Remote Sensing of Environment 114 (2010) 2337–2352

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

An effective assessment protocol for continuous geospatial datasets of forest characteristics using USFS Forest Inventory and Analysis (FIA) data

Rachel Riemann [a,*], Barry Tyler Wilson [b], Andrew Lister [c], Sarah Parks [a]

[a] U.S. Forest Service, Northern Research Station, 425 Jordan Road, Troy, NY 12180, USA
[b] U.S. Forest Service, Northern Research Station, St. Paul, MN, USA
[c] U.S. Forest Service, Northern Research Station, Newtown Square, PA, 19073, USA

ARTICLE INFO

Article history:
Received 9 October 2009
Received in revised form 6 May 2010
Accepted 7 May 2010

Keywords:
Accuracy
Biomass
Uncertainty
Geospatial datasets
FIA
Comparative assessment
Forest characteristics
Minnesota
New York
Moderate resolution datasets
Measures of agreement

ABSTRACT

Geospatial datasets of forest characteristics are modeled representations of real populations on the ground. The continuous spatial character of such datasets provides an incredible source of information at the landscape level for ecosystem research, policy analysis, and planning applications, all of which are critical for addressing current challenges related to climate change, urbanization pressures, and data requirements for monitoring carbon sequestration. However, the effectiveness of these applications is dependent upon the accuracy of the geospatial input datasets. A comprehensive set of robust measures is necessary to provide sufficient information to effectively assess the accuracy of these modeled geospatial datasets being produced. Yet challenges in the availability of reference data, in the appropriateness of assessment methods to dataset use, and in the completeness of assessment methods available have continued to hamper the timely and consistent application of map assessments. In this study we present a suite of assessments that can be used to characterize the accuracy of geospatial datasets of modeled continuous variables—an increasingly common format for modeling such attributes as proportion or probability of forestland as well as more traditionally continuous attributes such as leaf area index and forest biomass. It is a comparative accuracy assessment, in which each modeled dataset is compared to a set of reference data, recognizing both the potential for error in reference data, and probable differences in spatial support between the datasets. When used together, this proposed suite of assessments provides essential information on the type, magnitude, frequency and location of errors in each dataset. The assessments presented depend upon reference data with large sample sizes. The U.S. Forest Service (USFS) Forest Inventory and Analysis (FIA) database is introduced as an available reference dataset of sufficient sampling intensity to take full advantage of these assessments and facilitate their prompt application after modeled datasets are developed. We illustrate the application of this suite of assessments with two modeled datasets of forest biomass, in Minnesota and New York. The information provided by this suite of assessments substantially improves a user's ability to apply modeled geospatial datasets effectively and to assess the relative strengths and weaknesses of multiple datasets depicting the same forest characteristic.

Published by Elsevier Inc.

1. Introduction

Geospatial modeling with remotely sensed data is being used to produce geospatial datasets of continuous variables such as proportion of forest cover, biomass, leaf area index, and forest composition variables (e.g. Cohen & Goward, 2004; Ohmann & Gregory, 2002). Every one of those modeled datasets contains error, as both our knowledge and the information available for modeling is imperfect. This error can be the result of poor georeferencing of input datasets, a poor relationship between the attribute being modeled and the input datasets available, error in those input datasets, inappropriate assumptions made by the modeling technique used, and/or poor choice of parameters in the application of that technique. Errors can take the form of truncated distributions, loss of variability, and/or overestimation or underestimation of values, and these errors can vary by subpopulation, by region and by scale. In addition, the error will be some combination of both random or unsystematic error, and systematic error or bias. Such inaccuracies do not necessarily render a modeled dataset useless, but they do affect the uses and questions to which it can be applied. To support effective sharing and application of geospatial datasets, the Federal Geographic Data Commission developed guidelines for providing detailed information regarding how a geospatial dataset was produced (FGDC, 1998), and many widely available datasets provide this information. However, while this information on dataset

* Corresponding author. Tel.: +1 5182855607; fax: +1 5182855601.
E-mail address: rriemann@fs.fed.us (R. Riemann).

0034-4257/$ – see front matter. Published by Elsevier Inc.
doi:10.1016/j.rse.2010.05.010

development is useful for understanding a dataset's objectives, strengths, and limitations, it requires considerable experience to interpret, particularly with respect to dataset accuracy for a specific application.

Many measures of agreement have been developed for the assessment of geospatial datasets; however, most research has focused on measures for categorical variables such as land use or forest type classes, rather than continuous variables (Lunetta & Lyon, 2004). There remains an urgent need for methods to effectively, consistently, and efficiently determine the relative strengths and limitations of geospatial datasets of continuous variables to better inform the user community of their efficacy for various applications and to comparatively assess two datasets of the same attribute (Congalton, 2001; Congalton, 2004; Wardlow & Egbert, 2003). Without valid accuracy assessment we cannot know which modeled maps to use for estimates of biomass or deforestation or where red oak is dominant. As Strahler et al. (2006) pointed out, "without proper validation, any map, whether at global, regional, or local scale, remains an untested hypothesis."

Foody (2002) and Canters (1997) identified four characteristics of error that each provide invaluable information toward understanding the utility of a final geospatial product or toward iterative development of a new geospatial dataset: the *location of errors,* the *frequency of errors,* the *magnitude of errors,* and the *type/nature of errors*. This list reflects a clear understanding that error does indeed vary across a dataset, and that the frequency, magnitude, scale, and type of error present, in addition to its location, dramatically affect whether that error is tolerable or too high for a particular application. Yet few assessment protocols currently exist for effectively characterizing all of these components of error. The objective of this study is to build an assessment protocol that addresses all of these characteristics—a protocol that explicitly recognizes the variety of ways in which geospatial datasets are used and is sufficiently complete to inform most uses. In addition, the protocol includes measures of agreement that recognize the uncertainty in reference as well as modeled datasets, provide consistent and readily interpretable results, and are capable of being applied quickly after datasets are developed.

1.1. Characteristics of effective assessments

1.1.1. Including multiple types of assessment

Both the understanding and implementation of accuracy assessment or validation of modeled geospatial datasets has evolved considerably over the years. Whereas once a single percent-correctly-classified metric was considered sufficient, it is now recognized that limiting map assessments to a few simple measures does not provide the user with a full understanding of a map's quality or usefulness, particularly when producing geospatial datasets of land characteristics that may be used for multiple purposes and objectives (Laba et al., 2002; Stehman et al., 1997). Strahler et al. (2006) recommended that a suite of assessments be used for evaluating the quality of global land cover maps and pointed out that statistical observations, confidence maps, and qualitative-systematic accuracy reviews all contribute to our understanding of the quality of a particular geospatial dataset. Despite this understanding, however, the approach has not yet been regularly implemented, particularly with continuous datasets. Many continuous geospatial datasets are still frequently provided with only a single measure of agreement—e.g. a coefficient of determination (r^2) or root mean square error (RMSE) value. This is insufficient information with which to assess the efficacy of a geospatial dataset for a particular application as it provides no indication of the types of error present, or of any regional variation in that error, nor is it always in a form that can be readily compared across regions and datasets or clearly describes the magnitude of error.

1.1.2. Describing characteristics relevant to how the dataset will be used

Each measure of accuracy that has been employed over the years is sensitive to different features and evaluates different components of error (Foody, 2002). Thus, it is important when choosing assessment procedures to consider the likely uses of a geospatial product to ensure that the map assessment process is valid for those uses (Congalton & Plourde, 2000; Congalton, 2004). For some applications, this translates into an area-based assessment in which data summaries are compared for a set of meaningful analysis units (e.g., Holden et al., 2003; Lowell, 2001; White et al., 2005 or Blackard et al., 2008). In other instances, where modeled geospatial data are being used as input for subsequent environmental modeling or decision analysis, information on the spatial distribution of error is essential (e.g. Binaghi et al., 1999; DeGloria et al., 2001), and errors in modeled data distribution can be critical because of their effect on subsequent interpretation of high and low threshold levels (Moeur & Riemann Hershey, 1999). At other times, it is critical to understand the accuracy of a map's depiction of spatial variability because of the effect of local variability on the outcomes of forest management decisions (Smith et al., 1991). For a general purpose geospatial dataset, one for which a specific use has not been defined, a complete assessment should include all those characteristics that affect a user's ability to discern spatial features and patterns, accurately calculate summary statistics for local areas, utilize threshold values, or overlay it with other geospatial datasets at several potential application scales.

1.1.3. Describing the location of error

The effectiveness of many traditional assessments is limited by the presence and significance of regional variation in error. Even employing multiple relevant metrics may be of limited use when these metrics are applied only to the dataset as a whole, providing no information as to how that error varies across the landscape (Bastin et al., 2002). Error can also vary by subpopulations defined by an attribute, such as below and above an important percent-biomass or percent-forest threshold. In thematic accuracy assessment, information on individual class accuracy is available from the confusion matrix, and many studies have long recommended developing methods for effectively using this information (e.g., Aspinall, 2002; Goodchild et al., 1992; Pontius, 2002). Of equal importance is information about accuracy within different segments of a continuously modeled dataset, either representing the variation of accuracy in space or between subsets of the population (e.g., Fassnacht et al., 2006; Holden et al., 2003; Janssen & van der Wel, 1994; White et al., 2005).

1.1.4. Assessing across a range of scales

The effectiveness of many traditional assessments is also limited by the scale of assessment applied, since differences observed in area assessments frequently vary with the scale of assessment (e.g. Blackard et al., 2008; Nelson et al., 2009). Assessment results are typically presented at the scale at which the data are suspected of being the most accurate, or at which accuracy results were highest for the type of assessment being conducted, with the implication, in either case, that it is the recommended scale of application for that geospatial dataset. In contrast, calculating and comparing assessment results across a range of scales would provide assessment information for a wider variety of potential users, allowing each to determine if the level and type of disagreement at that scale is too high for their application.

1.1.5. Timely and consistent application of assessments

Timely availability of assessment results improves the speed with which geospatial datasets can be put into effective application. Many over the years have called for more funding and effort to be directed regularly into the validation of all geospatial datasets produced (e.g., Goodchild et al., 1992; Strahler et al., 2006), yet complete and timely

assessments are still not commonplace. Evidenced by the wealth of research that has gone into assessing the performance and accuracy of the 1992 National Land Cover Dataset (NLCD), it can take many years before a large-area dataset is fully understood even if it is as widely used as NLCD (e.g., Galbraith et al., 2003; Stehman et al., 2003; Smith et al., 2002, 2003; Wardlow & Egbert, 2003; Wickham et al., 2004; Yang et al., 2001). Improving our ability to effectively utilize existing data sources could dramatically improve the problem of timeliness.

Fassnacht et al. (2006) and others have observed that a variety of accuracy results can be obtained, often accidentally, when minor modifications are incorporated into the methods used to calculate those accuracies. Comparability between assessments conducted at different times would be improved with increased consistency in the methods and computation of comparative accuracy assessments. Comparability between assessments conducted on different datasets or different regions is also improved when the measures of agreement used are standardized or bounded, and not dependent on the magnitude of data values, units, or scale (Ji & Gallo, 2006).

In short, current assessment practices for geospatial products are incomplete and ineffective when they only rely on one or two measures, use measures that do not address the scales or the way in which the geospatial dataset will be used, apply only to the dataset as a whole and ignore regional and subpopulation differences, are not sufficiently timely, and/or do not lend themselves to comparability across datasets.

1.2. Challenges limiting the interpretation of geospatial data assessments

1.2.1. Mismatches in spatial support

In cases where assessments involve a comparison of values from an individual plot that occupies a small spatial footprint (like that of a ground inventory plot) with values that represent a larger area of land (like a MODIS pixel), results can be strongly affected by the spatial mismatch between the area covered by a ground plot (typically between 0.067 and 0.4 ha) vs. that covered by a 250-m pixel (6.25 ha) (e.g., Congalton, 2004; Zhu et al., 2000). This mismatch in sample unit size is particularly significant with discrete variables involving a high contrast between classes, and with continuous variables when environmental heterogeneity causes fine-scale spatial patterns (Mayaux & Lambin, 1995; Moody & Woodcock, 1994; Townshend et al., 1992; Riemann et al., 2000). In our view of accuracy assessment, a large pixel can be thought of as containing a potentially heterogeneous population from which a sample is taken using a small inventory plot. Thus comparative assessments in locally heterogeneous areas are most affected by this mismatch. To avoid this complication, some studies have attempted to minimize mixed areas in the assessment by using only single-condition plots—i.e. plots identified on the ground as occurring completely in a single forest type, land cover class, or standsize (e.g., Blackard et al., 2008). However, this may bias an accuracy assessment towards those homogeneous areas in which we tend to have more confidence (Hammond & Verbyla, 1996), or may cause an assessment to miss a substantial portion of the landscape in highly fragmented areas (Fassnacht et al., 2006). Other studies have suggested providing an associated dataset derived from plot and/or image data to indicate levels of expected local spatial variability present and thus potential difference due to differences in sample unit size (Hershey & Reese, 1999). Modeled uncertainties from geostatistical simulation are also based on the unexplained local spatial variability at the plot level (Hershey, 2000). In the last two approaches, the effects of local heterogeneity are not removed from an accuracy assessment, but instead additional information is provided to help users identify those areas where local estimates are likely to be strongly affected by differences in spatial support. Comparative assessments at the plot–pixel level will be most strongly affected by differences in sample unit size, with decreasing effect as the scale of assessment increases.

Comparisons between areal estimates derived from sample plots vs. modeled datasets are also affected by a mismatch of a second type—effectively a mismatch in "sampling intensity" when calculating area summaries. In other words, when comparing a ground plot dataset with a sampling intensity of one plot for every 3000 ha to a continuous modeled dataset at 250-m resolution which provides a modeled estimate for every 6.25 ha, the difference between 1:3000 ha and 1:6.25 ha is substantial. Thus in this example, for a 10,000 ha area, the number of individual values contributing to each mean value would be 1600 pixels for the 250 m modeled dataset, compared to only 3–4 ground inventory plots within the same area. Left uncorrected, differences in sampling intensity of this magnitude primarily affect assessments examining spatial variability and data distribution, and at the finer assessment scales.

1.2.2. Error and uncertainty in the reference dataset

Any reference dataset used in an assessment typically contains some error and uncertainty, whether that dataset was based on a sample of ground measurements or modeled at a finer resolution from a combination of remotely sensed, ground-based and other spatial data layers. For estimates from design-based inference such as from a plot-based sample, one can calculate a precision metric such as sampling error or confidence interval which is related to the variability of observations in the sample. For a model-based dataset, modeling techniques are increasingly able to produce uncertainty values that calculate indices of the strength of correlation between training data and predictor layers used in the modeling. In both cases, incorporating these known errors explicitly into the assessment improves the interpretability and applicability of accuracy assessment results (e.g. Holden et al., 2003; Ji & Gallo, 2006).

When assessment results are interpreted, any discrepancies between fine-scale observations from a sample of ground plots and the relatively continuous broader scale estimates, such as biomass for 250-m pixels, are due to a combination of real error in the modeled estimate, *and* uncertainty in the reference data, *and* artifacts associated with the different levels of spatial support being compared (Liu & Journel, 2009; Xu et al., 2009). To maximize the utility of an accuracy assessment, it is extremely valuable to clearly distinguish between these as much as possible.

2. Description of the assessment protocol

Based upon the gaps and recommendations observed above, we have identified a need for a specific suite of metrics that a) together effectively characterize the type, magnitude, frequency, and location of error in each dataset, b) account for or consider in the interpretation all mismatches in spatial support between modeled and reference datasets, c) account for uncertainty in both datasets, d) are ideally applied at more than one scale, and e) given sufficient validation data, can be quickly and consistently applied to maps of continuous estimates of some attribute. In order to accomplish this, the protocol we recommend relies on four types of assessment that incorporate both graphical and geographical visualizations and both qualitative and quantitative measures of agreement. When used together, these assessments provide the breadth of information necessary to effectively assess a geospatial dataset for general use. The selected assessments and measures of agreement are summarized in Table 1 and described below. Section 3.2 presents examples of each using test datasets of aboveground forest biomass.

2.1. Examining data distribution

An empirical cumulative distribution function (ecdf) describes the distribution of values present in a dataset and is an effective tool for comparing the distributions of modeled and reference datasets, a

Table 1
Summary of the types of assessment and measures of agreement used in the proposed protocol, with example scales.

Assessment type	Graphics and measures of agreement used	Purpose
Examine data distribution at several scales		
• Plot:pixel • 10 k hex (21,400 ha) • 30 k hex (78,100 ha) • 50 k hex (216,500 ha) • By subpopulation or specific region of interest	• Empirical cumulative distribution function (ecdf) (Fig. 3) • Kolmogorov-Smirnov statistic (KS)	• Identify differences in data distribution (e.g., 0's, max values, missing range of values…) • Metric summarizing the maximum distance between modeled and reference ecdfs
Examine overall agreement of area estimates at several scales		
• Plot:pixel • 10 k hex (21,400 ha) • 50 k hex (216,500 ha) • By subpopulation or specific region of interest	• Scatterplot with 1:1 line and geometric mean functional relationship (GMFR) regression line (Fig. 4) • Agreement Coefficient (AC)	• Visualize areas of relative overestimation and underestimation, and level of systematic and unsystematic differences present • Metric summarizing the level of agreement between the two datasets (range: 0–1; affected by both systematic and unsystematic error)
	• Systematic AC (AC_{sys})	• Metric summarizing the level of systematic agreement (proximity of GMFR regression line to 1:1 line)
	• Unsystematic AC (AC_{uns})	• Metric summarizing the level of unsystematic agreement (level of scatter about the GMFR regression line)
	• Difference between means • Root mean square error (RMSE)	• Metrics summarizing the difference between datasets in data units (e.g. Mg/ha) both overall (RMSE) and at one threshold of interest
	• Maps of plot- and model-based estimates by hex (Figs. 5, 6)	• Visual comparison of area estimates for assessing spatial patterns
Examine spatial and distribution patterns of local differences		
• 50 k hex (216,500 ha) scale	• Comparison of estimates vs. plot-based confidence interval (Fig. 8) • Choropleth map of modeled estimates with respect to plot confidence intervals (CI) (Fig. 9)	• Identify where modeled means fall with respect to reference means and confidence intervals • Spatially identify areas where local differences are within plot confidence intervals • Identify direction and approximate relative magnitude of local bias where local estimates are outside plot confidence intervals
Examine local variability		
• 50 k hex (216,500 ha) scale	• Choropleth maps of standard deviation of modeled estimates within 50 k hex area (Fig. 7)	• Identify difference in modeled and plot local variation—a characteristic that is often compromised in an effort to improve the accuracy of area means.

characteristic frequently affected, to varying degrees, by the modeling methods used. Assumptions about distribution may be built into the modeling technique used, resulting in a shift, often toward normality, in the modeled output. Similarly, techniques involving averaging commonly result in a loss of values at the tails of the distribution. Such differences can affect a modeled dataset's ability to accurately depict the very high, very low, or rare values that may be present in the real population and picked up in the reference dataset. Evaluation of dataset differences can be done qualitatively via a visual comparison of the ecdf plots. Of interest is the general shape of the curves and their distance from one another, the *y*-intercepts (that proportion of the dataset that has zero values), flat sections in the ecdf curves (missing classes of values), and at what value the ecdf plot reaches 100% (the max dataset value).

Differences in data distribution can also be quantified. For comparison of continuous data distributions, the Kolmogorov–Smirnov (*KS*) statistic quantifies the agreement between the distributions of two datasets, in terms of the maximum difference (*D*) in their empirical distribution functions. When comparing two samples with cumulative distribution functions $F(x)$ and $G(x)$, the *KS* statistic is defined as

$$D_{KS} = \max |F(x) - G(x)| \quad (1)$$

Kolmogorov–Smirnov is a robust statistic that makes no assumption about the distribution of data, and is independent of scale changes (e.g. Lopes et al., 2007). A measure that captures the average distance between the two curves might be even more desirable as a quantitative measure of agreement between two data distributions.

An ecdf assessment should be conducted both at the individual points where plots and pixels intersect, and for area estimates at several scales. Included should be one scale that is large enough to contain sufficient reference data such that the confidence interval associated with each estimate is acceptably small, plus one or two scales above and below that. The size required will depend upon the spatial intensity of reference data available and the spatial variability of the attribute in question. Choosing this scale can be based on a general threshold or based on an acceptable maximum or average confidence interval size. Comparison of ecdfs at coarser scales will be less affected by differences in spatial support than finer scales because of the number of plots and pixels contributing to the mean estimate.

2.2. Examining overall agreement of area estimates

Comparison of modeled pixel estimates to measurements from reference data points, or of area estimates calculated from the modeled vs. reference datasets, provides a way to evaluate several other types of relative error that may be present. Two characteristics of agreement evident in a scatterplot comparison that are worth capturing are the proportion of difference that is systematic disagreement or bias, and the proportion that is unsystematic or random difference (lack of precision) (Fig. 1). The reduced major axis (RMA) line (Mark & Church, 1977) or geometric mean functional relationship (GMFR) regression line (Draper & Smith, 1998; Ricker, 1984) can be used to describe the relationship between two datasets, *X* and *Y*. Unlike least squares regression, the GMFR is a symmetric regression model

$$\left(\hat{Y} = a + bX\right) \quad (2)$$

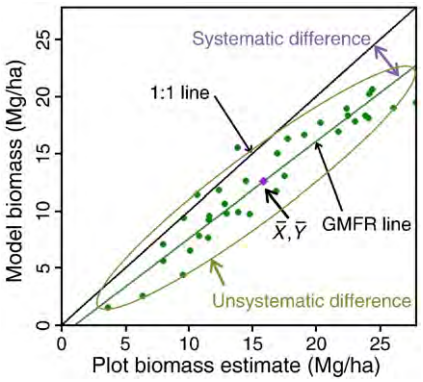

Fig. 1. Illustration of systematic and unsystematic differences in the relationship between two datasets.

that assumes both X and Y are subject to error, and thus the coefficients are determined by

$$a = \overline{Y} - b\overline{X} \tag{3}$$

and

$$b = \pm\left(\frac{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right)^{\frac{1}{2}} \tag{4}$$

In this equation, datasets X and Y can be interchanged without changing the GMFR line. Systematic difference can be described as that distance between the GMFR regression line and the 1:1 line. Unsystematic difference is the scatter about the GMFR regression line (Fig. 1). Overall agreement is affected by both systematic and unsystematic differences.

Many quantitative measures of agreement have been developed for the comparison of per-pixel or area estimates of continuous variables, and Ji and Gallo (2006) provide a brief but useful summary of many of the metrics that have been used to measure agreement between two such datasets. The metrics r^2 and RMSE are two that are commonly reported with continuous geospatial datasets as a measure of agreement, but they are not the best choice, particularly when used in isolation. To facilitate cross-region and cross-dataset comparisons, it is important that the metric be bounded (between fixed minimum and maximum values) and standardized to non-dimensional units so that the units of measurement of X and Y do not affect the measure. Without this, metric values vary with data unit and scale as well as with level of agreement, making comparison more difficult between regions with substantially different amounts of the attribute of interest. It may also be important that the metric be symmetric—i.e. that interchanging datasets X and Y in the assessment calculations will produce identical results. A metric that is symmetric allows for agreement comparisons among datasets derived from different sources that may each contain error—a typical situation in natural resource applications. The coefficient of determination (r^2) measures the proportion of the total data variation explained by the least squares regression model, which can be misleading when it is used to measure data agreement as it fails to measure the actual difference between the two datasets (Ji & Gallo, 2006). Root mean square error (RMSE) does measure the actual difference between two datasets, but it is neither standardized nor bounded, making comparisons of levels of agreement between regions characterized by high vs. low values difficult when RMSE is used on its own. Root mean squared percentage error (RMSPE) is a standardized version of RMSE, however it is not symmetric, and becomes unstable when X_i values are near zero, a common situation both with remotely sensed data and maps of forest characteristics (Ji & Gallo, 2006).

In this paper we use the agreement coefficient (AC) developed by Ji and Gallo (2006) for its ability to provide a symmetric, bounded, and non-dimensional measure of agreement. Ji and Gallo (2006) provide a full description, but briefly, AC is defined as

$$AC = 1 - \frac{SSD}{SPOD} \tag{5}$$

where SSD is the sum of square difference

$$SSD = \sum_{i=1}^{n}(X_i - Y_i)^2 \tag{6}$$

and SPOD is the sum of potential difference

$$SPOD = \sum_{i=1}^{n}\left(|\overline{X} - \overline{Y}| + |X_i - \overline{X}|\right)\left(|\overline{X} - \overline{Y}| + |Y_i - \overline{Y}|\right) \tag{7}$$

where $\overline{X}$ and $\overline{Y}$ are the mean values of datasets X and Y, respectively. The AC ranges from ≤ 0–1, where AC = 1 if agreement between X and Y is perfect, and values less than or equal to zero indicate no agreement. The agreement coefficient is symmetric because the identical value is obtained if X and Y are interchanged, making the assumption that neither dataset is more correct than the other. If we do believe one dataset is substantially more accurate than the other, then Willmott's Index of Agreement (d) can be used as a measure of agreement that is bounded and non-dimensional, but is not symmetric (Ji & Gallo, 2006; Willmott, 1981). Importantly, one can calculate systematic vs. unsystematic difference separately, for either d or AC. From Ji and Gallo (2006), the unsystematic sum of product-difference (SPD_u) is defined as

$$SPD_u = \sum_{i=1}^{n}\left(|X_i - \hat{X}_i|\right)\left(|Y_i - \hat{Y}_i|\right) \tag{8}$$

where $\hat{X}$ and $\hat{Y}$ are from the GMFR regression model, and SPD_s can be obtained by $SPD_s = SSD - SPD_u$. Using SPD_s and SPD_u in place of SSD in the AC equation, one can calculate systematic and unsystematic agreement coefficients by

$$AC_{sys} = 1 - \frac{SPD_s}{SPOD} \tag{9}$$

and

$$AC_{uns} = 1 - \frac{SPD_u}{SPOD} \tag{10}$$

where $AC_{sys} = 1$ if the GMFR line is perfectly in line with the 1:1 line, and $AC_{uns} = 1$ if all points fall directly on the GMFR line. For a user, systematic difference represents that difference which could be predicted from the other dataset by a simple linear model. Unsystematic difference represents those differences which appear to be random and unrelated to the reference dataset. Using measures of agreement that are standardized facilitates cross-site and cross-dataset comparisons, while additional use of measures of agreement in data units, such as RMSE or the difference between means, provides complementary information that can facilitate interpretation of the actual magnitude and potential impact of those differences.

Similar to evaluation of data distribution, assessments should be conducted at the same set of scales. Assessments of area estimates at finer scales will be more affected by differences in spatial support than at the coarser scales because of the number of plots and pixels contributing to the mean estimate.

2.3. Examining spatial and distribution patterns of local differences

Spatial variations in ground conditions, in the quality of input datasets, and in the local applicability of modeling techniques inevitably result in modeled geospatial datasets that vary in their level of error across the region of interest. Assessment and reporting of relative accuracy by region or by subpopulation often reveals important patterns of variation in the magnitude and direction of relative error.

This assessment should be conducted at that scale which is large enough to contain sufficient reference data for associated confidence intervals to be of an acceptable size, while also being small enough to provide a reasonable spatial depiction of regional variation across the area of interest—a compromise between precision and spatial distinction. This will likely be a scale at which unsystematic agreement, AC_{uns}, is high in the assessment of the overall agreement of area estimates. Although several measures of agreement are possible, we have chosen a simple difference in estimates of area means—essentially complementing the assessment in Section 2.2 by choosing one scale and examining the results by subpopulation or region and in comparison to the actual size of the reference data confidence interval at each location. This measure is in data units and thus the size of the confidence interval and magnitude of difference is readily interpretable. Results can also be depicted as being within or outside different confidence interval levels, provided that the size of that confidence interval is also presented and less precise estimates are regarded with less confidence. Presenting results as a choropleth map facilitates qualitative interpretation of the spatial pattern of dataset differences as well as the magnitude of differences between the datasets. Presenting the results in graph form provides readily interpretable information on the distribution of differences across the range of values.

2.4. Examining local variability

Another characteristic often compromised in modeled geospatial datasets is local variability. Loss of local variability can be an unintended side-effect of optimizing for the accuracy of local or global means, a result of a lack of input information at that scale, or it can be an intended modification in some output datasets to improve the interpretability of other spatial characteristics such as broad-scale trends. Meanwhile, other techniques such as most-similar-neighbor may preserve local spatial and/or attribute variability over other population characteristics (Moeur & Riemann Hershey, 1999). Whatever the situation, the level of local variability present impacts applications from the calculation of local texture or diversity, to the cost-effectiveness of management actions, to the choice of research sites.

Assessment of local variability should be conducted at the same scale as the assessment of local differences in the previous section. Presenting local variability as a choropleth map of local variance or standard deviation facilitates qualitative interpretation of the magnitude and spatial patterns of differences between the datasets. Local variability is one assessment that is highly affected by differences in spatial support. Procedures for matching sampling intensity are relatively straightforward, but differences in the sample unit size remain, and the local variability of 6-ha means cannot match the local variability of .06-ha ground inventory plots because they are describing a different scale. Nevertheless, local variability is a sufficiently important characteristic of modeled geospatial datasets to warrant its assessment as a description of the level of local spatial variability present in the modeled dataset and its relationship to the local variability present in the reference data.

2.5. Choosing a reference dataset for maps of forest characteristics—the FIA ground inventory data as a reference dataset

The assessments proposed in this paper depend upon a reference data source of sufficient extent and intensity to provide reasonably accurate summary statistics for comparison of agreement by region. The reference data source should be able to provide unbiased estimates (as in an equal probability sampling design), be in the same time frame as the modeled dataset, and collected in a consistent and well understood manner with known ground data accuracies. For modeled maps of forest characteristics, the USFS Forest Inventory and Analysis (FIA) database is a valuable resource. The FIA program has created a database that can be used to generate a set of reference data consisting of forest inventory information attached to georeferenced plots. These plot data are relatively current, collected in a standardized fashion, and distributed relatively uniformly across the continental United States, Hawaii, U.S. territories, and parts of Alaska. This database provides useful training and accuracy assessment data for large-area mapping applications (e.g., Blackard et al., 2008), providing design-based estimates that are relatively free of assumptions as compared with model-based estimates. Furthermore, the random location of FIA plots within cells formed by a hexagonal tessellation of the country ensures that the reference data provide an unbiased characterization of the entire study area in which they are used (Bechtold & Scott, 2005). This helps mitigate concerns that accuracy results may be affected by an optimistic or pessimistic ("conservative") bias due to sampling design (Hammond & Verbyla, 1996) or sample placement (Congalton & Plourde, 2000). One factor that needs to be considered when using FIA plots as reference data for accuracy assessment across all lands, however, is that tree information is currently not collected in areas that do not meet FIA's definition of forest: a minimum current or past tree stocking of 10%, minimum area of 0.405 ha, minimum width of 37 m, and not subject to a nonforest use that prevents normal tree regeneration. Therefore, in nonforest areas, FIA plots contain little or no information on vegetation characteristics, which will affect the accuracy assessment of those "nonforest" areas containing trees.

3. Application of the assessment protocol with two modeled biomass datasets in Minnesota and New York

In order to illustrate the value of this protocol we have applied it in the assessment of two modeled geospatial datasets of aboveground tree biomass. In this application we use USFS Forest Inventory and Analysis (FIA) ground plot data as the reference dataset. These data are readily available to researchers within FIA, and to those outside FIA via a Memorandum of Understanding (MOU) agreement or through FIA's Spatial Data Services Center. Protocols developed in this paper should be able to be readily incorporated into the services provided by the Spatial Data Services Center. A full assessment of these two modeled biomass datasets will be presented in another paper.

3.1. Data

3.1.1. Modeled datasets

To demonstrate the performance of these assessments we applied them to two different geospatial datasets of aboveground live forest biomass using two different techniques. Both techniques were modeled at a grain size of 250-m resolution, using Moderate Resolution Imaging Spectrometer (MODIS) composites and other ancillary information. Study areas were the states of Minnesota (MN) and New York (NY). Both states have over 16 million ac of forestland (Miles, 2010), but MN presents strong regionalization of forest and biomass values, while NY presents a greater proportion of forest area intermixed with nonforest.

The first dataset (Blackard et al., 2008) was created using the classification and regression tree software Cubist (http://www.rulequest.com, Quinlan, 1986, 1993) and included as predictor data sources: MODIS composites (Justice et al., 2002), Landsat Thematic Mapper image-derived National Land Cover Dataset (NLCD92,

Vogelmann et al., 2001), raster climate data, and topographic variables. Biomass was modeled for only those 250-m pixels previously identified as forestland using See5 (http://www.rulequest.com, Quinlan, 1986, 1993). We refer to this dataset hereafter as *bCUB* (Blackard CUBist) dataset (Fig. 2a).

The second modeled dataset of forest biomass was created using a procedure modified from the gradient nearest neighbor (GNN) technique developed by Ohmann and Gregory (2002). One primary difference from the methods outlined by Ohmann and Gregory (2002) is the use of MODIS composites from the entire growing season to take advantage of phenological differences between species (Wilson et al., submitted for publication). Hence this dataset is referred to as the *pGNN* (phenological-GNN) dataset in this study (Fig. 2b). This weighted nearest neighbor approach uses the 2nd–7th nearest neighbors and is moderated by the proportion of National Land Cover Dataset 2001 (NLCD2001) forest pixels within each modeled grid cell. This is in contrast to the bCUB dataset in which only forested pixels were modeled.

3.1.2. Reference dataset—the FIA ground inventory data

The ground inventory data used in the comparative accuracy assessment are from the annual FIA plots. FIA plots consist of a 0.067-ha (0.17-ac) plot cluster distributed over approximately 0.405 ha (1 ac). They are collected at a standard sampling intensity of 1 plot per 2428 ha (6000 ac), as in NY, or 1214 ha (3000 ac) in double-intensity states such as MN, resulting in 17,883 plots in MN and 5198 plots in NY after a full cycle. In states on a 5-year cycle, one fifth of the plots are measured each year, resulting in a full cycle of plots completed every 5 years. States on a longer cycle (such as NY for the first few years), will have less than a full cycle of plots available after 5 years. For this study 2979 plots were available for NY for the years 2001–2005, approximately 60% of the full cycle. The plot data generally consist of direct measurements of tree and site attributes and have a history of quality assurance behind them (U.S. Forest Service, 2003). We generated plot-level summaries of *aboveground live biomass*, which included biomass in live tree bole, wood, top, and limbs, for trees 2.54 cm (1 in.) diameter at breast height or larger, by summing tree-level biomass values generated from regional and species-specific allometric equations (Fig. 2c).

3.1.3. Handling the mismatches in sampling intensity and sample unit size

One could conceive of geospatial datasets such as the pGNN and bCUB as area samples effectively generating continuous datasets with

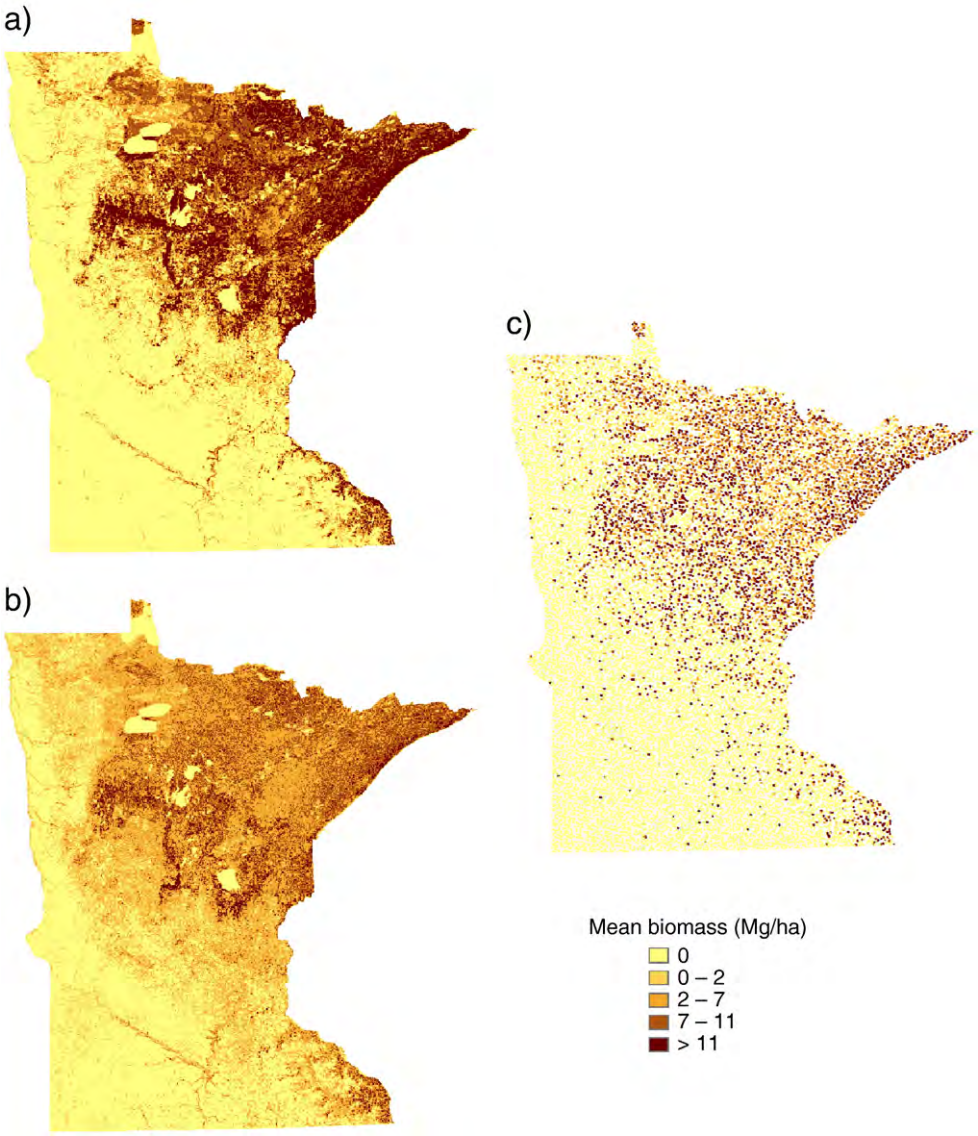

Fig. 2. Maps of aboveground tree biomass as depicted by the two moderate spatial resolution (250-m) modeled biomass datasets compared in this study for Minnesota — (a) the Blackard CUBist (bCUB) dataset, and (b) the phenological gradient nearest neighbor (pGNN) dataset, and (c) the 2001–2005 annual FIA plots.

an estimate at each location. Each location is equal to the grid cell size used in the modeling, which is typically the pixel size of one of the primary image data sources—here 6.25 ha (15.44 ac), for a "sampling intensity" of one modeled biomass value for every 6.25 ha. This is in contrast to the sampling intensity of the FIA plots at one plot for every 2428 ha. Differences in sample intensity (si) should be corrected before any assessments. In this study this was accomplished by subsetting a new dataset from the full modeled datasets of only those pixels that co-occurred with the FIA plots. This comparison dataset is distinguished from the original continuous dataset by the suffix "si"—thus the bCUB dataset being compared to the plots is referred to as *bCUBsi*, and the pGNN dataset used for the comparison as *pGNNsi*. The mismatch in sample unit size remains; however, the finer-scale assessments are interpreted with this knowledge, and their relative importance weighted accordingly.

3.2. Comparative assessments

The suite of assessments performed included all four types of comparisons (Table 1), two of which were performed at both point locations and area estimates at 3 scales.

3.2.1. Assessment scales used

Choice of scale is the result of a compromise between having sufficient reference data points within each unit area such that confidence intervals are adequately low, and having a sufficient number of area estimates from which to generate ecdfs, calculate GMFR regression lines and measures of agreement, and provide good spatial regionalization of error when mapping dataset differences. Additional consideration can be given to those scales at which the data are commonly applied and the scales of variation present within the dataset, if known. In this assessment we generated three polygon datasets of tessellated hexagons, providing three scales of equal area polygons. In the first dataset, hexes were generated around centroids spaced 10 km apart, resulting in hexes 8660 ha (21,400 ac) in size, and corresponding to the approximate size of the smallest counties nationally. These hexes contain 3–4 FIA plots under standard FIA sampling intensity (e.g., NY) and represent the minimum number of plots (4) used by FIA to define its strata for estimation (Scott et al., 2005). The remaining two datasets were generated with hex centroids separated by 30 km and by 50 km, resulting in hexes 78,100 and 216,500 ha in area, respectively. The 50 k hexes contain an average of 89 plots after a full cycle under standard FIA sampling intensity. In this study, only 60% of the full inventory cycle was available in NY, resulting in an average of 51 plots in each 50 k hex. In MN, where a double-intensity plot sample was available, there was an average of 178 plots per 50 k hex. At this scale, 90% confidence intervals ranged from +/–0 to 1.09 Mg/ha for means ranging from 0 to 7.89 Mg/ha in MN for aboveground tree biomass. In this assessment the 216,500 ha scale was chosen for those assessments conducted at just one scale.

3.2.2. Assessment of data distributions, at four scales

Fig. 3 presents the ecdfs of each of the three datasets (FIA plots, bCUBsi and pGNNsi) in MN (a–d) and NY (e–h), calculated at all four scales, and plotted together for comparison. At the plot:pixel scale in both MN and NY (Fig. 3a, e), the flat section of the line in the bCUB dataset indicates that entire classes of values are missing—between 0 and about 5 Mg/ha biomass in MN and between 0 and 11 Mg/ha in NY, almost certainly a difference created by the application of a nonforest mask in the modeling that effectively eliminated all areas with biomass <5 Mg/ha in MN. In contrast, the FIA plot ecdfs clearly indicate the presence of some areas at each level of aboveground tree biomass, even without having measured trees on nonforest plots. It is also apparent that the pGNN dataset, which was developed using an approach that models biomass at all pixels regardless of their forest/nonforest status, results in the largest number of pixels with greater than zero biomass, essentially predicting biomass into areas that were not measured in the FIA plot database and thus which we know nothing about from the reference dataset. Each of these effects could have been roughly predicted from knowledge of the modeling methods used, but an ecdf assessment summarizes the actual effect on the output dataset. Examining maximum predicted values at the 216,500 ha scale, the FIA plot dataset reaches its maximum close to 9 Mg/ha (Fig. 3d), whereas the bCUBsi dataset reaches that value at 90% of the dataset, indicating that 10% of the bCUBsi estimates at that scale in MN are above the maximum estimate in the reference dataset. At the same scale in NY (Fig. 3h) the maximum value in the bCUBsi area estimates is 22 Mg/ha while 27% of the area estimates calculated from the reference plot data are above that level, indicating a relative underestimation of biomass estimates in the bCUB dataset in NY at that scale. These differences may be due to the spatial pattern of biomass values in NY, in which high and low values are more spatially intermixed throughout the state than in MN where both high biomass and the tendency toward overestimation is more concentrated in the northeastern portion of the state.

The KS distance quantifies the maximum difference between each model ecdf and the ecdf of the FIA plot data. In this example, the KS distance is largest at the plot:pixel scale for the pGNNsi dataset in both MN and NY because of the much larger number of plots with very small biomass values, however it decreases to close to zero by the 78,100 ha scale of aggregation as the two distributions become more similar (Fig. 3 and Table 2). In contrast, the KS distance for the bCUBsi dataset in MN is lowest at the plot:pixel scale and increases with increasing scales of aggregation. Results for the bCUBsi dataset in NY are less clear, with KS staying approximately the same and then increasing at the 216,500 ha scale of aggregation to 0.286. Although not picking up the detail described above, the KS results support the visual conclusion from the ecdf plots that, between these two modeled datasets, except for perhaps the plot:pixel scale, the pGNNsi distribution is closest to the plots in terms of data distribution, and that the distributions become more similar with increasing scale. A metric capturing the mean difference over the entire distribution would be more specifically descriptive of data distribution agreement.

One important advantage of a multi-scale assessment is the clues it provides or does not provide regarding the presence of dataset disagreement over and above uncertainty and spatial support differences. For example, the fact that the pGNNsi dataset becomes more similar to the plot ecdf with increasing scale and is quite similar by the 78,100 ha scale, suggests that the differences present at the plot–pixel and 8600 ha scale may be more due to differences in spatial support and/or dataset uncertainty than real differences between the modeled dataset and the true population from which the FIA reference dataset is a sample. In contrast, differences between the reference and bCUBsi dataset persist throughout all scales examined, including those scales at which we are reasonably confident in the FIA plot estimates. This situation can be interpreted as evidence that the differences observed at all scales suggest real differences between the datasets in addition to reference data uncertainties and differences in spatial support.

3.2.3. Assessment of overall agreement of area estimates, at four scales

Fig. 4 presents the scatterplot comparison of modeled estimates vs. estimates derived from the FIA plots, for both test states and at all four scales. The 1:1 line and the GMFR regression lines for each model–plot relationship are added to aid visual interpretation, and the AC, AC_{sys}, and AC_{uns} (Ji & Gallo, 2006) were calculated for each modeled dataset. It was readily apparent that the GMFR line changed with increasing scale, and that the two modeled datasets under examination in this study each behaved differently with increasing scale. The GMFR line associated with the pGNNsi:FIA relationship moved closer to the 1:1 line as scale increased—also reflected in increasing AC_{sys} values. The

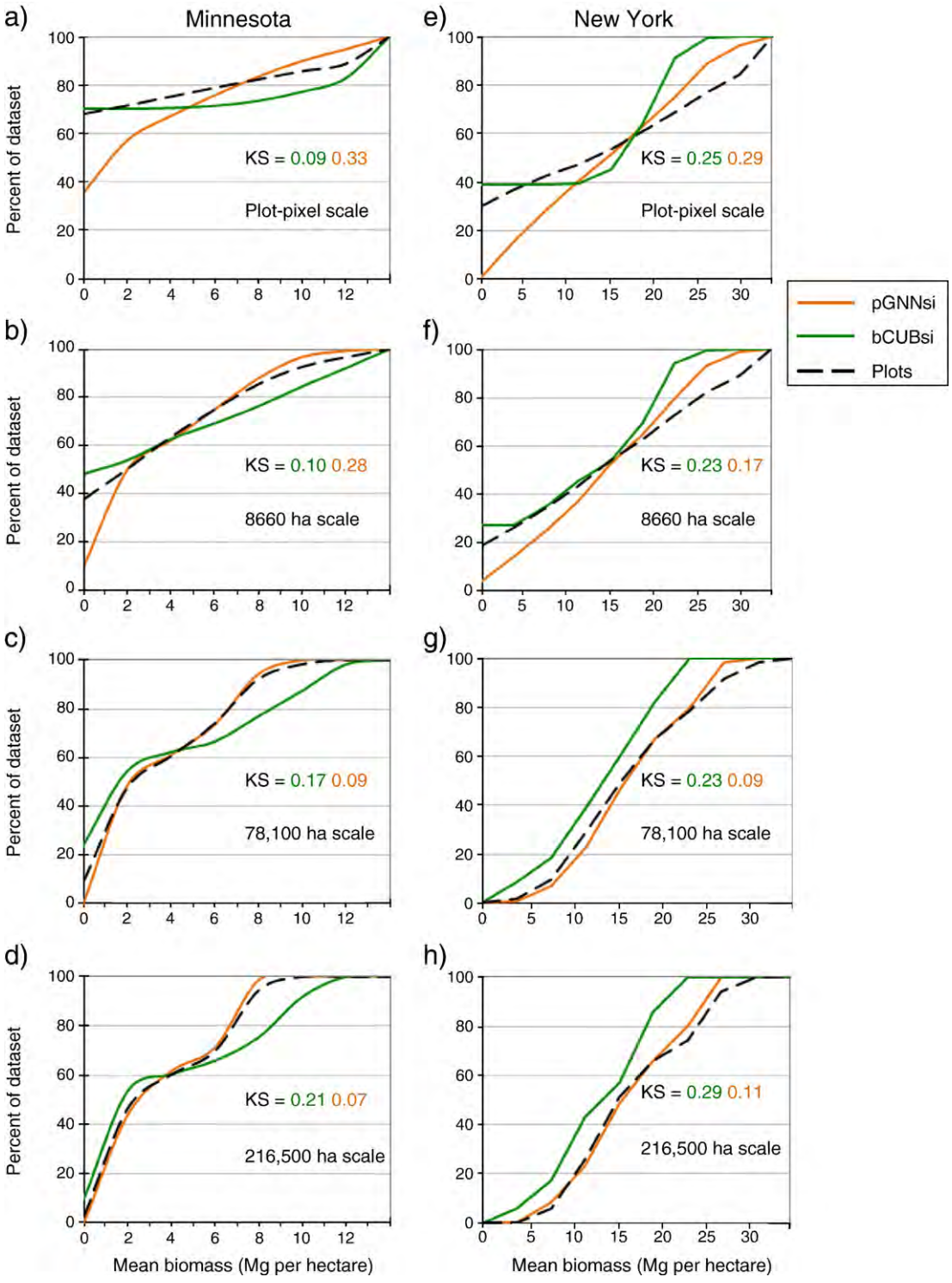

Fig. 3. Empirical cumulative distribution functions (ecdf) of each geospatial dataset in Minnesota (a–d) and New York (e–h) at all four scales: plot:pixel (a,e), 8660 ha (b,f), 78,100 ha (c,g), and 216,500 ha (d,h).

scatter about the GMFR line decreased with scale as expected, resulting in increasing AC_{uns} values with scale as well. Again, we interpreted the steady increase in AC_{sys} values all the way to 1.0 with increasing scale to be an indication that the differences present at the plot–pixel scale and 8660 ha scale were dominated by differences in spatial support (.067 ha vs. 6.25 ha) and uncertainty in the reference dataset, rather than differences between the modeled dataset and the true population from which the reference data are a sample. Put another way, the results of the assessment at the 78,100 and 216,500 ha scales at which we have reasonable confidence in estimates derived from the FIA plots provided little indication that the pGNN dataset contained substantial error, and if we assume that the mathematical relationship between the plots and the predictors developed for the modeling applies across scales, that confidence could be cautiously transferred down to the lower scales as well. From this assessment, the population depicted in the modeled dataset (pGNN) appeared to be very similar to the population on the ground—i.e. that sampled by the FIA plots. In contrast, AC_{sys} values for the

Table 2
Summary statistics comparing the bCUBsi and pGNNsi datasets at all four scales in Minnesota and New York.

	Minnesota		New York	
	bCUBsi	pGNNsi	bCUBsi	pGNNsi
plot:pixel	$n=17{,}883$		$n=2979$	
KS[a]	0.09	0.33	0.25	0.29
AC[b]	0.24	−0.33	0.22	−0.24
AC_{sys}[c]	0.99	0.75	0.85	0.75
AC_{uns}[d]	0.25	−0.08	0.37	0.01
Diff in means[e]	0.40	0.03	−2.62	−0.04
r^2[f]	0.43	0.37	0.34	0.37
RMSE[g]	4.85	4.81	11.66	11.01
8660 ha scale	#hexes = 2357 avg #plots per hex = 7		#hexes = 1282 avg #plots per hex = 2	
KS	0.10	0.28	0.23	0.17
AC	0.67	0.64	0.36	−0.01
AC_{sys}	0.96	0.97	0.86	0.81
AC_{uns}	0.71	0.67	0.51	0.18
Diff in means	0.41	−0.01	−2.72	0.15
r^2	0.71	0.72	0.43	0.44
RMSE	2.34	1.89	9.33	8.76
78,100 ha scale	#hexes = 230 avg #plots per hex = 64		#hexes = 118 avg #plots per hex = 18	
KS	0.17	0.09	0.23	0.09
AC	0.83	0.94	0.71	0.76
AC_{sys}	0.91	0.998	0.85	0.98
AC_{uns}	0.92	0.94	0.87	0.79
Diff in means	0.38	−0.06	−3.19	0.26
r^2	0.90	0.94	0.74	0.80
RMSE	1.47	0.70	4.60	2.91
216,500 ha scale	#hexes = 72 avg #plots per hex = 178		#hexes = 35 avg #plots per hex = 51	
KS	0.21	0.07	0.29	0.11
AC	0.86	0.999	0.80	0.89
AC_{sys}	0.91	0.998	0.84	0.99
AC_{uns}	0.95	0.98	0.96	0.90
Diff in means	0.37	−0.05	−3.30	0.10
r^2	0.94	0.98	0.90	0.90
RMSE	1.29	0.40	3.90	1.97

[a] Range: 0–1; 0 = perfect agreement.
[b] Range: ≤0–1; 1 = perfect agreement, ≤0 = no relationship.
[c] Range: 0–1; 1 = GMFR regression line is identical to the 1:1 line.
[d] Range: 0–1; 1 = no scatter; all points line up on the GMFR regression line.
[e] Positive = relative overestimation (Mg/ha).
[f] Range: 0–1; 1 = perfect agreement.
[g] range: 0–unbounded; 0 = perfect agreement.

modeled bCUB dataset decreased in MN with increasing scale and remained relatively low at all scales in NY. We interpret this type of behavior in the AC_{sys} values—decreasing or remaining steady off the 1:1 line—to be an indication that there is likely to be real error present over and above expected differences due to differences in spatial support and uncertainty in the reference dataset. This interpretation is stronger when the assessment includes scales at which there are sufficient ground plots to be reasonably confident of reference data estimates. The distance and direction of the GMFR line from the 1:1 line provides an approximate indication of the direction and magnitude of the difference between the datasets at that scale. And if the GMFR line crosses the 1:1 line, it indicates that the direction of relative over- and underestimation changes at a threshold mean biomass/ha value.

Table 2 also presents the r^2 and RMSE values for each dataset and scale. Our results support those found in Ji and Gallo (2006) that the r^2 value does not reflect the level of systematic difference present in the relationship and simply increases with increasing scale. Root mean square error (RMSE) values reflect the effects of both systematic and unsystematic difference and provide valuable information regarding differences between the datasets in data units, however it is impossible to tell, from the RMSE value alone, what proportion of the difference is due to systematic versus unsystematic disagreement.

Mapping the resulting estimates at each scale provides a quick and useful visualization of differences in the spatial pattern of area estimates. Figs. 5 and 6 present mapped results of the 8660 ha estimates for MN and NY, respectively. Differences between mean biomass estimates calculated from the sample-intensity-corrected modeled datasets (a and b), and those calculated from the FIA plot data (c) reflect a combination of factors: differences in spatial support (0.067 ha plots vs. 6.25 ha modeled pixels), uncertainty in the reference dataset, and error in the modeled dataset. Figs. 5d–e and 6d–e illustrate the effect of not making the sampling intensity correction. Both (d) and (e) appear more different from the FIA plot results than is in fact true if sampling intensities are matched. To generate a picture of the landscape as a dataset user, one would use the approach used in figure (d) and (e) to take advantage of the full population of modeled estimates (one for every 6.25 ha). However, to assess how different a modeled dataset is from the FIA plot dataset, one must make the comparison using the same sampling intensity, particularly when assessing local spatial variability (figures a and b) and data distribution at the finer spatial scales.

3.2.4. Assessment of differences in local means at the 216,500 ha scale

In the scatterplot comparisons across scales and in the assessment of change in AC_{sys} values with increasing scale, we were able to interpret whether there was likely to be error present in the modeled dataset over and above differences due to mismatches in spatial support or uncertainty in the reference dataset. We could also identify the direction of that potential error from the location of the GMFR line with respect to the 1:1 line. Examining this difference with respect to calculated confidence intervals associated with the reference data provides additional information to identify where the observed difference is likely to be real or whether it is less than the uncertainty associated with the sample-based estimate.

Fig. 7 plots differences in estimated biomass per hectare values within each 216,500-ha area in relation to the plot-based estimate and 90% confidence interval. From Fig. 7, it is readily apparent how the bCUBsi estimates fall substantially above even the 90% plot CI in most of the areas with mean plot biomass greater than 5.5 Mg/ha, and frequently fall below the 90% plot CI between plot biomass means of 1–3 Mg/ha. In contrast, the pGNNsi estimates generally fall within the 90% plot CI, and appear relatively evenly distributed above and below the plot-based estimates. Fig. 8 presents this information in map form, where the bCUBsi results show stronger regional patterns and higher levels of relative overestimation and underestimation than the pGNNsi dataset in both MN and NY. Mean biomass/ha values calculated from the pGNNsi dataset fall within the 95% confidence interval associated with the plot means for MN 88% of the time. Estimates of mean biomass calculated from the bCUBsi dataset for MN fall inside this 95% confidence interval 58% of the time. From Fig. 7 one can also readily identify how the plot CI values increase with higher biomass values, reflecting higher local variation in these areas. The magnitude of the overestimation or underestimation error can be identified only where uncertainty in the reference dataset has been identified as sufficiently small.

Expressing dataset differences in terms of plot-based confidence intervals explicitly integrates into the assessment the uncertainty in our reference dataset, simultaneously alerting us to our level of confidence in these estimates and where additional reference data would be most useful. Presenting the results in map form indicates spatially the direction and approximate magnitude of bias in the modeled datasets for each hex area, while Fig. 7 highlights where in the distribution of values the overestimation and underestimation are occurring.

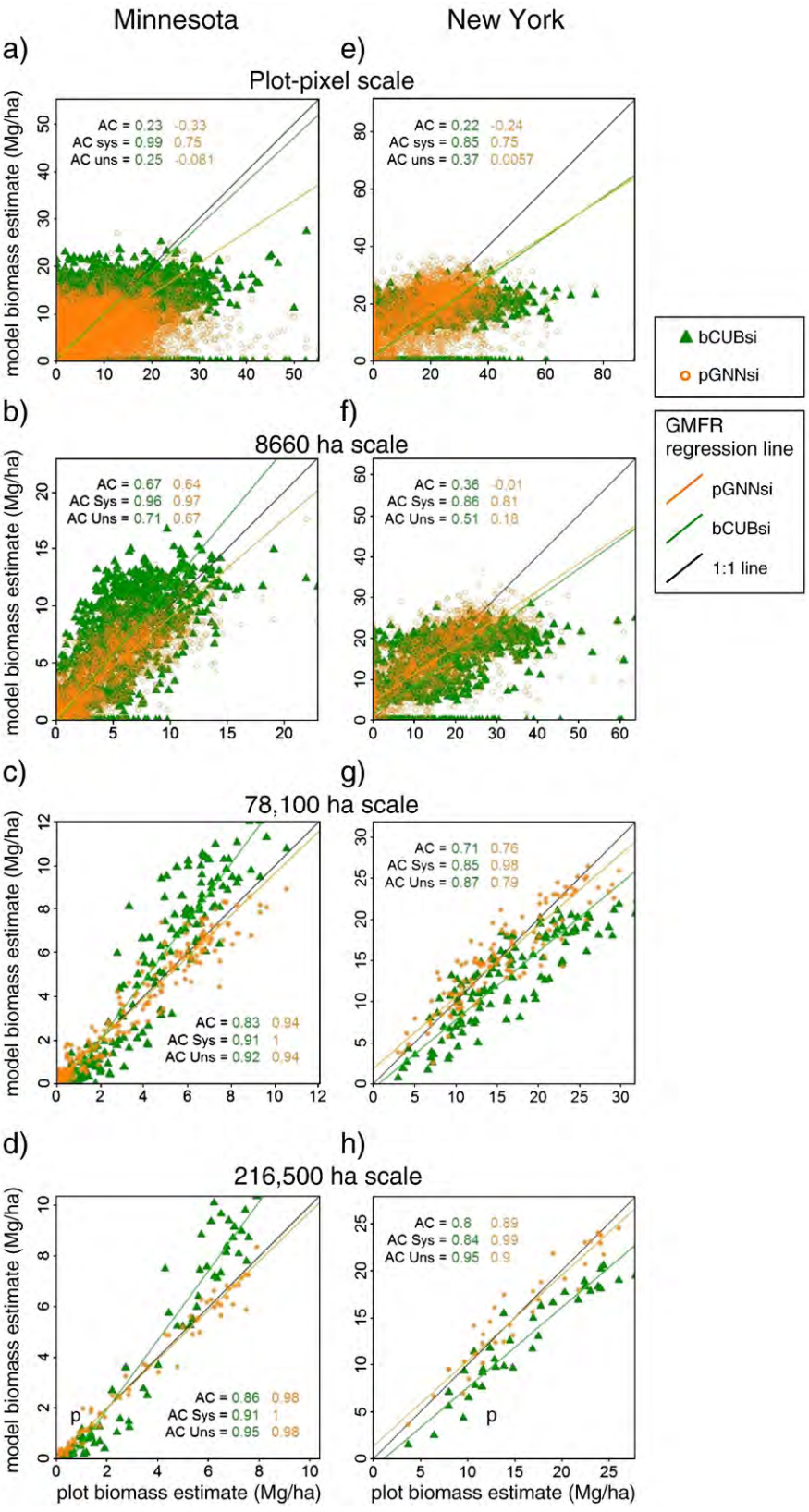

Fig. 4. Scatterplots of modeled estimates vs. plot-derived estimates for Minnesota (a–d) and New York (e–h) at all four scales: plot:pixel (a,e), 8660 ha (b,f), 78,100 ha (c,g) and 216,500 ha (d,h).

3.2.5. Assessment of differences in local variability at the 216,500 ha scale

In addition to biomass totals and means, local spatial variability is an important factor affecting both management and research applications and contributing to the uncertainty associated with FIA plot-based estimates in each area. Local spatial variability is a characteristic of spatial datasets that is sometimes compromised in an effort to improve area means. In this study we compared the local variability of each geospatial dataset in terms of the standard

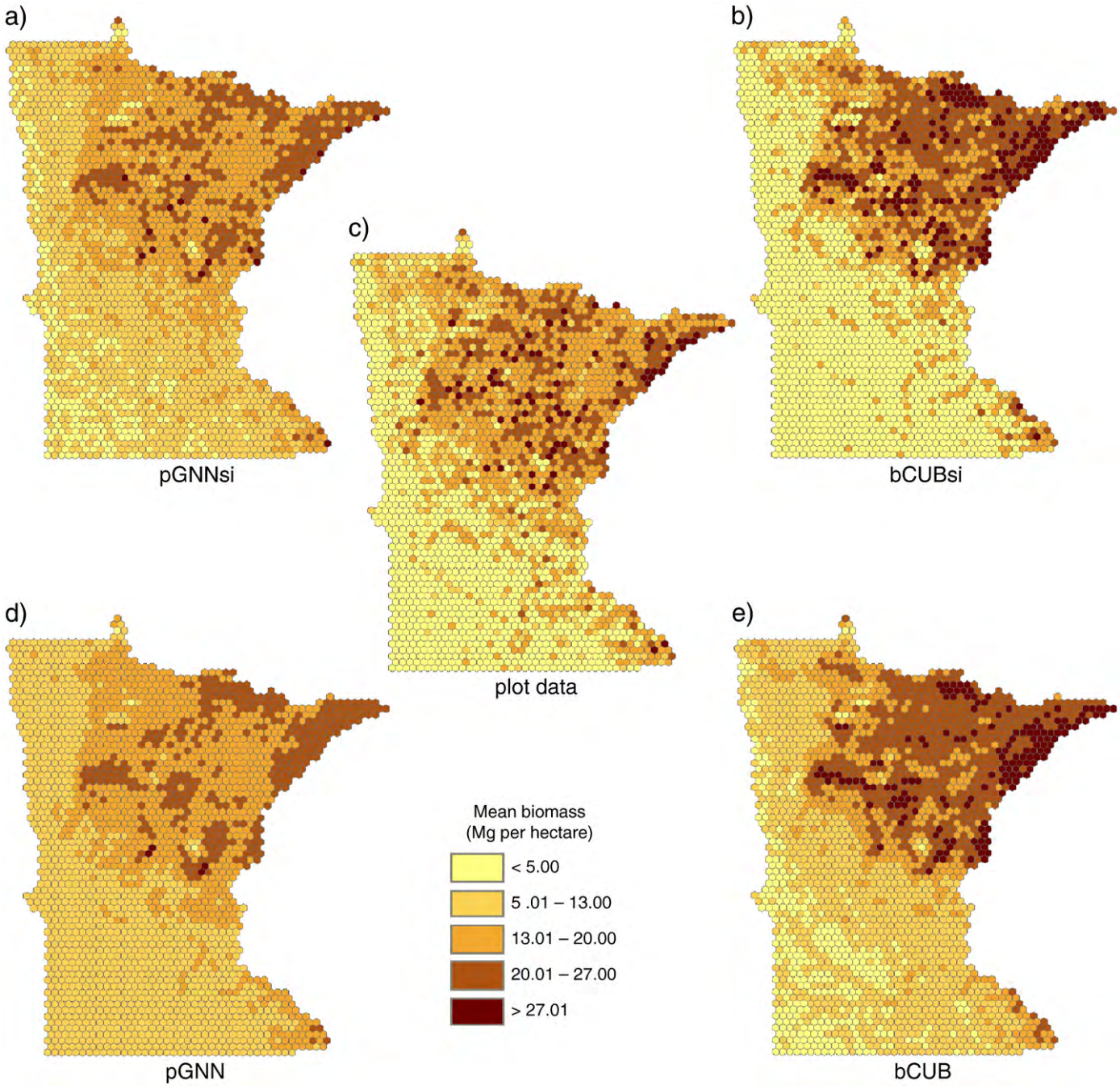

Fig. 5. Mapped results of area estimates at the 8660 ha scale for Minnesota, including summaries calculated from the full datasets (a) pGNN and (b) bCUB, summaries calculated from (c) the plot data, and summaries calculated from the sample-intensity corrected datasets (d) pGNNsi and (e) bCUBsi.

deviation of plot:pixel biomass per hectare values within each 216,500 ha area (Fig. 9). This statistic in particular is strongly affected by the mismatch in sample unit size—the 0.067 ha area measured by the FIA plot vs. the 6.25 ha area covered by the model input data and output estimate. In this example assessment, the local variability of the bCUBsi dataset was more similar to the FIA plots than was the pGNNsi dataset in either MN or NY, at this scale.

4. Discussion

4.1. How did this suite of assessments perform?

The suite of accuracy assessment metrics presented here provided a substantial amount of information on the nature, location, magnitude, and frequency of error in the geospatial datasets examined, picking up the four important characteristics of error identified by Foody (2002). From the information provided one could discern where the error was occurring, both spatially (Figs. 5–6 and 8–9) and in what portions of the distribution (Fig. 7). We could also begin to identify how much of this observed error was an artifact of differences in spatial support or reference dataset uncertainty and how much was probably the result of real differences between the modeled dataset and the true population on the ground. The assessment provided information regarding the magnitude and direction of that error across the range of distribution of biomass values (Figs. 4, 7) or in different regions of each state (Fig. 8). With these assessments we could identify the type of error—whether it was in terms of biomass estimates (Figs. 4–6 and 7–8), the frequency distribution of values (Fig. 3), or in the local spatial variability of the

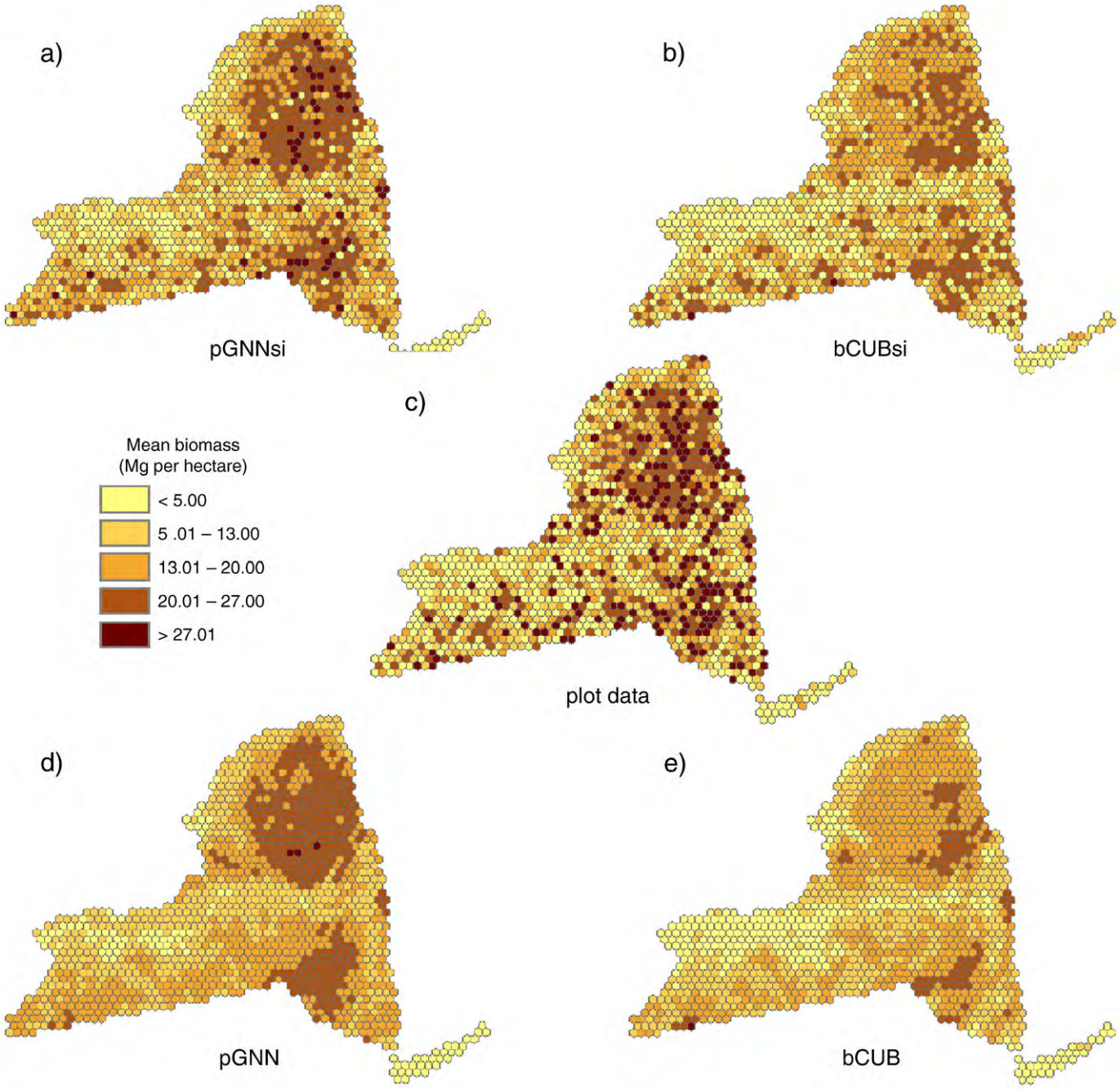

Fig. 6. Mapped results of area estimates at the 8660 ha scale for New York, including summaries calculated from the full datasets (a) pGNN and (b) bCUB, summaries calculated from (c) the plot data, and summaries calculated from the sample-intensity corrected datasets (d) pGNNsi and (e) bCUBsi.

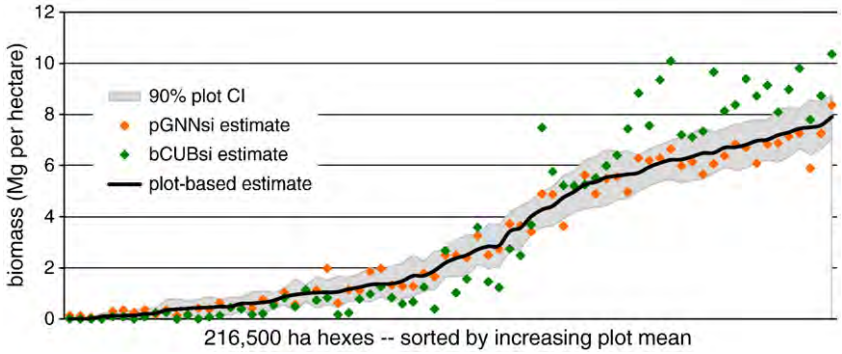

Fig. 7. Differences in modeled area estimates with respect to plot-based confidence intervals at the 216,500 ha (50 k hex) scale in Minnesota.

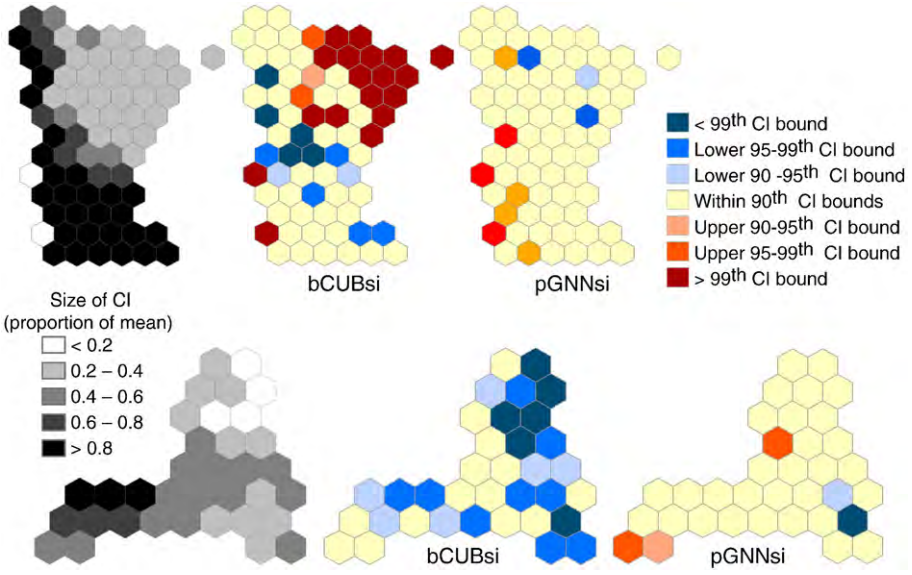

Fig. 8. Mapped differences in modeled area estimates at the 216,500 ha (50 k hex) scale with respect to plot-based confidence intervals.

modeled geospatial dataset (Figs. 5–6 and 9). We could identify at what scale(s) these errors persisted or changed, and what proportion of the relative error could be attributed to systematic vs. unsystematic differences (Fig. 4 and Table 2). Often a single error type, such as the tendency to overestimate the high biomass values and underestimate the low values in MN, appeared in several assessments such as the ecdf plots, the scatterplots, and the maps. When multiple metrics or metrics at multiple scales reported the same results, this redundancy of information allowed corroboration of the results.

If one had to choose just one scale at which to examine geospatial datasets using FIA plots, the 216,500 ha scale appears to provide useful and readily interpretable results. Even at 60% of current standard FIA sampling intensities, this size area captures an average of 50 plots per hex, which provides a robust, spatial (Fig. 8), graphic (Fig. 7), and statistical examination of where a modeled dataset is overestimating or underestimating the real population with respect to aboveground forest biomass. Similarly, ecdf and scatterplot assessments at this scale provide a reasonably clear assessment of dataset differences because of our confidence in the FIA plot estimates at this scale.

For areas smaller than 216,500 ha in NY and likely 78,100 ha in MN (due to double-intensity samples available there), the accuracy assessment becomes more comparative and some of the observed difference between model-based and FIA plot-based estimates will be due to differences in sample unit size mismatch and reference data uncertainty rather than real error in the model-based estimate. Assessments at these scales are informative, particularly if they match the scale of application, and can highlight and quantify artifacts in the dataset resulting from decisions made in dataset development, such as whether to use a forest/nonforest mask. However, at the same time they need to be interpreted in association with results from broader and finer scales. For example, when an error present at the 216,500 ha scale also occurs at finer scales, it is more likely to indicate that the difference observed at the finer scale contains a large component of real bias over and above uncertainty in the plot-based estimate. For assessments of 250-m modeled datasets, plot:pixel comparative assessments are less valuable than area comparisons and may be inappropriate when a large amount of local spatial variance is present (Xu et al., 2009). Whatever the reference dataset used, assessing

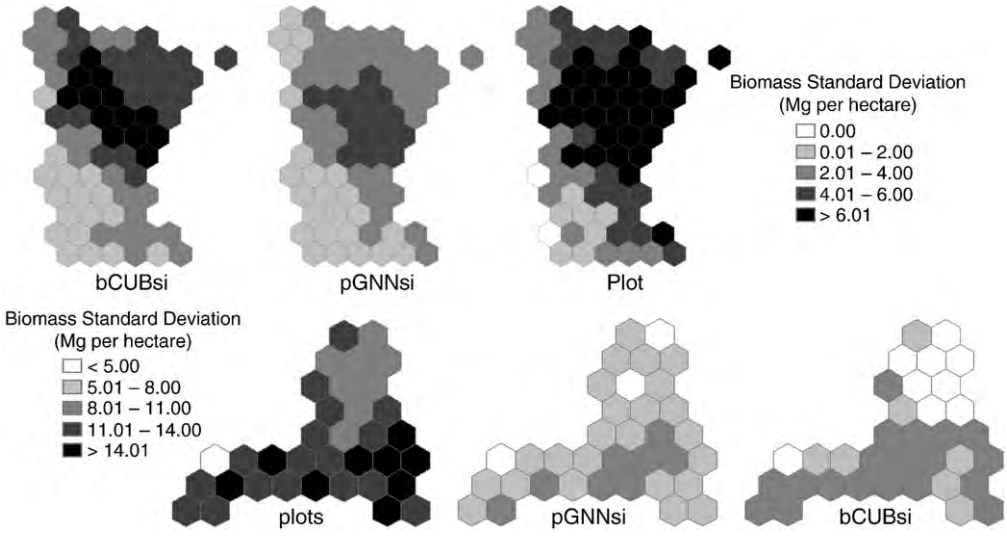

Fig. 9. Mapped results of local variability of plot (or sample pixel) values in each 216,500 ha area.

datasets at multiple scales appears to improve our ability to distinguish these artifacts, and increase our understanding of the source and persistence of each type of disagreement, bringing us closer to understanding the real error in the modeled dataset—its difference from the population on the ground—in addition to its relative error with respect to the reference dataset.

Addressing the need for timely accuracy assessment results, we were able to calculate these assessments relatively quickly and easily because we were using existing FIA plot data, corrected for the mismatch in sampling intensity. Using a ground plot reference dataset collected using an equal probability sampling design, with its ability to provide unbiased estimates and known sampling error, allowed us to clearly identify the levels of uncertainty present and use the reference dataset error in our assessments. A mismatch in sample unit size between the reference and modeled datasets is always an essential consideration in any assessment. By examining assessments at several scales, including one at which the plot confidence intervals are considered acceptable for this location and variable, one has a strong reference point from which to tease apart real difference from comparative uncertainty in the observed error. The protocol here makes use of several relatively basic assessments whose strength comes in their application together. This is important when a geospatial dataset can be very similar to FIA plot data in one respect (and report high comparative accuracies) and may differ substantially in another.

One aspect this study did not tackle was the *significance* of the errors, primarily because significance is defined in terms of a particular application. For example, Holden et al. (2003) examined the significance of forest/nonforest error with respect to its impact on stratified estimation, and Shao and Jianguo (2008) noted how landscape index values are altered by both the magnitude and spatial distribution of classification errors. Understanding what types of error affect a particular application, what magnitudes of error translate into a significant impact on the results, and whether the error is occurring over a critical or less important range of values is the next step in interpreting assessment results for a specific application.

4.2. The potential for modeled datasets to be more accurate than the reference data

FIA plots are a rich and robust source of data on forest composition and structure. Alone they are able to provide design-based, statistical estimates of the forest resource without assumptions and with standard errors. When this plot information is combined with remotely sensed imagery and other relevant geospatial datasets in a modeling framework, it allows for the estimation of forest characteristics at much finer scales. These modeled geospatial datasets thus have the potential to provide more information in some areas than the FIA plot estimates because they supplement the limits of sampling intensity with additional information at all unsampled locations. This can allow for increased spatial resolution of estimates of forest characteristics and potentially more accurate area estimates of biomass at scales where the standard error of plot estimates is too high (Ghosh & Rao, 1994; Rao, 2003). This is particularly likely where the predictor data are of sufficient resolution and quality, where there is a strong relationship between the forest characteristics and the predictor data being used, and assumptions are well-founded. Complete assessments such as the protocol presented in this paper begin to offer some insight into where a modeled dataset is consistent with FIA plot estimates at scales we have confidence in, and where it differs in ways that suggest it could be adding to what we know about the true population from the FIA data alone.

4.3. Additional benefit of model uncertainty layers

Some modeling techniques generate an uncertainty layer along with the modeled output, which can provide valuable per-pixel information both to the user community and to the iterative improvement of the dataset by its developers (Aspinall 2002; Blackard et al., 2008; Hershey, 2000; Strahler et al., 2006; Woodbury et al., 1998; Woodbury, 2003). These modeled uncertainty values will be affected largely by the quality of the input layers and the relationship between the features they describe and the variable being modeled (e.g. biomass). An accuracy assessment of these datasets may (and actually should if it's a quality uncertainty layer) corroborate the information provided in the uncertainty layer(s). If so, the uncertainty layer itself can be used to provide additional information on the fine-grained spatial distribution of error, and if not, the uncertainty layer should probably be questioned.

5. Conclusions

To use any geospatial dataset effectively, one must have a comprehensive and effective accuracy assessment that provides information regarding as many characteristics of the dataset as will be used in subsequent analyses or decision-making. To date, assessments of continuous variables have tended to focus on one or two metrics, like RMSE, which provide the user with little information when used alone. Complementing RMSE, the assessments and measures proposed in this paper provide information on the location, magnitude, frequency, and type of error in geospatial datasets of continuous variables. Many of these assessments are amenable to automation using scripting in commonly-used software like R or SAS. The ease of application of this suite of assessments and their general purpose coverage should facilitate their consistent application, addressing concern that methods for accuracy assessment must be consistent if they are to be readily comparable. With this protocol one should be able to assess the accuracy of any modeled dataset of continuous variables regardless of its resolution, with the only limit being the ability of the reference data (in this case FIA plots) to accurately characterize the landscape at small scales and in nonforest areas. In addition, these measures can be applied to many variables that have been traditionally modeled as categorical classes, such as “forest” vs. “nonforest,” but which are now often being more effectively and flexibly modeled and used in the form of a continuous variable such as percentage of forest (Blackard et al., 2008; Cohen et al., 2001). It is worth noting that reference data collected with sufficient sampling intensity using an equal probability sampling design can be most readily used with these assessments. The assessment framework provided here should help both researchers and managers improve and more effectively use modeled map products.

Acknowledgements

The authors wish to thank Ray Czaplewski, Linda Heath, Eileen Helmer, and four anonymous reviewers for their thoughtful reviews which substantially improved the manuscript. The authors also wish to thank everyone on FIA's Accuracy Assessment team for invaluable comments and contribution during concept development.

References

Aspinall, R. (2002). A land-cover data infrastructure for measurement, modeling, and analysis of land-cover change dynamics. *Photogrammetric Engineering and Remote Sensing*, *68*(10), 1101−1105.

Bastin, L., Fisher, P. F., & Wood, J. (2002). Visualizing uncertainty in multi-spectral remotely sensed imagery. *Computers & Geosciences*, *28*(3), 337−350.

Bechtold, W. A., & Patterson, C. T. (2005). The forest inventory and analysis plot design. In W. A. Bechtold, & P. L. Patterson (Eds.), *The enhanced forest inventory and analysis program–National sampling design and estimation procedures. Gen. Tech. Rep. SRS-80* (pp. 11−26). Asheville, NC: US Department of Agriculture, Forest Service, Southern Research Station.

Binaghi, E., Brivio, P., Ghezzi, P., & Rampini, A. (1999). A fuzzy set-based accuracy assessment of soft classification. *Pattern Recognition Letters*, *20*(9), 935−948.

Blackard, J. A., Finco, M. V., Helmer, E. H., Holden, G. R., Hoppus, M. L., Jacobs, D. M., et al. (2008). Mapping U.S. forest biomass using nationwide forest inventory data and moderate resolution information. *Remote Sensing of Environment*, *112*(4), 1658−1677.

Canters, F. (1997). Evaluating the uncertainty of area estimates derived from fuzzy land-cover classification. *PE&RS, Photogrammetric Engineering & Remote Sensing*, *63*, 403−414.

Cohen, W. B., Maiersperger, T. K., Spies, T. A., & Oetter, D. R. (2001). Modelling forest cover attributes as continuous variables in a regional context with Thematic Mapper data. *International Journal of Remote Sensing*, *22*(12), 2279−2310.

Cohen, W., & Goward, S. (2004). Landsat's role in ecological applications of remote sensing. *BioScience*, *54*(6), 535−545.

Congalton, R. G. (2001). Accuracy assessment and validation of remotely sensed and other spatial information. *International Journal of Wildland Fire*, *10*, 321−328.

Congalton, R. G. (2004). Putting the map back in map accuracy assessment. In R. S. Lunetta, & J. G. Lyon (Eds.), *Remote Sensing and GIS Accuracy Assessment* (pp. 1−11). Boca Raton, FL: CRC Press.

Congalton, R. G., & Plourde, L. C. (2000). Sampling methodology, sample placement, and other important factors in assessing the accuracy of remotely sensed forest maps. In G. B. M. Heuvelink, & M. J. P. M. Lemmens (Eds.), *Proceedings of the 4th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences* (pp. 117−124). The Netherlands: Delft University Press.

DeGloria, S. D., Laba, M., & Gregory, S. K. (2001). Labeled remotely sensed imagery: fuzzy and conventional accuracy assessment. Mapping the state of New York from Landsat TM. *GIM International*, *15*(2), 76-69.

Draper, N. R., & Smith, H. (1998). *Applied regression analysis.* New York: John Wiley and Sons.

Fassnacht, K., Cohen, W., & Spies, T. (2006). Key issues in making and using satellite-based maps in ecology: A primer. *Forest Ecology and Management*, *222*, 167−181.

Federal Geographic Data Committee. (1998). *FGDC-STD-001-1998. Content standard for digital geospatial metadata (revised June 1998)*. Washington, DC: Federal Geographic Data Committee.

Foody, G. M. (2002). Status of land cover classification accuracy assessment. *Remote Sensing of Environment*, *80*(1), 185−201.

Galbraith, J. M., Donovan, P. F., Smith, K. M., & Zipper, C. E. (2003). Using public domain data to aid in field identification of hydric soils. *Soil Science*, *168*(8), 563−575.

Ghosh, M., & Rao, J. N. K. (1994). Small area estimation: An appraisal. *Statistical Science*, *9*, 55−76.

Goodchild, M., Haining, R., & Wise, S. (1992). Integrating GIS and spatial data analysis: problems and possibilities. *International Journal of Geographical Information Science*, *6*(5), 407−423.

Hammond, T. O., & Verbyla, D. L. (1996). Optimistic bias in classification accuracy assessment. *International Journal of Remote Sensing, 17*(6), 1261−1266.

Hershey, R. R. (2000). Modeling the spatial distribution of ten tree species in Pennsylvania. In H. T. Mowrer, & R. G. Congalton (Eds.), *Quantifying spatial uncertainty in natural resources* (pp. 119−135). Chelsea, MI: Ann Arbor Press.

Hershey, R.R., and Reese, G. (1999). Creating a "first-cut" species distribution map for large areas from forest inventory data. Gen. Tech. Rep. NE-256. Radnor, PA: US Department of Agriculture, Forest Service, Northeastern Research Station.

Holden, G. R., Nelson, M. D., & McRoberts, R. E. (2003). Accuracy assessment of FIA's nationwide biomass mapping products: Results from the North Central FIA region. *Proceedings of the fifth annual forest inventory and analysis symposium. Gen. Tech. Rep. WO-69* (pp. 139−147). Washington, DC: U.S. Department of Agriculture Forest Service.

Janssen, L. L. F., & van der Wel, F. J. M. (1994). Accuracy assessment of satellite-derived land cover data—A review. *Photogrammetric Engineering and Remote Sensing*, *60*(4), 419−426.

Ji, L., & Gallo, K. (2006). An agreement coefficient for image comparison. *Photogrammetric Engineering and Remote Sensing*, *72*(7), 823−833.

Justice, C. O., Townshend, J. R. G., Vermote, E. F., Masuoka, E., Wolfe, R. E., Saleous, N., et al. (2002). An overview of MODIS Land data processing and product status. *Remote Sensing of Environment*, *83*, 3−15.

Laba, M., Gregory, S. K., Braden, J., Ogurcak, D., Hill, E., Fegraus, E., et al. (2002). Conventional and fuzzy accuracy assessment of the New York Gap Analysis Project land cover map. *Remote Sensing of Environment*, *81*, 443−455.

Liu, Y. S., & Journel, A. G. (2009). A package for geostatistical integration of coarse and fine scale data. *Computers & Geosciences*, *35*(3), 527−547.

Lopes, R. H. C., Reid, I., & Hobson, P. R. (2007). The two-dimensional Kolmogorov–Smirnov test. *XI International Workshop on Advanced Computing and Analysis Techniques in Physics Research (April 23–27, 2007)*. the Netherlands: Amsterdam.

Lowell, K. (2001). An area-based accuracy assessment methodology for digital change maps. *International Journal of Remote Sensing*, *22*(17), 3571−3596.

Lunetta, R. S., & Lyon, J. G. (2004). *Remote sensing and GIS accuracy assessment.* Boca Raton: CRC Press.

Mark, D. M., & Church, M. (1977). On the misuse of regression in earth science. *Mathematical Geology*, *9*(1), 63−75.

Mayaux, P., & Lambin, E. F. (1995). Estimation of tropical forest area from coarse spatial resolution data: a two-step correction function for proportional errors due to spatial aggregation. *Remote Sensing of the Environment.*, *53*, 1−15.

Miles, P.D. Tue Mar 23 15:13:39 CDT. (2010). Forest Inventory EVALIDator web-application version 4.01 beta. St. Paul, MN: U.S. Department of Agriculture, Forest Service, Northern Research Station.

Moeur, M., & Riemann Hershey, R. (1999). Preserving spatial and attribute correlation in the interpolation of forest inventory data. In K. Lowell, & A. Jaton (Eds.), *Spatial accuracy assessment–Land information uncertainty in natural resources* (pp. 419−430). Chelsea, MI: Sleeping Bear Press/Ann Arbor Press.

Moody, A., & Woodcock, C. E. (1994). Scale-dependent errors in the estimation of land-cover proportions—Implications for global land-cover datasets. *Photogrammetric Engineering and Remote Sensing*, *60*, 585−594.

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. *International Journal of Remote Sensing*, *30*(8), 1913−1940.

Ohmann, J. L., & Gregory, M. J. (2002). Predictive mapping of forest composition and structure with direct gradient analysis and nearest-neighbor imputation in coastal Oregon, U.S.A. *Canadian Journal of Forest Research*, *32*(4), 725−741.

Pontius, R. G. (2002). Statistical methods to partition effects of quantity and location during comparison of categorical maps at multiple resolutions. *Photogrammetric Engineering and Remote Sensing*, *68*(10), 1041−1049.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*, 81−106.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning.* San Mateo, CA: Morgan Kaufman Publishers.

Rao, J. N. K. (2003). *Small area estimation.* Hoboken, New Jersey: John Wiley and Sons.

Ricker, W. E. (1984). Computation and uses of central trend lines. *Canadian Journal of Zoology*, *62*, 1897−1905.

Riemann, R., Hoppus, M., & Lister, A. (2000). Using arrays of small ground sample plots to assess the accuracy of Landsat TM-derived forest cover maps. *Accuracy 2000: Proceedings of the 4th international symposium on spatial accuracy assessment in natural resources and environmental sciences*. The Netherlands: Amsterdam.

Scott, C. T., Bechtold, W. A., Reams, G. A., Smith, W. D., Westfall, J. A., Hansen, M. H., et al. (2005). Sample-based estimators used by the forest inventory and analysis national information management system. In W. A. Bechtold, & P. L. Patterson (Eds.), *The enhanced forest inventory and analysis program–National sampling design and estimation procedures. Gen. Tech. Rep. SRS-80* (pp. 27−42). Asheville, NC: US Department of Agriculture, Forest Service, Southern Research Station.

Shao, G., & Jianguo, W. (2008). On the accuracy of landscape pattern analysis using remote sensing data. *Landscape Ecology*, *23*(5), 505−511.

Smith, J., Prisley, S., & Weih, R. (1991). Considering the effect of spatial variability on the outcomes of forest management decisions. *Proceedings of GIS/LIS 91* (pp. 286−292). Atlanta, GA.

Smith, J. H., Wickham, J. D., Stehman, S. V., & Yang, L. (2002). Impacts of patch size and land-cover heterogeneity on thematic image classification accuracy. *Photogrammetric Engineering & Remote Sensing*, *68*(1), 65−70.

Smith, J. H., Stehman, S. V., Wickham, J. D., & Yang, L. M. (2003). Effects of landscape characteristics on land-cover class accuracy. *Remote Sensing of Environment*, *84*(3), 342−349.

Stehman, S. V., Arora, M., Kasetkasem, T., & Varshney, P. (2007). Estimation of fuzzy error matrix accuracy measures under stratified random sampling. *Photogrammetric Engineering & Remote Sensing*, *73*(2), 165−174.

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 Sensing of Environment*, *86*(4), 500−516.

Strahler, A. H., Boschetti, L., Foody, G. M., Friedl, M. A., Hansen, M. C., Herold, M., et al. (2006). *Global land cover validation: Recommendations for evaluation and accuracy assessment of global land cover maps. Report no. 25. Luxembourg: Global observation of forest and land cover dynamics (GOFC-GOLD)*.

Townshend, J. R. G., Justice, C. O., Gurney, C., & McManus, J. (1992). The impact of misregistration on change detection. *IEEE Transactions on Geoscience and Remote Sensing*, *30*, 1054−1060.

U.S. Forest Service. 2003. Forest inventory and analysis national core field guide, volume 1: field data collection procedures for phase 2 plots, Version 2.0. Washington, DC: US Department of Agriculture, Forest Service, Forest Inventory and Analysis.

Vogelmann, J. E., Howard, S., Yang, L., Larson, C., Wylie, B., & Van Driel, N. (2001). Completion of the 1990s National Land Cover Data set for the conterminous United States from Landsat Thematic Mapper data and ancillary data sources. *Photogrammetric Engineering and Remote Sensing*, *67*, 650−661.

Wardlow, B. D., & Egbert, S. L. (2003). A state-level comparative analysis of the GAP and NLCD land-cover data sets. *Photogrammetric Engineering & Remote Sensing*, *69*(12), 1387−1397.

White, M. A., Shaw, J. D., & Ramsey, R. D. (2005). Accuracy assessment of the vegetation continuous field tree cover product using 3954 ground plots in the south-western USA. *International Journal of Remote Sensing*, *26*(12), 2699−2704.

Wickham, J. D., Stehman, S. V., Smith, J. H., Wade, T. G., & Yang, L. (2004). A priori evaluation of two-stage cluster sampling for accuracy assessment of large-area land-cover maps. *International Journal of Remote Sensing*, *25*(6), 1235−1252.

Willmott, C. J. (1981). On the validation of models. *Physical Geography*, *2*(2), 184−194.

Wilson, B. T., Lister, A. J., & Riemann, R. (submitted for publication). A nearest-neighbor imputation approach to large area mapping of tree species distributions using field sampled data. *Remote Sensing of Environment*.

Woodbury, P. B. (2003). Dos and don'ts of spatially explicit ecological risk assessments. *Environmental Toxicology and Chemistry*, *22*(5), 977−982.

Woodbury, P. B., Smith, J. E., Weinstein, D. A., & Laurence, J. A. (1998). Assessing potential climate change effects on loblolly pine growth: A probabilistic regional modeling approach. *Forest Ecology and Management*, *107*, 99−116.

Xu, Y., Dickson, B., Hampton, H., Sisk, T., Palumbo, J., & Prather, J. (2009). Effects of mismatches of scale and location between predictor and response variables on forest structure mapping. *Photogrammetric Engineering and Remote Sensing*, *75*(3), 313−322.

Yang, L., Stehman, S., Smith, J., & Wickham, J. (2001). Thematic accuracy of MRLC land cover for the eastern United States. *Remote Sensing of Environment*, *76*(3), 418−422.

Zhu, Z., Yang, L., Stehman, S. V., & Czaplewski, R. L. (2000). Accuracy assessment for the U. S. Geological Survey regional land-cover mapping program: New York and New Jersey Region. *Photogrammetric Engineering and Remote Sensing*, *66*(12), 1425−1435.