GLOBAL CHANGE BIOLOGY
BIOENERGY

GCB Bioenergy (2010), doi: 10.1111/j.1757-1707.2010.01040.x

Using species distribution models to identify suitable areas for biofuel feedstock production

JASON M. EVANS*[1], ROBERT J. FLETCHER, JR.† and JANAKI ALAVALAPATI‡

*Department of Wildlife Ecology and Conservation, IFAS/University of Florida, Gainesville, FL 32611, USA, †Department of Wildlife Ecology and Conservation, 110 Newin-Ziegler Hall, PO Box 110430, University of Florida, Gainesville, FL 32611, USA, ‡Department of Forest Resources and Environmental Conservation, 313 Cheatham Hall, Virginia Tech University, Blacksburg VA 24061, USA

Abstract

The 2007 Energy Independence and Security Act mandates a five-fold increase in US biofuel production by 2022. Given this ambitious policy target, there is a need for spatially explicit estimates of landscape suitability for growing biofuel feedstocks. We developed a suitability modeling approach for two major US biofuel crops, corn (*Zea mays*) and switchgrass (*Panicum virgatum*), based upon the use of two presence-only species distribution models (SDMs): maximum entropy (Maxent) and support vector machines (SVM). SDMs are commonly used for modeling animal and plant distributions in natural environments, but have rarely been used to develop landscape models for cultivated crops. AUC, Kappa, and correlation measures derived from test data indicate that SVM slightly outperformed Maxent in modeling US corn production, although both models produced significantly accurate results. When compared with results from a mechanistic switchgrass model recently developed by Oak Ridge National Laboratory (ORNL), SVM results showed higher correlation than Maxent results with models fit using county-scale point inputs of switchgrass production derived from expert opinion estimates. However, Maxent results for an alternative switchgrass model developed with point inputs from research trial sites showed higher correlation to the ORNL model than the corresponding results obtained from SVM. Further analysis indicates that both modeling approaches were effective in predicting county-scale increases in corn production from 2006 to 2007, a time period in which US corn production increased by 24%. We conclude that presence-only methods are a powerful first-cut tool for estimating relative land suitability across geographic regions in which candidate biofuel feedstocks can be grown, and may also provide important insight into potential land-use change patterns likely to be associated with increased biofuel demand.

Keywords: biofuels, corn, ethanol, land-use change, species distribution models, switchgrass

Received 20 January 2010 and accepted 7 March 2010

Introduction

Concerns about domestic energy independence, greenhouse gas emissions from fossil fuels, and the health of rural economies have sparked great interest in the development of bioenergy supplies in the United States. Perhaps the most important policy initiative for promoting future US bioenergy production is the Renewable Fuels Standard (RFS) provision of the 2007 Energy Independence and Security Act. The RFS mandates increasing levels of liquid biofuel (e.g., bio-ethanol and biodiesel) use in the US over the next decade and a half, culminating in a biofuel target of 136 billion liters (36 billion gallons) by 2022 (Sissine, 2007). This goal represents five times the amount of biofuels produced by the US in 2007, almost all of which was comprised of corn-based ethanol. The RFS also stipulates that over half (approximately 80 billion liters) of the US biofuel portfolio in 2022 must be composed of 'advanced' biofuels, such as cellulosic ethanol and

Correspondence: Jason M. Evans, tel. (706) 542-2808, fax (706) 542-9301, e-mail: jevans@cviog.uga.edu

[1] Present address: Carl Vinson Institute of Government, University of Georgia, 201 N. Milledge Ave., Athens, GA 30602, USA.

biodiesel, which are not currently in large-scale commercial production (Sissine, 2007).

While the RFS clearly establishes biofuels as a primary pathway for alternative energy development in the US over the next decades, there is great controversy about the consequences of a move toward large-scale biofuel production. On the one hand, several recent studies suggest that some biofuel processes may offer important advantages over petroleum-based liquid fuels, including an increase in domestic energy supply and reduction of greenhouse gas emissions (e.g., Kim & Dale, 2005; Farrell *et al*., 2006; Sartori *et al*., 2006; Schmer *et al*., 2008). On the other hand, a number of other studies argue that large-scale biofuel production – particularly in the form of corn-based ethanol – could exacerbate existing socio-environmental problems such as consumptive use of freshwater (Evans & Cohen, 2009; Gerbens-Leenes *et al*., 2009), aquatic eutrophication (Donner & Kucharik, 2008), wildlife habitat loss (Fletcher *et al*., 2010), air pollution (Jacobson, 2009), and even increased greenhouse gas emissions (Crutzen *et al*., 2008; Fargione *et al*., 2008; Searchinger *et al*., 2008).

Underlying the concerns about biofuels are the overall land area and concomitant land-use changes that will be necessary to grow sufficient biomass feedstocks for meeting biofuel targets. Development of accurate models that delineate relative land suitability across the geographic extent of lands on which candidate feedstocks can be efficiently grown is a clear priority in efforts to better understand the potential benefits and risks from increased biofuel production (USEPA, 2009). Nonetheless, efforts to estimate the extent of areas suitable for production of biofuel feedstocks have been remarkably limited. The primary reason is that the input and production requirements are not well-quantified for many candidate species, meaning that very little data are available for traditional agronomic model construction and validation. Thus, investigators often rely on expert opinion and broad assumptions regarding land-use constraints to delineate the geographic suitability for prospective biofuel feedstocks (e.g., Milbrandt, 2005).

Here, we present results from a novel application of species distribution models (SDMs) to estimate landscape suitability for biofuel feedstocks. SDMs are commonly used for modeling animal and plant population distributions in the natural environment, but have rarely been used for the purpose of modeling cultivated crop species (Miller & Knouft, 2006). We first modeled the geographic distribution for corn (*Zea mays*) production. Corn provided a good test case because it is a crop that is already commercially produced, and thus has readily available data for both model construction (i.e., model training) and independent model testing at a national level. Two commonly used SDM techniques, maximum entropy (Maxent) and support vector machines (SVM), were evaluated in terms of their ability to predict both crop distribution (presence/absence) and relative intensity for corn yields across the US at a county-scale. Two different modeling approaches – one based upon point inputs obtained from expert opinion estimates, and the other based on production at field research sites – were then developed for switchgrass production using both Maxent and SVM. Results of these were then contrasted to an agronomic production model recently developed for switchgrass. Furthermore, an observed increase in corn production between 2006 and 2007, widely thought to be caused by growing demands for corn-based ethanol (see, e.g., Westcott, 2007), provided a unique opportunity to assess SDMs in terms of their ability to predict agricultural land-use change associated with an increase in biofuel production. To conclude, we discuss implications of SDM results in terms of land-use change, differential environmental impact of feedstock production across the landscape, and other potential applications.

SDMs and their applicability to biofuels

Models that use known location records to make predictions about habitat suitability and associated distribution of species are increasingly being used by ecologists and conservation scientists. Also commonly called habitat suitability and/or ecological niche models, the approach for building a SDM is basically threefold: (1) overlay point location records for a given species across GIS-based layers containing socio-environmental variables (e.g., climate, altitude, land-use, human population density, etc.); (2) apply a mathematical model to identify relationships between observed point locations and socio-environmental parameters at or near point locations; and (3) use the model results from step 2) to interpolate (and potentially extrapolate) relative habitat suitability across the landscape in which socio-environmental parameters are known.

Two general classes of SDMs are often distinguished based upon the type of data required. One common model type is known as 'presence-absence.' As the name implies, presence-absence models use discrete point location information for areas that a species is known to be present as well as for areas in which the species is known to be absent. Generalized linear models, generalized additive models, and classification and regression trees are some of the more commonly used presence-absence modeling approaches (Brotons *et al*., 2004). By contrast, 'presence-only' modeling approaches require location information only for known presence points and do not require explicit absence data. Commonly used presence-only models include Genetic Algorithm for Rule-set Prediction (GARP), Maxent, SVM, and climate envelope models (e.g., BIOCLIM).

Presence-only models have two unique attributes that would seem to make them more useful than presence-absence models for estimating landscape suitability for biofuel feedstock production. First, presence-absence models require absence data, which reflect areas unsuitable for a species, for model development. For many candidate biofuel feedstocks not yet widely grown, it would be inappropriate to treat localities without current production as unsuitable. Presence-only models avoid this problem because they do not require the explicit constraints indicated by absence data. Second, most presence-only models are designed to function well even when limited to very sparse presence data sets (Engler *et al*., 2004; Hernandez *et al*., 2006), meaning that useful geographic models of suitability can often be developed with very few presence point locations. This feature is particularly important because many prospective biofuel feedstocks currently are being grown in only a small number of test locations.

Based upon this reasoning, we hypothesized that presence-only SDMs, which to date have not been widely applied to cultivated crops (Miller & Knouft, 2006), could produce useful estimates of landscape suitability for biofuel feedstock production. Corn and switchgrass provide ideal cases for testing of this hypothesis due to the large amounts of extant information (detailed crop data in the case of corn, and a comprehensive agronomic model in the case of switchgrass) that can be used to evaluate suitability estimates derived through presence-only models.

Methods and materials

Presence-only models

We used two SDM approaches to develop suitability maps for U.S. corn and switchgrass production: Maxent and single-classification SVM (Phillips *et al*., 2006; Drake *et al*., 2006). Several additional models – including GARP, BIOCLIM, DOMAIN, and two-class support vector machines – were also tested in initial model runs for corn. However, we chose to not use these additional models when developing detailed corn and switchgrass analyses because they were found to perform significantly less well than Maxent and SVM, a finding that is generally consistent with other recent SDM model evaluations (e.g., Elith *et al*., 2006).

Although both Maxent and SVM require only presence point inputs for model fitting, a further distinction can be made in that single-class SVM derives suitability estimates only from socio-environmental data at presence localities (Drake *et al*., 2006), while Maxent also utilizes socio-environmental data from randomly selected 'background' locations (Phillips *et al*., 2006). Such background locations are commonly referred to as 'pseudo-absences,' and are used because they provide information regarding available socio-environmental gradients relative to point locations. Pseudo-absences can potentially increase model performance, particularly when using sparse point input data (Elith *et al*., 2006). However, some argue that pseudo-absences are theoretically suspect in the sense that they do not represent true absences, and may thus provide unjustified constraints that in some cases could deleteriously bias the model outputs (Drake *et al*., 2006).

Maxent. Maxent is a supervised machine learning method based on statistical mechanics (Jaynes, 1957). It uses the concept of maximum entropy to estimate suitability, which is generally thought to work well under conditions of sparse data (i.e., 'incomplete information'; Phillips *et al*., 2006). Potential suitability or predicted distributions are represented as an unknown probability distribution across all sites (study area). The algorithm constrains the probability distribution based on environmental data at presence locations and chooses a distribution of Maxent, i.e., the most unconstrained. Maxent uses different classes of linear and nonlinear 'features' for quantifying information in the data (i.e., relationships of presence locations with socio-environmental data; Phillips *et al*., 2006). The Maximum Entropy Species Distribution Modeling Version 3.2.1. package was used to run Maxent. We used recommendations in Phillips & Dudik (2008) for parameter tuning.

Single-classification SVMs. SVMs are another kind of supervised machine learning technique. This technique, based on single-classification (or 'one-class' classification), uses only presence locations to map input vectors to a higher dimensional space using kernel functions (Scholkopf *et al*., 2000).The objective is to minimize the multidimensional space that encapsulates presence data. Because SVM is not based on an underlying theoretical distribution, there are no assumptions regarding independence of data points. The solution is thus deterministic, thereby greatly reducing computation time as compared with some other presence-only modeling approaches (Drake *et al*., 2006). Use of single-classification SVM methods in ecological niche-modeling is explained in detail by Guo *et al*. (2005) and Drake *et al*. (2006). We used a radial-basis kernel function, which is generally thought to be superior for estimation and classification accuracy (e.g., Pirooznia & Deng, 2006). SVM was run using the OPENMODELLER DESKTOP 1.0.7 package (de Souza Munoz *et al.*, 2009).

Socio-environmental data layers

A total of 27detailed climate, environmental, and social layers covering the continental United States at a cell resolution of 9 × 9 km were originally compiled and considered for model runs. These layers were analyzed for correlations based on the respective values at 3000 randomly selected points, with a correlation cutoff ($r = 0.8$) used to select layers for use in final model runs.

Three primary climate variables were selected as the basis for correlative comparisons among socio-environmental variables: (1) mean annual temperature (USDA, 2009); (2) mean annual precipitation (USDA, 2009); and (3) mean diurnal temperature range (openModeller, 2008). Eight other variables with correlation results lower than the defined threshold were also selected for model runs, resulting in a total of 11 explanatory data layers for model training: (4) altitude (CGIAR-CSI, 2008); (5) mean annual temperature maximum (USDA, 2009); (6) mean annual temperature minimum (USDA, 2009); (7) mean precipitation warmest month (openModeller, 2008); (8) mean wind speed (openModeller, 2008); (9) county-scale major road density as defined by the United States Department of Transportation (data set obtained through University of California Berkeley, 2008); (10) standard deviation mean precipitation (openModeller, 2008); and (11) standard deviation mean temperature (openModeller, 2008). Rejected layers included 15 additional climate variables available through the openModeller (2008) package and county-scale population density (US Census Bureau, 2009). In addition, the road density layer was removed as an explanatory variable the from point-sparse switchgrass models due to apparent biasing in the point data set toward moderate road (and population) densities found near universities, where most switchgrass trials occurred.

Precipitation and temperature are dominant forcing functions for determining crop suitability, and thus are included as standard variables in models of crop distribution and production (Parry *et al.*, 2004; Fischer *et al.*, 2005). Wind speed is also used in many crop production models due to the close relationship of wind with landscape evapotranspiration rates and, by extension, soil water loss (Geerts *et al.*, 2006). We included altitude as an additional environmental variable due to the relatively independent effects of altitude on crop suitability in some locations (Mani *et al.*, 2007). The rationale for including road density is that the availability of sufficient road infrastructure in rural areas is likely to have important implications for biofuel feedstock production (Walsh, 1998).

We also note that two variables with clear importance for agricultural production were not included in the models: (1) soil quality and (2) availability of irrigation water. Omission of these variables was due to the current lack of readily available GIS data layers containing such information at the scale of the continental US. Although an ideal crop distribution model would include these two (and potentially other) variables, we proceeded under the hypothesis that available climate, altitude, and infrastructure layers would provide sufficient information for developing accurate estimates of relative crop suitability at a national scale.

Presence points

The data for development of the corn model were obtained from the 2006 National Agricultural Statistics Service (NASS) data for county-scale field corn production (data available at http://www.nass.usda.gov/QuickStats/Create_County_All.jsp; county-scale map of 2006 corn production data shown in Fig. 1). The pool of potential training points for the corn model was limited to a random selection containing two-third of all counties in the continental US ($n = 2026$). The remaining counties ($n = 1013$) were held out for independent validity testing with corn data.

The first switchgrass modeling approach (the 'point-intensive' switchgrass model hereafter) was similar to the corn model in that it used county-scale input points, but with the important difference that points were based upon expert opinion of potential production – not actual production as was the case for corn. The point-intensive switchgrass model was developed using expert assessments, as compiled by the National Renewable Energy Laboratory, of potential switchgrass biomass production on Conservation Reserve Program (CRP) lands for each U.S. county (Milbrandt, 2005). For consistency in model-training, the point-intensive switchgrass models were fit using the same pool of counties randomly selected for the corn model.

A second switchgrass model (the 'point-sparse' switchgrass model hereafter) was constructed from a relatively small number of presence points obtained through a literature search of switchgrass production field tests (supporting information Table S1). All available points were used for training the point-sparse model due to the limited number of switchgrass study sites ($n = 48$). Because such a point-sparse approach is likely to be the only feasible option for other prospective biofuel crops not currently produced at large scales, switchgrass presents an interesting opportunity to compare results from a point-sparse model with those from

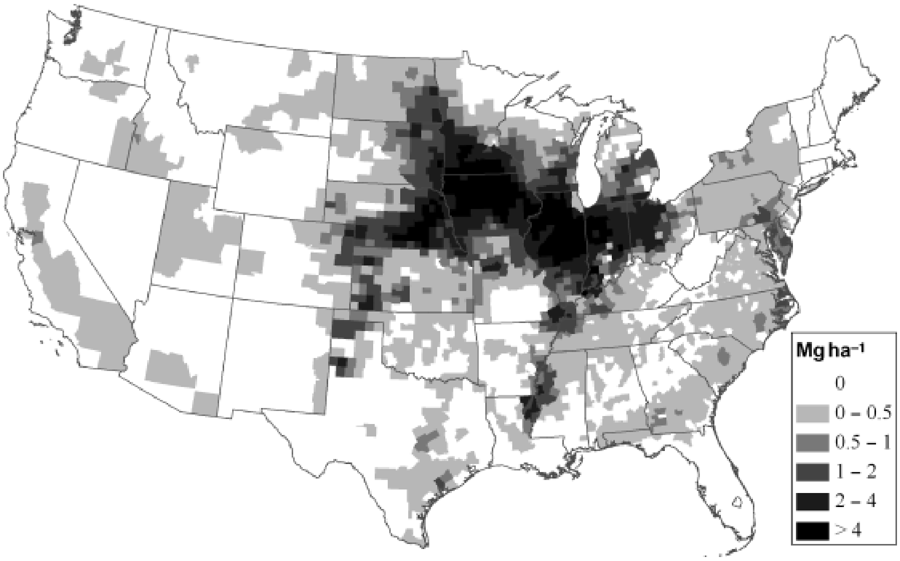

Fig. 1 Corn production intensity (as Mg ha^{-1} on the scale of total county area) across the United States in 2006. Source data obtained from the National Agricultural Statistics Service (http://www.nass.usda.gov/QuickStats/Create_County_All.jsp).

a much more point-intensive model based on expert opinion. Another possible mechanism for selecting switchgrass point locations would be to assemble species location data from herbaria and other natural history surveys (Graham *et al.*, 2004). However, we did not use this latter approach because such data would be more appropriate for modeling the naturalized range for switchgrass, and would not provide relevant information about the biomass production rates from switchgrass grown in monotypic stands.

Point location errors in corn and point-intensive switchgrass models. It must be noted that the point data sources for the corn and point-intensive switchgrass models are aggregated at a county scale. Because national-scale land cover data are not available for determining precise coordinates of crop locations at the county level, the geographic centroid of each county was used as the basis for presence points. This use of centroids, rather than specifically verified presence points, introduces some inherent error into the modeling inputs. Moreover, the possibility of significant error is greater for counties with relatively small amounts of crop production area, as there is less chance that the centroid represents a 'true' presence point than in counties with relatively large amounts of production. For example, there is a 1% probability that the centroid represents a true presence point for a county with 1% of its land area in corn production, whereas there is a 50% probability of the centroid serving as a true presence point for a county with 50% corn production by area.

However, the realized importance of this error in our models is greatly reduced by the fact that explanatory variables were aggregated to a 9 × 9 km cell size. In practice, the precision of presence points is only meaningful within this spatial scale, as the centroid effectively serves as an accurate proxy for a true presence point so long as a crop location also falls within the 9 × 9 km cell in which the centroid is contained. This spatial tolerance is particularly important for minimizing the realized importance of errors in smaller counties, which generally should have less heterogeneity in explanatory variables than larger counties. The realized effect of spatial errors is also reduced in the modeling process through the use of a production weighting procedure – described in more detail in 'Model weighting' – that gives less weight to counties with low areal crop production intensity (kg ha^{-1}) as averaged across the county's total land area.

Model weighting

Because of the large differences in feedstock production intensity at county scales, we used a method for weighting multiple model runs with subsets of the presence points that reflect different production intensity levels. For corn, all counties that showed production in the 2006 NASS data set were ranked according to their areal production intensity (kg ha^{-1}) across the county. For the point-intensive switchgrass model, all counties that showed some level of production were ranked according to an expert estimate of potential biomass yield (kg ha^{-1}) on conservation reserve program (CRP) lands (Milbrandt, 2005). For the point-sparse switchgrass model, each data point was ranked according to the average biomass output reported at the study site (supporting information Table S1).

For the corn and point-intensive switchgrass models, we selected out five progressively less restrictive sets of presence point classes based upon five equally spaced quantiles of areal production intensity: upper 20%, 40%, 60%, 80%, and all counties in which the feedstock was present (Fig. 2a). For the point-sparse switchgrass model, only two quantiles were used due to the small number of initial presence points available for the model ($n = 48$): upper 50%, and all points (Fig. 2b). Model runs were then made with points from each of these production intensity subsets.

We used results from the individual runs for each quantile to develop a final, production-weighted model of suitability for each feedstock. To do so, we weighted the model results for each quantile, based on the average production for that quantile, as $w_q = p_q / \sum p$, where w_q is the weight for suitability for quantile q, and p_q is the average production yields for data from quantile q. After making these weighted adjustments, we then

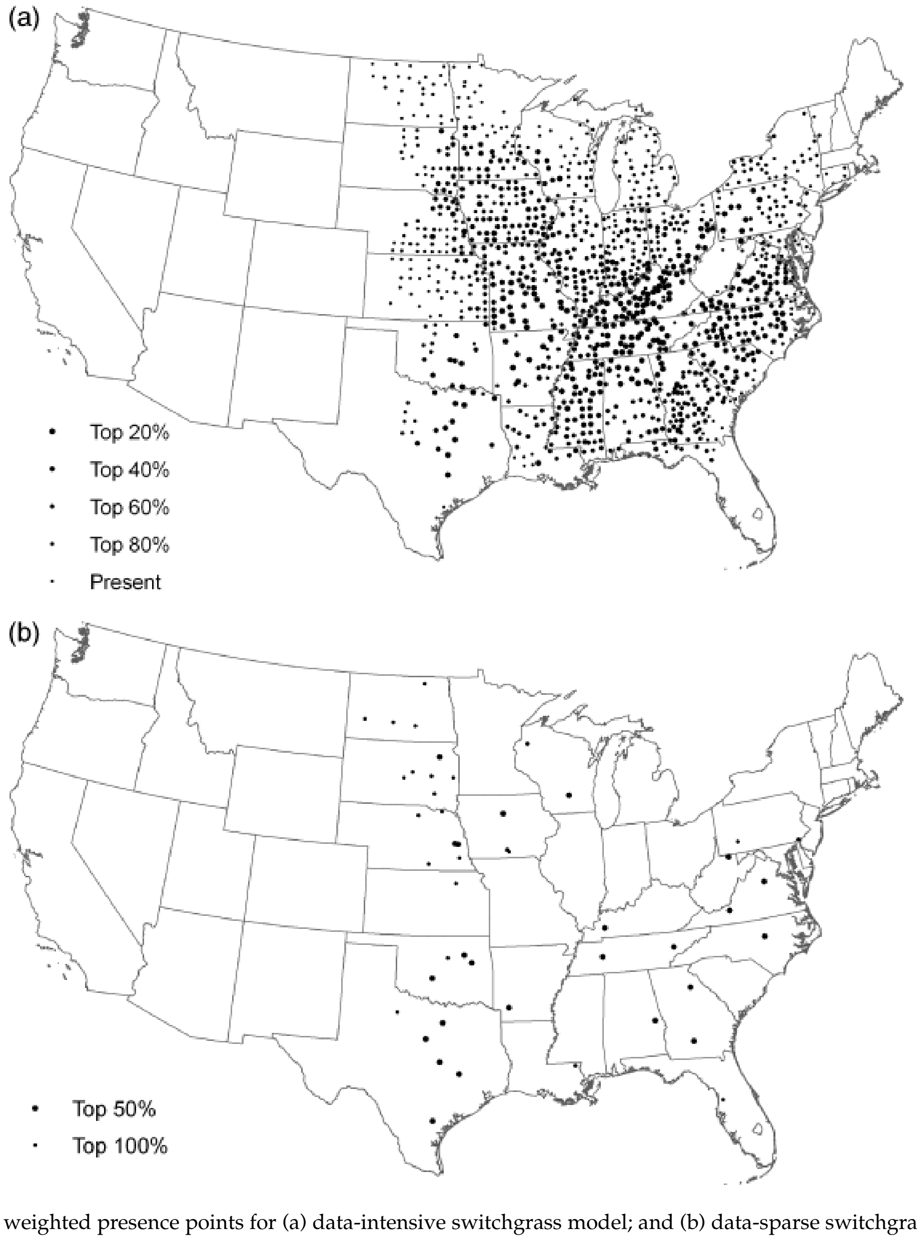

Fig. 2 Production weighted presence points for (a) data-intensive switchgrass model; and (b) data-sparse switchgrass model. Presence points for data-intensive model derived from expert opinion estimates in Milbrandt (2005). Presence points for data-sparse switchgrass model derived from study data listed in Table 1.

calculated a weighted average of individual quantile results to produce a production-weighted suitability (PWS hereafter) for each feedstock model. As such, higher weight is given to high production counties and less weight is given to low production counties in the aggregated PWS results. Because crop data are reported at a county scale, we then calculated a mean PWS value from model predictions within the boundary of each US county polygon.

As discussed above in 'Point location errors in corn and point-intensive switchgrass models', relatively low production intensity correspondingly increases the likelihood that the centroid is not a representative presence point. Although the primary rationale for using the weighting procedure is to develop a model that can reflect the different production intensities expected across the landscape, the production weighting procedure also has the effect of reducing the importance of location errors for corn models by giving less weight to low production counties in which the centroid would be less likely to represent a true presence point.

Assessing model performance

We contrasted mean PWS values to corn yields reported by NASS for 2006 and 2007 in test counties ($n = 1026$) that were not used in model development. To assess the predictive performance of corn suitability models, we used three measures: (1) area under the receiver operating curve (AUC) for predicting the likelihood of production within counties; (2) Kappa; and (3) rank correlation to determine if mean PWS correlated with county-level production yields.

AUC is a threshold independent measure of model performance (see Fielding & Bell, 1997). An AUC value of 0.5 indicates a model that predicts no better than chance, with higher values up to a maximum value of 1 indicating progressively better model performance. AUC is frequently used for interpreting whether models correctly predict actual distributions (Elith *et al*., 2006; Peterson *et al*., 2007), although it has received some recent criticism regarding its utility as a standalone evaluative metric (Austin, 2007; Lobo *et al*., 2008). Consequently, we also used another common performance metric, Kappa (Fielding & Bell, 1997). Kappa measures the proportion of correctly predicted points after the probability of chance agreement has been removed. For the Kappa statistic, we used a threshold cutoff value that maximized Kappa (Freeman & Moisen, 2008). Rank correlation was provided as an additional test for measuring the extent to which suitability results correlated with corn yields for test counties.

The dramatic increase in corn yields (~24%) observed in 2007 over 2006 (USDA, 2008) provided a unique opportunity to test if SDM results could be useful for predicting patterns of land-use change. To examine such potential land-use relationships, we first compared averaged PWS scores from both models with test county corn production data as divided into three classes: (1) greater corn production in 2007 relative to 2006; (2) less corn production in 2007 relative to 2006; and (3) no change in corn production between the 2 years (which for all relevant test counties meant that no corn was produced in either years). Secondly, we calculated the percentage of available land converted into corn for those test counties that did show an increase in corn production in 2007, and then compared this percentage with the respective mean PWS suitability results provided by both Maxent and SVM.

There currently are not large-scale farms growing switchgrass as a biofuel feedstock or corresponding data sets that could provide information about the landscape distribution of actual switchgrass production. Thus, it is not possible to assess the predictive accuracy of switchgrass models using true production data. Instead, we used a county-scale switchgrass agronomic model recently developed by the Oak Ridge National Laboratory (ORNL) (Fig. 5) as a basis for interpreting results from Maxent and SVM. The ORNL model is built upon a mechanistic agronomic approach that includes information about local climate, crop management, plant cultivar selection, and field trial production to provide county-scale estimates of optimal switchgrass yield per hectare (Gunderson *et al*., 2008). The data used in the ORNL model were derived from many of the same field trial studies used in the construction of our point-sparse model.

Given that the ORNL model does not necessarily represent true presence or absence in the landscape, use of AUC or Kappa as a method for assessing performance of SDMs would be inappropriate. However, correlative comparisons of SDM results with the ORNL model do provide a useful benchmark for determining whether the SDMs, which can be constructed using considerably less detailed input parameters than mechanistic agronomic models, might serve as an effective first-cut tool for estimating landscape suitability for other emergent biofuel feedstocks.

Results

Corn model comparisons

Qualitative comparison of Maxent and SVM results (Fig. 3) with a map of 2006 county-scale corn production (Fig. 2) suggests that both models capture the outline of the Midwestern corn belt in some detail. Secondary production areas along the Mississippi River lowlands,

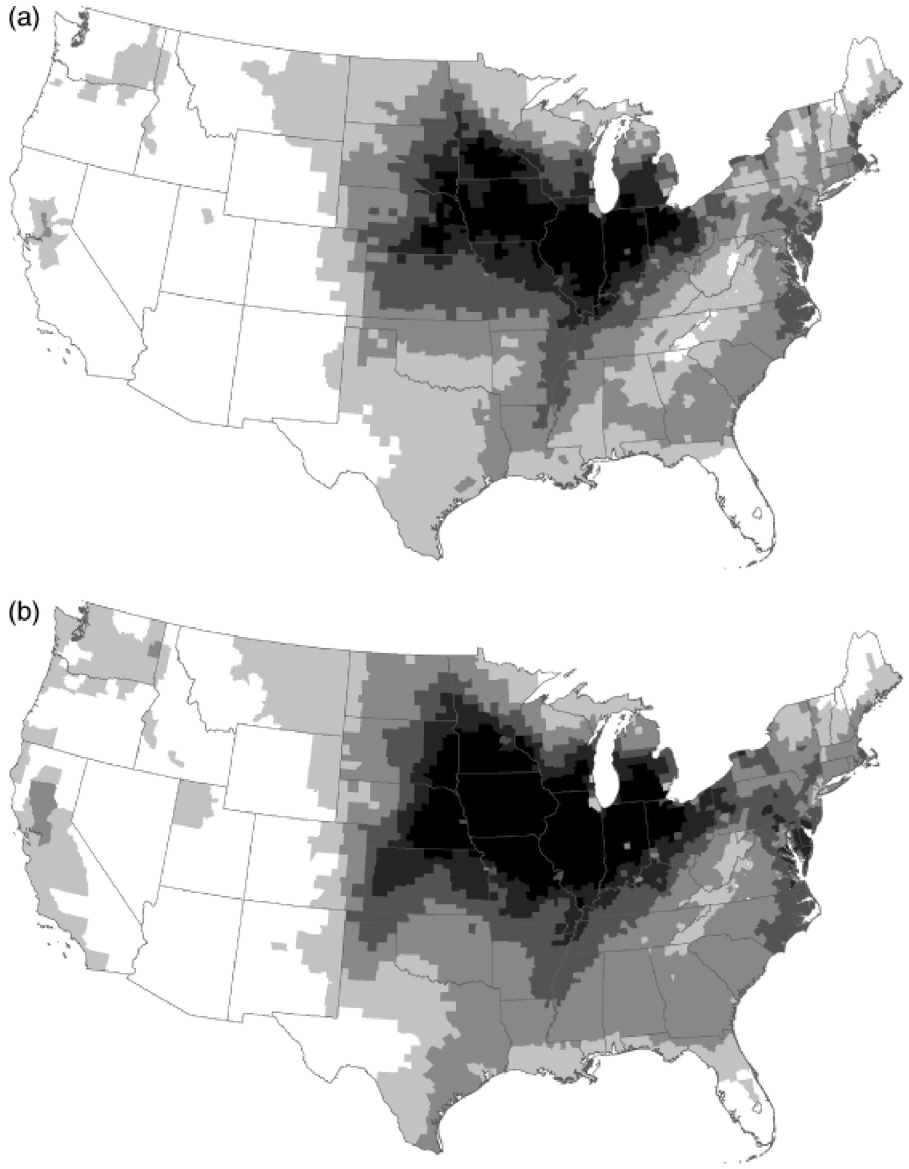

Fig. 3 Corn suitability estimates for (a) Maxent and (b) SVM. Results displayed using equivalent scale, with darker areas indicating higher suitability scores.

the southeastern coastal plain, and the California central valley are also evident in both maps. All test county validation statistics for corn models (Table 1) were slightly, but consistently, higher for SVM than for Maxent.

One important advantage provided by Maxent, however, is that variable contribution to model fit is provided with model output, whereas this information is not readily provided by SVM. Standard deviation mean temperature showed considerable importance for corn, as it contributed 32% of the information in the fitted model output (Table 2). Relatively large contributions (>5%) were also made by altitude, mean precipitation, mean annual temperature maximum, mean temperature, standard deviation mean precipitation, and mean diurnal temperature. Although other variables showed only minor overall contribution, it is important to note that high road density (1.7% of overall model fit) was clearly the primary factor for significantly reducing predicted suitability in major population centers (e.g., Chicago, IL, USA).

Switchgrass models

Maps for the point-intensive Maxent and SVM switchgrass models (Fig. 4) both show high suitability for much of the non-Appalachian southeast and the Midwest, marginal suitability for the northeast, and low suitability in the mountain west. Results in the Pacific west do diverge somewhat, as SVM shows marginal suitability throughout wide areas of this region, while Maxent shows very low suitability for all but a few isolated locations.

More dramatic visual differences are found in the point-sparse outputs for Maxent and SVM. The point-sparse Maxent model (Fig. 4b) is more optimistic than both point-intensive models (Fig. 4a and c) in terms of predicting high switchgrass suitability for much of eastern Texas, Oklahoma, and Kansas. This result likely can be explained by the point-sparse model being biased by the relatively high number of switchgrass study sites in the western plains (Fig. 1b) – a geographic distribution that reflects overall research interest in developing switchgrass as a biofuel feedstock in marginal crop land and more arid grassland areas (e.g., Varvel *et al*., 2008). However, the SVM point-sparse model (Fig. 4d) restricted high switchgrass suitability to extreme southeast Texas, and showed generally marginal suitability throughout much of the continental USA.

Table 1 Model comparison with test counties, US corn production

	AUC		Kappa		Rank correlation
Model	2006	2007	2006	2007	2006
SVM	0.873	0.860	0.564	0.534	0.787
Maxent	0.867	0.857	0.527	0.512	0.781

AUC, maxKappa calculated using the R package 'PresenceAbsence'.

Variable contribution analyses provided by Maxent indicated that altitude had the largest effect (32.9%) on the point-intensive model, while mean precipitation was the single most influential variable (36.9%) for the point-sparse switchgrass model (Table 2). Although the relative contributions vary, the top five explanatory variables for both switchgrass models were altitude, mean precipitation, mean temperature, mean annual temperature maximum, and mean precipitation warmest month. However, the importance of minor variables is demonstrated by road density in the point-intensive switchgrass model (2.3% contribution), which, similar to the corn model, had a clear influence in suppressing relative suitability results for major metropolitan areas (e.g., Chicago, IL, USA and Atlanta, GA, USA). This effect becomes particularly clear when examining point-sparse models that do not include road density as an explanatory variable (Fig. 4b and d), as reduced suitabilities are not observed near major metropolitan areas in these results for either Maxent or SVM.

Scatter plot comparisons (Fig. 5) show that the models rarely predicted high suitability in areas predicted to be low suitability from ORNL (i.e., the absence of points in the lower right portion of the graphs). However, the converse was not true (i.e., when ORNL predicted high suitabilities, a wide range of suitabilities were given by the models). For the point-intensive models (Fig. 5a and c), this discrepancy is largely explained by the ORNL projecting very high switchgrass yields in the Pacific northwest (Gunderson *et al*., 2008), a region that the Maxent point-intensive model shows as having very low (Fig. 4a) and the SVM model shows as having marginal suitability (Fig. 4c). A similar discrepancy for the Pacific Northwest holds for the comparison with the point-sparse Maxent model (Figs 4b and 5b). Further divergence is notable in the central plains areas of

Table 2 Variable contribution (%) to production weighted Maxent models

Variable	Corn	Point-intense switchgrass	Point-sparse switchgrass
Altitude	17.0	32.9	26.4
Mean diurnal temperature	5.2	3.1	6.0
Mean precipitation	15.5	20.6	36.9
Mean precipitation warmest month	1.2	9.2	4.1
Mean temperature	8.3	9.1	14.6
Mean temperature maximum	12.2	10.9	3.9
Mean temperature minimum	0.3	1.2	0.6
Mean wind speed	1.3	0.5	0.8
Road density	1.7	2.3	na
Standard deviation mean precipitation	5.3	3.3	3.6
Standard deviation mean temperature	32.2	6.9	3.1

Variable contributions from individual Maxent model runs were weighted according to the production weighting formula method described in 'Model weighting'.

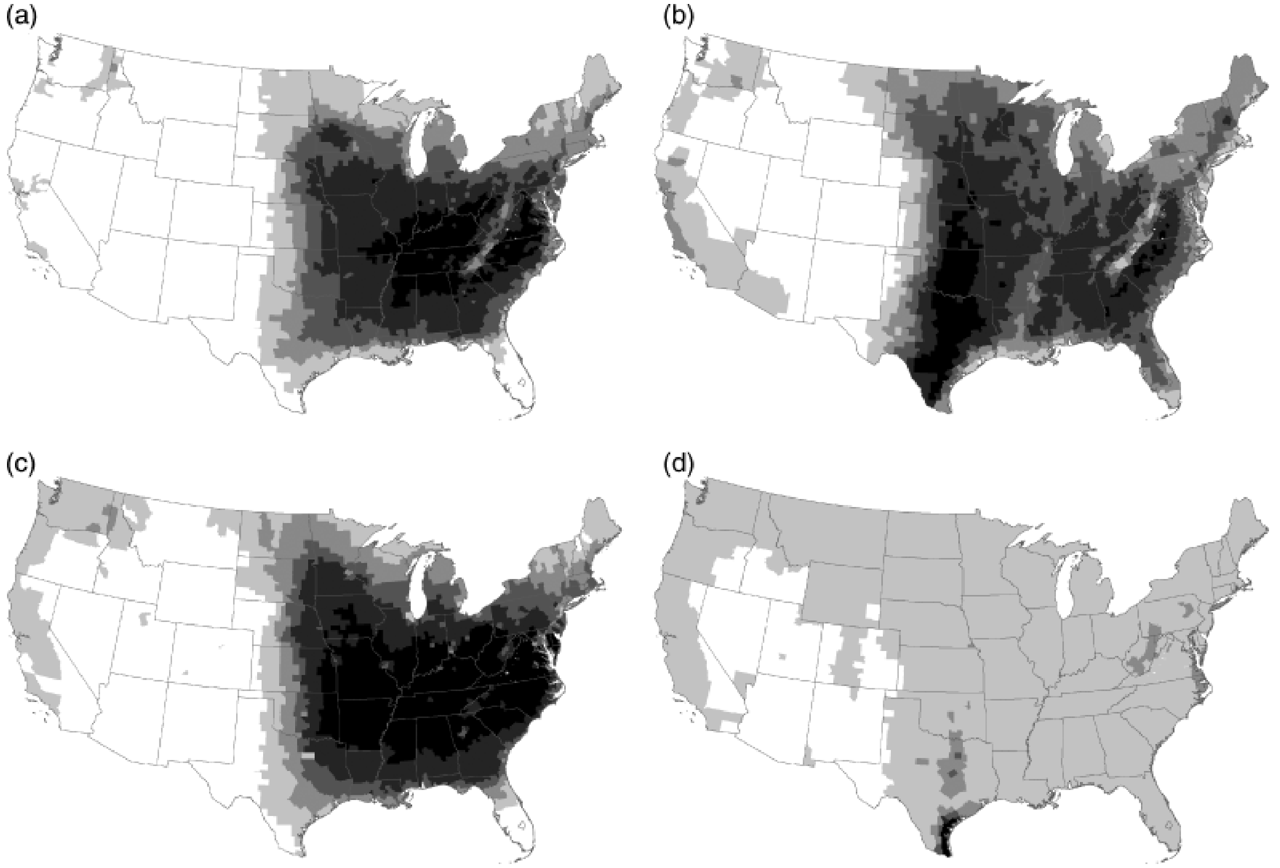

Fig. 4 Switchgrass suitability estimates for (a) Maxent point-intensive model; (b) Maxent point-sparse model; (c) SVM point-intensive model; and (d) SVM point-sparse model. Darker areas indicate higher suitability scores.

Texas, Oklahoma, and Kansas that the Maxent point-sparse model shows as having high suitability (Fig. 4b), but are shown as having low to moderate biomass potential by the ORNL model (Gunderson *et al*., 2008). As shown in the inset to Fig. 5d, the scatter plot for the point-sparse SVM switchgrass model for suitability values <0.33 is visually similar to the point-sparse Maxent switchgrass model (Fig. 5b) for all suitability values. However, the scatter plot shows poor correlation with the ORNL model for those counties, mostly located in eastern Texas (Fig. 4d), with SVM suitability results over 0.33.

Similar to the corn results, county-scale correlations with ORNL's switchgrass production model (Gunderson *et al*., 2008) were somewhat higher for the SVM point-intensive model than for the Maxent point-intensive model. By contrast, a similar comparison of the point-sparse models indicated a higher correlation for Maxent than SVM. Assuming that the ORNL model provides a good basis for comparison, these results are consistent with the corn results suggesting that SVM performs somewhat better than Maxent when fitting models for cultivated crops based on relatively large amounts of data inputs. At the same time, the point-sparse results suggest that important modeling gains may indeed be obtained using Maxent within the context of sparse data sets.

Landscape distribution of corn production increases

Several lines of evidence suggest that SDMs provided useful information about the distribution of corn production increases, and thus potentially land-use change, across the United States. First, validation with 2007 corn test data had similar AUC, Kappa, and rank correlations as the 2006 data (Table 1) used for model training. Second, the average predicted suitability estimated from both models was much greater for counties with increased production than for counties with lower production or no change in production from 2006 to 2007 (One-way ANOVA: $F_{2,1023} = 239.40$, $P<0.0001$; Fig. 6a). It is important to reiterate that counties with no change in production had zero hectares in corn for both years, consistent with the SDMs predicting the lowest

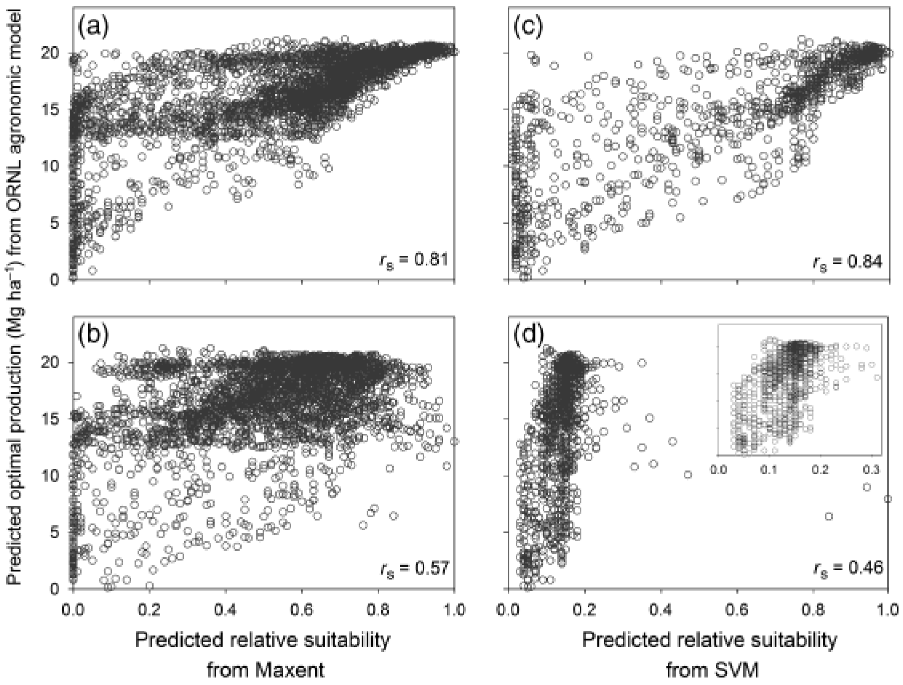

Fig. 5 Correlations between ORNL agronomic model and species distribution models that use contrasting data sets for model development. (a) Maxent intensive, (b) Maxent sparse, (c) SVM intensive, and (d) SVM sparse. Inset in (d) provides a finer resolution for predictions from SVM model <0.33. For all model comparisons, $P < 0.0001$.

average suitabilities for these counties. Third, the expansion in corn production for test counties in 2007 was significantly correlated with suitability results from both Maxent and SDMs (Fig. 6b and c).

Discussion

Implications for using SDMs on cultivated crop species

Our results indicate that presence-only models, which typically have been used for estimating habitat suitability for plant and animal species in natural areas, can also provide reasonable estimates of relative landscape suitability for cultivated crops. Moreover, we found that apparently accurate presence-only models could be fitted based upon three distinct types of input data: (1) weighted production records for a widely grown crop (corn); (2) weighted production estimates (based upon expert opinion) that are detailed across the United States, even though the crop is not yet widely grown as a biofuel feedstock (point-intensive switchgrass); and (3) weighted production data from relatively few study sites dispersed across the landscape of interest (point-sparse switchgrass). Such data-input flexibility may be a particularly attractive feature for managers looking to develop first-cut landscape suitability assessments of biofuel feedstocks, particularly because SDMs generally are already known for their low cost, ease of use, and transportability across spatial scales.

It should also be noted that models for other feedstocks do not necessarily have to be derived from presence points in the landscape of interest (e.g., the continental United States), but potentially could also be fitted to presence points from remote locations in which they are currently cultivated or found to thrive in natural areas. The ready availability of detailed global climate and environmental data correspondingly allows for development of models that extrapolate suitability into areas far beyond those from which presence points were drawn. Estimating changes in species distribution over time due to global climate change and predicting the large-scale spread of non-native invasive species represent similar types of applications in which presence-only models are being increasingly employed (Peterson, 2003; Araujo *et al*., 2005; Drake *et al*., 2006). Aside from the utility of such global models for assessing landscape suitability for candidate crops, results could also be used for development of risk profiles for non-native biofuel feedstocks that have been identified as potential invasive species (see Raghu *et al*., 2006; Barney & DiTomaso, 2008).

Although it is notable that models have strong predictive performance even when restricted to climate, altitude, and road infrastructure as explanatory

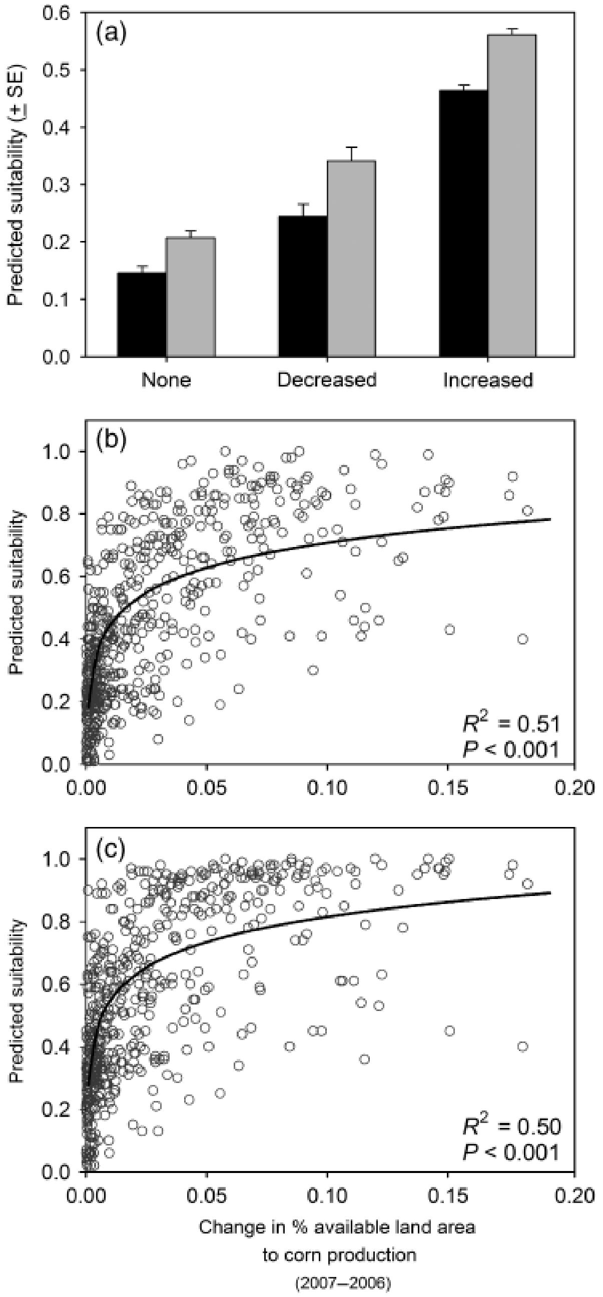

Fig. 6 Both models [(a) black-Maxent, gray-SVM] predicted greater suitability for corn production for test counties that increased production from 2006 to 2007 than for counties that decreased production. Predictions were lowest for counties that did not change in production, which were counties with no production in either year. Furthermore, both models [(b) Maxent, (c) SVM] were highly correlated with the amount of corn expansion in test counties. The predicted lines comes from regression models fitting predicted suitability as a function of the log of percent change of land area potentially available (i.e., not in production in 2006). A log-transformed model fit the data better than a linear model (Maxent: $\mathrm{AIC}_{\mathrm{linear}} = -174.1$, $\mathrm{AIC}_{\mathrm{log}} = -273.2$; SVM: $\mathrm{AIC}_{\mathrm{linear}} = -88.5$, $\mathrm{AIC}_{\mathrm{log}} = -210.7$).

variables, one potential limitation in our approach, as noted in 'Presence points', was omission of variables such as soil quality and irrigation availability that are not currently available at a national level. We suspect that model performance was strong in the absence of this information because other explanatory variables were likely correlated with soil quality and irrigation availability (e.g., mean precipitation). Inclusion of detailed explanatory data layers – such as irrigation, soils, slope, and potentially others – at more local spatial scales in which they are available could be expected to result in models that have greater spatial accuracy and precision than those presented here. For example, very high model accuracy (test AUCs > 0.945) was shown for a Maxent model constructed to predict occurrence of several invasive plant species in riparian areas along Nebraska's North Platte River using local environmental layers assembled at a 30 m cell resolution (Hoffman *et al.*, 2008). Similarly precise models of biofuel feedstocks and other novel crops may be achieved by obtaining presence locations for successful field trials, and then fitting models to an appropriate set of fine-scale explanatory variables.

However, we do caution that such presence-only models should not be viewed as a ready substitute for traditional agronomic production models. Instead, we suggest that presence-only models might be seen as a 'filter' for identifying the most promising (or, in some cases, least promising) new crop candidates for a given region. Results presence-only models might then, for example, facilitate allocation of advanced research toward particular crop species that show high relative suitability across large areas, and away from crops predicted to have much more limited distributions.

Model choice

The slightly better performance of SVM relative to Maxent for corn production was somewhat surprising. Although we are not aware of direct comparisons of Maxent to SVM, several recent comparisons showed Maxent to be generally more accurate than other widely used presence-only algorithms (Elith *et al.*, 2006; Hernandez *et al.*, 2006; Phillips *et al.*, 2006). Overall, our results suggest that, at least for applications with large amounts of presence data, SVM does provides a reasonable alternative to Maxent that has the added advantage of avoiding assumptions associated with use of pseudo-absences (Drake *et al.*, 2006). However, the Maxent program does have the important advantage of readily providing detailed information about the relative contribution of input variables – information that is not currently provided with software that runs SVM presence-only models. Moreover, Maxent appears to have performed considerably better than SVM for point-sparse switchgrass models. This result is consistent with other studies suggesting that the Maxent algorithm has particularly high utility for applications in which presence data are limited (e.g., Hernandez *et al*., 2006; Phillips *et al*., 2006). Given that both models require similar data inputs and produce results within relatively short amounts of running time (all model runs were completed in <5 min), we suggest that future applications may be best served by using both Maxent and SVM. In general, it is likely that comparison of similarities and differences between results obtained from these two (and potentially other) modeling approaches will provide more interpretive confidence than the standalone results of one model (see, e.g., Araujo & New, 2007).

Implications for assessing environmental impacts of biofuels

There is great interest in developing land-use change models for anticipated increases biofuel production, both to understand the upper bounds of fuel supply potential and the extent of negative environmental consequences that may be associated with large-scale use of biofuels (Donner & Kucharik, 2008). While the modeling results presented here do have some interesting implications for addressing such land-use questions, it is important to stress that there are limitations as to how the model output should be interpreted. A fundamental point worth stressing is that the model results provide a prediction of relative suitability at the scale of the explanatory variables (9 × 9 km), rather than a prediction of absolute suitability. This limitation is apparent in Fig. 7, which shows the land area contained within a range of suitability thresholds (>0.2 to >0.9 at a 0.1 interval) for each of the corn and switchgrass model results (Fig. 7). As a basis for comparison, up to 43 million corn hectares has been forecast for maximum biofuel production scenarios in 2015 (e.g., Donner & Kucharik, 2008), while the US Department of Energy (DOE) estimates that 15–20 million hectares of land could be converted into cellulosic feedstock production by 2020 (Perlack *et al*., 2005). With the exception of the point-sparse SVM switchgrass model, such results may appear to suggest that these land area requirements could be met only by including lands with relative suitability scores that are greater than 0.8 or even 0.9. However, such a straightforward interpretation of the model outputs would be based on at least two unreasonable assumptions: (1) all lands contained within a 9 × 9 km cell would indeed be suitable and available for biofuel crop production and (2) all suitable lands would be dedicated to one crop (e.g., in the case of corn, excluding coverage for rotations with soybean, wheat, or fallow). Thus, these aggregated threshold area calculations should be more cautiously interpreted as providing insight into how the models are distributing relative crop suitability scores across the landscape, rather than as a spatially precise measure of lands that would truly be available for biofuel production.

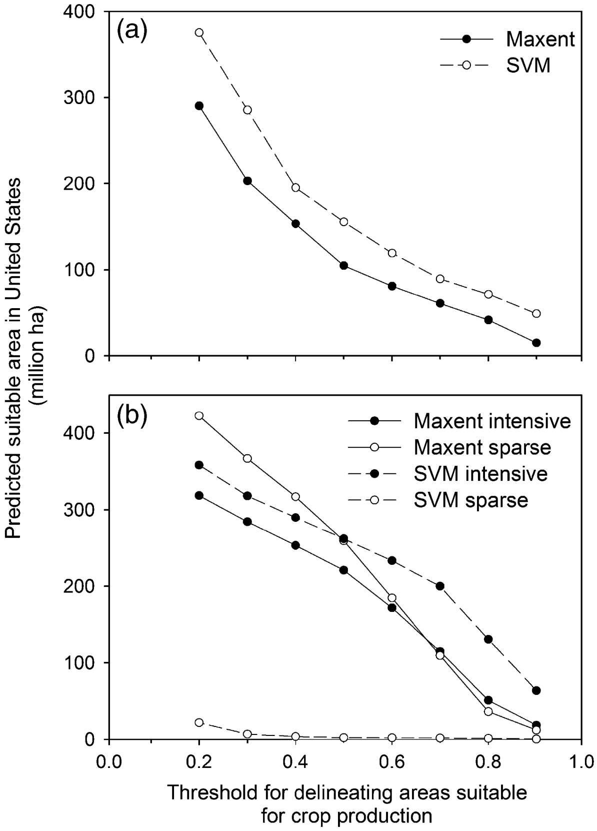

Fig. 7 The predicted area (in millions of hectares) suitable for (a) corn and (b) switchgrass production, as a function of changes in assumptions regarding the threshold for predictions of relative suitability from models.

While this limitation is important to note, we reiterate that both Maxent and SVM were successful in predicting relative patterns of land-use change associated with the major increase in corn production observed in 2007 (Fig. 6). This finding provides some confidence for potentially using presence-only suitability results as an aggregated input (i.e., substituted for individual climate, environmental, and social data layers) in models for forecasting more detailed land-use changes from biofuel production. Aggregation of major explanatory variables into one suitability metric may allow for inclusion of more detailed data for other critically important aspects of land-use change models, such as vagaries in crop/biofuel market forces and understanding agent (farmer) decision trees (see, e.g., Scheffran & BenDor, 2009). At the very least, additional exploratory research is warranted to determine whether SDM results may indeed provide advantages over current environmental input approaches used in land-use change models.

More speculatively, our results may suggest that areas with low predicted suitabilities could generally require higher agronomic inputs, and thus have potentially higher environmental consequences, than those with high suitabilities. For example, farm-scale corn yields in the Southeast US – an area that both Maxent and SVM showed as having low to marginal suitability – can often exceed those of the Midwest corn belt, but only when cultivated under relatively intense fertilizer, irrigation, and pesticide regimes (Evans & Cohen, 2009). Similarly, low switchgrass suitabilities were noted by both Maxent and SVM for the Pacific Northwest, an area in which large-scale production of this feedstock is likely predicated on careful irrigation and field water management (Gunderson *et al.*, 2008). However, more detailed analysis is necessary to determine if such qualitative comparisons hold at a more aggregated level.

Conclusions

We examined the performance of presence-only SDMs, which typically have been used to model plant and animal distributions in the natural environment, as a tool for modeling relative land suitability for biofuel feedstocks. The results indicated good performance by both Maxent and SVM in predicting landscape distribution and observed production increases for corn. Results from both SDMs for switchgrass also showed significant correlation with an agronomic switcghrass production model recently developed by ORNL (Gunderson *et al.*, 2008). We suggest that these results represent an initial, but promising, step into the use of SDMs for modeling biofuel feedstocks and other cultivated crops. Assessment of invasive species risks, potential patterns of land-use change, and landscape distribution of environmental resource demands – all of which are research and management priorities for biofuel crops – are further suggested applications for SDMs moving forward.

Acknowledgements

We give many thanks to the Wildlife Habitat Policy Research Program (WHPRP) of the National Council for Science and the Environment (NCSE) for grant-funds that supported this project. We also thank L.M. Baskaran and others at Oak Ridge National Laboratory (ORNL) for providing us with data files and detailed information about their switchgrass modeling programs. A. Pendleton, J. Kjer, N. Gentry, and M. Wietlisbach were invaluable work assistants on this project. Thoughtful comments by M. Acevedo and anonymous peer reviewers greatly improved the final manuscript.

References

Araujo MB, New M (2007) Ensemble forecasting of species distributions. *Trends in Ecology and Evolution*, **22**, 42–47.

Araujo MB, Pearson RG, Thuiller W, Erhard M (2005) Validation of species-climate impact models under climate change. *Global Change Biology*, **11**, 1504–1513.

Austin M (2007) Species distribution models and ecological theory: a critical assessment and some possible new approaches. *Ecological Modelling*, **200**, 1–19.

Barney JN, DiTomaso JM (2008) Nonnative species and bioenergy: are we cultivating the next invader? *Bioscience*, **58**, 64–70.

Brotons L, Thuiller W, Araujo MB, Hirzel AH (2004) Presence-absence versus presence-only modeling results for predicting bird habitat suitability. *Ecography*, **27**, 437–448.

CGIAR-CSI (2008) SRTM 90m Digital Elevation Data. Available at: http://srtm.csi.cgiar.org/ (accessed January 2009).

Crutzen PJ, Mosler AR, Smith KA, Winiwarter W (2008) N2O release from agro-biofuel production negates global warming reduction by replacing fossil fuels. *Atmospheric Chemistry and Physics*, **8**, 389–395.

de Souza Munoz ME, De Giovanni R, de Siqueira MF *et al.* (2009) openModeller: A generic approach to species' potential distribution modeling. *GeoInformatica*, doi: 10.1007/s10707-009-0090-7.

Donner SD, Kucharik CJ (2008) Corn-based ethanol production compromises goal of reducing nitrogen export by the Mississippi River. *Proceedings of the National Academy of Sciences USA*, **105**, 4513–4518.

Drake JM, Randin C, Guisan A (2006) Modelling ecological niches with support vector machines. *Journal of Applied Ecology*, **43**, 424–432.

Elith J, Graham CH, Anderson RP *et al.* (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29**, 129–151.

Engler R, Guisan A, Rechsteiner L (2004) An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. *Journal of Applied Ecology*, **41**, 263–274.

Evans JM, Cohen MJ (2009) Regional water resource implications of bioethanol production in the southeastern United States. *Global Change Biology*, **15**, 2261–2273.

Fargione J, Hill J, Tilman D, Polasky S, Hawthorne P (2008) Land clearing and the biofuel carbon debt. *Science*, **319**, 1235–1238.

Farrell AE, Plevin RJ, Turner BT, Jones AD, O'Hare M, Kammen DM (2006) Ethanol can contribute to energy and environmental goals. *Science*, **311**, 506–508.

Fielding AH, Bell JF (1997) A review of methods for the assessment of prediction errors in convervation presence/absence models. *Environmental Conservation*, **24**, 38–49.

Fischer G, Shar M, Tubiello FN, van Velhuizen H (2005) Socio-economic and climate change impacts on agriculture: an integrated assessment, 1990–2080. *Philosophical Transaction of the Royal Society Biological Sciences*, **360**, 2067–2083.

Fletcher RJ, Robertson BA, Evans J, Doran PJ, Alavalapati JRR, Schemske DW (2010) Biodiversity in the era of biofuels: risks and opportunities. *Frontiers in Ecology and the Environment*, doi: 10.1890/090091.

Freeman EA, Moisen GG (2008) A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. *Ecological Modelling*, **217**, 48–58.

Geerts S, Raes D, Garcia M, Del Castillo C, Buytaert W (2006) Agro-climatic suitability mapping for crop production in the Bolivian altiplano: a case study for quinoa. *Agricultural and Forest Meteorology*, **139**, 399–412.

Gerbens-Leenes W, Hoekstra AY, van der Meer TH (2009) The water footprint of bioenergy. *Proceedings of the National Academy of Sciences USA*, **106**, 10219–10223.

Graham CH, Ferrier S, Huettman F, Moritz C, Peterson AT (2004) New developments in museum-based informatics and application in biodiversity analysis. *Trends in Ecology and Evolution*, **19**, 497–503.

Gunderson CA, Davis EB, Hager HI *et al*. (2008) *Exploring Potential U.S. Switchgrass Production for Lignocellulosic Ethanol*. Oak Ridge National Laboratory, Oak Ridge, TN.

Guo Q, Kelly M, Graham CH (2005) Support vector machines for predicting distribution of sudden oak death in California. *Ecological Modelling*, **182**, 75–90.

Hernandez PA, Graham CH, Master LL, Albert DL (2006) The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography*, **29**, 773–785.

Hoffman JD, Narumalani S, Mishra DR, Merani P, Wilson RG (2008) Predicting potential occurrence and spread of invasive plant species along the North Platte River, Nebraska. *Invasive Plant Science and Management*, **1**, 359–367.

Jacobson MZ (2009) Review of solutions to global warming, air pollution, and energy security. *Energy and Environmental Science*, **2**, 148–173.

Jaynes ET (1957) Information theory and statistical mechanics. *Physical Review*, **106**, 620–630.

Kim S, Dale BE (2005) Environmental aspects of ethanol derived from no-tilled corn grain: nonrenewable energy consumption and greenhouse gas emissions. *Biomass and Bioenergy*, **28**, 475–489.

Lobo JM, Jimenez-Valverde A, Real R (2008) AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, **17**, 145–151.

Mani I, Kumar P, Panwar JS, Kant K (2007) Variation in energy consumption in production of wheat-maize with varying altitudes in hilly regions of himachal Pradesh, India. *Energy*, **32**, 2336–2339.

Milbrandt A (2005) *A Geographic Perspective on the Current Biomass Resource Availability in the United States*. Technical Report NREL/TP-560-39181. United States Department of Energy, National Renewable Energy Laboratory, Golden, CO.

Miller AJ, Knouft JH (2006) GIS-based characterization of the geographic distributions of wild and cultivated populations of Mesoamerican fruit tree *Spondias purpuria* (Anacardiaceae). *American Journal of Botany*, **93**, 1757–1767.

openModeller (2008) Climate data. Available at: http://openmodeller.sourceforge.net/index.php?option=com_content&task=blogcategory&id=13&Itemid=16 (accessed January 2009).

Parry ML, Rosenzweig C, Iglesias A, Livermore M, Fischer G (2004) Effects of climate change on global food production under SRES emissions and socio-economic scenarios. *Global Environmental Change*, **14**, 53–67.

Perlack RD, Wright LL, Turhollow AF, Graham RL, Stokes BJ, Erbach DC (2005) *Biomass as Feedstock for a Bioenergy and Bioproducts Industry: The Technical Feasibility of a Billion-ton Annual Supply. ORLN/TM-2005/66*. Oak Ridge National Laboratory, Oak Ridge, TN.

Peterson AT (2003) Predicting the geography of species' invasions via ecological niche modeling. *The Quarterly Review of Biology*, **78**, 419–433.

Peterson AT, Pares M, Eaton M (2007) Transferability and model evaluation in ecological niche modeling: a comparison of GARP and maxent. *Ecography*, **30**, 550–560.

Phillips SJ, Anderson RP, Schapire RE (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231–259.

Phillips SJ, Dudik M (2008) Modeling of species distribution with maxent: new extensions and a comprehensive evaluation. *Ecography*, **31**, 161–175.

Pirooznia M, Deng Y (2006) SVM classifier – a comprehensive java interface for support vector machine classification of microarray data. *BMC Bioinfomatics*, **7** (Suppl 4), S25.

Raghu S, Anderson RC, Daehler CC, Davis AS, Wiedenmann RN, Simberloff D, Mack RN (2006) Adding biofuels to the invasive species fire. *Science*, **313**, 1742.

Sartori F, Lal R, Ebinger MH, Parrish DJ (2006) Potential soil carbon sequestration and CO_2 offset by dedicated energy crops in the USA. *Critical Reviews in Plant Sciences*, **25**, 441–472.

Scheffran J, BenDor T (2009) Bioenergy and land use: a spatial-agent dynamic model of energy crop production in Illinois. *International Journal of Environment and Pollution*, **39**, 4–27.

Schmer MR, Vogel KP, Mitchell RB, Perrin RK (2008) Net energy of cellulosic ethanol from switchgrass. *Proceedings of the National Academy of Sciences USA*, **105**, 464–469.

Scholkopf B, Smola AJ, Williamson RC, Bartlett PL (2000) New support vector algorithms. *Neural Computation*, **12**, 1207–1245.

Searchinger T, Heimlich R, Houghton RA *et al*. (2008) Use of US croplands for biofuels increases greenhouse gases through emissions from land-use change. *Science*, **319**, 1238–1240.

Sissine F (2007) Energy Independence and Security Act of 2007: A Summary of Major Provisions. Congressional Research Service, Washington. Available at: http://energy.senate.gov/public/_files/RL342941.pdf (accessed June 2009).

University of California Berkeley (2008) Berkeley/Penn urban & environmental modeler's datakit. Available at: http://dcrp.ced.berkeley.edu/research/footprint/index.php?option=com_frontpage&Itemid=1 (accessed January 2009).

US Census Bureau (2009) Population estimates. Available at: http://www.census.gov/popest/counties/ (accessed January 2009).

USDA (2008) 2007 corn crop a record breaker, USDA reports. Available at: (http://www.nass.usda.gov/Newsroom/2008/01_11_2008.asp) accessed June 2009.

USDA (2009) Geospatial Gateway. Available at: http://datagateway.nrcs.usda.gov/NextPage.aspx (accessed January 2009).

USEPA (2009) EPA Lifecycle Analysis of Greenhous Gas Emission from Renewable Fuels. EPA-420-F-09-024. Environmental Protection Agency, Office of Transportation and Air Quality, Washington. Available at: http://www.epa.gov/oms/renewablefuels/420f09024.pdf (accessed June 2009)

Vapnik VN (1995) *The Nature of Statistical Learning Theory.* Springer, New York.

Varvel GE, Vogel KP, Mitchell RB, Follett RF, Kimble JM (2008) Comparison of corn and switchgrass on marginal soils for bioenergy. *Biomass and Bioenergy*, **32**, 18–21.

Walsh ME (1998) U.S. bioenergy crop economic analyses. *Status and needs. Biomass and Bioenergy*, **14**, 341–350.

Westcott PC (2007) Ethanol Expansion in the United States: How Will the Agricultural Sector Adjust? FDS-07D-01. United States Department of Agriculture, Washington. Available at: http://www.ers.usda.gov/Publications/FDS/2007/05May/FDS07D01/fds07D01.pdf (accessed December 2009).

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1. Locations and mean dry matter yields for point-sparse switchgrass model

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.