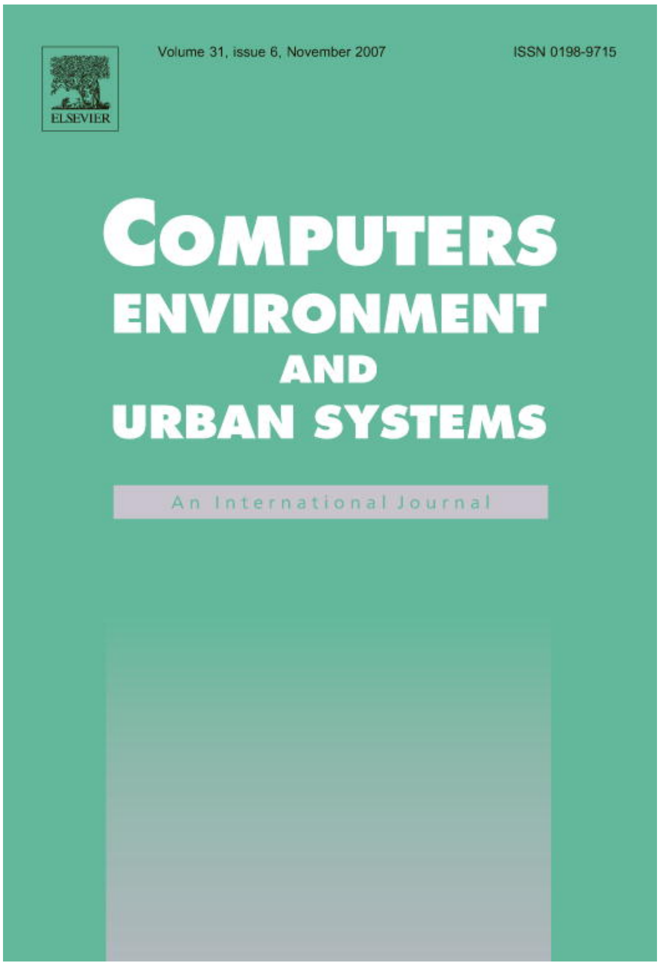

Volume 31, issue 6, November 2007

ISSN 0198-9715

ELSEVIER

COMPUTERS ENVIRONMENT AND URBAN SYSTEMS

An International Journal

Available online at www.sciencedirect.com
ScienceDirect
Computers, Environment and Urban Systems
31 (2007) 689–710

Computers, Environment and Urban Systems

www.elsevier.com/locate/compenvurbsys

Modeling the risk for a new invasive forest disease in the United States: An evaluation of five environmental niche models

Maggi Kelly *, Qinghua Guo [1], Desheng Liu [2], David Shaari [3]

Geospatial Imaging and Informatics Facility, Department of Environmental Sciences, Policy and Management, University of California, Berkeley, 137 Mulford Hall #3114, Berkeley, CA 94720-3114, USA

Received 25 May 2006; accepted in revised form 23 October 2006

Abstract

Efforts to model the potential habitat and risk for spread of invasive diseases such as Sudden Oak Death (SOD) are important for disease regulation and management. However, spatially referenced risk models using identical data can have differing results, making decision-making based on the mapped results problematic. We examined the results from five spatial risk models generated from common input parameters, and investigated model agreement for mapping risk for the causal pathogen for SOD, *Phytophthora ramorum* across the conterminous United States. We examined five models: Expert-driven Rule-based, Logistic Regression, Classification and Regression Trees, Genetic Algorithms, and Support Vector Machines. All models were consistent in their prediction of some SOD risk in coastal California, Oregon and Washington states, and in the northern foothills of the Sierra Nevada Mountains in California, and in an east–west oriented band including eastern Oklahoma, central Arkansas, Tennessee, Kentucky, northern Mississippi, Alabama, Georgia and South Carolina, parts of central North Carolina, and eastern Virginia, Delaware and Maryland states. The SVM model was the most accurate model, and had several advantages over the other models. Although theoretical in nature, this paper presents results that have practical, applied value

* Corresponding author. Tel.: +1 510 642 7272; fax: +1 510 642 1477.
E-mail address: mkelly@nature.berkeley.edu (M. Kelly).

[1] Present address: School of Engineering, University of California, Merced, P.O. Box 2039, Merced, CA 95344, USA.

[2] Present address: Department of Geography, The Ohio State University, 1036 Derby Hall, 154 North Oval Mall, Columbus, OH 43210, USA.

[3] Present address: California Department of Fish and Game, 4949 Viewridge Ave., San Diego, CA 92123, USA.

doi:10.1016/j.compenvurbsys.2006.10.002

for managers and regulators of this disease, and discusses common challenges in modeling invasive species niches over large scales.

Keywords: Environmental niche modeling; Sudden oak death; Geographic information systems

1. Introduction

An invasive pathogen *Phytophthora ramorum* is the causal agent for a new disease called "Sudden Oak Death" that has reached epidemic levels in hardwood and mixed hardwood forests in 14 counties in central coastal California and one county in southern Oregon (Fig. 1). In the United States the disease is confined to the west coast, however the potential for the disease to spread to other areas is high, and modeling its potential environmental niche across a broad geographic scope is important for disease monitoring and management.

The disease has killed tens of thousands of oak and tanoak trees (*Quercus agrifolia*, *Quercus kelloggii* and *Quercus parvula* var.*shrevei* and *Lithocarpus densiflorus*) and affects more than 25 "foliar hosts" – plant species that experience non-fatal foliar symptoms (McPherson et al., 2005; Rizzo, 2003; Rizzo & Garbelotto, 2003). The foliar hosts, particularly California bay laurel (*Umbellularia californica*), play an important role in short-range pathogen dispersal within forests: their leaves serve as reservoirs for the pathogen spores which, given advantageous moisture and temperature conditions, can be spread to soil and leaf litter via rain, and short distances through a forest by wind-driven rain (Davidson, Rizzo, Garbelotto, Tjosvold, & Slaughter, 2002; Davidson & Shaw, 2003; Davidson, Wickland, Patterson, Falk, & Rizzo, 2005; Rizzo & Garbelotto, 2003). Other foliar hosts play a similar role to a lesser degree. At larger scales, human activity might play a role in pathogen movement. Researchers have discussed the possibility of movement of affected soil material through recreation activities (e.g., hiking, biking) or on vehicle tires (Cushman & Meentemeyer, 2005; Davidson & Shaw, 2003; Tjosvold, Chambers, Davidson, & Rizzo, 2002), and likely long-range dispersal mechanisms include trade in ornamental plants, including the foliar hosts *Rhododendron*,*Camellia* and *Viburnum* (Davidson & Shaw, 2003; Englander & Tooley, 2003; Rizzo, 2003), and trade in seasonal garlands, wreaths and Christmas trees, which can be made from foliar hosts (Davidson & Shaw, 2003).

The concern about SOD spread is not only theoretical: recent inspections of US nurseries have documented the transport of potentially infected ornamental plants from a California wholesale nursery to over 700 garden centers in 39 states in 2005, and confirmed positive samples of *P. ramorum* in nurseries in Oregon, Washington and British Columbia (Stokstad, 2004). Given that similar environmental conditions and susceptible plant species are found elsewhere in North America (for example, two eastern US oak trees, northern red oak (*Q. rubra*) and pin oak (*Q. palustris*) have been proven susceptible to *P. ramorum* (Rizzo, 2003)), there is considerable concern about potential spread of the pathogen to other forests nationwide (Cree, 2003; Davidson & Shaw, 2003; Gottschalk, Morin, & Liebhold, 2002; Rizzo, 2003; Tooley & Kyde, 2003).

Predictive habitat distribution models, or environmental niche models (ENM), are increasingly recognized as important tools that can support our understanding of biotic

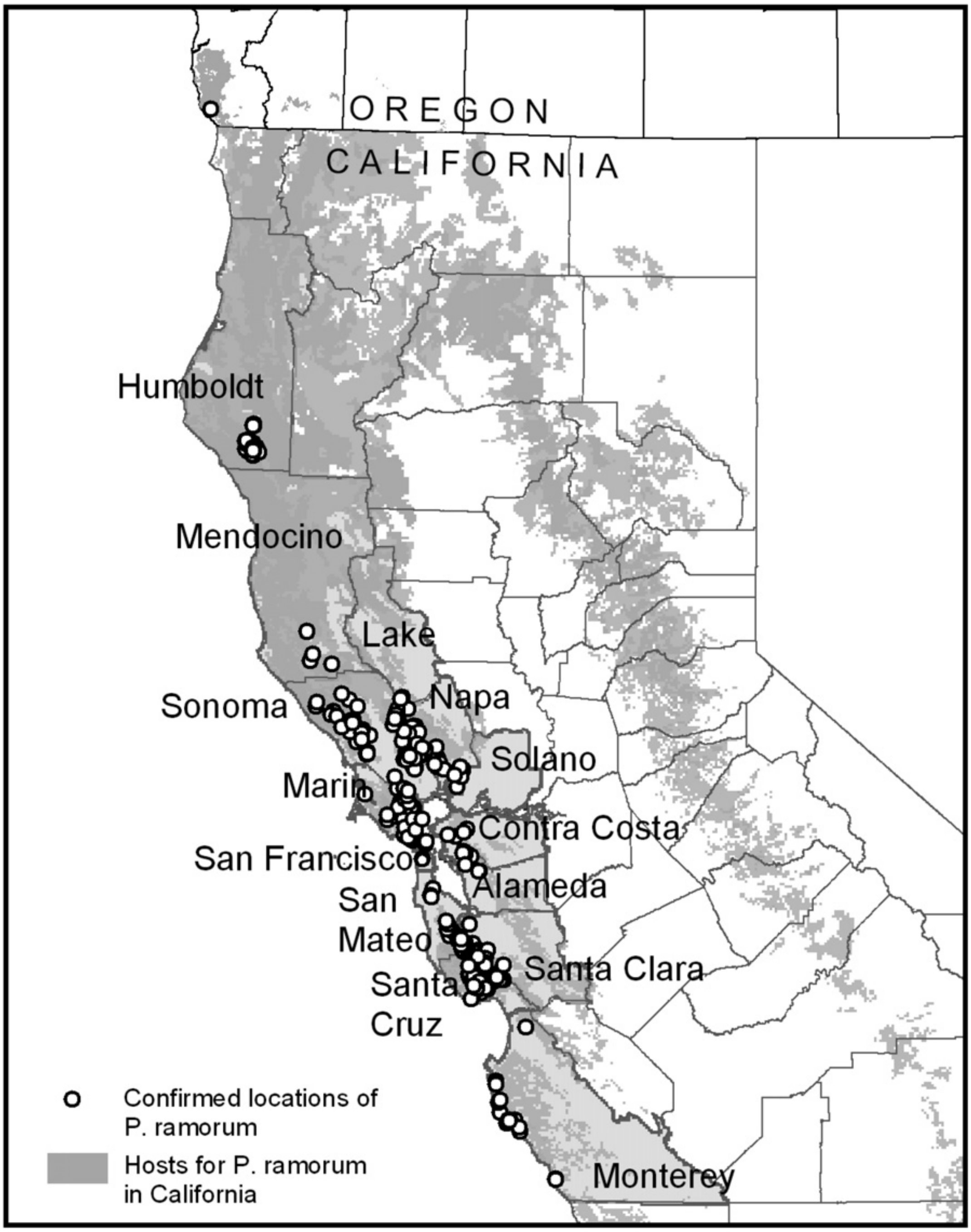

Fig. 1. Current distribution of *Phytophthora ramorum* in California and Oregon.

invasions and diseases (Costa, Peterson, & Beard, 2002; Guo, Kelly, & Graham, 2005; Higgins, Richardson, & Cowling, 2001; Jules, Kauffman, Ritts, & Carroll, 2002; Venette & Cohen, 2006), historical habitats and climate change impacts (Iverson & Prasad, 1998; Clark, Rose, Levine, & Hardgrove, 2001), biodiversity and speciation mechanisms (Rushton, Omerod, & Kerby, 2004; Graham, Ferrier, Huettman, Moritz, & Peterson, 2004), as well as aid in setting natural resource, conservation and species management priorities (Felicisimo, Frances, Fernandez, Gondalez-Diez, & Varas, 2002; Fleishman, Nally, Fay, & Murphy, 2001; Illoldi-Rangel, Sanchez-Cordero, & Peterson, 2004; Kelly, Fonseca, & Whitfield, 2001; McShea et al., 2005; Mladenoff, Sickley, & Wydeven, 1999; Raxworthy et al., 2003; Sperduto & Congalton, 1996; Zaniewski, Lehmann, & Overton, 2002).

Efforts to model the potential habitat for *P. ramorum* and Sudden Oak Death (SOD) are important for disease regulation and management, and have been used at landscape-, regional- and nationwide-scales. For example, in California, researchers used an expert knowledge driven Rule-based Geographic Information System (GIS) model to predict SOD risk based on plant host, temperature and moisture data (Meentemeyer, Rizzo, Mark, & Lotz, 2004), and this model has been used to guide sampling, aerial surveys, and other statewide monitoring efforts. In addition, Venette and Cohen (2006) used a Rule-based model and expert knowledge to characterize suitable regional climate for *P. ramorum* across the United States. Rule-based models such as these use expert knowledge rather than statistical inference, and thus the predictor ecological variables used are known *a priori*. Other spatially referenced ecological niche models (such as logistic regression and classification and regression trees) use presence data to train a statistically based model with the aim of revealing interactions between data that represent ecological niches. The variety of techniques used for ecological niche modeling is growing (Guisan & Zimmermann, 2000), and there has been a corresponding increase the spatial modeling literature in work that compares results from different models (Guisan & Zimmermann, 2000; Huang & Lees, 2004; Manel, Dias, & Ormerod, 1999; Manel, Dias, Buckton, & Ormerod, 1999; Muñoz & Felicísimo, 2004). These efforts show that spatially referenced models using identical data often have differing results, due in part to: (1) the fact that models can be either parametric or non-parametric with varying reliance on explanatory variable distribution, (2) user-defined weightings placed on variables can differ by analyst, and (3) the different methods used for generating absence data for input to models can influence results. Therefore, it is often necessary to implement several niche models to determine whether the prediction results will depend on methods used (Graham, Moritz, & Williams, 2006). In addition, when modeling a biotic invasion, one is necessarily using samples drawn from a geographically constrained area to model over a larger geographic scope. This necessarily increases uncertainties, and enhances the differences between models in the results. In this research, we used a collection of common and newer model types to map risk for SOD across the conterminous United States, and evaluate model performance. We examined five classes of models that differed in terms of parametric and non-parametric requirements, the necessity for presence and/or absence data, and whether or not the explanatory variables were determined *a priori* or revealed during the model process. Our spatial models included: (1) Rule-based Expert-driven GIS overlay, (2) Logistic regression (LR), (3) Classification and Regression Trees (CART), (4) Genetic Algorithms (GA), and (5) Support Vector Machines (SVM).

1.1. Model descriptions

Rule-based spatial risk models use research data and expert input, rather than statistical inference to determine the importance of predictor variables (Meentemeyer et al., 2004). Predictor variables are given weights based on importance, and all weighted variables are manipulated using algebraic or Boolean logic operations in a GIS overlay procedure to produce a mapped output (Franklin, 1995). This method is straightforward, intuitively understandable, and not computationally intensive. Rule-based models have been used in numerous ecological and natural resource management applications. Site-selection applications benefit from this approach because experts can weigh the relative importance of variables in locating areas that meet a suite of criteria. For example, an expert-driven overlay process has been used to site wetland restoration projects (Llewellyn et al., 1995; McCauley & Jenkins, 2005; Russell, Hawkins, & O'Neill, 1997), aquaculture sites (Arnold, White, Norris, & Berrigan, 2000; Buitrago, Rada, Hernandez, & Buitrago, 2005; Karthik, Suri, Saharan, & Biradar, 2005) and to predict habitat for plant and animal species using expert input (Kampichler, Barthel, & Wieland, 2000; Petit et al., 2003; Pyke, 2005). The ability to harness expert opinion in a series of mathematical equations or logical "if" "then" statements is one of the cited advantages of the GIS overlay approach. For example, Parra, Graham, and Freile (2004) found sites for potential riparian vegetation restoration based on land cover, wetness, proximity to water, and size constraints. These "layers" were summed through an overlay process to yield a prioritization of potential restoration sites. Pyke (2005) modifies the technique to allow fuzzy membership in various output results, and site suitability for a California salamander was based on a series of logical propositions using input data layers such as land cover, road density, agriculture, urban growth, and movement parameters. Despite ease of use and flexibility, the overlay method does not take advantage of the strengths of statistical inference in deriving models, as do the next four models evaluated.

One of the most commonly used model for inferring ecological niches over space using presence and absence data is Logistic Regression (LR) (Franklin, 1995). LR is a variation of ordinary regression which is used when the dependent (response) variable is binary and represents the occurrence or non-occurrence of some outcome events, usually coded as '0' or '1', and the independent (input) variables are continuous, categorical, or both. Resulting probabilities can be mapped over space for an easily understood cartographic representation of modeled distribution. LR is a powerful parametric method for ecological niche modeling, and has been used in a variety of ecological modeling and conservation examples. McShea et al. (2005) used a LR model to identify environmental variables that were significant predictors of Eld's deer in Southeast Asia in support of region-wide species conservation efforts. Kelly et al. (2001) used LR to map the area potentially utilizable for seagrass colonization using bathymetry and disturbance information. Since seagrass patches can migrate over the seafloor, the seagrass area mapped directly at any one time can underestimate the actual habitat area; LR provided a more comprehensive picture of the habitat. Felicisimo et al. (2002) modeled the potential for six different forest types in northern Spain based on topographic variable and proximity to marine influences in order to better plan for forest management in the region. Mladenoff et al. (1999) used LR to guide wolf re-colonization efforts, and estimated the amount and spatial configuration of potential wolf habitat in the northeastern US.

The third model examined is relatively new, but is increasingly being used in ecological modeling and classification applications. Classification and Regression Trees (here called CART for simplicity) are a non-parametric alternative to parametric techniques such as LR (De'ath & Fabricius, 2000) and Linear Discriminant Analysis (LDA) (Feldesman, 2002). The method is increasing in popularity among researchers analyzing multivariate data, as it requires no advance variable selection, its results are invariant to transformations such as log transforms, it can use any combination of categorical and continuous predictor variables, it can handle missing data (Feldesman, 2002), and it has the ability to capture hierarchical and non-linear relationships and expose interactions among predictor variables (Clark & Pregibon, 1993; De'ath & Fabricius, 2000; Kelly & Meentemeyer, 2002; Michaelsen, Schimel, Friedl, Davis, & Dubayah, 1994). The tree models are developed by recursively partitioning the response variable into increasingly homogeneous binary subsets based on critical thresholds in predictor variables. The split chosen is the one that most reduces the average impurity in the resulting bins (Breiman, Friedman, Olshen, & Stone, 1984; De'ath & Fabricius, 2000; Venables & Ripley, 2002). The resulting "trees" are often displayed graphically, and are easy to understand as a series of if/then conditions, but they can be complex to render cartographically (Muñoz & Felicısimo, 2004).

Vayssières, Plant, and Allen-Diaz (2000) compared CART to generalize linear models for predicting the distribution of three major oak species in California. They found the CART models performed significantly better than the regression models, and noted the suitability of the "trees" to deal with complex environmental data which involves interactions and non-linearities. Fabricius and De'ath (2001) also comment on this: they used CART to examine the relationship between a kind of marine algae and visibility, slope and sediment exposure on coral reefs. Kelly and Meentemeyer (2002) used CART to reveal landscape-scale environmental controls distribution and spread of Sudden Oak Death mortality in a park in California facing epidemic levels of the disease. The method was able to use data from numerous formats, and revealed important interactions between environmental variables in defining the niche for the disease.

The fourth method evaluated, Genetic Algorithm (GA) modeling, is a an evolutionary computing system that has documented capabilities for delineating ecological niches and geographical distributions of species (Anderson, Lew, & Peterson, 2003; Raxworthy et al., 2003; Stockwell & Peters, 1999a, 1999b; Stockwell, 1999). The method use genetic algorithms to predict the potential distribution of a species by generating a set of rules. First, occurrence points are divided evenly into training and test data sets, and initial rules describing niches are developed by choosing a method from a set of possibilities. For example in the software we used (Desktop GARP (Genetic Algorithm for Rule-set Production) (Stockwell, 2006)) there are four types of rules implemented: atomic, logistic regression, bioclimatic envelope, and negated bioclimatic envelope rules. The atomic rule uses a single value of a variable (e.g., if max temp = 32 °C); the logistic regression uses linear logit equations; the bioclimatic rule encloses the range of the variables in a climate "envelope" (e.g., if the max temp is >25° and <32°); negated bioclimatic rules are similar to the bioclimatic rules, but they also allow negation (i.e. the rule applies outside of the range indicated) (Stockwell, 2006). We used all the rules in the GARP process, which will select the best fit rules for the model. Predictive accuracy of the model is then evaluated using test data. Rules "evolve" through an iterative process in which operational concepts similar to evolutionary biology (such as mutation and crossover) are employed; these might consist of random perturbations to rule structure, or additional rules may be produced. The change in predictive accuracy of a rule from one iteration to the next is used to evaluate whether a particular rule should be incorporated into the model, and the algorithm runs until convergence. As such, the GA method represents a superset of other approaches, and has several advantages: it is less susceptible to local maxima, it is able to handle various data formats (continuous and discrete), and should always have greater predictive ability than any one of the possible algorithms (logistic regression or bioclimatic rules) when used alone (Godown & Peterson, 2000).

GA models have been used in the exploration of potential niches for species of concern using voucher museum data (Graham et al., 2004), and other occurrence data. For example, Raxworthy et al. (2003) report significant ability to predict chameleon distribution in Madagascar using museum occurrence data and numerous land cover and climate data layers as inputs to a GA model. Illoldi-Rangel et al. (2004) used GA modeling to predict potential geographic distribution for 17 mammal species in Mexico, using museum occurrence data and climate, vegetation and topography. Foci of richness of endangered bird species in the US were successfully modeled using a GA method with Breeding Bird Survey data (Godown & Peterson, 2000), aiding avian protection efforts. The method has also been successful in predicting potential niche of diseases or disease vectors from regional monitoring data (Costa et al., 2002).

Finally, Support Vector Machines are a new generation of learning algorithms that can perform binary classification (pattern recognition) and real valued function approximation (regression estimation) tasks. SVM have been developed on a solid base of statistical learning theory and are designed especially to provide high flexibility for approximating class boundaries while avoiding over-fitting phenomena (Guo et al., 2005). Functionally, SVM seek to find an optimal hyperplane with the maximal margin separating presence and absence training point classes (this is called a "two-class" case), or a series of hyperplanes around presence training points (this is called a "one-class" case) in multidimensional space (Cristianini & Scholkopf, 2002; Huang, Davis, & Townshend, 2002). These multidimensional classes are then used to map a niche across a landscape. SVM are able to handle non-linear and categorical data, they make no assumption about the probability density of the data, and are competitive with the best available machine learning algorithms in classifying high-dimensional datasets. SVMs are new tools, and as yet not commonly used in ecological niche modeling; one example is Guo et al. (2005), who used one- and two-class SVM to model SOD risk in California, and the authors comment on the utility of one-class SVMs in cases where presence-only data is available.

These models differ in numerous ways (characteristics for these five models are summarized in Table 1), but all can be used in a spatial framework, using geographically referenced data "layers" as explanatory inputs. For example, with the exception of the Rule-based model, the models are inferential, but some are parametric and some are non-parametric. All inferential models with the exception of one-class SVM require presence and absence data to train the model, the SVM provides a means to generate an environmental niche based on presence data alone (called a "one-class" case). The outputs also differ: the Rule-based model produces a ranked mapped output that can change depending on how input values are weighted and combined; the LR model is

Table 1
Characteristics of the five classes of models used

Model name	Absence data required?	Parametric/non-parametric	Important variable selection	Output
Rule-based	No	Non-parametric	*a priori*	Ranked
Logistic Regression	Yes	(semi-) Parametric	Through training	Probability
CART[a]	Yes	Non-parametric	Through training	P/A based on # runs
GA[b]	Yes	Both	Through training	P/A based on # runs
SVM[c]	No (one-class) Yes (two-class)	Non-parametric	Through training	P/A based on # runs

[a] Classification and Regression Tree.

[b] Genetic Algorithm.

[c] Support Vector Machines.

deterministic and produces one map of probability; and the CART, GA and SVM produce binary presence/absence maps that can be combined after multiple runs.

2. Methods

We developed five spatial models using common nationwide spatial data. All spatial data were maintained in Albers Conical Equal Area projection (NAD83, GRS80, meters, Parallels: 29.5°N, 45.5°N, Central Meridian: 96°W, Origin: 23°N). All explanatory variables were maintained in GRID format, the training data was originally a point shapefile, and transformed to either x, y locations or a binary grid for use in models. SVM, CART and LR required this data in x, y format and GARP required it in gridded format; the Rule-based model did not require training data.

2.1. Database development

2.1.1. Explanatory ecological variables

Physical data. Physical variables included topography and climate: we used Digital Elevation Model (DEM) and DAYMET weather and climatologically modeled raster surfaces gridded at 1 km to summarize physical conditions for the pathogen. DAYMET is an assortment of climate raster surfaces interpolated from ground-based meteorological stations (interpolation considers station density, elevation, daylight, and incident solar radiation) on a daily basis over an 18 year period (1980–1997) yielding 1-km resolution data (Thornton, Running, & White, 1997). Data were downloaded (http://www.daymet.org/) as grids of 18-year means or as monthly means over the 18-year period. The primary climate surfaces utilized in this modeling project included total annual precipitation, total annual frost days, average minimum temperature, average maximum temperature and average maximum august temperature. Numerous other variables were evaluated early in the project, and discarded. Topography was derived from GTOPO30, a global digital elevation model (DEM) with a horizontal grid spacing of 30 arc sec (approx. 1 km) that was derived from multiple raster and vector sources. The United States portion was derived from USGS DEM data United States Geological Survey National Elevation Dataset (U.S.G.S., 1999b). BIL data were downloaded (http://edc.usgs.gov/products/elevation/gtopo30/gtopo30.html) and imported into Arc/Info GRIDs using ArcTools (ESRI, 2004), and mosiacked to create a seamless 1-km resolution dataset.

Host/vegetation data. We had a considerable challenge finding a detailed vegetation map for the conterminous US with sufficient floristic and spatial detail to allow modeling. We explored the utility of four datasets: (1) National Land Cover Data (NLCD), (2) USGS digital tree range maps, (3) FIA Percent basal area estimates, and (4) EPA Ecoregion Level 3 data, and ended up using the NLCD and FIA data to refine the modeled results, and the EPA Level 3 Ecoregion data to ensure that the areas of high risk had a strong ecological rationale. The USGS digital tree range maps were not used.

The first among these was the National Land Cover Data (NLCD) dataset, a 21-category classification derived primarily from Landsat Thematic Mapper (TM) imagery from 1992 (Vogelmann, Sohl, & Howard, 1998; Vogelmann, Sohl, Campbell, & Shaw, 1998; Vogelmann et al., 2001). The NLCD classification supplies high spatial resolution (30-m) but poor floristic detail, with only three general vegetation categories relevant to our project: deciduous forest (areas dominated by trees where 75% or more of the tree species shed foliage simultaneously in response to seasonal change), evergreen forest (areas dominated by trees where 75% or more of the tree species maintain their leaves all year) and mixed forest (areas dominated by trees where neither deciduous nor evergreen species represent more than 75% of the cover present). Each conterminous state's NLCD image was downloaded (http://landcover.usgs.gov/natllandcover.asp) and converted into an Arc/Info GRID. Hosts for *P. ramorum* exist in all of the three forest categories, but we decided to enhance the hardwood forests, and created a 1-km gridded vegetation dataset depicting "hardwood density" based on the percent of deciduous and mixed forest 30-m cells found within each resampled 1-km cell.

The second dataset we investigated was the digital tree range maps for North America created by the USGS for a vegetation climate modeling study (U.S.G.S., 1999a), which provided more floristic detail, but was coarse in spatial detail. This product was based upon a series of tree range maps assembled by Elbert L. Little, Jr. in the 1970s as the "Atlas of the United States Trees" (Little, 1971, 1976, 1977; U.S.G.S., 1999a), and digitized by the USGS. Of the 58 digital oak species maps created by USGS (http://climchange.cr.usgs.gov/data/atlas/little/), we determined that 34 were potentially susceptible to SOD. Digital versions of these 34 oak species maps were then combined with digital maps of 12 other tree range maps of species found to either be directly susceptible to the pathogen or to be related (i.e., in the same taxonomic Genus as a susceptible species). The 46 tree range maps were then combined to form a "hardwood diversity index" map by summing the number of susceptible species per pixel. It should be emphasized that the hardwood diversity index as calculated for this study was limited to a portion of the tree range maps made available by the USGS, and contains only a minimal number of shrub or understory species. It is only intended to represent areas within the US that potentially contain high numbers of susceptible SOD host species (both foliar and terminal hosts) with the recognition that there are many more species not included.

The third vegetation layer examined was provided by the USFS Northeastern Research Station (Gottschalk et al., 2002). In this product, Forest Inventory and Analysis (FIA) plot data for the eastern US were used to calculate the percentage of forest basal area composed of the red and live oak groups, and these points were kriged to create a continuous raster surface for the eastern United States. Percent basal area estimates were adjusted for forest density using the NLCD dataset.

Finally, we also used Environmental Protection Agency Ecoregion Level 3 data (ECOMAP, 1993; Omernik, 1987, 1995) as a post-model "screen" to evaluate the ecological rationale for model results. Ecoregions are defined to be areas within which geology, physiography, vegetation, climate, soils, land use, wildlife, and hydrology are similar. Ecoregion Level 3 descriptions include information on climatic regime, topography and predominant vegetation alliances, with information on dominant species. A shapefile containing the Level 3 Ecoregions for the conterminous US was downloaded (http://www.epa.gov/), brought into the ArcMap GIS and overlayed on the model results.

All raster layers (climactic variables, host data, and topographic data) were resampled to 1-km resolution and clipped with a detailed US boundary vector layer in which coastal islands were removed.

2.1.2. Presence/absence data

Four of the five models (GA, LR, CART and SVM) required both presence and absence data for model training (we ran a two-class SVM model). The use of *P. ramorum* presence data is straightforward. In California, presence of *P. ramorum* is determined by either the California Department of Food and Agriculture or University of California at Davis using identical isolation techniques. The protocol involves growing mycelia of *Phytophthora* species from sampled leaf tissue, and identification of *P. ramorum* based on morphological characteristics (Davidson, Werres, Garbelotto, Hansen, & Rizzo, 2003). Samples are collected in the field, and in most cases, Global Positioning Systems (GPS) locations are recorded on a Pest Record Form indicating location of the sample. The complete dataset of *P. ramorum* occurrences is maintained by the University of California at Berkeley OakMapper project (Kelly, Tuxen, & Kearns, 2004). Four hundred and eighty seven confirmed SOD locations were obtained from this database. Of these points, many were redundant in that the field personnel recorded one GPS point for an entire forest stand. These redundant points were removed from the dataset, leaving 266 points. To avoid pseudo-replication, we retained only one point per 1 km^2 area (the modeling resolution), leaving 169 points for the predictive models. This point shapefile was converted to a binary grid (for modeling with GARP), or used to generate a table of presence and absence locations with corresponding climactic variables (for modeling with SVM, CART and LR). Since the prediction was taking place across the United States and we had a limited number of points, a subset of the presence points was not removed for testing the accuracy of the models.

We chose not to use absence data from the field collections for three reasons. First, correct isolation of *P. ramorum* can depend on sampling technique and timing (Davidson et al., 2003), and the process can yield false negatives. Second, the pathogen is an invasive one; there are areas that had been sampled and negative for the pathogen in the past that are now infested. Consequently, absence data are often unreliable or meaningless for modeling invasive species (Hirzel, Hausser, Chessel, & Perrin, 2002; Hirzel, Helfer, & Metral, 2001). Finally, the disease appears to be patchy across a landscape due to landscape heterogeneity and possible plant resistance (Kelly & Meentemeyer, 2002; Rizzo, 2003), thus accurate negative sample can be taken from an infested site. Given these considerations, we generated 'pseudo-absence' data for the models in the following manner. We created a zone of infestation within California consisting of the infested counties, and their border counties. We then generated random "pseudo-absence points" ($n = 169$) from locations in California outside this zone using a random point generator written in Visual Basic Application for ArcGIS (Fig. 2). These pseudo-absence points were generated 100 times for 100 different model runs. These pseudo-absence points were then used for LR, SVM, and CART. Desktop GARP, which we used for GA model, has a built-in function to generate pseudo-absence points. The slight differences in generating pseudo-absence points between GARP and other models (LR, SVM, CART) may attribute to some of the prediction differences described in later section.

2.2. Model development

We first developed our nationwide Rule-based model using similar input data to those used in the California model (Meentemeyer et al., 2004) for the conterminous United States. Climate variables for six winter months were parameterized and placed into weighted classes in accordance with the methods of Meentemeyer et al. (2004). Two different coarse-resolution vegetation maps (hardwood diversity and hardwood density) were used as a surrogate for the detailed vegetation map used in the California modeling case.

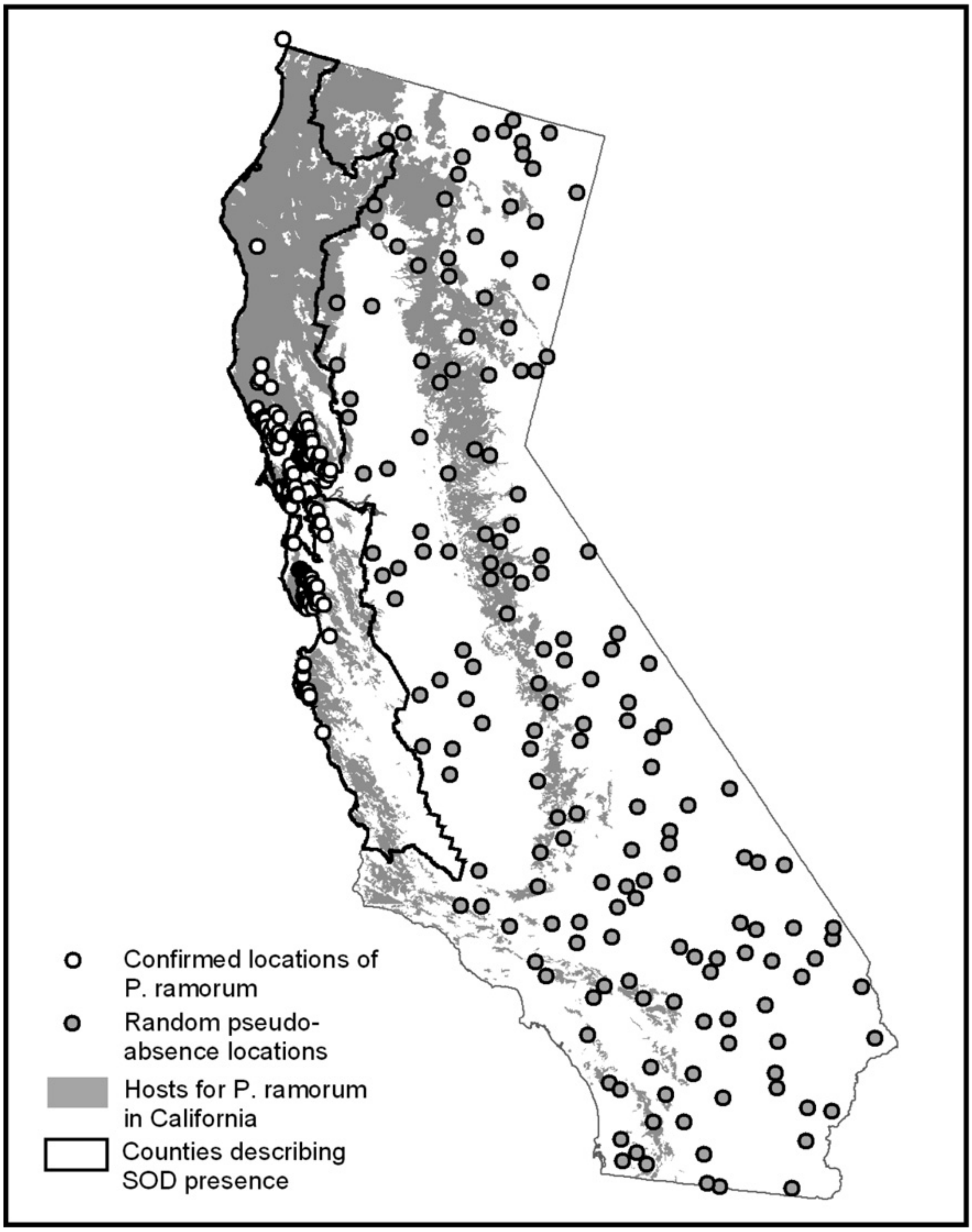
Confirmed locations of P. ramorum
Random pseudo-absence locations
Hosts for P. ramorum in California
Counties describing SOD presence

Fig. 2. Location of *Phytophthora ramorum* occurrences and pseudo-absence locations used in the modeling.

It became clear at this stage that the topographic data and the vegetation information would not be useful as direct model inputs in the subsequent modeling exercises. Topography did not appear to be a strong control on disease at larger scales; as Rizzo and Garbelotto (2003) point out, the disease exists in niches from sea level to 800 m, and slope and aspect appeared insignificant at the national scale. Also, detailed vegetation maps are not available for every state, and the coarseness of our nationwide data, either in spatial or floristic terms was problematic. We thus considered climatic niches to be the dominant controller for *P. ramorum* establishment, and interpreted the results according to logical associations of vegetation.

Our CART model utilized Splus v. 6.2 for Windows. We generated 100 classification "trees" using 100 unique pseudo-absence point distributions and the 169 SOD presence points. Each "tree" was pruned to its appropriate size by examining a plot of deviance and tree complexity (Feldesman, 2002), and the resulting tree models were implemented in ArcInfo using Arc Macro Language (AMLs) as 100 binary "presence" or "absence" maps. These were summed to produce a 0–100 scored presence map. We then developed a LR equation using Splus v. 6.2 © for Windows © and mapped the probabilities (0–100%) in ArcInfo. Desktop GARP (Stockwell, 2006; Stockwell & Peters, 1999b) software was used for the application of the GA model. 100 model runs were performed (the model generates its own "pseudo-absence" points for each run). Finally, we developed SVM models using Matlab© and LIBSVM software (Chuang & Lin, 2001). Cross-validation was used for each of the 100 runs to optimize parameter selection.

The initial model inputs (with the exception of the Rule-based model) were limited to climate variables, since we had decided the topographic data provided no discriminating information with respect to the environmental niche of *P. ramorum* at the continental scale, and the host data was too floristically or spatially coarse to be of use. The four predictive models (excluding the Rule-based model) were run with a range of input predictor variables including temperature, precipitation, humidity and radiation. Precipitation total, frost days, average maximum temperature and average minimum temperature were determined to be the most important variables in predicting the niche for *P. ramorum*.

2.2.1. Risk weightings

All model results were normalized for visual display purposes using the following technique. The mean value was calculated and then boundaries were set for plus or minus two standard deviations. Any values above or below were reclassed to the minimum or maximum of the 95th percentile. This grid was then rescaled from 0 to 100, and classed into five classes of risk: 0–20% low, 20–40% medium low, 40–70% medium, 70–90% medium high, and 90–100% high.

2.3. Accuracy assessment

A simple metric of model accuracy was assessed using the original 169 training points. A point intersect tool was used with each model grid, and the percentage of points falling into each of the five risk classes (high, moderately high, medium, moderately low, and low) was calculated. Accuracy of the models outside of CA and OR could not be attempted, as there are no positive cases in the US of *P. ramorum* outside of the west coast of the US.

2.4. Model combination and filtering

We combined the results from the five models together to create a final map based on model agreement. First, we added together the five (un-classed) model results per pixel, giving each of the five models equal weighting. We then used the same rescaling method as described above. Second, in an effort to eliminate areas of non-hardwood forest in the final risk maps, we filtered the combined model results through the NLCD and FIA red oak basal area vegetation maps by multiplying the map by each vegetation map rescaled from 0 to 1. We also examined model results in relation to EPA Level 3 Ecoregion data in order to ensure that the areas of high risk had a strong ecological rationale. We did not include the USGS vegetation map in this exercise, as the spatial fidelity of the product seemed problematic. The representation of hardwood diversity in the southeast US may not be accurate due to the relatively small number of tree ranges used.

3. Results

The results from each individual model are shown in Fig. 3. The Rule-based model shows a high risk for SOD spread across the southeastern US from eastern Texas to Virginia including Florida, with risk declining to the north and west. The model also shows moderately high risk in the coastal northwest, and risk in the northern foothills of the Sierra Nevada Mountains in California. While this model is a copy of that provided by the Meentemeyer et al. (2004) model, our different results in California are likely due to qualitative differences in input data and spatial resolution. The LR results show a broadly similar pattern to the Rule-based model, with less risk on the west coast, and less overall risk in the southeast: the model constrains the highest risk to the southern states of Louisiana, Alabama, and Mississippi with risk declining in a northerly and easterly direction

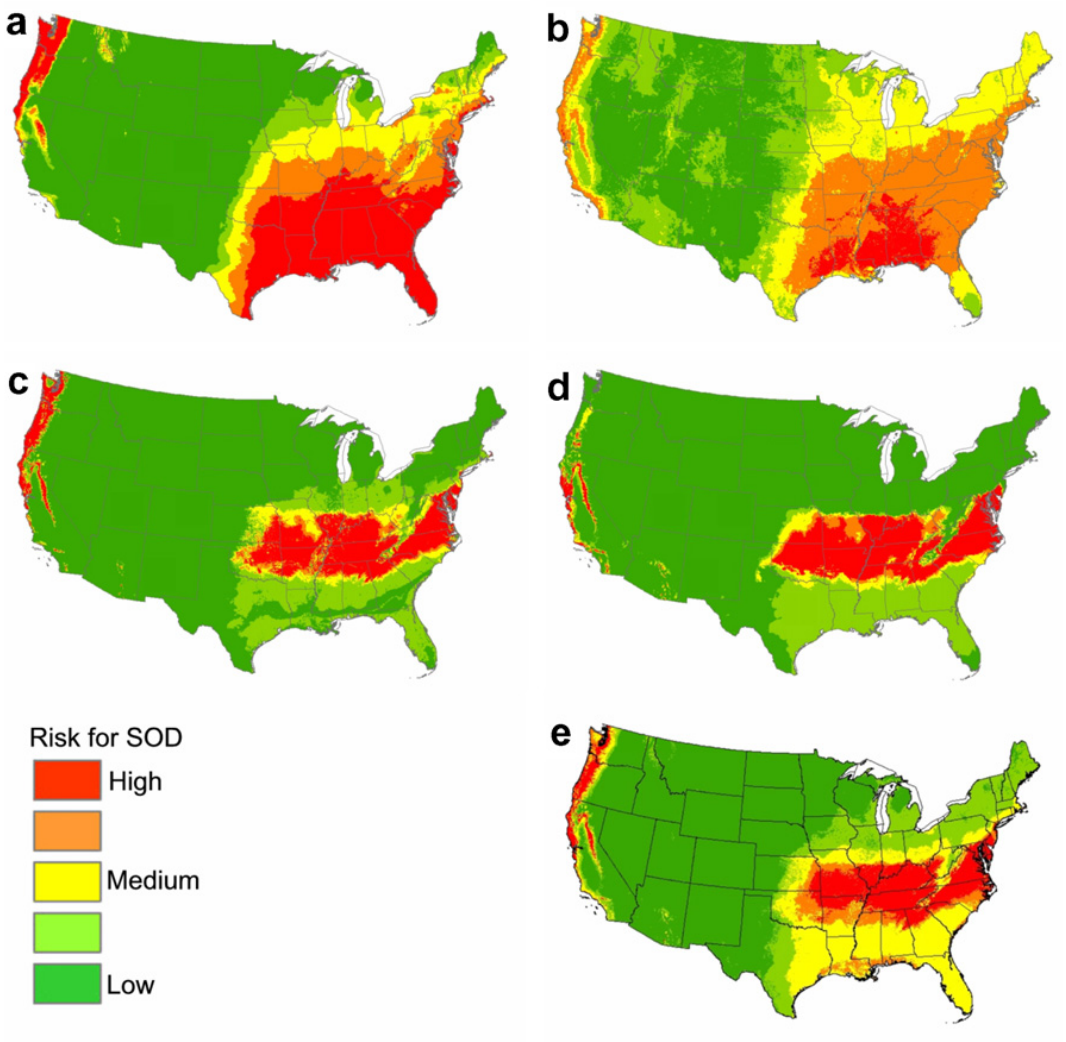

Fig. 3. Risk for Sudden Oak Death in the conterminous United States from five spatially referenced models: (a) Rule-based, (b) Logistic Regression, (c) Classification Tree, (d) Genetic Algorithm, and (e) Support Vector Machine.

from there. The LR formula determined that precipitation total, average minimum temperature and average maximum temperature were important in risk prediction. CART and GA results are similar, with a band of highest risk occurring throughout the middle southeast, from Oklahoma in the west to Virginia and North Carolina in the east. Both models found precipitation total, frost days, and average maximum temperature to be the most important predictors. The SVM result captures some of the patterning from all the other models, with risk in a west–east band across the southeast (as with the CART and GA models), and with some moderate risk for the disease in the southern states (as with the Rule-based and LR models). The SVM algorithm predicted risk as a result of precipitation total, frost days and average maximum temperature.

The combined model output (with each model given equal weight) show that the models agree on an area of high-risk for SOD in roughly 500,000 km^2 of the conterminous United States (Fig. 4). All models predict risk for the disease in coastal California and in the northern foothills of the Sierra Nevada mountains in California. Away from the west coast, SOD risk appears in all models across an east–west oriented band including the hardwood forests of Oklahoma, Arkansas, Tennessee, Kentucky, and in the northern portions of Mississippi, Alabama, Georgia and South Carolina, parts of central North Carolina, eastern Virginia, Delaware and Maryland.

We filtered the combined model with two vegetation datasets, and evaluated it with EPA Ecoregion Level 3 data (Fig. 5). Use of the NLCD data constrains the overall risk somewhat, with higher risk areas remaining in hardwood forests in Arkansas, Tennessee, southern Kentucky and northern Alabama. A more limited map of potential risk is depicted when the final combined model is filtered with the FIA data red oak basal area map; higher risk clusters are scattered across the southeast, with the largest clusters found in southern Missouri and northern Arkansas in the Ozark Mountains. Interpretation of the final map with the EPA Ecoregion Level 3 data is more useful. The high-risk area includes portions of several ecoregions, including the Piedmont ecoregion of North Carolina, South Carolina, Georgia, Alabama and Mississippi, a transitional area between the mostly mountainous ecoregions of the Appalachians to the northwest and the rela-

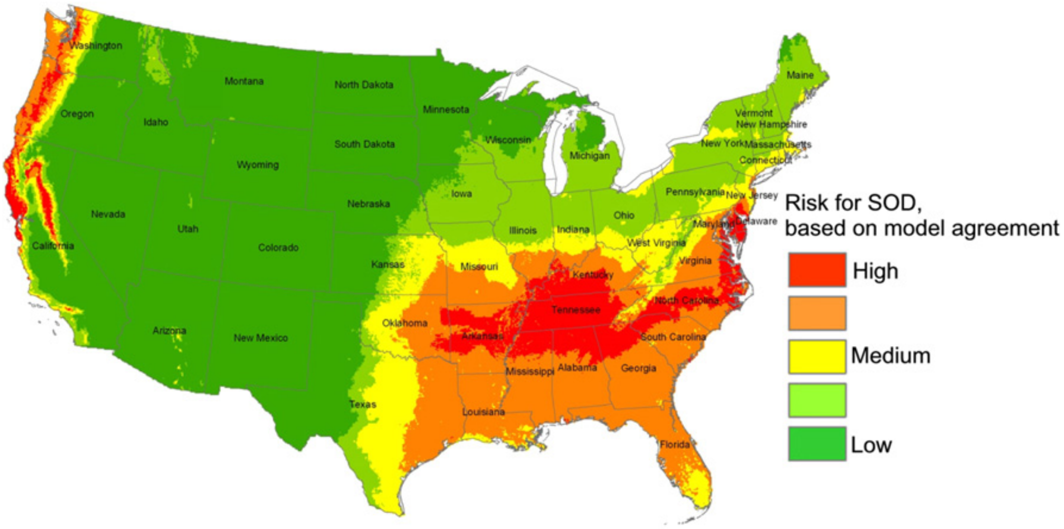

Fig. 4. Risk for Sudden Oak Death in the conterminous United States based on agreement between five spatially referenced models.

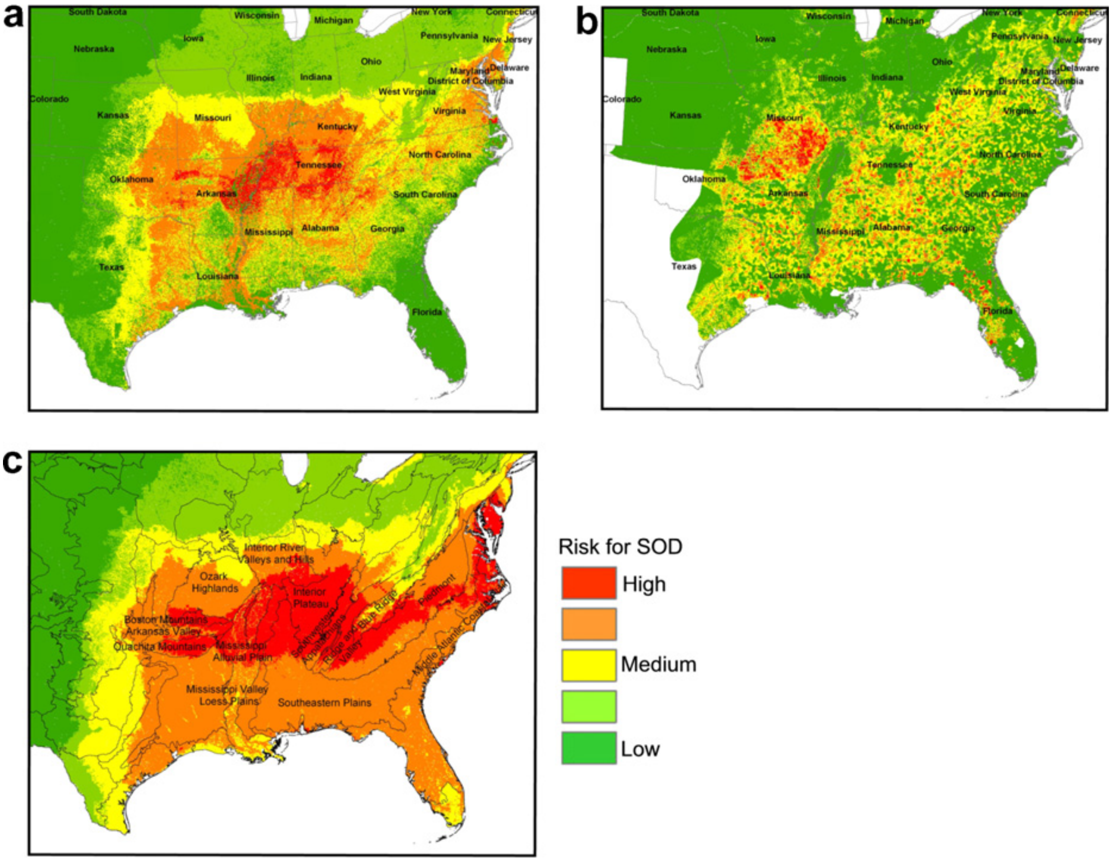

Fig. 5. Southeastern US, showing the climatic niche results constrained by vegetation data: (a) NLCD, (b) FIA red oak basal area, and EPA Level 3 Ecoregions.

tively flat coastal plain to the southeast. Much of this region has reverted to successional pine and hardwood woodlands, with an increasing conversion to an urban and suburban land cover (ECOMAP, 1993; Omernik, 1987, 1995). There is also predicted risk for SOD in hardwood forests of the Southwestern Appalachians in Tennessee, and in the primarily oak–hickory forests of the Interior Plateau in Kentucky and Tennessee. Eastern Oklahoma and central Arkansas shows high risk, in the oak–hickory–pine forests of the Ouachita Mountains, and in the red oak, white oak, and hickory dominated forests of the Boston Mountains, and in the predominantly oak forests of the Southern Ozarks (ECOMAP, 1993; Omernik, 1987, 1995). Coastal Maryland and Delaware, part of the Middle Atlantic Coastal Plain are climactically susceptible, and have forests at risk in riparian areas.

We cannot assess accuracy for the models in areas outside of California and southern Oregon with field data (as there are no positive cases of *P. ramorum* outside of CA and OR), and accuracy of the models varied widely within that area (Table 2). The most accurate model (defined as the number of positive samples falling into the highest risk area) was the SVM model, with 98.8% of sample points falling into the highest risk area in California and southern Oregon. The other newer modeling tools also performed well (CART and GA had 96% and 95% respectively of confirmed occurrences of *P. ramorum* in high risk areas). The Rule-based model had poorer results, with no positive *P. ramorum* samples falling in the modeled high-risk area; most points (61.5%) fell in the moderately high risk areas on the west coast. The LR was the poorest performer overall; only 5% of positive *P. ramorum* occurrences were mapped on high risk areas, 26.3% of points fell into the

Table 2
Accuracy assessment of all models, using 169 presence points from California and Oregon: number represents percentage of training points found in each risk class

Model type						
Risk level	Rule-based	LR	CART	GA	SVM	Combination model
High	0	5.5	96	95	98.8	89.3
	61.5	20.8	0.5	2	0	8.9
Medium	38.5	46.3	1.8	2	0	1.8
	0	24.9	0.5	1	0	0
Low	0	2.5	1.2	0	1.2	0

moderately high area, and well over half of positive *P. ramorum* occurrence points fell in the medium-risk and lower-risk categories. The final combination model had a high overall accuracy (89.3% of locations of confirmed presence of *P. ramorum* were found in modeled high risk areas, and 98.2% of points were found in both high and moderately high risk areas), but was slightly poorer than the GA, CART and SVM models.

4. Discussion and conclusions

The difference in model results can be explained by a number of factors. First, we have a geographically constrained training sample, and are modeling across a large geographic scope. With such a small training sample (as is the case with a new biotic invasion), variance is large, and we would expect the models have different results. Second, there are general trends in the models that can be examined in addition. The results are easily split into three broad patterns: (1) the Rule-based and LR models, which tended to predict larger distributions of risk (across the southern states of the United States east of the Mississippi River with risk declining northwards) and had lower overall accuracies, (2) the GA and CART models, which predicted a tighter geographic band of risk (running from the Ozarks east along the Appalachian foothills) with higher accuracies; and (3) the SVM model, which had patterns from both other sets (displaying the tight band of west–east risk through the mid southeast similar to the GA and CART models, and also predicting more risk through the southeast coastal plain, similar to the rule-based and LR models) and the highest accuracy overall. Model mechanics, and in particular, how each model deals with overfitting given a small sample size, explain much of these differences.

Each of the five models we examined deals with overfitting differently, but each of the three machine learning models (GA, CART and SVM) have explicit tools for minimizing overfitting. For example, CART avoids overfitting by pruning the "trees" so that the "tree" structure is not overly large and complex. GA avoids overfitting by implementing a cross-validation approach, in which the whole dataset is divided into training and testing data. SVM minimizes classification errors and at the same time, constrains the model complexity to avoid overfitting. In contrast, the LR model is less useful as it attempt to minimize empirical errors based on training data from a small area (e.g., California) without considering overall model generalization (across the US), and likely under-fits the data. The Rule-based model is a special case; it does not technically "fit" a model as it is not inferential and so has no mechanism to avoid overfitting, but it could be described as developing a model from expert knowledge. Indeed, our accuracy assessment suggests that the Rule-based model and LR model did not capture the complexity of data because of their simpler rule structure or parametric function form whereas the GA, CART, and SVM models were able to derive more structure from the small training sample.

The SVM model deserves more discussion. It was the most accurate model and it captured some of the results from all models, demonstrating capacity without overfitting. The model is also known to generalize unseen data well, which is exactly the case early in a biotic invasion. In light of these reasons, we suggest that SVM is a robust, accurate and easily implemented choice for modeling the potential niche of a biotic invasion when multiple model comparisons are not possible.

Despite their differences, these kinds of predictive environmental niche model results are useful for the management of Sudden Oak Death in several ways. The results can be used to raise concern about the potential for SOD establishment and spread in the southeastern US. There is clearly a large area potentially at high-risk for the disease (around 500,000 km^2); given that the pathogen can be dispersed via wind-driven rain and that there have been past shipments of possibly infected stock to nurseries in the area, we should be concerned about hurricanes and other storm events potentially moving the pathogen. In addition, the risk maps can be combined with refined state vegetation data to target monitoring efforts or overflights, as has been the case in California (Meentemeyer et al., 2004). Many of the states in high-risk areas have floristically and spatially precise spatial data that can be used in this way. Finally, the maps can also be used to target areas for public outreach efforts. A considerable amount of early reconnaissance of the disease in California was facilitated through public awareness of the disease, its symptoms, and its potential dramatic effects (Carlsen, 2003; Kelly, 2001; Kelly & Tuxen, 2003).

There are other issues raised by this work that are common to other efforts modeling invasive species dynamics. First, because there are no wildland cases of SOD outside of California and Oregon, none of these models can be adequately assessed for accuracy. This is unfortunate, but a common situation when modeling invasive species (Muñoz & Felicísimo, 2004). Three of our five initial models, and the two combination models fit well in coastal California and Oregon, but that fact helps little in the areas of concern outside of the west coast. Several of the models allow for some form of cross-fold validation tools for assessing accuracy, but these tools can be problematic due to the small number of training samples, and their concentrated distribution (Graham et al., 2004). A small number of spatially concentrated samples are typical in cases of invasive organisms in the introductory phase.

Second, the generation of pseudo-absence data must also be examined. We do not have reliable negatives for *P. ramorum*, so we used a common method for generation of pseudo-absence data, constraining the pool of possible absence points to be taken from outside the zone of infestation. The lack of scientific knowledge regarding the ecological and physiological boundaries of the pathogen (outside of a laboratory environment), the unknown probabilities of pathogen absence due to competition, lack of dispersal or lack of detection, and its long-distance dispersal potential (via ornamental species in nurseries) make any delineation of presence/absence zones within potential host ranges a somewhat arbitrary process. Experiments with pseudo-absence data generated within the zone of infestation resulted in models that over-predicted the risk of the disease. Clearly, the choice of pseudo-absence data generation can have an influence on the end-product. Another possible method of approaching this problem would be to use a model requiring only presence data, such as one-class Support Vector Machine (SVM) or Environmental Niche

Factor Analysis (ENFA); these models can be used to directly map an environmental niche (Guo et al., 2005), or to create a predicted probability surface outside of which 'pseudo' absence points can be generated (Muñoz & Felicísimo, 2004).

Third, while environmental niche models can capture similarities to potential niches outside of a current distribution, one obvious problem with the fundamental assumptions of this predictive model is that all presence and absence points are located within the ecological and climatic confines of a select range of tree habitats in California, that might not be adequately modeled. Factors such as coastal influences of the Pacific Ocean, weather regimes of the west coast, orographic climatic impacts of the Coast Ranges and Sierra Nevada Mountains, and the biogeography of hardwood and conifer distributions are possibly not adequately modeled using these input data, influencing predictions of SOD risk elsewhere in the country.

Finally, the host data available for the entire United States was the largest limiting factor in our modeling exercise; all nationwide vegetation layers we used had significant drawbacks. Climate, vegetation and topography are common inputs for these types of ecological niche models (Parra et al., 2004), but different models have different levels of success with various combinations. For example, Parra et al. (2004) had trouble using remotely sensed vegetation indices as a proxy for vegetation cover, and relied more on climatic variables for modeling bird habitat in the Andes. We faced (as do all others modeling with vegetation data at this scale) a trade-off between spatial and floristic detail. Specifically, the NLCD data was the most spatially comprehensive layer, with compete coverage at a high spatial resolution; however, specific floristic detail was absent, and the vegetation classes were much too broad to be of great use in the modeling exercise. The USGS layer and the EPA Level 3 ecoregion layers have sufficient floristic detail, but they are tremendously course in resolution. Finally, the FIA product only covered the east coast area, and thus could not be used in the models that required training. A similar west coast product is not currently available. Although the United States is a geospatial-data-rich country, floristically detailed coverages that are consistent across the US are needed to improve the quality of this work. The requirements of consistency and detail are what limited our use of existing datasets. That said, detailed host data on a state-by-state framework can be used in areas of high risk.

This work examined common ecological niches for *P. ramorum*, but an investigation of the human component to disease establishment and spread should also be considered. The locations of wholesale and retail nurseries to which infested stock has been sent is a necessary future important component to this research. Although theoretical in nature, the results of this paper have practical, applied value for managers and regulators of this disease.

Acknowledgement

This work was supported through funding from the USDA Forest Service Sudden Oak Death Science Program through a grant to M. Kelly.

References

Anderson, R. P., Lew, D., & Peterson, A. T. (2003). Evaluating predictive models of species' distributions: criteria for selecting optimal models. *Ecological Modelling, 162*, 211–232.

Arnold, W. S., White, M. W., Norris, H. A., & Berrigan, M. E. (2000). Hard clam (*Mercenaria* spp.) aquaculture in Florida, USA: geographic information system applications to lease site selection. *Aquacultural Engineering, 23*(1–3), 203–231.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees*. New York: Chapman and Hall.

Buitrago, J., Rada, M., Hernandez, H., & Buitrago, E. (2005). A single-use site selection technique, using GIS, for aquaculture planning: choosing locations for mangrove oyster raft culture in Margarita Island, Venezuela. *Environmental Management, 35*(5), 544–556.

Carlsen, S. (2003). Sudden Oak Death in Marin County: a case study of community impacts. Available from http://www.apsnet.org/online/SOD.

Chuang, C.-C., & Lin, C.-J., 2001. LIBSVM – A library for Support Vector Machines.

Clark, L. A., & Pregibon, D. (1993). Tree-based models. In J. M. Chambers & T. J. Hastie (Eds.), *Statistical models*. London: Chapman & Hall, Inc.

Clark, M. E., Rose, K. A., Levine, D. A., & Hardgrove, W. W. (2001). Predicting climate change effects on Appalachian trout: combining GIS and individual-based modeling. *Ecological Applications, 11*(1), 161–178.

Costa, J., Peterson, A. T., & Beard, C. B. (2002). Ecologic niche modeling and differentiation of populations of Triatoma brasiliensis neiva, 1911, the most important Chagas' disease vector in northeastern Brazil (*hemiptera*, *reduviidae*, *triatominae*. *American Journal of Tropical Medicine and Hygiene, 67*(5), 516–520.

Cree, L. (2003). Risk assessment: a tool for decision making. Available from http://www.apsnet.org/online/SOD.

Cristianini, N., & Scholkopf, B. (2002). Support vector machines and kernel methods – The new generation of learning machines. *Ai Magazine, 23*(3), 31–41.

Cushman, J. H., & Meentemeyer, R. K. (2005). *The role of humans in the dispersal and spread of Phytophthora ramorum*. Paper presented at the Second SOD Science Symposium, Monterey, CA.

Davidson, J.M., & Shaw, C.G. (2003). Pathways of movement for *Phytophthora ramorum*, the causal agent of Sudden Oak Death. Available from http://www.apsnet.org/online/SOD.

Davidson, J. M., Rizzo, D. M., Garbelotto, M., Tjosvold, S., & Slaughter, G. W. (2002, October 22–25, 2001). *Phytophthora ramorum and sudden oak death in California: II. Transmission and survival*. Paper presented at the Fifth Symposium on Oak Woodlands, San Diego, CA.

Davidson, J. M., Werres, S., Garbelotto, M., Hansen, E. M., & Rizzo, D. M. (2003). Sudden Oak Death and associated diseases caused by *Phytophthora ramorum*. Unpublished manuscript.

Davidson, J. M., Wickland, A. C., Patterson, H., Falk, K., & Rizzo, D. M. (2005). Transmission of *Phytophthora ramorum* in mixed-evergreen forests in California. *Phytopathology, 5*, 587–596.

De'ath, G., & Fabricius, K. E. (2000). Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology, 81*(11), 3178–3192.

ECOMAP (1993). *National hierarchical framework of ecological units*. Washington, DC: USDA Forest Service.

Englander, L., & Tooley, P. (2003). Plants hosts in the nursery industry: how might the movement of plants in the nursery industry contribute to the spread of *Phytophthora ramorum* to new areas? Sudden Oak Death Online Symposium.

ESRI (2004). *ArcGIS software*. Environmental Systems Research Institute, Inc.

Fabricius, K., & De'ath, G. (2001). Environmental factors associated with the spatial distribution of crustose coralline algae on the Great Barrier Reef. *Coral Reefs, 19*, 303–309.

Feldesman, M. R. (2002). Classification trees as an alternative to Linear Discriminant Analysis. *American Journal of Physical Anthropology, 119*, 257–275.

Felicisimo, A. M., Frances, E., Fernandez, J. M., Gondalez-Diez, A., & Varas, J. (2002). Modeling the potential distribution of forests with a GIS. *Photogrammetric Engineering & Remote Sensing, 68*(5), 455–462.

Fleishman, E., Nally, R. M., Fay, K. P., & Murphy, D. D. (2001). Modeling and predicting species occurrence using broad-scale environmental variables: an example with butterflies of the Great Basin. *Conservation Biology, 15*(6), 1674–1685.

Franklin, J. (1995). Predictive vegetation mapping: geographic modelling of biospatial patterns in relation to environmental gradients. *Progress in Physical Geography, 19*(4), 474–499.

Godown, M. E., & Peterson, A. T. (2000). Preliminary distributional analysis of US endangered bird species. *Biodiversity and Conservation, 9*(9), 1313–1322.

Gottschalk, K. W., Morin, R. S., & Liebhold, A. M. (2002). *Potential susceptibility of eastern forests to Sudden Oak Death, Phytophthora ramorum*. Paper presented at the USDA Forest Service Forest Health Monitoring (FHM) Conference.

Graham, C. H., Ferrier, S., Huettman, F., Moritz, C., & Peterson, A. T. (2004). New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology and Evolution, 19*(9), 497–503.

Graham, C. H., Moritz, C., & Williams, S. E. (2006). Habitat history improves prediction of biodiversity in rainforest fauna. *Proceedings of the National Academy of Sciences of the United States of America, 103*, 632–636.

Guisan, A., & Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling, 135*, 147–186.

Guo, Q., Kelly, M., & Graham, C. (2005). Support vector machines for predicting distribution of Sudden Oak Death in California. *Ecological Modelling, 128*(1), 75–90.

Higgins, S. I., Richardson, D. M., & Cowling, R. M. (2001). Validation of a spatial simulation model of a spreading alien plant population. *Journal of Applied Ecology, 38*, 571–584.

Hirzel, A. H., Hausser, J., Chessel, D., & Perrin, N. (2002). Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? *Ecology, 83*, 2027–2036.

Hirzel, A. H., Helfer, V., & Metral, F. (2001). Assessing habitat-suitability models with a virtual species. *Ecological Modelling, 145*(2–3), 111–121.

Huang, C., Davis, L. S., & Townshend, J. R. G. (2002). An assessment of support vector machines for land cover classification. *International Journal of Remote Sensing, 23*(4), 725–749.

Huang, Z., & Lees, B. G. (2004). Combining non-parametric models for multisource predictive forest mapping. *Photogrammetric Engineering & Remote Sensing, 70*(4), 415–425.

Illoldi-Rangel, P., Sanchez-Cordero, V., & Peterson, A. T. (2004). Predicting distributions of Mexican mammals using ecological niche modeling. *Journal of Mammalogy, 85*(4), 658–662.

Iverson, L. R., & Prasad, A. M. (1998). Predicting abundance of 80 tree species following climate change in the Eastern United States. *Ecological Monographs, 68*(4), 465–485.

Jules, E. S., Kauffman, M. J., Ritts, W. D., & Carroll, A. L. (2002). Spread of an invasive pathogen over a variable landscape: a nonnative root rot on Port Orford Cedar. *Ecology, 83*(11), 3167–3181.

Kampichler, C., Barthel, J., & Wieland, R. (2000). Species density of foliage-dwelling spiders in field margins: a simple, fuzzy rule-based model. *Ecological Modelling, 129*(1), 87–99.

Karthik, M., Suri, J., Saharan, N., & Biradar, R. S. (2005). Brackish water aquaculture site selection in Palghar Taluk, Thane district of Maharashtra, India, using the techniques of remote sensing and geographical information system. *Aquacultural Engineering, 32*(2), 285–302.

Kelly, M. (2001). Community involvement needed in monitoring sudden oak death in California. *Oaks 'n' folks, 17*(1), 1–2.

Kelly, N. M., Fonseca, M., & Whitfield, P. (2001). Predictive mapping for management and conservation of seagrass beds in North Carolina. *Aquatic Conservation: Marine and Freshwater Ecosystems, 11*(6), 437–451.

Kelly, M., & Meentemeyer, R. K. (2002). Landscape dynamics of the spread of Sudden Oak Death. *Photogrammetric Engineering & Remote Sensing, 68*(10), 1001–1009.

Kelly, M., & Tuxen, K. (2003). WebGIS for monitoring "sudden oak death" in coastal California. *Computers, Environment and Urban Systems, 27*(5), 527–547.

Kelly, M., Tuxen, K., & Kearns, F. (2004). Geospatial informatics for management of a new forest disease: sudden oak death. *Photogrammetric Engineering and Remote Sensing, 70*(1), 1001–1004.

Little, E. L. (1971). Atlas of United States trees, Vol. 1, conifers and important hardwoods (200 maps No. Miscellaneous Publication 1146): US Department of Agriculture.

Little, E. L. (1976). Atlas of United States trees, Vol. 3, minor Western hardwoods (290 maps No. Miscellaneous Publication 1314): US Department of Agriculture.

Little, E. L. (1977). Atlas of United States trees, Vol. 4, minor Eastern hardwoods (230 maps No. Miscellaneous Publication 1342): US Department of Agriculture.

Llewellyn, D. W., Shaffer, G. P., Craig, N. J., Creasman, L., Pashley, D., Swain, M., et al. (1995). A decision-support system for prioritizing restoration sites on the Mississippi River alluvial plain. *Conservation Biology, 10*(5), 1446–1455.

Manel, S., Dias, J. M., Buckton, S. T., & Ormerod, S. J. (1999). Alternative methods for predicting species distribution: an illustration with Himalayan river birds. *Journal of Applied Ecology, 36*, 734–747.

Manel, S., Dias, J.-M., & Ormerod, S. J. (1999). Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: a case study with a Himalayan river bird. *Ecological Modelling, 120*, 337–347.

McCauley, L. A., & Jenkins, D. G. (2005). GIS-based estimates of former and current depressional wetlands in an agricultural landscape. *Ecological Applications, 15*(4), 1199–1208.

McPherson, B. A., Wood, D. L., Švihra, P., Kelly, N. M., Storer, A. J., & Standiford, R. B. (2005). Sudden oak death in California: disease progression in oaks and tanoaks. *Journal of Forest Ecology and Management, 213*(1–3), 71–89.

McShea, W. J., Koy, K., Clements, T., Johnson, A., Vongkhamheng, C., & Aung, M. (2005). Finding a needle in the haystack: regional analysis of suitable Eld's deer (*Cervus eldi*) forest in Southeast Asia. *Biological Conservation, 125*, 101–111.

Meentemeyer, R., Rizzo, D., Mark, W., & Lotz, E. (2004). Mapping the risk of establishment and spread of sudden oak death in California. *Forest Ecology & Management, 200*(1–3), 195–214.

Michaelsen, J., Schimel, D., Friedl, M., Davis, F. W., & Dubayah, R. C. (1994). Regression tree analysis of satellite and terrain data to guide vegetation sampling and surveys. *Journal of Vegetation Science, 5*, 673–686.

Mladenoff, D. J., Sickley, T. A., & Wydeven, A. P. (1999). Predicting gray wolf landscape recolonization: logistic regression models vs. new field data. *Ecological Applications, 9*(1), 37–44.

Muñoz, J., & Felicísimo, Á. M. (2004). Comparison of statistical methods commonly used in predictive modelling. *Journal of Vegetation Science, 15*, 285–292.

Omernik, J. M. (1987). Ecoregions of the conterminous United States. Map (scale 1:7,500,000). *Annals of the Association of American Geographers, 77*(1), 118–125.

Omernik, J. M. (1995). Ecoregions: A spatial framework for environmental management. In W. S. Davis & T. P. Simon (Eds.), *Biological assessment and criteria: tools for water resource planning and decision making* (pp. 49–62). Boca Raton, FL: Lewis Publishers.

Parra, J. L., Graham, C. C., & Freile, J. F. (2004). Evaluating alternative data sets for ecological niche models of birds in the Andes. *Ecography, 27*, 350–360.

Petit, S., Chamberlain, D., Haysom, K., Pywell, R., Vickery, J., Warman, L., et al. (2003). Knowledge-based models for predicting species occurrence in arable conditions. *Ecography, 26*(5), 626–640.

Pyke, C. R. (2005). Assessing suitability for conservation action: prioritizing interpond linkages for the California tiger salamander. *Conservation Biology, 19*(2), 492–503.

Raxworthy, C. J., Martinez-Meyer, E., Horning, N., Nussbaum, R. A., Schneider, G. E., Ortega-Huerta, M. A., et al. (2003). Predicting distributions of known and unknown reptile species in Madagascar. *Nature, 426*(18/25 December), 837–841.

Rizzo, D. M. (2003). Sudden Oak Death: host plants in forest ecosystems in California and Oregon. Available from http://www.apsnet.org/online/SOD.

Rizzo, D. M., & Garbelotto, M. (2003). Sudden oak death: endangering California and Oregon forest ecosystems. *Frontiers in Ecology and the Environment, 1*(5), 197–204.

Rushton, S. P., Omerod, S. J., & Kerby, G. (2004). New paradigms for modelling species distributions? *Journal of Applied Ecology, 41*, 193–200.

Russell, G., Hawkins, C., & O'Neill, M. (1997). The role of GIS in selecting sites for riparian restoration based on hydrology and land use. *Restoration Ecology, 5*(4S), 56–68.

Sperduto, M. B., & Congalton, R. G. (1996). Predicting rare orchid (small whorled pogonia) habitat using GIS. *Photogrametric Engineering and Remote Sensing, 62*(11), 1269–1279.

Stockwell, D. R. (1999). Genetic Algorithms II. In A. H. Fielding (Ed.), *Machine learning methods for ecological applications* (pp. 123–144). Boston: Kluwer Academic Publishers.

Stockwell, D. (2006). DesktopGarp. [Retrieved September 20, 2006. Available from http://www.lifemapper.org/desktopgarp/.

Stockwell, D., & Peters, D. (1999a). The GARP modelling system: problems and solutions to automated spatial prediction. *International Journal of Geographical Information Science, 13*(2), 143–158.

Stockwell, D. R. B., & Peters, D. P. (1999b). The GARP modelling system: problems and solutions to automated spatial prediction. *International Journal of Geographic Information Systems, 13*, 143–158.

Stokstad, E. (2004). Plant pathology: nurseries may have shipped sudden oak death pathogen nationwide. *Science, 303*(5666), 1959.

Thornton, P. E., Running, S. W., & White, M. A. (1997). Generating surfaces of daily meteorological variables over large regions of complex terrain. *Journal of Hydrology, 190*, 214–251.

Tjosvold, S. A., Chambers, D. L., Davidson, J. M., & Rizzo, D. M. (2002). *Incidence of Phytophthora ramorum inoculum found in soil collected from a hiking trail and hikers' shoes in a California park*. Paper presented at the Sudden Oak Death Science Symposium, Monterey, CA.

Tooley, P. W., & Kyde, K. L. (2003). Susceptibility of some eastern oak species to sudden oak death caused by *Phytophthora ramorum*. *Phytopathology, 93*, S84.

U.S.G.S. (1999a). Digital representation of "Atlas of United States Trees" by E.L. Little Jr.: US Geological Survey.

U.S.G.S. (1999b). *National elevation database (USGS)*. Sioux Falls, SD: US Geological Survey.

Vayssières, M. P., Plant, R. E., & Allen-Diaz, B. H. (2000). Classification trees: An alternative non-parametric approach for predicting species distributions. *Journal of Vegetation Science, 11*, 679–694.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed). New York: Springer.

Venette, R. C., & Cohen, S. D. (2006). Potential climatic suitability for establishment of *Phytophthora ramorum* within the contiguous United States. *Forest Ecology and Management, 231*, 18–26.

Vogelmann, J. E., Howard, S. M., Yang, Y., Larson, C. R., Wylie, B. K., & Van Driel, N. (2001). Completion of the 1990s National Land Cover Data Set for the conterminous United States from Landsat Thematic Mapper Data and ancillary datasources. *Photogrammetric Engineering and Remote Sensing, 67*, 650–652.

Vogelmann, J. E., Sohl, T. L., Campbell, P. V., & Shaw, D. M. (1998). Regional land cover characterization using Landsat Thematic Mapper data and ancillary data sources. *Environmental Monitoring and Assessment, 51*, 415–428.

Vogelmann, J., Sohl, T., & Howard, S. (1998). Regional characterization of land cover using multiple sources of data. *Photogrammetric Engineering and Remote Sensing, 64*(1), 45–57.

Zaniewski, A. E., Lehmann, A., & Overton, J. M. C. (2002). Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. *Ecological Modelling, 157*, 261–280.