Phylogeny and the prediction of tree functional diversity across novel continental settings

Nathan G. Swenson1*, Michael D. Weiser2, Lingfeng Mao1, Miguel B. Araújo3,4,5, José Alexandre F. Diniz-Filho6, Johannes Kollmann7, David Nogués-Bravo4, Signe Normand8,12, Miguel A. Rodríguez9, Raúl García-Valdés10, Fernando Valladares3,11, Miguel A. Zavala9 and Jens-Christian Svenning12

ABSTRACT

Aim Mapping the distribution and diversity of plant functional traits is critical for projecting future changes to vegetation under global change. Maps of plant functional traits, however, are scarce due very sparse global trait data matrices. A potential solution to this data limitation is to utilize the known levels of phylogenetic signal in trait data to predict missing values. Here we aim to test existing phylogenetic comparative methods for imputing missing trait data for the purpose of producing continental-scale maps of plant functional traits.

Location North America and Europe.

Methods Phylogenetic imputation models and trait data from one continent were used to predict the trait values for tree species on the other continent and to produce trait maps. Predicted maps of trait means, variances and functional diversity were compared with known maps to quantify the degree to which predicted trait values could estimate spatial patterns of trait distributions and diversity.

Results We show that the phylogenetic signal in plant functional trait data can be used to provide robust predictions of the geographical distribution of tree functional diversity. However, predictions for traits with little phylogenetic signal, such as maximum height, are error prone. Lastly, trait imputation methods based on phylogenetic generalized least squares tended to outperform those based on phylogenetic eigenvectors.

Main conclusions It is possible to predict patterns of functional diversity across continental settings with novel species assemblages for most of the traits studied for which we have no direct trait information, thereby offering an effective method for overcoming a key data limitation in global change biology, macroecology and ecosystem modelling.

Keywords Forest ecology, imputation, plant biodiversity, phylogeny, temperate forest, trait biogeography.
INTRODUCTION

Theoretical and empirical ecological investigations suggest that strong linkages exist between plant functional diversity and ecosystem function (Tilman et al., 1997; Loreau et al., 2001). The distribution of functional diversity across a variety of spatial scales is therefore of fundamental interest to ecosystem modellers. Quantifying the continental-scale distribution of plant functional diversity has, however, been particularly challenging due to limitations in the available species trait data (Reich, 2005; Swenson & Weiser, 2010; Swenson et al., 2012). This lack of information has led ecosystem modellers to characterize vegetation types using a few plant functional types, leading to coarse and potentially inaccurate projections of ecosystem function under global climate change (Purves & Pacala, 2008; van Bodegom et al., 2012).

The most obvious obstacle to estimating the continental-scale distribution of plant functional diversity is the requirement for species-level functional trait data that are linked to performance for thousands of species distributed across vast areas, as well as specific knowledge about how such traits are directly or indirectly linked to ecosystem function or persistence. It may require many years to collect such data, even in less diverse temperate floras, and much longer in highly diverse tropical floras (Swenson, 2013; Umaña et al., 2015). A potentially powerful and more easily employed alternative or stopgap measure is to take advantage of phylogenetic signal in functional traits (i.e. the tendency for closely related species to have similar trait values) to estimate the function of individual species. Plant ecologists have demonstrated a large degree of phylogenetic signal in global-scale studies of plant functional traits (e.g. Moles et al., 2005; Swenson & Enquist, 2007), suggesting that reasonable estimates of trait values for species that are absent in global databases may be possible based on their phylogenetic position. Specifically, phylogenetic imputation, in which a model of trait evolution is applied to a phylogeny to estimate the missing trait values for species, holds tremendous promise (Swenson, 2014a). However, these methods have not yet been applied to large plant trait databases nor have they been used to predict the spatial distribution of multiple traits across continents or to predict the distribution of functional diversity itself.

Here, we show that phylogenetic information can be used to generate robust predictions of the distribution of individual functional traits and the overall functional diversity of tree assemblages on continental scales. The analyses focus on using phylogenetic generalized least squares (pGLS) regression and phylogenetic eigenvector regression to evaluate phylogenetic signal in available trait data from one continent and to estimate the functional trait values of individual species on another continent based upon their phylogenetic position (Martins & Hansen, 1997; Garland & Ives, 2000; Swenson, 2014a,b). The analyses were conducted using the geographical distribution of tree species in eastern North America and Europe, a phylogenetic tree of these species and data for four key functional traits (leaf size, maximum height, seed mass and wood density) for all species. The specific questions we ask are: (1) can the mean and variance of individual traits and multivariate functional diversity of tree species on one continent be predicted by simply knowing the traits and phylogenetic positions of a different set of species on a different continent; (2) does a lack of detailed phylogenetic information within genera greatly hinder predictive models; and (3) do alternative phylogenetic regression models, such as those built using phylogenetic eigenvectors, provide robust predictions of the distribution and diversity of plant function across continents?

MATERIALS AND METHODS

Geographical data

Geographical range maps for 273 eastern North American and 121 European tree species were used in this study (we defined a ‘tree’ as any free-standing woody plant with a maximum height greater than 10 m). Tree species in these two regions that did not have trait data available in the literature were not included in the study. The eastern North American tree range maps were downloaded from the United States Geological Survey (http://esp.cr.usgs.gov/data/little/) and gridded into 1° squares. The European tree range data were digitized from the Atlas Flora Europaeae (http://www.loumus.fi/english/botany/afe/) and were gridded using the atlas’s map grid system where grid cells are 50 km² on average. The two tree floras used are well known for their compositional similarity, making them a probable ‘best case scenario’ for phylogenetic imputation. Specifically, 72.7% of the genera in our European data set are in the North American data set and 25.2% of the North American genera are in our European data set.

Phylogenetic tree

A single phylogenetic tree was generated for this study using the eco-informatics software Phylocom (Webb & Donoghue, 2005). Specifically, we used the Phylocom R20100701.new backbone phylogeny and our species list to produce a phylogeny. Generally, the degree of relatedness between species within genera was left unresolved using this approach (i.e. all congeneric species pairs were treated as equally related). We used this approach to generate the phylogenetic tree because it is likely to be the approach most widely employed by ecologists in the future attempting to predict trait data on continental scales, particularly in geographical regions where DNA sequences for most species are unavailable (e.g. tropical floras).

Trait data

This study utilized data for four traits that indicate where a species falls along the spectrum of plant ecological strategies (e.g. Grubb, 1977; Dolph & Dilcher, 1980; Chave et al., 2009; Moles et al., 2009). These traits were also used because they are widely available, allowing for model testing. The traits we
considered were maximum height, seed mass, wood density and leaf size, and were recorded for every species (i.e. there were no missing trait values for any species or trait). The maximum height data came from the literature where we recorded the absolute largest value reported (Britton & Shafer, 1923; Poulton, 1976) and the United States Department of Agriculture PLANTS database (http://plants.usda.gov). The wood density data came from the global wood density database published by Chave et al. (2009) and from additional literature sources (Iatsenko-Khmelevski, 1954; Bosshard, 1974). Leaf area was estimated as the product of the reported leaf length, leaf width and 0.70 to account for leaf tapering. This calculation has recently been shown to produce values that are highly correlated with the known area of leaves (Kraft et al., 2008) and represents a pragmatic approach for estimating leaf area for hundreds of species from the literature. For some species, the leaf length and/or width was not available in the literature and was recorded by N.G.S. using herbarium specimens in the Gray Herbarium at Harvard University and the Michigan State University Herbarium. Because the degree of leaf shrinkage across these taxa was not known and leaves could not be rehydrated we retained the dry dimensions. We expect that this introduced error is minimal given the total variation in leaf size in our data set and would probably bias towards weaker predictions. Seed mass was recorded from the Kew Millennium Seed Database (http://data.kew.org/sid/) and the PLANTS database. An additional 15 species had their seed masses quantified using seeds stored with herbaria sheets at the Michigan State University Herbarium by N.G.S. The maximum height, leaf size and seed mass data were all log transformed for the downstream analyses given their highly skewed global distributions.

**Phylogenetic generalized least squares regression**

We used pGLS regression to model the trait data for species on one continent given their phylogenetic position and the phylogenetic distribution of traits for species on the second continent. A pGLS regression can incorporate the phylogenetic non-independence of data points by assuming a phylogenetic error structure given a model of trait evolution. In the simplest case, a Brownian motion model of trait evolution can be assumed in which the error structure takes the form of an untransformed phylogenetic variance–covariance (VCV) matrix where the diagonal is the root to tip distance and the off-diagonal elements are the amounts of shared branch length between two taxa. This basic model can become more flexible by fitting a model of trait evolution given the data by transforming the phylogenetic VCV matrix and finding the transformation that best fits the data (Swenson, 2014a,b). For example, if the data have no evident phylogenetic signal (i.e. non-independence) the transformation of the off-diagonal values in the VCV matrix that would best fit the data would be to multiply the values by zero. Similarly, if the data are best explained by a Brownian motion model the transformation that would best fit the data would be to multiply the off-diagonal elements by one. The values by which the off-diagonal elements are multiplied are referred to as \( \lambda \). We utilized maximum likelihood to estimate the \( \lambda \) values (Pagel, 1999; Freckleton et al., 2002) using the R package ‘caper’ (http://caper.r-forge.r-project.org/) for each trait on each continent and generated a GLS regression model for that trait using the estimated phylogenetic error structure (i.e. the transformed phylogenetic VCV matrix; Swenson, 2014a,b). This model and the transformed VCV matrix containing all species on both continents were then used to predict the trait values of species on the other continent given the model from first continent. To assess the degree to which the predicted species-level values were related to the known values we regressed the predicted values against the known values.

Next, the predicted values were then used to quantify the mean and variance of traits in map grid cells on each continent as well as the multivariate functional dispersion (FDIs) and functional richness (FDis) in those grid cells. The FDis is the mean distance of each species to the centroid of the multivariate trait space and the FDis is the volume of the multivariate trait space that an assemblage occupies (Laliberte & Legendre, 2010). These values were then compared with the known values using a regression.

A simple alternative to estimating the most likely \( \lambda \) values for a given trait dataset and phylogeny is just to assume that traits evolve under a Brownian motion model. For example, a Brownian motion model could be assumed where the phylogenetic VCV matrix is left untransformed (i.e. \( \lambda = 1 \)). We generated these models for each trait on each continent and used the models and an untransformed phylogenetic VCV matrix containing all species to predict the trait values on the other continent. As with the previous analysis, we then regressed predicted trait values for species against their known values. Then, the predicted values were used to quantify the mean and variance of traits in map grid cells on each continent as well as the multivariate FDIs and FDis in those grid cells.

**Phylogenetic eigenvector regression**

In addition to the two pGLS approaches used to predict trait values, we utilized phylogenetic eigenvectors to predict trait values, which assume no model of trait evolution – Brownian motion or otherwise. To accomplish this, a phylogenetic distance matrix was computed from the phylogeny and used in a principal coordinate analysis to generate phylogenetic eigenvectors (Diniz-Filho et al., 1998; Ramirez et al., 2008; Diniz-Filho et al., 2011). The number of phylogenetic eigenvectors produced is equal to the number of species minus one. A subset of eigenvectors must be selected for phylogenetic eigenvector regression because the use of all eigenvectors leads to model saturation (Rohlf, 2001). We utilized an iterative search for the subset of eigenvectors that reduces the largest amount of autocorrelation in the residuals (Griffith &
Peres-Neto, 2006; Diniz-Filho et al., 2012). Specifically, as new eigenvectors were added to the model for a single trait on a single continent, residual autocorrelation was recalculated and the iterative search stopped until the residual autocorrelation calculated using Moran’s I was less than 0.05. The selected eigenvector values for species on one continent were then used as independent variables in a multiple linear model with the data for a single trait from the same continent as the dependent variable. This model was then projected onto the values for the species on the other continent from the same subset of eigenvectors. This process was repeated for each trait to produce predicted trait values on one continent given the trait data on the other continent and their phylogenetic eigenvector positions. The R package ‘PVR’ was used for all phylogenetic eigenvector analyses (http://cran.r-project.org/web/packages/PVR/). Again, the predicted species-level trait values were regressed onto the known values through the origin and the coefficient of determination was recorded. Next, the predicted trait values derived from this phylogenetic eigenvector approach were then used to quantify the mean and variance of traits in map grid cells on each continent as well as the multivariate FDIs and FRic in those grid cells. These values were then compared with the known values.

Prediction error and climate

Deviations of the predicted map grid cell values from the ‘known’ values may be linked to climate. We therefore performed a series of ad hoc tests in which we first quantified the deviation of the predicted values from the known values (i.e. known value minus the predicted value) and correlated these values with four climatic variables for the same grid cell. Specifically, we used Pearson correlations to evaluate the relationships between the deviations and mean annual temperature, temperature seasonality, annual precipitation and precipitation seasonality using climate maps from the WorldClim database (Hijmans et al., 2005) at a resolution of 2.5°.

RESULTS

We utilized three phylogenetic imputation methods to predict the trait values of species in one region (eastern North America or Europe) based upon their phylogenetic position and the traits and the phylogenetic position of species in the other region. We used the predicted values to map the mean and variance of each trait and to estimate two multivariate functional diversity indices in the map grid cells in each region. We began by testing the pGLS regression with a fit model of trait evolution. The predicted trait means, FDIs and FRic in map grid cells in the projection region based on trait information in the calibration region and phylogenetic information were typically highly correlated ($r^2 > 0.60$; Table 1, Figs 1 & 2). The predicted trait variances in map grid cells were also highly correlated with the known variances ($r^2 > 0.60$; Table 1). However, predictions of the mean and variance of maximum height values for the map grid cells were far weaker ($r^2 \sim 0.1–0.3$) indicating that the lability in the evolution of this trait prevented strong predictions even when $\lambda$ was estimated and used to fit the model. The geographical locations that were the most difficult to predict in Europe were typically in the south-east (Figs 1 & 2). Similarly, the more species-rich south-eastern portion of eastern North America was the region hardest to predict, probably due to the higher number of congeners and the greater number of species that may be distantly related from the dataset used to build the statistical model.

To explore whether alternative phylogenetic prediction frameworks provided similarly strong predictions we took two additional approaches. First, we did not use maximum likelihood to estimate $\lambda$ values in the pGLS model. Rather, we used the observed phylogenetic VCV matrix in the pGLS model, effectively assuming a $\lambda$ value of one (i.e. Brownian
Figure 1 The known (left) and predicted (middle) trait means in map grid cells for European trees. Deviations (right) where the predicted values were subtracted from the known value are also plotted. The top row is mean maximum height (log m), the second row is mean leaf size (log cm²), the third row is mean seed mass (log g) and the fourth row is mean wood density (g cm⁻³). The predicted values were generated by fitting a model of trait evolution for maximum height, leaf size, seed mass and wood density for eastern North American trees and using that model to predict the trait values of European tree species based on their phylogenetic position. The colour legends are provided on the right side of each row with the top legend corresponding to the maps in the first two columns (i.e. the trait means) and the bottom legend to the map in the last column (i.e. the deviations).
motion trait evolution) for every trait dataset (Table 2). Second, we utilized a phylogenetic eigenvector regression that does not fit a model of trait evolution (Table 3). The results from both pGLS approaches were qualitatively similar (Tables 1 & 2) where strong predictions were possible for most traits, with the notable exception of maximum height. The phylogenetic eigenvector predictions were less robust, with some traits having strong predictions (e.g. leaf area and seed mass); wood density and maximum height predictions were less strong (Table 3).

Lastly, we quantified the correlation between four climatic variables and deviations of the predicted values from known values for map grid cells on both continents. We found that deviations were nearly always correlated with the four climatic variables (Tables S1 & S2 in the Supporting Information). The correlations were generally stronger for temperature-related variables than for precipitation-related variables. The geographical signature in the deviations for Europe can be seen in Fig. 1, indicating that in the study system the major deviations generally occur at the extremes of latitude.

**DISCUSSION**

Mapping the distribution and diversity of plant functional trait on continental scales is a fundamental goal in biogeography and ecosystem ecology (Reich, 2005; Swenson & Weiser, 2010; Swenson et al., 2012). A key limitation to progress is that most large plant trait databases are highly sparse (Kattge et al., 2011) so probably making most efforts at functional trait mapping prone to large error. While waiting for more data to accumulate, a pragmatic way forward may be to impute or estimate the missing trait values in existing databases. These estimates could be strengthened by incorporating phylogenetic information (Swenson, 2014a; Schrodt et al., 2015). This is because some plant functional traits of

---

**Figure 2** The multivariate functional richness (FRic) quantified using the known trait data (left) and the predicted trait data (right) for European trees. The predicted values were generated by fitting a model of trait evolution for maximum height, leaf size, seed mass and wood density for eastern North American trees and using that model to predict the trait values of European tree species based on their phylogenetic position. The known and predicted FRic values are highly correlated ($r^2 = 0.964$) with a lower than expected root mean squared error (RMSE = 1.944; $P < 0.05$).

**Table 1**

<table>
<thead>
<tr>
<th>Functional Richness</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.54 - 4.83</td>
</tr>
<tr>
<td>4.84 - 7.46</td>
</tr>
<tr>
<td>7.47 - 10.09</td>
</tr>
<tr>
<td>10.10 - 12.33</td>
</tr>
<tr>
<td>12.34 - 14.51</td>
</tr>
<tr>
<td>14.52 - 16.88</td>
</tr>
</tbody>
</table>

---

Global Ecology and Biogeography, 26, 553–562, © 2017 John Wiley & Sons Ltd
Table 2 In this table we do not estimate a model of trait evolution, rather we assume a Brownian motion model of trait evolution ($\lambda = 1$) and phylogenetic generalized least squares. The predicted trait values and species distribution maps were then used to calculate the predicted mean and variance of each trait value and the predicted multivariate functional dispersion and functional richness value in map grid cells on each continent. The predicted mean, variance, functional dispersion and functional richness values were regressed onto the known values. The table gives the intercept and slope of each regression with their standard errors (SE) and the $r^2$.

<table>
<thead>
<tr>
<th>Map grid cell value</th>
<th>Intercept</th>
<th>SE</th>
<th>Slope</th>
<th>SE</th>
<th>$r^2$</th>
<th>Intercept</th>
<th>SE</th>
<th>Slope</th>
<th>SE</th>
<th>$r^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean maximum height (m)</td>
<td>1.13</td>
<td>0.01</td>
<td>0.162</td>
<td>0.01</td>
<td>0.18</td>
<td>1.33</td>
<td>0.01</td>
<td>0.026</td>
<td>0.01</td>
<td>0.11</td>
</tr>
<tr>
<td>Variance maximum height (m)</td>
<td>0.02</td>
<td>&lt; 0.00</td>
<td>−0.068</td>
<td>0.01</td>
<td>0.14</td>
<td>0.02</td>
<td>&lt; 0.00</td>
<td>−0.118</td>
<td>0.01</td>
<td>0.18</td>
</tr>
<tr>
<td>Mean leaf size (cm$^2$)</td>
<td>−0.09</td>
<td>0.01</td>
<td>1.174</td>
<td>0.01</td>
<td>0.88</td>
<td>0.17</td>
<td>0.01</td>
<td>0.604</td>
<td>0.01</td>
<td>0.77</td>
</tr>
<tr>
<td>Variance leaf size (cm$^2$)</td>
<td>−0.03</td>
<td>&lt; 0.00</td>
<td>0.942</td>
<td>0.01</td>
<td>0.93</td>
<td>−0.02</td>
<td>0.01</td>
<td>0.631</td>
<td>0.01</td>
<td>0.88</td>
</tr>
<tr>
<td>Mean seed mass (g)</td>
<td>0.13</td>
<td>&lt; 0.00</td>
<td>0.794</td>
<td>&lt; 0.00</td>
<td>0.98</td>
<td>0.28</td>
<td>0.01</td>
<td>0.907</td>
<td>0.01</td>
<td>0.91</td>
</tr>
<tr>
<td>Variance seed mass (g)</td>
<td>0.10</td>
<td>0.01</td>
<td>0.641</td>
<td>0.01</td>
<td>0.88</td>
<td>0.68</td>
<td>0.02</td>
<td>0.316</td>
<td>0.02</td>
<td>0.41</td>
</tr>
<tr>
<td>Mean wood density (g cm$^{-3}$)</td>
<td>0.01</td>
<td>&lt; 0.00</td>
<td>0.973</td>
<td>0.01</td>
<td>0.93</td>
<td>−0.03</td>
<td>0.01</td>
<td>1.076</td>
<td>0.02</td>
<td>0.99</td>
</tr>
<tr>
<td>Variance wood density (g cm$^{-3}$)</td>
<td>0.01</td>
<td>&lt; 0.00</td>
<td>0.117</td>
<td>0.01</td>
<td>0.26</td>
<td>0.01</td>
<td>&lt; 0.00</td>
<td>0.243</td>
<td>0.02</td>
<td>0.30</td>
</tr>
<tr>
<td>Functional dispersion</td>
<td>0.46</td>
<td>0.02</td>
<td>0.767</td>
<td>0.01</td>
<td>0.44</td>
<td>0.07</td>
<td>0.02</td>
<td>1.121</td>
<td>0.01</td>
<td>0.88</td>
</tr>
<tr>
<td>Functional richness</td>
<td>3.67</td>
<td>0.09</td>
<td>0.619</td>
<td>0.01</td>
<td>0.67</td>
<td>4.16</td>
<td>0.15</td>
<td>0.786</td>
<td>0.01</td>
<td>0.81</td>
</tr>
</tbody>
</table>

interest are known to have a phylogenetic signal in global datasets (e.g. Moles et al., 2005; Swenson & Enquist, 2007). The goal of the present work was to implement and test the ability of phylogenetic imputation methods to predict the distribution and diversity of plant functional traits on continental scales.

Here, we have shown that robust predictions of individual trait distributions and the overall functional diversity within map grid cells can be predicted among novel continental settings simply by taking advantage of the phylogenetic signal in trait data from another continent. The three approaches to phylogenetic imputation used here all were able to predict a large amount of the variance in trait distributions at the species and map grid cell levels (Tables 1–3). However, the two pGLS regression-based approaches explained more variance than phylogenetic eigenvector regression-based methods. Thus, even when setting aside conceptual debates regarding eigenvector approaches (e.g. Rohlf, 2001), these methods tended to perform well, but not as well as pGLS. The two pGLS approaches yielded similar results in this study, but it is expected that in many cases trait evolution will not as closely approximate a Brownian motion model and the pGLS approach fitting a $\lambda$ value will be more reliable. More work is needed, using larger empirical and simulated datasets, (Swenson, 2014a) to confirm or reject this general recommendation.

The phylogenetic imputation methods were able to make strong predictions of the spatial distribution of traits, but the

Table 3 We used phylogenetic eigenvector regression using the trait data from one continent to predict the trait values for species on the other continent. The predicted trait values and species distribution maps were then used to calculate the predicted mean and variance of each trait value and the predicted multivariate functional dispersion and functional richness value in map grid cells on each continent. The predicted mean, variance, functional dispersion and functional richness values were regressed onto the known values through the origin. This table shows the intercept and slope of each regression with their standard errors (SE) and the $r^2$.

<table>
<thead>
<tr>
<th>Map grid cell value</th>
<th>Intercept</th>
<th>SE</th>
<th>Slope</th>
<th>SE</th>
<th>$r^2$</th>
<th>Intercept</th>
<th>SE</th>
<th>Slope</th>
<th>SE</th>
<th>$r^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean maximum height (m)</td>
<td>1.26</td>
<td>0.01</td>
<td>0.06</td>
<td>0.01</td>
<td>0.11</td>
<td>1.33</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.08</td>
</tr>
<tr>
<td>Variance maximum height (m)</td>
<td>0.01</td>
<td>&lt; 0.00</td>
<td>−0.05</td>
<td>0.01</td>
<td>0.11</td>
<td>0.01</td>
<td>&lt; 0.00</td>
<td>−0.06</td>
<td>0.01</td>
<td>0.11</td>
</tr>
<tr>
<td>Mean leaf size (cm$^2$)</td>
<td>−0.17</td>
<td>0.01</td>
<td>1.16</td>
<td>0.01</td>
<td>0.75</td>
<td>0.27</td>
<td>0.01</td>
<td>0.53</td>
<td>0.02</td>
<td>0.63</td>
</tr>
<tr>
<td>Variance leaf size (cm$^2$)</td>
<td>0.02</td>
<td>0.01</td>
<td>0.76</td>
<td>0.01</td>
<td>0.46</td>
<td>0.41</td>
<td>0.02</td>
<td>0.43</td>
<td>0.02</td>
<td>0.46</td>
</tr>
<tr>
<td>Mean seed mass (g)</td>
<td>0.36</td>
<td>&lt; 0.00</td>
<td>0.73</td>
<td>&lt; 0.00</td>
<td>0.96</td>
<td>1.03</td>
<td>0.01</td>
<td>0.43</td>
<td>0.02</td>
<td>0.52</td>
</tr>
<tr>
<td>Variance seed mass (g)</td>
<td>−0.08</td>
<td>0.01</td>
<td>0.60</td>
<td>0.01</td>
<td>0.81</td>
<td>0.65</td>
<td>0.03</td>
<td>0.18</td>
<td>0.01</td>
<td>0.11</td>
</tr>
<tr>
<td>Mean wood density (g cm$^{-3}$)</td>
<td>0.22</td>
<td>0.01</td>
<td>0.60</td>
<td>0.02</td>
<td>0.51</td>
<td>0.03</td>
<td>0.02</td>
<td>0.96</td>
<td>0.04</td>
<td>0.93</td>
</tr>
<tr>
<td>Variance wood density (g cm$^{-3}$)</td>
<td>&lt; 0.00</td>
<td>&lt; 0.00</td>
<td>0.04</td>
<td>0.01</td>
<td>0.15</td>
<td>0.01</td>
<td>&lt; 0.00</td>
<td>−0.06</td>
<td>0.01</td>
<td>0.09</td>
</tr>
<tr>
<td>Functional dispersion</td>
<td>0.09</td>
<td>0.03</td>
<td>0.85</td>
<td>0.01</td>
<td>0.44</td>
<td>0.21</td>
<td>0.03</td>
<td>0.92</td>
<td>0.02</td>
<td>0.69</td>
</tr>
<tr>
<td>Functional richness</td>
<td>0.44</td>
<td>0.15</td>
<td>0.90</td>
<td>0.01</td>
<td>0.65</td>
<td>−1.26</td>
<td>0.56</td>
<td>1.88</td>
<td>0.02</td>
<td>0.64</td>
</tr>
</tbody>
</table>
northern- and southernmost portions of both regions were
where the methods performed worst (Figs 1 & 2). This is
particularly evident when we consider the strong relation-
ships between temperature variables and the deviation of pre-
dicted values from known values (Tables S1 & S2). One
reason for this may be the tendency of the methods to
under-represent trait divergences due to habitat differences
within a clade and an over-averaging of trait data leading to
higher deviations in more extreme climates within clades.
Future work may be able to remedy this bias by either incor-
porating climatic information into the species-level trait pre-
dictions or adjusting predicted species-level values in map
grid cells or the assemblage-level trait or diversity values
based upon climate, but such work is beyond the scope of
the present paper. A second reason is that these regions con-
tain a greater number of species from different parts of the
phylogenetic tree. In other words, the distance between a
data point used to build the statistical model and a species
in these regions will increase. This is particularly the case
when building a model on one continent and projecting it to
another where it is likely that many genera in the species-
rich regions on the continent to be predicted are not found
on the continent used to build the model.

This study focused on four functional traits commonly
used in trait-based ecology and readily available in the litera-
ture. Two of these traits, wood density and seed mass, are
known to have a great deal of phylogenetic signal (Moles
et al., 2005; Swenson & Enquist, 2007), meaning that phylo-
genetic imputation methods are likely to be very successful.
Indeed, we found this to be the case at the species and map
grid cell levels (Tables 1–3). The phylogenetic signal in the
other two traits, maximum height and leaf size, has not been
as well scrutinized in the literature at global scales. Maximum
height was found to have much less phylogenetic signal than
the other traits, but leaf size had a similar degree of phylo-
genetic signal to seed mass and wood density (Table 1). The
outcome of this was that predictions of maximum height dis-
tributions were far less reliable than those of leaf size distri-
butions (Tables 1–3).

Considerations for future implementation of
phylogenetic imputation

It may seem surprising that our phylogenetically based
approach is able to predict the observed geographical pat-
terns so strongly. We expect that some of this success is due
to the fact that the two tree floras are very similar in their
familial and generic compositions. Thus, the average phylo-
genetic distance between a training trait data point and a pre-
dicted trait data point is relatively low and represents
perhaps a best-case scenario. In other words, projecting the
traits of another flora with a very different phylogenetic com-
position (e.g. the Amazon) from European data would be
likely to result in much more error. Indeed, we found evi-
dence of this to a smaller degree when we consider that less
variation in eastern North America could be predicted using
the smaller European flora (e.g. Table 1). Additionally, the
methods used are regressions and extrapolations of these
models, so will more likely than not introduce large errors.
In the present study, the bounds of the data in each region
are roughly similar, but if one region lacked, for example,
gymnosperms there would be a highly increased potential for
error. Taken together, future work will have to closely con-
sider the phylogenetic compositions of the training data set
and the species set to be predicted. Some of the potential for
error could be mitigated by using the largest trait datasets
available (e.g. Kattge et al., 2011; Schrotz et al., 2015) such
that phylogenetic extrapolation does not occur and the pre-
dicted trait values can stay within reasonable bounds.

Another consideration arising from this study is that we
only considered four traits that, while being of interest to
ecologists, do not represent the entirety of the traits that
ecologists are interested in mapping. For example, earth sys-
tem modellers are likely to be more interested in leaf gas
exchange rates that may be highly variable within families
and genera (i.e. have little phylogenetic signal; see van Bode-
gom et al., 2012). Such traits may approximate the situation
we encountered with maximum height where predictions are
not as strong, and this would propagate error once aggre-
gated into things like global dynamic vegetation models.
Thus, an important question will be the degree to which the
error introduced via phylogenetic imputation is less or more
than the error introduced by lumping species into a few dis-
crete functional types.

Next, the present study found strong relationships between
climate and deviations from predictions. Each of the meth-
ods used could incorporate climatic information by quantifying
the average climate for each species and using this
information as an additional independent variable in the
model, such that phylogenetic signal and trait–climate rela-
tionships are simultaneously used to predict missing trait val-
ues. It is expected that such models will strengthen trait
predictions, particularly when phylogenies with no resolution
within genera are utilized. An alternative approach could be
adjusting post hoc the grid cell values for assemblages by cli-
mate, but this approach may be more arbitrary and unreli-
able. More detailed future models may also seek to model
population-level response to climate hierarchically, which
may help refine predictions of traits that are very sensitive
to local abiotic conditions (e.g. gas exchange). However, to our
knowledge, such phylogenetically explicit methods that model
trait evolution along branch lengths have not yet been
developed.

Lastly, it is worth highlighting again that the proposed
methods are meant to serve as a pragmatic approach to esti-
mat ing trait values given the current circumstances. Without
a doubt we would prefer that trait values were actually mea-
ured than predicted, and future trait collection campaigns,
particularly in under-sampled regions like the tropics, should
remain a priority. Further, as previously noted (see Swenson,
2014a), while the biases or errors introduced by phylogenetic
imputation may be tolerable on very large scales, using
imputed values for local-scale studies or community ecology would be likely to introduce levels of error that would not be tolerable. Thus, we are not recommending the use of these methods for trait-based community ecology.

CONCLUSIONS

In recent years plant ecologists and evolutionary biologists have made tremendous advances by generating and analysing large plant trait databases (Kattge et al., 2011) and large phylogenetic trees (Webb & Donoghue, 2005). We suggest that these advances can now be leveraged to produce phylogenetically based predictions of the continental-scale distribution and the diversity of plant function, even into areas with novel sets of species. This predictive power will be crucial in a future where climate change and species introductions will increasingly generate novel assemblages. Importantly these predictions may be the most pragmatic way for ecosystem modellers to incorporate functional diversity within and among map grid cells into their models and move beyond using a singular plant functional type to represent all vegetation within a region, and to do so even for less-studied regions with many species for which we have little direct trait information.

ACKNOWLEDGEMENTS

N.G.S. and L.F. were supported by a US National Science Foundation Advances in Bioinformatics Innovation Grant (DBI 1262475). M.D.W. was supported by a US National Science Foundation grant (EF-1065844). J.-C.S. was supported by the European Research Council (ERC-2012-StG-310886-HIST-FUNC), the Danish Council for Independent Research | Natural Sciences (12-125079), and Aarhus University and Aarhus University Research Foundation under the AU IDEAS programme (via the Center for Informatics Research on Complexity in Ecology, CIRCE). M.B.A. was supported through the Imperial College London’s Grand Challenges in Ecosystems and Environment initiative. M.A.Z. was supported by grant FUNDIVER (MINECO, Spain; CGL2015-69186-C2-2-R). J.A.F.D.-F. was supported by several grants from CNPq. S.N. was supported by the Danish Council for Independent Research – Natural Sciences (10-085056) and the Villum Foundation’s Young Investigator Programme (VKR023456). M.A.R. was supported by a Spanish Ministry of Economy and Competitiveness grant (CGL2013–476 48768-P). D.N.B. thanks ‘Det Frie Forskningsrads Forskerkarriere Program Sapere Aude’. D.N.B. thanks the Danish National Research Foundation for support to the Center for Macroecology, Evolution and Climate.

REFERENCES


**SUPPORTING INFORMATION**

Additional supporting information may be found in the online version of this article at the publisher’s web-site:

Table S1 Pearson correlation coefficients for eastern North America between the deviation of the predicted values from the known value where the deviation is calculated as the predicted subtracted from the observed.

Table S2 Pearson correlation coefficients for Europe between the deviation of the predicted values from the known value where the deviation is calculated as the predicted subtracted from the observed.

**BIOSKETCH**

Nathan G. Swenson is a plant biologist interested in the distribution of biodiversity. His work focuses on patterns of species and functional and phylogenetic diversity through space and time and the mechanisms that underlie them.

Editor: Adam Algar