

Assessing map accuracy from a suite of site-specific, non-site specific, and spatial distribution approaches.

Mark D. Nelson^{a,*}, James D. Garner^a, Brian G. Tavernia^b, Stephen V. Stehman^c,
Rachel I. Riemann^d, Andrew J. Lister^e, Charles H. Perry^a

^a USDA Forest Service, Northern Research Station, 1992 Folwell Avenue, St. Paul, MN 55108, USA

^b National Audubon Society, Western Water Initiative, 231 West 800 South, Ste. E., Salt Lake City, UT 84101, USA

^c SUNY College of Environmental Science and Forestry, 322 Bray Hall, 1 Forestry Drive, Syracuse, NY 13210, USA

^d USDA Forest Service, Northern Research Station, 425 Joran Road, Troy, NY 12180, USA

^e USDA Forest Service, Northern Research Station, 3460 Industrial Drive, York, PA 17402, USA

ARTICLE INFO

Keywords:

Assessment framework
Map validation

ABSTRACT

Assessing a map's accuracy and efficacy requires careful consideration and calculation of a comprehensive set of measures, particularly when characteristics in geospatial datasets may be used for addressing multiple objectives. We describe a map accuracy assessment framework for categorical geospatial data over large areas that utilizes site-specific (plot-pixel), non-site specific (area estimates), and comparative accuracy (spatial and statistical distributions) methods. This comprehensive framework spans the range of approaches applied in most other studies, maximizing opportunities for both within- and between-study comparisons. The analysis framework incorporated recommended good practices for sampling design, response design, analysis protocols, and design-based inference for providing rigorous estimates of accuracy and precision. Further partitioning of disagreement allows users to assess the relative importance of spatial location versus total abundance of map classes to map error; we introduced a ratio of allocation disagreement to quantify disagreement for comparing relative importance of each across datasets. Assessment of fuzzy membership with linguistic scales may have minimal effect on overall accuracy, but can provide substantial improvement in user's accuracy and producer's accuracy for uncommon classes. We adapted non-site specific approaches to provide metrics and insights into spatial variability of map accuracy of thematic classes. Comparing estimates based on reference data and co-located samples of map data may reveal potential issues with reference data. We demonstrated example applications all components of the assessment framework for thematic maps of early successional forest and other land cover classes across the western Great Lakes region, USA.

1. Introduction

The utility and reliability of map-based information is dependent upon a map's quality and intended application. Assessing a map's accuracy and efficacy requires careful consideration and calculation of a comprehensive set of measures (Olofsson et al., 2014; Riemann et al., 2010; Tavernia et al., 2016), particularly when characteristics in geospatial datasets may be used for addressing multiple objectives (Laba et al., 2002). Map data of different measurement levels (e.g., nominal, ordinal, interval, ratio; Stevens, 1946) require assessments based on a system of measurement frameworks (Chrisman, 1998), not constrained

to a single standard metric or a single scale. We describe a map accuracy assessment framework for categorical geospatial data over large areas that utilizes site-specific (plot-pixel), non-site specific (area estimates), and comparative accuracy (spatial and statistical distributions) methods. Creating this framework involved assembling and adapting procedures for estimating two complementary measures - accuracy (agreement) and error (disagreement), in a multi-scale and spatially explicit manner. The proposed framework includes a suite of metrics that support a comprehensive understanding of map quality.

* Corresponding author.

E-mail addresses: Mark.D.Nelson@usda.gov (M.D. Nelson), James.Garner1@usda.gov (J.D. Garner), Brian.Tavernia@audubon.org (B.G. Tavernia), SVStehma@esf.edu (S.V. Stehman), Rachel.Riemann@usda.gov (R.I. Riemann), Andrew.Lister@usda.gov (A.J. Lister), Charles.H.Perry@usda.gov (C.H. Perry).

<https://doi.org/10.1016/j.rse.2021.112442>

Received 14 December 2020; Received in revised form 2 March 2021; Accepted 5 April 2021

Available online 16 April 2021

0034-4257/© 2021 Published by Elsevier Inc.

2. Assessment framework

Good practices for estimating area and assessing accuracy of land change are recommended in Olofsson et al. (2014). Our framework incorporates these good practices, augmented with additional metrics for conducting both site-specific and non-site-specific assessments, and an adaptation of the comparative accuracy assessment protocol of Riemann et al. (2010), which we refer to as a spatial distribution assessment (Table 1).

2.1. Reference data – sampling and response designs

Statistically rigorous map accuracy assessments require careful consideration of three basic components: 1) a sampling design used to select the reference sample, 2) a response design used to obtain the reference classification for each sampling unit, and 3) estimation and analysis procedures (Stehman and Czaplewski, 1998). In accuracy assessment, pixels, polygons, or blocks can all be used as the spatial unit of a response design (Olofsson et al., 2014). The choice of spatial unit also has implications on the sampling design (Stehman and Wickham, 2011). A reference data source is used to implement a response design for producing a reference classification that is compared to the map

classification. See Olofsson et al. (2014) for discussion of response designs with spatial units of unequal size or heterogeneous classes.

Olofsson et al. (2014) recommend design-based inference (Särndal et al., 1992) as the framework within which accuracy assessment design and associated area estimation is conducted. The chosen sampling design governs which estimators are appropriate for generating area and precision estimates, and how reference data should be distributed within the mapped area. The sampling design and response design should have sufficient spatial and temporal resolution to accurately label reference sample units. Selected estimators should be unbiased and account for inclusion probabilities associated with the sampling design. One advantage of correctly implemented probability sampling designs is that uncertainty estimates can be interpreted through the lens of sampling theory, with metrics such as sampling errors and confidence intervals having universal meaning. Reference data obtained for other purposes may be used when based on a probability sampling design satisfying good practice guidelines. An expansive review of the sampling designs typically applied to accuracy assessment is provided in Stehman (2009).

Map accuracy assessment requires quality reference data that should be considerably more accurate than the map being evaluated. For large area mapping projects, reference data typically are obtained via manual photo-interpretation of aerial imagery (Wickham et al., 2017;

Table 1

Assessment framework of site-specific, non-site specific, and spatial distribution metrics and their sources. Abbreviations are shown in parentheses. Number within brackets following reference is equation number from corresponding reference. Equations are replicated from their cited source. Terms are defined below, following Olofsson et al. (2014) as default, with deviations and additions noted specific to each reference.

Site-specific metric	Reference	Equation
Error Matrix	Olofsson et al. (2014) ^a [4]	$p_{ij} = W_i \frac{n_{ij}}{n_i}$
Overall Accuracy (OA)	Olofsson et al. (2014) [1]	$O = \sum_{j=1}^q p_{jj}$
User's Accuracy (UA)	Olofsson et al. (2014) [2]	$U_i = p_{ii}/p_{i\cdot}$
Producer's Accuracy (PA)	Olofsson et al. (2014) [3]	$P_j = p_{jj}/p_{\cdot j}$
Standard Error (SE) of OA	Olofsson et al. (2014) [5]	$\hat{V}(\hat{O}) = \sum_{j=1}^q W_i^2 \hat{U}_i (1 - \hat{U}_i) / (n_i - 1)$
Standard Error (SE) of UA	Olofsson et al. (2014) [6]	$\hat{V}(\hat{U}_i) = \hat{U}_i (1 - \hat{U}_i) / (n_i - 1)$
Standard Error (SE) of PA	Olofsson et al. (2014) [7]	$\hat{V}(\hat{P}_j) = \frac{1}{\hat{N}_{\cdot j}^2} \left[\frac{N_{\cdot j}^2 (1 - \hat{P}_j)^2 \hat{U}_j (1 - \hat{U}_j)}{n_j - 1} + \hat{P}_j^2 \sum_{i \neq j} N_{i \cdot}^2 \frac{n_{ij}}{n_i} \left(1 - \frac{n_{ij}}{n_i} \right) / (n_i - 1) \right]$
Quantity Disagreement (QD)	Pontius and Millones (2011) ^b [3]	$Q = \frac{\sum_{g=1}^J q_g}{2}$
Allocation Disagreement (AD)	Pontius and Millones (2011) [5]	$q_g = (\sum_{i=1}^J p_{ig}) - (\sum_{j=1}^J p_{gj}) $; $a_g = 2 \min (\sum_{i=1}^J p_{ig}) - p_{gg}, (\sum_{j=1}^J p_{gj}) - p_{gg} $
Ratio of AD/QD	–	$A = \frac{\sum_{g=1}^J a_g}{2}$
Linguistic Scale of Fuzzy Measures	Gopal and Woodcock (1994), Woodcock and Gopal (2000)	$\frac{A}{Q}$
Non-site specific metric	Reference	See text description.
Error matrix	Olofsson et al. (2014) [4]	Equation $p_{ij} = W_i \frac{n_{ij}}{n_i}$
Class Area	Olofsson et al. (2014) [9]	$\hat{p}_{\cdot k} = \sum_{i=1}^q W_i \frac{n_{ik}}{n_i}$
Standard Error (SE) of Area Estimate	Olofsson et al. (2014) [10]	$\hat{S}(\hat{p}_{\cdot k}) = \sqrt{\frac{\sum_i W_i^2 \frac{n_{ik} (1 - \frac{n_{ik}}{n_i})}{n_i - 1}}{\sum_i W_i^2 \frac{n_{ik} (1 - \frac{n_{ik}}{n_i})}{n_i - 1}}} = \sqrt{\frac{\sum_i W_i \hat{p}_{ik} - \hat{p}_{\cdot k}^2}{\sum_i W_i \frac{n_{ik} (1 - \frac{n_{ik}}{n_i})}{n_i - 1}}}$
Spatial distribution metric	Reference	Equation
Kolmogorov-Smirnov statistic (KS)	Riemann et al. (2010) ^c [1]	$D_{KS} = \max F(x) - G(x) $
Agreement Coefficient (AC)	Riemann et al. (2010) [5]	$AC = 1 - \frac{SSD}{SPOD}$
		$SSD = \sum_{i=1}^n (X_i - Y_i)^2$
		$SPOD = \sum_{i=1}^n (\bar{X} - \bar{Y} + X_i - \bar{X}) (\bar{X} - \bar{Y} + Y_i - \bar{Y})$
Systematic AC (AC _{sys})	Riemann et al. (2010) [9]	$AC_{sys} = 1 - \frac{SSD_s}{SPOD}$
		$SPD_s = SPD - SPD_u$
Unsystematic AC (AC _{uns})	Riemann et al. (2010) [10]	$AC_{uns} = 1 - \frac{SSD_u}{SPOD}$
		$SPD_u = \sum_{i=1}^n (X_i - \bar{X}_i) (Y_i - \bar{Y}_i)$

^a Where p_{ij} represents the proportion of area for the population that has map class i and reference class j ; W_i is the proportion of area mapped as class i ; from a population error matrix of q classes.

^b Where q_g is the quantity disagreement for an arbitrary category g , which differs from the definition of q in Olofsson et al. (2014).

^c Where $F(x)$ and $G(x)$ represent two samples with cumulative distribution functions. SSD is the sum of square difference, SPOD is the sum of potential difference, $\bar{X}$ and $\bar{Y}$ are the mean values of datasets X and Y , respectively.

Zimmerman et al., 2013) or human-interpretations of satellite-based imagery (Cohen et al., 2017; Cohen et al., 2010). Alternatively, reference data may be based on inventory datasets containing ground-based observations like those produced by the USDA Forest Service Forest Inventory and Analysis (FIA) program. Cost, data availability, operational constraints, or other factors often dictate the reference source. For many land-cover attributes, ground-based (in situ) observations are expected to be more accurate than image-based reference data at ground plot scale. Representativeness of a ground plot will depend on its size relative to map pixel spatial resolution, temporal and spatial registration between plot and pixel, and on the local variability of a map attribute. In contrast, one might design image interpreted reference data to better align with the majority condition at pixel scale, but its accuracy will be dependent on image characteristics and interpreter experience. Although reference classifications are assumed to be of higher quality than map classifications (Olofsson et al., 2014), reference data are rarely perfect, and remote sensing studies of land change can accrue large error from even small amounts of error in the ground reference data (Foody, 2010).

2.2. Accuracy assessment

We describe a map accuracy assessment framework, with metrics grouped into three categories: 1) site-specific, e.g., plot-pixel error matrices, 2) non-site specific, e.g., area estimates, and 3) spatial distribution, based on statistical comparisons across equal area polygon estimation units.

2.2.1. Site-specific

Standard metrics of site-specific accuracy include OA - the proportion of observations classified correctly (agreement), UA of each map class (the complement of commission error), and PA of each map class (the complement of omission error) (Congalton, 1991). These estimates are based on comparison of discrete map classes with corresponding reference data, tabulated in an error matrix, also referred to as a confusion matrix (Card, 1982). See Table 1 for a list of these, and all subsequent metrics and their abbreviations.

Assessment of map classification accuracy is confounded by uncertainty in both thematic aspects (e.g., land cover classification error), and geospatial aspects (e.g., location error), of both map and validation data (Cheng et al., 2001; Hagen, 2003). Some map classes are difficult to distinguish from one another due to their similar biophysical characteristics, leading to similar spectral reflectance patterns when maps are produced from optical imagery. Temporal misregistration may occur when map and reference data are not collected simultaneously and class changes occur at a particular location in either the ground or the imagery data between observation dates. Even when both map and reference observations are labeled correctly, errors in coordinate locations for either map or reference data can result in spatial misregistration between map and reference observations. Pontius and Millones (2011) proposed a method for separating map classification error into two sources: quantity disagreement, which relates to less than perfect match in class proportions, and allocation disagreement, which relates to the less than optimal match in spatial allocation of classes. Disagreement (the inverse of agreement) can be better understood by partitioning into quantity and allocation disagreement (Pontius and Millones, 2011).

2.2.2. Fuzzy sets

Some thematic classes that fall along a continuum, like forest age classes, may have map classification errors that are not ecologically important, e.g., 6–10 vs. 11–15 year old forest providing similar habitat quality for some wildlife species. In such classification systems, thematic uncertainty can be further analyzed using fuzzy membership approaches; fuzzy set is defined as a class with a continuum of grades of membership (Zadeh, 1965). The assumption underlying fuzzy set theory is that the transition from membership to nonmembership is seldom a

step function but instead is better represented as a gradual but specifiable change from membership to nonmembership (Gopal and Woodcock, 1994). Map-validation mismatches are not completely wrong when they have some similarity with other classes; some mismatches are more or less acceptable than others. Thus, a binary ‘correct/incorrect’ map-validation choice is overly conservative, failing to accommodate class pairs that are similar enough to be counted as ‘correct’ for some map applications. Various methods have been described for incorporating fuzzy sets into estimates of classification accuracy (Foody, 1992; Gopal and Woodcock, 1994; Sarmiento et al., 2015; Stehman et al., 2007; Woodcock and Gopal, 2000; Zadeh, 1963, 1965). The good practices recommended in Olofsson et al. (2014) accommodate fuzzy set classification by allowing matches between a map label and either primary or secondary reference labels.

2.2.3. Linguistic scales for fuzzy sets

Gopal and Woodcock (1994) describe a method for applying fuzzy sets by assigning a membership scale to each class-class pair and then defining which membership classes to include as ‘correct’ observations. To create the membership scale, Gopal and Woodcock (1994) defined five linguistic classes: (1) Absolutely Wrong, (2) Understandable but Wrong, (3) Reasonable or Acceptable Answer, (4) Good Answer, and (5) Absolutely Right. Using these five classes, Gopal and Woodcock (1994) defined two functions for assigning class-class combinations as being correct:

- MAX function defines a match for the linguistic class with maximum correct membership, in this case, including only those classes that are absolutely right.
- RIGHT function defines a match when the linguistic class is equal to or greater than a pre-defined threshold.

Woodcock and Gopal (2000) demonstrated this approach when assessing accuracy of a map with six vegetation classes in California, USA. They refer to the MAX and RIGHT functions as “fuzzy measures”, and assigned MAX to only class 5, i.e., diagonal cells used to estimate percent correctly classified in a traditional error matrix, and RIGHT to any linguistic score greater than or equal to 3, with experts assigning a rating for each possible map label. Examples of other applications of the fuzzy set approach include Gill et al. (2000), Laba et al. (2002), and Wickham et al. (2017). Likewise, Mayaux et al. (2006) applied a “similarity matrix” approach to assessing accuracy of the 2000 Global Land Cover product. In that application, off-diagonal errors were quantified for thematic proximity in the legend definition, and the similarity values were added to the diagonal of the matrix.

2.2.4. Non-site specific

2.2.4.1. Area estimates. Utility of thematic map data for supporting land management decisions depends not only upon site-specific accuracies (OA, UA, PA), and their corresponding SE's, but also upon the bias and precision of per-class area estimates. If map-based area estimates are produced by simply counting pixels in a wall-to-wall map, such estimates have zero sampling error (perfect precision) because all population elements (pixels) are included. However, estimates obtained by counting pixels likely are biased because individual pixel predictions have model errors that propagate to population parameters based on summaries of those pixels. To overcome these issues, Olofsson et al. (2014) recommends that area estimates not be based on pixel counts within map classes, but instead, should be based on the reference classification. Reference data are assumed to have superior quality than is attributable to the map classification and given a probability sampling design (like simple random sampling), reference-based estimators are unbiased. Precision of reference-based area estimates may be lower than desired due to budgetary or operational constraints, but, if constraints

are addressed, can be improved either by increasing reference sample size; or by applying a post-stratified estimator, which may be obtained from the error matrix via Eq. (10) in Olofsson et al. (2014). For simple random sampling, the post-stratified estimator generally has better precision than other estimators evaluated by Stehman (2013). If area estimation is a primary objective, stratified sampling is also commonly implemented (Stehman, 2013), although this approach may preclude the use of existing inventory data obtained under other sampling designs.

2.2.4.2. Comparing area estimates. A common approach for assessing non-site specific accuracy is comparison of map-based estimates with reference-based estimates, e.g., from field inventory plots (Häme et al., 2001; Kennedy and Bertolo, 2002; Nelson, 2005; Päivinen et al., 2001; Schuck et al., 2003). Estimates of uncertainty (e.g., confidence intervals) are available for estimates produced from probability sample-based inventory datasets, but traditionally not from modeled geospatial datasets (McRoberts, 2011). Therefore, a common approach is to compare map-based population estimates to confidence intervals surrounding probability sample-based estimates.

2.2.5. Spatial distribution

2.2.5.1. Distribution of map error. Comparisons of map and reference estimates among geographic subareas reveal statistical and spatial patterns of consistency or inconsistency. However, such comparisons are confounded when estimation units have varying sample sizes of reference data, resulting in confidence intervals of varying widths, which precludes comparisons among estimation units. Riemann et al. (2010) overcomes this issue of variable sample sizes when presenting a suite of comparative accuracy metrics for assessing the type, magnitude, frequency, and location of errors in modeled geospatial datasets. These metrics require samples of sufficient and consistent size across a spatially balanced design and are computed for an array of subareas, or estimation units. The selection of number, area, and geographic delineation of subareas should reflect the intended use of the map. One could apply this approach to geographic units representing pre-established analysis units or management areas, e.g., administrative or ownership boundaries, but such polygons typically vary in area, limiting their application in some statistical approaches. When selecting or defining subareas, one compromises between having sufficient reference data within each subarea and having a sufficient number of subarea estimates to calculate accuracy metrics and display spatial variability (Riemann et al., 2010).

Among other accuracy metrics, Riemann et al. (2010) demonstrated the Kolmogorov-Smirnov statistic (KS; Feller, 1948) and the non-parametric Agreement Coefficient (AC; Ji and Gallo, 2006). These metrics are based on comparisons between data collected on probability sample-based inventory plots and a corresponding sample of modeled predictions from map pixels. In this context, KS is a metric summarizing the maximum distance between modeled and reference empirical cumulative distribution functions (ECDFs), treated as an index of agreement rather than as a traditional statistical test. AC provides a metric summarizing the level of agreement between the two datasets (range: 0–1), which is affected by both systematic (bias) and unsystematic (random) error, estimated as Systematic AC (AC_{sys}), and Unsystematic AC (AC_{uns}), respectively.

In an application of Riemann et al. (2010) methods across the eastern United States, Riemann et al. (2012) used hexagonal arrays for representing equal land areas, with each hexagon containing approximately equal numbers of probability sample-based inventory plots (and corresponding map pixels) – with the exception of “edge” hexagons which were excluded from analyses due to their smaller and variable polygon areas. To overcome the challenges of dealing with edge hexagons, Lister and Scott (2009) developed a “space-filling curves” method for using a

geographic information system (GIS) and fractals to create a sampling frame from which spatially balanced sample locations can be selected from each equal-area estimation unit, leaving no portion of a study area unsampled. This method compared favorably with the Generalized Random Tessellation Stratified Design method described in Stevens Jr. and Olsen (1999) and Stevens and Olsen (2004) but with added advantages of relative transparency, intuitive clarity, and ease of implementation within GIS (Lister and Scott, 2009).

3. Application of the assessment framework

In the following example applications, we apply all components of the assessment framework described above (Table 1) to obtain a suite of accuracy and area estimates for a thematic map of early successional forest and other land cover classes across the western Great Lakes region, USA. We report and compare results for individual metrics, then present insights provided by the assessment framework.

3.1. Study area and data

The example application was conducted within portions of Bird Conservation Regions (BCRs) 12 (Boreal Hardwood Transition) and 23 (Prairie Hardwood Transition) (<http://nabci-us.org/bcrs.htm>) within the western Great Lakes states of Minnesota, Wisconsin, and Michigan, USA (Fig. 1). BCRs define ecologically distinct regions with similar bird communities, habitats, and resource management issues. The combination of BCRs and States delineating this study area corresponds with the Upper Great Lakes Young Forest Initiative (YFI), which promotes conservation of American woodcock (*Scolopax minor*) habitat within the region (Cooper, 2008; Kelley Jr. et al., 2008). Based on analyses of FIA data, about 13.7% of forest land across Minnesota, Wisconsin, and Michigan is 0–20 years of age (Miles, 2015), a stage referred to as early successional forest (ESF). During the five-year period 2008–2012, the average annual rate of land use change among all classes was 0.67%; forest land area increased at an average annual rate of 0.63% (Miles, 2015). During this same period, average annual rate of tree cover change resulting from tree harvest, tree damage, or canopy disturbance was 3.4%, which is approximately five times the rate of forest land use change (Miles, 2015). Statewide conditions and trends are similar to those within BCR portions of each state because BCRs 12 and 23 are based on ecological conditions that coincide approximately with the predominantly forested regions of each state.

The 43.3 million ha study area, excluding Great Lakes water bodies, contains approximately 481 million pixels and 16,485 FIA reference sample plots. Fig. 2 portrays the spatial distribution of water, nonforest, and forest classes, and subcategories of forest by type and age classes.

See Tavernia et al. (2016) for detailed description and historical context of the study area.

3.1.1. Thematic map classification of land cover, forest type, and forest age

Using a multi-source map integration approach, we created a 30-m spatial resolution geospatial dataset with pixel labels consisting of forest age and land cover class combinations (Garner et al., 2016b; Garner et al., 2015; Tavernia et al., 2016). A forest canopy disturbance Vegetation Change Tracker (VCT) algorithm (Huang et al., 2010), growing season Landsat Time Series Stacks (1987–2010), and winter image enhancement (VCTw) described in Stueve et al. (2011) were used to model classes of forest, nonforest, and water. Within forest class, stand age was inferred from year of canopy disturbance (Thomas et al., 2011) and aggregated into four 5-year age classes (1–5, 6–10, 11–15, 16–20 years) and one class of persistent forest (>20 years). Prior analyses suggested that OA of NLCD2011 maps with categories aggregated into three classes (forest, nonforest, water) was higher than OA for a comparable map of three classes produced from the VCTw algorithm. Therefore, NLCD2011 (Homer et al., 2015) was selected as a land cover ‘basemap’ and VCTw age classes were assigned to corresponding

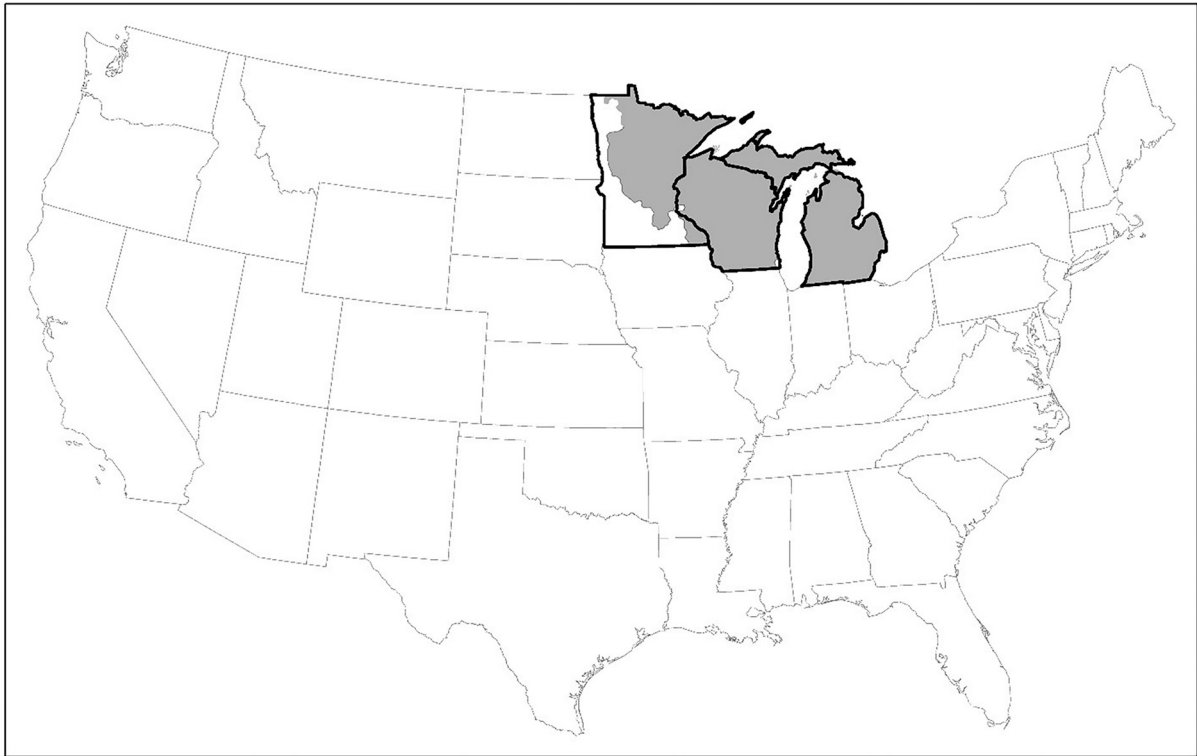


Fig. 1. Study area, comprising Bird Conservation Regions 12 and 23 (gray shading) within the states of Michigan, Wisconsin, and Minnesota, USA.

NLCD2011 pixels in deciduous, evergreen, and mixed forest and woody wetlands classes (Garner et al., 2016b; Garner et al., 2015).

NLCD2011 forest is defined as areas dominated by trees generally greater than 5 m tall, and greater than 20% of total vegetation cover (<https://www.mrlc.gov/data/legends/national-land-cover-database-2011-nlcd2011-legend>). Forest is classified as deciduous if more than 75% of the tree species shed foliage simultaneously in response to seasonal change, as evergreen if more than 75% of the tree species maintain their leaves all year, and as mixed if neither deciduous nor evergreen species are greater than 75% of total tree cover. NLCD2011 woody wetlands are areas where forest or shrubland vegetation accounts for greater than 20% of vegetative cover and the soil or substrate is periodically saturated with or covered with water.

Narrow forest slivers of 1 or 2 pixel width with no corresponding age data were assigned to the age class of adjacent forest pixels; larger patches of forest pixels with no age data were assigned to >20 year old age class. The rationales for this decision were that: 1) larger patches likely represented distinct conditions, not edge slivers resulting from spatial misregistration artifacts (addressed above), 2) >88% of forest land area is >20 years of age, based on reference data, so older forest age class is the most likely, and 3) we chose to err on the side of omission (reducing commission) when assessing ESF as potential wildlife habitat.

We reclassified almost half of NLCD2011 pixels of class 52-shrub/scrub and 71-grassland/herbaceous to “other forest” class where corresponding VCTw pixels were classified as forest, under the assumption that such pixels represent recently disturbed, regenerating forest (Wickham et al., 2017). Remaining NLCD2011 pixels of classes 52 and 71 not associated with VCTw forest pixels were aggregated into non-forest class, as were all other NLCD2011 nonforest pixels, with the exception of class 11-open water, which was retained as a separate class. The resulting dataset contains 23 unique thematic classes, consisting of water, nonforest, 20 combinations of forest type and age, and one other forest class for which neither type nor age were modeled. We applied a four-pixel minimum mapping unit (MMU) to all classes within the dataset; each pixel within a patch containing fewer than four pixels was

independently dissolved into its geographically nearest neighboring patch using ‘clump’ and ‘eliminate’ functions within Erdas Imagine software package. A hierarchical classification system was applied to each pixel, emphasizing either general land cover, forest type, or forest age. This dataset was described and a preliminary assessment was conducted previously using a subset of procedures used in this study (Garner et al., 2016a; Garner et al., 2015). Hereafter, we treat the hierarchical components as three map classifications (Fig. 2).

3.1.2. Reference dataset – FIA ground inventory samples

The FIA plot data (Woudenberg et al., 2010) provided existing reference data for a spatially balanced probability-based sample of sufficient intensity and spatial and thematic resolution to satisfy the recommendations of Olofsson et al. (2014). FIA employs a probability-based sample of permanent, fixed-radius cluster plots that are re-measured over 5–10 year cycles. At least one plot is randomly located within each ~2400-ha hexagon in a hexagonal sampling array (Reams et al., 2005). A subset of sample plots, geographically distributed by interpenetrating panels, are measured each year, typically at a rate of 14–20% in the eastern United States and 10% in the western United States, resulting in full inventory cycles of five-seven and ten years, for the East and West, respectively. At the time of this study, all FIA data were measured on a five-year cycle within the study area. Nonforest FIA plots are not field-visited in the East, but land use attributes at the plot center are recorded based on remote sensing approaches. Forested FIA plots are field visited, and established as a set of four subplots within a cluster plot design. Each subplot is circular with fixed radius of 7.32 m. Three peripheral subplots surround one central subplot, with subplot centers located at azimuths of 0, 120, and 240, and subplot centers separated by 36.68 m (120 ft) from the central subplot. FIA typically refers to the central subplot as subplot 1, which is the only subplot where Global Positioning System (GPS) coordinates are recorded at subplot center; the centers of three peripheral subplots are determined using field measurements of distance and azimuth from the central subplot center point.

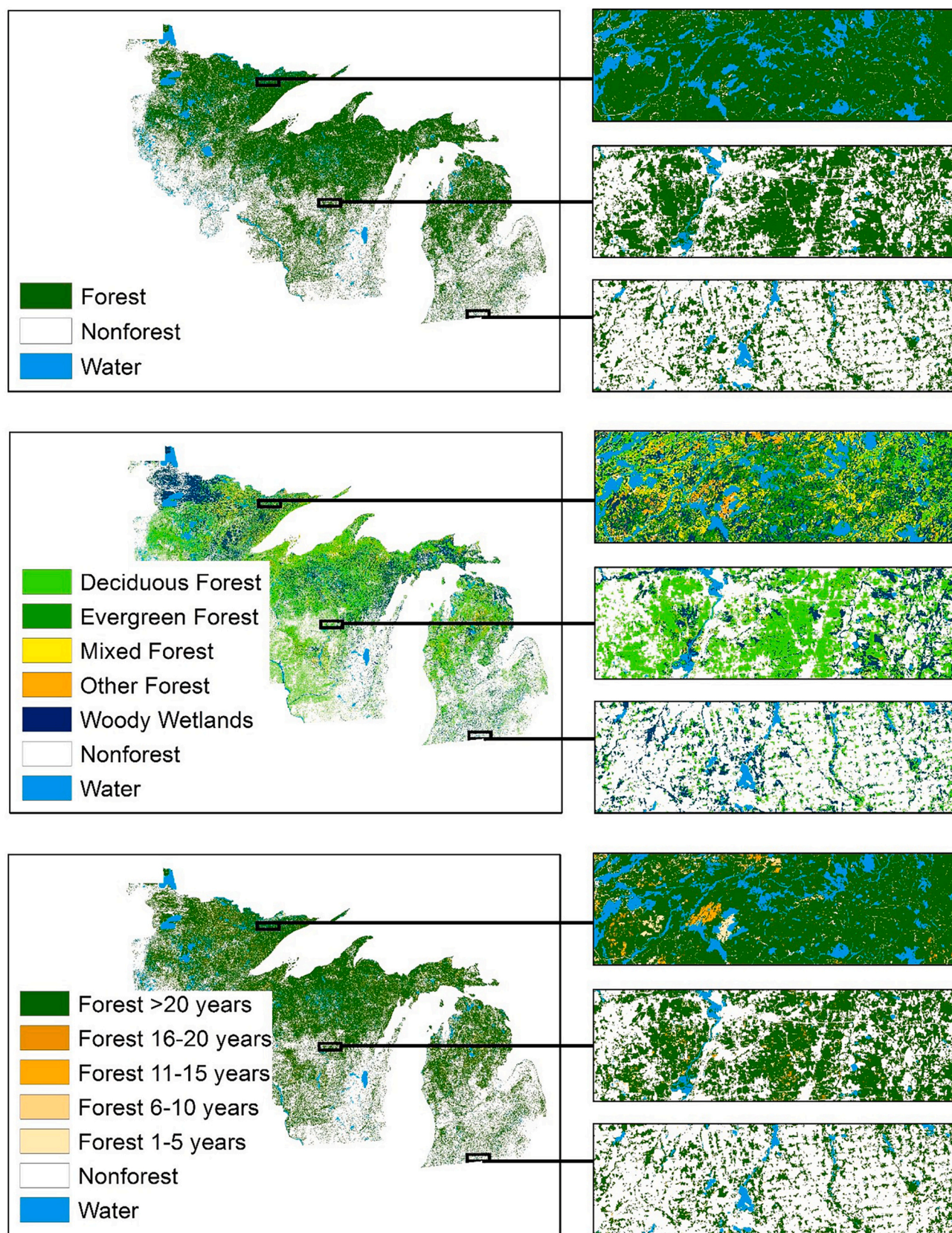


Fig. 2. Map classification hierarchy levels emphasizing land cover (upper), forest type (middle), and forest age (lower) in Bird Conservation Regions 12 and 23 within the states of Michigan, Wisconsin, and Minnesota, USA, with a set of three rectangular insets for each classification from northern Minnesota, central Wisconsin, and southern Michigan.

FIA sampling intensity varies among inventory years, and varies among and within three states in this study area. We constrained our reference data to include only base intensity (one plot per approximately 2400 ha) plots to remove the effects of variable sample intensity on resulting population estimates and associated estimates of uncertainty in non-site specific and comparative accuracy assessments. FIA plots from Minnesota, Wisconsin, and Michigan collected in the 2008–2012 inventory cycle – this period spans 2010, which is the most recent year of Landsat imagery included in the VCTw-based forest age map – were overlaid on the ESF map; plots falling in one of the Great Lakes or outside the study area were excluded. To improve spatial alignment between FIA plots and pixels, we implemented the method of McRoberts (2011) by using fixed radius circular central subplots (0.0169 ha) having a single condition, resulting in 16,485 reference plots. McRoberts et al. (2018) reported that four subplots of the same FIA plot are largely correlated, meaning little information is lost by considering only the central subplot, and using only the central subplot resulted in less residual variation around model predictions than when using all four subplots separately.

FIA land use classes were assumed to represent land cover reference classes for this application, even though land use and land cover are not synonymous (Comber, 2008a, 2008b). Reference plots were labeled as being one of three land cover classes: forest, nonforest, or water. Each forest plot was assigned a reference label corresponding to one of four NLCD2011 forest type classes. Forest plots with hydric physiographic class code (FIA attribute PHYSCLCD) were assigned to woody wetlands class, and mesic and xeric forest plots were assigned to the deciduous, evergreen, or mixed forest class based on tree species composition. NLCD2011 definitions for these three forest classes are based on minimum tree cover thresholds. Because tree canopy cover data were not available for FIA reference data we computed an importance value (IV) for each plot, scaled from 0 to 100, from an equally weighted combination of tree density and basal area for both tree types: conifer and deciduous. Next, we chose a threshold of $IV = 75$ to determine whether a plot should be labeled as deciduous or evergreen class; if neither tree type's IV exceeded this threshold, plots were assigned to the mixed forest class. IV could not be computed for plots having no recorded trees with diameter at breast height (dbh; 1.37 m) ≥ 12.7 cm; these plots were assigned to the 'other forest' type class. We assumed such plots represented recently disturbed, regenerating forest canopy – hence no large trees, similar to the definition for 'other forest' pixels described above for the geospatial dataset.

Forest age classes were based on field crew assignment of age (FIA attribute FLDAGE), adjusted to account for differences between year of plot measurement and last year of imagery in the LTSS used to process VCTw (Age – Measurement Year – Final Image Year), then aggregated into 5-year classes of 1–5, 6–10, 11–15, 16–20, and persisting forest class for age > 20 years. For example, a plot measured in year 2012, with age 7 years, would be relabeled as age 5 years ($7 - (2012 - 2010) = 5$), and assigned to class '1–5 years'. Nonstocked forest plots (having FIA FLDAGE = 0) were assigned to the 1–5 year old age class under an assumption that most nonstocked forest in this study area has forest structural characteristics typical of post-canopy disturbance, i.e., ESF.

3.2. Accuracy assessment

Following the example of FIA reference data used in Thomas et al. (2011), we assumed that the FIA reference classification described above was considerably more accurate than each of three map classifications being evaluated, per recommended good practices in Olofsson et al. (2014). Map accuracy and area estimates were assessed for three maps: general land cover classes, forest type classes, and forest age classes (Fig. 2).

3.2.1. Site-specific

Site-specific tests were conducted following eight good practices recommended in Olofsson et al. (2014), expanded to include additional metrics listed in Table 1.

3.2.1.1. Error matrices. We produced error matrices utilizing the FIA sample of base intensity, single-condition central subplots as reference plots. Counts of plot-pixel class-class assignments were arranged into a matrix with map data in rows and reference data in columns and one matrix was produced for each of the maps. Counts in these matrices were used to estimate weighted proportions for every cell in corresponding error matrices, with cell weighting based on the proportion of area mapped in each map class (Olofsson et al., 2014; Eq. (4)).

3.2.1.2. Overall, user's, and producer's accuracies. Estimates of OA, UA, and PA were obtained using the error matrices described above, and Olofsson et al. (2014) Eqs. (1), (2), and (3), respectively, with the sample-based estimator of matrix cell proportions from Olofsson et al. (2014) Eq. (4) substituting for the population estimators within formulae (1), (2), and (3). Standard errors (SE) of accuracy estimates were obtained using Olofsson et al. (2014) Eqs. (5), (6), and (7) for OA, UA, and PA, respectively, reported as 95% confidence intervals ($\pm 1.96 \times SE$).

3.2.1.3. Quantity disagreement and allocation disagreement. Total disagreement is computed as 1 minus OA, which was further partitioned into quantity disagreement and allocation disagreement following eqs. (3) and (5), respectively, in Pontius and Millones (2011), based on estimated weighted proportions in error matrices, as described above for good practices in Olofsson et al. (2014).

3.2.1.4. Fuzzy membership of forest age classes. We adapted the linguistic approach described in Gopal and Woodcock (1994) and applied in Woodcock and Gopal (2000) to assign fuzzy measures that correspond to each cell in the forest age error matrix (Table 2); fuzzy measures were not applied to error matrices of general land cover or forest types. Although fuzzy sets can apply to both map and reference data, we applied a conservative approach by associating fuzzy measures only with reference labels per Olofsson et al. (2014); map labels retained their original "hard classes". Furthermore, we adapted the fuzzy measure concept to rank similarity among ordinal classes rather than correctness within each class. We produced a fuzzy membership matrix with seven rows and seven columns, corresponding with the 7×7 error matrix for the forest age classification system (Table 3). This matrix was used as a guide for collapsing error matrix values into classes corresponding with those from Table 2. We retained the 'MAX' class, but separated the 'RIGHT' class of Gopal and Woodcock (1994) into two categories to represent differing levels of similarity among different pairs of 5-year age classes, described below: RIGHT3 and RIGHT4 (Tables 2, 3). This approach is similar to Mayaux et al. (2006), but rather than quantifying similarity, our fuzzy matrices were simplified into 'all or nothing' acceptance of fuzzy matches via two alternative linguistic scales.

Adjacent 5-year age classes (e.g., 6–10 years vs. 11–15 years) were assumed to have high similarity with each other and were assigned an index fuzzy measure value of 4. Young forest age classes separated by one class from another 5-year forest age class (e.g., 6–10 years vs. 16–20 years) were labeled as 3, under the assumption that they provided a "reasonable or acceptable" answer under the linguistic scale. Although the oldest 5-year age class (16–20 years) was assumed to be highly similar to persisting forest of age 21–25 years, a large majority of persisting forest is older than 25 years, which we assumed to be less similar to 16–20 year class; therefore, this pairing was assigned an index fuzzy

Table 2

Linguistic scale and fuzzy reference labeling protocol, adapted from Gopal and Woodcock (1994). Descriptions are quoted from Gopal and Woodcock (1994). Rectangles indicate which linguistic classes are included under two labeling systems.

Linguistic score	Linguistic class	Description	Label (Gopal and Woodcock, 1994)	Label (this study)
1	<i>Absolutely Wrong</i>	This answer is absolutely unacceptable.		
2	<i>Under-standable but Wrong</i>	Not a good answer. There is something about the site that makes the answer understandable but there is clearly a better answer. This answer would pose a problem for users of the map.		
3	<i>Reasonable or Acceptable Answer</i>	Maybe not the best possible answer but it is acceptable; this answer does not pose a problem to the user if it is seen on the map.	RIGHT	RIGHT3
4	<i>Good Answer</i>	Would be happy to find this answer given on the map.		RIGHT4
5	<i>Absolutely Right</i>	No doubt about the match.	MAX	MAX

Table 3

Linguistic scale of fuzzy measures (Gopal and Woodcock, 1994; Woodcock and Gopal, 2000): 1 - Absolutely Wrong, 2 - Understandable but Wrong, 3 - Reasonable or Acceptable Answer, 4 - Good Answer, 5 - Absolutely Right.

Map	Reference						
	Forest >20 years	Forest 16–20 years	Forest 11–15 years	Forest 6–10 years	Forest 1–5 Years	Nonforest	Water
Forest >20 years	5	3	2	2	2	1	1
Forest 16–20 years	3	5	4	3	2	2	1
Forest 11–15 years	2	4	5	4	3	2	1
Forest 6–10 years	2	3	4	5	4	2	1
Forest 1–5 years	2	2	3	4	5	3	1
Nonforest	1	2	2	2	3	5	1
Water	1	1	1	1	1	1	5

measure value of 3. We also assigned a value of 3 to the pairing of 1–5 year forest and nonforest class because the youngest forest age class often has ground vegetation with minimal tree canopy cover during regeneration, similar to grassland or agricultural vegetation that predominates the nonforest class within this study area. Water was assumed to be highly dissimilar from every terrestrial class, so observations that paired water with any other class were labeled 1. Likewise, nonforest was assumed to be highly dissimilar from persisting forest age class, which also resulted in a label of 1. All other class pairs received a similarity index value of 2 (Table 3).

We produced three complementary estimates of UA and PA, one using the traditional error matrix for each of our 3 maps, synonymous with including only class 5 fuzzy measures (MAX). For age class, we also produced one including all diagonals in class 5 plus class 4 (RIGHT4), and one including classes 5, 4, and 3 (RIGHT3).

3.2.2. Non-site specific

3.2.2.1. Area estimates. Non-site specific accuracy was assessed by comparing per-class map-based estimates from pixel counts with reference-based estimates from sample points. Reference-based area estimates and corresponding estimates of standard errors were based on FIA reference data, post-stratified by map class areas, using Olofsson et al. (2014) Eqs. (9) and (10), respectively, with error estimates reported as 95% confidence intervals. The variance estimator for the post-stratified estimator is the same as that used for a stratified sampling design based on a conditional inference argument, where the conditioning is on the observed sample sizes in each (post)stratum (Särndal et al., 1992, Eqs. (7), (10), (13)). Map-based estimates were compared to 95% confidence intervals surrounding reference-based estimates.

3.2.3. Spatial distribution

We adapted the suite of comparative accuracy metrics from Riemann et al. (2010) for application to categorical datasets by estimating the

percentage of each thematic class within equal-area-sized polygons. Prevalence of national and state boundaries and Great Lakes shorelines resulted in a relatively large proportion of partial hexagons along edges of our study area, resulting in unequal sample sizes and loss of comparability between estimation units, so we implemented a space-filling curves approach (Lister and Scott, 2009), as described above, to produce equal-area estimation units that encompass all inland area, excluding Great Lakes waters, resulting in 101 polygons, each of approximately 433,000-ha. Equal-area polygon borders were constrained to match boundaries of States (MI, WI, MN) and BCRs (12,23), enabling analyses for sub-populations.

Reference- and map-based area estimates were produced for each equal-area polygon, for each thematic class, for each of the three maps. Reference estimates for each estimation unit were obtained using the FIA reference plot data, under an assumption of simple random sampling, post-stratified by class area, with error estimates reported as 95% confidence intervals for each equal-area polygon. A sample of pixels that corresponded with the FIA reference plot locations within each equal area polygon were used to produce map sample-based area estimates, also under an assumption of simple random sampling (but not post-stratified), following Riemann et al. (2010). Per-polygon, per-class reference- and map-based estimates were used to estimate KS, AC, AC_{sys} , and AC_{uns} , using formulae (1), (5), (9), and (10), respectively, in Riemann et al. (2010). The KS distance quantifies the maximum difference between each map- and reference-based ECDF. The inclusion of AC_{sys} , and AC_{uns} allows for further partitioning of AC. Higher values of AC_{uns} (random or unsystematic uncertainty) or AC_{sys} (bias, or systematic uncertainty) indicate less uncertainty whereas lower values indicate more uncertainty associated with each subcomponent of AC. When classes contain no reference plots or no sample map pixels within a class, no comparison could be made due to absence of confidence intervals surrounding reference-based estimates. We assigned special cartographic categories to portray these two situations in maps portraying spatial distributions of errors

4. Results

4.1. Accuracy assessment

4.1.1. Site-specific

4.1.1.1. Error matrices. Error matrices for each of the three maps are presented in [Tables 4, 5, and 6](#), with accompanying estimates of OA, PA, UA, area, and standard errors of accuracy and area (reported as $\pm 95\%$ CI's).

4.1.1.2. Overall, user's, and producer's accuracies. Overall accuracy for the land cover map was 90% (± 0.5), with PA and UA both ranging among map classes from about 88 to 92% ([Table 4](#)). The forest type classification had an OA of 73% (± 0.6), with per-class accuracies varying substantially among forest types ([Table 5](#)). Among forest classes, 'other forest' performed worst (PA: 16%, UA: 22%), and woody wetlands (PA: 77%) and deciduous forest (UA: 89%) performed best. Overall accuracy was 85% (± 0.5) for the forest age classification ([Table 6](#)). Persisting forest accuracies were high (PA: 91%, UA: 84%), but accuracies for 5-year age classes of young forest were low, ranging from 16 to 28% PA and 36–44% UA. For comparison, the four 5-year age classes in [Table 6](#) were aggregated into one young forest age class (1–20 years), resulting in estimated OA increasing only marginally to 86%; per-class accuracies for the 1–20 year-old age class increased substantially for both PA (61%) and UA (90%), compared to accuracies for individual 5-year age classes reported above.

4.1.1.3. Quantity disagreement and allocation disagreement. Disagreement for three classifications are reported in [Table 7](#). Allocation disagreement (AD) was substantially larger than quantity disagreement (QD) for all three classifications, with AD/QD ratios of 8.0 for land cover, 3.1 for forest type, and 2.6 for forest age classifications.

4.1.1.4. Fuzzy membership of forest age classes. Classification accuracies for three linguistic scales of forest age are shown in [Table 8](#). OA varied slightly among three linguistic scales; PA and UA for water, nonforest, and > 20 year-old forest showed negligible differences, but per-class accuracies for 5-year age classes varied substantially when incorporating fuzzy measures of linguistic scales. Most of the improvement for the middle two 5-year age classes resulted when adapting MAX (traditional error matrix) to the RIGHT4 linguistic scale, with only moderate additional gain when moving between the RIGHT4 and RIGHT3 linguistic scales. In contrast, the youngest and oldest 5-year age classes had only minor improvements when moving from MAX to RIGHT4, and

substantial additional gains when moving between RIGHT4 and RIGHT3 linguistic scales ([Table 8](#)).

4.1.2. Non-site specific

Reference-based stratified estimates of class area with 95% confidence interval half-widths ($SE * 1.96$), and map-based estimates from total pixel class area and from pixel sample-based estimates corresponding to reference plot locations are reported for land cover ([Table 4](#)), forest type ([Table 5](#)), and forest age ([Table 6](#)). Percent difference between map class areas and reference-based estimates, and between map sample-based estimates and reference-based estimates ranged from -3 to 2% for land cover classes, -29 to 72% for forest types, and -64 to 9% for forest age classes ([Table 9](#)). Water was the only class where map area fell within the 95% confidence interval surrounding the corresponding reference-based estimate. Map sample pixel-based estimates fell outside 95% confidence intervals surrounding reference-based estimates for every class

4.1.3. Spatial distribution

Reference-based area estimates for all equal-area polygons were greater than zero for forest and nonforest land cover classes, deciduous forest type, and persistent forest age class; all other map classes had one or more polygons with reference-based per-class area estimates of zero. Consequently, ordered distributions for some classes portray gray error bars surrounding reference-based estimates spanning the entire range of estimates, but other classes contained no reference plots in their respective classes for specific polygons resulting in an estimate of area = 0 and no error bars. Some of these cases also had map sample-based estimates of class area = 0, and some had map sample-based estimates >0 ([Fig. 3](#)).

Based on the map sample data, the maximum areal extent of any single land cover class within a 433,000-ha polygon was about 400,000 ha for forest, about 375,000 ha for nonforest, and about 100,000 ha for water class, indicating that both forest and nonforest classes can be locally dominant, but in different polygons, whereas water class (excluding open water of Great Lakes) never comprises a majority of surface area in any polygon. For all three land cover classes, $KS < 0.08$, indicating that reference- and map sample-based estimates track each other closely ([Table 9](#)). AC, AC_{sys} , and AC_{uns} values exceed 0.97 for forest and nonforest classes, and exceed 0.87 for water class, indicating very little disagreement between map sample- and reference-based estimates of land cover classes from either systematic or unsystematic sources. Values for the water class are slightly lower ([Table 9](#)). The majority of map sample-based area estimates fell within 95% confidence intervals associated with the reference-based estimates at 433,000-ha scale for forest and nonforest land cover classes, but only for about

Table 4

Overall, producer's, and user's accuracies; and estimates of class area from all map pixels, from a sample of map pixels corresponding to reference plots (summary columns), and from reference plots post-stratified by map area (summary row), with $\pm 95\%$ Confidence Interval half-widths (in parentheses), for three general land cover classes – forest (F), nonforest (NF), and water (W), based on combined NLCD2011-VCTw map within Bird Conservation Regions 12 and 23 in Minnesota, Wisconsin, and Michigan, USA.

Map class	Reference class			Map total	Map area (millions of ha)	Map sample estimate of area (millions of ha)	Sample plots (#)	User's accuracy
	F	NF	W					
F	0.465	0.050	0.003	0.518	22.408	21.539	8207	0.897 (0.007)
NF	0.041	0.391	0.003	0.435	18.798	19.479	7422	0.900 (0.007)
W	0.001	0.004	0.042	0.048	2.059	2.247	856	0.889 (0.021)
Reference total	0.507	0.446	0.048	1				
Stratified estimate of area (millions of ha)	21.920 (0.194)	19.280 (0.197)	2.065 (0.065)					
Reference Plots (#)	8081	7554	850				16,485	
Producer's accuracy	0.917 (0.005)	0.878 (0.007)	0.886 (0.021)				Overall accuracy	0.898 (0.005)

Table 5

Overall, producer's, and user's accuracies; and estimates of class area from all map pixels, from a sample of map pixels corresponding to reference plots (summary columns), and from reference plots post-stratified by map area (summary row), with $\pm 95\%$ Confidence Intervals (in parentheses), for five forest types – deciduous (FD), evergreen (FE), mixed (FM), other (FO), and woody wetlands (WW); and two general land cover classes – nonforest (NF) and water (W) based on combined NLCD2011-VCTw map within Bird Conservation Regions 12 and 23 in Minnesota, Wisconsin, and Michigan, USA.

Map class	Reference class							Total	Map area (millions of ha)	Map sample estimate of area (millions of ha)	Sample plots (#)	User's accuracy
	FD	FE	FM	FO	WW	NF	W					
FD	0.189	0.008	0.012	0.013	0.007	0.025	0.001	0.255	11.036	10.637	4053	0.741 (0.014)
FE	0.006	0.022	0.008	0.002	0.004	0.002	0.000	0.043	1.863	1.740	663	0.511 (0.038)
FM	0.017	0.006	0.008	0.001	0.003	0.001	0.000	0.037	1.601	1.480	564	0.222 (0.034)
FO	0.011	0.005	0.002	0.006	0.001	0.001	0.000	0.025	1.095	1.094	417	0.221 (0.040)
WW	0.037	0.011	0.011	0.006	0.070	0.021	0.002	0.158	6.813	6.588	2510	0.447 (0.020)
NF	0.018	0.005	0.002	0.008	0.007	0.391	0.003	0.435	18.798	19.479	7422	0.900 (0.007)
W	0.000	0.000	0.000	0.000	0.000	0.004	0.042	0.048	2.059	2.247	856	0.889 (0.021)
Total	0.278	0.059	0.044	0.036	0.091	0.446	0.048	1				
Stratified estimate of area (millions of ha)	12.010 (0.227)	2.535 (0.141)	1.891 (0.133)	1.537 (0.121)	3.953 (0.163)	19.274 (0.196)	2.065 (0.065)					
Reference plots (#)	4427	927	690	576	1461	7554	850				16,485	
Producer's accuracy	0.681 (0.011)	0.376 (0.025)	0.188 (0.027)	0.157 (0.027)	0.770 (0.020)	0.878 (0.007)	0.886 (0.021)				Overall accuracy	0.729 (0.006)

Table 6

Overall, producer's, and user's accuracies; and estimates of class area from all map pixels, from a sample of map pixels corresponding to reference plots (summary columns), and from reference plots post-stratified by map area (summary row), with $\pm 95\%$ Confidence Intervals (in parentheses), for five forest age classes (F) and two general land cover classes – nonforest (NF) and water (W) based on combined NLCD2011-VCTw map within Bird Conservation Regions 12 and 23 in Minnesota, Wisconsin, and Michigan, USA.

Map class	Reference class							Total	Map area (millions of ha)	Map sample estimate of area (millions of ha)	Sample plots (#)	User's accuracy
	F > 20	F 16–20	F 11–15	F 6–10	F 1–5	NF	W					
F > 20 years	0.407	0.009	0.009	0.006	0.005	0.047	0.003	0.487	21.084	20.180	7689	0.836 (0.008)
F 16–20 years	0.002	0.003	0.001	0.000	0.000	0.000	0.000	0.006	0.251	0.265	101	0.436 (0.097)
F 11–15 years	0.002	0.001	0.003	0.001	0.000	0.001	0.000	0.007	0.316	0.297	113	0.407 (0.091)
F 6–10 years	0.002	0.000	0.002	0.003	0.001	0.001	0.000	0.009	0.391	0.407	155	0.361 (0.076)
F 1–5 years	0.003	0.000	0.001	0.001	0.003	0.001	0.000	0.009	0.366	0.391	149	0.356 (0.077)
NF	0.030	0.003	0.004	0.003	0.002	0.391	0.003	0.435	18.798	19.479	7422	0.900 (0.007)
W	0.001	0.000	0.000	0.000	0.000	0.004	0.042	0.048	2.059	2.247	856	0.889 (0.021)
Total	0.447	0.016	0.019	0.014	0.011	0.446	0.048	1				
Stratified estimate of area (millions of ha)	19.323 (0.213)	0.696 (0.081)	0.814 (0.087)	0.621 (0.075)	0.467 (0.065)	19.278 (0.197)	2.065 (0.065)					
Reference plots (#)	7098	262	305	237	179	7554	850				16,485	
Producer's accuracy	0.912 (0.006)	0.157 (0.034)	0.158 (0.034)	0.228 (0.045)	0.279 (0.056)	0.878 (0.007)	0.886 (0.021)				Overall accuracy	0.852 (0.005)

half of all polygons for water class.

Maximum per-polygon area varied considerably among forest types (Fig. 3). Statistical distributions of reference- and corresponding map sample-based estimates show generally similar patterns for forest types, but with larger divergence than for land cover classes. KS distances for all forest type classes were larger than 0.10 (Table 9; Fig. 3), and AC

values ranged from 0.22 for other forest to 0.80 for deciduous forest (Table 9, Fig. 3). Except for the other forest class, all forest types had AC_{uns} values exceeding 0.74, and AC_{sys} values exceeding 0.91.

Map sample-based estimates of forest types were more frequently below 95% confidence intervals associated with the reference-based estimates for deciduous, evergreen, mixed, and other forest classes,

Table 7

Overall accuracy, allocation disagreement, and quantity disagreement for three classifications based on combined NLCD2011-VCTw map within BCR 12 and 23 in Minnesota, Wisconsin, and Michigan, USA. Standard errors of estimates are reported for OA in Tables 4, 5, and 6. No standard error estimates were available for AD or QD.

	Land cover	Forest type	Forest age
Overall accuracy	0.898	0.728	0.852
Allocation disagreement	0.091	0.205	0.107
Quantity disagreement	0.011	0.066	0.041
AD/QD ratio	8.0	3.1	2.6

Table 8

Producer's and user's accuracies (in percent) for five forest age classes and two general land cover classes, under three fuzzy set linguistic scales, based on combined NLCD2011-VCTw map within Bird Conservation Regions 12 and 23 in Minnesota, Wisconsin, and Michigan, USA.

Class	Producer's accuracy			User's accuracy		
	MAX	RIGHT4	RIGHT3	MAX	RIGHT4	RIGHT3
Forest >20 years	91	91	92	84	84	85
Forest 16–20 years	16	23	83	44	55	93
Forest 11–15 years	16	29	32	41	64	66
Forest 6–10 years	23	34	36	36	61	66
Forest 1–5 years	28	32	53	36	48	67
Nonforest	88	88	88	90	90	90
Water	89	89	89	89	89	89

but estimates for woody wetlands were substantially above confidence intervals for a few polygons (Fig. 3), which is consistent with Table 9. Scatterplots for evergreen forest and other forest classes reveal apparent artifacts of sample intensity (one plot per 2400-ha), showing several estimates falling along lines associated with a y-axis values of 0 ha, 2400 ha, and 4800 ha.

For forest age classes, maximum per-polygon areal extent occurred in persisting forest class, at nearly 400,000 ha, with map sample-based estimates overestimating relative to reference-based estimates (Tables 6, 9). Area estimates for four classes of young forest never exceeded 8% of any equal area polygon, with most map sample-based estimates

Table 9

Percent difference between map class area vs. reference-based area estimates and map sample-based estimates vs. reference-based area estimates across the entire study area; and metrics of KS, AC, AC_{sys}, and AC_{uns} based on 101 equal-area polygons, for land cover class, forest type, and forest age classifications in portions of Minnesota, Wisconsin, and Michigan, USA, within bird conservation regions 12 and 23.

Classification	Class	Map class difference (percent) ^a	Map sample difference (percent) ^b	KS ^c	AC ^d	AC _{sys} ^e	AC _{uns} ^f
Land cover	Forest	2.2	−1.7	0.07	0.97	1.00	0.97
	Nonforest	−2.5	1.0	0.06	0.98	1.00	0.98
	Water	−0.3	8.8	0.08	0.87	0.99	0.88
Forest type	Forest - Deciduous	−8.1	−11.4	0.15	0.80	0.96	0.84
	Forest - Evergreen	−26.5	−31.4	0.19	0.68	0.93	0.75
	Forest - Mixed	−15.3	−21.7	0.22	0.65	0.91	0.74
	Forest - Other	−28.8	−28.8	0.35	0.50	1.01	0.49
	Woody Wetlands	72.3	66.6	0.26	0.74	0.98	0.76
Forest age	Forest >20 years	9.1	4.4	0.16	0.94	0.99	0.95
	Forest 16–20 years	−64.0	−61.9	0.41	0.12	0.44	0.68
	Forest 11–15 years	−61.2	−63.5	0.33	0.09	0.42	0.67
	Forest 6–10 years	−37.0	−34.5	0.20	0.27	0.68	0.59
	Forest 1–5 years	−21.5	−16.2	0.12	0.64	0.08	0.29

^a Map class difference = ((Map class area – Reference sample plot-based estimate) / Reference sample plot-based estimate) * 100. Bold italics font represents map estimate within 95% CI of reference sample-based estimate.

^b Map sample difference = ((Map sample pixel-based estimate – Reference sample plot-based estimate) / Reference sample plot-based estimate) * 100. No sample pixel based estimates fell within 95% CI of reference sample-based estimates; therefore, this column contains no values with bold italics font.

^c Range: 0–1; 0 = perfect agreement.

^d Range: ≤ 0–1; 1 = perfect agreement, ≤ 0 = no relationship.

^e Range: 0–1; 1 = GMFR regression line is identical to the 1:1 line.

^f Range: 0–1; 1 = no scatter; all points line up on the GMFR regression line.

substantially smaller than reference-based estimates (Tables 6, 9). Among forest age classes, KS distance was smallest for the youngest age class (Table 9). Persisting forest age class had close alignment between 1:1 lines and reduced major axis (RMA) lines (Mark and Church, 1977), with corresponding high values of AC, AC_{sys}, and AC_{unsys}. Values were substantially lower for all 5-year age classes of young forest (Table 9). Map sample-based estimates of forest age classes had more variable patterns of over- and under-estimation compared to reference-based estimates. Similar to the results for forest types, forest age class also exhibited sampling artifacts as suggested by multiple estimates of 2400 ha.

Fig. 4 portrays spatial distributions of comparisons between map sample-based estimates of forest type classes with respect to 95 and 99% CI's of reference-based estimates (per Riemann et al., 2010), which correspond to points and confidence intervals, respectively, displayed in Fig. 3 (right column). With the exception of deciduous forest type, Fig. 4 includes two additional map categories representing polygons having zero of that class on the reference plots, the majority of which also have zero sampled map pixels in that same class (Fig. 4). For forest type classes, the polygons in which there is no reference data for a particular class are spatially clustered along the southern edge of the study area, where forest area is sparse. Spatial distributions of over- or under-estimates, and map-based estimates falling within reference-based confidence intervals, varied among forest types. Mixed forest class showed the fewest and woody wetlands showed the greatest number of deviations. Polygons showing over- or under-estimation of deciduous forest class usually show the reverse pattern for evergreen, and vice versa, but not always (Fig. 4).

5. Discussion

The intended application of a map determines the relevant accuracy metrics. Because actual map uses may differ from those originally intended, we recommend conducting a comprehensive assessment, rather than producing only one or two basic metrics. We reviewed and refined procedures for assessing map accuracy, class area estimates, and statistical and spatial patterns of errors and applied them in an analysis framework. Both thematic and geospatial errors can increase disagreement between map and reference observations, reducing estimated classification accuracy. To reduce such artifacts, we expanded site-specific metrics to partition disagreement and include fuzzy

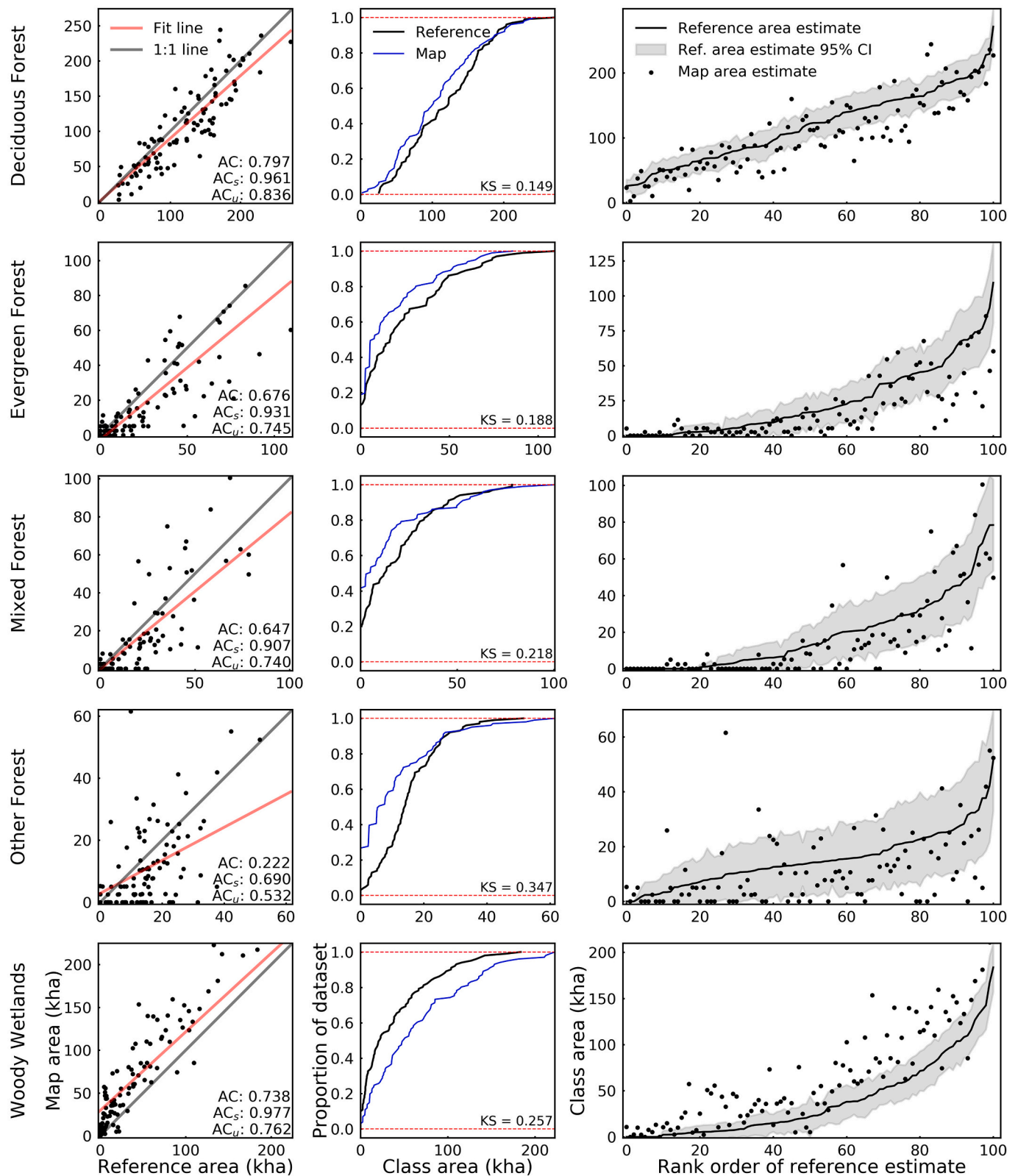


Fig. 3. Scatterplots of map sample- vs. reference-based area estimates (kha; thousand hectares) and AC, AC_{sys} (labeled AC_s) and AC_{unsys} (labeled AC_u) (left column); empirical cumulative distribution function (ECDF) and KS (middle column); and distribution of map sample- and reference-based area estimates with reference-based confidence intervals – sorted by reference-based estimates (right column), for deciduous, evergreen, mixed, other, and woody wetland forest type classes, at 433,000-ha scale, in portions of Minnesota, Wisconsin, and Michigan, USA within bird conservation regions 12 and 23.

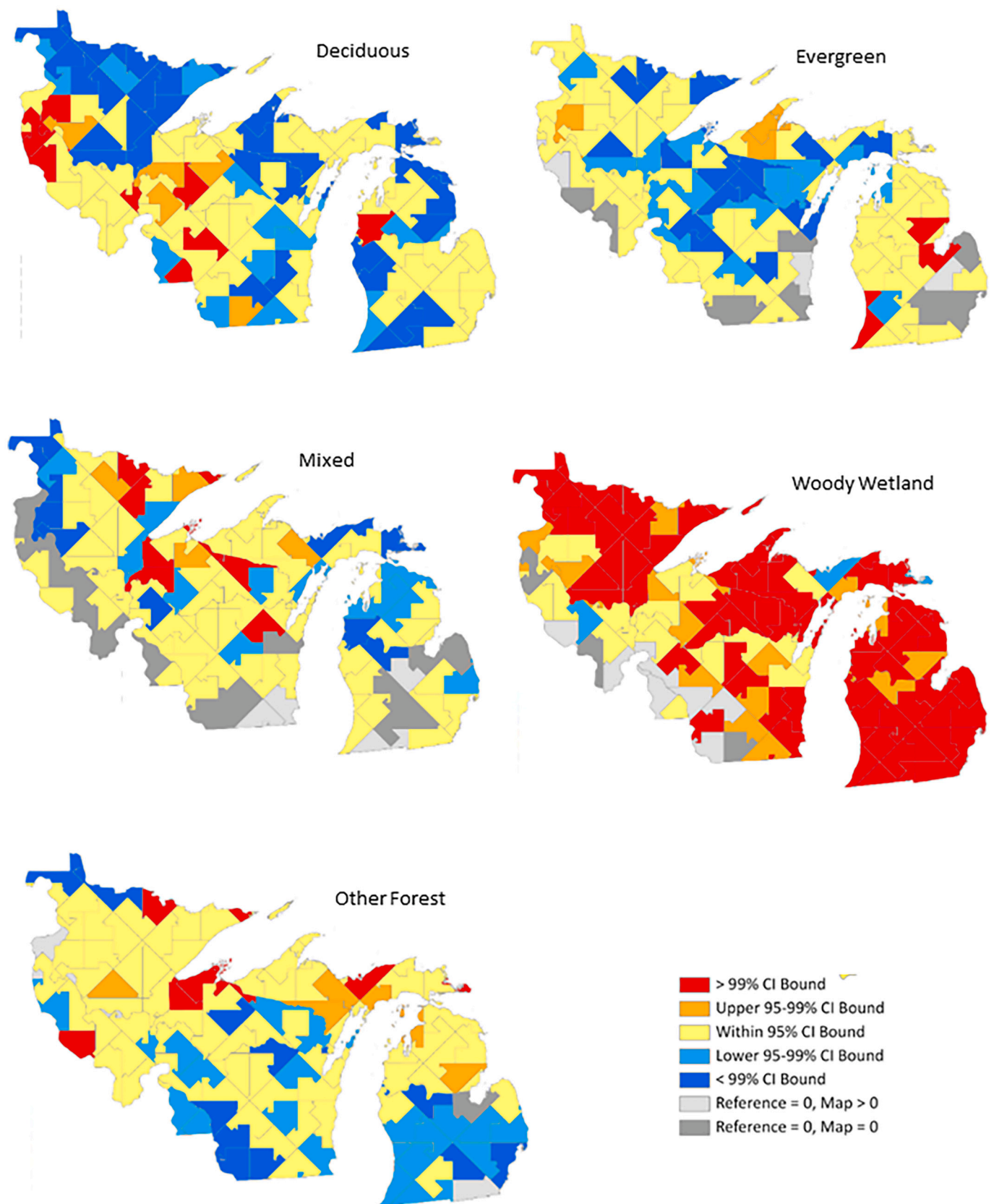


Fig. 4. Mapped comparisons of modeled area estimates relative to reference-based confidence intervals 99%, 95–99%, 0–95%, no reference data, or no data for either reference or map, for forest type classes in portions of Minnesota, Wisconsin, and Michigan, USA within bird conservation regions 12 and 23 at 433,000-ha scale.

membership, and we added non-site specific and spatial distribution metrics.

This comprehensive framework spans the range of approaches applied in most other studies, maximizing opportunities for both within- and between-study comparisons. We are unaware of published estimates from another study with which to compare the full suite of our results. Nevertheless, our primary objective was to inform appropriate use of a given map, to reveal deficiencies in map and reference classes, and to identify opportunities for improvement. We discuss the utility of this framework by describing its application to the Upper Great Lakes Woodcock and Young Forest Initiative (YFI) (Cooper, 2008) - the original intended use of these maps, and to other potential applications.

5.1.1. Site-specific

- Estimates of accuracy (OA, PA, UA) provided metrics of the agreement between map and reference observations for a classified map and across classified maps following recommended good practices in Olofsson et al. (2014).
- Procedures were implemented to reduce spatial, temporal, and definitional differences between map and reference observations.
- Including QD and AD metrics (Pontius and Millones, 2011) in our framework added a practical approach for partitioning sources of disagreement; we also introduced the ratio of AD/QD for comparing relative disagreement among studies.
- Inclusion of fuzzy membership with linguistic scales helped account for uncertainties inherent in image or reference labels but gains in OA resulting from fuzzy membership are not necessarily comparable among some studies due to different populations of interest, initial OAs, and other factors.

Site-specific assessments determine map accuracy at the scale of individual map elements (e.g., pixels), summarized across a geographic area of interest. We utilized an existing inventory dataset as reference data (FIA), which met good practices recommendations according to Olofsson et al. (2014). Confidence intervals associated with estimates of accuracy and area were quite small for many classes due to the relatively large sample size of reference data. Even the rarest classes of young forest had at least 100 reference observations, and their 95% confidence intervals were relatively small, suggesting that results can be applied with reasonable confidence. OA's were encouragingly high for maps of land cover and forest ages, but lower for forest types.

Based on relatively high OA for the forest age map, and high PA and UA for the mature forest age class, one might conclude that the forest age map has sufficient accuracy to support YFI needs. However, accuracies for individual ESF 5-year age classes were very low, and these low accuracies cause doubt about whether the forest age map is useful for assessing ESF habitats. We suspect that both map classification errors, and imperfect reference labels and locations contributed to low accuracies for ESF age classes, and that reported accuracies are conservative estimates, similar to the conclusions of Tavernia et al. (2016). The following points support this conclusion.

We assumed that the FIA reference classification was considerably more accurate than the map classification, per good practice recommendations of Olofsson et al. (2014). Supporting this assumption is anecdotal evidence from FIA field staff that FIA reference observations of stand age are reliable for youngest forest ages, becoming less reliable with increasing age, particularly in forest originating from natural regeneration (James Blehm, personal communication). However, Bell et al. (2015) suggested that accuracy assessments may underestimate map predictive performance when reference plot area is smaller than pixel resolution. Such is the case when using FIA central subplots with 30-m spatial resolution pixels, as was done for this study.

Furthermore, spatial misregistration between map and reference locations (pixel-plot) can reduce map accuracies, even when both map and reference labels are correct. McRoberts (2010) reported that up to half of the FIA subplots could be associated with an incorrect 30-m Landsat pixel due to the type of GPS receivers used in this region to measure FIA plot location coordinates. This may partially explain why Thomas et al. (2011) reported higher accuracies of mapped forest disturbance when using reference data from human interpretation of Landsat and higher resolution imagery than for FIA plot-based reference data. For future applications of FIA reference data, we recommend incorporating averaged locations to improve both reference locations (McRoberts et al., 2018) and age class assignment (Cohen et al., 2010).

We examined additional metrics to further understand these site-specific accuracies. For all three maps, less disagreement was attributed to QD, and more was attributed to AD. We computed a simple ratio of allocation disagreement to quantity disagreement (AD/QD) to compare the relative importance of disagreement sources among different classification systems. For example, Schmitt-Harsh et al. (2013) reported AD and QD for four models of their 5-class system, from which we computed AD/QD ratios, which were smaller than ratios in our 7-class and 3-class systems. This suggests that the number of classes, alone, does not explain relative size of AD/QD ratios. We recommend further investigation into the effects of number of classes and associated thematic and spatial characteristics on AD, QD, and AD/QD ratios, both within and between studies.

Map attribution of forest age is affected by timing of summer growing season image acquisition relative to timing of forest stand initiation, and by unavailability of imagery in some years due to cloud cover or other image anomalies. Estimation of fuzzy membership accuracy resulted in substantial improvement in PA and UA of ESF age classes when moving from MAX to RIGHT4. The RIGHT3 scale produced additional gains in PA and UA for the 1–5 and 16–20 years age classes. Thus, RIGHT4 detected substantial disagreement between 5-year age classes and their temporally adjacent 5-year age classes (1–5 vs. 6–10, 6–10 vs. 11–15, and 11–15 vs. 16–20); RIGHT3 revealed additional disagreement between 1 and 5 year old class and nonforest class, and between 16 and 20 year old class and > 20 year old forest age class. These results suggest that aggregating two or more 5-year age classes may affect higher per-class accuracies. For example, Tavernia et al. (2018) aggregated four 5-year age classes into a single ESF class (1–20 years) to assess habitat representativeness of American woodcock singing ground surveys. For this study, aggregating four 5-year age classes in the error matrix into a single ESF class (1–20 years) increased OA only slightly, a difference that likely is statistically insignificant. However, ESF class accuracies for the aggregated 1–20 year old class increased substantially over UAs and PAs reported for individual 5-year classes.

OA in some studies increased substantially under fuzzy membership assessment (Gill et al., 2000; Stehman et al., 2003; Wickham et al., 2017; Woodcock and Gopal, 2000), which contrasts with the minimal improvement in OA observed in our study. We assigned linguistic scales not only to forest age classes, but also nonforest and water classes, which were substantially more abundant than most age classes, resulting in estimates of OA that included all cover classes. In contrast, Gill et al. (2000) constrained their fuzzy membership accuracy results to the subset of map classes representing forest size and cover classes, omitting other land covers (nonforest, water). Furthermore, OAs in our study were larger than for other studies, likely due in part to the high PAs and UAs for older forest, nonforest, and water, with OA's minimally affected by the relatively smaller contributions from rare young forest classes.

Recall that non-stocked forest reference plots were labeled as 1–5 year old age class, under the assumption that such plots are regenerating forest. Conditions of forest canopy disturbance and subsequent regeneration contribute to a substantial share of discrepancies between NLCD and FIA forest classes (Nelson et al., 2020). We attempted to reduce these discrepancies by reclassifying NLCD grassland/herbaceous and

shrub/scrub classes as other forest where an independent forest disturbance map classified pixels as forest. Reference age was based on average age of trees, not on date of stand initiation. Trees can be >1 years of age when planted as seedlings, and natural regeneration often results in seedling establishment over multiple years, resulting in reference age labels possibly associated with a multi-year span of regeneration. In addition, narrow slivers of forest pixels without corresponding age data were assigned to the age class of adjacent forest pixels, and larger patches of forest pixels with no age class were assigned to >20 year old age class. For future applications, we recommend assessing impacts on map accuracy of these and other procedures.

5.1.2. Non-site specific

- Comparisons of map class areas to confidence intervals surrounding reference-based estimates of class area provided a non-site specific approach to assessing map accuracy for the study area as a whole; consistency varied among classes.
- Understanding the non-site specific accuracy of map class area estimates helps inform conservation decision making, like estimating abundance of ESF habitat for planning and management activities.

We reported pair-wise comparisons of area estimates and corresponding accuracies, between: 1) map class-based estimates and reference sample-based estimates, and 2) map sample-based estimates and reference sample-based estimates. Although pairwise comparisons were generally similar, random errors associated with misclassified pixels may have gone undetected when omissions and commissions largely offset each other.

Given our large sample size, we had expected that reference sample-based estimates would be closer to map sample-based estimates than to total map class areas because data for sample-based estimates were spatially co-located on the same number of observations, which should remove random effects of sample selection. This expectation was met for some classes, but not others. For water class, for example, the deviation between map sample- and reference sample-based estimates was substantially larger than the difference between map class- and reference sample-based estimates.

Two potential explanations are considered for why reference sample-based estimates were closer to map class-based estimates than to map sample-based estimates. First, recall that for error matrices based on good practice recommendations of Olofsson et al. (2014), each cell represents weighted class proportions based on counts of plot-pixel class-class assignments, with cell weighting based on the proportion of area mapped in each map class; reference sample-based area estimates are post-stratified estimates, with stratum weights obtained from total map class areas. Thus, stratum weighting from map class areas may affect reference sample-based estimates. Second, the relatively higher inconsistency between map sample- and map class-based estimates suggests that sample pixels, and therefore reference sample plots, may not have been representative of water map class proportions. We recommend further evaluation of both explanations.

The datasets in our example application have been used to support YFI needs by assessing metrics of landscape pattern associated with potential habitat for American woodcock (Tavernia et al., 2016), and spatial representativeness of American woodcock singing ground survey (SGS) routes (Tavernia et al., 2018). As noted previously, Olofsson et al. (2014) recommends using reference data to estimate class area. But in the case of both landscape metrics (e.g., patch, edge), and small area estimates (vicinity surrounding SGS routes) reference data either lacked appropriate attributes, or were of insufficient sample size; thus, the ESF map was the only available source of information. By estimating overall and per-class accuracies (site-specific), then comparing per-class reference- vs. map-based area estimates (non-site specific), we were able to assess the likely utility of map-based information for applications where reference data were unavailable. For the YFI, these results informed a

compromise approach for aggregating 5-year age class into one 20-year age class having much higher per-class accuracy, while still addressing the ESF habitat requirements for American woodcock. Similarly, national-level forest resource assessments often encounter similar challenges, prompting the use of both sample-based inventory data and map-based estimates (Nelson et al., 2020). In these cases, non-site specific assessments provide insights beyond those obtained only from site-specific assessments.

5.1.3. Spatial distribution

- We adapted approaches in Riemann et al. (2010) to assess statistical and spatial distributions of thematic map class errors among geographic subareas.
- Incorporating a space-filling curve allowed us to apply these approaches to equal-area polygons while removing artifacts in regular systematic sampling arrays.
- Map classifications with locally rare classes require a compromise of sample size, equal-area polygon size, or both.
- Mapped differences and ECDF graphs highlight variations in statistical distributions that are not evident from values of KS or AC alone.

Riemann et al. (2010) introduced a protocol for comparative accuracy assessment of continuous geospatial datasets, from which we adapted KS, AC, AC_{sys} , AC_{uns} , and ordered statistical distributions for application to thematic datasets. These metrics provide information on the type, magnitude, frequency and location of errors in each dataset. ECDF plots and mapped spatial distributions help further interpret results drawn from KS and AC metrics.

The following example illustrates how a map user can compare multiple metrics when assessing map quality and ascribe more or less confidence to a particular class in the context of specific geographic subregions and class thresholds. Compared to woody wetlands, mixed forest class appears to have higher accuracy based on KS results, but lower accuracy based on AC results. Mapped differences reveal apparently random deviations for mixed forest, but consistent over-estimation of map vs. reference estimates for woody wetlands. The map sample-based ECDF for woody wetlands was always below the reference-based ECDF line with relatively consistent deviations throughout the distribution of the two lines. Compared to woody wetlands, the average ECDF deviations for mixed forest are smaller (thus, smaller KS value), but deviations are inconsistent, and map sample- and reference-based lines cross each other because the shape of the map sample-based ECDF is irregular, reversing the over/underestimation relationships when per-polygon estimates exceed about 40,000 ha. A lesson from this example is that relative accuracy within a map class can vary among spatial subareas, which is revealed in figures for both statistical and spatial distributions of errors. When prioritizing local management decisions, one can consider both the spatial variability in class abundance, and also the varying uncertainties in those estimates.

The inclusion of AC_{sys} and AC_{uns} allowed for further partitioning of AC, suggesting that random uncertainty (unsystematic) and bias (systematic) are relatively lower for general land cover classes, moderate for forest type classes, and higher for 5-year age classes. AC_{sys} was larger than AC_{uns} for most classes, suggesting that bias was less important than random uncertainty, with the exception of three young forest age classes, where bias was more important. A high AC_{uns} value corroborates that the selected scale of equal area polygons contain sufficient reference data for associated confidence intervals to be of an acceptable width (Riemann et al., 2010), a condition which is met for only some map classes in this study. Furthermore, artifacts of sparse sample intensity in young forest classes, which are rare, was also revealed by linear alignment of per-polygon estimates at increments of 2400 ha, which is the average area represented by each sample plot. Both observations suggest that larger equal area polygons may be needed to better address rare classes when using this sample density.

The approach adapted from Riemann et al. (2010) requires a study area to be partitioned into subareas of non-overlapping equal area polygons, each delineating an estimation unit in which similar numbers of reference samples are located. Our use of a space-filling curve equal-area polygon framework (Lister and Scott, 2009) resulted in approximately equal sample sizes per polygon while also covering the full extent of our study area. We selected a polygon size of 433,000-ha across 101 estimation units as a compromise for obtaining suitable sample sizes for rare map classes within most polygons, for total number of polygons from which KS and AC are computed, and for portraying useful information on spatial variability of errors. Compared to Riemann et al. (2010), our selected polygon scale was twice as large as their coarsest scale (216,500-ha), and substantially larger than their finest scale (8660-ha). Whereas Riemann et al. (2010) assessed modeled estimates of a continuous variable (biomass) within a single class (forest) which was minimally affected by 'nodata' values, our three classifications each contained from three to seven classes, some of which were rare sub-categories of forest class, resulting in 'nodata' polygons even at the relatively coarse polygon scale of 433,000-ha.

In contrast with results for the study area as a whole, comparisons within equal-area polygons of 433,000-ha scale resulted in some cases of map-based estimates for every map class falling within 95% confidence intervals of corresponding reference-based estimates. At least two plausible explanations are that, 1) some subareas have more accurate classifications compared to OA for the study area as a whole, or, 2) smaller sample sizes corresponding with smaller estimation units result in higher standard errors and widths of associated confidence intervals, thereby reducing the likelihood per-polygon estimates are statistically significantly different. Indeed, Riemann et al. (2010) stated that some of the differences observed when using relatively smaller equal-area polygons with fewer reference sample plots will be due to increased uncertainty in reference-based estimates rather than real error in map-based estimates. We recommend development of additional approaches to further separate the effects of classification error from artifacts of class abundance and associated sample size.

Riemann et al. (2010) recommended analyzing error distribution at multiple scales from which additional understanding can be obtained about appropriate scales of estimation. We did not follow this recommendation for our study because smaller polygons would have resulted in larger confidence intervals and more numerous 'nodata' cases for rare classes; larger polygons would have benefited from more reference plots per polygon, with smaller confidence intervals and fewer 'nodata' cases, but with an insufficient number of polygon observations with which to perform statistical analyses of KS, AC, AC_{sys} , AC_{uns} , and ordered statistical distributions. Our 16,485 reference sample points were nominally sufficient for assessing rare classes within 101 equal area polygons at 433,000-ha scale across three states of our study area.

Calculation of these metrics may be hindered by data limitations and unusual situations. The following issues were noted in our example application.

There were a few instances for the water class where equal area polygons had sample-based estimates of area > 0 for both reference and map, but with no confidence intervals surrounding the post-stratified reference-based estimates. These cases are attributed to occurrences in which map sample- and reference-based observations for a class are identical within an individual polygon. When using Olofsson et al. (2014) Eq. (9) for estimating per-class areas, and the number of observations in reference class i that are in map class k (n_{ik}) is equal to the total number of observations in reference class i (n_i), then their ratio = 1, meaning that the class estimate reduces to the stratum weight (W_i), i.e., the per-polygon total pixel area for that class. We recommend further investigation and development of treatments for such instances.

Likewise, when using Olofsson et al. (2014) Eq. (10) to estimate the per-polygon, per-class proportion of area standard error and resulting confidence intervals, the numerator = 0 when map and reference classes are identical, resulting in no error estimate for that class. For a handful

of cases, we observed no confidence intervals surrounding post-stratified reference-based estimates for the water class, even though the class area estimate was > 0 . These unusual cases appear to occur when map accuracy for a class is 100% and number of reference samples for that class is relatively low. When such issues are rare, map users may choose to omit those cases from analysis, or assign them to special cases. Larger sample sizes likely would reduce or eliminate this anomaly of having no confidence interval for a class estimate.

Except for the unusual situation of perfect agreement, effects of insufficient per-class sample sizes likely are less concerning for classes with higher accuracies, and more concerning for classes with lower accuracies. Delineating smaller equal area polygons for assessing management-scale estimation units, with corresponding smaller sample sizes, likely would exacerbate this problem under the current sampling design, or would require a modification to sample intensity. For MN and WI portions of our study, FIA reference data are available at double sampling intensity, but only at single intensity in MI. To obtain equal sample sizes across the entire study area we defaulted to the single sample intensity that is available everywhere. However, given equal area polygons that are also constrained to state boundaries, it would be possible to evaluate each state separately, with different sample intensities in each state.

Many uses of 30-m spatial resolution maps are based not on individual pixels, but on aggregations of adjacent like pixels known as 'patches', 'features', 'objects', or 'stands'. For example, Tavernia et al. (2016) used the maps described in this study to obtain landscape metrics associated with patches of ESF within forest age classes for YFI assessment; these patches were comprised of multiple pixels, but reported map accuracies pertain to pixels, not to resulting patches. Thus, there is need to further refine the accuracy assessment framework to better align map accuracy assessments with the scale of features used in map analysis. Post-processing procedures may have partially alleviated pixel-patch mismatches. We applied a four-pixel minimum mapping unit, meaning that in the final map, every pixel is immediately adjacent to at least three like neighbors, resulting in a minimum patch size of about 0.4 ha. Other studies have applied MMUs of different sizes to various classes within the same map to correspond with differing levels in confidence between static and change classes (Thomas et al., 2011). Map-based estimates are affected by underlying spatial resolution, and aggregation of map pixels to coarser spatial resolution can affect changes in estimates (Nelson et al., 2009), but spatial aggregation was not performed in this study. This map accuracy assessment framework can be applied during multiple phases of a map's production to determine, for example, the effects on accuracies of applying MMUs, spatial resolution and aggregation, reclassifying pixels based on manual revisions or inclusion of ancillary data, and refinements to reference data.

6. Conclusions

Although some of the procedures in our framework require more reference data and have more constraints than for standard site-specific assessments like OA, PA, and UA, inclusion of additional metrics enables a more comprehensive understanding of map quality for a given map classification. Therefore, we recommend the inclusion of more, rather than fewer assessment procedures.

A robust framework with multiple assessments benefits both map producers and map users. Map producers can identify potential issues with quality and sample size limitations of reference data, partition sources of error, uncover thematic and spatial variability in class accuracies, and identify targeted improvements to map production. Map users can benefit by better understanding a dataset's strengths and weaknesses, choosing which thematic classes to aggregate, making more informed choices about which datasets to use where, and improving land management decision making. In conclusion, we recommend the following:

- To obtain and report a comprehensive statement of map quality and utility for multiple potential map uses, we recommend including more, rather than fewer validation metrics when conducting accuracy assessments.
- An analysis framework incorporating recommended good practices for sampling design, response design, analysis protocols, and design-based inference (Olofsson et al., 2014) provides rigorous estimates of accuracy and precision.
- Partitioning disagreement (Pontius and Millones, 2011) allows map users to assess the relative importance of spatial location versus total abundance of map classes to map error; we introduced a ratio of AD/QD for comparing relative importance of each across datasets.
- Assessment of fuzzy membership with linguistic scales (Gopal and Woodcock, 1994) allows map users to discount map–reference mismatches between pairs of classes that are deemed similar for a particular resource need; we demonstrated that this approach may have minimal effect on OA, while providing substantial improvement in UA and PA for uncommon classes.
- Non-site specific approaches provide metrics and insights into spatial variability of map accuracy. Comparing estimates based on reference data and co-located samples of map data may reveal additional issues with reference data.
- Adapting or adding metrics may further improve the map accuracy assessment framework.

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.

Acknowledgements

We thank the USDA Forest Service Northern Research Station and the USFWS's Webless Migratory Game Bird Program for funding. We also thank Gretchen Moisen, Jody Vogeler, Dacia Meneguzzu, and Randy Morin for their reviews and suggestions.

References

- Bell, D.M., Gregory, M.J., Roberts, H.M., Davis, R.J., Ohmann, J.L., 2015. How sampling and scale limit accuracy assessment of vegetation maps: a comment on Loehle et al. (2015). *For. Ecol. Manag.* 358, 361–364.
- Card, D.H., 1982. Using known map category marginal frequencies to improve estimates of thematic map accuracy. *Photogramm. Eng. Remote Sens.* 48, 431–439.
- Cheng, T., Molenaar, M., Lin, H., 2001. Formalizing fuzzy objects from uncertain classification results. *Int. J. Geogr. Inf. Sci.* 15, 27–42.
- Chrisman, N.R., 1998. Rethinking levels of measurement for cartography. *Cartogr. Geogr. Inform. Syst.* 25, 231–242.
- 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. *Rem. Sens. Environ.* 114, 2911–2924.
- Cohen, W.B., Healey, S.P., Yang, Z., Stehman, S.V., Brewer, C.K., Brooks, E.B., Gorelick, N., Huang, C., Hughes, M.J., Kennedy, R.E., Loveland, T.R., Moisen, G.G., Schroeder, T.A., Vogelmann, J.E., Woodcock, C.E., Yang, L., Zhu, Z., 2017. How similar are forest disturbance maps derived from different landsat time series algorithms? *Forests* 8, 98.
- Comber, A.J., 2008a. Land use or land cover? *J. Land Use Sci.* 3, 199–201.
- Comber, A.J., 2008b. The separation of land cover from land use using data primitives. *J. Land Use Sci.* 3, 215–229.
- Congalton, R.G., 1991. A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sens. Environ.* 37, 35–46.
- Cooper, T.R., 2008. Stepping Down the Habitat Goals of the American Woodcock Conservation Plan for the Upper Great Lakes Woodcock and Young Forest Initiative. Association of Fish and Wildlife Agencies, pp. 1–15. Unpublished Report to the Woodcock Task Force.
- Feller, W., 1948. On the Kolmogorov-Smirnov Limit Theorems for Empirical Distributions, pp. 177–189.
- Foody, G.M., 1992. A fuzzy sets approach to the representation of vegetation continua from remotely sensed data: an example from lowland heath. *Photogramm. Eng. Remote. Sens.* 58, 221–225.
- Foody, G.M., 2010. Assessing the accuracy of land cover change with imperfect ground reference data. *Remote Sens. Environ.* 114, 2271–2285.
- Garner, J.D., Nelson, M.D., Tavernia, B.G., Perry, C.H., Housman, I.W., 2015. Mapping forest canopy disturbance in the Upper Great Lakes, USA. In: Stanton, S.M., Christensen, G.A. (Eds.), *Pushing Boundaries: New Directions in Inventory Techniques and Applications: Forest Inventory and Analysis (FIA) Symposium 2015*. U.S. Department of Agriculture, Forest Service, Pacific Northwest Research Station, Portland, OR, pp. 363–367 [CD-ROM].
- Garner, J.D., Nelson, M.D., Tavernia, B.G., Housman, I.W., Perry, C.H., 2016a. Early Successional Forest and Land Cover Geospatial Data of the Upper Midwest. Forest Service Research Data Archive, Fort Collins, CO.
- Garner, J.D., Nelson, M.D., Tavernia, B.G., Housman, I.W., Perry, C.H., 2016b. Early Successional Forest and Land Cover Geospatial Data of the Upper Midwest. Data Product Updated 07 April 2016. Forest Service Research Data Archive, Fort Collins, CO. <https://doi.org/10.2737/RDS-2016-0001>.
- Gill, S.J., Milliken, J., Beardsley, D., Warbington, R., 2000. Using a mensuration approach with FIA vegetation plot data to assess the accuracy of tree size and crown closure classes in a vegetation map of northeastern California. *Remote Sens. Environ.* 73, 298–306.
- Gopal, S., Woodcock, C.E., 1994. Theory and methods for accuracy assessment of thematic maps using fuzzy sets. *Photogramm. Eng. Remote. Sens.* 60, 181–188.
- Hagen, A., 2003. Fuzzy set approach to assessing similarity of categorical maps. *Int. J. Geogr. Inf. Sci.* 17, 235–249.
- Häme, T., Stenberg, P., Andersson, K., Rauste, Y., Kennedy, P., Folving, S., Sarkeala, J., 2001. AVHRR-based forest proportion map of the Pan-European area. *Remote Sens. Environ.* 77, 76–91.
- Homer, C.G., Dewitz, J.A., Yang, L., Jin, S., Danielson, P., Xian, G., Coulston, J.W., Herold, N.D., Wickham, J.D., Megown, K.A., 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.
- Huang, C., Goward, S.N., Masek, J.G., Thomas, N., Zhu, Z., Vogelmann, J.E., 2010. An automated approach for reconstructing recent forest disturbance history using dense Landsat time series stacks. *Remote Sens. Environ.* 114, 183–198.
- Ji, L., Gallo, K., 2006. An agreement coefficient for image comparison. *Photogramm. Eng. Remote. Sens.* 72, 823–833.
- Kelley Jr., J.R., Williamson, S., Cooper, T.R., 2008. American woodcock conservation plan: a summary of and recommendations for woodcock conservation in North America. In: Rahm, J. (Ed.), *Compiled by the Woodcock Task Force, Migratory Shore and Upland Game Bird Working Group*. Association of Fish and Wildlife Agencies, p. 168.
- Kennedy, P., Bertolo, F., 2002. Mapping sub-pixel forest cover in Europe using AVHRR data and national and regional statistics. *Can. J. Remote. Sens.* 28, 302–331.
- Laba, M., Gregory, S.K., Braden, J., Ogurcak, D., Hill, E., Fegraus, E., Fiore, J., DeGloria, S.D., 2002. Conventional and fuzzy accuracy assessment of the New York Gap Analysis Project land cover map. *Remote Sens. Environ.* 81, 443–455.
- Lister, A.J., Scott, C.T., 2009. Use of space-filling curves to select sample locations in natural resource monitoring studies. *Environ. Monit. Assess.* 149, 71–80.
- Mark, D.M., Church, M., 1977. On the misuse of regression in earth science. *J. Int. Assoc. Math. Geol.* 9, 63–75.
- Mayaux, P., Eva, H., Javier, G., Strahler, A.H., Herold, M., Agrawal, S., Naumov, S., De Miranda, E.E., Di Bella, C.M., Ordoyne, C., Kopin, Y., Roy, P.S., 2006. Validation of the global land cover 2000 map. *IEEE Trans. Geosci. Remote Sens.* 44, 1728–1739.
- McRoberts, R.E., 2010. The effects of rectification and Global Positioning System errors on satellite image-based estimates of forest area. *Remote Sens. Environ.* 114, 1710–1717.
- McRoberts, R.E., 2011. Satellite image-based maps: scientific inference or pretty pictures? *Remote Sens. Environ.* 115, 715–724.
- McRoberts, R.E., Chen, Q., Walters, B.F., Kaisershot, D.J., 2018. The effects of global positioning system receiver accuracy on airborne laser scanning-assisted estimates of aboveground biomass. *Remote Sens. Environ.* 207, 42–49.
- Miles, P.D., 2015. Tue Sep 01 12:58:04 MDT 2015. Forest Inventory EVALIDATOR web-application version 1.6.0.02. U.S. Department of Agriculture, Forest Service, Northern Research Station, St. Paul, MN [Available only on internet: <http://apps.fs.fed.us/Evalidator/evalidator.jsp>].
- Nelson, M.D., 2005. Satellite remote sensing for enhancing national forest inventory. In: *Forest Resources*. University of Minnesota, St. Paul, Minnesota, p. 167.
- Nelson, M.D., McRoberts, R.E., Holden, G.R., Bauer, M.E., 2009. Effects of satellite image spatial aggregation and resolution on estimates of forest land area. *Int. J. Remote Sens.* 30, 1913–1940.
- Nelson, M.D., Riitters, K.H., Coulston, J.W., Domke, G.M., Greenfield, E.J., Langner, L.L., Nowak, D.J., O'Dea, C.B., Oswalt, S.N., Reeves, M.C., Wear, D.N., 2020. Defining the United States Land Base: A Technical Document Supporting the USDA Forest Service 2020 RPA Assessment. U.S. Department of Agriculture, Forest Service, Northern Research Station, Madison, WI, p. 70.
- 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.
- Päivinen, R., Lehtikoinen, M., Schuck, A., Häme, T., Väättäin, S., Kennedy, P., Folving, S., 2001. Combining Earth Observation Data and Forest Statistics. European Forest Institute, Joint Research Center-European Commission, Joensuu, Finland, p. 101.
- Pontius, R.G., Millones, M., 2011. Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. *Int. J. Remote Sens.* 32, 4407–4429.
- Reams, G.A., Smith, W.D., Hansen, M.H., Bechtold, W.A., Roesch, F.A., Moisen, G.G., 2005. The Forest Inventory and Analysis sampling frame. pp. 11–26. In: W. A. Bechtold and P. L. Patterson (Eds.), *The enhanced Forest Inventory and Analysis*

- Program - national sampling design and estimation procedures. U.S. Department of Agriculture Forest Service, Asheville, NC, p. 85.
- Riemann, R., Wilson, B.T., Lister, A., Parks, S., 2010. An effective assessment protocol for continuous geospatial datasets of forest characteristics using USFS Forest Inventory and Analysis (FIA) data. *Remote Sens. Environ.* 114, 2337–2352.
- Riemann, R., Wilson, B.T., & Lister, A.J. (2012). Application of an assessment protocol to extensive species and total basal area per acre datasets for the eastern coterminous United States. In R.S.L. Morin, Greg C., comps. (Ed.), *Moving from status to trends: Forest Inventory and Analysis (FIA) symposium 2012* (pp. [CD-ROM]: 161–174). Baltimore, MD: U.S. Department of Agriculture, Forest Service, Northern Research Station.
- Sarmiento, P., Fonte, C.C., Dinis, J., Stehman, S.V., Caetano, M., 2015. Assessing the impacts of human uncertainty in the accuracy assessment of land-cover maps using linguistic scales and fuzzy intervals. *Int. J. Remote Sens.* 36, 2524–2547.
- Särndal, C.E., Swensson, B., Wretman, J., 1992. *Model Assisted Survey Sampling*. Springer-Verlag, New York, NY.
- Schmitt-Harsh, M., Sweeney, S.P., Evans, T.P., 2013. Classification of coffee-forest landscapes using Landsat TM imagery and spectral mixture analysis. *Photogramm. Eng. Remote Sens.* 79, 457–468.
- Schuck, A., Päivinen, R., Häme, T., Van Brusselen, J., Kennedy, P., Folving, S., 2003. Compilation of a European forest map from Portugal to the Ural mountains based on earth observation data and forest statistics. *Forest Policy Econ.* 5, 187–202.
- Stehman, S.V., 2009. Sampling designs for accuracy assessment of land cover. *Int. J. Remote Sens.* 30, 5243–5272.
- Stehman, S.V., 2013. Estimating area from an accuracy assessment error matrix. *Remote Sens. Environ.* 132, 202–211.
- Stehman, S.V., Czaplewski, R.L., 1998. Design and analysis for thematic map accuracy assessment: fundamental principles. *Remote Sens. Environ.* 64, 331–344.
- 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.
- Stehman, S.V., Wickham, J.D., Smith, J.H., Yang, L., 2003. Thematic accuracy of the 1992 National Land-Cover Data for the eastern United States: statistical methodology and regional results. *Remote Sens. Environ.* 86, 500–516.
- Stehman, S.V., Arora, M.K., Kasetkasem, T., Varshney, P.K., 2007. Estimation of fuzzy error matrix accuracy measures under stratified random sampling. *Photogramm. Eng. Remote. Sens.* 73, 165–173.
- Stevens, D.L., Olsen, A.R., 2004. Spatially balanced sampling of natural resources. *J. Am. Stat. Assoc.* 99, 262–278.
- Stevens Jr., D.L., Olsen, A.R., 1999. Spatially restricted surveys over time for aquatic resources. *J. Agric. Biol. Environ. Stat.* 4, 415–428.
- Stevens, S.S., 1946. On the theory of scales of measurement. *Science* 103, 677–680.
- Stueve, K.M., Housman, I.W., Zimmerman, P.L., Nelson, M.D., Webb, J.B., Perry, C.H., Chastain, R.A., Gormanson, D.D., Huang, C., Healey, S.P., Cohen, W.B., 2011. Snow-covered Landsat time series stacks improve automated disturbance mapping accuracy in forested landscapes. *Remote Sens. Environ.* 115, 3203–3219.
- Tavernia, B.G., Nelson, M.D., Garner, J.D., Perry, C.H., 2016. Spatial characteristics of early successional habitat across the upper Great Lakes states. *For. Ecol. Manag.* 372, 164–174.
- Tavernia, B.G., Nelson, M.D., Rau, R., Garner, J.D., Perry, C.H., 2018. American woodcock singing-ground survey sampling of forest type and age. *J. Wildl. Manag.* 82, 1794–1802.
- Thomas, N.E., Huang, C., Goward, S.N., Powell, S., Rishmawi, K., Schleeuwis, K., Hinds, A., 2011. Validation of North American Forest Disturbance dynamics derived from Landsat time series stacks. *Remote Sens. Environ.* 115, 19–32.
- Wickham, J., Stehman, S.V., Gass, L., Dewitz, J.A., Sorenson, D.G., Granneman, B.J., Poss, R.V., Baer, L.A., 2017. Thematic accuracy assessment of the 2011 National Land Cover Database (NLCD). *Remote Sens. Environ.* 191, 328–341.
- Woodcock, C.E., Gopal, S., 2000. Fuzzy set theory and thematic maps: accuracy assessment and area estimation. *Int. J. Geogr. Inf. Sci.* 14, 153–172.
- Woudenberg, S.W., Conkling, B.L., O'Connell, B.M., LaPoint, E.B., Turner, J.A., Waddell, K.L., 2010. *The Forest Inventory and Analysis Database: Database description and users manual version 4.0 for Phase 2*. U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station, Fort Collins, CO, p. 336.
- Zadeh, L.A., 1963. Outline of a new approach to the analysis of complex or imprecise concepts. *IEEE Trans. Syst. Man Cybernet.* 3, 28–45.
- Zadeh, L.A., 1965. Fuzzy sets. *Inf. Control.* 8, 338–353.
- Zimmerman, P.L., Housman, I.W., Perry, C.H., Chastain, R.A., Webb, J.B., Finco, M.V., 2013. An accuracy assessment of forest disturbance mapping in the western Great Lakes. *Remote Sens. Environ.* 128, 176–185.