Science of Remote Sensing 9 (2024) 100119

Contents lists available at ScienceDirect

Science of Remote Sensing

journal homepage: www.sciencedirect.com/journal/science-of-remote-sensing

Testing temporal transferability of remote sensing models for large area monitoring

Steven K. Filippelli [a,*], Karen Schleeweis [b], Mark D. Nelson [c], Patrick A. Fekety [a], Jody C. Vogeler [a]

[a] Natural Resource Ecology Laboratory, Colorado State University, Fort Collins, CO 80523-1499, USA
[b] USDA Forest Service, Rocky Mountain Research Station, Riverdale, UT 84405, USA
[c] USDA Forest Service, Northern Research Station, 1992 Folwell Avenue, St. Paul, MN 55108, USA

ARTICLE INFO

Keywords:
Temporal transfer
Temporal cross-validation
Time series
Multiscale
Small area estimation
Model-based
Canopy cover

ABSTRACT

Applying remote sensing models outside the temporal range of their training data, referred to as temporal model transfer, has become common practice for large area monitoring projects that extrapolate models for hindcasting or forecasting to time periods lacking reference data. However, the development of appropriate validation methods for temporal transfer has lagged behind its rapid adoption. Breaking temporal transfer's assumption of temporal consistency in both remote sensing and reference data and their relationship to each other could lead to biased pixel-level predictions and small area estimators, compromising the operational validity of large area monitoring products. Few studies using temporal transfer have evaluated its effects on model accuracy at the pixel/plot level, and the propensity for biased small area estimators and trends resulting from temporal transfer remains unexplored. We present a framework for evaluating temporal transferability by combining temporal cross-validation with a multiscale map assessment to aid in identifying where and when biased model predictions could scale to small area estimates and affect predicted trends.

This validation framework is demonstrated in a case study of annual percent tree canopy cover mapping in Michigan. We tested and compared temporal transferability of random forest models of canopy cover derived from 2010 to 2016 systematic dot-grid photo-interpretations at Forest Inventory and Analysis plots with Landsat spectral indices fit with the LandTrendr temporal segmentation algorithm serving as the primary predictor variables. The temporal cross-validation error (mean RMSE = 13.9% cover) was higher than the common validation approach of considering all time periods of testing data together (RMSE = 13.6% cover) and lower than models trained and tested within the same year (mean RMSE = 14.2% cover). However, the bias of model predictions and small area estimators for individual years was higher with temporal transfer models than when applying models within the same year as their training data. We also evaluated how training models using different temporal subsets and with and without LandTrendr fitting affected predictions of Michigan's 1984–2020 predicted annual mean cover. The mean cover from LandTrendr-based models followed expected and consistent trends and had less difference between models trained with different temporal subsets (max difference = 5.8% cover). While those from Landsat had high interannual variations and greater difference between temporal models (max difference = 11.2% cover). The results of this case study demonstrate that evaluation of temporal transferability is necessary for establishing the operational validity of large area monitoring products, even when using time series algorithms that improve temporal consistency.

1. Introduction

Remote sensing of land attributes and their change has been undergoing a paradigm shift from two-date change detection to land monitoring, where maps are generated at regular time intervals such as annually (Woodcock et al., 2020). Land monitoring can provide many advantages over single period change detection including alerting decision makers to important recent changes, enabling quantification of trends, and reducing temporal mismatches in analysis with other datasets. The advantages and increased feasibility of land monitoring has led

* Corresponding author.
E-mail address: steven.filippelli@colostate.edu (S.K. Filippelli).

https://doi.org/10.1016/j.srs.2024.100119
Received 17 October 2023; Received in revised form 29 December 2023; Accepted 18 January 2024
Available online 23 January 2024

to a proliferation of programs that aim to provide easy access to maps depicting various land attributes that are generated with a low latency at national to global scales, moderate resolution (10–30 *m*), and high frequency (Boryan et al., 2011; Brown et al., 2020, 2022; Defourny et al., 2019; Friedl et al., 2022; Gong et al., 2020; Hansen et al., 2013; Healey et al., 2018; Karra et al., 2021; Marston et al., 2022; Melton et al., 2021; Potapov et al., 2022; Zanaga et al., 2022). However, land monitoring may require extrapolation of models outside the time period of the training data (i.e., temporal transfer) because of the limited availability of reference data (Botkin et al., 1984; Foody et al., 2003; Minter, 1979; Woodcock et al., 2001). The rapid development of land monitoring programs and products has outpaced the development of appropriate validation frameworks for temporal transfer, which may lead to maps that are unknowingly biased and subsequently causing false inferences.

Developing and validating annual and historical land cover type and change maps under a monitoring paradigm requires reference data for training and testing that are considerably more accurate than the map classification being evaluated (Olofsson et al., 2014). An annual reference dataset enables validation of annual maps and derived maps of change (Cohen et al., 2010; Pengra et al., 2020; Schleeweis et al., 2020; Schroeder et al., 2007, 2014; Thomas et al., 2011; Zhao et al., 2018). For some land attributes, reference data can be obtained from visual interpretation of the same image data used in the mapping process, such as with the TimeSync tool (Cohen et al., 2010). However, it would be impossible to apply this image interpretation process for gathering reference data for most continuous attributes such as forest biomass (Duncanson et al., 2021), evapotranspiration (García-Santos et al., 2022), or soil surface properties (Mulder et al., 2011). In the case of continuous attributes, reference data typically needs to be collected *in situ* or, in some cases, derived from remote sensing data that are limited in time and space, like lidar measurements of forest height. Developing and validating frequently generated maps from a temporally limited reference dataset therefore requires making assumptions about the applicability of a remote sensing model in other time periods.

Applying a remote sensing model outside the spatial or temporal extent of model development has been referred to as “signature extension” (Minter, 1979), “spectral extendibility” (Botkin et al., 1984), “generalization” (Woodcock et al., 2001), and “model transferability” (Foody et al., 2003). In this manuscript we refer to the extrapolation of models outside the spatiotemporal extent of the model development as “model transfer”. The primary assumption underlying model transfer is that the relationship defined by the model between the target attribute (e.g., land cover, forest biomass, etc.) and the remotely sensed data remains the same in time periods or areas outside the spatiotemporal extent of the training data. In spatial model transfer, a model developed from training and remotely sensed data in one region is applied to remotely sensed data in another region outside the spatial extent of the training data (Foody et al., 2003), but which is expected to have similar biological or physical characteristics. Temporal model transfer involves applying a model to a time period outside the temporal extent of the training data. Model transfer is common in application areas where reference data are limited and thus done out of necessity.

Some of the first attempts at model transfer were with Landsat imagery for mapping agriculture as part of the Large Area Crop Inventory Experiment (LACIE) in the 1970s (Minter, 1979) and the AgRISTARS program of the 1980s (NASA, 1983). In the early 2000s the idea of model transfer received renewed attention as it became more technically feasible to work with multiple Landsat scenes for large area monitoring (Cohen et al., 2001; Foody et al., 2003; Pax-Lenney et al., 2001; Woodcock et al., 2001). Some of these tests showed either spatial or temporal model transfer reduced map accuracy. The use of model transfer began to show more promising results once radiometric normalization (Canty et al., 2004; Schroeder et al., 2006), atmospheric correction (Vermote et al., 1997), and cloud masking (Irish et al., 2006) were more thoroughly addressed, and these advances in preprocessing along with temporal compositing and time series algorithms could be applied to the opened Landsat archive (Wulder et al., 2012, 2022).

The success of these advances in pre-processing has in part led to the rise of land monitoring and use of temporal transfer, but there are still numerous factors that reduce the temporal consistency of remote sensing data, reference data, and their relationship to each other, which can hinder temporal transfer. For example, optical remote sensing is also affected by differences in the spectral response across sensors (Roy et al., 2016), variations in the Bi-Directional Reflectance Function (BRDF) over repeat visits (Nagol et al., 2015; Roy et al., 2020), topography (Hantson and Chuvieco, 2011; Riano et al., 2003; Sola et al., 2016), and rare sensor issues and anomalies (“Landsat Known Issues,” n. d.). More importantly, actual changes in biophysical properties not related to the attribute of interest may also hinder temporal transfer, such as interannual variations in phenology (Melaas et al., 2013) and non-seasonal temporary changes in soil moisture (Muller and Décamps, 2001) or vegetation vigor (Ji and Peters, 2003). Reductions in the effectiveness of temporal transfer can also come from changes in the nature and quality of reference data over time due to changes in definitions, protocols, technician competence, instrumentation, or dependent data sources (e.g., Westfall and Patterson, 2007). Furthermore, in an era of rapid environmental change, the relationship between remote sensing data and land attributes may not remain consistent, such as with shifting trends in phenology due to climate change (Cleland et al., 2007). These changes could make common remote sensing metrics less reliable predictors when applied over time.

Even with the above challenges of temporal consistency in both remote sensing and reference data and the growing prevalence of monitoring applications, few studies or remote sensing programs directly address potential issues with temporal transferability. Some Landsat-based studies have tested transferability by training and testing on different spatial or temporal subsets of data (Kim et al., 2014; Majumdar et al., 2020; Nelson et al., 2011; Zhong et al., 2014; Zhou et al., 2020), but the approaches used can vary greatly between studies. When implemented, temporal cross-validation can lead to a better understanding of temporal transfer, such as higher performance of phenological metrics over temporal composites (Zhong et al., 2014), or that error may increase with greater differences in the training and testing data in either average value or the time lag between them (Zhou et al., 2020). Many recent examples of evaluating model transferability also come from the use of lidar and photogrammetric point clouds for modeling forest attributes because of the concern that differences in collection parameters and point cloud characteristics between collections could lead to biased estimates (Coops et al., 2021; Domingo et al., 2019; Fekety et al., 2015, 2018; Mauro et al., 2021; Navarro et al., 2020). Whether in optical or point cloud-based applications, these examples limited their analysis of temporal transfer to model or map error without further examining the consequences of temporal transfer on the spatial distribution of errors or on the temporal trends estimated from the resulting maps.

There is a critical need to develop a common set of practices for evaluating model and map errors with temporal transfer and to understand the effects of these errors on subsequent spatiotemporal analyses. As the monitoring paradigm of remote sensing progresses and both national and global map programs transition to more frequent releases of widely used data sets, there will likely be more map users that analyze and draw conclusions from maps that were developed from the temporal transfer of remote sensing models. One of the most common forms of map analysis is to generate summary statistics for an area of interest through spatial aggregations, also known as zonal statistics. These aggregate statistics may be functionally equivalent to small area estimates made in a model-based inference framework (Rao and Molina, 2015), even if it is not explicitly stated as such. A key tenet of model-based inference is that the estimators for population parameters, like the mean biomass of a study area for example, depend upon correct model specification to be unbiased (Särndal et al., 1992). Even seemingly reasonable models can lead to heavily biased estimators (Hansen et al., 1983). Failing to acknowledge or evaluate the potential for increased bias or decreased precision with temporal transfer could therefore lead to false inferences being drawn from maps, which may have undesirable consequences for decision making.

Remote sensing scientists are faced with the conundrum of needing to apply temporal transfer to the wealth of remote sensing time series data for establishing historical baselines, providing near real time predictions, and quantifying change, while simultaneously having insufficient reference data for conducting traditional map validations and properly quantifying uncertainty across a full time series. In lieu of traditional map validation, the question then becomes: given the reference data available, what can we know about the potential for error to be introduced through temporal transfer? The primary objective of this study is to introduce a new quality assessment method for temporal transfer and evaluating the effects of temporal transfer errors on downstream analyses. The diagnostics we illustrate are built on the basic framework of combining temporal cross-validation (Roberts et al., 2017) with an adapted set of multiscale assessments from Riemann et al. (2010) to test if temporal model transfer leads to model misspecification and potentially biased model-based small area estimators. This framework has the advantage of emulating the spatial aggregations often performed by map users and highlighting areas where such small area estimates may be biased because of temporal transfer.

We applied this assessment method to a case study of mapping percent tree canopy cover in Michigan using Landsat time series with three years of photo-interpreted reference data in a six-year time span. This case study represents a relatively conservative scenario for testing temporal transfer in that it covers a short time period that is expected to have little change in canopy cover. We hypothesize that when compared to a base model (i.e., training/testing acquired in the same time step), a temporally transferred model would yield higher map error at a pixel/plot level and would result in model-based small area estimates that deviate more from unbiased design-based estimators obtained from a probabilistic reference sample. Identifying an increase in map error and biased population-level estimators with this relatively conservative scenario could indicate that there is even greater potential for model misspecification when applying temporal transfer over a longer time span with more variable conditions. Also, time series algorithms for remote sensing data, such as LandTrendr and BFAST (Verbesselt et al., 2010), are becoming more widely adopted and may lead some to believe that they address all issues with temporal transferability. We tested this assumption and evaluated the degree of improvement in temporal transferability by comparing the use of the annual Landsat medoid composite values to the fitted values from LandTrendr segmentation in modeling of canopy cover. Our evaluations illustrate the need to test the temporal transfer of remote sensing models when used in large area monitoring.

2. Methods

The case study described below demonstrates how to test the effectiveness of temporal transfer for modeling and mapping canopy cover across a time series to support efforts for generation of both hindcasted maps and low latency updates in large area monitoring applications with limited reference data. We developed models of canopy cover for the state of Michigan, USA, from annual Landsat composites and a series of topographic variables. Percent tree canopy cover, hereafter "canopy cover", is defined here as the percent of vertical projection of live tree crowns over a horizontal ground area (Jennings et al., 1999). Reference data on canopy cover may be difficult and expensive to obtain, even with photo-interpretation of very high spatial resolution imagery, and it would not be possible for time periods lacking such imagery. In section 2.1, we present the background and basic approach for modeling and mapping canopy cover in Michigan. In section 2.2, we explain how this modeling approach is extended for evaluating temporal transferability using temporal cross-validation with blocking on the time period of interest as described by Roberts et al. (2017). In section 2.3, we describe a method for evaluating the effects of temporal transfer on spatial aggregation at multiple scales based on a modification of the multiscale evaluation methods presented by Riemann et al. (2010). How we evaluate the effects of temporal model transfer on temporal trends in average canopy cover is briefly explained in section 2.4.

2.1. Case study: Michigan canopy cover

2.1.1. Study area

For this study, canopy cover was mapped in the state of Michigan in the upper Midwest of the United States (Fig. 1). Michigan is mostly in Köppen class Dfb (Kottek et al., 2006), which has a humid continental climate with warm summers, freezing cold winters, and consistent levels of precipitation throughout the year. Most of the forest land lies within the upper peninsula and northern half of the lower peninsula, with the southern lower peninsula being comprised largely of croplands and built-up areas. As of 2016, 62% of forest land was privately owned (USDA Forest Service, 2023). Stands consisting of maple, beech, and birch trees were the most common forest type group in 2016, making up 30% of all forest land (USDA Forest Service, 2023). Forest area and biomass have both increased since 1980, and live aboveground biomass increased 6.6% between 2010 and 2016 forest inventories covering the time period of our reference data (USDA Forest Service, 2023). Michigan forests typically have rotation ages of 50–100 years (Michigan Department of Natural Resources, 2008) and modern stand replacing fire return intervals are in the hundreds of years (Cleland et al., 2004). The mean area of fast disturbance per year as a percent of the state was 0.49% for our study time frame of 2010–2016 and 0.42% for 1985 to 2019 according to the Landscape Change Monitoring System dataset (USDA Forest Service, 2021).

2.1.2. Data

Very high spatial resolution imagery was photo-interpreted to generate the reference dataset of canopy cover to be used in modeling (Fig. 2). A sample of USDA Forest Service Forest Inventory and Analysis (FIA) plots in Michigan ($n = 1273$) were photo-interpreted using 1 *m* resolution 2010 aerial imagery from the USDA National Agriculture

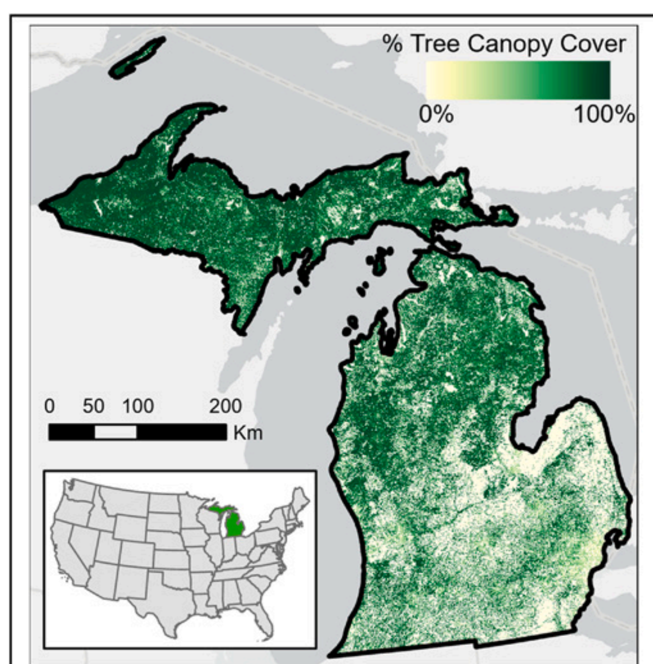

Fig. 1. NLCD 2016 Tree Canopy Cover for the state of Michigan, USA.

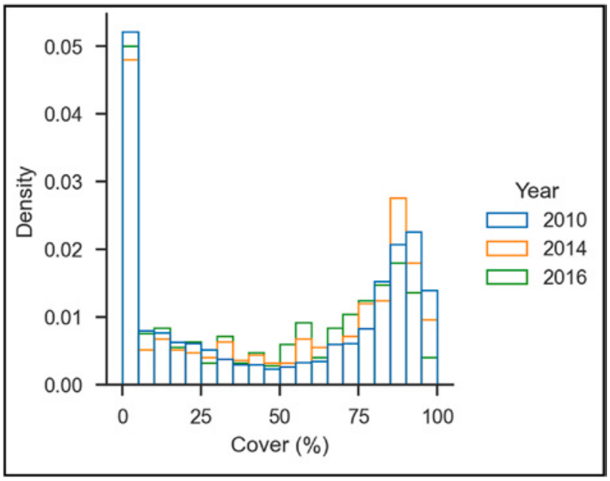

Fig. 2. Density of plots in percent cover bins for each inventory year.

Imagery Program (NAIP) to aid in creation of the 2011 National Land Cover Database (NLCD) Tree Canopy Cover product (Coulston et al., 2012). This initial reference dataset was supplemented with photo-interpretation of an additional simple random subsample of 500 FIA plots using 1 *m* NAIP imagery acquired in 2014 and another random subsample of 500 FIA plots using 0.6 *m* NAIP imagery acquired in 2016. In total 2273 plots were photo-interpreted. We used FIA plot locations to leverage pre-existing photo-interpretations and have a probabilistic sample, but no *in-situ* FIA data were used in this study. Photo-interpretation was done with a dot-grid of 109 dots that were arranged in a single 43.9 *m* radius circular area encompassing the entire FIA plot. The percent of dots labeled as tree canopy was used as canopy cover for the plot.

A time series of annual Landsat composite images was prepared in Google Earth Engine. Landsat 4, 5, 7, and 8 Collection I Tier I Level 2 surface reflectance images were compiled for the summer of each year (July 1st to September 10th) from 1984 to 2020 and filtered for clouds, cloud shadows, snow, and water using the provided Pixel Quality Assurance band and the Temporal Dark Outlier Mask (TDOM) algorithm (Housman et al., 2018). We temporally composited the images in each year by calculating the medoid across all bands (Flood, 2013). A set of spectral indices were calculated for each temporal composite including the Normalized Difference Vegetation Index (Rouse et al., 1974), Normalized Burn Ratio (Key and Benson, 2006), the Tasseled-Cap Brightness, Greenness, Wetness (Crist, 1985; Huang et al., 2002), and Tasseled-Cap Angle (Powell et al., 2010). The LandTrendr algorithm (Kennedy et al., 2010, 2018) was used to temporally segment the annual composite time series for each pixel in Google Earth Engine, and the fitted values were taken from the segmentation for canopy cover modeling and mapping. The LandTrendr temporal segmentation also filled missing pixels in the Landsat time series, which were primarily due to masking of water, and this led to slightly different sample sizes in each year between Landsat and LandTrendr modeling.

A series of topographic variables calculated from the U.S. Geological Survey's 10 *m* National Elevation Dataset resampled to 30 *m* using bicubic interpolation were also included in canopy cover modeling. These included the elevation in meters, aspect in degrees, the percent slope, the cosine of aspect, sine of aspect, and the cosine and sine of aspect multiplied by the percent slope (Stage, 1976). More complex topographic variables included the solar-radiation aspect index (Roberts and Cooper, 1989), the topographic position index calculated at scales of 90 and 990 m (Weiss, 2001), the continuous heat load index (Theobald et al., 2015), and the topographic diversity index (Theobald et al., 2015).

All remote sensing data was resampled such that each 30 *m* pixel represented the mean of the 3 × 3 window (90 *m*) surrounding it to approximately match the scale of the 43.9 *m* radius FIA plots for model development and mapping. Additional details on data preparation are provided in the supplemental methods sections 1.1 and 1.2.

2.1.3. *Canopy cover modeling*

Annual maps of canopy cover were created with a two-step process of regression modeling of percent tree canopy cover and then reclassifying non-forest as 0% canopy cover with an annual forest/non-forest mask. First, random forest regression models of canopy cover were generated from the remote sensing variables, which were taken from the same year used for photo-interpretation of canopy cover for each FIA plot. Next, a classification model of forest and non-forest land cover was developed to generate a set of annual forest masks to be applied to the canopy cover predictions to reduce overestimation of canopy cover in croplands and for use in subsequent analyses of forest areas. We applied a threshold to photo-interpretation values of canopy cover for the same set of FIA plots to designate them as forest (i.e., ≥10% canopy cover) and non-forest (i.e., <10% canopy cover). The forest masks were then generated from a random forest classification model of forest and non-forest using this reference data as the response variable and the same set of remote sensing predictor variables used in the canopy cover model. Details on the forest mask creation can be found in supplemental section 1.3. Both the canopy cover models and the forest mask model used the random forest algorithm (Breiman, 2001) with 500 trees grown to full depth and with the square root of the total number of variables being randomly selected for consideration in splitting at each node.

All maps were then evaluated for accuracy by comparison to the reference sample at the plot level by their root-mean-square-error (RMSE), RMSE as a percent of the mean observed value (RMSE%), the proportion of variance explained by the model (R^2), and the mean of residuals which we refer to as "bias". These values are reported after application of the forest masks. This approach to modeling and evaluating maps at the plot level was applied to the full set of training data as a basis of comparison to the temporal transferability and temporal cross-validation approach described below in section 2.2.

2.2. *Temporal cross-validation*

2.2.1. *Temporal blocking scheme*

Testing of model temporal transferability was performed using a temporal cross-validation scheme. We perform temporal cross-validation with data splitting (i.e., blocking) over years as part of model evaluation because failing to account for dependence structures of the data when performing cross-validation leads to overly optimistic error assessments, particularly when the intended application of the model is extrapolation, such as to new time periods (Roberts et al., 2017). We use a leave-one-block-out approach where a cross-validation fold matches the block or time unit of interest (e.g., one year), as recommended by Roberts et al. (2017) for maximizing the use of the training data. Having the block size match the cross-validation fold also allows for the examination of error in particular time periods as well as the variance of error across years. This scheme also aligns with our objective of extrapolating the model to any year in the time series of remote sensing data without reference data rather than applying solely to the beginning of the time series (i.e., hindcasting) or the end of the time series (i.e., forecasting).

2.2.2. *Evaluating model specification in temporal transfer*

A poorly specified model may not adequately represent the true data generating process, which can lead to poor predictions and false inferences. Misspecification of parametric models, like an ordinary least squares linear model, may be caused by omission of important variables, inclusion of irrelevant variables, incorrect model form, and a failure to meet model assumptions, such as having homoscedastic residuals (Ramsey, 1969). While nonparametric models, like random forest, benefit from having fewer assumptions and having the form of the model learned from the data, there are still factors that can lead to poor specification. These factors may include omission of important variables, inclusion of irrelevant variables, inadequate training data, and poor selection of hyperparameters. The primary means for checking model specification is applicable to both parametric and nonparametric models, which is to compare predictions to observations with regression diagnostics, such as scatterplots, residual plots, and fit statistics such as RMSE, R^2, and bias. If a model is correctly specified, a regression of its predictions on the observations will ideally have an intercept of 0 and slope of 1 (McRoberts, 2011).

Temporal transferability places the additional constraint of needing correct model specification to hold in new time periods that may lack reference data. This means it is also necessary to specify a model that can accommodate the expected range of variation in both the predictor and response variables over time and ensuring that their relationship as captured by the model remains consistent. The additional factors that can lead to poor specification for temporal transferability are a lack of consistency through time in the reference data, the predictor variables, or their relationship with each other. To test whether a model is likely to remain valid across time, the same regression diagnostics may be applied to a withheld time period in temporal cross-validation using any available multi-temporal data.

We employed this approach by developing separate models trained and tested on different temporal subsets of the reference data. Models trained and tested on the full set of reference years together are referred to as the All-Time Validation (ATV) models, which represent the common validation approach of selecting test sets at random without regard to temporal dependence structures in the data. We also evaluated models by training and testing within a single year, called Same-Time Validation (STV), which we expected to have the most accurate predictions. For the STV and ATV models, accuracy statistics were calculated with the out-of-bag predictions from the canopy cover random forest model (i.e., the mean prediction across trees when a data point was excluded from training). Out-of-bag error is an unbiased indicator of generalization error that has a lower computational cost than cross-validation (Breiman, 1996). Models were also evaluated for temporal transferability using a leave-one-year-out temporal cross-validation

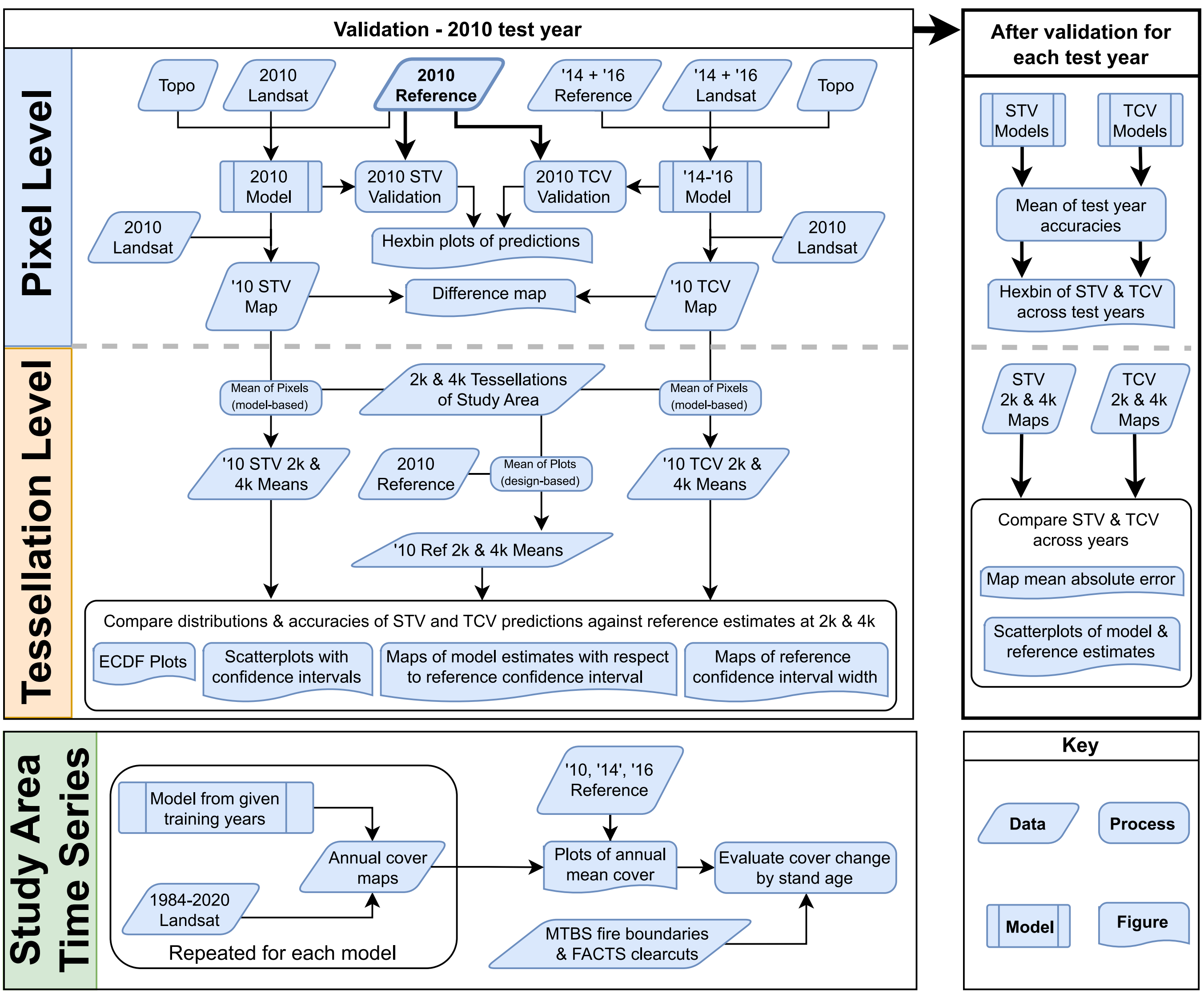

Fig. 3. Flow chart of analysis for comparisons between same-time-validation (STV) models and temporal cross-validation (TCV) models at the pixel/plot level, tessellation level, and entire study area. Validation and comparisons at the pixel level and 2 k and 4 k tessellation levels are shown for 2010 being the test year, but this process is repeated for each year in a leave-one-year-out temporal cross-validation scheme. Maps and statistics are then aggregated across test years for a multi-year and multiscale evaluation.

scheme (TCV) as described in the previous section (Fig. 3). When summarizing the STV and TCV accuracy statistics for multiple years, we calculated the mean of the statistics across years weighted by the sample size in each year. These temporal transferability assessments were conducted at the scale of individual plots and at two scales of spatial aggregation as part of a multiscale accuracy assessment described in section 2.3.

The TCV summary statistics were compared to the STV statistics to evaluate the potential for error introduced by temporal transfer. Our comparisons illustrate a conservative real-world scenario for evaluating transferability because models were applied and tested within a short time frame of six years in an area where little change is expected to have occurred (see section 2.1.1). Thus, witnessing substantial changes in the maps for this time period and seeing differences in model error between the STV and TCV results would indicate significant errors from model transfer. Errors from temporal transfer would likely be worse for applications using data with greater time lags that may have greater changes in the remote sensing or reference data sources over time.

2.3. Multiscale evaluation

In addition to standard evaluations of model accuracy at the plot and pixel level, we perform a multiscale evaluation following the approach of Riemann et al. (2010) to validate use of the maps for model-based small area estimation. One of the most common applications of remote sensing-based maps is to conduct aggregations over areas of interest, such as estimating the mean or total of a population (i.e., small area estimation). It follows that an evaluation of remote sensing models should then also be performed for this intended application. Multiscale evaluations of continuous variables like in Riemann et al. (2010) can reveal spatial patterns in the correspondence of estimates derived from mapped predictions and from the reference sample at multiple scales of aggregation. We perform a modified version of this multiscale evaluation that uses small area estimation methods (Rao and Molina, 2015) to compare the map-based and reference-based mean canopy cover estimated within each cell of the tessellated study area. The comparisons are made through analyses of their empirical cumulative distribution functions, scatter plots of estimates and associated summary statistics, and maps illustrating the estimates and their confidence intervals as described in detail below.

2.3.1. Tessellation scheme

We tessellated the study area by applying a variant of the K-means clustering approach to a grid of points covering the study area (Krivoruchko and Gribov, 2020), which was implemented with the Generate Subset Polygons tool in the ArcGIS Pro software (ESRI, 2022). This process created compact polygon cells that fell completely within the study area and had an approximately equal area and number of reference plots. Scale “2 k” had 76 cells with a mean area of 1990 km^2 (standard deviation = 22 km^2) and a mean of 33.2 plots in each cell (standard deviation = 4.7 plots). Scale “4 k” had 38 cells with a mean area of 3981 km^2 (standard deviation = 43 km^2) and a mean of 66.4 plots in each cell (standard deviation = 7.6 plots).

2.3.2. Model-based versus design-based estimators

Small area estimation methods (Rao and Molina, 2015) were used to estimate the mean canopy cover for each tessellation cell and the entire study area. We estimated mean canopy cover from the sample of reference plots using design-based estimators and from the model-predicted canopy cover map using model-based estimators. Design-based inference is a classic statistical framework of survey sampling that is commonly used in national forest inventories, such as FIA’s estimates of forest area or timber volume (Bechtold and Patterson, 2005). Design-based estimators have the advantage of being unbiased because of randomization in the sampling design (i.e., design-unbiased). Model-based inference is an alternative inferential framework where a model is used with auxiliary information (e.g., Landsat data) to predict the value of each population unit (i.e., pixel), and population units are then aggregated to form model-based predictions of population parameters like the mean of an area (Cochran, 1977; Särndal et al., 1992). This approach is functionally equivalent to the zonal statistics or spatial aggregations performed by map users. Model-based inference is used for small area estimation (Rao and Molina, 2015) when the sample size is too small to obtain reasonably small confidence intervals with a direct design-based estimator. While the model used for the unit-level predictions in small area estimation are often linear mixed models like the Empirical Best Linear Unbiased Predictor (EBLUP) (Rao and Molina, 2015), it is also possible to use non-parametric models such as nearest neighbors or decision tree-based models (McRoberts et al., 2022). Comparisons of design-based estimators with model-based or other types of estimators are common in the remote sensing literature, particularly for forest applications (Gregoire, 1998; McRoberts, 2011; McRoberts et al., 2007; Menlove and Healey, 2020; Ståhl et al., 2016).

The direct design-based estimator for the mean canopy cover for each tessellation cell ($\widehat{\mu}_d$) was calculated as the Horvitz-Thompson estimate using the sample of n reference plots intersecting the cell, with each having canopy cover values of y_i:

$$\widehat{\mu}_d = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Instead of using a traditional variance estimator, we used 1000 bootstrap replicates of the mean to obtain bias corrected and accelerated confidence intervals (Efron, 1987) because they are likely to be more reliable when used with small samples from populations that have a non-normal distribution (Carpenter and Bithell, 2000), which was likely the case for many cells at the 2 k scale.

For model-based inference, the random forest model described in section 2.1.3 served as a unit-level model linking the remote sensing predictor variables (x_i) to predictions of canopy cover for each population unit ($\widetilde{y}_i$):

$$\widetilde{y}_i = f(x_i)$$

The model-based synthetic estimator for the mean canopy cover of a tessellation cell or the study area is then:

$$\widehat{\mu}_m = \frac{1}{N}\sum_{i=1}^{N} \widetilde{y}_i$$

where N is the number of population units (i.e., pixels) in the area of interest. It is expected that if the correct model is specified then the predicted mean ($\widehat{\mu}_m$) will match that of the true mean (μ_m), and the estimator is then model-unbiased. However, Hansen et al. (1983) demonstrated how even seemingly reasonable models can lead to substantial bias and misleading uncertainty estimates. We evaluated model specification at the pixel/plot level by comparing predictions against observations as described in sections 2.1.3 and 2.2.2. At the population level, model-based estimates are expected to be very similar but not necessarily the same as the design-based estimates because they rely on different inferential frameworks. Others have used the rule of thumb that model-based estimates should lie within two standard errors of the design-based estimates (McRoberts, 2010; McRoberts et al., 2007). According to McRoberts et al. (2007), the estimates could be very different either because 1) the model-based estimator is biased, 2) the sample data used in the design-based estimator is not representative of the population because of issues in the study design or implementation, or 3) an unlikely population has been realized from the superpopulation. Through the multiscale diagnostics in the following section, we illustrate the difference between the model-based predictions and the design-based estimates with the expectation that the greater the separation between these values, the more likely it is that the model is not correctly specified and the model-based estimator is biased. We do not calculate the precision of model-based estimators because it is not necessary for evaluating potential bias, and meaningful comparisons of precision with that of design-based estimators is hindered by their use of different inferential frameworks (McRoberts et al., 2007).

2.3.3. Multiscale diagnostics

We adapted a series of diagnostics from Riemann et al. (2010) for evaluating model specification across scales with a focus on identifying the potential for *bias* resulting from temporal transfer. To visualize the agreement between estimates of mean canopy cover at each scale, we show scatterplots of the model-based estimates derived from remote sensing (STV and TCV predictions) against the design-based estimates derived from the reference sample with their 90% confidence intervals along with the RMSE, RMSE%, R^2, and bias statistics. We also compared the empirical cumulative distribution functions (ECDF) of mean canopy cover estimates from each of the methods and calculated the two-sample Kolmogorov-Smirnov (KS) statistic to give the maximum distance between the ECDF for each of the model-based estimates and the design-based estimates.

Spatial patterns in predicted canopy cover and their potential error were visualized through a choropleth map of the tessellation where the color of each cell portrays the model-based predicted canopy cover value, and the outline of each cell portrays the location of that prediction with regard to the design-based 90% and 95% confidence intervals for mean canopy cover from the reference sample. These maps combine two different assessment methods from Riemann et al. (2010) into a single visualization so both the canopy cover prediction and its potential error can be easily assessed together. Bias corrected and accelerated confidence intervals (Efron, 1987) were used to illustrate uncertainty in these maps and for the scatterplots since the 2 k scale had a relatively small sample size for each cell in individual years and the underlying distribution of canopy cover is likely not normal. The width of the confidence intervals from the reference sample is shown in an additional choropleth map to illustrate the relative uncertainty of design-based mean canopy cover estimate. For cells with wide confidence intervals, it may be prudent to examine the absolute difference between the model-based and design-based estimate or consider increasing the cell size to encompass more sample units. The above maps were created for an individual year. To indicate the general potential for error across a longer time span when using model-based small area estimates derived from temporal transfer, we created another choropleth map that compares the mean absolute difference between the model-based predictions and the design-based estimates of mean canopy cover at the 2 k scale from the STV, TCV, and ATV models. The individual year differences between these estimation methods were averaged across all three years with reference data.

For each of the above multiscale evaluation methods, we show results for the 2010 estimates since that year has the largest sample size and yields the narrowest confidence intervals for the reference sample estimates. Results for other years and all years together can be found in the supplementary materials.

2.4. Canopy cover trends

Canopy cover models developed from each of the temporal subsets were used to map canopy cover for the study area from 1984 to 2020. We illustrate the change in canopy cover over the time series for the entirety of the forested area of Michigan through line plots of mean canopy cover. In these line plots we compare the estimates derived from three sources – 1) model-based estimates from the models trained with different temporal subsets of reference data; 2) design-based estimates derived from the photo-interpreted reference plots in 2010, 2014 and 2016; 3) NLCD Tree Canopy Cover products in 2011 and 2016. For the design-based estimates we calculated the 95% confidence intervals of mean canopy cover and the percent area following the FIA population estimation procedures in Bechtold and Patterson (2005). These line plots are used to illustrate the extent to which models based on training data from different time periods can lead to different temporal trends or biases. We also compare these temporal trends between estimates derived from LandTrendr-based models and Landsat-based models to show whether LandTrendr does in fact lead to greater temporal stability when aggregating across large areas. The plausibility of trends in the LandTrendr-based canopy cover maps was evaluated by comparing the change in canopy cover for mature stands and young stands. Stands were considered mature if they were 25 years or older from the last clearing disturbance as determined by fires from the Monitoring Trends in Burn Severity (MTBS) dataset or clearcuts from the Forest Service Activity Tracking System (FACTS) database.

3. Results

3.1. Pixel-level comparisons

When combining model predictions across years, models that used the Landsat images processed with the LandTrendr temporal segmentation algorithm yielded higher accuracy than using the base Landsat composites for both STV and TCV (Fig. 4), which had lower RMSEs by 2.1% and 3% canopy cover, respectively. Because of higher performance of the LandTrendr-based models, most of the subsequent analysis, such as the multiscale evaluations, were done solely with the LandTrendr-based canopy cover maps. The LandTrendr TCV (Fig. 4) had slightly worse validation statistics (RMSE = 13.9%, bias = −3.7, $R^2 = 0.86$) than the LandTrendr ATV, which had an RMSE of 13.6%, bias of −3.0, and R^2 of 0.87. The LandTrendr STV had a slightly higher mean RMSE (14.2%) than the TCV (Fig. 4), but its absolute bias (−3.5) was lower. The lower accuracy of LandTrendr STV may stem from having relatively few training points in 2014 and 2016 as compared to the two years of training data in each TCV model. Examining model accuracy for individual years (Table 1) shows that LandTrendr TCV had more variation in bias values of −5.46, −2.62, and −0.54 in 2010, 2014, and 2016, respectively, when compared to the STV bias values of −3.39, −3.59, and −3.69 for the same years. All models showed a compression towards the mean observed value, with high values tending to be slightly underestimated and low values tending to be overestimated (Fig. 4).

Maps of differences between the LandTrendr STV and TCV predictions for each year reveal that for forested areas the TCV model tended to produce much lower estimates of canopy cover than STV in 2010 but higher estimates in 2014 and 2016 across many areas (Fig. 5). Given that the LandTrendr STV bias was consistently low in each year (−3.39 to −3.69), these spatial results illustrate how maps generated from models using training data from other time periods (i.e., the TCV models) may exhibit more variable bias in some times and areas than what is reflected by the overall TCV bias across years of −3.7% canopy cover (Fig. 4) or even the individual year bias values (Table 1).

3.2. Multiscale evaluation

We summarized the canopy cover predictions from the STV- and TCV-based maps at different scales of aggregation (2 k, 4 k, and study wide) and compared the predictions to each other and to the reference data. The STV- and TCV-based maps of 2010 tessellated mean canopy cover show the same general patterns at both scales (Fig. 6), such that the pixel level differences seen in Fig. 5a are not readily distinguishable. Most of the model-based predictions were within the 90% confidence interval of the design-based estimate of mean canopy cover, and the areas with predictions exceeding the 90% confidence interval were different between scales and not concentrated in any region. For the 2010 reference data tessellated at the 2 k scale (Fig. 6 f), the mean width of the confidence intervals was 27% canopy cover (mean of 70% when scaled as a percent of the cell's canopy cover value), while at 4 k (Fig. 6 c) the average confidence interval width was 20% (scaled mean of 49%). In the southern and eastern portions of the state, the mean canopy cover was low and variance was high due to tree patches of varying densities

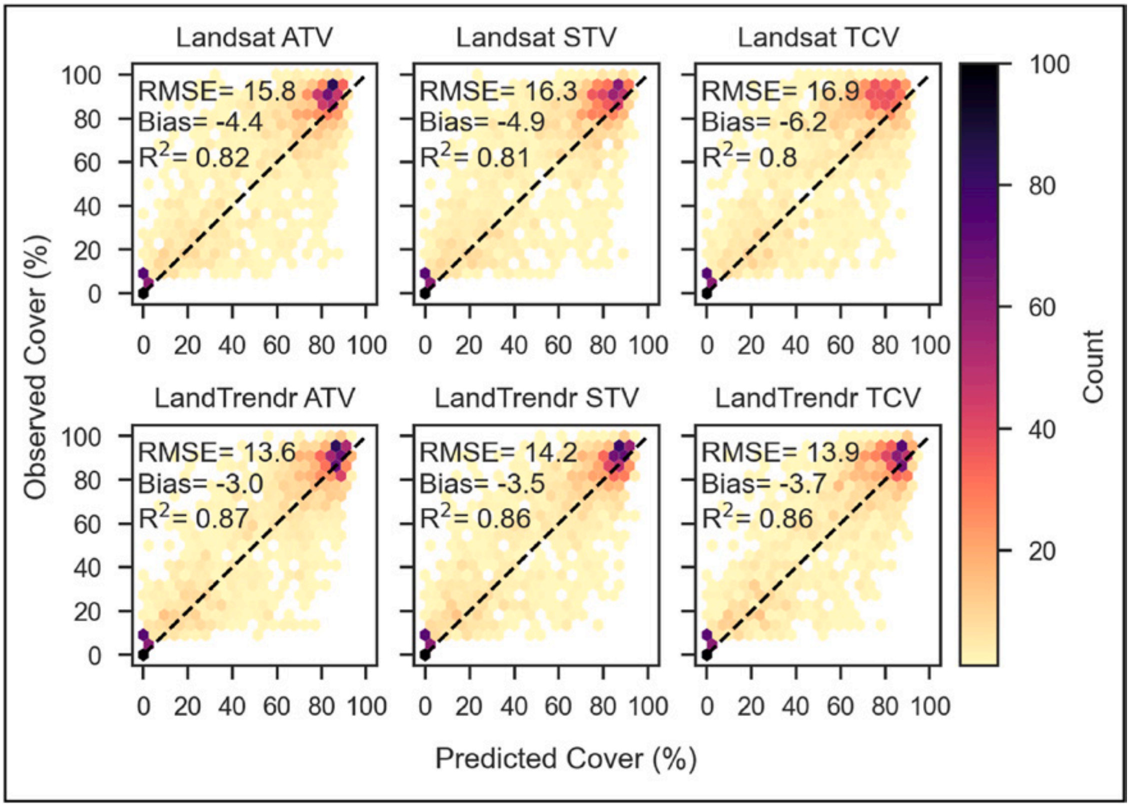

Fig. 4. Observed and predicted cover for models using either the original Landsat predictors or those fit with LandTrendr. Results were combined from individual year models which were either trained and tested on the same year (STV) or tested on the year excluded from model training in a temporal cross-validation scheme (TCV). R^2, bias, and RMSE statistics are the weighted average of the individual year values for the STV and TCV models.

Table 1
Individual year results of models using Landsat composites or LandTrendr fitted predictors evaluated with predictions from the same-time validation (i.e., trained and tested in same year) or temporal cross-validation (i.e. tested on withheld year and trained using other years).

Predictor	Model	Year	N test	R^2	RMSE	Bias
Landsat	STV	2010	1225	0.82	16.47	−4.44
Landsat	STV	2014	495	0.81	16.26	−5.41
Landsat	STV	2016	498	0.80	15.91	−5.56
Landsat	TCV	2010	1225	0.80	17.51	−8.30
Landsat	TCV	2014	495	0.78	17.39	−5.84
Landsat	TCV	2016	498	0.82	15.01	−1.53
LandTrendr	STV	2010	1234	0.87	14.06	−3.39
LandTrendr	STV	2014	497	0.84	14.68	−3.59
LandTrendr	STV	2016	499	0.84	14.25	−3.69
LandTrendr	TCV	2010	1234	0.87	14.21	−5.46
LandTrendr	TCV	2014	497	0.86	14.02	−2.62
LandTrendr	TCV	2016	499	0.87	12.90	−0.54

dispersed amongst the abundant cropland; in these areas, confidence intervals were up to 174% and 100% of the cell's mean canopy cover at 2 k and 4 k, respectively.

The model-based prediction along with the design-based estimate and confidence interval can also be seen non-spatially in Fig. 7 scatterplots for comparing the STV and TCV results at different canopy cover ranges. Fig. 7 also shows that when aggregating to mean canopy cover in the tessellations, the absolute average difference between the design-based estimator (i.e., observed canopy cover) and both of the 2010 STV and TCV model-based estimators at both scales is slightly lower (i.e., less negative bias) than in the pixel level comparisons to the reference data (Table 1) and RMSE decreases. The difference between the TCV and STV mapped predictions is better captured by a direct comparison at each scale (Fig. S3), which shows that in 2010 TCV canopy cover in forest areas was lower by 3.84% at the pixel level, 3.14% at 2 k, and 3.15% at 4 k on average. The RMSE of 2010 TCV predictions is also higher (2 k RMSE = 8.83%, 4 k RMSE = 6.97%) than predictions from STV (2 k RMSE = 7.78%, 4 k RMSE = 5.56%) when aggregating to both scales despite their nearly identical accuracy at a pixel level (TCV RMSE = 14.06%, STV RMSE = 14.21%; Table 1). Differences between the different estimation methods (STV, TCV and design-based) for specific canopy cover ranges can be more clearly seen in ECDF plots (Fig. S4), which show that TCV models resulted in fewer cells with over 50% canopy cover than either the design-based estimates or the STV predictions at either scale.

Figs. 6 and 7 highlight the 2010 results, which had the most reference points to compare against, but the multiscale comparisons can also be aggregated across time in various ways to examine overall temporal transferability from multiple training and testing sets. For example, Fig. 8 compares the mean absolute difference between model-based predictions and design-based estimates across years for the ATV, STV, and TCV models. The overall difference between models across the cells was fairly small, but the TCV predictions have higher absolute difference between the estimation methods (mean difference = 9.43% canopy cover) in more cells of the heavily forested northern peninsula than the STV (mean difference = 7.88% canopy cover) or ATV (mean difference = 8.69% canopy cover) estimates. Scatterplots of observations and predictions at different scales can also be extended across years (Fig. S5). Other diagnostics such as the ECDF plots (e.g., Fig. S4) may be less suitable for application across time periods because they would encompass changes between years.

3.3. *Temporal trends*

We examined the effect of temporal transferability on long term trend estimates of mean canopy cover for Michigan from 1984 to 2020 b

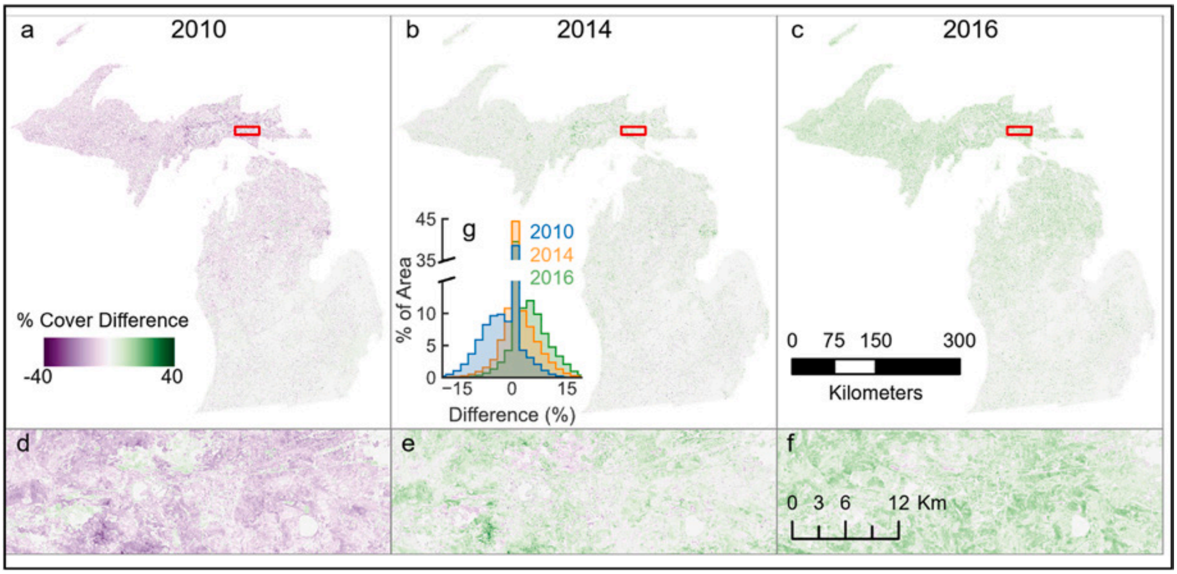

Fig. 5. Difference in canopy cover maps created by subtracting the same-time validation maps (STV) from the temporal cross-validation maps (TCV) for 2010 (a), 2014 (b), 2016 (c). An area of the upper peninsula with large differences is shown below each map (d, e, f). Subfigure g shows the histogram of each map with normalized counts.

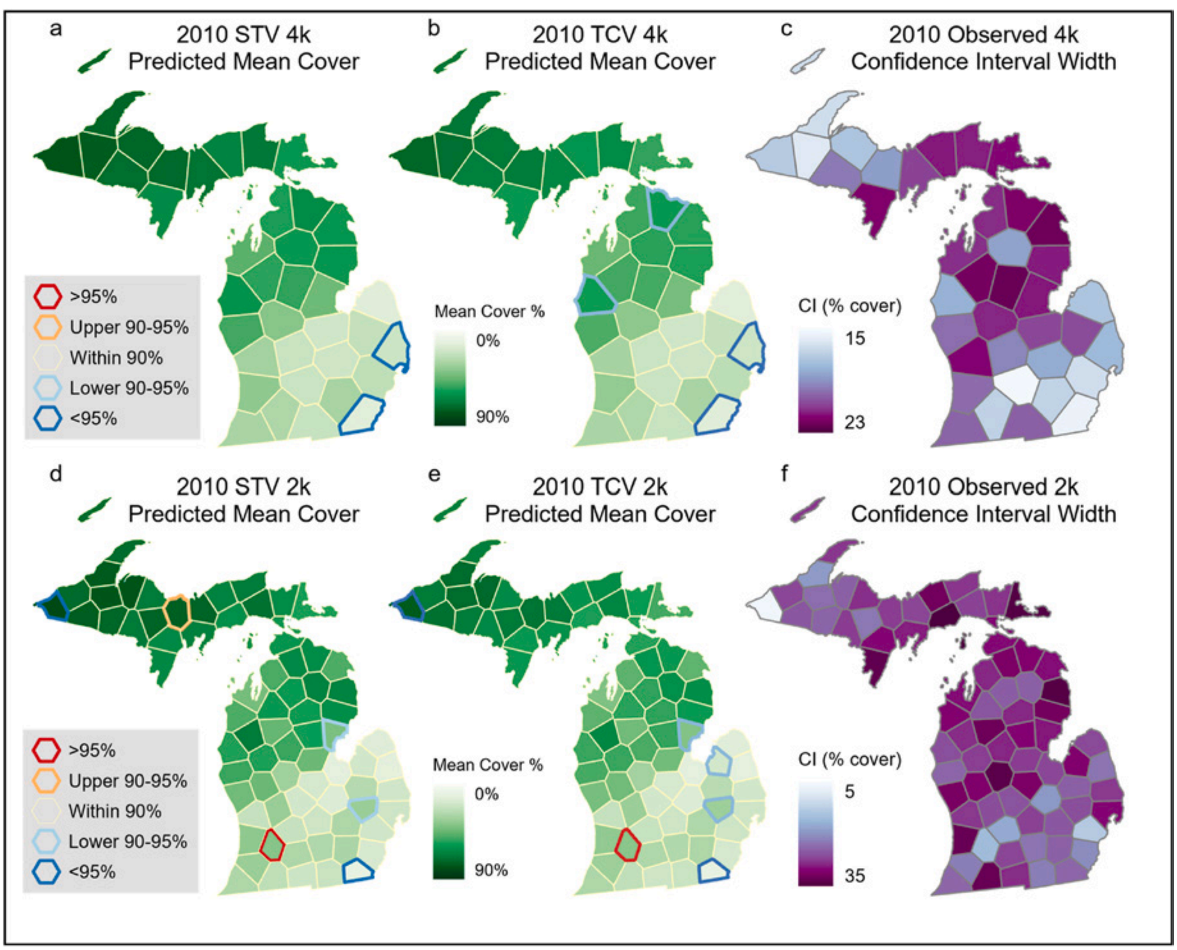

Fig. 6. Mean canopy cover in 2010 from same-time validation (STV) and temporal cross-validation (TCV) predictions at the 2 k (d,e) and 4 k (a,b) tessellation scales. The location of the model-based mean canopy cover with respect to the confidence interval bounds derived from the reference design-based estimates are shown in a, b,d and e as colored cell outlines. The width of the 90% confidence intervals is shown in c and f.

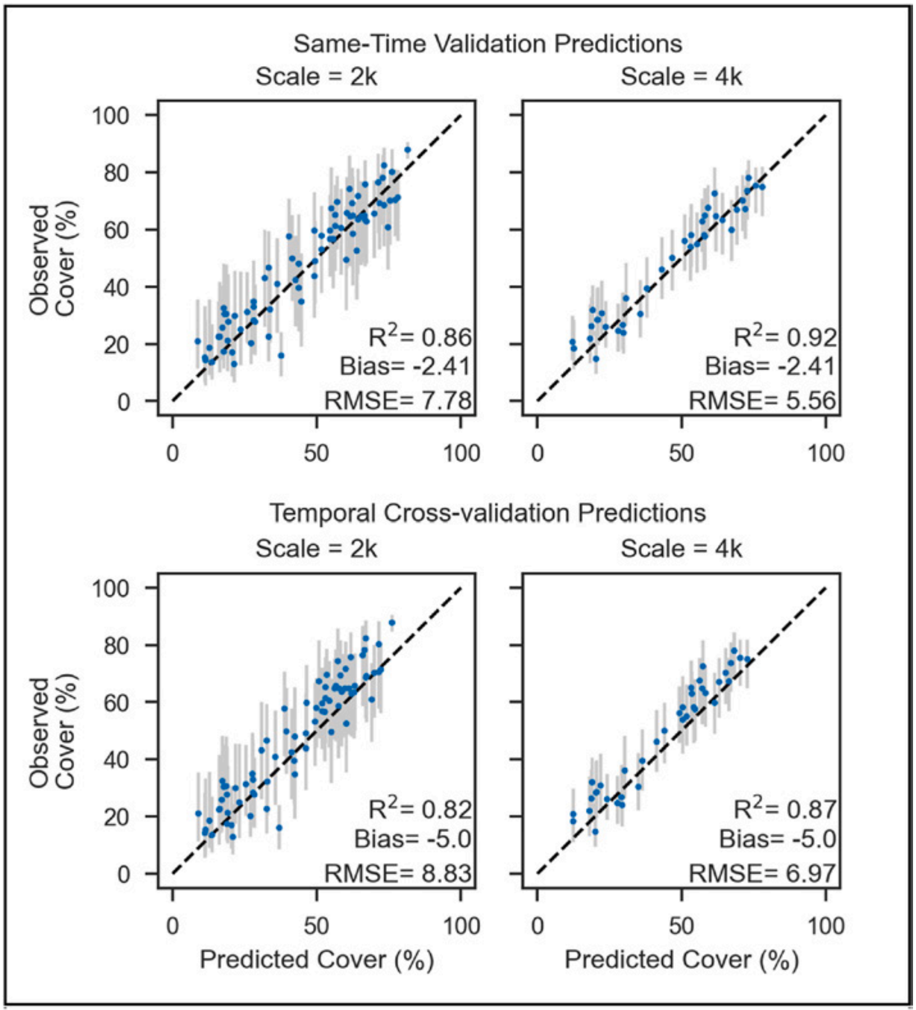

Fig. 7. Comparison of predicted and observed mean canopy cover in 2010 using predictions from the same-time validation models (STV) and the temporal cross-validation models (TCV) for 2 k and 4 k scale tessellations. The 90% confidence intervals for the design-based estimates are shown as grey lines.

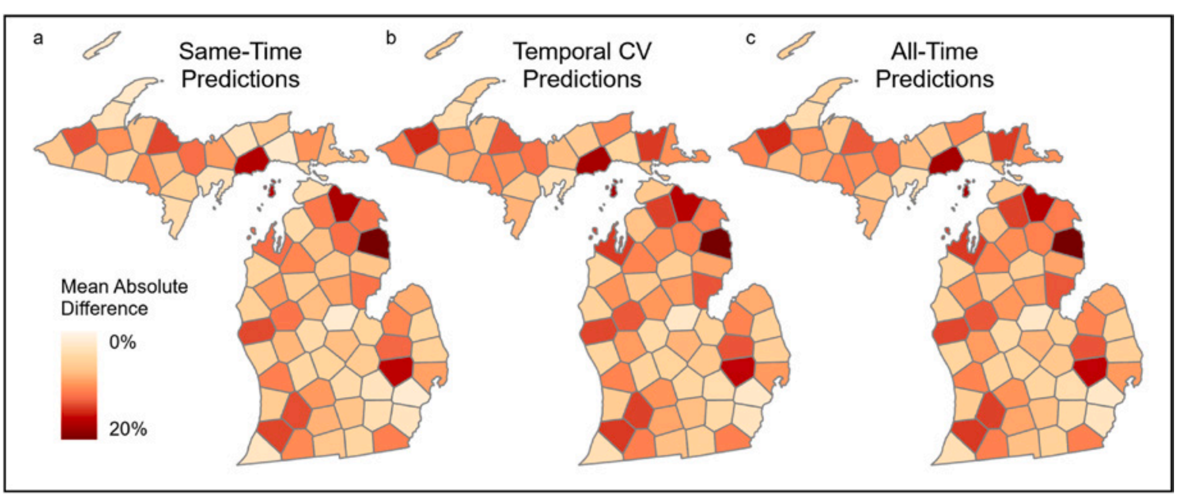

Fig. 8. The absolute difference between the model-based predictions and design-based estimates of mean canopy cover at 2 k was averaged across the three years (2010, 2014, 2016) for the same-time models (a), temporal-cross validation models (b) and the model using training data from all years together (c).

y comparing the results from training with different temporal subsets of data. Note that estimates of canopy cover extrapolated well beyond the temporal extent of the training data should be regarded with caution, and that we show the predicted trends mainly to illustrate the effects of temporal transfer and LandTrendr on the trend lines. When using LandTrendr data, the models built with different temporal subsets showed a nearly identical trend in mean canopy cover (Fig. 9a) with a fairly consistent difference of 5.3% canopy cover between the model

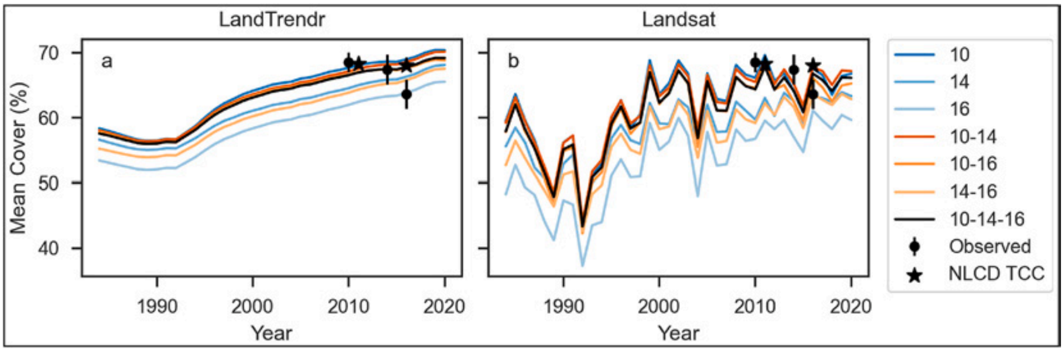

Fig. 9. The time series of model-based predictions of mean annual canopy cover of forested areas in Michigan derived from the maps that were produced using different temporal subsets during model training (i.e. 10-14-16 uses training data from 2010, 2014, and 2016) are shown as colored line plots. Results are shown separately for models that used LandTrendr data (a) and Landsat data (b). The design-based estimates of mean canopy cover from the photo-interpreted sample are shown as black dots with vertical lines representing its 90% confidence interval. Model-based predictions from NLCD TCC are shown as stars.

that yielded the highest estimate (2010) and the model that yielded the lowest estimate (2016). The mean canopy cover trend lines from the LandTrendr-based estimates were also fairly smooth with the model using all years of training data (10-14-16) showing a steady increase from a low of 56.0% in 1989 to high of 69.2% in 2019. By contrast, the trend lines from the Landsat-based models (i.e., without pixel-level LandTrendr temporal segmentation) varied wildly, with the all-years model (10-14-16) rapidly dropping from 62.0% in 1985 to a low of 43.3% in 1992 and then rising again to a near high of 67.0% in 1999. With the Landsat-based estimates, the different temporal subsets also showed a little more variation between their trend lines. For example, the maximum difference between the Landsat models with highest trend (2010) and the model with the lowest trend (2016) was 11.2% in 1984, which is 5.4% more than the maximum difference between the LandTrendr models.

Observed pixel-level trends in canopy cover were supported by external evidence on stand age by obtaining stand initiation date for fires and harvests in 1985 and later using the MTBS and FACTS databases, respectively. Areas with young stands (i.e., <25 years old) increased in canopy cover more than mature stands (i.e., ≥25 years old) by 8.9% on average when looking across 5-year increments from 1985 to 2020. The distinction between young and old stands held for each of the 5-year increments (Fig. S6), suggesting that trends in canopy cover were predicted consistently across the time series.

4. Discussion

The results of our above case study revealed that even when temporal cross-validation shows little bias or even higher precision at the pixel/plot level, temporal transfer may cause model-based small area estimators to deviate more from unbiased design-based estimators. The most likely explanation for the difference between the STV and TCV results is that the model-based estimators using temporally transferred models were indeed slightly biased for some cells. As Hansen et al. (1983) concluded, even seemingly reasonable models can lead to biased model-based estimators. Model development with a balanced or random sample, as in our study, is expected to reduce the potential for bias (Hansen et al., 1983; Nedyalkova and Tillé, 2012), but biased model-based estimators may still have arisen for a couple of reasons. For one, random forest models are sensitive to the distribution of the training data for the entire study area (Zhang and Lu, 2012), and this could have led to biased pixel-level predictions and model-based estimators in subpopulations (i.e., tessellation cells) having a different canopy cover distribution. Secondly, this principle may have also applied to the slight differences in the distribution of canopy cover between the time period used for model development and the time period of the maps used in model-based estimation (Fig. 2). Most applications of model-based estimators to remote sensing-based time series maps are likely to fall under these same circumstances, which suggests that instances of bias in small area estimators from map users performing zonal statistics are likely quite common.

Biases in small area estimators translate to consistent differences in predicted trends, as we found when training models with different temporal subsets (Fig. 9). Furthermore, atmospheric correction alone may not be sufficient to provide consistent estimates of trends when applying a model across a remote sensing time series, as evidenced by the high interannual variation in Fig. 9b. Time series algorithms like LandTrendr or the use of multi-year metrics are likely necessary for reliable trend estimates to account for sources of interannual variation in remote sensing data not related to the attribute of interest. Given the results of the multiscale assessment and these trend comparisons, there is a need to adopt standard approaches for testing temporal transfer of models at multiple scales. The potential for increased bias and uncertainty from temporal transfer should also be conveyed to map users.

According to Rykiel (1996), one of the most important aspects of ecosystem modeling is to establish a model's operational validity according to prespecified criteria. For this reason, the temporal transfer assessment framework and diagnostics we propose establish the model's operational validity by emulating two of the most common applications by map users — to perform spatial aggregations and examine temporal trends. The criteria for determining a model's suitability for temporal transfer depends on the sensitivity of the application to expected bias. For example, if our goal was to estimate the direction of the trend in canopy cover for Michigan from 1990 to present, our LandTrendr-based model would likely be deemed suitable for temporal transfer because of the low variability in trends of mean canopy cover between models built from different temporal subsets of training data (Fig. 9a). If the model is well specified, confidence in the trend would ultimately be determined by the uncertainty of the model-based estimators of mean canopy cover in both time periods. In this way, the diagnostics based on our framework aid in determining if a model is well specified enough for temporal transfer for a specific purpose. Below we discuss how this approach applies to three aspects of evaluating operational use of remote sensing models for large area monitoring: testing temporal transferability, small area estimation, and examining temporal trends.

For testing the implicit assumption that model error remains constant with temporal transfer, validation using withheld time periods of reference data is likely the only option. However, in the rare studies where temporal validation has been performed, it has been done through a mix of approaches that don't necessarily correspond to the objectives for applying the model. Often specific years are arbitrarily selected or combined for training and testing (Majumdar et al., 2020; Nelson et al., 2011; Zhong et al., 2014), or an exhaustive combination of different lengths and lags of time periods are used (Zhou et al., 2020).

This study demonstrates the benefits of temporal cross-validation with a blocking scheme that tests on a single withheld unit of time and trains with the remaining data (i.e., leave-one-time-out) to support both hindcasting and forecasting (Bergmeir and Benítez, 2012; Roberts et al., 2017). Temporal blocking, preferably with each block representing the time period of interest, maximizes use of the data for training and testing (Roberts et al., 2017). Temporal blocking also enables results to be investigated for a single time period (e.g., Fig. 6), summarized across multiple time periods (e.g., Fig. 8), and compared between time periods (e.g., Fig. 5 and Table 1). We compared the results of this temporal cross-validation strategy against the results of validating within the same time period as the training data (i.e., either 1 year at a time or random splitting across the full dataset) because it can provide at least some indication of the degree to which differences in model error are due to extrapolating to new time periods (Blasi et al., 2021; Smith et al., 2023; Vulova et al., 2021).

The framework we propose for testing temporal transfer of models incorporates small area estimation at multiple scales with model-based inference because it is likely the most common analytical application of raster maps, despite the potential for bias with model-based estimators. Model-assisted estimators that use both a probability sample and auxiliary information like maps are design-unbiased and may yield low uncertainty (McRoberts, 2010; Næsset et al., 2020; Ståhl et al., 2016), but raster maps are rarely distributed with the required reference data. Plus, model-assisted estimators would not be applicable for time periods completely lacking reference data. Model-based estimators have been compared to design-based estimators for checking model specification (McRoberts, 2011; McRoberts et al., 2007), but making this comparison for multiple areas and across scales using an approach similar to Riemann et al. (2010) enabled us to identify regions where the model may produce less accurate estimates (Fig. 6). For instance, we saw that the locations of model-based over/under-estimation changed slightly and became less common when going from 2 k to 4 k, despite narrower per-cell confidence intervals at 4 k. The lower frequency of over/under-estimation at 4 k could be due to localized biases being averaged out over larger areas as an example of the modifiable areal unit problem (Dark and Bram, 2007). Extending this multiscale analysis to compare temporal cross-validation with the STV estimates then revealed where the model may be applicable in new time periods. Thus, combining elements of scale and temporal transfer reveals more about model specification for small area estimates in new time periods than either analysis method would alone.

Generating unbiased small area estimators for new time periods is of even greater importance for maps created at high frequencies because another common application is examining trends for areas and individual pixels. Trend analyses are already incorporated as tools in the online platforms for major map products like the Landscape Change Monitoring System Data Explorer (https://apps.fs.usda.gov/lcms-viewer/). We examined trends in mean canopy cover for forest areas in Michigan and found wide interannual swings when using the maps based on Landsat composites without LandTrendr (Fig. 9b), which are likely caused by interannual spectral variations not related to actual changes in canopy cover. This finding is reflected in other studies comparing predictions derived from raw remote sensing values against those derived from spectral values fit with a time series algorithm (Moisen et al., 2016) or where smoothing is applied to the initial predictions (Powell et al., 2010). Spurious interannual variation illustrates what is perhaps the greatest problem with monitoring rather than direct change mapping — the error between two maps is compounded with post-classification change analysis and can lead to gross overestimates or underestimates of change (Coppin and Bauer, 1996). When measurement error exceeds the degree of real change, monitoring can give the false impression of high interannual variation or real changes may be mistakenly regarded as noise. Conversely, models like random forest tend to compress predictions towards the mean observed value (Zhang and Lu, 2012), which could lead to underestimates of real change when compounded across years. For temporal transfer, combining temporal cross-validation with small area estimation likely serves as the best proxy for checking the suitability of a model for post-classification comparison or trend analysis. Checking the plausibility of trends against outside information is another means to identify possible model misspecification with temporal transfer. For example, our comparison of canopy cover changes based on stand age (Fig. S6) confirmed that the models at least reflect expected trends based on stand succession.

While the framework we propose provides a means for checking possible model misspecification, there is still a need to expand upon this approach and, if possible, develop statistical methods that can actually quantify the additional uncertainty that comes with temporal transfer. After identifying model misspecification, the next step would be further investigation into its potential causes, which is beyond the scope of this study. As discussed in the introduction, there are many possible underlying reasons for a lack of temporal consistency in remote sensing data (Nagol et al., 2015; Roy et al., 2016; Sola et al., 2016), reference data (Westfall and Patterson, 2007), or their relationship to each other (Cleland et al., 2007; Melaas et al., 2013). Research and adoption of systematic ways to prevent, identify, and quantify these specific problems would be valuable. As one example, to guard against bias in model transfer, it would be wise to examine how predictor values of new times or areas exceed the range of values captured by the model such as with the dissimilarity index of Meyer and Pebesma (2021). This analysis could identify when and where more reference data are needed to fill gaps in model training and also convey to users where models are extrapolating. Our framework could also be extended to include confidence intervals for the model-based small area estimates, which is now more efficient for non-parametric models and feasible for map users (McRoberts et al., 2022). Bayesian model-based estimators that utilize geostatistical models offer several advantages over traditional approaches, including the ability to propagate errors through multiple stages of modeling to generate uncertainty estimates at both the unit and population level while accounting for spatial autocorrelation among reference plots (Babcock et al., 2018). Given the flexibility of these Bayesian models, they may also hold promise for properly quantifying uncertainty when extrapolating a model to new time periods and incorporating that uncertainty into credible intervals for individual population units or small area estimates. This would simplify communicating the uncertainty of temporal transfer to map users and enable error propagation into subsequent analyses.

5. Conclusion

Although, “no single combination of validation tests will be applicable across the diverse range of models and their uses” as noted by Mayer and Butler (1993), models should be evaluated against criteria that can be used to determine their acceptable use for a stated purpose and within a given context (Rykiel, 1996). While remote sensing has rapidly advanced towards continuous land monitoring, validation methods and criteria have failed to keep up. Evaluations often do not match the broad spatiotemporal contexts or purpose for which the models are designed, or they neglect spatial and temporal dependencies.

In this work, we present a framework for evaluating remote sensing models intended for large area monitoring, where extrapolation to time periods outside the domain of the training data becomes necessary. This framework was adapted from the temporal block cross-validation approach advocated by Roberts et al. (2017) and an adaptation of the spatial multiscale assessment techniques introduced by Riemann et al. (2010). Application of this framework to a case study of mapping percent tree canopy cover in Michigan revealed that temporal transfer of remote sensing models may lead to biased model-based estimators, even with transfer over short time lags. Our diagnostic tests also revealed that errors differ over spatial extents and scales. When temporally transferring models, use of a time series algorithm, such as LandTrendr, may yield more accurate predictions and increase the temporal consistency of small area estimates, but it should not be assumed to resolve all issues. Examining the potential for model misspecification with temporal transfer is still prudent. These practices serve as a steppingstone towards more robust uncertainty evaluation of remote sensing products generated through the temporal transfer of models and intended for a variety of common end user applications.

CRediT authorship contribution statement

Steven K. Filippelli: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Karen Schleeweis:** Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Writing – review & editing. **Mark D. Nelson:** Conceptualization, Funding acquisition, Methodology, Writing – review & editing. **Patrick A. Fekety:** Conceptualization, Methodology, Writing – review & editing. **Jody C. Vogeler:** Conceptualization, Funding acquisition, Methodology, Project administration, Supervision, Writing – review & editing.

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.

Data availability

Canopy cover maps for Michigan are available upon request. Canopy cover maps for the Great Lakes region, including Michigan, were created following very similar methods and are available through the USDA Forest Service Research Data Archive at: https://doi.org/10.2737/RDS-2023-0043.

Acknowledgements

This work was supported under USDA Forest Service Joint Venture Agreement 19-JV-11242305-066 and by the Great Lakes Restoration Initiative. The Forest Inventory and Analysis Program data used in this publication was made possible, in part, through the assistance of FIA Spatial Data Services and an Agreement from the United States Department of Agriculture's Forest Service (FS) (MTA #20-RD-11242305-034). We would like to thank Andrew Lister for advice on creating tessellations.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.srs.2024.100119.

References

Babcock, C., Finley, A.O., Andersen, H.-E., Pattison, R., Cook, B.D., Morton, D.C., Alonzo, M., Nelson, R., Gregoire, T., Ene, L., Gobakken, T., Næsset, E., 2018. Geostatistical estimation of forest biomass in interior Alaska combining Landsat-derived tree cover, sampled airborne lidar and field observations. Rem. Sens. Environ. 212, 212–230. https://doi.org/10.1016/j.rse.2018.04.044.

Bechtold, W.A., Patterson, P.L., 2005. The Enhanced Forest Inventory and Analysis Program - National Sampling Design and Estimation Procedures (No. SRS-GTR-80). U.S. Department of Agriculture, Forest Service, Southern Research Station, Asheville, NC. https://doi.org/10.2737/SRS-GTR-80.

Bergmeir, C., Benítez, J.M., 2012. On the use of cross-validation for time series predictor evaluation. Inf. Sci. 191, 192–213. https://doi.org/10.1016/j.ins.2011.12.028.

Blasi, M., Bartomeus, I., Bommarco, R., Gagic, V., Garratt, M., Holzschuh, A., Kleijn, D., Lindström, S.A.M., Olsson, P., Polce, C., Potts, S.G., Rundlöf, M., Scheper, J., Smith, H.G., Steffan-Dewenter, I., Clough, Y., 2021. Evaluating predictive performance of statistical models explaining wild bee abundance in a mass-flowering crop. Ecography 44, 525–536. https://doi.org/10.1111/ecog.05308.

Boryan, C., Yang, Z., Mueller, R., Craig, M., 2011. Monitoring US agriculture: the US department of agriculture, national agricultural statistics Service, cropland data layer program. Geocarto Int. 26, 341–358. https://doi.org/10.1080/10106049.2011.562309.

Botkin, D.B., Estes, J.E., MacDonald, R.M., Wilson, M.V., 1984. Studying the earth's vegetation from space. Bioscience 34, 508–514. https://doi.org/10.2307/1309693.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Breiman, L., 1996. Out-of-bag Estimation.

Brown, C.F., Brumby, S.P., Guzder-Williams, B., Birch, T., Hyde, S.B., Mazzariello, J., Czerwinski, W., Pasquarella, V.J., Haertel, R., Ilyushchenko, S., Schwehr, K., Weisse, M., Stolle, F., Hanson, C., Guinan, O., Moore, R., Tait, A.M., 2022. Dynamic World, Near real-time global 10 m land use land cover mapping. Sci. Data 9, 251. https://doi.org/10.1038/s41597-022-01307-4.

Brown, J.F., Tollerud, H.J., Barber, C.P., Zhou, Q., Dwyer, J.L., Vogelmann, J.E., Loveland, T.R., Woodcock, C.E., Stehman, S.V., Zhu, Z., Pengra, B.W., Smith, K., Horton, J.A., Xian, G., Auch, R.F., Sohl, T.L., Sayler, K.L., Gallant, A.L., Zelenak, D., Reker, R.R., Rover, J., 2020. Lessons learned implementing an operational continuous United States national land change monitoring capability: the Land Change Monitoring, Assessment, and Projection (LCMAP) approach. Remote Sensing of Environment, Time Series Analysis with High Spatial Resolution Imagery 238, 111356. https://doi.org/10.1016/j.rse.2019.111356.

Canty, M.J., Nielsen, A.A., Schmidt, M., 2004. Automatic radiometric normalization of multitemporal satellite imagery. Rem. Sens. Environ. 91, 441–451. https://doi.org/10.1016/j.rse.2003.10.024.

Carpenter, J., Bithell, J., 2000. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Stat. Med. 19, 1141–1164. https://doi.org/10.1002/(SICI)1097-0258(20000515)19:9<1141::AID-SIM479>3.0.CO;2-F.

Cleland, D.T., Crow, T.R., Saunders, S.C., Dickmann, D.I., Maclean, A.L., Jordan, J.K., Watson, R.L., Sloan, A.M., Brosofske, K.D., 2004. Characterizing historical and modern fire regimes in Michigan (USA): a landscape ecosystem approach. Landsc. Ecol. 19, 311–325. https://doi.org/10.1023/B:LAND.0000030437.29258.3c.

Cleland, E.E., Chuine, I., Menzel, A., Mooney, H.A., Schwartz, M.D., 2007. Shifting plant phenology in response to global change. Trends Ecol. Evol. 22, 357–365. https://doi.org/10.1016/j.tree.2007.04.003.

Cochran, W.G., 1977. Sampling techniques. In: Wiley Series in Probability and Mathematical Statistics, 3d ed. Wiley, New York.

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. Int. J. Rem. Sens. 22, 2279–2310. https://doi.org/10.1080/01431160121472.

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. https://doi.org/10.1016/j.rse.2010.07.010.

Coops, N.C., Tompalski, P., Goodbody, T.R.H., Queinnec, M., Luther, J.E., Bolton, D.K., White, J.C., Wulder, M.A., van Lier, O.R., Hermosilla, T., 2021. Modelling lidar-derived estimates of forest attributes over space and time: a review of approaches and future trends. Rem. Sens. Environ. 260, 112477 https://doi.org/10.1016/j.rse.2021.112477.

Coppin, P.R., Bauer, M.E., 1996. Digital change detection in forest ecosystems with remote sensing imagery. Rem. Sens. Rev. 13, 207–234.

Coulston, J.W., Moisen, G.G., Wilson, B.T., Finco, M.V., Cohen, W.B., Brewer, C.K., 2012. Modeling percent tree canopy cover: a pilot study. Photogramm. Eng. Rem. Sens. 78, 715–727.

Crist, E.P., 1985. A TM Tasseled Cap equivalent transformation for reflectance factor data. Rem. Sens. Environ. 17, 301–306. https://doi.org/10.1016/0034-4257(85)90102-6.

Dark, S.J., Bram, D., 2007. The modifiable areal unit problem (MAUP) in physical geography. Prog. Phys. Geogr. Earth Environ. 31, 471–479. https://doi.org/10.1177/0309133307083294.

Defourny, P., Bontemps, S., Bellemans, N., Cara, C., Dedieu, G., Guzzonato, E., Hagolle, O., Inglada, J., Nicola, L., Rabaute, T., Savinaud, M., Udroiu, C., Valero, S., Bégué, A., Dejoux, J.-F., El Harti, A., Ezzahar, J., Kussul, N., Labbassi, K., Lebourgeois, V., Miao, Z., Newby, T., Nyamugama, A., Salh, N., Shelestov, A., Simonneaux, V., Traore, P.S., Traore, S.S., Koetz, B., 2019. Near real-time agriculture monitoring at national scale at parcel resolution: performance assessment of the Sen2-Agri automated system in various cropping systems around the world. Rem. Sens. Environ. 221, 551–568. https://doi.org/10.1016/j.rse.2018.11.007.

Domingo, D., Alonso, R., Lamelas, M.T., Montealegre, A.L., Rodríguez, F., de la Riva, J., 2019. Temporal transferability of pine forest attributes modeling using low-density airborne laser scanning data. Rem. Sens. 11, 261. https://doi.org/10.3390/rs11030261.

Duncanson, L., Armston, J., Disney, M., Avitabile, V., Barbier, N., Calders, K., Carter, S., Chave, J., Herold, M., MacBean, N., 2021. Aboveground woody biomass product validation good practices protocol. Land Product Validation Subgroup (Working Group on Calibration and Validation, Committee on Earth Observation Satellites).

Efron, B., 1987. Better bootstrap confidence intervals. J. Am. Stat. Assoc. 82, 171–185.

ESRI, 2022. ArcGIS Pro.

Fekety, P.A., Falkowski, M.J., Hudak, A.T., 2015. Temporal transferability of LiDAR-based imputation of forest inventory attributes. Can. J. For. Res. 45, 422–435. https://doi.org/10.1139/cjfr-2014-0405.

Fekety, P.A., Falkowski, M.J., Hudak, A.T., Jain, T.B., Evans, J.S., 2018. Transferability of lidar-derived basal area and stem density models within a northern Idaho ecoregion. Can. J. Rem. Sens. 44, 131–143. https://doi.org/10.1080/07038992.2018.1461557.

Flood, N., 2013. Seasonal composite Landsat TM/ETM+ images using the medoid (a multi-dimensional median). Rem. Sens. 5, 6481–6500. https://doi.org/10.3390/rs5126481.

Foody, G.M., Boyd, D.S., Cutler, M.E.J., 2003. Predictive relations of tropical forest biomass from Landsat TM data and their transferability between regions. Rem. Sens. Environ. 85, 463–474. https://doi.org/10.1016/S0034-4257(03)00039-7.

Friedl, M.A., Woodcock, C.E., Olofsson, P., Zhu, Z., Loveland, T., Stanimirova, R., Arevalo, P., Bullock, E., Hu, K.-T., Zhang, Y., Turlej, K., Tarrio, K., McAvoy, K., Gorelick, N., Wang, J.A., Barber, C.P., Souza, C., 2022. Medium spatial resolution mapping of global land cover and land cover change across multiple decades from Landsat. Frontiers in Remote Sensing 3.

García-Santos, V., Sánchez, J.M., Cuxart, J., 2022. Evapotranspiration acquired with remote sensing thermal-based algorithms: a state-of-the-art review. Rem. Sens. 14, 3440. https://doi.org/10.3390/rs14143440.

Gong, P., Li, X., Wang, J., Bai, Y., Chen, B., Hu, T., Liu, X., Xu, B., Yang, J., Zhang, W., Zhou, Y., 2020. Annual maps of global artificial impervious area (GAIA) between 1985 and 2018. Rem. Sens. Environ. 236, 111510 https://doi.org/10.1016/j.rse.2019.111510.

Gregoire, T., 1998. Design-based and model-based inference in survey sampling: appreciating the difference. Can. J. For. Res. 28, 1429–1447. https://doi.org/10.1139/x98-166.

Hansen, M.C., Potapov, P.V., Moore, R., Hancher, M., Turubanova, S.A., Tyukavina, A., Thau, D., Stehman, S.V., Goetz, S.J., Loveland, T.R., 2013. High-resolution global maps of 21st-century forest cover change. Science 342, 850–853.

Hansen, M.H., Madow, W.G., Tepping, B.J., 1983. An evaluation of model-dependent and probability-sampling inferences in sample surveys. J. Am. Stat. Assoc. 78, 776–779.

Hantson, S., Chuvieco, E., 2011. Evaluation of different topographic correction methods for Landsat imagery. Int. J. Appl. Earth Obs. Geoinf. 13, 691–700. https://doi.org/10.1016/j.jag.2011.05.001.

Healey, S.P., Cohen, W.B., Yang, Z., Kenneth Brewer, C., Brooks, E.B., Gorelick, N., Hernandez, A.J., Huang, C., Joseph Hughes, M., Kennedy, R.E., Loveland, T.R., Moisen, G.G., Schroeder, T.A., Stehman, S.V., Vogelmann, J.E., Woodcock, C.E., Yang, L., Zhu, Z., 2018. Mapping forest change using stacked generalization: an ensemble approach. Rem. Sens. Environ. 204, 717–728. https://doi.org/10.1016/j.rse.2017.09.029.

Housman, I.W., Chastain, R.A., Finco, M.V., 2018. An evaluation of forest health insect and disease survey data and satellite-based remote sensing forest change detection methods: case studies in the United States. Rem. Sens. 10, 1184. https://doi.org/10.3390/rs10081184.

Huang, C., Wylie, B., Yang, L., Homer, C., Zylstra, G., 2002. Derivation of a tasselled cap transformation based on Landsat 7 at-satellite reflectance. Int. J. Rem. Sens. 23, 1741–1748. https://doi.org/10.1080/01431160110106113.

Irish, R.R., Barker, J.L., Goward, S.N., Arvidson, T., 2006. Characterization of the Landsat-7 ETM+ automated cloud-cover assessment (ACCA) algorithm. Photogramm. Eng. Rem. Sens. 72, 1179–1188.

Jennings, S.B., Brown, N.D., Sheil, D., 1999. Assessing forest canopies and understorey illumination: canopy closure, canopy cover and other measures. Forestry 72, 59–74.

Ji, L., Peters, A.J., 2003. Assessing vegetation response to drought in the northern Great Plains using vegetation and drought indices. Rem. Sens. Environ. 87, 85–98. https://doi.org/10.1016/S0034-4257(03)00174-3.

Karra, K., Kontgis, C., Statman-Weil, Z., Mazzariello, J.C., Mathis, M., Brumby, S.P., 2021. Global land use/land cover with Sentinel 2 and deep learning. In: 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS. Presented at the 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, pp. 4704–4707. https://doi.org/10.1109/IGARSS47720.2021.9553499.

Kennedy, R.E., Yang, Z., Cohen, W.B., 2010. Detecting trends in forest disturbance and recovery using yearly Landsat time series: 1. LandTrendr — temporal segmentation algorithms. Rem. Sens. Environ. 114, 2897–2910. https://doi.org/10.1016/j.rse.2010.07.008.

Kennedy, R.E., Yang, Z., Gorelick, N., Braaten, J., Cavalcante, L., Cohen, W.B., Healey, S., 2018. Implementation of the LandTrendr algorithm on Google Earth engine. Rem. Sens. 10, 691. https://doi.org/10.3390/rs10050691.

Key, C.H., Benson, N.C., 2006. Landscape assessment (LA). In: Lutes, Duncan C., Keane, Robert E., Caratti, John F., Key, Carl H., Benson, Nathan C., Sutherland, Steve, Gangi, Larry, J. (Eds.), 2006. FIREMON: Fire Effects Monitoring and Inventory System. (No. RMRS-GTR-164-CD), General Technical Report (GTR). US Department of Agriculture, Forest Service, Rocky Mountain Research Station, Fort Collins, CO.

Kim, D.-H., Sexton, J.O., Noojipady, P., Huang, C., Anand, A., Channan, S., Feng, M., Townshend, J.R., 2014. Global, Landsat-based forest-cover change from 1990 to 2000. Rem. Sens. Environ. 155, 178–193. https://doi.org/10.1016/j.rse.2014.08.017.

Kottek, M., Grieser, J., Beck, C., Rudolf, B., Rubel, F., 2006. World map of the Köppen-Geiger climate classification updated. Meteorol. Z. 15, 259–263.

Krivoruchko, K., Gribov, A., 2020. Distance metrics for data interpolation over large areas on Earth's surface. Spatial Statistics 35, 100396. https://doi.org/10.1016/j.spasta.2019.100396.

Landsat Known Issues [WWW Document], n.d. URL https://www.usgs.gov/core-science-systems/nli/landsat/landsat-known-issues (accessed November.30.21).

Majumdar, S., Smith, R., Butler, J.J., Lakshmi, V., 2020. Groundwater withdrawal prediction using integrated multitemporal remote sensing data sets and machine learning. Water Resour. Res. 56, e2020WR028059 https://doi.org/10.1029/2020WR028059.

Marston, C., Rowland, C.S., O'Neil, A.W., Morton, R.D., 2022. Land Cover Map 2021 (10m Classified Pixels, GB). NERC EDS Environmental Information Data Centre. https://doi.org/10.5285/a22baa7c-5809-4a02-87e0-3cf87d4e223a.

Mauro, F., Hudak, A.T., Fekety, P.A., Frank, B., Temesgen, H., Bell, D.M., Gregory, M.J., McCarley, T.R., 2021. Regional modeling of forest fuels and structural attributes using airborne laser scanning data in Oregon. Rem. Sens. 13, 261. https://doi.org/10.3390/rs13020261.

Mayer, D.G., Butler, D.G., 1993. Statistical validation. Ecological Modelling, Theoretical Modelling Aspects 68, 21–32. https://doi.org/10.1016/0304-3800(93)90105-2.

McRoberts, R.E., 2011. Satellite image-based maps: scientific inference or pretty pictures? Rem. Sens. Environ. 115, 715–724. https://doi.org/10.1016/j.rse.2010.10.013.

McRoberts, R.E., 2010. Probability- and model-based approaches to inference for proportion forest using satellite imagery as ancillary data. Rem. Sens. Environ. 114, 1017–1025. https://doi.org/10.1016/j.rse.2009.12.013.

McRoberts, R.E., Næsset, E., Saatchi, S., Quegan, S., 2022. Statistically rigorous, model-based inferences from maps. Rem. Sens. Environ. 279, 113028 https://doi.org/10.1016/j.rse.2022.113028.

McRoberts, R.E., Tomppo, E.O., Finley, A.O., Heikkinen, J., 2007. Estimating areal means and variances of forest attributes using the k-Nearest Neighbors technique and satellite imagery. Rem. Sens. Environ. 111, 466–480. https://doi.org/10.1016/j.rse.2007.04.002.

Melaas, E.K., Friedl, M.A., Zhu, Z., 2013. Detecting interannual variation in deciduous broadleaf forest phenology using Landsat TM/ETM+ data. Rem. Sens. Environ. 132, 176–185. https://doi.org/10.1016/j.rse.2013.01.011.

Melton, F.S., Huntington, J., Grimm, R., Herring, J., Hall, M., Rollison, D., Erickson, T., Allen, R., Anderson, M., Fisher, J.B., Kilic, A., Senay, G.B., Volk, J., Hain, C., Johnson, L., Ruhoff, A., Blankenau, P., Bromley, M., Carrara, W., Daudert, B., Doherty, C., Dunkerly, C., Friedrichs, M., Guzman, A., Halverson, G., Hansen, J., Harding, J., Kang, Y., Ketchum, D., Minor, B., Morton, C., Ortega-Salazar, S., Ott, T., Ozdogan, M., ReVelle, P.M., Schull, M., Wang, C., Yang, Y., Anderson, R.G., 2021. OpenET: filling a critical data gap in water management for the western United States. JAWRA Journal of the American Water Resources. https://doi.org/10.1111/1752-1688.12956. Association n/a.

Menlove, J., Healey, S.P., 2020. A comprehensive forest biomass dataset for the USA allows customized validation of remotely sensed biomass estimates. Rem. Sens. 12, 4141. https://doi.org/10.3390/rs12244141.

Meyer, H., Pebesma, E., 2021. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. Methods Ecol. Evol. 12, 1620–1633. https://doi.org/10.1111/2041-210X.13650.

Michigan Department of Natural Resources, 2008. Michigan State Forest Management Plan. Michigan Department of Natural Resources, Forest, Mineral, and Fire Management Division and Wildlife Division, Lansing, Michigan.

Minter, T.C., 1979. Methods of extending crop signatures from one area to another. In: Proceedings of Technical Sessions, Volumes 1 and 2. Presented at the LACIE Symposium, NASA. Johnson Space Center, Houston, Texas.

Moisen, G.G., Meyer, M.C., Schroeder, T.A., Liao, X., Schleeweis, K.G., Freeman, E.A., Toney, C., 2016. Shape selection in Landsat time series: a tool for monitoring forest dynamics. Global Change Biol. 22, 3518–3528. https://doi.org/10.1111/gcb.13358.

Mulder, V.L., de Bruin, S., Schaepman, M.E., Mayr, T.R., 2011. The use of remote sensing in soil and terrain mapping — a review. Geoderma 162, 1–19. https://doi.org/10.1016/j.geoderma.2010.12.018.

Muller, E., Décamps, H., 2001. Modeling soil moisture–reflectance. Rem. Sens. Environ. 76, 173–180. https://doi.org/10.1016/S0034-4257(00)00198-X.

Næsset, E., McRoberts, R.E., Pekkarinen, A., Saatchi, S., Santoro, M., Trier, Ø.D., Zahabu, E., Gobakken, T., 2020. Use of local and global maps of forest canopy height and aboveground biomass to enhance local estimates of biomass in miombo woodlands in Tanzania. Int. J. Appl. Earth Obs. Geoinf. 93, 102138 https://doi.org/10.1016/j.jag.2020.102138.

Nagol, J.R., Sexton, J.O., Kim, D.-H., Anand, A., Morton, D., Vermote, E., Townshend, J. R., 2015. Bidirectional effects in Landsat reflectance estimates: is there a problem to solve? ISPRS J. Photogram. Remote Sensing, Global Land Cover Mapping and Monitor. 103, 129–135. https://doi.org/10.1016/j.isprsjprs.2014.09.006.

NASA, 1983. AgRISTARS: Agriculture and Resources Inventory Surveys through Erospace Remote Sensing, Research Report, Fiscal Year 1982 (Technical Memorandum No. E83-10318). National Aeronautics and Space Administration, Houston, TX.

Navarro, J.A., Tomé, J.L., Marino, E., Guillén-Climent, M.L., Fernández-Landa, A., 2020. Assessing the transferability of airborne laser scanning and digital aerial photogrammetry derived growing stock volume models. Int. J. Appl. Earth Obs. Geoinf. 91, 102135 https://doi.org/10.1016/j.jag.2020.102135.

Nedyalkova, D., Tillé, Y., 2012. Bias-robustness and efficiency of model-based inference in survey sampling. Stat. Sin. 777–794.

Nelson, M.D., Healey, S.P., Moser, W.K., Masek, J.G., Cohen, W.B., 2011. Consistency of forest presence and biomass predictions modeled across overlapping spatial and temporal extents. Math. Comput. For. Nat. Resour. Sci. 3, 102–113.

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. Rem. Sens. Environ. 148, 42–57. https://doi.org/10.1016/j.rse.2014.02.015.

Pax-Lenney, M., Woodcock, C.E., Macomber, S.A., Gopal, S., Song, C., 2001. Forest mapping with a generalized classifier and Landsat TM data. Rem. Sens. Environ. 77, 241–250. https://doi.org/10.1016/S0034-4257(01)00208-5.

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

Potapov, P., Hansen, M.C., Pickens, A., Hernandez-Serna, A., Tyukavina, A., Turubanova, S., Zalles, V., Li, X., Khan, A., Stolle, F., Harris, N., Song, X.-P., Baggett, A., Kommareddy, I., Kommareddy, A., 2022. The global 2000-2020 land cover and land use change dataset derived from the Landsat archive: first results. Frontiers in Remote Sensing 3.

Powell, S.L., Cohen, W.B., Healey, S.P., Kennedy, R.E., Moisen, G.G., Pierce, K.B., Ohmann, J.L., 2010. Quantification of live aboveground forest biomass dynamics with Landsat time-series and field inventory data: a comparison of empirical modeling approaches. Rem. Sens. Environ. 114, 1053–1068. https://doi.org/10.1016/j.rse.2009.12.018.

Ramsey, J.B., 1969. Tests for specification errors in classical linear least-squares regression analysis. J. Roy. Stat. Soc. B 31, 350–371.

Rao, J.N.K., Molina, I., 2015. Small Area Estimation. John Wiley & Sons, Incorporated.

Riano, D., Chuvieco, E., Salas, J., Aguado, I., 2003. Assessment of different topographic corrections in Landsat-TM data for mapping vegetation types (2003). IEEE Trans. Geosci. Rem. Sens. 41, 1056–1061. https://doi.org/10.1109/TGRS.2003.811693.

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. Rem. Sens. Environ. 114, 2337–2352. https://doi.org/10.1016/j.rse.2010.05.010.

Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J.J., Schröder, B., Thuiller, W., Warton, D.I., Wintle, B.A., Hartig, F., Dormann, C.F., 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40, 913–929. https://doi.org/10.1111/ecog.02881.

Roberts, D.W., Cooper, S.V., 1989. Concepts and Techniques of Vegetation Mapping. General Technical Report INT-US Department of Agriculture, Forest Service, Intermountain Research Station (USA).

Rouse, J.W., Haas, R.H., Schell, J.A., Deering, D.W., 1974. Monitoring vegetation systems in the Great plains with ERTS. In: Proceedings. Presented at the 3rd Earth Resource Technology Satellite (ERTS) Symposium, pp. 48–62.

Roy, D.P., Kovalskyy, V., Zhang, H.K., Vermote, E.F., Yan, L., Kumar, S.S., Egorov, A., 2016. Characterization of Landsat-7 to Landsat-8 reflective wavelength and normalized difference vegetation index continuity. Remote Sensing of Environment, Landsat 8 Science Results 185, 57–70. https://doi.org/10.1016/j.rse.2015.12.024.

Roy, D.P., Li, Z., Zhang, H.K., Huang, H., 2020. A conterminous United States analysis of the impact of Landsat 5 orbit drift on the temporal consistency of Landsat 5 Thematic Mapper data. Rem. Sens. Environ. 240, 111701 https://doi.org/10.1016/j.rse.2020.111701.

Rykiel, E.J., 1996. Testing ecological models: the meaning of validation. Ecol. Model. 90, 229–244. https://doi.org/10.1016/0304-3800(95)00152-2.

Särndal, C.-E., Swensson, B., Wretman, J., 1992. Model Assisted Survey Sampling, first ed. Springer Science & Business Media, New York, NY.

Schleeweis, K.G., Moisen, G.G., Schroeder, T.A., Toney, C., Freeman, E.A., Goward, S.N., Huang, C., Dungan, J.L., 2020. US national maps attributing forest change: 1986–2010. Forests 11, 653. https://doi.org/10.3390/f11060653.

Schroeder, T.A., Cohen, W.B., Song, C., Canty, M.J., Yang, Z., 2006. Radiometric correction of multi-temporal Landsat data for characterization of early successional forest patterns in western Oregon. Rem. Sens. Environ. 103, 16–26. https://doi.org/10.1016/j.rse.2006.03.008.

Schroeder, T.A., Cohen, W.B., Yang, Z., 2007. Patterns of forest regrowth following clearcutting in western Oregon as determined from a Landsat time-series. For. Ecol. Manag. 243, 259–273. https://doi.org/10.1016/j.foreco.2007.03.019.

Schroeder, T.A., Healey, S.P., Moisen, G.G., Frescino, T.S., Cohen, W.B., Huang, C., Kennedy, R.E., Yang, Z., 2014. Improving estimates of forest disturbance by combining observations from Landsat time series with U.S. Forest Service Forest Inventory and Analysis data. Rem. Sens. Environ. 154, 61–73. https://doi.org/10.1016/j.rse.2014.08.005.

Smith, H.D., Dubeux, J.C.B., Zare, A., Wilson, C.H., 2023. Assessing transferability of remote sensing pasture estimates using multiple machine learning algorithms and evaluation structures. Rem. Sens. 15, 2940. https://doi.org/10.3390/rs15112940.

Sola, I., González-Audícana, M., Álvarez-Mozos, J., 2016. Multi-criteria evaluation of topographic correction methods. Rem. Sens. Environ. 184, 247–262. https://doi.org/10.1016/j.rse.2016.07.002.

Stage, A.R., 1976. An expression for the effect of aspect, slope, and habitat type on tree growth. For. Sci. 22, 457–460.

Ståhl, G., Saarela, S., Schnell, S., Holm, S., Breidenbach, J., Healey, S.P., Patterson, P.L., Magnussen, S., Næsset, E., McRoberts, R.E., Gregoire, T.G., 2016. Use of models in large-area forest surveys: comparing model-assisted, model-based and hybrid estimation. Forest Ecosystems 3. https://doi.org/10.1186/s40663-016-0064-9.

Theobald, D.M., Harrison-Atlas, D., Monahan, W.B., Albano, C.M., 2015. Ecologically-relevant maps of landforms and physiographic diversity for climate adaptation planning. PLoS One 10, e0143619. https://doi.org/10.1371/journal.pone.0143619.

Thomas, N.E., Huang, C., Goward, S.N., Powell, S., Rishmawi, K., Schleeweis, K., Hinds, A., 2011. Validation of North American Forest Disturbance dynamics derived from Landsat time series stacks. Rem. Sens. Environ. 115, 19–32. https://doi.org/10.1016/j.rse.2010.07.009.

USDA Forest Service, 2023. Forest Inventory EVALIDator Web-Application.

USDA Forest Service, 2021. Landscape Change Monitoring System Conterminous United States.

Verbesselt, J., Hyndman, R., Newnham, G., Culvenor, D., 2010. Detecting trend and seasonal changes in satellite image time series. Rem. Sens. Environ. 114, 106–115. https://doi.org/10.1016/j.rse.2009.08.014.

Vermote, E.F., Tanre, D., Deuze, J.L., Herman, M., Morcette, J.-J., 1997. Second simulation of the satellite signal in the solar spectrum, 6S: an overview. IEEE Trans. Geosci. Rem. Sens. 35, 675–686. https://doi.org/10.1109/36.581987.

Vulova, S., Meier, F., Rocha, A.D., Quanz, J., Nouri, H., Kleinschmit, B., 2021. Modeling urban evapotranspiration using remote sensing, flux footprints, and artificial intelligence. Sci. Total Environ. 786, 147293 https://doi.org/10.1016/j.scitotenv.2021.147293.

Weiss, A., 2001. Topographic position and landforms analysis. In: Poster Presentation, ESRI User Conference. San Diego, CA.

Westfall, J.A., Patterson, P.L., 2007. Measurement variability error for estimates of volume change. Can. J. For. Res. 37, 2201–2210. https://doi.org/10.1139/X07-082.

Woodcock, C.E., Loveland, T.R., Herold, M., Bauer, M.E., 2020. Transitioning from change detection to monitoring with remote sensing: a paradigm shift. Remote Sensing of Environment, Time Series Analysis with High Spatial Resolution Imagery 238, 111558. https://doi.org/10.1016/j.rse.2019.111558.

Woodcock, C.E., Macomber, S.A., Pax-Lenney, M., Cohen, W.B., 2001. Monitoring large areas for forest change using Landsat: generalization across space, time and Landsat sensors. Remote Sens. Envrion. Lett. 7 78, 194–203. https://doi.org/10.1016/S0034-4257(01)00259-0.

Wulder, M.A., Masek, J.G., Cohen, W.B., Loveland, T.R., Woodcock, C.E., 2012. Opening the archive: how free data has enabled the science and monitoring promise of Landsat. Remote Sensing of Environment, Landsat Legacy Special Issue 122, 2–10. https://doi.org/10.1016/j.rse.2012.01.010.

Wulder, M.A., Roy, D.P., Radeloff, V.C., Loveland, T.R., Anderson, M.C., Johnson, D.M., Healey, S., Zhu, Z., Scambos, T.A., Pahlevan, N., Hansen, M., Gorelick, N., Crawford, C.J., Masek, J.G., Hermosilla, T., White, J.C., Belward, A.S., Schaaf, C., Woodcock, C.E., Huntington, J.L., Lymburner, L., Hostert, P., Gao, F., Lyapustin, A., Pekel, J.-F., Strobl, P., Cook, B.D., 2022. Fifty years of Landsat science and impacts. Rem. Sens. Environ. 280, 113195 https://doi.org/10.1016/j.rse.2022.113195.

Zanaga, D., Van De Kerchove, R., Daems, D., De Keersmaecker, W., Brockmann, C., Kirches, G., Wevers, J., Cartus, O., Santoro, M., Fritz, S., 2022. ESA WorldCover 10 M 2021 V200.

Zhang, G., Lu, Y., 2012. Bias-corrected random forests in regression. J. Appl. Stat. 39, 151–160. https://doi.org/10.1080/02664763.2011.578621.

Zhao, F., Huang, C., Goward, S.N., Schleeweis, K., Rishmawi, K., Lindsey, M.A., Denning, E., Keddell, L., Cohen, W.B., Yang, Z., Dungan, J.L., Michaelis, A., 2018. Development of Landsat-based annual US forest disturbance history maps (1986–2010) in support of the North American Carbon Program (NACP). Rem. Sens. Environ. 209, 312–326. https://doi.org/10.1016/j.rse.2018.02.035.

Zhong, L., Gong, P., Biging, G.S., 2014. Efficient corn and soybean mapping with temporal extendability: a multi-year experiment using Landsat imagery. Rem. Sens. Environ. 140, 1–13. https://doi.org/10.1016/j.rse.2013.08.023.

Zhou, B., Okin, G.S., Zhang, J., 2020. Leveraging Google Earth Engine (GEE) and machine learning algorithms to incorporate in situ measurement from different times for rangelands monitoring. Rem. Sens. Environ. 236, 111521 https://doi.org/10.1016/j.rse.2019.111521.