Costanza et al. Forest Ecosystems (2018) 5:8
DOI 10.1186/s40663-017-0123-x

Forest Ecosystems

RESEARCH Open Access

Classifying forest inventory data into species-based forest community types at broad extents: exploring tradeoffs among supervised and unsupervised approaches

Jennifer K. Costanza[1*], Don Faber-Langendoen[2], John W. Coulston[3] and David N. Wear[4]

Abstract

Background: Knowledge of the different kinds of tree communities that currently exist can provide a baseline for assessing the ecological attributes of forests and monitoring future changes. Forest inventory data can facilitate the development of this baseline knowledge across broad extents, but they first must be classified into forest community types. Here, we compared three alternative classifications across the United States using data from over 117,000 U.S. Department of Agriculture Forest Service Forest Inventory and Analysis (FIA) plots.

Methods: Each plot had three forest community type labels: (1) "FIA" types were assigned by the FIA program using a supervised method; (2) "USNVC" types were assigned via a key based on the U.S. National Vegetation Classification; (3) "empirical" types resulted from unsupervised clustering of tree species information. We assessed the degree to which analog classes occurred among classifications, compared indicator species values, and used random forest models to determine how well the classifications could be predicted using environmental variables.

Results: The classifications generated groups of classes that had broadly similar distributions, but often there was no one-to-one analog across the classifications. The longleaf pine forest community type stood out as the exception: it was the only class with strong analogs across all classifications. Analogs were most lacking for forest community types with species that occurred across a range of geographic and environmental conditions, such as loblolly pine types. Indicator species metrics were generally high for the USNVC, suggesting that USNVC classes are floristically well-defined. The empirical classification was best predicted by environmental variables. The most important predictors differed slightly but were broadly similar across all classifications, and included slope, amount of forest in the surrounding landscape, average minimum temperature, and other climate variables.

Conclusions: The classifications have similarities and differences that reflect their differing approaches and objectives. They are most consistent for forest community types that occur in a relatively narrow range of environmental conditions, and differ most for types with wide-ranging tree species. Environmental variables at a variety of scales were important for predicting all classifications, though strongest for the empirical and FIA, suggesting that each is useful for studying how forest communities respond to of multi-scale environmental processes, including global change drivers.

Keywords: Big data, Correspondence analysis, Dominant species, Forest communities, Global change, Hierarchical classification, Indicator species, Random forests, Species assemblages

* Correspondence: jennifer_costanza@ncsu.edu
[1]Department of Forestry and Environmental Resources, North Carolina State University, 3041 Cornwallis Rd., Research Triangle Park, Raleigh, NC 27709, USA
Full list of author information is available at the end of the article

Background

The identity and composition of tree species in a forest community can affect the ecological functions and ecosystem services that forests provide. For example, in many temperate forest communities in the eastern United States (U.S.), recent decreases in the abundance of oak species have been associated with impacts on biodiversity, wildlife habitat, and water quantity (Fralish 2004; Nowacki and Abrams 2008; Hanberry 2013; Hiers et al. 2014; Caldwell et al. 2016). Because changes in tree species composition can affect forest functions in these ways, characterizing species composition of existing forest communities is important for understanding the functions of those communities (Tierney et al. 2009; Thompson et al. 2013). And, as forests change in response to climate and land use, characterization and classification of their species assemblages can aid in understanding reference conditions, monitoring changes in species composition over time, and detecting early warning signs of vulnerability to those global changes (Tierney et al. 2009). While much of climate change impact and vulnerability research emphasizes the responses of individual species, focusing on communities allows incorporation of the relative occurrence of species–that is, the dominance (or conversely, the evenness) of species (Hildebrand et al. 2008; Levine et al. 2017). The relative occurrence of species in a forest ecosystem can influence species interactions, and is also likely to be influenced by environmental changes (Kardol et al. 2010; Le Roux et al. 2014). For example, in a recent study in the middle and eastern U.S., knowing which tree species were dominant within forest community types and quantifying the potential threats to those species were critical for determining the vulnerability of forest communities to climate change (Brandt et al. 2017). Because global change drivers are expected to affect the distribution of species from local to broad extents, characterizing species-based forest community types at those broad extents will be especially critical for monitoring and projecting the effects of global change on forest communities.

An increasing number of data sources are becoming available for understanding, characterizing and classifying forest communities across broad extents. Field-observed vegetation data (plot data), which contain information on species occurrence, abundance, and structure, are a particularly important source of information for characterizing community composition (Franklin et al. 2017). One source of vegetation plot data is national forest inventories, which provide forest community information on a systematic sampling grid at relatively fine grains (average plot sizes of 0.04 ha globally, Liang et al. 2016) and relatively large extents, often across regions or countries. All of these plot data are becoming more widely accessible at broad extents and can thus be used for regional, national, and even global studies (Dengler et al. 2011; Peet et al. 2012; Liang et al. 2016). With the increased availability of species-based forest inventory plot data comes the opportunity to explore various methods for classifying these data into species-based community types. Understanding the relative merits of those methods will be critical for informing potential users of those classifications.

As with any classification problem, techniques for classifying forest inventory plot data into forest communities can range from supervised to unsupervised. In a purely supervised classification approach, each plot is labeled with a pre-determined forest community type based on the characteristics observed at the location. Supervised approaches are often done using a decision rule or expert knowledge. In contrast, in a purely unsupervised, empirical, or "data-driven" classification, inventory data are partitioned into community type classes based on the relative similarity of their plot characteristics without regard to any published authority (Costanza et al. 2017). Because no classification of forest community types will be suitable for all research questions, it is important to understand the strengths and weaknesses of each. However, comparisons of alternative forest community classifications, especially among supervised and unsupervised approaches across broad extents, have rarely been done (but see Tichý et al. 2014 for comparisons using local data).

Here, we compared three species-based classifications of forest community types using forest inventory data across the United States (Table 1). The three classifications range from supervised to unsupervised. Inventory plot data came from the Forest Inventory and Analysis (FIA) database, which is produced by the U.S. Department of Agriculture (USDA) Forest Service. The FIA program applies a nationally-consistent sampling design, with one permanent plot established for every 2428 ha of land (Bechtold and Patterson 2005). Data collected by field crews include the diameter and species of every tree in each plot. More detail about plot design and data collection can be found in the Methods section. The three classifications of FIA plots we compared were: (1) FIA forest type groups (hereafter, "FIA" classes) (2) forest macrogroups from the U.S. National Vegetation Classification ("USNVC") (3) empirically-derived types ("empirical"). Each of these classifications is based, at least in part, on species composition within each plot. The FIA classes are based on the cover of dominant species (Eyre 1980), and are assigned to the classes using a supervised key developed by FIA staff. The USNVC classes are based on the composition of all species, plus environmental and disturbance data, and were developed using a combination of local vegetation data and expert

Table 1 The main attributes of the classifications used here. See text for further details and explanation

Name	Classification	Primary uses	Origin	Primary criteria	# classes included here (conterminous U.S., eastern U.S.)	Supervised / unsupervised	Extent (for this project)[a]
Empirical	Empirical classification	Assessment and projection of tree species assemblage change over time based on FIA data	Recent: published in 2017 (Costanza et al. 2017)	All native tree species, excluding the most rare	29, 17	Unsupervised	Conterminous U.S.
FIA	FIA forest type groups	Forest inventory and assessment based on FIA data	Forest types developed in the 1980s (Eyre 1980)	Emphasis on dominant tree species	28, 17	Supervised	Conterminous U.S.
USNVC	USNVC Macrogroups	Vegetation inventory, monitoring, assessment, mapping, conservation planning	Iterative: version 2.0 published in 2016 (usnvc.org); tree key recently applied to FIA plots (Menard et al. 2017)	All plant species, plus environmental conditions and disturbance; a tree-based key was used to as sign FIA plots to USNVC classes.	N/A, 25	A combination of supervised and unsupervised	Eastern U.S.

[a]Extent reflects the extent of the classifications we used in this paper; FIA and USNVC classes exist for the entire U.S. and its territories

opinion (Franklin et al. 2015; USNVC 2016). The USNVC classes were applied to FIA plots using a supervised key developed by FIA and NatureServe staff (Menard et al. 2017). The empirical classes are the result of an unsupervised cluster analysis of all tree species in FIA plots (Costanza et al. 2017).

Because the methods used in these approaches differ, the resulting classifications will differ, and we examine some of the differences here. Understanding how these approaches lead to differences or similarities in the resulting classifications will be critical for those who wish to use these classifications to examine forest-related questions. In addition, each of these classifications have been used already to study recent and potential future changes to forest communities (e.g., Iverson and Prasad 2001; Palmquist et al. 2014; Costanza et al. 2017) and each has the potential to support additional research on forest change. While an assessment of the performance of these classifications in models of forest change was beyond the scope of this paper, we here examine how well the classifications corresponded with local and broad-scale environmental variables. As climate and environmental conditions change, an approach that produces classes with high fidelity to climate and environmental conditions will likely be useful in detecting future forest community changes.

We used a database of FIA plots that systematically cover the conterminous United States. Each plot was labeled with the three classifications to ask the following questions:

- When the classifications are compared pairwise, which classes in each classification have high or low fidelity to each other, both in terms of the plots assigned to each and in terms of average species composition?
- Which classifications are best characterized by indicator and dominant species?
- Which classifications are predicted well by environmental variables, and which environmental variables are important for predicting them?

The results will provide information to potential users of these specific classifications, and will point to broader insights into the differences, similarities, and best uses of forest community type classifications that result from different approaches and developed for different objectives. Those insights will be critical as the global availability of large forest and vegetation data sets increases.

Methods

Forest inventory data

Forest plot observations from across the conterminous U.S. were extracted from the FIA database (FIADB version 6.1; O'Connell et al. 2016). The FIA program uses a nationally-consistent, sample-based statistical design to quantify forest conditions across the U.S., and is the primary source for information about the status and trends of U.S. forest resources (Smith 2002). The FIA program samples all forest and other land uses, with one permanent plot established for every 2428 ha of land (Bechtold and Patterson 2005). FIA plots consist of four

7.2 m fixed radius subplots (0.067 ha or 67 m^2 each), with three subplots spaced 36.6 m apart in a triangular arrangement and one subplot in the center (O'Connell et al. 2016). Data collected for forested plots by field crews include the diameter and species of every tree stem in each plot. To protect sensitive plot information, especially on privately owned lands, the publicly available FIA database contains plot location information that has been altered slightly from the true location. We used actual plot locations to extract environmental variables at plot locations (described below), but show altered locations in Fig. 1. To ensure that we included the full set of plots for each state in the conterminous U.S., we selected the set of plots for each state that was used by FIA to produce a recent evaluation of forest conditions. For most states, the data represented evaluations from 2013, but for some states, evaluations were from 2012, and the evaluation for one state (Tennessee) was from 2009.

Because FIA plots have fixed points and layout, they can straddle, for example, multiple forest types, and to track those differences, multiple "conditions" can be identified within a plot. Each condition within a plot represents all or a portion of one or more subplots, and can have a different forest type and forest type group assigned to it (O'Connell et al. 2016). Therefore, to ensure that our sample units were each labeled with only a single forest type group and ensure our sample units represent a relatively homogenous tree species assemblage, we used the conditions within plots as our unit for classification. The FIA database contains information on each condition and the proportion of the plot it represents. For each plot, we used the single condition that made up the greatest proportion of the plot. In the case of ties, we picked one condition at random. We refer to these conditions as "plots" throughout this paper. We excluded plots labeled as being of planted origin. We also excluded plots labeled with FIA forest type groups "Exotic Hardwoods" or "Exotic Softwoods" because the empirical classification was based on native U.S. species only. We also excluded plots labeled as "Nonstocked" because those typically have few, if any, trees and we are specifically interested here in tree species composition.

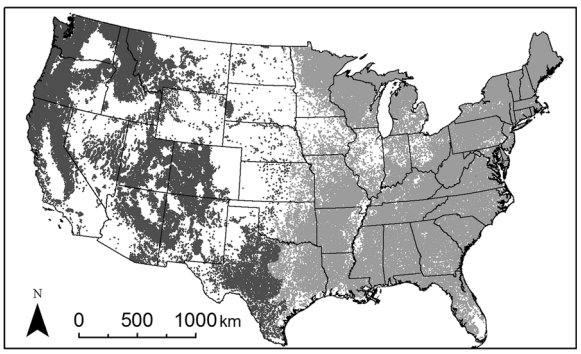

Fig. 1 Plots included in the eastern U.S. (locations shown in light gray) and outside the eastern U.S. (dark gray). Plots are spaced at approximately one per 2428 ha of forest land, but due to the scale of the map, in some areas, individual plots cannot be distinguished

For assigning plots to empirical classes, calculating species composition similarity, and analyzing indicator and dominant species, we needed information on species composition within each plot. For each plot, we assembled data on relative importance value by species for each plot we included. A species' relative importance value is the average of its relative abundance and relative basal area compared to other species in the plot. Trees greater than 2.54 cm (1 in.) diameter at breast height were used to calculate importance value. The most rare species (those occurring on less than 250 plots) were excluded to avoid biasing the analysis toward extremely rare species (McCune and Grace 2002), and nonnative species were excluded because those species may respond to different environmental drivers than native species. We used the same data set of relative importance by tree species for the comparisons among classifications as well as indicator and dominant species analysis described below (see Costanza et al. 2017 for a list of all common and scientific names of tree species used here).

The three classifications and their assignment to FIA plots

The FIA class assignment for each plot came directly from the FIA database. Personnel from the USDA Forest Service assign forest types and forest type groups to its FIA plots using a decision-tree approach based on the relative stocking values of tree species in the plot, which are primarily a function of basal area (Arner et al. 2003). The combination of related species that comprises the highest stocking values is used to assign most of the forest types, and related forest types are aggregated into forest type groups. The FIA forest types are based on a number of sources, including a published set of types and groups that emphasize the cover of the dominant species and are largely derived from a classification by Eyre (1980). Thus, the FIA classification is a supervised approach, and the FIA groupings are static and well-recognized. The FIA forest types and forest type groups are widely used in maps of forest resources, ecosystem services, pest impacts and risk, and to inform initial conditions in forest landscape dynamics modeling studies (Ruefenacht et al. 2008; DeSantis et al. 2013; Duveneck et al. 2015; Healey et al. 2016).

The USNVC is a federal vegetation classification standard (FGDC 2008) that was developed independently of the FIA program, and applies to all vegetation. In the USNVC, forest types are classified at multiple

levels using both tree and non-tree species, growth forms and ecology (the EcoVeg approach of Faber-Langendoen et al. 2014). Types have been recognized based on a combination of local to regional empirical data, literature, and expert judgement (Franklin et al. 2015; USNVC 2016). USNVC types have largely been formally described and published (USNVC 2016; http://www.usnvc.org), but the classification is dynamic and open to peer-reviewed changes. The classification has used been to understand the distribution of vegetation types locally to nationally (Hoagland 2000; Matthews et al. 2011; Belote et al. 2017), and by land management agencies for conservation planning (Franklin et al. 2015). Although the USNVC standard is national, the classification is based on vegetation plot data from a variety of sources, along with ancillary environmental characteristics such as soil wetness and disturbance information (Franklin et al. 2015).

Recently, a supervised decision tree, or key, was developed to assign FIA plots to one level of the hierarchy, the macrogroup level, in the eastern U.S. (Menard et al. 2017). The macrogroup is a mid-scale unit (5th level, approximately at the scale of FIA forest type groups) and is based on sets of diagnostic and dominant plant species and growth forms that reflect biogeographic differences in composition and sub-continental to regional differences in ecological factors (FGDC 2008). The macrogroup classes were developed based on ancillary local data and expert opinion for all vegetation types, not just forests, but the supervised key applies them to tree data in the FIA database. Expansion of the key to the entire country has been proposed but not yet implemented. Because the bulk of FIA data only contain tree species data and limited ecological data, the key is necessarily restricted to those data, though the ecoregion within which each plot is located is also used as a proxy for non-tree species and ecological factors. The USNVC classes were assigned to FIA plots based on the relative importance value information for each species described above, including trees greater than 2.54 cm (1 in.) diameter at breast height within the plot. Although the key assigns FIA plots to both "cultural" (plantations and other planted stands) and "natural" (non-planted) types, only plots falling into one of the natural macrogroups were included here.

The third classification of FIA plots was a recently developed set of empirically-derived forest community types (Costanza et al. 2017). The empirical classification used an unsupervised hierarchical method to cluster tree species composition information within FIA inventory plots across the conterminous U.S. (Costanza et al. 2017). The clusters were based on the relative importance value for species in each plot, using the same set of species described above and trees greater than 2.54 cm (1 in.) diameter at breast height (Costanza et al. 2017).

Thus, unlike the FIA classes, which rely most strongly on one, two or three dominant species, the empirical classes are defined based on the composition of a larger group of species in a plot. This classification was developed to facilitate the study of forest dynamics over time in response to global change scenarios (Costanza et al. 2017).

The empirical classification is hierarchical, and any level of the hierarchy with any number of clusters can be chosen to fit the objectives of a given study. We used the level with 29 clusters spanning the conterminous U.S. because that was one level that we previously identified as optimal based on standard diagnostic metrics for hierarchical clusters (Costanza et al. 2017). From our previous work, we had a set of cluster "seeds" that had been assigned to one of the 29 empirical clusters (Costanza et al. 2017). We used the relative importance values from those seeds as the training set in *k*-nearest neighbor classification algorithm to assign the current set of FIA plots to those 29 clusters. In other words, we assigned each FIA plot in the current study to one of the 29 clusters based on the similarity of its relative importance values to those in the data set that was used to develop the empirical clusters.

The FIA and empirical classifications apply to all FIA plots in the conterminous U.S.; however, the key for USNVC macrogroups has only been developed for FIA plots in the eastern U.S. Therefore, we assigned all three of the alternative forest community type labels for each plot in the eastern U.S., and two labels for plots outside the eastern U.S. (Fig. 1). A set of 70,425 forested plots in the eastern U.S. and a total of 117,813 forested plots spanning the conterminous U.S. were used.

Correspondence among classifications

We compared the classifications to determine the degree of correspondence among classes. We completed four total two-way comparisons: three in the eastern U.S., among the USNVC, the FIA, and the empirical classifications, and one across the entire U.S., between the FIA and empirical classifications. In each case, we determined the relative similarities among classes in the pairwise comparison, both in terms of the frequency distributions of plots in the classes and in terms of the average species composition of classes. We first constructed two-way cross-tabulated contingency tables showing the frequency distributions of plots in the classifications, and input those tables into correspondence analysis. Correspondence analysis is similar to principal components analysis but is used for analyzing tables of count data (Greenacre 2013). It provides a graphical representation of the relationship between two categorical variables via an examination of how classes load onto

important correspondence axes. Specifically, it can point out the classes that are most strongly associated with one another.

For each class in each classification, we also calculated the average species composition in terms of relative species importance values across all FIA plots that were assigned to the class; that is, we identified the centroid of each class in multivariate species space. Importantly, although each classification scheme used slightly different information from the FIA database to label plots (for example, the FIA classes used relative stocking, etc.), the same relative importance value by species as described above was used to compare the classes here. For classes in each classification, the centroids of relative importance values were used to construct matrices of pairwise similarities between classes, using 1 - Bray-Curtis dissimilarity. We examined those similarities and visualized them using heat maps. For all pairwise comparisons, we omitted classes from each classification that had fewer than ten plots in them to simplify comparisons because initial tests of correspondence analysis indicated that those small classes had disproportionately large effects on the results.

Indicator and dominant species

Because the classification techniques differed in the extents to which they relied on dominant or characteristic tree species, we examined how well each classification could be characterized by indicator and dominant species. We conducted indicator species analysis and calculated species dominance values for all classes (Dufrene and Legendre 1997; Frieswyk et al. 2007). We used the same set of tree species as described above, with nonnative and the most rare species removed. Indicator species analysis finds both dominant and non-dominant species that have high specificity and fidelity to a given class using a permutation test to assign an indicator value and associated p-value to those species (Dufrene and Legendre 1997) and any set of alternative classifications can be compared based on those metrics (Dufrene and Legendre 1997). Classifications ranked highly by indicator species analysis will have at least one significant indicator species associated with each class, higher total numbers of indicator species overall, larger sums of all significant indicator values, and lower p-values on average. We used those metrics in our comparison of the classifications here. We standardized all indicator species metrics that were based on sums or counts by dividing by the number of classes in each classification.

We also calculated a species dominance index (SDI) for every species in each class, following Frieswyk et al. (2007). Dominant species are those that comprise a large proportion of the species assemblage in a plot, and thus the dominant species in a species assemblage are often responsible for much of the assemblages' ecological functioning and ecosystem services (Frieswyk et al. 2007; Hildebrand et al. 2008; Le Roux et al. 2014). The SDI ranges from 0 to 1, with 1 being a perfectly dominant species within a given class. A species with a high SDI value for a given class will tend to: (1) have high canopy cover on average across all plots within the class, (2) occur with few or no other species in plots within the class, and/or (3) only occur in a small number of plots within the class but tend to have high cover where it occurs (Frieswyk et al. 2007). The SDI is typically based on canopy cover, but here we use relative importance values (described above) in place of cover. We averaged the SDI values for the top most dominant species in every class within the classifications to determine which classifications tended to produce classes with more dominant species.

Predicting the classifications using environmental variables

We used the nonparametric random forest algorithm (Breiman 2001; Cutler et al. 2007) to determine how well each classification scheme could be predicted using environmental variables alone, and to examine the most important environmental predictors for each classification scheme. Each of the classification schemes was used as the response variable in a separate random forest model. The random forest algorithm fits many classification trees to a data set using a subset of predictors and a bootstrap sample of the data, then combines the results (Prasad et al. 2006; Cutler et al. 2007). In the random forest algorithm, the data not used to construct each classification tree, called the out-of-bag observations, are used to determine the error rate of each tree. Variable importance is based on the difference between each tree's error rate and a permutation of the tree without the given variable (Cutler et al. 2007, 2012).

As predictor variables for random forest models, we compiled spatial data on local and landscape soil characteristics, recent historical climate, landscape, and topography from ancillary sources (see Additional file 1 for a list of all variables used). We overlaid FIA plots on these spatial data layers to extract the values for plot locations. We focused on variables that have been shown to be important in other recent studies of tree distributions, vulnerability, and climate change (e.g., Iverson et al. 2008; Rogers et al. 2017). Soil variables were taken from the 10-m resolution Gridded Soil Survey Geographic Database (gSSURGO; Soil Survey Staff 2017a), where available. We used variables for the dominant component in the topmost horizon in gSSURGO data. For a few counties in the eastern U.S. and several places in the western

U.S., gSSURGO spatial data were not available, and we substituted data from the polygon-based State Survey Geographic Dataset (STATSGO2; Soil Survey Staff 2017b). We also included landscape Productivity Index and Drainage Index as measures of landscape-scale soil properties of each site (Schaetzl et al. 2009, 2012). We used average recent historical climate (1979–2015) data from the Multivariate Adapted Constructed Analogs (MACA) v2 METDATA data set (Abatzoglou 2013) to derive 19 bioclimatic variables (Hijmans et al. 2005), as well as annual averages of potential evapotranspiration and solar radiation. We also used the MACA data to calculate growing degree days (Sork et al. 2010), and annual summed moisture index (Koch and Coulston 2015). A set of seven topographic variables were derived from the 30-m resolution National Elevation Dataset (Gesch et al. 2002). Three landscape condition metrics were also used as predictors. One was ecological landscape condition, which assesses the level of human stressors surrounding a plot (Hak and Comer 2017). The other two were forest area density metrics, which measure the area and contagion of forests in a 65.6-ha window surrounding each FIA plot based on the 2011 National Land Cover Database (Riitters and Wickham 2012; Homer et al. 2015).

Because of the large number of trees grown, the random forest algorithm is not as sensitive as other algorithms to overfitting (Breiman 2001; Evans et al. 2011). However, to minimize the number of highly-correlated predictors and aid in interpretation of model results, we reduced slightly the number of variables within each set of predictors (climate, landscape, landscape soil, local soil, and topographic), using Pearson correlation statistics and principal component analysis. We included variables that were not highly correlated with others in the same set (absolute value of Pearson statistic >0.7), and which loaded highest on the principal components (those that together accounted for a cumulative variance of 75%). The result was a set of 22 variables that were input into random forest models (Table 2).

We ran random forest models for the FIA and empirical classifications across the U.S., as well as all three classifications in the east. We set the number of trees in each random forest model to 2000 and used all other defaults in the randomForest package in R (Liaw and Wiener 2002). We also used the default output from randomForest which assigns a predicted class to every data point based on the majority of votes across all trees in the forest. The random forest algorithm aims to minimize the overall error rate across all classes; thus, we report the average out-of-bag error rates. We also report Cohen's Kappa statistic of interrater reliability here for each classification (Cohen 1960).

All data manipulation and analysis was done using R 3.4.0 (R Core Team 2017) and the following contributed packages: ca (Nenadic and Greenacre 2007), class (Venables and Ripley 2002), factoextra (Kassambara and Mundt 2017), irr (Gamer et al. 2012), labdsv (Roberts 2016), randomForest (Liaw and Wiener 2002), RColorBrewer (Neuwirth 2014), raster (Hijmans 2016), and tidyverse (Wickham 2017), and vegan (Oksanen et al. 2016).

Table 2 Variables used as predictors in random forest models

Category	Variable	Description
Climate	Bio2	Mean diurnal range (Mean of monthly (max temp - min temp))
	Bio6	Minimum temperature of coldest month
	Bio8	Mean temperature of wettest quarter
	Bio15	Precipitation seasonality (coefficient of variation)
	Bio16	Precipitation of wettest quarter
	Bio18	Precipitation of warmest quarter
Landscape	Landsc. cond.	Landscape condition
	Prop. forest	Forest area density
Landscape soil	DI	Soil drainage index
	PI	Soil productivity index
Local soil	Bedrock depth	Depth to bedrock (cm
	Pct. clay	Percent clay (< 0.002 mm size)
	Pct. org. Matt.	Organic matter content (% by weight)
	Pct. sand	Percent sand (0.05–2.0 mm size)
	Pct. silt	Percent silt (0.02–0.05 mm size)
	pH	Soil pH
	Sieve 10	Percent soil passing sieve No. 10 (coarse)
Topographic	East	Easting: sin(aspect)
	Elevation	Elevation
	North	Northing: cos(aspect)
	Slope	Slope
	TPI	Topographic Position Index

See Additional file 1 for a list of all variables considered as well as data sources for all

Results

For the empirical classification, 17 of the 29 classes occurred in the set of eastern U.S. FIA plots. For the FIA classification, 17 of 28 classes occurred in the eastern U.S. The eastern plots contained 25 USNVC macrogroups. See Additional file 2 for lists of all classes.

Correspondence among classifications

We completed four two-way comparisons of plot frequencies and species composition: three in the eastern U.S. and one nationwide. We only included classes with greater than ten plots for this portion of the analysis, which resulted in eliminating 11 plots from the national

data set and 21 plots from the eastern U.S. data set. In each of the four comparisons, some classes had relatively good analogs in terms of species composition and cross-tabulated frequencies of plots and some did not. Here, we point out some of the classes with high and low correspondence in each case using three summary methods: biplots from the first three dimensions of each correspondence analysis (Figs. 2 and 3), heat maps of species similarity (Fig. 4), and the full contingency tables of cross-tabulated plot frequencies (Additional file 3).

The comparison between the empirical and USNVC classifications in the eastern U.S. shows some broad similarities, but few one-to-one analogs in terms of plot frequencies. Biplots from correspondence analysis indicate a few cases in which one empirical class was associated with several USNVC classes (Fig. 2a). For example, the balsam fir-quaking aspen empirical class was associated with five USNVC classes. In this case, all the 1082 plots in the North American Boreal Conifer Poor Swamp USNVC class were classified in the balsam fir-quaking aspen empirical class, but that empirical class also included plots in several other USNVC classes (Additional file 3). These differences likely occurred because the same tree species exist in both upland and wetland boreal and sub-boreal forests. The slash pine-longleaf pine empirical class was associated with the Pond-Cypress Basin Swamp and Longleaf Pine Woodland USNVC classes in terms of plot frequencies, again crossing upland and wetland USNVC classes. The loblolly pine-sweetgum empirical class was associated with

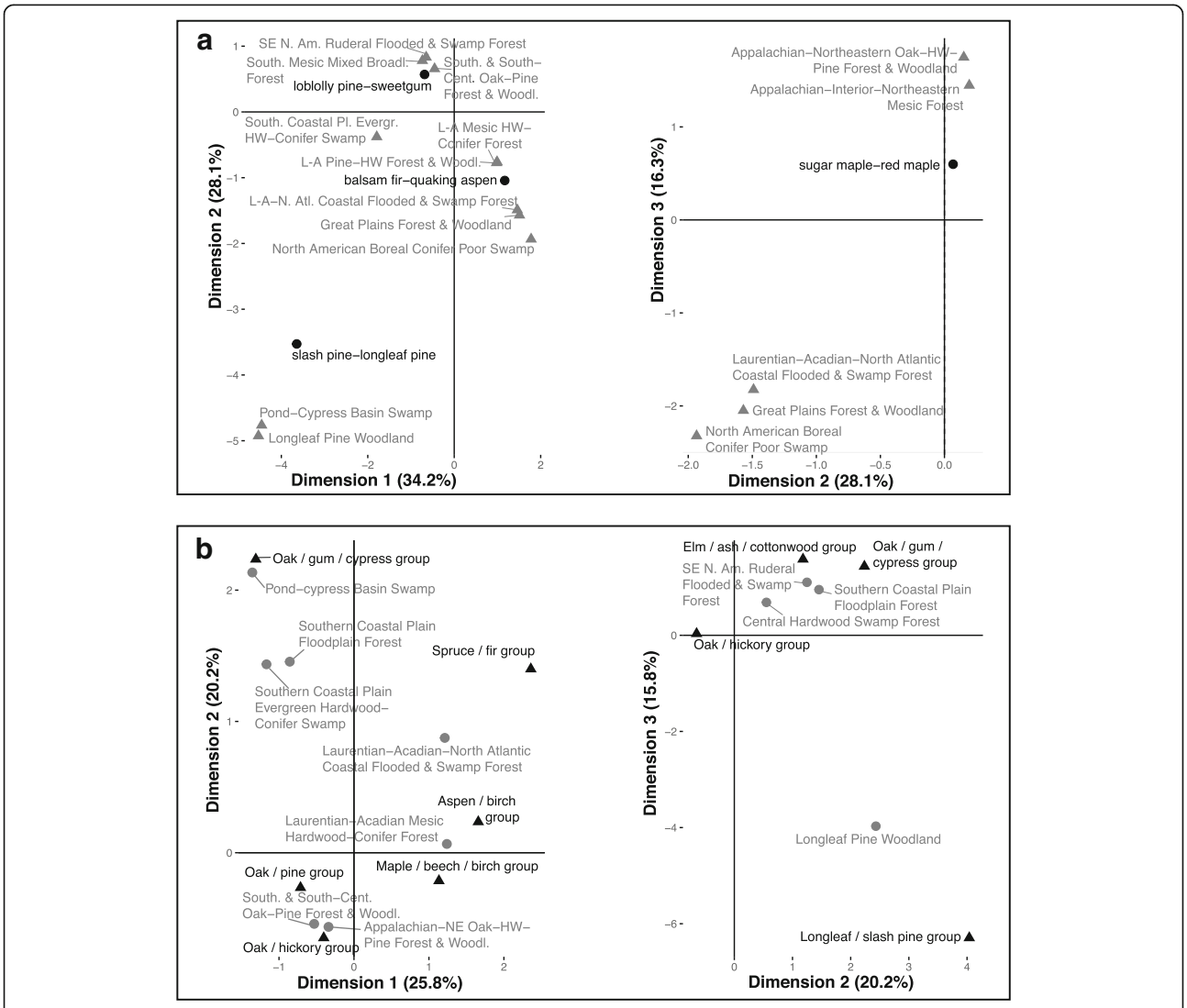

Fig. 2 Biplots from dimensions 1 and 2 (left) and 2 and 3 (right) from correspondence analysis: (**a**) Empirical (black) and USNVC classes (gray); (**b**) FIA (black) and USNVC classes (gray). Only the classes that loaded the highest on these dimensions are plotted in each case. Row standardization was used in these plots, and numbers in parentheses indicate the percentages of inertia explained by each dimension. Class names have been abbreviated to fit in some cases, and the full names of all classes can be found in Additional file 2

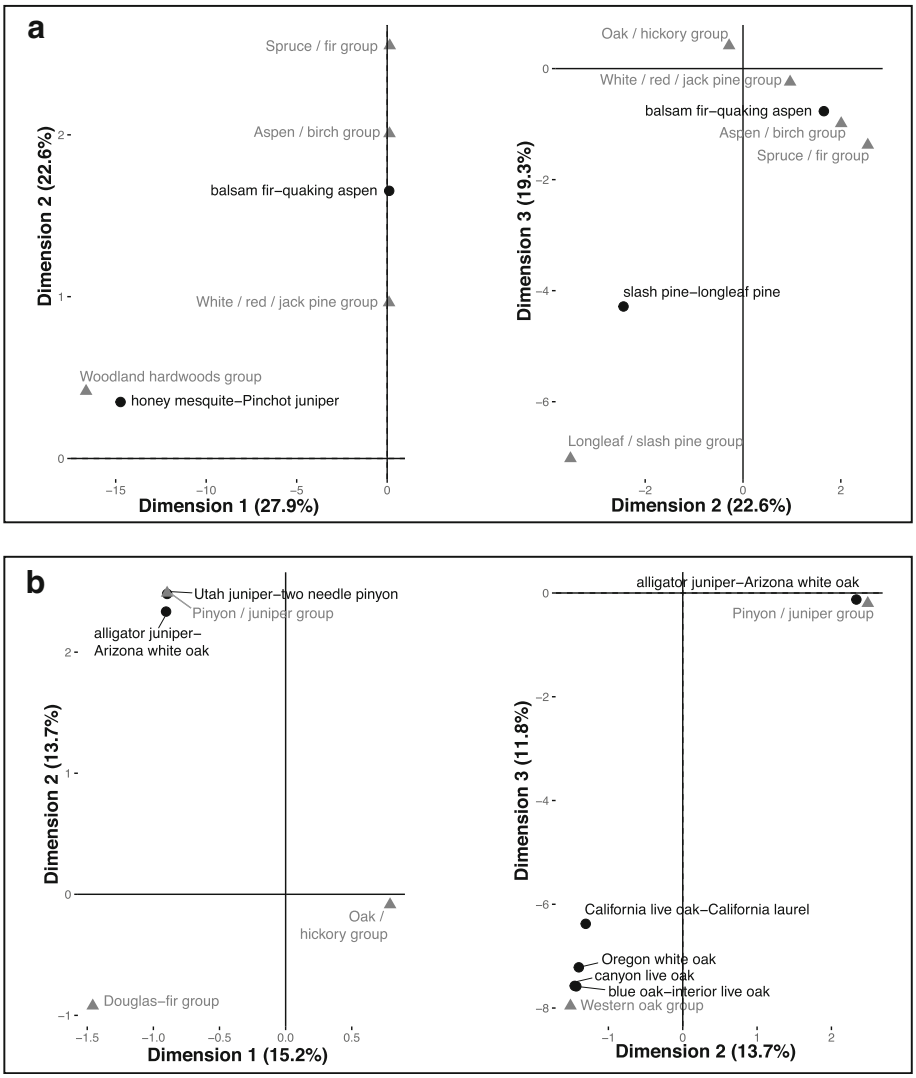

Fig. 3 Biplots from dimensions 1 and 2 (left) and 2 and 3 (right) from correspondence analysis: (**a**) Empirical (black) and FIA classes (gray) in the eastern U.S.; (**b**) Empirical (black) and FIA classes (gray) across the U.S. Only the classes that loaded the highest on these dimensions are plotted in each case. Row standardization was used in these plots, and numbers in parentheses indicate the percentages of inertia explained by each dimension. Class names have been abbreviated to fit in some cases, and the full names of all classes can be found in Additional file 2

the Southern & South-Central Oak-Pine Forest & Woodland, the Southeastern North American Ruderal Flooded & Swamp Forest, and the Southern Mesic Mixed Broadleaf Forest USNVC classes, thus crossing, upland, wetland and ruderal forest classes. In terms of average species composition, the highest similarity among pairs of classes in the empirical and USNVC classifications was between the slash pine-longleaf pine empirical class and the Longleaf Pine Woodland USNVC class (Fig. 4a). Despite having low correspondence in terms of plot frequencies, the loblolly pine-sweetgum empirical and Southern Mesic Mixed Broadleaf Forest USNVC classes were relatively similar in species composition (Fig. 4a). The sugar maple-red maple empirical class had a relatively high average similarity and was associated in terms of plot frequencies with two USNVC classes: the Appalachian-Interior-Northeastern Mesic Forest and the Appalachian-Northeastern Oak-Hardwood-Pine Forest & Woodland (Figs. 2a and 4a). Several empirical and USNVC classes had low similarities in species composition to all classes in the other classification. Those included the honey mesquite-Pinchot juniper and chokecherry-Pacific dogwood empirical classes, and the North American Boreal Conifer Poor Swamp USNVC class, likely because those empirical classes were largely determined by western species. Interestingly, in contrast to the high correspondence in terms of plot frequencies, the similarity in species composition between the balsam fir-quaking aspen empirical and the North American Boreal Conifer Poor Swamp USNVC class was low.

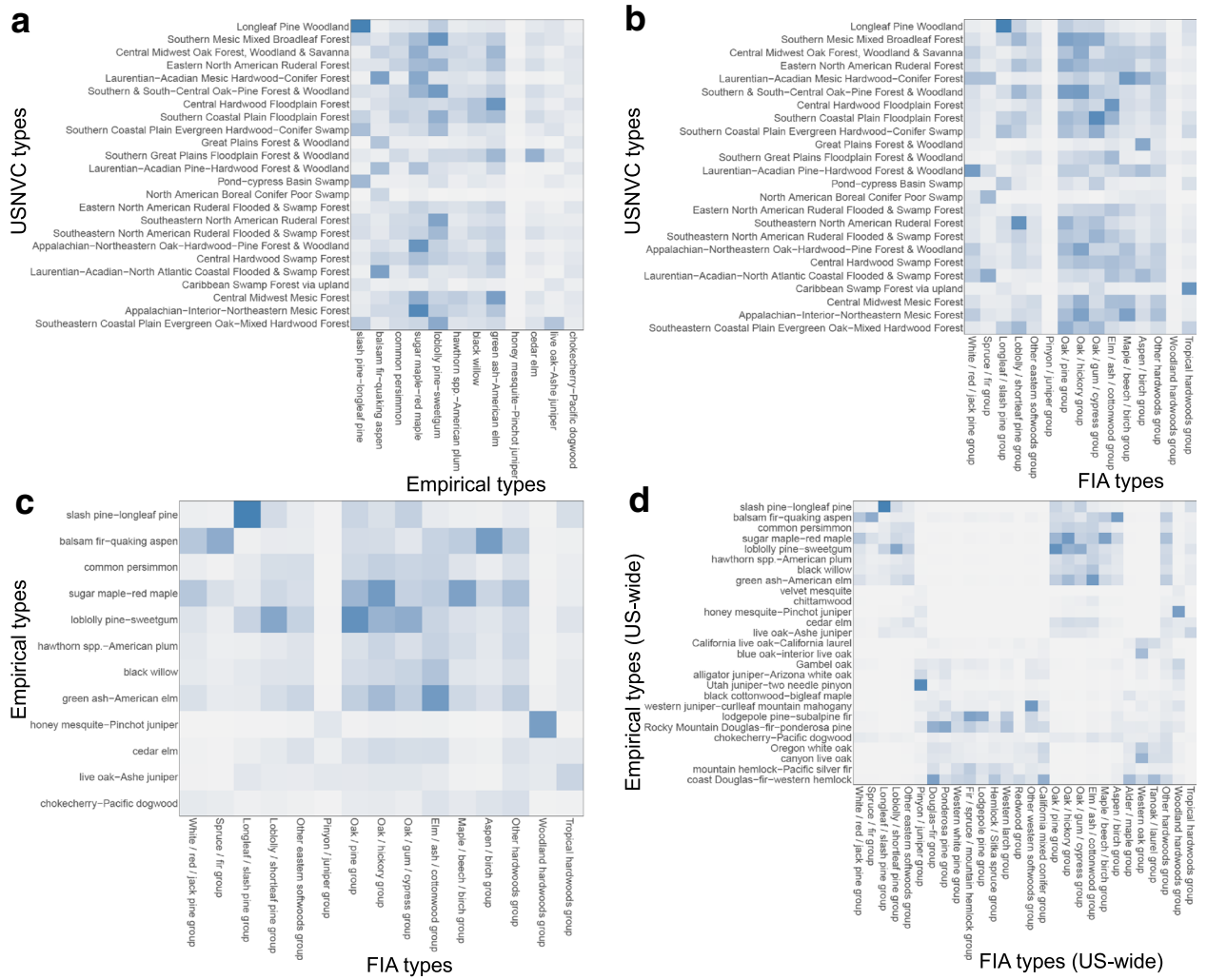

Fig. 4 Heat maps showing similarity in average species composition from each pairwise comparison of classifications: (**a**) empirical and USNVC classes; (**b**) FIA and USNVC classes; (**c**) FIA and empirical classes in the eastern U.S.; (**d**) FIA and empirical classes across the U.S. Similarities range from 0 (low) to 1 (high), with darker blues representing higher values. In each case, classes are ordered by their codes in Additional file 2

The comparison between the FIA and USNVC classifications in the eastern U.S. also shows some broad similarities, but few one-to-one analogs in terms of plot frequencies. Biplots from correspondence analysis show a few cases in which classes from the two classifications were associated on a one-to-many or many-to-many basis (Fig. 2b). The oak/pine and oak/hickory FIA classes were associated with two USNVC classes: the Southern & South-Central Oak-Pine Forest & Woodland, and the Appalachian-Northeastern Oak-Hardwood-Pine Forest & Woodland. Here the USNVC separates two classes based on overall floristic and biogeographic differences, whereas the two FIA classes emphasize the mixed conifer-deciduous versus pure deciduous overstories. The aspen/birch and spruce/fir FIA classes were also both associated with two USNVC classes: the Laurentian-Acadian-North Atlantic Coastal Flooded & Swamp Forest, and the Laurentian-Acadian Mesic Hardwood-Conifer Forest, thus crossing upland and wetland USNVC classes, but distinguishing deciduous from conifer. There were two other strong associations between one FIA class and three USNVC classes. The only one-to-one association in terms of plot frequencies was between the longleaf/slash pine FIA class and the Longleaf Pine Woodland USNVC class (Fig. 2b, Additional file 3). Average species similarities among classes showed the highest pairwise similarity was between the Longleaf/slash pine FIA class and the Longleaf Pine Woodland USNVC class (Fig. 4b). The Maple/beech/birch FIA class had relatively high similarity on average with two USNVC classes: Laurentian-Acadian Mesic Hardwood-Conifer Forest, and Appalachian-Interior-Northeastern Mesic Forest. Other high average similarities in species composition

occurred between the Southern Coastal Plain Floodplain Forest USNVC and Oak/gum/cypress FIA classes, and between the Laurentian-Acadian Pine-Hardwood Forest & Woodland USNVC and White/red/jack pine FIA classes. Two FIA classes had low similarities with all USNVC classes, namely the Pinyon/juniper and woodland hardwoods classes. These classes were largely based on western tree species. Other FIA classes had moderate similarity with several USNVC classes. For example, the loblolly/shortleaf pine group was moderately similar in species composition to four USNVC classes including Southern Mesic Mixed Broadleaf Forest. All the USNVC classes were at least moderately similar to a minimum of one FIA class.

In the eastern U.S., the comparison between the empirical and FIA classifications shows both broad similarity and one-to-one analogs for some classes. The honey mesquite-Pinchot juniper empirical class was associated with the woodland hardwoods FIA class, indicating that those two classes were good analogs in terms of plot frequencies (Fig. 3a, Additional file 3). The balsam fir-quaking aspen empirical class was associated with three FIA classes – the spruce/fir, aspen/birch and white/red/jack pine – suggesting that the empirical class is more heterogeneous than the FIA class. The slash pine-longleaf pine empirical class was associated with the longleaf/slash pine FIA class. When comparing the FIA and empirical classes in the eastern U.S. in terms of species composition, similar analogs stand out. Three pairs of classes had the highest similarities: (1) the balsam fir-quaking aspen empirical and aspen/birch FIA classes; (2) the slash pine-longleaf pine empirical and longleaf/slash pine FIA classes; and (3) the honey mesquite-Pinchot juniper empirical and woodland hardwoods FIA classes (Fig. 4c). Six empirical classes and five FIA classes in the eastern U.S. had low similarity in species composition on average with all classes in the other classification. Examples of those classes are the black willow and cedar elm empirical classes, and the other eastern softwoods and other hardwoods FIA classes. The largest empirical classes in terms of numbers of plots had relatively high similarities in species composition with more than one FIA class. For example, the loblolly pine-sweetgum empirical class was most similar in species composition to the oak/pine FIA class, but it also had relatively high similarities with other FIA classes, including the loblolly/shortleaf pine and oak/gum/cypress FIA groups.

When the empirical and FIA classifications were compared for FIA plots across the U.S., there were necessarily some similarities to the eastern U.S. comparison because eastern U.S. plots were included in the full set of U.S. plots. However, there were also several important differences, and we focus on those differences here. There were only two sets of classes that were associated with one another in terms of the first three dimensions of correspondence analysis, and all included classes that did not occur in the east (Fig. 3b). The Utah juniper-two needle pinyon and alligator juniper-Arizona white oak empirical classes were associated with the pinyon/juniper FIA class, and three empirical groups containing western oak species were associated with the western oak FIA class. In terms of average species composition, there were many pairs of classes with low to moderate similarity, and only a few cases of relatively high similarity (Fig. 4d). The highest similarities overall were those described above for the eastern U.S., but the Utah juniper-two needle pinyon empirical class also had high similarity in average species composition to the pinyon/juniper FIA group.

Indicator and dominant species

When considering the FIA and empirical classifications across the U.S., optimal values of metrics from indicator species analysis were split between the two (Table 3). Both had at least one significant indicator species for every class, leaving zero classes without an indicator species. The FIA classification had lower average *p*-values and more significant indicator species overall, while the empirical classification had a higher sum of all indicator values. When considering just the inventory plots in the eastern U.S., the USNVC classification had the best values of three of

Table 3 Results from indicator species analysis, species dominance, and random forest models for all classifications

Region	Classification	Indicator species metrics				Species dominance	Random forest metrics	
		Sum of indicator vals.[a]	Avg. of *p*-values	Num. significant indicator vals.[a]	Prop. classes without significant indicator vals.	Mean of largest SDI (std. dev.)	Kappa	Avg. out-of-bag error
U.S.	FIA	1.01	**0.01**	**6.00**	**0**	0.47 (0.17)	0.59	35.74%
	Empirical	**1.21**	0.05	3.76	**0**	**0.61 (0.21)**	**0.74**	**22.80%**
East	USNVC	0.60	**0.07**	**3.32**	**0**	0.36 (0.16)	0.55	40.25%
	FIA	0.87	0.08	3.24	0.12	0.42 (0.17)	0.48	38.55%
	Empirical	**0.88**	0.19	1.35	0.24	**0.59 (0.24)**	**0.64**	**24.90%**

[a]The sum of indicator values and the number of significant indicator values reported here have been standardized by the number of classes in each classificationBold indicates the best value for each metric in each region

the four indicator species metrics, whereas for the fourth (the sum of indicator values) the empirical and FIA classifications had the best value. The species dominance analysis shows that on average, the species with the highest SDI values per class were found in the empirical classifications in the U.S. as well as in the east. However, the standard deviations of SDI indicate large overlap among the distributions across all classifications.

Predicting the classifications using environmental variables

The average out-of-bag error rates and Kappa statistics for all classifications show that the set of environmental variables included here best predicted the empirical classification, either when considering plots across the conterminous U.S. or the subset of eastern U.S. plots (Table 3; full confusion matrices are in Additional file 5). The empirical classification across the U.S. had the lowest error rate and highest Kappa value, and the empirical classification applied to eastern plots had the second lowest and highest values, respectively. Metrics for the FIA classes across the U.S. were third best overall, but metrics for the USNVC were better than those of the FIA classes for eastern plots.

A somewhat similar set of variables tended to be among the most important for predicting all classifications (Fig. 5). The mean temperature of the coldest month, slope, and the proportion of forest in the landscape were the top three most important for predicting the empirical and FIA classifications, whether for all plots or just eastern plots (Fig. 5a-b, d-e). Mean temperature of the coldest month, precipitation of the warmest quarter, and precipitation seasonality were the most important variables in predicting the USNVC classification (Fig. 5c). Soil and other climate variables tended to be next most important in all cases. Landscape soil variables and aspect-related topographic variables tended to be least important for predicting all classifications. For the empirical and FIA classifications, the order of variable importance changed slightly when considering all plots versus eastern U.S. plots, but the most and least important variables stayed relatively constant.

Discussion

No classification of forest community types will be suitable for all research questions. To shed light on the best uses of alternative classifications, we compared three species-based forest type classifications, ranging from supervised to unsupervised, using a national set of forest inventory data. Across the classifications, some notable patterns emerged from the comparisons of plot frequencies and species similarity. Those patterns can be interpreted in view of the classification methods and objectives for which each classification has been developed.

Despite the different methods and objectives for each classification, the high correspondence among the longleaf pine classes in each provided one notable case of alignment among the classifications. All the classifications included one longleaf pine class (the slash pine-longleaf pine class in the empirical classification, longleaf/slash pine in FIA, and Longleaf Pine Woodland in USNVC). Pairwise comparisons showed that many plots were classified into these classes, as well as high similarities in average species composition for those classes. In addition, the indicator species of these classes were nearly the same, with longleaf pine, slash pine, and turkey oak as indicators in all cases, and two other species as indicators of the empirical class (Additional file 4). This correspondence is likely because of the relatively low tree diversity and geographically restricted range of longleaf pine trees, coupled with the distinct nature of longleaf pine communities. Longleaf pine forest communities are currently restricted to places in the southeastern U.S. that have been frequently burned (Frost 2006). Longleaf pine trees tend to be dominant in those communities, along with slash pines toward the southern part of the community's range (Burns and Honkala, 1990). Longleaf trees themselves are rarely found in other forest communities either as dominant species or co-occurring with other tree species. Thus, the three classifications, despite their different methods, were each able to identify longleaf pine forest community types.

In contrast, there were many instances in which forest classes that did not correspond well across classifications. As an important example, there was less correspondence than we would have expected among loblolly pine classes. Loblolly pine forests are one of the most ubiquitous in the southeastern U.S. Therefore, examining the differences among classifications in the ways those forests are will be important for informing potential uses of the classifications in the region. Loblolly pine trees currently occur across a wide range of environmental and human disturbance conditions, from mesic bottomland sites to drier upland sites across the coastal plain and Piedmont regions of the southeastern U.S., and in forests established on abandoned agricultural fields (Burns and Honkala, 1990). This likely explains why the loblolly pine-sweetgum empirical class was highly associated in terms of plot frequencies with three USNVC classes that have a range of disturbance histories and geographic affinities, including the Southeastern North American Ruderal Forest USNVC class, which had loblolly pine as an indicator species (Additional file 4). In contrast, the empirical and FIA classifications each do include a class with loblolly pine in the name. While the two classes corresponded relatively well in terms of plot frequencies (Additional file 3), and

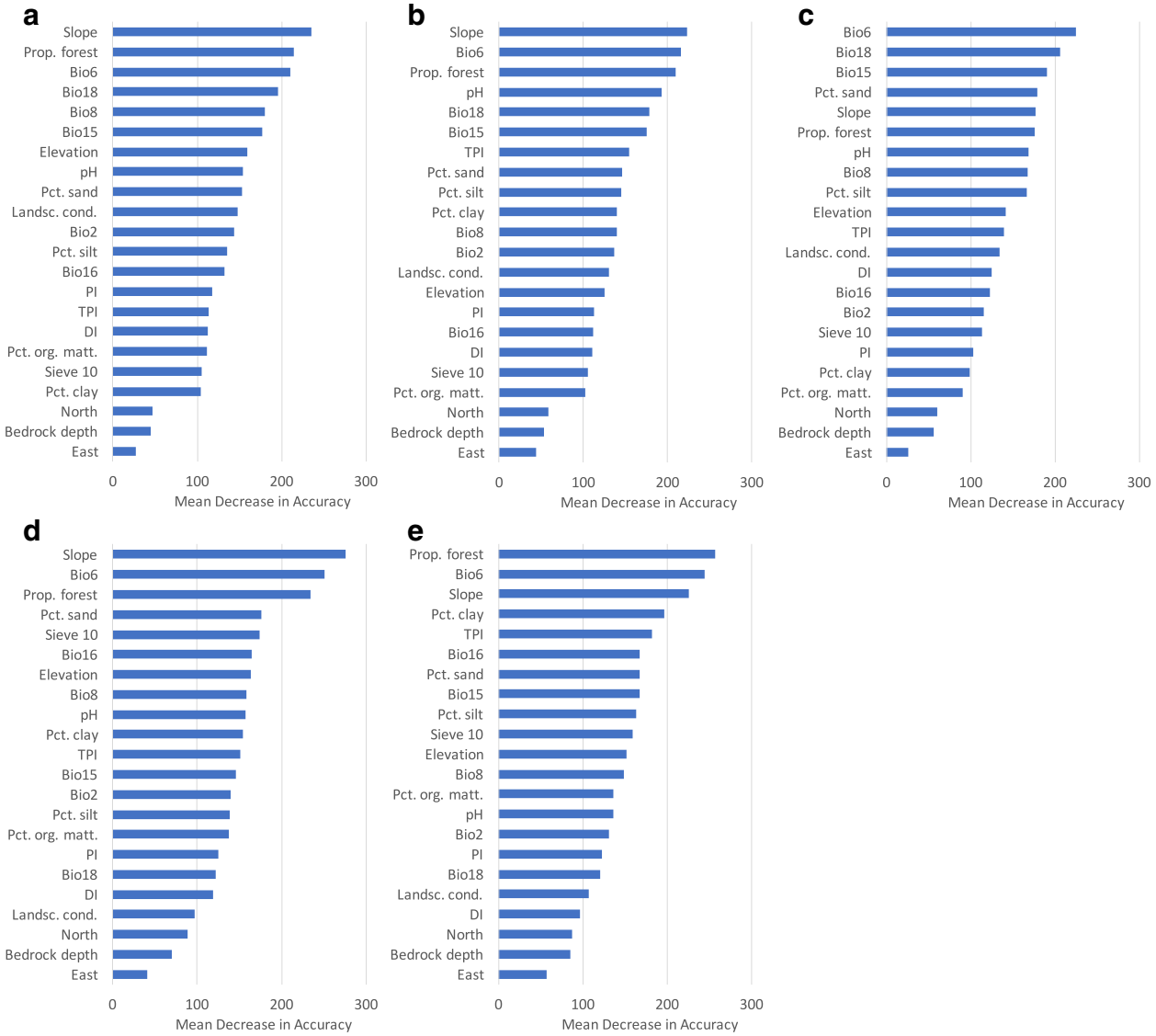

Fig. 5 Variable importance plots for each random forest model: (**a**) FIA classes in the eastern U.S.; (**b**) empirical classes in the eastern U.S.; (**c**) USNVC classes; (**d**) FIA classes across the U.S.; (**e**) empirical classes across the U.S. See Table 2 for a description of each variable

moderately well in terms of species composition, the loblolly pine-sweetgum empirical class was most similar in species composition to the oak/pine FIA class. In terms of indicator species, the empirical loblolly pine-sweetgum class had ten significant indicator species, including loblolly pine, while the FIA class had six (Additional file 4). Therefore, our results suggest that the empirical class is more inclusive than the FIA class. Because it is based on dominant species, the FIA class is likely capturing places where loblolly pines are dominant or co-dominant, while the empirical class also includes places where loblolly occurs with other species.

Similarly, in other cases, a one-to-many correspondence among classes signifies a broader concept in one classification scheme than in others. For example, in the eastern U.S., the balsam fir-quaking aspen empirical class was highly associated in terms of plot frequencies with three FIA classes and four USNVC classes, indicating that the empirical classification included a broader definition of that forest type than the other two. Outside the eastern U.S., the western oak FIA class was highly associated with four empirical classes that all included oak species, underscoring the fact that the FIA class included forest types with any number of oak species.

Results from random forest models highlight the degree to which each classification overall corresponds to the selected environmental conditions. Kappa statistics and error rates show that the empirical classification, whether across the U.S. or in the eastern portion of the country, was best predicted by environmental variables.

Because the empirical classification did not explicitly incorporate environmental conditions, and because the USNVC classification is based in part on environmental conditions and ecological settings, it is somewhat surprising that prediction of the empirical classification was best. However, there are at least two potential explanations for this. First, the USNVC macrogroups are based on tree, non-tree, ecological and biogeographic characteristics, and disturbance history (FGDC 2008). The set of environmental variables used here may not address all those factors equally well. For example, several USNVC classes are defined largely by their wetland status (e.g. swamp forest classes) or disturbance status (e.g. ruderal forest types). Thus, variables related to wetlands and land-use history are likely most important for predicting those forest types. The landscape condition variable (Hak and Comer 2017) provides the best indicator of land-use effects but it was only moderately important in the USNVC random forest model. Indeed, the misclassification rates of many of the ruderal and wetland classes in the USNVC model are high (Additional file 5), supporting the idea that those classes are not well predicted by the model.

A second reason for the relatively strong performance of the random forest model for predicting the empirical classification is the fact that the empirical classification was developed based on similarities among plots in recent forest inventory data (Costanza et al. 2017). Conversely, the FIA classes are largely based on a well-recognized set of forest types (Eyre 1980) that were developed a few decades ago but they may not adequately represent all of the forest communities that currently exist in the database. Considering this, it is not surprising that recent environmental data would predict the empirical classes better than the FIA classes.

In addition, the random forest model results point to the multiple scales at which the environment influences forest communities. In each of the models, climate and topographic variables were all among the top five most important variables, and soil and landscape variables were among the top five in most cases. All of those variables affect the ecological processes that influence species composition in ecological communities at a variety of scales, from climate at the broadest extents to soil variables locally (Shmida and Wilson 1985; McGill 2010). Random forest results overall suggest that the classifications can be used to examine forest community-environment relationships across a range of scales. However, for the USNVC classification, in contrast with the other classifications, the top three most important variables for predicting the USNVC were all regional climate variables, and specifically, the mean temperature of the coldest month, the precipitation of the warmest quarter, and precipitation seasonality were the most important. This suggests that relatively broad climate drivers may be relatively more important for driving the distributions of the USNVC macrogroups than the other classifications. While we lacked data on some fine-scale drivers like microclimate and hydrologic processes for the full set of plots used here, further work could examine how those types of drivers differ among these classifications using a smaller subset of plots for which those data are available. However, the relationships between all of the classifications and environmental variables at a range of scales suggests that these classifications could be useful for informing research related to the emerging field of macrosystems ecology, which focuses on understanding ecological processes and patterns at broad extents, while emphasizing hierarchies and multiple scales (Heffernan et al. 2014; Rose et al. 2016).

When taken together, the results here highlight the similarities and differences among these classifications and provide critical information to potential users. The empirical classification highlights assemblages of tree species, including dominant species, and corresponds well with broad environmental conditions. Therefore, when the aim is to relate forest community types across broad extents to climate or environmental characteristics, the empirical classification may be best. Indeed, the empirical classification was specifically developed for studies related to global change across the U.S. (Costanza et al. 2017). Because it is data driven, the empirical classification can both classify forest plots into types independently over time periods, and compare the response of those types to changing environmental conditions over those same time periods. In the FIA classification, classes are characterized well by indicator species, especially when viewed across the entire U.S. However, in terms of indicator species metrics and random forest metrics, both the empirical and FIA classifications saw reduced performance when just plots in the eastern U.S. were considered, suggesting that they may not be as effective in accounting for the full diversity of tree species patterns when subset to smaller regions. In addition, both the empirical and FIA classifications included some large, broad assemblages that did not have analogs in the other classifications. Lower levels of the empirical classification hierarchy as well as FIA forest types (not type groups) may show higher correspondence. A comparison of the two, along with a lower level of the USNVC hierarchy, would shed light on whether and how those finer species assemblages match.

The USNVC was represented better by indicator species than the other two classifications when considering eastern plots only. This suggests that the strength of this classification is in identifying distinct assemblages of species in the eastern U.S. The USNVC classes were not

as well predicted as the empirical classification by the environmental predictors used here. However, this is likely because the multi-factor approach used by the USNVC based on species composition, disturbance history, and ecological settings does not correspond to a single set of drivers across the region. Users of the USNVC classification who want to relate macrogroups to environmental variables could explore environmental predictors for subsets of the classes, or identify alternative predictors, such as those that do capture land-use and disturbance history, and wetland conditions.

As the availability of large plot databases and the computing power required to work with them increases, future work should compare alternative classifications of these data sets. For example, the Global Forest Biodiversity Initiative (GFBI, http://www.gfbinitiative.org) has synthesized data from nearly 800,000 permanent forest vegetation plots in 44 countries for analysis of forest community ecological function and structure (Liang et al. 2016). Classifying this global plot data set into species-based forest community types set could illuminate the worldwide geographic patterns of forest communities and their environmental drivers, and could be useful for global change studies. Indeed, development of global species-based vegetation classification hierarchies based on the same principles and design as the USNVC are in progress (Faber-Langendoen et al. 2014), and could be applied to global forest plot data. Comparison with one or more unsupervised classifications of forest community types would shed light on whether the differences among classification approaches seen in this paper apply globally, and better inform users about the most appropriate forest community type classifications for global analysis.

A full assessment of the utility of these classifications for studies of global change drivers and forest communities was beyond the scope of this study. However, as forest ecosystems respond to global change drivers like climate change, some species may decline while others may expand their ranges (Iverson et al. 2008; Zhu et al. 2014; Fei et al. 2017). One key issue to be addressed in classification is how to handle expected changes to the forest classes that comprise a forest community classification. This is particularly important if the goal is to use the classifications to monitor forest change and detect new forest community types. Throughout this paper, we have contrasted unsupervised and supervised classifications as a starting point for how to address forest classification for large forest inventory data sets. In that regard, the semi-supervised classification method proposed by Tichý et al. (2014) has high potential for allowing classes to be compared and adjusted based on changes in forest conditions over time. In this method, a classification first formally creates a set of existing classes in a supervised or unsupervised mode and simultaneously identifies new units among unassigned sites in an unsupervised mode. This semi-supervised method has potential for adding classes to any of the classifications presented here. We note that the USNVC is open to redefining or proposing new forest types over time, through a dynamic peer review process (FGDC 2008). In addition, the empirical classification will be used as the basis for modeling change in future forest conditions under global change scenarios for the USDA Forest Service's 2020 Resources Planning Act Assessment (https://www.fs.fed.us/research/rpa/). Therefore, developing methods to accommodate changes in forest community types within these classifications will be critical for ensuring their utility in global change studies.

Conclusion

Understanding the differences and similarities among alternative approaches to classifying forest community types across broad extents is important, especially in the face of global change drivers. The three classifications we examined here show the strongest similarity for forest types that are relatively restricted in range and vary little in tree species composition. But the classifications differ in other cases, such as for wide-ranging, general forest types. Those general patterns point to the differences among classifications in terms of their approach and objectives. The empirical classification was predicted best by environmental variables, and, along with the FIA classification, was highly related to a set of environmental variables measured at a range of scales. Coarse-grain climate variables were the most important for predicting the USNVC classification, but other variables related to the local environment were moderately important for that classification. Therefore, these classifications will all have utility for studies of multi-scale environmental correlates of forest communities, including under global change scenarios.

Additional files

Additional file 1: All variables considered in random forest analysis. (XLSX 13 kb)

Additional file 2: List of all classes in all classifications used in this analysis. (XLSX 12 kb)

Additional file 3: Full contingency tables of cross-tabulated plot frequencies used in correspondence analysis. (XLSX 20 kb)

Additional file 4: Confusion matrices from random forest analysis. (XLSX 27 kb)

Additional file 5: List of indicator species for all classifications. (XLSX 27 kb)

Acknowledgments

We thank K. Riitters for spatial data extraction and the forest area density and contagion data, J. Hak for landscape condition data, R. Li for assistance with querying the FIA database, and K. Nimerfro for assignment of macrogroups to FIA plots. Two anonymous referees provided helpful comments that strengthened the manuscript.

Funding
Funding for this work came from the USDA Forest Service Resources Planning Act Assessment, via an agreement with North Carolina State University.

Authors' contributions
All authors conceived of the idea. JKC performed analysis and wrote the manuscript. DFL assisted with data analysis and edited the manuscript. JWC and DNW secured funding and edited the manuscript. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Forestry and Environmental Resources, North Carolina State University, 3041 Cornwallis Rd., Research Triangle Park, Raleigh, NC 27709, USA. [2]NatureServe, 4600 N. Fairfax Dr., 7th Floor, Arlington, VA 22203, USA. [3]Southern Research Station, USDA Forest Service, Blacksburg, VA, USA. [4]Southern Research Station, USDA Forest Service, Raleigh, NC, USA.

Received: 1 September 2017 Accepted: 14 December 2017
Published online: 06 February 2018

References

Abatzoglou JT (2013) Development of gridded surface meteorological data for ecological applications and modelling. Int J Climatol 33(1):121–131. https://doi.org/10.1002/joc.3413

Arner SL, Woudenberg S, Waters S, Vissage J, Maclean C, Thompson M, Hansen M (2003) National Algorithms for Determining Stocking Class, Stand Size Class, and Forest Type for Forest Inventory and Analysis Plots. http://www.fia.fs.fed.us/library/field-guides-methods-proc/docs/National%20algorithms.doc . Accessed 15 Jan 2015

Bechtold WA, Patterson PL (2005) The enhanced Forest inventory and analysis program — National Sampling Design and estimation procedures. USDA Gen Tech Rep SRS 80:85

Belote RT, Dietz MS, Jenkins CN, McKinley PS, Irwin GH, Fullman TJ, Leppi JC, Aplet GH (2017) Wild, connected, and diverse: building a more resilient system of protected areas. Ecol Appl 27(4):1050–1056

Brandt LA, Butler PR, Handler SD, Janowiak MK, Shannon PD, Swanston CW (2017) Integrating science and management to assess Forest ecosystem vulnerability to climate change. J Forest 115(3):212–221. https://doi.org/10.5849/jof.15-147

Breiman L (2001) Random forests. Mach Learn 45(1):5–32. https://doi.org/10.1023/A:1010933404324

Burns RM, Honkala BH (1990) Silvics of North America: volume 1. Conifers. Agriculture Handbook 654, USDA Forest Service, Washington, p 877

Caldwell PV, Miniat CF, Elliott KJ, Swank WT, Brantley ST, Laseter SH (2016) Declining water yield from forested mountain watersheds in response to climate change and forest mesophication. Glob Chang Biol 22(9):2997–3012. https://doi.org/10.1111/gcb.13309

Cohen J (1960) A coefficient of agreement for norminal scales. Educ Psychol Meas XX:37–46

Costanza JK, Coulston JW, Wear DN (2017) An empirical, hierarchical typology of tree species assemblages for assessing forest dynamics under global change scenarios. PLoS One 12(9):e0184062. https://doi.org/10.1371/ journal.pone.0184062

Cutler A, Cutler DR, Stevens JR (2012) Random forests. In: Zhang C, Ma Y (eds) Ensemble machine learning: methods and applications. Springer, New York, pp 157–175

Cutler DR, Edwards TC, Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ (2007) Random forests for classification in ecology. Ecology 88(1):2783–2792. https://doi.org/10.1890/07-0539.1

Dengler J, Jansen F, Glöckler F, Peet RK, Cáceres MD, Chytrý M, Ewald J, Oldeland J, Lopez-Gonzalez G, Finckh M, Mucina L, Rodwell JS, Schaminée JHJ, Spencer N (2011) The global index of vegetation-plot databases (GIVD): a new resource for vegetation science. J Veg Sci 22(4):582–597. https://doi.org/10.1111/j.1654-1103.2011.01265.x

DeSantis RD, Moser WK, Li RH, Wear DN, Miles PD (2013) Modeling the effects of emerald ash borer on forest composition in the Midwest and Northeast United States. USDA For Serv Gen Tech Rep NRS-112, North Res Station. pp 1-28

Dufrene M, Legendre P (1997) Species Assamblages and indicator species: the need for a flexible Asymetrical approach. Ecol Monogr 67:345–366

Duveneck MJ, Thompson JR, Wilson BT (2015) An imputed forest composition map for New England screened by species range boundaries. Forest Ecol Manag 347:107–115

Evans JS, Murphy MA, Holden ZA, Cushman SA (2011) Modeling species distribution and change using random Forest. In: Drew CA, Wiersma YF, Huettmann F (eds) Predictive species and habitat modeling in landscape ecology: concepts and applications. Springer Science+Business Media, New York, pp 139–159

Eyre FH (1980) Forest cover types of the United States and Canada. Society of American Foresters, Washington

Faber-Langendoen D, Keeler-Wolf T, Meidinger D, Tart D, Hoagland B, Josse C, Navarro G, Ponomarenko S, Saucier JP, Weakley A, Comer P (2014) EcoVeg: a new approach to vegetation description and classification. Ecol Monogr 84:533–561. https://doi.org/10.1890/13-2334.1

Fei SL, Desprez JM, Potter KM, Jo I, Knott JA, Oswalt CM (2017) Divergence of species responses to climate change. Sci Adv 3(5):e1603055. https://doi.org/10.1126/sciadv.1603055

FGDC (2008) National Vegetation Classification Standard, version 2 FGDC-STD-005-2008. Federal Geographic Data Committee. Reston, Virginia, USA, pp 55 (+ Appendices). Available at: https://www.fgdc.gov/standards/projects/FGDC-standards-projects/vegetation/NVCS_V2_FINAL_2008-02.pdf Accessed 19 Dec 2017

Fralish JS (2004) The keystone role of oak and hickory in the central hardwood forest. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station. Gen Tech Rep SRS-73:78–87

Franklin J, Serra-Diaz JM, Syphard AD, Regan HM (2017) Big data for forecasting the impacts of global change on plant communities. Glob Ecol Biogeogr 26:6–17. https://doi.org/10.1111/geb.12501

Franklin S, Comer P, Evens J, Ezcurra E, Faber-Langendoen D, Franklin J, Jennings M, Josse C, Lea C, Loucks O, Muldavin E, Peet R, Ponomarenko S, Roberts D, Solomeshch A, Keeler-Wolf T, Kley JV, Weakley A, McKerrow A, Burke M, Spurrier C (2015) How a national vegetation classification can help ecological research and management. Front Ecol Environ 13(4):185–186. https://doi.org/10.1890/15.WB.006

Frieswyk CB, Johnston CA, Zedler JB (2007) Identifying and characterizing dominant plants as an indicator of community condition. J Great Lakes Res 33(sp3):125–135. https://doi.org/10.3394/0380-1330(2007)33

Frost CC (2006) History and future of the longleaf pine ecosystem. In: Jose S, Jokela E, Miller D (eds) The longleaf pine ecosystems: ecology, management, and restoration. Springer, New York, pp 9–48

Gamer M, Lemon J, Fellows I, Singh P (2012) Irr: various coefficients of Interrater reliability and agreement. R package version 0.84. https://CRAN.R-project.org/package=irr . Accessed 30 Aug 2017

Gesch D, Oimoen M, Greenlee S, Nelson C, Steuck M, Tyle D (2002) The National Elevation Dataset. Photogramm Eng Rem S 68:5–11

Greenacre M (2013) The contributions of rare objects in correspondence analysis. Ecology 94(1):241–249. https://doi.org/10.1890/11-1730.1

Hak JC, Comer PJ (2017) Modeling landscape condition for biodiversity assessment—application in temperate North America. Ecol Indic 82:206–216. https://doi.org/10.1016/j.ecolind.2017.06.049

Hanberry BB (2013) Changing eastern broadleaf, southern mixed, and northern mixed forest ecosystems of the eastern United States. Forest Ecol Manag 306:171–178. https://doi.org/10.1016/j.foreco.2013.06.040

Healey SP, Raymond CL, Blakey Lockman I, Hernandez AJ, Garrard C, Huang CQ (2016) Root disease can rival fire and harvest in reducing forest carbon storage. Ecosphere 7(11):e01569. https://doi.org/10.1002/ecs2.1569

Heffernan JB, Soranno PA, Angilletta MJ, Buckley LB, Gruner DS, Keitt TH, Kellner JR, Kominoski JS, Rocha AV, Xiao JF, Harms TK, Goring SJ, Koenig LE, McDowell WH, Powell H, Richardson AD, Stow CA, Vargas R, Weathers KC (2014) Macrosystems ecology: understanding ecological patterns and processes at continental scales. Front Ecol Environ 12(1):5–14. https://doi.org/10.1890/130017

Hiers JK, Walters JR, Mitchell RJ, Varner JM, Conner LM, Blanc LA, Stowe J (2014) Ecological value of retaining pyrophytic oaks in longleaf pine ecosystems. J wildlife. Manage 78(3):383–393

Hijmans RJ (2016) Raster: geographic data analysis and modeling. R package version 2.5-8. https://CRAN.R-project.org/package=raster . Accessed 1 June 2017

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. Int J Climatol 25(15):1965–1978

Hildebrand H, Bennet DM, Cadotte MW (2008) Consequences of dominance: a review of evenness effects on local and regional ecosystem processes. Ecology 89(6):1510–1520

Hoagland B (2000) The vegetation of Oklahoma : a classification for landscape mapping and conservation planning. Southwest Nat 45:385–420

Homer C, Dewitz J, Yang LM, Jin SM, Danielson P, Xian G, Coulston J, Herold N, Wickham J, Megown K (2015) Completion of the 2011 National Land Cover Database for the conterminous United States-representing a decade of land cover change information. Photogramm Eng Rem S 81(5):345–354

Iverson LR, Prasad AM (2001) Potential changes in tree species richness and Forest Community types following climate change. Ecosystems 4(3):186–199. https://doi.org/10.1007/s10021-001-0003-6

Iverson LR, Prasad AM, Matthews SN, Peters M (2008) Estimating potential habitat for 134 eastern US tree species under six climate scenarios. Forest Ecol Manag 254:390–406. https://doi.org/10.1016/j.foreco.2007.07.023

Kardol P, Campany CE, Souza L, Norby RJ, Weltzin JF, Classen AT (2010) Climate change effects on plant biomass alter dominance patterns and community evenness in an experimental old-field ecosystem. Glob Chang Biol 16(10): 2676–2687. https://doi.org/10.1111/j.1365-2486.2010.02162.x

Kassambara A, Mundt F (2017) Factoextra: extract and visualize the results of multivariate data analyses. R package version 1.0.4

Koch FH, Coulston JW (2015) One-year (2013), three-year (2011–2013), and five-year (2009–2013) drought maps for the conterminous United States. In: Potter KM, Conkling BL (eds) Forest health monitoring: National Status, trends, and analysis 2014. Gen. Tech. Rep. SRS-GTR-209. US Department of Agriculture, Forest Service, Southern Research Station, Asheville, pp 57–71

Le Roux PC, Pellissier L, Wisz MS, Luoto M (2014) Incorporating dominant species as proxies for biotic interactions strengthens plant community models. J Ecol 102(3):767–775. https://doi.org/10.1111/1365-2745.12239

Levine JM, Bascompte J, Adler PB, Allesina S (2017) Beyond pairwise mechanisms of species coexistence in complex communities. Nature 546:56–64. https://doi.org/10.1038/nature22898

Liang JJ, Crowther TW, Picard N, Wiser S, Zhou M, Alberti G, Schulze ED, McGuire AD, Bozzato F, Pretzsch H, de-Miguel S, Paquette A, Herault B, Scherer-Lorenzen M, Barrett CB, Glick HB, Hengeveld GM, Nabuurs GJ, Pfautsch S, Viana H, Vibrans AC, Ammer C, Schall P, Verbyla D, Tchebakova N, Fischer M, Watson JV, HYH C, Lei XD, Schelhaas MJ, Lu HC, Gianelle D, Parfenova EI, Salas C, Lee E, Lee B, Kim HS, Bruelheide H, Coomes DA, Piotto D, Sunderland T, Schmid B, Gourlet-Fleury S, Sonke B, Tavani R, Zhu J, Brandl S, Vayreda J, Kitahara F, Searle EB, Neldner VJ, Ngugi MR, Baraloto C, Frizzera L, Balazy R, Oleksyn J, Zawila-Niedzwiecki T, Bouriaud O, Bussotti F, Finer L, Jaroszewicz B, Jucker T, Valladares F, Jagodzinski AM, Peri PL, Gonmadje C, Marthy W, O'Brien T, Martin EH, Marshall AR, Rovero F, Bitariho R, Niklaus PA, Alvarez-Loayza P, Chamuya N, Valencia R, Mortier F, Wortel V, Engone-Obiang NL, Ferreira LV, Odeke DE, Vasquez RM, Lewis SL, Reich PB (2016) Positive biodiversity–productivity relationship predominant in global forests. Science 354: aaf8957-1-aaf8957-12. https://doi.org/10.1126/science.aaf8957

Liaw A, Wiener M (2002) Classification and regression by randomForest. R News 2:18–22

Matthews ER, Peet RK, Weakley AS (2011) Classification and description of alluvial plant communities of the piedmont region, North Carolina, USA. Appl Veg Sci 14:485–505

McCune B, Grace JB (2002) Analysis of ecological communities. MjM Software Design, Gleneden Beach

McGill BJ (2010) Matters of scale. Science 328:575–576. https://doi.org/10.1126/science.1188528

Menard S, Faber-Langendoen D, Nelson M (2017) Integrating the U.S. National Vegetation Classification in the U.S. Forest Service FIA Program. Report prepared for USFS-FIA program, Arlington, p 104

Nenadic O, Greenacre M (2007) Correspondence analysis in R, with two- and three-dimensional graphics: the ca package. J Stat Softw 20(3):48202. https://doi.org/10.18637/jss.v020.i03

Neuwirth E (2014) RColorBrewer: ColorBrewer palettes. R package version 1:1–2

Nowacki GJ, Abrams MD (2008) The demise of fire and "mesophication" of forests in the eastern United States. Bioscience 58(2):123–138

O'Connell BM, Conkling BL, Wilson AM, Burrill EA, Turner JA, Pugh SA, Christensen G, Ridley T, Menlove J (2016) The Forest inventory and analysis database: database description and user guide for phase 2 (version 6.1). US Department of Agriculture Forest Service, Washington, DC, USA.

Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Szoecs E, Wagner H (2016) Vegan: community ecology package. R package version 2.4-0. https://CRAN.R-project.org/package=vegan . Accessed 1 July 2017

Palmquist KA, Peet RK, Weakley AS (2014) Changes in plant species richness following reduced fire frequency and drought in one of the most species-rich savannas in North America. J Veg Sci 25(6):1426–1437. https://doi.org/10.1111/jvs.12186

Peet R, Lee M, Jennings M, Faber-Langendoen D (2012) VegBank – a permanent, open-access archive for vegetation-plot data. Biodivers Ecol 4:233–241. https://doi.org/10.7809/b-e.00080

Prasad AM, Iverson LR, Liaw A (2006) Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems 9:181–199. https://doi.org/10.1007/s10021-005-0054-1

R Core Team (2017) R: A Language and Environment for Statistical Computing. https://www.r-project.org/ Accessed 10 Aug 2017

Riitters KH, Wickham JD (2012) Decline of forest interior conditions in the conterminous United States. Sci Rep 2:653. https://doi.org/10.1038/srep00653

Roberts DW (2016) Labdsv: ordination and multivariate analysis for ecology. R package version 1.8-0. https://CRAN.R-project.org/package=labdsv . Accessed 15 June 2017

Rogers BM, Jantz P, Goetz SJ (2017) Vulnerability of eastern US tree species to climate change. Glob Chang Biol 23(8):3302–3320. https://doi.org/10.1111/gcb.13585

Rose KC, Graves RA, Hansen WD, Harvey BJ, Qiu JX, Wood SA, Ziter C, Turner MG (2016) Historical foundations and future directions in macrosystems ecology. Ecol Lett 20(2):147–157. https://doi.org/10.1111/ele.12717

Ruefenacht B, Finco MV, Nelson MD, Czaplewski R, Helmer EH, Blackard JA, Holden GR, Lister AJ, Salajanu D, Weyermann D, Winterberger K (2008) Conterminous U.S. and Alaska Forest type mapping using Forest inventory and analysis data. Photogramm Eng Rem S 74(11):1379–1388. https://doi.org/10.14358/PERS.74.11.1379

Schaetzl RJ, Krist FJ, Miller BA (2012) A taxonomically based ordinal estimate of soil productivity for landscape-scale analyses. Soil Sci 177(4):288–299. https://doi.org/10.1097/SS.0b013e3182446c88

Schaetzl RJ, Krist FJ, Stanley K, Hupy CM (2009) The natural soil drainage index: an ordinal estimate of long-term soil wetness. Phys Geogr 30(5):383–409. https://doi.org/10.2747/0272-3646.30.5.383

Shmida A, Wilson MV (1985) Biological determinants of species diversity. J Biogeogr 12:1–20. https://doi.org/10.2307/2845026

Smith WB (2002) Forest inventory and analysis: a national inventory and monitoring program. Environ Pollut 116:S233–S242. https://doi.org/10.1016/S0269-7491(01)00255-X

Soil Survey Staff (2017a) The gridded soil survey geographic (gSSURGO) database. United States Department of Agriculture, Natural Resources Conservation Service. https://gdg.sc.egov.usda.gov/ . Accessed 15 Mar 2017

Soil Survey Staff (2017b) Soil survey geographic (STATSGO2) database. United States Department of Agriculture, Natural Resources Conservation Service. https://sdmdataaccess.sc.egov.usda.gov. Accessed 15 Mar 2017

Sork VL, Davis FW, Westfall R, Flint A, Ikegami M, Wang HF, Grivet D (2010) Gene movement and genetic association with regional climate gradients in California valley oak (Quercus Lobata nee) in the face of climate change. Mol Ecol 19(17):3806–3823. https://doi.org/10.1111/j.1365-294X.2010.04726.x

Thompson ID, Guariguata MR, Okabe K, Bahamondez C, Nasi R, Heymell V, Sabogal C (2013) An operational framework for defining and monitoring Forest degradation. Ecol Soc 18(2):art20. https://doi.org/10.5751/ES-05443-180220

Tichý L, Chytrý M, Botta-Dukát Z (2014) Semi-supervised classification of vegetation: preserving the good old units and searching for new ones. J Veg Sci 25(6):1504–1512. https://doi.org/10.1111/jvs.12193

Tierney GL, Faber-Langendoen D, Mitchell BR, Shriver WG, Gibbs JP (2009) Monitoring and evaluating the ecological integrity of forest ecosystems. Front Ecol Environ 7(6):308–316. https://doi.org/10.1890/070176.

USNVC (2016) USNVC [United States National Vegetation Classification] Database, v2.0. Federal Geographic Data Committee, Vegetation Subcommittee, Washington DC. http://www.usnvc.org. Accessed 10 Aug 2017

Venables WN, Ripley BD (2002) Modern applied statistics with S, fourth. Springer, New York

Wickham H (2017) Tidyverse: easily install and load "Tidyverse" packages. R package version 1.1.1. http://tidyverse.tidyverse.org . Accessed 1 Aug 2017

Zhu K, Woodall CW, Ghosh S, Gelfand AE, Clark JS (2014) Dual impacts of climate change: forest migration and turnover through life history. Glob Chang Biol 20(1):251–264. https://doi.org/10.1111/gcb.12382