Biodivers Conserv (2017) 26:2223–2249
DOI 10.1007/s10531-017-1354-4

ORIGINAL PAPER

Macro-scale assessment of demographic and environmental variation within genetically derived evolutionary lineages of eastern hemlock (*Tsuga canadensis*), an imperiled conifer of the eastern United States

Anantha M. Prasad[1] · **Kevin M. Potter**[2,3]

Received: 22 August 2016 / Revised: 2 March 2017 / Accepted: 18 April 2017 /
Published online: 25 April 2017

Abstract Eastern hemlock (*Tsuga canadensis*) occupies a large swath of eastern North America and has historically undergone range expansion and contraction resulting in several genetically separate lineages. This conifer is currently experiencing mortality across most of its range following infestation of a non-native insect. With the goal of better understanding the current and future conservation potential of the species, we evaluate ecological differences among populations within these genetically defined clusters, which were previously inferred using nuclear microsatellite molecular markers from 58 eastern hemlock populations. We sub-divide these clusters into four genetic zones to differentiate putative north-central, north-east and southeast (SE) and southwest evolutionary lineages in eastern hemlock. We use demographic data (relative abundance, mortality, and seedling regeneration) from the Forest Inventory Analysis program in conjunction with environmental data to model how these lineages respond to current and future climatic gradients. Ecologically meaningful relationships are explored in the intraspecific context of hemlock abundance distribution and then related to genetic variation. We also assess hemlock's colonization likelihood via a long distance dispersal model and explore its future genetic and ecological conservation potential by combining the future suitable habitats with colonization likelihoods. Results show that future habitats under climate change will markedly

Communicated by Danna J. Leaman.

Electronic supplementary material The online version of this article (doi:10.1007/s10531-017-1354-4) contains supplementary material, which is available to authorized users.

✉ Anantha M. Prasad
aprasad@fs.fed.us

Kevin M. Potter
kpotter@ncsu.edu

[1] Northern Research Station, USDA Forest Service, Delaware, OH 43065, USA

[2] Department of Forestry and Environmental Resources, North Carolina State University, Raleigh, NC, USA

[3] Southern Research Station, USDA Forest Service, Research Triangle Park, NC 27709, USA

decline for eastern hemlock. The remaining areas with higher habitat quality and colonization potential are confined to the SE, the genetic zone nearest the species' putative glacial refugia, pointing to the need to focus our conservation efforts on this ecologically and genetically important region.

Keywords Genetic variation · Environmental variation · Intraspecific variation · Genetic zones · Evolutionary lineages · Climate change

Introduction

Eastern hemlock (*Tsuga canadensis* [L.] Carr.) is a widespread tree species in the eastern United States and Canada, occupying multiple forest types and preferring loamy, cool, moist, well-drained, and acidic soils throughout its range (Godman and Lancaster 1990). This slow-growing conifer is threatened by an invasive insect pest, hemlock woolly adelgid (HWA, *Adelges tsugae* Annand), which has caused widespread mortality in many regions, so far negatively impacting approximately 50% of the hemlock ecosystems within the United States (McClure et al. 2003; United States Department of Agriculture, Forest Service 2015). Because eastern hemlock plays a vital role in maintaining water quality and is a component of unique floral and faunal assemblages (Heard and Valente 2009; Spaulding and Rieske 2010; Ford and Vose 2007), its declining status is the subject of diverse studies aimed at mitigating the effects of HWA via genetic, biological, silvicultural and chemical intervention (Cheah et al. 2004; Ward et al. 2004; Jetton et al. 2008; Montgomery et al. 2009). Of particular interest is the need to identify resilient and adaptable provenances of hemlock to restore populations degraded by HWA and other human-induced stressors like climate change (Schaberg et al. 2008).

Like most outcrossing, wind-pollinated trees, eastern hemlock is distributed widely in the eastern United States and demonstrates morphological variation at multiple scales (Olson and Nienstaedt 1957; Nienstaedt and Olson 1961), including two distinct types with respect to growth rate, morphology, and macroclimatic variation (Kessell 1979). Linking geographically structured morphological variation of eastern hemlock to genetic variation to unravel patterns of phenotypic traits is a challenge because both genomic and epigenomic factors interact with the environment to influence the biological phenotype (Lou 2014). For example, population differentiation can be caused by environmentally induced phenotypic plasticity responses and by genetically induced adaptation effects, as well as by an interaction between the two (Andrew et al. 2010). Non-additive genotype by environment (GxE) differentiation among populations at a macroscale can be reflected by within and among population means of demographic, genetic, and environmental variables (Zenni et al. 2014). Also, in a conservation setting, GxE interactions can lead to poor seed-choices for restoration and can complicate the design of provenance studies—but can also offer opportunity to select specific genotypes for specific environments (Murillo 2001; Baltunis et al. 2010).

Because of these considerations, the conservation of eastern hemlock requires an understanding of range-wide population genetic structure including distribution of genetic variation within and among populations. An in-depth study tracking these factors was previously performed using 13 highly polymorphic nuclear microsatellite loci on 60 populations throughout its range (Potter et al. 2012). This study used spatially explicit

Bayesian clustering to identify the potential existence of four distinct evolutionary lineages, each putatively associated with a different Pleistocene glacial refuge in the southeastern US, from which followed a post-glacial movement to the Northeast and the Great Lakes region. The existence of these lineages was also supported by other studies using allozyme markers (Potter et al. 2008) and seven haploid chloroplast DNA loci (Lemieux et al. 2011). Information from such within-species molecular-marker derived genetic lineages has the potential to provide more realistic predictions of habitat suitability than do projections conducted across the entirety of a species (Gotelli and Stanton-Geddes 2015). For example, delineation of potential climate niches for within-species evolutionary groups can provide managers insight into managing genetic variability to encourage the best fit on local landscapes under expected future climatic conditions (Shinneman et al. 2016).

Our main purpose in this paper is to further the knowledge gained by genetic studies by combining it with macro-scale intraspecific demographic and environmental models (Prasad 2015) to better understand the conservation potential of the species by taking both genetic and environmental variation into account (Alsos et al. 2012). Specifically, we aim to determine the degree to which different evolutionary lineages of eastern hemlock are associated with different environmental factors, and to then use this knowledge to prioritize populations for both in situ and ex situ conservation and for restoration. In the process, we examine various indices that highlight the diversity of these evolutionary lineages along genetic, ecological, and demographic gradients (Gotelli and Stanton-Geddes 2015). In our study, we are not focusing on the phenotypic differences along environmental gradients (because we can't assess these with neutral genetic markers like microsatellites) but rather the *abundance differences* (treated as rough surrogates for fitness—see Nagaraju et al. 2013) correlated with *suitable habitats*.

Methods

The current effort builds on two aspects of the previous study (Potter et al. 2012)—(a) the genetic indices of population diversity and differentiation, and (b) population clusters depicting potential evolutionary lineages. We then add an ecological dimension by evaluating the role of environmental factors to assess the abundance of eastern hemlock within genetic lineages. In the process, we divided the abundance distribution of hemlock into four genetically derived and geographically separated intraspecific regions based on the genetically inferred population clusters to conduct further environmental and demographic analysis using information from U.S. Forest Inventory and Analysis plots, which are sampled at a much higher spatial intensity (Smith 2002; Bechtold and Scott 2005; Woudenberg et al. 2010). We looked at demographic and genetic variation among and within the zones and constructed decision-tree based ensemble models to identify differences in abundance patterns among and within these zones under current and future climate conditions (Prasad et al. 2016). We also evaluated realistic future migration and conservation possibilities via a spatially explicit model which relies on historical migration rates to simulate current colonization likelihoods (Prasad et al. 2013). All statistical analyses, and most GIS analyses, were conducted in R 3.3.2 (R Core Team 2016), and maps analyzed and produced in Qgis 2.18.3 (Quantum GIS Development Team 2016) and ArcGIS 10.3.1 (ESRI 2015).

Derivation of genetic variables

Foliage samples were collected from 60 eastern hemlock populations (consisting of 1180 trees) as described in Potter et al. (2012), which represented most of the hemlock range in the eastern United States; these samples subsequently were genotyped for 13 highly polymorphic microsatellite markers developed for eastern hemlock (Shamblin et al. 2008) and Carolina hemlock (Josserand et al. 2008). All but two of the populations encompassed 20 sampled trees (the two exceptions were very small disjunct populations with 15 and 5 samples), with sampled trees spaced at least 100 m apart to avoid the sampling of neighbors that might be closely related. All stands were natural, not planted.

Our first objective was to evaluate how well both genetic diversity indices (derived from the molecular markers) and the environmental variables can predict the abundance of hemlock. This investigation sheds light on whether population differentiation measures, together with climatic gradients, can explain the abundance of hemlock to any significant degree. To accomplish this, the coordinates of 58 eastern hemlock populations (consisting of $\sim$1100 trees in the United States) were buffered at a distance of 9 km (using the function gBuffer in rgeos package in R) to ensure non-intersection with other buffers. The buffers around the sampled populations were used to extract abundance, climatic, topographic, and edaphic variables, and a predictive model then was developed to test the importance of the genetic and environmental variables in predicting hemlock abundance within the buffer.

Response variables

Abundance was estimated by averaging the average diameter of the hemlock stems on the FIA plots because this dominance metric represents a species-specific measure of absolute abundance in contrast to relativized measures of abundance like importance values, which incorporate interactions of other species in the plot (Pulliam 2000; Soberon and Peterson 2005; Godsoe 2010). We also derived percent mortality (PM, based on whether the sampled tree was alive or dead, averaged over the plot) and seedling count (SC, number of seedlings <1 inch in diameter and at least 12 inches tall, averaged over the plot) from the FIA data for the four zones. These three demographic measures together provide a rough measure of the "overall fitness" of the species (Nagaraju et al. 2013).

Models

We developed a model to test how well molecular marker data and environmental variables can predict hemlock abundance. This model was based on a decision tree based ensemble method called RandomForest (RF) (Breiman 2001). RF has been used extensively to make reliable predictions for multidimensional data that are nonlinear and exhibit interactions (Lawler et al. 2006; Prasad et al. 2006; Cutler et al. 2007).

However, in order to predict the FIA-derived abundance of hemlock based on environmental data for the entire range as well as the genetic clusters, a more sophisticated multi-model ensemble (henceforth, MME) approach was used in which four decision-tree based models were combined to arrive at an average prediction. The four models were computed in R and reflect robust statistical learning algorithms with a proven track record: RandomForest (randomForest package in R) (Breiman 2001), extremely randomized trees (extraTrees package in R) (Geurts et al. 2006), stochastic gradient boosting (gbm package

in R) (Friedman 2002), and regularized gradient boosting (xgboost in R) (Chen and He 2015; Chen and Guestrin 2016). We used repetitive resampling via the caret package in R to optimize the tuning parameters of these four predictive models (Kuhn 2008).

The random forest approach is a very well-known solution to the overfitting problem in individual decision trees by building a forest of decision trees from bootstrapped data and further de-correlating it by randomly choosing a subset of predictors to reduce variance and maintain a favorable bias-variance ratio. The extremely randomized forest (ERF) approach takes this one step further by randomizing the feature selection (split and threshold). While RF chooses the 'best' split at each node, ERF creates p splits randomly (i.e., independently of the response variable, p being the subset of predictors randomly chosen in each node) and then the split with the best gain (mean squared error for regression) is chosen. The rationale for ERF is that by randomizing the selection of split, the variance is reduced even further compared to the RF. However, ERF typically uses the entire learning sample instead of the bootstrapped sample to grow the trees in order to reduce bias (Geurts et al. 2006). The R package gbm implements the well-known stochastic boosting algorithm (Friedman 2002). Boosting is a method of iteratively converting weak learners to a strong one (in our case using decision trees). Boosting initially builds a base learner after examining the data and then reweights observations that have higher errors. Stochastic gradient boosting reduces variance by shrinkage (regularization) and stochasticity in which each newer iteration of the decision tree model learns from the previous one (minimizing residuals weighted by the previous model's errors based on a shrinkage parameter). Xgboost is another slightly different approach to boosting which differs in the way regularization is implemented, improving on its ability to control overfitting (Chen and Guestrin 2016).

The rationale behind using an ensemble of models is to leverage the strengths of these powerful learning algorithms (whose strengths and weaknesses depend on the training set) to build an average prediction model that better reflects the consensus (Jones and Cheung 2015; Martre et al. 2015). Because demographic, genetic, and environmental data have inherent interactions and nonlinearity, the MME approach typically performs better than an individual model due to the averaging out of errors (Zhang et al. 2015). Furthermore the MME approach further resists overfitting the data and maintains a favorable variance-bias ratio (Hastie et al. 2009), which is critical in minimizing the prediction error while extrapolating current prediction to future climates.

Genetic-environmental analysis

We first wanted to assess the variance explained, and also the relative importance of genetic and environmental variables in predicting abundance across the 58 eastern hemlock populations (see the "Derivation of genetic variables" section). In order to do that we used three versions of the RandomForest model (Tables 1, 2) using three different sets of variables: (a) climate variables only (growing season aridity index [GSAI], January temperatures [TJAN], May–September temperature [TMAYSEP] and May–September precipitation [PMAYSEP]); (b) genetic and HWA presence variables only (mean alleles per locus [A], unique alleles [A_U], effective number of alleles [A_E], percent polymorphic loci [P_P], observed heterozygosity [H_O], expected heterozygosity [H_E], mean pairwise chord distance from all other populations [MEAN_D_C], inbreeding coefficient [F_{IS}], proportion of the genetic variance contained in the population relative to the total genetic variance across all possible pairwise population comparisons [MEAN_F_{ST}], and whether the population was in a county infested by hemlock wooly adelgid [HWA]) by 2010 (Potter et al.

Table 1 The predictors used in the local and global model with sources

Climate[a]	
TJAN/tjan	Mean January temperature (°C)
TMAYSEP/tmaysep	Mean May–September temperature (°C)
PMAYSEP/pmaysep	May–September precipitation (mm)
GSAI/gsai	Growing season aridity index (May–September)
Elevation[b]	
ELVMAX/elvmax	Maximum elevation (m)
ELVSD/elvsd	Elevation standard deviation
Soil[c]	
CLAY/clay	Percent clay (<0.002 mm)
OM/om	Organic matter content (% by weight)
pH/ph	Soil pH
SIEVE10/sieve10	Percent passing sieve no. 10 (coarse)
SIEVE200/sieve200	Percent passing sieve no. 200 (fine)

[a] Data for the period 1981–2010 from (PRISM Climate Group), GCM data from NEX-DCP30 (Thrasher et al. 2013)

[b] From the NASA's Shuttle Radar Topography Mission provided at a resolution of 3″ (Guth 2006). We calculated the maximum value and standard deviation at 10 and 20 km^2 grids

[c] From Natural Resource Conservation Service's County Soil Survey Geographic (SSURGO) database (NRCS 2009). Data was processed by (Peters et al. 2013) and aggregated to 10 and 20 km^2 grids

Table 2 Explanation of the acronyms for the genetic indices

A	Mean number of alleles present in each population across the 13 microsatellite loci
A_U	Number of unique (private) alleles found in the population (i.e., occurring only in that population)
P_P	Percent polymorphic loci (percent of the 13 microsatellite loci for the population that have more than one allele)
H_O	Observed heterozygosity (observed proportion of heterozygotes [loci for which there are two different alleles], averaged over loci)
H_E	Expected heterozygosity (heterozygosity expected if the population is in Hardy–Weinberg equilibrium, a null hypothesis for testing whether populations have undergone certain genetic processes)
A_E	Effective number of alleles (based on the expected heterozygosity)
MEAN_F_{ST}	Mean pairwise F_{ST} value between a population and all other populations. F_{ST} is the proportion of the total genetic variance contained in a subpopulation relative to the total genetic variance. (Values can range from 0 to 1. High F_{ST} implies a considerable degree of differentiation among populations)
F_{IS}	Inbreeding coefficient (proportion of the genetic variance in the subpopulation contained in an individual). High F_{IS} indicates a considerable degree of inbreeding
MEAN_D_C	Mean pairwise genetic chord distance for a population with all other 59 populations. High values indicate high levels of genetic differentiation
HWA	Whether hemlock woody adelgid present (0 | 1)

These were derived from Potter et al. (2012)

2012); and (c) both climate and genetic variables. The climate variables were chosen to reflect the most parsimonious set that best explained variance based on previous screening experience (Prasad et al. 2016). The genetic variables were chosen to represent a wide variety of metrics that describe within-population genetic diversity and between-population genetic differentiation. They were all standard measures of genetic diversity based on allele size calls at the 13 polymorphic microsatellite loci, calculated using FSTAT version 2.9.3.2 (Goudet 1995), Arlequin 3.0 (Excoffier et al. 2005), and PHYLIP 3.6 (Felsenstein 2005). For more details, see Table 2.

Genetic zones

Another objective of our analyses was to evaluate the demographic, genetic, and environmental differences among the geographic zones determined based on the molecular-marker-inferred evolutionary lineages (Potter et al. 2012). TESS 2.3.1 (Chen et al. 2007), a spatially explicit Bayesian clustering model, inferred the number and composition of genetic clusters in eastern hemlock using the microsatellite genotypes for each of the ~1180 individual trees. This clustering model generates a large number of Markov Chain Monte Carlo simulations for multiple runs across a set of possible cluster numbers (K = 2

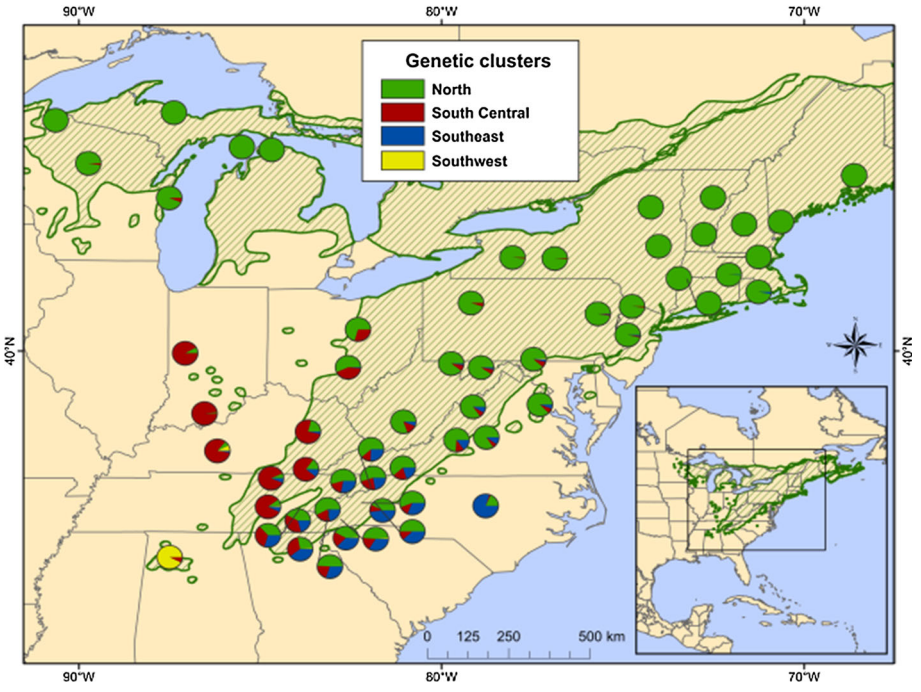

Fig. 1 The existence of distinct genetic clusters was consistent with range contraction and expansion in eastern hemlock as inferred by TESS Bayesian clustering approach (Chen et al. 2007). The figure shows the proportion within each eastern hemlock population of inferred ancestry from the genetic clusters (one for each *color*) suggesting that species was mostly confined to three or four separate glacial refuges in the Southeastern United States—since these clusters contain the all the ancestry except the Southwest cluster, which is rather unique (Potter et al. 2012)

to K = 12 in this case), with the mean deviance information criterion (DIC) results across runs used to select the most likely K. In this case, TESS inferred the existence of four genetic clusters, with all four detected in the southern part of the range but only one common in the north (Fig. 1). Informed by the proportion of these genetic clusters within each of 58 eastern hemlock populations in the United States and by the distribution of FIA plots, we delineated four genetic zones for further analyses (Fig. 2). The zones were delineated to match the extent of the TESS clusters to the degree possible (along with density of FIA plots and the microsatellite-populations—Fig. 2), though it is important to note that admixture among genetic clusters within populations made the delineations somewhat subjective in some cases. For example, because the original clustering analysis suggested high levels of admixture between the North and the Southeast clusters (Potter et al. 2012), we created analysis zones that widely separated an area of highest Southeast cluster prevalence (the SE zone) from areas consisting almost entirely of the North genetic cluster. These North genetic cluster areas were divided into two zones, one in the Northeastern states (the NE zone) and one in the Northcentral states of Michigan and Wisconsin (the NC zone), because of the wide geographic separation of these areas. Areas of highest prevalence of the South-Central genetic cluster were separated into a south-western (SW) analysis zone. The isolated Southwest cluster in Alabama in the original TESS analysis (Fig. 2) was not included in our analyses due to the paucity of FIA plots containing hemlock in the area.

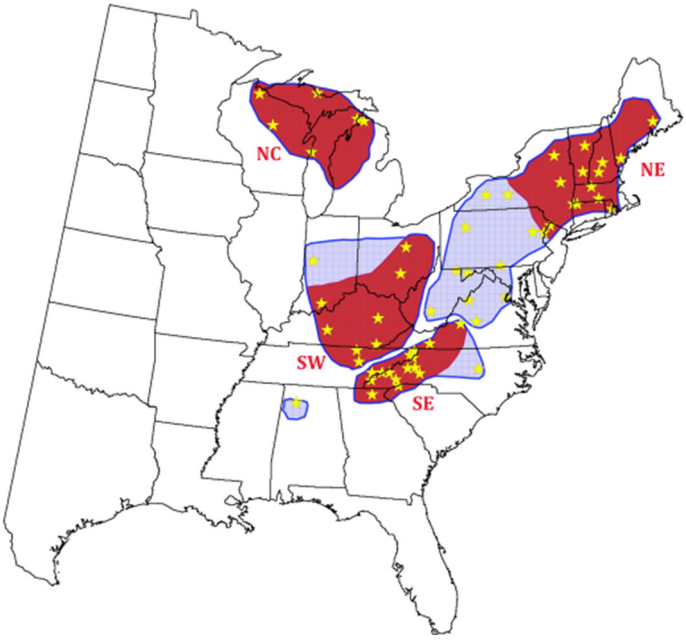

Fig. 2 The original polygonized TESS clusters (*blue outline*) are subset to Northcentral (NC), Northeast (NE), Southwest (SW) and Southeast (SE) zones to reflect FIA plots density and geographic separation (the *red* areas). We discount the original extreme Southwest outlier cluster because of the small size and few FIA plots. The *yellow stars* are the sampled microsatellite populations

Genetic-zone analysis

The differences in demographic, genetic, and environmental variables among the genetic zones were compared using Tukey's honest significance test (Tukey HSD) in conjunction with ANOVA to find means that are significantly different from each other. The observations from the zones are assumed to be independent within and among the zones, normally distributed, with homogeneity of variance. We also assume spatial independence among the zones because the populations are not highly genetically related and there is wide separation between the zones except for the SE and SW.

For a comparative model analysis of these genetic zones, we created an abundance raster (average diameter rescaled to 0–100) at 20 km resolution based on FIA plots and assembled environmental variables (climate, soil, and topographic) (Table 1). The global model involved prediction of eastern hemlock across its entire range for current and future climates. Because of the importance of the SE zone as putative refugia, it was modelled separately using a buffer of 200 km to test how this zone might behave under future climates. Both the global and local SE models were based on the MME approach described previously.

To assess how the climatic space varied with abundance among the genetic zones, we also prepared niche maps that display average diameter based on growing season precipitation and January temperatures (Online Supplementary Figs. A2, A3). We also evaluated how hemlock fared when combining future predicted habitats of the global model with colonization probabilities calculated from a spatially explicit model assuming an optimistic migration rate of 50 km/century, which is at the high end of the estimated migration rates for some eastern US tree species (McLachlan and Clark 2004; Yansa 2006; Svenning and Skov 2007; Dobrowski 2011).

Global and local models

For the global model and the local models, we drew on our previous experience with ensemble models (Prasad et al. 2006, 2016; Iverson et al. 2008). We screened the predictor variables that explained most of the variation in the model and chose a parsimonious set of 11 variables (Tables 1, 2). We found these variables to be the most relevant and ecologically important for explaining the variation in abundance at a resolution of 20 km, and were least affected by collinearity. The selected climatic variables address two key temperature-based constraints to tree species ranges—growing season temperature (May–September) and the temperature of the coldest month (January). An aridity index, which is the ratio of May to September precipitation to the potential evapotranspiration index (Thornthwaite and Mather 1957), along with growing season precipitation, adequately captures the moisture constraint (Bucklin et al. 2015; Pederson et al. 2015). We retained two elevation-based variables (maximum elevation and elevation standard deviation averaged over the 20 km pixels) because these variables are useful in distinguishing between low and high elevation areas and between those populations that prefer habitats with rugged or gentle terrain. The current climate data were obtained for the period 1981–2010 from the PRISM climate group (PRISM 2017; Daly et al. 2008). For future climate, we used "harsh" and "mild" scenarios describing predicted future greenhouse gas emissions. We chose the Hadley Global Environment Model (HAD, Jones et al. 2011), and a representative greenhouse gas emissions concentration pathway (RCP 8.5) (Meinshausen et al. 2011) as the "harsh" scenario because it depicted the largest increase in temperature

illustrating a scenario of business as usual; it was chosen to estimate the upper end of potential changes in climate. For the "mild" scenario, we chose the Community Climate Systems Model (CCSM4—Gent et al. 2011) under the representation concentration pathway of RCP 4.5 to illustrate the utility of greenhouse gas mitigation efforts.

We also derived a mapped decision tree by extracting the tree rules based on the 20 km resolution decision tree (pruned to nine nodes based on cross-validation) and mapping them to a 10 km abundance map. This helps us understand the predictor rules that drive the distribution of eastern hemlock at a higher resolution based on the rules at a coarser resolution (Prasad 2015).

Dispersal model

A spatial simulation model (SHIFT) was used calculate colonization likelihoods based on dispersal under current fragmented landscapes. SHIFT runs at a 1 km resolution and simulates dispersal via a fat-tailed inverse power function and calculates future colonization likelihoods based on historical tree migration rates and current fragmented landscapes (Schwartz 1993; Iverson et al. 2004; Prasad et al. 2013). The two main parameters affecting the migration are the calibration constant (C) that simulates the historical migration and the generation time of eastern hemlock (with a selected value of three to span approximately 100 years to coincide roughly with future climate habitat predictions obtained from the RandomForest model).

The range boundaries for eastern hemlock are derived from the latest FIA annualized data. A technique based on ESRI's (2015) "delineate built-up areas" tool and the aggregated abundance of FIA plots was used to generate an estimate of the 'core' species boundary (Peters et al. 2013). This 'core' boundary, based on the current species distribution, was used to depict the source region across which propagules are launched in the SHIFT model. Forested cells from within the core species boundary ("source") are modeled to contribute propagules into forested cells outside the boundary ("sink"). The source strength is a function of both the propagule production and dispersal capability across the boundary. Locations with higher species abundance and closer proximity to the boundary will create the highest colonization probabilities near the current boundary. Sinks are forests or woodlots of varying degrees of fragmentation of forest (defined as 1-km pixels with at least 10% forest in the United States Geological Survey's 2006 National Land Cover Data) that provide possible locations for the propagules to colonize under current and future climates. The predicted suitable habitats of eastern hemlock provide the future ($\sim$2100) suitability of the sink habitats based on projected abundance.

SHIFT is calibrated through trial runs to achieve migration rates ranging from approximately $<$25 to $>$100 km per century (depending on criteria of model runs) under high forest availability (80% cover, approximately representing the landscape under pre-settlement conditions), but with the current level of species abundance. We used a migration rate of 50 km/century for eastern hemlock to test how this optimistic assumption would affect the colonization of suitable habitats.

We classified the predictions of suitable habitats (from the abundance models) and the colonization likelihoods (from the SHIFT model) to assess combinations of habitat-quality (HQ) classes (Low, Medium and High) and colonization-likelihood (CL) classes (Low, Medium and High). The resulting classification results in nine combinations of (HQ–CL) classes (Low–Low, Low–Med, Low–High, Med–Low, Med–Med, Med–High, High–Low, High–Med and High–High) which define the degree of suitability for future conservation

and restoration efforts by combining future suitable habitats with the colonization potential of these habitats.

Results

Population genetic analysis

Our first objective was to investigate geographic differences in the genetic diversity indices previously calculated from the 58 populations throughout the range of eastern hemlock (Fig. 1) in the eastern US. This evaluation uses range-wide population-level genetic indices derived from microsatellite data to assess whether they can account for differential abundance of hemlock by itself and in association with climatic variables. The assumption here is that the buffered FIA plot locations (see "Methods" section) can be used to construct models to assess eastern hemlock abundance using (a) climate variables, (b) genetic variables and (c) both climate and genetic variables (Supplementary Material Appendix 1, Fig. A1). Climate alone provides a good fit, explaining 45% of the variance in abundance, and with GSAI, TMAYSEP and TJAN providing most of the explanatory power (Supplementary Material Appendix 1, Fig. A1a). Interestingly, using only genetic variables explained an even higher proportion of the abundance variance (54%) with MEAN_D_C, MEAN_F_{ST}, H_E, A_E and A explaining bulk of the variation (Supplementary Material Appendix 1, Fig. A1b). When both climate and genetic variables were combined, the variance explained increased to 61% with MEAN_D_C and GSAI explaining a major portion (40%) of it (Supplementary Material Appendix 1, Fig. A1c).

Differences among genetic zones

The differences in demographic and environmental variables among the four genetic zones—north central (NC), northeast (NE), southeast (SE) and southwest (SW)—(see Fig. 2) that were derived from the original TESS genetic clusters are shown in Tables 3 and 4 (see "Methods" section). The mean values of demographic variables (average diameter—AD, percent mortality—PM, seedling count—SC) vary among the zones

Table 3 The differences in the mean values of the demographic variables average diameter (AD), percent mortality (PM) and seedling count (SC) among the genetic zones

Tess clusters	No. of FIA plots	AD[a]*	PM[b]*	SC[c]*
NC	1035	10.6	7.8	2.1
NE	1943	8.6	4.7	4.2
SE	464	7.1	4.3	1.5
SW	215	7.5	6.6	2.9

The asterisk shows that the mean was significant at $p \leq 0.05$. The significant differences in mean between zone combinations are also shown

NC Northcentral, *NE* Northeastern, *SE* Southeastern, *SW* Southwestern

[a] (NE–NC, SE–NC, SW–NC, SE–NE, SW–NE)

[b] (NE–NC, SE–NC)

[c] (NE–NC, SE–NE)

Table 4 The differences in the mean values for the predictor (explanatory) variables among the genetic zones

Tess clusters	No. of FIA plots	GSAI	TJAN	TMAYSEP	PMAYSEP	ELVMAX	ELVSD	CLAY	OM	PH	SIEVE10	SIEVE200
NC	1035	0.87	−9.41	15.82	436.3	365.2	5.0	11.3	17.6	6.2	86.8	38.1
NE	1943	1.03	−7.29	16.66	521.8	324.6	13.7	10.2	8.2	5.5	75.3	41.1
SE	464	1.15	1.84	19.88	619.8	823.6	29.1	18.1	1.4	5.1	75.7	42.4
SW	215	0.96	1.02	20.99	556.1	469.8	23.7	25.6	1.3	5.1	74.0	52.1

All were significant at $p \leq 0.05$

(significance tested via both ANOVA and Kruskal–Wallis test), which are further analyzed using the Tukey HSD test to assess which zone combinations of the means are significantly different (Table 3). AD is significantly different between all pairs of zones except SW–SE. All three variables were significantly different between NE and NC, while the PM mean different is significant for SE and NC, and mean SC is significantly different between SE and NE. The differences among means for all environmental variables are also significant (Table 4). The NE and NC zones are fairly similar with respect to climatic variables, while the SE and the SW zones are warmer (TJAN and TMAYSEP), in line with the latitude of the zones. The SE zone has the highest elevation (ELVMAX) and the most rugged terrain (ELVSD). Among the edaphic variables, organic matter content (OM) is highest for NC and quite high for NE compared to the low values for SE and SW; however, CLAY is much higher in the SE and SW compared to NE and NC. The soils are more basic in the NC with pH of 6.2 compared to much lower values in the rest of the zones.

We also compared the genetic diversity metrics and HWA presence across these zones to assess evidence of population differentiation, although we had to rely on the much sparser subset of the 58 hemlock population locations sampled for microsatellite analysis (Table 5). Again, significant differences exist for most of these variables among the genetic zones, except for A_U, F_{IS} and HWA. Again, the Tukey HSD test was used after ANOVA to assess which zone combinations are significant for these variables (Table 5). Worth noting is that the SE zone, which may be nearest to the location of one or more of eastern hemlock's putative glacial refugia, showed the highest values of A, P_P, H_E and A_E. Also, as expected, populations within the two most widely geographically separated zones, the SW and NC, had the highest mean level of pairwise differentiation with all other populations (Mean F_{ST} and Mean D_C).

It is clear from the above analyses that the distribution of hemlock shows patterns of genetic and environmental differentiation that are significant among these four genetic zones, which paves the way for a more robust model-based analysis of hemlock abundance among these zones along environmental gradients.

The global abundance model

The global (entire eastern US distribution) MME approach (Fig. 3) shows overall decrease in suitability of eastern hemlock by ~2100—especially for the harsh climate scenario (the overall R-square was 0.53). The boosting models (gbm and xgboost packages in R) had ELVSD (21%), TMAYSEP (20%), TJAN (16%) and CLAY (10%) as the four most important variables while the randomization based models (RandomForest and extraTrees packages in R) had TMAYSEP (23%), TJAN (16%), ELVSD (12%) and GSAI (9%) as the four most important variables. Both sets of analyses highlight that temperature is the most influential predictor overall (36% for boosting models and 48% for random forest models). ELVSD was important too, underscoring that hemlock is more likely to grow in the cool, moist locations found in the cooler regions of topographically rugged terrain. Because of the importance of climate, and especially temperature, the distribution under future climates shows decreases in suitable habitat for eastern hemlock under both the harsh (HAD RCP8.5) scenario and mild (CCSM4 RCP4.5) scenario, albeit less so under the mild scenario (Fig. 3). The habitat suitability also decreases inside the range of current hemlock distribution. The future high quality suitable habitats are mainly confined to the northeast and some portions of the southeast for the mild climate scenario, while the best ones are only in the "LowMedium" class for the harsh scenario. While global models are good at assessing the overall pattern of habitat changes for the entire distribution of eastern

Table 5 The mean value of the genetic indices for the four genetic zones

Tess clusters	No. of populations	A*	AU	PP*	HO*	HE*	AE*	FIS	MEAN_FST*	MEAN_DC*	HWA
NC	6	4.80	0.83	93.58	0.53	0.56	2.26	0.04	0.0560	0.0425	1.0
NE	15	5.10	0.33	98.97	0.57	0.58	2.39	0.04	0.0475	0.0378	1.7
SE	15	5.30	0.20	99.49	0.55	0.61	2.55	0.09	0.0482	0.0415	1.9
SW	8	4.43	0.38	96.15	0.48	0.53	2.16	0.09	0.1074	0.0569	1.5

See Table 2 for the acronyms. The asterisk indicates that the mean was significant at $p \leq 0.05$. We also show cluster-combinations for which the means were significant

NC Northcentral, *NE* Northeastern, *SE* Southeastern, *SW* Southwestern

A: SW–NE, SW–SE, PP: NE–NC, SE–NC, HO: SW–NE, SW–SE, SW–SE, HE: SW–NE, SW–SE, AE: SE–NC, SW–NE, SW–SE, MEAN_FST: SW–NC, SW–NE, SW–SE, MEAN_DC: SW–NC, SW–NE, SW–SE

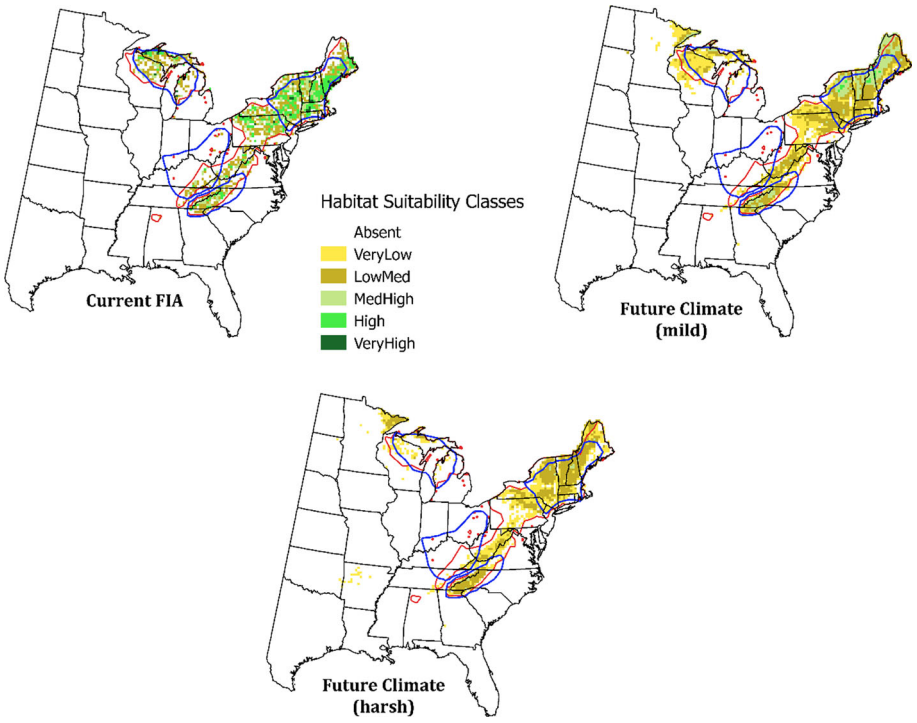

Fig. 3 The global multi-model ensemble outputs for the entire eastern hemlock habitats using Forest Inventory Analysis (FIA) plots aggregated to 20 km. *Top left* is actual FIA based distribution, *top right* is future mild climate habitats and *bottom* is the future harsh climate habitats (~2100). The high and the very high classes are absent under future climate scenarios pointing to a decrease in abundance under future climates

hemlock abundance, it is useful to assess how the MME approach for the buffered SE zone would differ from the global pattern. We are most interested in the SE zone in spite of the fact that the Northeast region has more suitable habitats. This is because it is very likely that the Pleistocene glacial refuges were in the Southeastern US based on the distributional pattern of the genetic clusters (Fig. 1) and therefore this region qualifies to be a main focus for genetic conservation efforts. The model output for the SE (overall R-square of 0.56) depicts movement of the suitable habitats towards the northeast for both the harsh and the mild scenario, although the habitat quality falls mostly under the very low and low-medium suitability classes (Fig. 4). The mild scenario also shows some locations east of the SE zone, albeit of very low suitability. ELVMAX (35%), ELVSD (16%), TJAN (14%) and TMAYSEP (9%), PMAYSEP (5%) are the most important variables that together explain 79% of the total variation for combined the boosted and randomized models. Topography (ELVMAX and ELVSD) becomes very important in the SE zone, accounting for 51% of the total variation.

The mapped decision tree (Fig. 5) shows the distribution of abundance based on the predictor rules of a single decision tree pruned to nine nodes based on cross-validation. This is useful for understanding patterns of differentiation of abundance in the predictor space. The predominant abundance distribution in the NE zone is defined by node 9 (magenta) which has the highest average diameter of 12.34 inches. The NC zone is

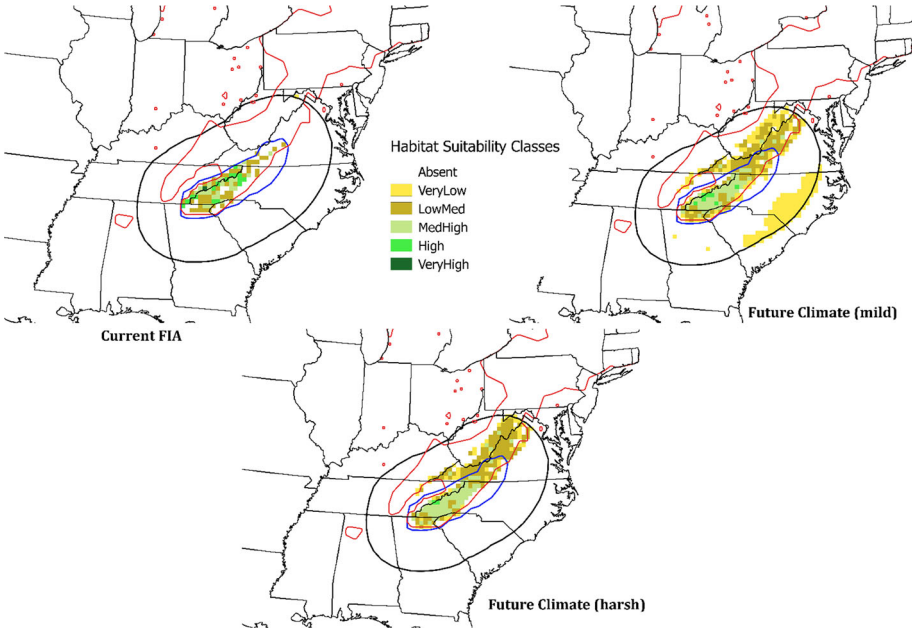

Fig. 4 The model ensemble results for the Southeast genetic zone, buffered to a distance of 200 km. The *red line* shows the current FIA distribution, the *blue line* the SE zone and the *black thicker line* the limit of the buffer

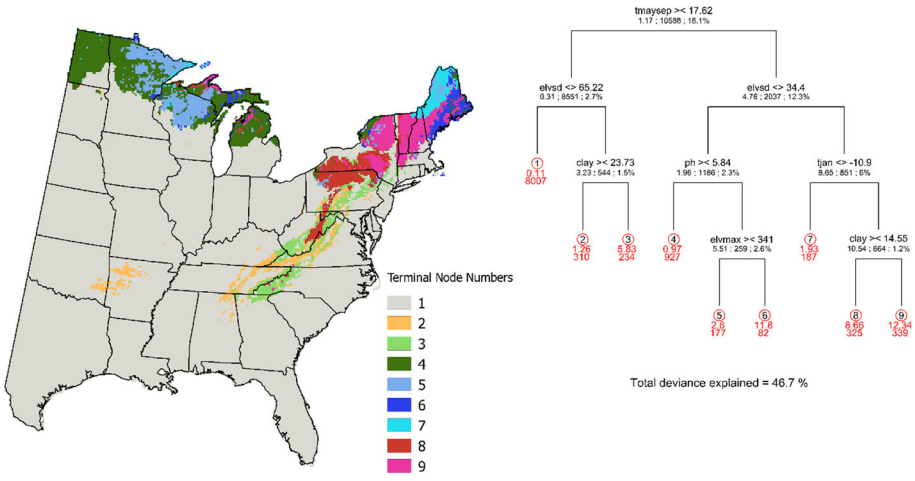

Fig. 5 Mapped decision tree. The terminal node numbers (*numbers inside rounded circles*) in the decision tree correspond to the classes in the map legend. In order to understand what predictor rules are driving the distribution of hemlock, we trace the decision tree rules from the terminal node. For example, the rule that corresponds to the highest abundance predicted (terminal node number 9 (*magenta*), average diameter of 12.34) has the rule TMAYSEP < 17.62 C and ELVSD > 34.4 and TJAN > -10.9 C and CLAY $< 14.55\%$. The numbers below the terminal nodes correspond to the average response (in this case average diameter) and the number below that is the number of cells in the map (339 for class 9)

dominated by node 5 (light blue, with average diameter of 2.6 inches) and node 4 (dark green, with average diameter of only 0.97 inch). The predictor rules defining these distributions can be traced within the decision tree (Fig. 5). The area with highest abundance (12.34 inches in average diameter) tends to be on the cooler side (TMAYSEP < 17.62 C), but not too cold (TJAN > -10.9 C)—with rugged terrain (ELVSD > 34.4) and not too clayey soils (CLAY $< 14.55\%$).

In order to understand how the variation in temperature and precipitation (TJAN and PMAYSEP) affects the abundance across the zones, we constructed what we term the niche diagrams (Supplementary Material Appendix 1, Figs. A2, A3). Abundance varies within the range of climatic possibilities across these zones with the NC zone being the driest and coldest compared to the SE zone, which is the warmest and wettest. The more detailed versions of these diagrams (Supplementary material Appendix 1, Fig. A3) depict the variations with greater clarity, highlighting areas of highest abundance (the narrow bands) for these climatic variables. For example, within the SE genetic zone, there is a large area of higher abundance when January temperature is low (-2 to 0 °C approximately) and the May–September precipitation is between 750 and 850 mm.

Future habitat quality and colonization potential

Whilst the analysis so far shows how the abundance of hemlock decreases, especially under the future "harsh" climate, this in effect depicts changes only in suitable habitats without considering the migration potential of the species. By running the SHIFT model for the east-wide distribution of hemlock, we were able to calculate the colonization likelihoods of suitable habitats by 2100. Even under an optimistic migration rate of 50 km/century, the colonization likelihoods drop off rapidly with distance (Fig. 6). Running SHIFT for the entire eastern hemlock distribution (Fig. 6, left) and for just the genetic zones (Fig. 6, right) reveals some interesting patterns. These maps show that most of the movement (including gene exchange between genetic zones) outside the current range of hemlock happens between the SE and the SW zones; however, there also is potential for

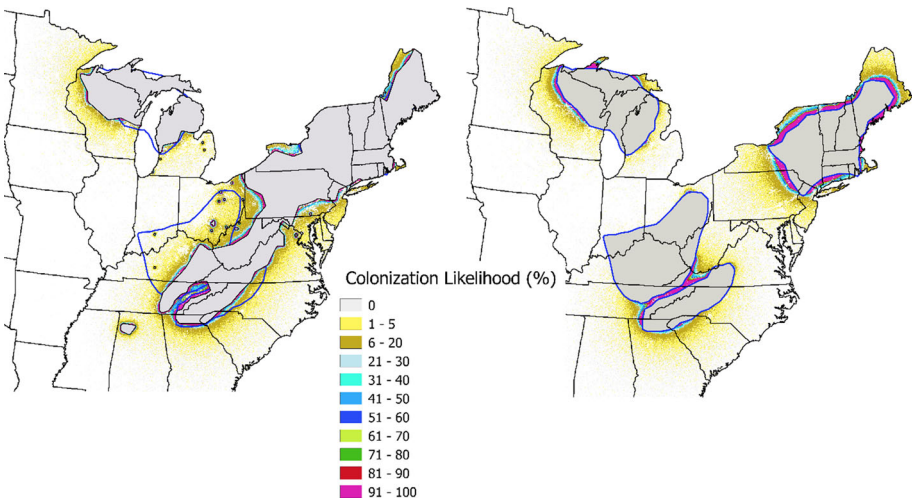

Fig. 6 The SHIFT model showing colonization likelihoods in ~2100 for eastern hemlock. The entire eastern hemlock distribution (*left*) and only the genetic zones (*right*)

enhanced exchange of genes between the NE region and the northeastern segment of the SW zone in the future.

In order to assess the overall future scenarios of eastern hemlock considering both habitat quality and colonization potential of these habitats, we intersected and reclassified the colonization likelihoods and the suitable habitats into three classes, and recombined them to obtain nine classes of habitat-quality and colonization likelihood (HQ–CL) for the entire east-wide distribution and for the genetic zones (Fig. 7). Even though the overall suitability is quite limited, regions of higher suitability exist between the SE and SW genetic bands. We again emphasize that the SE zone has the maximum diversity of inferred ancestral groups and therefore is highly valuable from a seed conservation perspective. Currently however, this region is already occupied by eastern hemlock, although exploitable niches may exist in these regions, particularly those with higher potential for gene-exchange.

Discussion

Preserving the genetic diversity of the threatened eastern hemlock, in situ and ex situ via sound conservation and management decisions, requires an integrated management approach that combines information about environmental and genetic lineages. In this study, we applied intraspecific demographic and environmental models to assess the degree

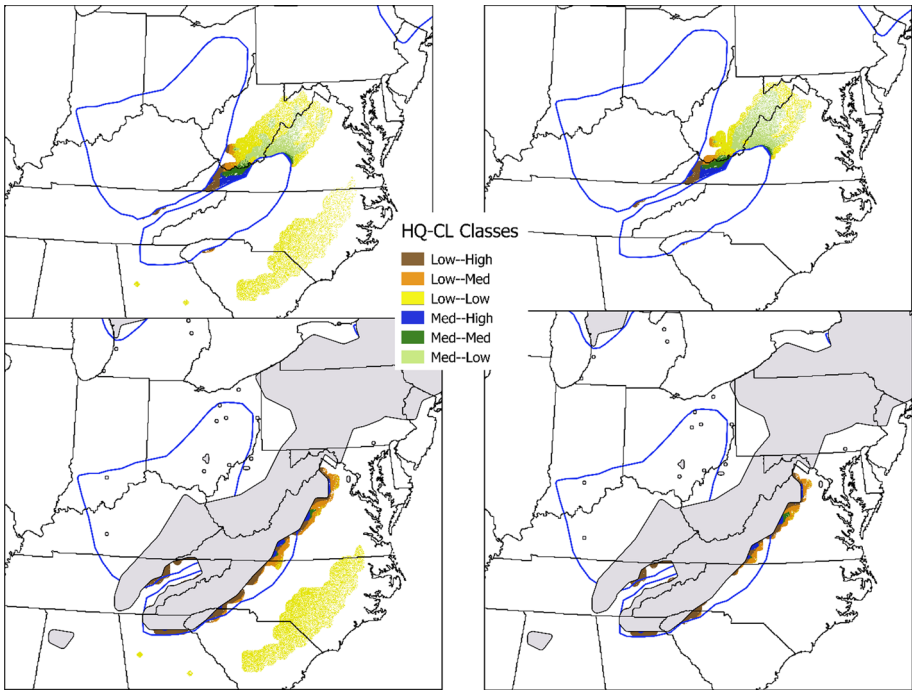

Fig. 7 The future habitats (HQ) for the "mild" (*left column*) and "harsh" (*right column*) scenarios are combined with colonization likelihoods (CL) of eastern hemlock (~2100) to illustrate the colonization likelihood of future suitable habitats (only 5 of 9 combined HQ–CL classes are present) for the entire eastern hemlock distribution (*bottom two*) and just the four genetic zones (*top two*)

to which eastern hemlock evolutionary lineages are associated with different environmental factors, and then projected how those lineages may respond differently to changing climatic conditions (Thomassen et al. 2010). These evolutionary lineages were likely generated by long-term biogeographic processes, including Pleistocene-era isolation, which may have resulted in differential adaptation among groups of populations in response to different sets of environmental pressures.

The results of our analyses show that (1) both genetic diversity and climate variables are associated with variation in the population-level abundance of eastern hemlock; (2) demographic, environmental, and genetic diversity differences exist among the four genetic zones we delineated for eastern hemlock based on the previous molecular marker results; (3) locations with suitable environmental conditions for eastern hemlock are expected to shift, in response to both harsh and mild climate change projections, but the situation is less dire when considering genetic zones separately; and (4) the colonization of hemlock beyond its existing distribution is likely to be very limited with climate change except for narrow bands near the NE and some near the SE.

Not surprisingly, given eastern hemlock's association with cool and humid climates (Godman and Lancaster 1990), abundance was positively associated with growing season precipitation [PMAYSEP] and growing season aridity index [GSAI] (with higher numbers indicating lower aridity), and negatively associated with May–September temperature [TMAYSEP] and January temperature [TJAN] (Appendix 1, Fig. A4 in Online Supplementary Material). The relationships, as expected, are in some cases non-linear as, for example, TJAN. Surprisingly, genetic variables explained a higher proportion of abundance variance than did the environmental variables. The relationships between abundance and expected heterozygosity (H_E), effective alleles (A_E), and allelic richness (A) were all mainly positive, while the relationships between abundance and both mean pairwise genetic distance (MEAN_D_C) and mean pairwise genetic differentiation (MEAN_F_{ST}) declined rapidly (Appendix 1, Fig. A5 in Online Supplementary Material). This is in keeping with previous work (Potter et al. 2012) that demonstrated that peripheral populations of eastern hemlock, which tend to consist of trees in low abundance, had significantly lower levels of genetic variation and were likely to be genetically distinct from interior range populations. The strong associations we detected between abundance and genetic indices based on buffered microsatellite population samples (R-square = 0.54, Supplementary Material Appendix 1, Fig. A1b) as well as climatic variables (R-square = 0.45, Supplementary Material Appendix 1, Fig. A1a) set the stage for the next step of evaluating whether significant differences in demography, genetic indices, and environmental variables exist among eastern hemlock genetic zones.

Our second set of results in fact demonstrates that significant differences exist among the four genetic zones, which were delineated using the microsatellite results and the location of FIA plot-level data. Because the large sample size was perhaps the main reason for the statistical significance in a majority of the cases, we have tried to interpret the result cautiously within the context of "ecological" significance. The means of all 11 environmental variables show some significant differences among the zones, although some values tend to be similar (for example pH for zones SE and SW and SIEVE10 for NE and SE) (Table 4);—however, barring a few exceptions, the zones are ecologically distinct. Specifically, the SE zone (encompassed within the Southern Appalachian Mountains) had the highest growing-season aridity index (indicating lower aridity) and precipitation, mean January temperatures, and the maximum elevation as well as elevational variation (Table 4). The Southwestern zone (encompassing southern Indiana and Ohio, eastern Kentucky, northeastern Tennessee, and western portions of Virginia and West Virginia)

was somewhat similar, but with the highest growing season mean temperature, lower elevation, and higher soil clay content. The Northeastern and North Central zones were both cooler and drier than the southern zones, with the North Central zone (in the Great Lakes States of Michigan and Wisconsin) the most extreme in these variables. Both also had soils that were less acidic. Significantly, eastern hemlock average tree diameter was the highest in this zone (Table 3), which is not unexpected given that this is the only zone which has not yet been effected by the hemlock woolly adelgid (Morin et al. 2011). More surprisingly, however, this region had the highest percent mortality and the second lowest seedling count, suggesting that eastern hemlock in this region may not currently be structurally sustainable (Cale et al. 2014) for reasons unrelated to HWA. Despite having been subjected to HWA infestation for the longest time, the Northeast zone (encompassing the New England States and parts of New York, Pennsylvania and New Jersey) had the highest seedling count and a relatively high average diameter and relatively low percent mortality. This was consistent with recent broad-scale analyses of eastern hemlock demographic trends in the region, which suggest that increasing tree density associated with the past century of reforestation and succession currently may be overwhelming the negative impacts of HWA (Trotter et al. 2013). It is not likely the result of increased hemlock advanced regeneration following HWA-caused tree mortality, however, as research has found an inverse correlation between HWA density and hemlock seedling density, perhaps as a result of increased deer browsing (Preisser et al. 2011). Meanwhile, mortality was the lowest in the Southeast zone, an area where HWA would be expected to cause higher mortality rates in hemlock because HWA itself does not experience as much overwinter mortality as in the colder north (Trotter and Shields 2009). The southern infestations may be recent enough, however, for the impacts on hemlock abundance to not have fully manifested themselves yet (Trotter et al. 2013). This pattern of abundance, along with the fact that seedling count was the lowest in the Southeastern zone, is consistent with the finding that the regional impacts of HWA on its host are surprisingly slow, but that the insect may be altering regional forest succession pathways (Morin and Liebhold 2015).

Meanwhile, we also detected significant genetic diversity differences among the zones, relying on the sparser sample of 58 locations across the range of eastern hemlock (Table 5). Several of the genetic diversity metrics (A, P_P, H_E, and A_E) were highest in the Southeast, a finding which has led to previous inferences that this is the area most likely to be near the location of at least one Pleistocene refuge for eastern hemlock (Potter et al. 2008, 2012; Lemieux et al. 2011). Populations near such glacial refugia are expected to have higher levels of genetic diversity than do the areas colonized from these locations (Hewitt 2000; Comes and Kadereit 1998). At the same time, we found that the most geographically isolated zones (Southwest and North Central) were also characterized by the highest levels of population-level differentiation (Mean FST and Mean DC). This also is expected, given previous inference about the location of eastern hemlock Pleistocene refugia and about the subsequent movement of the species from those locations with climate warming. Specifically, populations in the Southwest zone are likely to have descended from ancestors in a relatively isolated glacial refuge to the west of the Southern Appalachian Mountains, while the ancestors of the North Central populations appear to have moved north from a refuge near the Southern Appalachians before turning west through the Great Lakes region (Potter et al. 2012). Eastern hemlocks in these two zones, then, have become highly isolated from the hemlocks in other zones, leading to genetic differentiation of populations within each of the regions (Provan and Bennett 2008).

By establishing environmental and genetic differences among the four zones, these preceding results underscore the utility of generating projections of how climate change in

the coming decades may affect eastern hemlock separately within each of the genetic zones. For example, our range-wide analysis of environmental suitability for eastern hemlock (as a single undifferentiated unit) predicts a loss of suitable area under both mild and harsh future climate projections (Fig. 3), although under harsh climate, the loss is more widespread and intensive. When the analysis is limited to a single genetic zone (SE), however, the predictions are not as dire: While less area falls within the "highly suitable" category compared to the present, more area is categorized as having "MediumHigh" or "LowMedium" suitability, under both mild and harsh projections (Fig. 4). Additionally, the mapped decision tree (Fig. 5), which explains 47% of the deviance (goodness-of-fit of the model compared to the null model), can be used to understand the rules driving the distribution of the species as a function of abundance, targeting areas of higher vulnerability. For example, for the SE genetic zone, the predominant rule is depicted by node three (green) that has an average diameter of 5.83 based on rule: TMAYSEP $>$ 17.62 C and ELVSD $>$ 65.22 and CLAY $<$ 23.73%. Further, the niche maps (Supplementary Material Appendix 1, Figs. A2, A3) help in understanding small scale variations in abundance based on climatic variables within the genetic zones. All these results could guide both existing in situ and ex situ conservation activities (that is, determining where to focus seed collection efforts and to preserve remaining populations of hemlock), as well as informing future replanting/assisted migration efforts for the species (that is, deciding where genetic material collected from a certain place could be planted and vice versa; see below). Such efforts have been under way since 2003, with the USDA Forest Service and Camcore, an international tree breeding and conservation program at North Carolina State University, focusing on conserving the genetic resources of both eastern hemlock and Carolina hemlock (*Tsuga caroliniana*) through the collection of seed, from hundreds of mother trees, placed into long-term storage or planted in conservation orchards in Chile, Brazil, and North Carolina (Jetton et al. 2013; Oten et al. 2014). Specifically, the findings of the current analyses predict that the suitability of environmental conditions will decline across much of the distribution of eastern hemlock, but that gene conservation activities, especially seed collections, will particularly need to focus on the SW and NC genetic zones, where suitable habitat may be scarce even under a mild future scenario (Fig. 3). The highly genetically diverse SE genetic zone, meanwhile, may undergo less of a decline in habitat suitability (except under a harsher climate scenario), but is likely to continue to experience high levels of HWA mortality and will therefore require a continued emphasis on both seed collection and on chemical, silvicultural and biological efforts to control the exotic insect (Cheah et al. 2004; Ward et al. 2004; Jetton et al. 2008). Additionally, the findings of these analyses could also inform studies, such as common garden experiments, that test the suitability of genetic lineages to altered climatic conditions in high risk areas (Shinneman et al. 2016).

We additionally assessed the colonization potential of eastern hemlock around 2100, by treating the species both as a single undifferentiated unit (Fig. 6, left), and as a set of genetically differentiated zones (Fig. 6, right). In both cases, when the areas of high habitat-quality and high colonization likelihood (HQ–CL) are combined under future climatic conditions, the possibilities are quite restricted (Fig. 7), with those areas for the genetic zones occurring mostly within areas that are currently occupied by eastern hemlock, but outside our delineation of those zones. Clearly, the colonization potential of eastern hemlock is very limited even under a generous migration potential of 50 km/century, emphasizing that it likely would not be an effective and successful conservation strategy to rely on the species to migrate by itself to future environmentally suitable locations (e.g., those delineated in Fig. 3). Within the climate change context, and

separate from eastern hemlock's susceptibility to the invasive hemlock woolly adelgid, comprehensive ex situ seed collection (Jetton et al. 2013) combined with some form of assisted migration (Pedlar et al. 2012; Iverson and McKenzie 2013) may be required to move the species into newly suitable locations in an attempt to maintain it as a meaningful component of eastern forests. Matching seed sources with appropriate environmental conditions would be a critical step in this process (Potter and Hargrove 2012), a task that could be assisted with an understanding of how intra-species evolutionary lineages are associated differentially with environmental conditions (e.g., Shinneman et al. 2016). Information such as that presented in Fig. 4 can help identify areas of future suitability which the trees from a given genetic zone may not be able to reach unassisted, for example north along the Appalachian chain from the SE genetic zone.

We note that some limitations exist in the combination of demographic data and environmental information with genetic diversity measurements derived from neutral molecular markers. First, a mismatch in data resolution is a major impediment in incorporating genetic data into range-wide assessments of climate change susceptibility in tree species. Given the expense and time required to conduct a range-wide molecular marker study, especially for a species like eastern hemlock with a large distribution, it is currently not possible to generate genetic measurements at the same resolution at which environmental and even demographic data (in the case of the U.S. Forest Inventory and Analysis program) are available. The genetic information will, therefore, not have the spatial precision of the other data sets, although aggregating analyses to larger scales can help to some extent when the number of populations sampled for genetic information is large enough. Among other things, this means that the analyses are not sensitive to isolated disjunct populations, which may have high numbers of rare alleles (Gapare et al. 2005) and heightened levels of genetic differentiation (Eckert et al. 2008), except to indicate whether the locations of these populations are likely to have suitable environmental conditions for the species in the future. On the other hand, matching relatively fine scale environmental and demographic information with coarser-scale genetic diversity data may enable landscape-genetics analyses that focus on processes at relatively fine spatial scales (Manel et al. 2003), including the development of models predicting locations of populations with high genetic diversity. These models could then be used to guide efforts to prioritize locations for conservation management and for seed collections.

Second, variation in neutral molecular markers like microsatellites generally does not reflect adaptive genetic variation within populations or individuals. Instead, it can serve as an indicator of the existence of within-species evolutionary lineages that, as a result of broad-scale phylogeographic processes, potentially underwent differential natural selection. While such information is important and informative, as well as relatively straightforward to attain and interpret, a renewed focus is needed on quantitative genetic studies to determine if and how species will adapt to changing conditions (Kramer and Havens 2009). In eastern hemlock, this could include research to identify molecular markers, such as single nucleotide polymorphisms (SNPs), that are associated with important fitness-related traits such as adaptation to climatic and other environmental conditions (Kirk and Freeland 2011; Eckert et al. 2013; Pais et al. 2017). The results presented here lay the foundation for such future work, which could play an important role in the genetic conservation of a widespread and ecological important tree species that faces multiple threats to its existence across much of its distribution. Also, because our study tackles the widespread global problem of tree conservation under multiple threats by combining genetic analysis with environmental modelling, our approach can be applied to any region where such a problem exists, and where the underlying data are available. The methods and techniques used here

should of course be modified and fine-tuned to local and regional conditions, but we provide a broad template for such studies that can be generalized across continents.

In summary, we found that separate eastern hemlock zones, associated with different evolutionary lineages within the species, exhibit differences in demography, environmental associations, and genetic variation. Under changing climate, locations with suitable environmental conditions for eastern hemlock are expected to shift, but the colonization of hemlock beyond its existing distribution is likely to be very limited. The results we present, however, can help guide on-the-ground conservation activities, such as prioritizing areas for climate-change-oriented ex situ seed collections (including the SW and NC genetic zones) and for areas (such as the SE genetic zone) where efforts to control the exotic hemlock wooly adelgid are likely to be most necessary.

Acknowledgements The authors thank Louis Iverson and Bill Hargrove for their advice and assistance, and to the two anonymous reviewers for their valuable comments. The authors also express their gratitude to the Forest Inventory and Analysis field crew members for their efforts to collect the data used in this study. This research was supported in part through Cost Share Agreements 14-CS-11330110-042 and 15-CS-11330110-067 between the between the U.S. Department of Agriculture, Forest Service, Southern Research Station, and North Carolina State University.

References

Alsos IG, Ehrich D, Thuiller W, Eidesen PB, Tribsch A, Schonswetter P, Lagaye C, Taberlet P, Brochmann C (2012) Genetic consequences of climate change for northern plants. Proc R Soc B 279:2042–2051

Andrew RL, Wallis IR, Harwood CE, Foley WJ (2010) Genetic and environmental contributions to variation and population divergence in a broad-spectrum foliar defence of Eucalyptus tricarpa. Ann Bot 105:707–717

Baltunis BS, Gapare WJ, Wu HX (2010) Genetic parameters and genotype by environment interaction in radiata pine for growth and wood quality traits in Australia. Silvaec Genet 59:113–124

Bechtold WA, Scott CT (2005) The forest inventory and analysis plot design. In: Bechtold WA, Patterson PL (eds) The enhanced forest inventory and analysis program—national sampling design and estimation procedures., Technical Report SRS-80, GeneralUnited States Department of Agriculture, Forest Service, Southern Research Station, Asheville, pp 27–42

Breiman L (2001) Random forests. Mach Learn 45:5–32

Bucklin DN, Basille M, Benscoter AM, Brandt LA, Mazzotti FJ, Romañach SS, Speroterra C, Watling JI (2015) Comparing species distribution models constructed with different subsets of environmental predictors. Divers Distrib 21:23–35

Cale JA, Teale SA, West JL, Zhang LJI, Castello DR, Devlin P, Castello JD (2014) A quantitative index of forest structural sustainability. Forests 5:1618–1634

Cheah C, Montgomery ME, Salom S, Parker B, Skinner M, Costa S (2004) Biological Control of Hemlock Woolly Adelgid. United States Department of Agriculture, Forest Service, Forest Health Technology Enterprise Team, Morgantown

Chen T, Guestrin C (2016) XGBoost?: reliable large-scale tree boosting system. arXiv:1603.02754 [cs.LG], http://arxiv.org/pdf/1603.02754v1

Chen T, He T (2015) Higgs Boson discovery with boosted trees. JMLR 42:69–80

Chen C, Durand E, Forbes F, Francois O (2007) Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study. Mol Ecol Notes 7:747–756

Comes HP, Kadereit JW (1998) The effect of quaternary climatic changes on plant distribution and evolution. Trends Plant Sci 3:432–438

Cutler DR, Edwards TC, Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ (2007) Random forests for classification in ecology. Ecology 88:2783–2792

Daly C, Halbleib M, Smith JI, Gibson WP, Doggett MK, Taylor GH, Curtis J, Pasteris PP (2008) Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. Int J Climatol 28:2031–2206

Dobrowski SZ (2011) A climatic basis for microrefugia: the influence of terrain on climate. Glob Change Biol 17:1022–1035

Eckert CG, Samis KE, Lougheed SC (2008) Genetic variation across species' geographical ranges: the central-marginal hypothesis and beyond. Mol Ecol 17:1170–1188
Eckert AJ, Wegrzyn JL, Liechty JD, Lee JM, Cumbie WP, Davis JM, Goldfarb B, Loopstra CA, Palle SR, Quesada T, Langley CH, Neale DB (2013) The evolutionary genetics of the genes underlying phenotypic associations for loblolly pine (Pinus taeda, Pinaceae). Genetics 195:1353–1372
ESRI (2015) ArcGIS [GIS software]. Version 10.3.1. Environmental Systems Research Institute, Redlands, CA
Excoffier L, Laval G, Schneider S (2005) Arlequin ver. 3.0: an integrated software package for population genetics data analysis. Evol Bioinform 1:47–50
Felsenstein J (2005) PHYLIP (phylogeny inference package), version 3.6. Department of Genome Sciences, University of Washington, Washington, DC
Ford CR, Vose JM (2007) *Tsuga canadensis* (L.) Carr. mortality will impact hydrologic processes in southern Appalachian forest ecosystems. Ecol Appl 17:1156–1167
Friedman JH (2002) Stochastic gradient boosting. Comput Stat Data Anal 38:367–378
Gapare WJ, Aitken SN, Ritland CE (2005) Genetic diversity of core and peripheral Sitka spruce (Picea sitchensis (Bong.) Carr) populations: implications for conservation of widespread species. Biol Cons 123:113–123
Gent PR, Danabasoglu G, Donner LJ, Holland MM, Hunke EC, Jayne SR, Lawrence DM, Neale RB, Rasch PJ, Vertenstein M, Worley PH, Yang ZL, Zhang M (2011) The community climate system model, version 4. J Clim 24:4973–4991
Geurts P, Ernst D, Wehenkel L (2006) Extremely randomized trees. Mach Learn 63:3–42
Godman RM, Lancaster K (1990) Eastern Hemlock. In: Burns RM, Honkala BH (eds) Silvics of North America: 1 conifers. United States Department of Agriculture Forest Service, Washington, DC
Godsoe W (2010) I can't define the niche but I know it when I see it: a formal link between statistical theory and the ecological niche. Oikos 119:53–60
Gotelli NJ, Stanton-Geddes J (2015) Climate change, genetic markers and species distribution modelling. J Biogeogr 42:1577–1585
Goudet J (1995) FSTAT (Version 1.2): a computer program to calculate F-statistics. J Hered 86:485–486
Guth PL (2006) Geomorphometry from SRTM: comparison to NEDPhotogrammetric engineering and remote sensing 72:269–277
Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning. Springer, New York, p 745
Heard MJ, Valente MJ (2009) Fossil pollen records forecast response of forests to hemlock woolly adelgid invasion. Ecography 32:881–887
Hewitt GM (2000) The genetic legacy of the Quaternary ice ages. Nature 405:907–913
Iverson LR, McKenzie D (2013) Tree-species range shifts in a changing climate, detecting, modeling, assisting. Landsc Ecol 28:879–889
Iverson LR, Schwartz MW, Prasad AM (2004) How fast and far might tree species migrate under climate change in the eastern United States? Glob Ecol Biogeogr 13:209–219
Iverson LR, Prasad AM, Matthews SN, Peters M (2008) Estimating potential habitat for 134 eastern US tree species under six climate scenarios. For Ecol Manage 254:390–406
Jetton RM, Whittier WA, Dvorak WS, Potter KM (2008) Status of ex situ conservation efforts for Carolina and eastern hemlock in the southeastern United States. In: Onken B, Reardon R (eds) Proceedings of the fourth symposium on Hemlock Woolly adelgid in the Eastern United States. USDA Forest Service, Hartford, pp 81–89
Jetton RM, Whittier WA, Dvorak WS, Rhea JR (2013) Conserved ex situ genetic resources of eastern and Carolina hemlock: eastern North American conifers threatened by the hemlock woolly adelgid. Tree Plant Notes 56:59–71
Jones MC, Cheung WWL (2015) Multi-model ensemble projections of climate change effects on global marine biodiversity. ICES J Mar Sci 72:741–752
Jones CD, Hughes JK, Bellouin N, Hardiman SC, Jones GS, Knight J, Liddicoat S, O'Connor FM, Andres RJ, Bell C, Boo KO, Bozzo A, Butchart N, Cadule P, Corbin KD, Doutriaux-Boucher M, Friedlingstein P, Gornall J, Gray L, Halloran PR, Hurtt G, Ingram WJ, Lamarque JF, Law RM, Meinshausen M, Osprey S, Palin EJ, Parsons-Chini L, Raddatz T, Sanderson MG, Sellar AA, Schurer A, Valdes P, Wood N, Woodward S, Yoshioka M, Zerroukat M (2011) The HadGEM2-ES implementation of CMIP5 centennial simulations. Geosci Model Dev 4:543–570
Josserand SA, Potter KM, Echt CS, Nelson CD (2008) Isolation and characterization of microsatellite markers for Carolina hemlock (*Tsuga caroliniana*). Mol Ecol Resour 8:1371–1374
Kessell S (1979) Adaptation and dimorphism in Eastern Hemlock, *Tsuga canadensis* (L.) Carr. Am Nat 113:333–350

Kirk H, Freeland JR (2011) Applications and Implications of neutral versus non-neutral markers in molecular ecology. Int J Mol Sci 12:3966–3988
Kramer AT, Havens K (2009) Plant conservation genetics in a changing world. Trends Plant Sci 14:599–607
Kuhn M (2008) Building predictive models in R using the caret package. J Stat Softw 28:1–26
Lawler JJ, White D, Neilson RP, Blaustein AR (2006) Predicting climate-induced range shifts: model differences and model reliability. Glob Change Biol 12:1568–1584
Lemieux MJ, Beaulieu J, Bousquet J (2011) Chloroplast DNA polymorphisms in eastern hemlock: range-wide genogeographic analyses and implications for gene conservation. Can J For Res 41:1047–1059
Lou X (2014) Gene–gene and gene–environment interactions underlying complex traits and their detection. Biomed Biostat Int J 1:1–8
Manel S, Schwartz MK, Luikart G, Taberlet P (2003) Landscape genetics: combining landscape ecology and population genetics. Trends Ecol Evol 18:189–197
Martre P, Wallach D, Asseng S, Ewert F, Boote KJ, Ruane AC, Peter J, Cammarano D, Hatfield JL, Rosenzweig C, Aggarwal PK, Angulo C, Basso B, Bertuzzi P (2015) Multimodel ensembles of wheat growth: many models are better than one. Glob Change Biol 21:911–925
McClure MS, Salom SM, Shields KS (2003) Hemlock woolly adelgid. Forest Health Technology Enterprise Team, United States Department of Agriculture, Forest Service, Morgantown
McLachlan JS, Clark JS (2004) Reconstructing historical ranges with fossil data at continental scales. For Ecol Manage 197:139–147
Meinshausen M, Smith SJ, Calvin K, Daniel JS, Kainuma MLT, Lamarque JF, Matsumoto K, Montzka SA, Raper SCB, Riahi K, Thomson A, Velders GJM, van Vuuren DPP (2011) The RCP greenhouse gas concentrations and their extensions from 1765 to 2300. Clim Change 109:213–241
Montgomery ME, Bentz SE, Olsen RT (2009) Evaluation of hemlock (*Tsuga*) species and hybrids for resistance to *Adelges tsugae* (Hemiptera: Adelgidae) using artificial infestation. J Econ Entomol 102:1247–1254
Morin RS, Liebhold AM (2015) Invasions by two non-native insects alter regional forest species composition and successional trajectories. For Ecol Manage 341:67–74
Morin RS, Oswalt SN, Trotter RT, Liebhold AW (2011) Status of hemlock in the Eastern United States, Forest Inventory and Analysis factshee. US Department of Agriculture Forest Service, Southern Research Station, Asheville
Murillo O (2001) Genotype by environment interaction and genetic gain on. Agron Costarric 25:21–31
Nagaraju SK, Gudasalamani R, Barve N, Ghazoul J, Narayanagowda GK, Ramanan US (2013) Do ecological niche model predictions reflect the adaptive landscape of species? A test using *Myristica malabarica* Lam, an endemic tree in the Western Ghats, India. PLoS ONE 8:1–13
Nienstaedt H, Olson JS (1961) Effects of photoperiod and source on seedling growth of eastern hemlock. For Sci 7:81–96
NRCS (Natural Resources Conservation Service) (2009) Soil survey geographic (SSURGO). http://soildatamart.nrcs.usda.gov/State.aspx. Accessed between August 2009 and November 2010
Olson JS, Nienstaedt H (1957) Photoperiod and chilling control growth of hemlock. Science 125:492–494
Oten KLF, Merkle SA, Jetton RM, Smith BC, Talley ME, Hain FP (2014) Understanding and developing resistance in hemlocks to the hemlock woolly adelgid. Southeast Nat 13:147–167
Pais AL, Whetten RW, Xiang Q (2017) Ecological genomics of local adaptation in Cornus florida L. by genotyping and sequencing. Ecol Evol 7:441–465
Pederson N, Amato AWD, Dyer JM, Foster DR, Goldblum D, Hart JL, Hessl AE, Iverson LR et al (2015) Climate remains an important driver of post-European vegetation change in the eastern United States. Glob Change Biol 21:2105–2110
Pedlar JH, McKenney DW, Aubin I, Beardmore T, Beaulieu J, Iverson L, O'Neill GA, Winder RS, Ste-Marie C (2012) Placing forestry in the assisted migration debate. Bioscience 62:835–842
Peters MP, Matthews SN, Iverson LR, Prasad AM (2013) Delineating generalized species boundaries from species distribution data and a species distribution model. Int J Geogr Inf Sci 28:1547–1560
Potter KM, Hargrove WW (2012) Determining suitable locations for seed transfer under climate change, a global quantitative model. New Forest 43:581–599
Potter KM, Dvorak WS, Crane BS, Hipkins VD, Jetton RM, Whittier WA, Rhea R (2008) Allozyme variation and recent evolutionary history of eastern hemlock (*Tsuga canadensis*) in the southeastern United States. New Forest 35:131–145
Potter KM, Jetton RM, Dvorak WS, Hipkins VD, Rhea R, Whittier WA (2012) Widespread inbreeding and unexpected geographic patterns of genetic variation in eastern hemlock (*Tsuga canadensis*), an imperiled North American conifer. Conserv Genet 13:475–498

Prasad AM (2015) Macroscale intraspecific variation and environmental heterogeneity: analysis of cold and warm zone abundance, mortality, and regeneration distributions of four eastern US tree species. Ecol Evol 5:5033–5048

Prasad AM, Iverson LR, Liaw A (2006) Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems 9:181–199

Prasad AM, Gardiner JD, Iverson LR, Matthews SN, Peters M (2013) Exploring tree species colonization potentials using a spatially explicit simulation model: implications for four oaks under climate change. Glob Change Biol 19:2196–2208

Prasad AM, Iverson LR, Matthews SN, Peters MP (2016) A multistage decision support framework to guide tree species management under climate change via habitat suitability and colonization models, and a knowledge-based scoring system. Landsc Ecol. doi:10.1007/s10980-016-0369-

Preisser EL, Miller-Pierce MR, Vansant J, Orwig DA (2011) Eastern hemlock (*Tsuga canadensis*) regeneration in the presence of hemlock woolly adelgid (*Adelgis tsugae*) and elongate hemlock scale (*Fiorinia externa*). Can J For Res 41:2433–2439

PRISM Climate Group (2017) Oregon State University, http://prism.oregonstate.edu

Provan J, Bennett KD (2008) Phylogeographic insights into cryptic glacial refugia. Trends Ecol Evol 23:564–571

Pulliam HR (2000) On the relationship between niche and distribution. Ecol Lett 3:349–361

R Core Team (2016) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/

Schaberg PG, DeHayes DH, Hawley GJ, Nijensohn SE (2008) Anthropogenic alterations of genetic diversity within tree populations: implications for forest ecosystem resilience. For Ecol Manage 256:855–862

Schwartz MW (1993) Modelling effects of habitat fragmentation on the ability of trees to respond to climatic warming. Biodivers Conserv 2:51–61

Shamblin BM, Faircloth BC, Josserand SA, Nelson D, Nairn CJ (2008) Microsatellite markers for eastern hemlock (*Tsuga canadensis*). Mol Ecol Resour 8:1354–1356

Shinneman DJ, Means RE, Potter KM, Hipkins VD (2016) Exploring climate niches of ponderosa pine haplotypes in the western United States: insight into evolutionary history and implications for future conservation. PLoS ONE 11(3):e0151811

Smith WB (2002) Forest inventory and analysis: a national inventory and monitoring program. Environ Pollut 116:S233–S242

Soberon J, Peterson AT (2005) Interpretation of models of fundamental ecological niches and species' distributional areas. Biodivers Inform 2:1–10

Spaulding HL, Rieske LK (2010) The aftermath of an invasion: structure and composition of Central Appalachian hemlock forests following establishment of the hemlock woolly adelgid, *Adelges tsugae*. Biol Invasions 12:3135–3143

Svenning JC, Skov F (2007) Could the tree diversity pattern in Europe be generated by postglacial dispersal limitation? Ecol Lett 10:453–460

Quantum GIS Development Team (2016). Quantum geographic information system. Open source geospatial foundation project. http://qgis.osgeo.org. Accessed 17 Feb 2017

Thomassen HA, Cheviron ZA, Freedman AH, Harrigan RJ, Wayne RK, Smith TB (2010) Spatial modelling and landscape-level approaches for visualizing intra-specific variation. Mol Ecol 19:3532–3548

Thornthwaite C, Mather J (1957) Instructions and tables for computing potential evapotranspiration and the water balance. Publ Climatol 10:185–311

Thrasher B, Xiong J, Wang W, Melton F, Michaelis A, Nemani R (2013) Downscaled climate projections suitable for resource management. Trans Am Geophys Union 94:321–323

Trotter RT, Shields KS (2009) Variation in winter survival of the invasive hemlock woolly adelgid (Hemiptera, Adelgidae) Across the Eastern United States. Environ Entomol 38:577–587

Trotter RT, Morin RS, Oswalt SN, Liebhold AM (2013) Changes in the regional abundance of hemlock associated with the invasion of hemlock woolly adelgid (*Adelges tsugae Annand*). Biol Invasions 15:2667–2679

USDA Forest Service (2015) Forest health protection—Hemlock Woolly Adelgid, distribution maps. https://www.na.fs.fed.us/fhp/hwa/maps/distribution.shtm. Accessed 13 Feb 2017

Ward JS, Montgomery ME, Cheah C, Onken BP, Cowles RS (2004) Eastern Hemlock forests: guidelines to minimize the impacts of Hemlock Woolly Adelgid. United States Department of Agriculture, Forest Service, Morgantown

Woudenberg SW, Conkling BL, O'Connell BM, LaPoint EB, Turner JA, Waddell KL (2010) The forest inventory and analysis database: database description and user's manual version 4.0 for Phase 2. General Technical Report RMRS-GTR-245, USDA Forest Service, Rocky Mountain Research Station, Fort Collins, Colorado

Yansa CH (2006) The timing and nature of Late Quaternary vegetation changes in the northern Great Plains, USA and Canada: a re-assessment of the spruce phase. Q Sci Rev 25:263–281

Zenni RD, Lamy JB, Lamarque LJ, Porte AJ (2014) Adaptive evolution and phenotypic plasticity during naturalization and spread of invasive species: implications for tree invasion biology. Biol Invasions 16:635–644

Zhang L, Liu S, Sun P, Wang T, Wang G, Zhang X, Wang L (2015) Consensus forecasting of species distributions: the effects of niche model performance and niche properties. PLoS ONE 10:1–18