2.3) are divided into 3 components. They are: 1) Level II and Level I accuracy estimates; 2) accuracy estimates for Level I change, and; 3) comparison of accuracy estimates for NLCD2016 with those reported for NLCD2011 (Wickham et al., 2017).

3. Results

3.1. Level II and Level I individual date accuracy estimates for the NLCD2016 database

All results in this section are for the NLCD2016 database so we refer only to the year of the land cover component and omit the preceding NLCD2016 database identifier. OA was 86.4% and 86.5% at Level II for 2016 and 2011, respectively, when agreement was based on a match between the map label and either the primary or alternate reference label (Tables 4 and 5). When the definition of agreement included the primary reference label only, Level II OA declined to 72.1% for both 2016 and 2011. OA standard errors were < 1% regardless of classification level and definition of agreement. UA values for dominant (by area) classes 41, 42, 52, 71, and 82 approached or exceeded 90% when the definition of agreement included the alternate reference label. Level II UA and PA for the 2016 and 2011 exceeded 70% for all classes except UA for wetland classes (90 and 95) and PA for open and low-density urban (21 and 22). The absolute value of map bias was greater than 1% for classes 21, 41, 42, 52, 81, and 90.

Level I OA was 90.6% and 90.5%for 2016 and 2011, respectively, when the definition of agreement included the alternate reference label and Level I OA declined to 79.0% for 2016 and 78.8% for 2011 when agreement did not include the alternate reference label (Table 6). UA and PA approached or exceeded 90% for all Level I classes except barren (UA and PA), urban and grassland (PA only) and wetland (UA only). UA and PA increased substantially for barren (30), shrubland (50), grassland (70), and wetland (90) when the agreement definition included the alternate reference label. Class-specific accuracies for the Level I urban class (20) tended to be much higher than its constituent individual Level II subclasses (21, 22, 23, and 24) accuracies, suggesting that a substantial portion of urban misclassification was within the Level II subclasses (compare Tables 4 and 5 with Table 6). Forest (40) and its Level II subclasses (41, 42, and 43) and agriculture (80) and its Level II subclasses (81 and 82) followed the same pattern. For the 2016 cropland, nearly 50% of the commission error (complement of UA) was attributable to a reference classification of pasture (Table 4).

Level II OA estimates for 2006 and 2001, which were based on the reference dataset collected for the NLCD2011 assessment (Wickham et al., 2017), were about 3% to 5% lower than those for 2016 and 2011 depending on the agreement definition (Tables 7 and 8). When agreement was defined as a match between the map label and the primary reference label only, Level II OA values for 2006 and 2001 were 66.9% and 67.2%, respectively, whereas the respective Level II OA estimates for 2016 and 2011 were 72.1% and 72.8%. Although class-specific differences in UA and PA estimates varied, those for 2016 and 2011 tended to be higher than their 2006 and 2001 counterparts regardless of agreement definition. The same pattern was apparent for Level I for OA,

Table 5
Agreement between map and reference labels for NLCD2016 2011 for the continental United States at Level II of the classification hierarchy. See Table 4 for an explanation of contents. OA = 72.8% (± 0.9%); OAa = 86.8% (± 0.7%).

	Reference																				
Map	11	21	22	23	24	31	41	42	43	52	71	81	82	90	95	Total	UA	UAa	MB	n
11	1.6610	0.0090					0.0263				0.0145	0.0145	0.0118	0.0290	0.0879	1.8541	90 (2)	96 (1)	0.11	232
21	0.0119	2.1628	0.3318	0.0474	0.0237		0.1006	0.0948	0.0237	0.0593	0.0119	0.0593	0.0356			2.9626	73 (3)	88 (2)	−2.05	251
22		0.5034	0.6611	0.2547	0.0121		0.0303	0.0061		0.0061	0.0061	0.0182	0.0121			1.5102	43 (3)	75 (3)	0.00	249
23		0.0519	0.1514	0.3418	0.1255		0.0130				0.0043		0.0043			0.6923	49 (4)	79 (3)	−0.01	160
24		0.0212	0.0202	0.0444	0.1584	0.0030				0.0010	0.0010					0.2492	63 (3)	75 (3)	−0.07	247
31	0.0108	0.0475		0.0113		0.5909	0.0108	0.0108		0.1840	0.1516	0.0108	0.0118			1.0404	57 (5)	76 (4)	0.10	109
41		0.1569	0.0389				8.0600	0.2219	0.6853	0.2900	0.0423	0.1627	0.0037	0.0389		9.7006	83 (2)	90 (2)	−2.31	315
42		0.1889				0.0462	0.1916	10.2734	0.4757	0.5445	0.2018			0.0462		11.9682	86 (2)	93 (2)	−1.51	446
43	0.0256	0.1026	0.0258				0.8956	0.5354	1.9581	0.0836	0.0767					3.7034	53 (4)	75 (4)	0.01	163
52		0.5718	0.1090	0.0003		0.1648	0.3970	1.5628	0.0753	15.2642	3.7413	0.4223	0.2264	0.0653		22.6005	68 (2)	87 (2)	3.61	753
71		0.0852	0.0043			0.1384	0.5983	0.2355	0.0320	2.2491	9.2466	1.8162	0.3292		0.0058	14.7406	66 (4)	89 (2)	0.68	650
81	0.0064	0.4798	0.1048		0.0003	0.0003	0.4766	0.0554	0.0021	0.1108	0.2118	4.3069	0.7321	0.0006	0.0564	6.5442	66 (4)	76 (4)	−1.26	311
82	0.0006	0.5212	0.0584			0.0006	0.1149	0.0606		0.0582	0.1734	0.8642	14.5819	0.0009	0.1159	16.5509	88 (2)	93 (1)	0.52	408
90		0.0569		0.0003			1.0553	0.4231	0.4294	0.0846	0.0914			2.0151	0.2878	4.4439	45 (4)	61 (4)	2.10	174
95	0.0271	0.0490	0.0013				0.0392	0.0003	0.0098	0.0558	0.0899	0.1285	0.0815	0.1430	0.8136	1.4389	57 (4)	70 (4)	0.07	161
Tot.	1.7432	5.0082	1.5071	0.7002	0.3200	0.9442	12.0097	13.4801	3.6913	18.9910	14.0646	7.8035	16.0305	2.3391	1.3673	100.0000				
TSE	0.0512	0.3381	0.1519	0.0525	0.0293	0.1522	0.4227	0.4236	0.2826	0.6401	0.6416	0.5053	0.4060	0.1979	0.1444					
PA	95 (2)	43 (3)	44 (5)	49 (4)	50 (5)	63 (10)	67 (2)	76 (2)	54 (4)	80 (2)	66 (2)	55 (3)	91 (1)	86 (3)	60 (6)					
PAa	96 (2)	62 (4)	70 (6)	83 (4)	76 (7)	73 (9)	78 (2)	84 (3)	77 (4)	94 (1)	90 (1)	78 (3)	93 (1)	96 (2)	74 (6)					
n	213	451	213	174	191	73	461	588	179	614	448	349	433	108	134					4629

Table 6
Level I agreement between map and reference labels for NLCD2016 2016 and NLCD2016 2011. See Table 4 for an explanation of contents.

NLCD2016 2016: OA = 79.0% (±0.8%) and OAa = 90.6% (±0.6%)

	Reference												
Map	10	20	30	40	50	70	80	90	Total	UA	UAa	MB	n
10	1.5870	0.0088	0.0003	0.0266		0.0088	0.0188	0.0799	1.7300	92 (2)	99 (1)	0.05	221
20	0.0121	5.0097	0.0028	0.2659	0.0561	0.0130	0.1324	0.0005	5.4925	91 (1)	94 (1)	−2.22	1202
30	0.0003	0.0543	0.6007	0.0324	0.1729	0.1513	0.0231	0.0003	1.0355	58 (5)	77 (4)	0.09	106
40	0.0256	0.6215	0.0462	23.0271	1.0105	0.2843	0.2108	0.0949	25.3207	91 (1)	96 (1)	−3.41	883
50		0.6636	0.1638	1.9994	15.3574	3.9002	0.6348	0.0548	22.7740	67 (2)	86 (2)	3.79	679
70		0.0982	0.0802	0.7078	1.9944	9.3697	2.1554	0.0185	14.4242	65 (3)	90 (2)	0.34	496
80	0.0020	1.1537	0.0542	0.7580	0.1769	0.2352	20.7657	0.1083	23.2541	89 (1)	94 (1)	−0.96	670
90	0.0541	0.1057		1.9172	0.2191	0.1191	0.2702	3.2836	5.9690	55 (3)	68 (3)	2.33	372
Total	1.6811	7.7156	0.9482	28.7343	18.9873	14.0815	24.2112	3.6408	100.0000				
TSE	0.0490	0.3586	0.1459	0.5221	0.6459	0.6342	0.5291	0.2180					
PA	94 (2)	65 (3)	63 (9)	80 (1)	81 (2)	67 (2)	86 (1)	90 (3)					
PAa	97 (2)	77 (3)	73 (9)	88 (1)	95 (1)	91 (1)	93 (1)	96 (2)					
n	208	1202	75	1108	565	474	759	238					4629

NLCD2016 2011: OA = 78.8% (±0.8%) and OAa = 90.5% (±0.6%)

	Reference												
Map	10	20	30	40	50	70	80	90	Total	UA	UAa	MB	n
10	1.6610	0.0090		0.0263		0.0145	0.0263	0.1169	1.8541	90 (2)	96 (1)	0.11	232
20	0.0119	4.9119	0.0030	0.2685	0.0663	0.0233	0.1295		5.4143	91 (1)	95 (1)	−2.12	907
30	0.0108	0.0589	0.5909	0.0216	0.1840	0.1516	0.0226		1.0404	57 (5)	76 (4)	0.10	109
40	0.0256	0.5131	0.0462	23.3004	0.9180	0.3174	0.1664	0.0851	25.3722	92 (1)	96 (1)	−3.81	924
50		0.6813	0.1648	2.0349	15.2644	3.7410	0.6488	0.0653	22.6005	68 (2)	87 (2)	3.61	753
70		0.0895	0.1384	0.8659	2.2491	9.2456	2.1464	0.0058	14.7406	63 (3)	89 (2)	0.73	650
80	0.0070	1.1644	0.0009	0.7097	0.1690	0.3331	20.5372	0.1738	23.0951	89 (1)	94 (1)	−0.79	719
90	0.0271	0.1075		1.9571	0.1404	0.1813	0.2100	3.2595	5.8828	55 (3)	68 (3)	2.18	335
Total	1.7432	7.5358	0.9442	29.1842	18.9913	14.0078	23.8871	3.7063	100.0000				
TSE	0.0512	0.3539	0.1522	0.5155	0.6401	0.6396	0.5246	0.2261					
PA	95 (2)	65 (3)	63 (10)	80 (1)	80 (2)	66 (2)	86 (1)	88 (3)					
PAa	96 (2)	78 (3)	73 (9)	87 (1)	95 (1)	90 (2)	94 (1)	95 (2)					
n	213	1030	73	1228	615	444	784	242					4629

UA, and PA between the two older (2001 and 2006) and two younger (2011 and 2016) land cover components in the NLCD2016 database (Table 9). Improved reference data quality cannot be discounted as a contributing factor to the differences in OA, UA, and PA between the two older (2001 and 2006) and two younger (2011 and 2016) land cover components in the NLCD 2016 database (see Discussion).

The patterns of misclassification were consistent across the nominal years of land cover dates and classification levels. The highest rates of misclassification occurred between shrubland (50) and grassland (70), ranging from 6% to 9% across the four nominal dates of land cover (Tables 6 and 9). Another notable example was the misclassification between the three upland forest classes (41, 42, 43) and woody wetland (90). The amount (percentage of map area) of misclassification between the upland forest classes and the woody wetlands class was much greater than the amount of misclassification between woody wetlands and emergent wetlands (95). Classification hierarchies, such as the one used here, are constructs that can be modified to address different objectives (Stehman and Wickham, 2020). For example, for the objective of mapping rangeland extent (e.g., Hewitt and Onsager, 1983), it would be reasonable to reconfigure the NLCD legend so that shrubland and grassland were Level II classes comprising a Level I rangeland class. *Re*-configuring NLCD2016 2016 so that shrubland and grassland form a Level I rangeland class changes OA from 79% (Table 6) to 85%, PA from 81% (shrubland) and 67% (grassland) to 93% for rangeland, and UA from 67% (shrubland) and 65% (grassland) to 82% for rangeland when the definition of agreement includes only the primary reference label.

Map bias tended to be more extreme for the 2016 and 2011 land cover components of the NLCD 2016 database than the 2006 and 2001 components (Tables 4–9), which may be attributable to the different reference datasets used. Map biases with absolute values of 1% or greater summed to greater than 12% for the 2016 and 2011 land cover components, whereas their 2006 and 2001 counterparts summed to less than 8%. Forest and shrubland map bias exceeded +3.5% (overestimate relative to reference classification) for the 2016 and 2011 components of the NLCD2016 database, whereas the highest map bias absolute value for the 2006 and 2001 components was less than 2% . Map biases for deciduous forest were less than −2% (underestimate relative to reference classification) and greater that +1% for woody wetland regardless of the era of the NLCD2016 land cover components.

3.2. *Change accuracies*

Overall accuracy of the binary change versus no change classification was about 96% (agreement based on only the primary reference labels) for all three change periods, 2001–2006, 2006–2011, and 2011–2016 (Table 10). These high OA values were expected because of the dominance of the no change class. UA for the change class ranged from 39.5% (2011 — 2016) to 45.6% (2001 — 2006) and PA ranged from 35.6% (2001 — 2006) to 47.4% (2011–2016). UA was higher than PA for the two earlier change periods but reversed so that UA was lower than PA for 2011–2016.

The estimated area of 2011–2016 land-cover change based on the reference sample data was 3.04% (± 0.3%) of the 7,792,667 km^2 area of the conterminous United States (see Table 3 caption), or 236,897 km^2, an area nearly the size of the state of Minnesota. The corresponding change column marginal totals for 2006–2011 and 2001–2006 using the NLCD2011 reference sample assessment (Wickham et al., 2017; p. 337) were 3.58% (± 0.2%) and 3.54% (± 0.2%), respectively, which translate to 278,977 km^2 and 275,860 km^2. Pooling over the 15-year period yields a per annum estimate of change area of about 52,872 km^2, an area almost equal to the size of the state of West Virginia. The estimated areas of change include silvicultural rotations and areas that may have changed more than once across 2001–2016 (Homer et al., 2020).

The corresponding change row marginal totals were 3.65%

Table 7
Agreement between map and reference labels for NLCD2016 2006 for the continental United States at Level II of the classification hierarchy based on reference data collected for accuracy assessment of the NLCD2011 database (Wickham et al., 2017). Reference samples of the NLCD class perennial ice & snow (12) were collected for the assessment of the NLCD2011 database and therefore reported here. Cell entries of 0.0000 indicate a percentage < 0.0005. See Table 4 for an explanation of contents. OA = 66.9% (± 0.7%); OAa = 83.6% (± 0.5%).

	Reference																				
Map	11	12	21	22	23	24	31	41	42	43	52	71	81	82	90	95	Total	UA	UAa	MB	n
11	1.5750		0.0152			0.0118	0.0206					0.0220		0.0079	0.0157	0.0131	1.6813	93 (2)	96 (1)	−0.20	192
12		0.0037					0.0022										0.0059	62 (17)	87 (12)	0.00	8
21	0.0041		1.3942	0.5908	0.0896	0.0003	0.0004	0.1748	0.1763	0.0283	0.1476	0.1048	0.0865	0.1012	0.0204		2.9196	48 (4)	58 (4)	−0.65	335
22			0.3954	0.7499	0.3024	0.0023	0.0102	0.0057	0.0396		0.0051	0.0031	0.0113	0.0111		0.0002	1.5362	49 (4)	75 (3)	−0.18	315
23	0.0004		0.0407	0.1286	0.2912	0.1462	0.0061	0.0040	0.0011		0.0003			0.0002			0.6186	47 (4)	79 (3)	−0.19	266
24			0.0149	0.0067	0.0248	0.1792	0.0045						0.0026				0.2326	77 (4)	83 (3)	−0.13	187
31	0.0321		0.0039	0.0016	0.0039	0.0015	0.3824	0.0047	0.0627		0.1225	0.1764	0.0175	0.0008	0.0093	0.0016	0.8209	47 (6)	59 (6)	−0.04	159
41	0.0234		0.1598	0.0010			0.0146	7.9689	0.4563	0.4456	0.3898	0.1049	0.1507	0.0669	0.1000	0.0331	9.9151	80 (2)	89 (1)	−2.58	660
42	0.0030		0.0931	0.0011		0.0000	0.0001	0.3374	9.4654	0.4481	1.5427	0.1197	0.0002	0.0000	0.0754	0.0005	12.0867	78 (2)	89 (1)	−0.70	1037
43			0.0554	0.0049	0.0003			1.5342	1.0292	0.8634	0.1703	0.0127	0.0386		0.0301		3.7391	23 (3)	58 (3)	1.84	367
52	0.0480		0.2518	0.0658	0.0347	0.0002	0.2735	0.5375	0.9337	0.0470	15.1945	4.7127	0.4539	0.0429	0.0035	0.0014	22.6013	67 (2)	87 (1)	0.25	1223
71	0.0286		0.3375	0.0010	0.0004		0.1316	0.2439	0.0858	0.0163	4.1276	7.4506	1.6206	0.4469	0.0003	0.0595	14.5507	51 (2)	83 (2)	1.12	1152
81	0.0285		0.4426	0.1067	0.0171	0.0248	0.0004	0.5014	0.0961		0.2228	0.3231	4.4020	0.8691	0.0452	0.0979	7.1779	61 (2)	75 (2)	−0.96	631
82	0.0238		0.3137	0.0378	0.0242	0.0007	0.0001	0.1732	0.0127	0.0004	0.0688	0.1975	1.2182	13.8894	0.0369	0.0657	16.0633	86 (1)	93 (1)	0.53	912
90	0.0580		0.0260	0.0248	0.0202			0.9337	0.4212	0.0533	0.2706	0.0743	0.0085	0.0363	2.3969	0.1317	4.4556	54 (3)	70 (3)	1.44	329
95	0.0585		0.0221	0.0000			0.0093	0.0717	0.0104		0.0891	0.1300	0.1247	0.0622	0.2784	0.7390	1.5953	46 (4)	59 (4)	0.45	227
Tot.	1.8835	0.0037	3.5664	1.7209	0.8088	0.3670	0.8561	12.4911	12.7906	1.9023	22.3518	13.4319	8.1353	15.5350	3.0122	1.1436	100.0000				
TSE	0.0867	0.0015	0.2270	0.1332	0.0756	0.0348	0.1039	0.3214	0.3227	0.1638	0.5367	0.5025	0.3442	0.3231	0.1703	0.0966					
PA	84 (3)	100 (0)	39 (3)	44 (4)	36 (4)	49 (5)	45 (6)	64 (2)	74 (2)	45 (4)	68 (2)	55 (2)	54 (2)	89 (1)	80 (3)	64 (5)					
PAa	89 (3)	100 (0)	60 (3)	60 (4)	63 (5)	71 (8)	65 (7)	79 (2)	84 (1)	84 (3)	87 (1)	85 (2)	73 (2)	93 (1)	90 (2)	82 (4)					
n	215	5	601	322	251	214	140	874	1184	169	1257	859	651	874	223	161					8000

Table 8
Agreement between map and reference labels for NLCD2016 2001 for the continental United States at Level II of the classification hierarchy based on reference data collected for accuracy assessment of the NLCD2011 database. See Table 4 for an explanation of contents. $OA_1 = 67.2\%$ ($\pm 0.7\%$); OAa = 83.7% ($\pm 0.5\%$).

	Reference																				
Map	11	12	21	22	23	24	31	41	42	43	52	71	81	82	90	95	Total	UA	UAa	MB	n
11	1.6116		0.0269				0.0118			0.0008	0.0046	0.0043	0.0079		0.0399	0.0387	1.7466	92 (2)	95 (2)	−0.14	201
12		0.0037					0.0022										0.0059	62 (17)	87 (12)	0.00	8
21	0.0040		1.3288	0.5818	0.0844			0.1828	0.1826	0.0280	0.1770	0.1057	0.0876	0.1087	0.0202		2.8916	46 (4)	56 (4)	−0.54	226
22			0.4050	0.7187	0.2738	0.0106	0.0100	0.0043	0.0394		0.0086	0.0020	0.0127	0.0087			1.4937	48 (4)	74 (3)	−0.13	212
23			0.0346	0.1168	0.2585	0.1372	0.0049	0.0078					0.0039	0.0003			0.5640	46 (4)	79 (3)	−0.15	183
24			0.0086	0.0076	0.0230	0.1579	0.0045					0.0026					0.2041	77 (3)	84 (3)	−0.10	165
31	0.0275		0.0093	0.0016	0.0039	0.0015	0.3770	0.0047	0.0627		0.1250	0.1836	0.0175	0.0008	0.0093	0.0016	0.8259	46 (6)	59 (6)	0.04	159
41	0.0234		0.1610	0.0038		0.0009	0.0100	8.1914	0.4538	0.4667	0.2986	0.1278	0.1480	0.0671	0.1000	0.0331	10.0856	81 (2)	89 (1)	−2.60	770
42	0.0030		0.1093	0.0012				0.3281	9.6967	0.4232	1.6054	0.1056	0.0292	0.0010	0.0769	0.0005	12.3801	78 (2)	89 (1)	−0.81	1168
43			0.0554	0.0051	0.0003		0.0003	1.5197	1.0519	0.8434	0.1787	0.0394	0.0387		0.0301		3.7628	22 (3)	57 (3)	1.85	396
52	0.0216		0.2742	0.0309	0.0029	0.0005	0.2345	0.5399	0.8050	0.0794	15.1765	4.8059	0.4110	0.0438	0.0016	0.0014	22.4293	68 (2)	88 (1)	0.32	1205
71	0.0160		0.2612	0.0011	0.0010		0.1267	0.1506	0.2567	0.0112	4.0329	7.2397	1.5980	0.3980	0.0011	0.0560	14.1502	51 (2)	82 (2)	0.83	1109
81	0.0285		0.3994	0.0926	0.0181	0.0004		0.5646	0.1502	0.0017	0.2084	0.3907	4.4893	0.9116	0.0453	0.0979	7.3986	61 (2)	75 (2)	−0.81	732
82	0.0279		0.3094	0.0380	0.0267			0.1974	0.0130	0.0004	0.0407	0.0846	1.2624	13.9842	0.0283	0.0821	16.0950	87 (1)	94 (1)	0.48	915
90	0.0580		0.0301	0.0248	0.0202			0.9541	0.4487	0.0625	0.2008	0.0721	0.0044	0.0366	2.4364	0.1030	4.4516	55 (3)	71 (3)	1.41	326
95	0.0630		0.0221				0.0052	0.0355	0.0317		0.0531	0.1565	0.0973	0.0576	0.2514	0.7417	1.5149	49 (4)	62 (4)	0.36	225
Total	1.8845	0.0037	3.4352	1.6240	0.7128	0.3089	0.7870	12.6808	13.1924	1.9172	22.1104	13.3206	8.2078	15.6185	3.0404	1.1558	100.0000				
TSE	0.0845	0.0015	0.2236	0.1279	0.0661	0.0210	0.0960	0.3214	0.3237	0.1640	0.5340	0.5004	0.3450	0.3214	0.1704	0.0993					
PA	86 (3)	100 (0)	39 (3)	44 (4)	36 (4)	51 (3)	48 (7)	65 (2)	74 (2)	44 (4)	69 (2)	54 (2)	55 (2)	90 (1)	80 (3)	64 (5)					
PAa	90 (2)	100 (0)	61 (4)	61 (4)	65 (5)	80 (4)	70 (7)	80 (2)	82 (1)	82 (3)	87 (1)	84 (2)	74 (2)	93 (1)	90 (2)	81 (4)					
n	217	5	444	256	181	179	127	965	1352	190	1203	857	708	930	227	159					8000

Table 9
Agreement between map and reference labels for NLCD2016 2006 and NLCD2016 2001 for the continental United States at Level I of the classification hierarchy based on reference data collected for the NLCD 2011 product suite. See Table 4 for an explanation of contents.

NLCD2016 2006: OA = 75.4% (± 0.7%); OAa = 89.2% (± 0.5%)

	Reference												
Map	10	20	30	40	50	70	80	90	Total	UA	UAa	MB	n
10	1.5787	0.0269	0.0228			0.0220	0.0079	0.0288	1.6872	94 (2)	96 (1)	−0.20	200
20	0.0046	4.3571	0.0213	0.4298	0.1530	0.1079	0.2129	0.0205	5.3070	82 (2)	87 (2)	−1.16	1103
30	0.0321	0.0109	0.3824	0.0674	0.1225	0.1764	0.0183	0.0109	0.8209	47 (6)	59 (6)	−0.04	159
40	0.0263	0.3157	0.0147	22.5485	2.1028	0.2373	0.2565	0.2390	25.7408	88 (1)	95 (1)	−1.44	2064
50	0.0480	0.3527	0.2735	1.5182	15.1945	4.7127	0.4968	0.0049	22.6013	67 (2)	87 (2)	0.25	1223
70	0.0286	0.3389	0.1316	0.3460	4.1276	7.4506	2.0675	0.0598	14.5507	51 (3)	83 (2)	1.12	1152
80	0.0524	0.9677	0.0005	0.7839	0.2917	0.5207	20.3787	0.2457	23.2412	88 (1)	93 (1)	−0.43	1543
90	0.1165	0.0931	0.0093	1.4902	0.3598	0.2043	0.2316	3.5461	6.0509	59 (3)	74 (2)	1.90	556
Total	1.8872	6.4631	0.8561	27.1840	22.3518	13.4319	23.6703	4.1557	100.0000				
TSE	0.0867	0.2473	0.1039	0.3895	0.5367	0.5025	0.3608	0.1860					
PA	84 (3)	67 (2)	45 (6)	83 (1)	68 (2)	55 (2)	86 (1)	85 (2)					
PAa	89 (3)	81 (2)	65 (7)	91 (1)	88 (1)	85 (2)	92 (1)	96 (1)					
n	220	1388	140	2227	1257	859	1525	384					8000

NLCD2016 2001: OA = 75.7% (± 0.7%); OAa = 89.2% (± 0.5%)

	Reference												
Map	10	20	30	40	50	70	80	90	Total	UA	UAa	MB	n
10	1.6153	0.0269	0.0140	0.0008	0.0046	0.0043	0.0079	0.0786	1.7526	92 (2)	94 (2)	−0.14	209
20	0.0040	4.1471	0.0194	0.4449	0.1856	0.1104	0.2218	0.0202	5.1533	80 (2)	85 (2)	−0.93	786
30	0.0275	0.0162	0.3770	0.0675	0.1250	0.1836	0.0183	0.0109	0.8259	46 (6)	59 (6)	0.04	159
40	0.0263	0.3369	0.0103	22.9750	2.0827	0.2729	0.2839	0.2406	26.2286	88 (1)	95 (1)	−1.56	2334
50	0.0216	0.3086	0.2345	1.4243	15.1765	4.8059	0.4548	0.0030	22.4293	68 (2)	88 (1)	0.32	1205
70	0.0160	0.2634	0.1267	0.4185	4.0329	7.2397	1.9960	0.0571	14.1502	51 (2)	82 (2)	0.83	1109
80	0.0564	0.8846		0.9272	0.2491	0.4752	20.6475	0.2535	23.4936	88 (1)	93 (1)	−0.33	1647
90	0.1210	0.0972	0.0052	1.5324	0.2539	0.2286	0.1959	3.5323	5.9664	59 (3)	75 (2)	1.77	551
Total	1.8882	6.0809	0.7870	27.7904	22.1104	13.3206	23.8263	4.1963	100.0000				
TSE	0.0846	0.2383	0.0960	0.3917	0.5340	0.5004	0.3573	0.1874					
PA	86 (3)	68 (2)	48 (7)	83 (1)	69 (2)	54 (2)	87 (1)	84 (2)					
PAa	91 (2)	83 (2)	71 (7)	90 (1)	89 (1)	85 (2)	92 (1)	94 (1)					
n	222	1060	127	2507	1203	857	1638	386					8000

Table 10
NLCD2016 accuracy of binary change versus no change for the agreement definition that includes matches between the map label and primary label only. See Table 4 for an explanation of contents. Results for 2006–2011 and 2001–2006 are based on reference sample data collected for NLCD2011 (Wickham et al., 2017).

Map	No Change	Change	Total	UA	MB
2011–2016. OA = 96.2% (±0.3%)					
No Change	94.7579	1.5975	96.3554	98.3 (0.3)	−0.61
Change	2.2041	1.4405	3.6446	39.5 (1.5)	0.61
Total	96.9620	3.0380			
TSE	0.2815	0.2815			
PA	97.7 (0.06)	47.4 (4.4)			
2006–2011. OA = 96.2% (±0.3%)					
No Change	94.5795	2.2571	96.8366	97.7 (0.2)	0.42
Change	1.8401	1.3233	3.1634	41.8 (3.2)	−0.42
Total	96.4196	3.5804			
TSE	0.2367	0.2367			
PA	98.1 (0.2)	37.0 (2.9)			
2001–2006. OA = 96.2% (±0.3%)					
No Change	94.9641	2.2780	97.2421	97.7 (0.2)	0.78
Change	1.4992	1.2587	2.7579	45.6 (3.5)	−0.78
Total	96.4633	3.5367			
TSE	0.2336	0.2336			
PA	98.4 (0.2)	35.6 (3.0)			

(2011–2016), 3.16% (2006–2011), and 2.76% (2001–2006) of map area (see Table 3 caption). Across the three change reporting periods, the area of change for the conterminous United States summed to 10.16% based on the reference (column) classification and summed to 9.57% based on the map (row) classification, so the total change estimated from the reference data exceeded the area of change mapped by NLCD2016. For the NLCD2016 database, a positive map bias for 2011–2016 (0.61%) indicated that change area was overestimated when compared to the reference classification, whereas negative map biases for 2006–2011 (−0.42%) and 2001–2006 (−0.78%) indicated change area was underestimated when compared to the reference classification.

For the class-specific accuracies of the change products (Table 11), UA and PA of the Level I no change (stable) themes closely mimic the accuracy of the classes for each land cover-year component. This result reflects the fact that change is relatively rare so the population of pixels in each no change class is almost the same as the population for that land cover class in each of the individual years. For the change themes, 2011–2016 forest loss was the most accurately mapped theme, for which UA and PA were both ~75% when agreement included the primary and alternate labels. Additionally, PA was at least 70% for 2011–2016 change for water loss, grassland loss, and grassland gain when agreement included the primary and alternate labels. UA and PA tended to be higher for the 2011–2016 change period than the 2006–2011 and 2001–2006 change periods (Table 11). Also, UA for the 2011–2016 change period was noticeably more precise than for the two earlier change periods. The average of standard errors (SE) for UA for the 2011–2016 change period was about 5%, whereas the average SE for the two earlier change periods (combined) was about 9%. However, there was little difference in the precision of PA between the 2011–2016 change period and the two earlier periods. The NLCD2016 reference sample was used for the 2011–2016 change period, whereas the NLCD2011 reference sample was used for the two earlier change periods, so differences between the 2011–2016 change period and the two earlier periods may be attributable to the use of different reference datasets.

Table 11
NLCD2016 user's (UA) and producer's (PA) accuracies (%) and (standard errors) for loss, gain, and no change (no Δ) themes where agreement was defined as a match between map and either primary or alternate reference labels. The 2011–2016 results are based on the reference data collected for this assessment (NLCD2016) and the 2006–2011 and 2001–2006 results are based on the reference dataset collected for assessment of NLCD2011 (Wickham et al., 2017).

	UA			PA		
Themes	2011–2016	2006–2011	2001–2006	2011–2016	2006–2011	2001–2006
Water loss	46 (9)	43 (21)	30 (11)	77 (11)	56 (23)	61 (13)
Water gain	60 (14)	41 (11)	41 (20)	60 (16)	75 (11)	40 (21)
Urban gain	64 (3)	61 (9)	66 (4)	51 (20)	48 (13)	41 (9)
Forest loss	74 (3)	59 (5)	67 (6)	75 (8)	67 (6)	56 (5)
Forest gain	35 (3)	51 (7)	41 (7)	52 (9)	41 (6)	40 (8)
Shrubland loss	27 (3)	34 (6)	30 (27)	48 (9)	36 (6)	27 (8)
Shrubland gain	29 (3)	49 (7)	38 (7)	59 (9)	52 (7)	44 (7)
Grassland loss	34 (3)	46 (7)	36 (7)	80 (11)	65 (9)	59 (10)
Grassland gain	52 (3)	38 (5)	52 (6)	73 (10)	71 (8)	80 (7)
Agriculture loss	27 (4)	34 (12)	11 (10)	20 (12)	29 (12)	17 (5)
Agriculture gain	33 (3)	5 (3)	10 (6)	23 (8)	12 (8)	21 (12)
Water, no Δ	98 (1)	96 (2)	95 (2)	96 (2)	97 (1)	93 (3)
Urban, no Δ	94 (1)	87 (2)	85 (2)	83 (3)	88 (2)	89 (2)
Forest, no Δ	96 (1)	95 (1)	95 (1)	90 (1)	93 (1)	93 (1)
Shrubland, no Δ	86 (2)	87 (1)	88 (1)	96 (1)	92 (1)	93 (1)
Grassland, no Δ	90 (2)	83 (2)	82 (2)	94 (1)	91 (1)	91 (1)
Agriculture, no Δ	94 (1)	93 (1)	93 (1)	96 (1)	94 (1)	94 (1)

3.3. Comparison of NLCD2016 database and NLCD2011 database

Land cover accuracy for the 2011 land cover component from NLCD2016 database was compared to its counterpart in the NLCD2011 database using the NLCD2016 reference sample. Differences in accuracies between the two versions of the 2011 land cover component are more likely attributable to differences in classification methods under the assumption that reference data quality is constant across the NLCD2016 reference sample dataset. Overall accuracies for NLCD2016 2011 were 4% - 7% greater than for NLCD2011 2011 depending on the classification hierarchy and agreement definition (Table 12). The gains in OA translated into nearly uniform gains in UA and PA. Declines in UA and PA were rare.

We used the NLCD2011 reference dataset to compare NLCD2016 and NLCD2011 change accuracies for the 2001–2006 and 2006–2011 time periods. OA of the binary change / no change classification were virtually the same for the two NLCD databases (NLCD2011 = 96.6% (both periods); NLCD2016 = 95.9% (2001–2006), 96.2% (2006–2011)). The two NLCD databases differed considerably for UA and PA for the change class. NLCD2016 had lower UA than NLCD2011 (difference was 14.9% in 2006–2011 and 8.5% in 2001–2006) but NLCD2016 had higher PA than NLCD2011 (difference was 10.4% in 2006–2011 and 11.2% in 2001–2006). For both NLCD2016 and NLCD2011, UA of change was greater than PA of change, but the imbalance between UA and PA was much greater for NLCD2011.

NLCD2016 had higher UA and PA for the class-specific no change themes of water, urban, forest, and agriculture (Table 13). The improvement of NLCD2016 over NLCD2011 for UA of no change in urban was 10% for both 2006–2011 and 2011–2016. NLCD2016 had lower UA and PA for the shrub and grass no change themes. For the change themes, NLCD2016 tended to have lower UA but higher PA than NLCD2011. One exception to this pattern was that UA of urban gain was 15% higher for NLCD2016 than UA for NLCD2011. Standard errors for UA and PA of the rare change themes are still relatively large, so relatively large differences between NLCD2011 and NLCD2016 UA and PA estimates may be partly attributable to sampling variability. The general pattern of greater PA and lower UA for NLCD2016 relative to NLCD2011 would be robust to sampling variability, as it reflects the trade-off that improving PA is typically accompanied by a decrease in UA (e.g., avoiding change omission errors by mapping more change may create more commission errors). NLCD2016 mapped a larger area of change than did NLCD2011, as map change of NLCD2016 was 3.16% for 2006–2011 and 2.76% for 2001–2006 compared to NLCD2011 which had 1.68% for 2006–2011 and 1.60% for 2001–2006.

Table 12
Difference between NLCD2016 2011 and NLCD2011 2011 class-specific accuracies at Level II and Level I for agreement defined as match between the map label and either primary or alternate reference label. The label V2 (version 2 of the 2011 land cover component) refers to NLCD2016 2011; Diff = NLCD2016 2011 — NLCD2011 2011.

Level II				
	Pri + Alt			
	UA		PA	
Class	V2	Diff	V2	Diff
11	98	5	96	15
21	88	26	62	2
22	75	−2	70	14
23	79	−1	83	19
24	75	−1	76	4
31	76	15	73	−8
41	90	5	78	−3
42	93	5	84	5
43	75	6	77	12
52	87	5	94	6
71	89	7	90	3
81	76	12	78	10
82	93	7	93	5
90	61	3	96	10
95	70	10	74	2
OA	72	7		
OAa	86	6		
Level I				
10	96	4	96	6
20	95	16	78	9
30	76	14	73	5
40	96	2	87	2
50	87	5	85	−6
70	89	7	90	6
80	94	4	94	6
90	68	3	94	5
OA	79	5		
OAa	90	4		

4. Discussion

Monitoring assumes accurate measurement, and the NLCD program operates in a mode of continuous effort toward product improvement to advance its goal of land cover monitoring. Overall accuracy of

Table 13
Comparison of accuracy (%) of NLC2016 with NLCD2011 based on differences in accuracy of change themes for 2001 — 2006 and 2006 — 2011 (positive differences indicate NLCD2016 accuracy was higher than NLCD2011 accuracy). Comparisons are based on the reference dataset collection for assessment of NLCD2011 (Wickham et al., 2017) and agreement is defined using only the primary reference class labels.

	UA		PA	
Themes	2006–2011	2001–2006	2006–2011	2001–2006
Water loss	−2	−23	26	−5
Water gain	−17	2	−1	14
Urban gain	15	−1	12	0
Forest loss	−8	−7	7	10
Forest gain	−10	−11	9	4
Shrubland loss	−12	0	6	8
Shrubland gain	−4	−3	4	6
Grassland loss	−11	−1	12	12
Grassland gain	−15	−2	17	28
Agriculture loss	2	−12	12	−2
Agriculture gain	−10	−4	−5	−2
Water, no Δ	7	6	2	1
Urban, no Δ	10	10	4	4
Forest, no Δ	2	2	1	1
Shrubland, no Δ	−3	−3	−4	−2
Grassland, no Δ	−2	−3	−6	−7
Agriculture, no Δ	1	2	2	2

NLCD2016 2011 was about 5% higher than OA for NLCD2011 2011 when agreement was defined as a match between the map and either the primary or alternate reference label. The 5% increase equates to an area about the size of California. In other words, the NLCD2016 database has accurately mapped an additional 404,000 km^2 compared to the NLCD2011 database for the nominal 2011 year at Level II of the classification hierarchy. Continued methodological innovations (Jin et al., 2019; Yang et al., 2018) appear to be leading to product improvement.

The plausible assertion that methodological innovation has been a foundation of NLCD product improvement derives from the trends in reported overall accuracies of the individual eras (Table 1). The accuracy trends for land cover change do not support such an assertion. Across the NLCD accuracy assessments involving land cover change (Wickham et al., 2013, 2017), UA has been static and only forest loss has been 70% or greater. Similarly, PA had been static and lower than UA until the improvement reported here for 2011–2016 (e.g., grassland loss and gain). One plausible explanation is that the same effort applied to methodological innovation for mapping has not been applied to the response design component of accuracy assessment.

The response design is the set of protocols for assignment of reference land cover labels (sensu Stehman and Czaplewski, 1998). Numerous studies have quantified variability in reference label assignment (i.e., reference data quality) and its potential impact on map accuracy (Foody, 2009, 2010; 2013; Mann and Rothley, 2006; Pengra et al., 2020; Powell et al., 2004). As summarized by Foody (2002) and Stehman and Foody (2019), reference data error may produce biased estimates of accuracy and area and the bias can be substantial even when reference data error is minimal. One outcome of these research efforts has been development of guidelines for collecting high-quality (i.e., the best possible) reference data (Olofsson et al., 2014, p. 55). NLCD accuracy assessments have implemented and contributed to the development of these guidelines (see references in Olofsson et al., 2014). Further advance of such guidelines is likely because of the reliance on and interest in planet-wide monitoring of land cover dynamics (Rindfuss et al., 2004; Turner II et al., 2007; e.g., Song et al., 2018). A logical pathway for advancement of response design protocols and guidelines may be research focused explicitly on what is the best we can do and how do we get there, i.e., a reference data quality ideal. Research focused explicitly on defining a reference data ideal would advance established good practices (Olofsson et al., 2014)

Interdisciplinary research with cognitive psychologists may be a component of the research needed to support establishment of a reference data quality ideal. Van Coillie et al. (2014) evaluated reference data accuracy against a suite of personality and cognitive factors and found that demographic (e.g., age, sex), psychological (e.g., conscientiousness), cognitive factors (e.g., visual memory span), and external factors (e.g., fatigue) influenced the accuracy of visual image interpretation. The variability across psychological and cognitive factors reinforces intuitive knowledge that all humans are not equally capable of accomplishing a given task. Some are likely to be more capable at reference data collection than others. The difference in OA between the two earlier (2001 and 2006) and two later (2011 and 2016) components of the NLCD2016 database may reflect how best practices can fall short of the reference data quality ideal. The reference data used to assess the 2001 and 2006 components were obtained during the NLCD2011 database accuracy assessment (Wickham et al., 2017). These NLCD2011 reference data were collected by a team of four interpreters, whereas the reference data used for 2011 and 2016 components of the NLCD2016 database assessment were collected by a subset of the previous team of four. Because of personnel changes between the NLCD2016 and NLCD2011 reference data collection efforts, differences in cognitive abilities, experience, and other factors cannot be discounted as contributing to differences in OA between the two earlier and two later dates of the NLCD2016 database.

There are undoubtedly many other factors that would need to be considered (e.g., Bianchetti, 2016; Bianchetti and MacEachren, 2015) if a reference data quality ideal were to be developed. We have highlighted only one of the likely foundational elements (human capability) on which a reference data quality ideal would be supported. We introduced the concept of a reference data quality ideal to further emphasize the importance of consistent high-quality reference data to an operational program focused on remotely sensed land cover monitoring, change, and trends. The loss and gain trends from assessment of the NLCD2006 database (Wickham et al., 2013) through NLCD 2016 can be summarized accurately as poor overall except for urban gain and forest loss (notwithstanding some improvement reported here for 2011–2016). With Level II and Level I single-date overall accuracies of 87% and 91%, respectively, for the 2011 and 2016 components of the NLCD2016 database, it is plausible that mapping methodologies may be at an asymptotic maximum beyond which substantial increases in estimated accuracy cannot be achieved. It may be that advances in the response design component of accuracy assessment and particularly further effort toward defining and realizing a reference data quality ideal are needed to improve accuracies of NLCD land cover change.

5. Conclusion

NLCD is a widely recognized and widely utilized database (Deering, 2014). It is used in modeling and assessment studies and to inform policy. Statistically rigorous assessments of data accuracy (Stehman and Czaplewski, 1998) are essential to continued widespread use of NLCD. The results reported herein support the conclusion that NLCD's emphasis on methodological advancement has led to improved thematic accuracies of the individual dates (e.g., 2001, 2006, 2011, and 2016) of the land cover component of the database. Nevertheless, NLCD's overarching goal of being a national land cover monitoring program (Homer et al., 2020; Yang et al., 2018) is currently hindered by the modest to poor agreement between reference- and map-derived land cover change. Further research on the response design components of accuracy assessment as they relate to land cover change may be a fruitful path forward to help NLCD more fully realize its land cover monitoring objective.

Credit author statement

JW conducted analysis, and wrote and edited the paper; SVS conducted analysis, and wrote and edited the paper; DS collected reference

data, and wrote and edited the paper; LG collected reference data, and wrote and edited the paper; JD supervised reference data collection and edited the paper.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This document has been reviewed by the U.S. Environmental Protection Agency, Office of Research and Development, and approved for publication. The views expressed in this paper are those of the authors and do not necessarily reflect the views or policies of the US Environmental Protection Agency. We also wish to thank Don Ebert (EPA), and anonymous reviewers for their thoughtful comments on earlier versions of the paper. Funding support (SS) was provided via contract G17AC00237 between SUNY-ESF and USGS.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.rse.2021.112357.

References

Song, X.-P., Hansen, M.C., Stehman, S.V., Potapov, P.V., Tyukavina, A., Vermote, E.F., Townshend, J.R., 2018. Global land change from 1982 to 2016. Nature 560, 639–643. https://doi.org/10.1038/s41586-018-0411-9.

Bianchetti, R.A., 2016. Describing the problem-solving strategies of expert image interpreters using graphical knowledge elicitation methods. GIScience Remote Sens. 53, 561–577. https://doi.org/10.1080/15481603.2016.1196424.

Bianchetti, R.A., MacEachren, A.M., 2015. Cognitive themes emerging from air photo interpretation texts published to 1960. ISPRS Int. J. Geo Inf. 4, 551–571.

Cochran, W.G., 1977. Sampling Techniques, 3rd ed. John Wiley & Sons, New York.

Deering, C.A., 2014. The National Land Cover database project: the story of its impact. In: GSA Annual Meeting, Vancouver, British Columbia, Canada, 21 October, 2014.

Foody, G.M., 2002. Status of land cover classification accuracy assessment. Remote Sens. Environ. 80, 185–201.

Foody, G.M., 2009. The impact of imperfect ground reference data on the accuracy of land cover change estimation. Int. J. Remote Sens. 30, 3275–3281. https://doi.org/10.1080/01431160902755346.

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

Fry, J., Xian, G., Jin, S., Dewitz, J., Homer, C., Yang, L., Barnes, C., Herold, N., Wickham, J., 2011. Completion of the 2006 national land cover database for the conterminous United States. Photogramm. Eng. Remote. Sens. 77, 858–863.

Gopal, S., Woodcock, C., 1994. Theory and methods for accuracy assessment of thematic maps using fuzzy sets. Photogramm. Eng. Remote. Sens. 60, 181–188.

Hansen, M.C., Loveland, T.R., 2012. A review of large area monitoring of land cover change using Landsat data. Remote Sens. Environ. 122, 66–74. https://doi.org/10.1016/j.rse.2011.08.024.

Hewitt, G.B., Onsager, J.A., 1983. Control of grasshoppers on rangeland in the United States — a perspective. J. Range Manag. 36, 202–207.

Homer, C., Huang, C., Yang, L., Wylie, B., Coan, M., 2004. Development of a 2001 National Land Cover Database for the United States. Photogramm. Eng. Remote. Sens. 70, 829–840.

Homer, C., Dewitz, J., Fry, J., Coan, M., Hossain, N., Larson, C., Herold, N., McKerrow, A., VanDriel, J., Wickham, J., 2007. Completion of the 2001 National Land Cover Database for the conterminous United States. Photogramm. Eng. Remote. Sens. 73, 337–341.

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

Homer, C., Dewitz, J., Jin, S., Danielson, P., Gass, L., Funk, M., Wickham, J., Stehman, S., Auch, R., Riitters, K., 2020. Conterminous land cover change patterns 2001-2016 from the 2016 National Land Cover Database. ISPRS J. Photogramm. Remote Sens. 162, 184–199. https://doi.org/10.1016/j.isprsjprs.2020.02.019.

Jin, S., Yang, L., Danielson, P., Homer, C., Fry, J., Xian, G., 2013. A comprehensive change detection method for updated the National Land Cover Database to 2011. Remote Sens. Environ. 132, 159–175. https://doi.org/10.1016/j.rse.2013.01.012.

Jin, S., Homer, C., Yang, L., Danielson, P., Dewitz, J., Li, C., Zhue, Z., Xian, G., Howard, D., 2019. Overall methodology design for the United States National Land Cover Database 2016 products. Remote Sens. 11, 2971. https://doi.org/10.3390/rs11242971.

Mann, S., Rothley, K.D., 2006. Sensitivity of Landsat/IKONOS accuracy comparison to errors in photointerpreted reference data and variations in test point sites. Int. J. Remote Sens. 27, 5027–5036. https://doi.org/10.1080/01431160600784291.

Olofsson, P., Foody, G.M., Herald, 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.

Pengra, B.W., Stehman, S.V., Horton, J.A., Dockter, D.J., Schroeder, T.A., Yang, Z., Cohen, W.B., Healey, S.P., Loveland, T.R., 2020. Quality control and assessment of interpreter consistency of annual land cover reference data in an operational national monitoring program. Remote Sens. Environ. 238, 111261. https://doi.org/10.1016/j.rse.2019.111261.

Powell, R.L., Matzke, N., de Souza Jr., C., 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, 221–234. https://doi.org/10.1016/j.rse2003.12.007.

Rindfuss, R.R., Walsh, S.J., Turner II, B.L., Fox, J., Mishra, V., 2004. Developing a science of land change: challenges and methodological issues. P. Natl. Acad. Sci. USA 104, 20666–20671. https://doi.org/10.1073/pnas.0401545101.

Särndal, C.E., Swensson, B., Wretman, J., 1992. Model-Assisted Survey Sampling. Springer-Verlag, New York.

Stehman, S.V., 2001. Statistical rigor and practical utility in thematic map accuracy assessment. Photogramm. Eng. Remote. Sens. 67, 727–734.

Stehman, S.V., 2014. Estimating area and map accuracy for stratified random sampling when the strata are different from the map classes. Int. J. Remote Sens. 35, 4923–4939.

Stehman, S.V., Czaplewski, R.L., 1998. Design and analysis for thematic map accuracy assessment: fundamental principles. Remote Sens. Environ. 64, 331–344. https://doi.org/10.1016/S0034-4257(98)00010-8.

Stehman, S.V., Foody, G.M., 2019. Key issues in rigorous accuracy assessment of land cover products. Remote Sens. Environ. 231, 111199. https://doi.org/10.1016/j.rse.2019.05.018.

Stehman, S.V., Wickham, J.D., 2011. Pixels, blocks of pixels, and polygons: choosing a spatial unit for thematic accuracy assessment. Remote Sens. Environ. 115, 3044–3055. https://doi.org/10.1016/j.rse.2011.06.007.

Stehman, S.V., Wickham, J.D., 2020. A guide for evaluating and reporting map data quality: Affirming Shao et al. "Overselling overall map accuracy misinforms about research reliability". Landsc. Ecol. 35, 1263–1267.

Stehman, S.V., Wickham, J.D., Smith, J.H., Yang, L., 2003. Thematic accuracy of the 1992 National Land-Cover Data (NLCD) for the eastern United States: statistical methodology and regional results. Remote Sens. Environ. 86, 500–516. https://doi.org/10.1016/S0034-4257(03)00128-7.

Turner II, B.L., Lambin, E.F., Reenberg, A., 2007. The emergence of land change science for global environmental change and sustainability. P. Natl. Acad. Sci. USA 104, 20666–20671. https://doi.org/10.1073/pnas.0704119104.

Van Coillie, F.M.B., Gardin, S., Anseel, F., Duyck, W., Verbeke, L.P.C., De Wulf, R.R., 2014. Variability of operator performance in remote sensing image interpretation: the importance of human and external factors. Int. J. Remote Sens. 35, 754–778. https://doi.org/10.1080/01431161.2013.873152.

Wickham, J.D., Stehman, S.V., Fry, J.A., Smith, J.H., Homer, C.G., 2010. Thematic accuracy of the NLCD 2001 land cover for the conterminous United States. Remote Sens. Environ. 114, 1286–1296. https://doi.org/10.1016/j.rse.2010.01.018.

Wickham, J., 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., Stehman, S.V., Gass, L., Dewitz, J.A., Sorenson, D.G., Granneman, B.J., Poss, R.V., Baer, L.A., 2017. The 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.

Yang, L., Jin, S., Danielson, P., Homer, C., Gass, L., Bender, S.M., Case, A., Costello, C., Dewitz, J., Fry, J., Funk, M., Granneman, B., Liknes, G.C., Rigge, M., Xian, G., 2018. A new generation of the United States National Land Cover Database: requirements, research priorities, design, and implementation strategies. ISPRS J. Photogramm. Remote Sens. 146, 108–123. https://doi.org/10.1016/j.isprsjprs.2018.09.006.