Tonini et al.: Niche Modeling of Invasive Termites

Environmental Entomology: Population Ecology

Francesco Tonini
University of Florida, Fort Lauderdale Research and Education Center, 3205 College Avenue,
Davie, Florida, 33314, U.S.A.
Phone: +1 954-577-6392
Fax: +1 954-424-6851
Email: ftonini@ufl.edu

Predicting the Geographical Distribution of Two Invasive Termite Species from Occurrence Data

FRANCESCO TONINI[1], FABIO DIVINO[2], GIOVANNA JONA LASINIO[3], HARTWIG H. HOCHMAIR[1], RUDOLF H. SCHEFFRAHN[1]

[1]University of Florida, Fort Lauderdale Research and Education Center, 3205 College Avenue, Davie, Florida, 33314, U.S.A.

[2]Division of Physics, Computer Science, and Mathematics, University of Molise, Contrada Fonte Lappone, 86090, Pesche (IS), Italy.

[3]DSS, University of Rome "La Sapienza", P.le Aldo Moro 5, 00185 Rome, Italy.

Abstract

Predicting the potential habitat of species under both current and future climate change scenarios is crucial for monitoring invasive species and understanding a species' response to different environmental conditions. Frequently, the only data available on a species is the location of its occurrence (presence-only data). Using occurrence records only, two models were used to predict the geographical distribution of two destructive invasive termite species, *Coptotermes gestroi* (Wasmann) and *Coptotermes formosanus* Shiraki. The first model uses a Bayesian linear logistic regression approach adjusted for presence-only data while the second one is the widely used maximum entropy approach (Maxent). Results show that the predicted distributions of both *C. gestroi* and *C. formosanus* are strongly linked to urban development. The impact of future scenarios such as climate warming and population growth on the biotic distribution of both termite species was also assessed. Future climate warming seems to affect their projected probability of presence to a lesser extent than population growth. The Bayesian logistic approach outperformed Maxent consistently in all models according to evaluation criteria such as model sensitivity and ecological realism. The importance of further studies for an explicit treatment of residual spatial autocorrelation and a more comprehensive comparison between both statistical approaches is suggested.

Keywords: Bayesian logistic modeling, Maxent, presence-only data, subterranean termite, species distribution models

Introduction

The use of statistical models to predict a species' potential habitat has seen a growing interest during the past two decades given the importance of monitoring endangered or invasive species and understanding a species' response to different environmental conditions (Guisan and Thuiller 2005). Such models are often referred to as habitat models, ecological niche models, or species distribution models (SDMs) (Elith and Leathwick 2007) and have been applied to a variety of fields such as ecology, conservation, and biogeography. SDMs attempt to model the species-environment relationships by using sites of known occurrence (presence data) and, sometimes, non-occurrence (absence data) together with environmental variables recorded over the whole study area. In most cases, records from atlases, herbaria, or museum databases only contain information on a species' incidental observations (Franklin 2009). A fundamental limitation of presence-only datasets is that the prevalence of a species, i.e. the proportion of occupied sites across the study area, is unknown. In recent years, several statistical methods have been proposed for modeling these types of datasets, such as inhomogeneous Poisson process (IPP), (Warton and Sheperd 2010, Chakraborty et al. 2011), and maximum entropy (Maxent) (Phillips et al. 2004, Phillips et al. 2006). Other approaches use presence-absence models by assuming random samples chosen from the region of interest (background samples) as absences (also called "pseudo-absences") (Elith et al. 2006). However, this assumption has been shown to have substantial problems of model specification, interpretation, and implementation (Warton and Sheperd 2010).

In this work, a recently developed Bayesian logistic regression model adjusted for presence-only data (Divino et al. 2011, Divino et al. 2013) and the widely used maximum entropy approach were used to predict the current and future biotic distributions of two major invasive termite pests within the state of Florida: the Asian subterranean termite (AST), *Coptotermes gestroi* (Wasmann), and the Formosan subterranean termite (FST), *Coptotermes formosanus* Shiraki. The Bayesian approach used herein has only been tested on artificial data prior to this study (Divino et al. 2013).

The highly invasive AST and FST are, or will become, the most destructive subterranean termites in areas of suitable climate, causing severe damage to wood in service (Evans et al. 2013). AST is endemic to southeast Asia and it is currently found mostly in tropical areas (Li et al. 2009). FST is probably endemic to southern China and is found primarily in subtropical and temperate regions (Li et al. 2009). AST and FST are only known to occur sympatrically in Taiwan, Florida, and Hawaii (Li et al. 2010). AST was first found in Florida in 1996 (Dade County) and is a more recent invasive species compared to FST, discovered in Florida in 1980 (Broward County) (Scheffrahn 2013). Both species are now well established pests in Florida.

Regional predictions of the potential habitat of the two termite species under both current and future climate scenarios are currently lacking in the available literature. A single recent study attempted to predict the ecological niche of AST on a global scale using mostly coarse-precision occurrence data derived from the literature (Li et al. 2013). However, the reliability of such predictions could be affected by the excessive extent of the study area used for both model calibration and estimation, given the small amount of available occurrence data.

The format of this paper is as follows. The study area, data, variables, and modeling approaches used are described in the Materials and Methods section. Results and their interpretation are then presented, followed by a final discussion on the advantages and limitations of the models tested herein.

Materials and Methods

Study Area and Species Data

Florida was selected as a common study area for both AST and FST in order to compare the performance of two different statistical approaches under the same environmental conditions. Termite collection localities, including winged reproductives and/or nonvolent foragers, were taken from the University of Florida Termite Collection at the Fort Lauderdale Research and Education Center. Winged reproductives were taken from within infested structures, and therefore in close proximity to their foraging nest mates and stationary nests. Geographical coordinates of 280 and 411 separate land-based colonies of AST (1996-2012) and FST (1985-2012), respectively, were used in this study (Fig. 1). A few records representing boat infestations (Scheffrahn and Crowe 2011) were excluded. All database samples were collected less than 40m from buildings by R.H.S., pest control professionals, property owners, entomologists, and others interested in species-level identification. About 95% of foraging caste samples were collected within 5-10 m or inside the structures themselves.

Figure 1–caption at the end of file

The study area was divided into roughly 38,000 2-km grid cells and all termite observations falling within a given cell were aggregated to a single point. After aggregation, a total of 65 and 160 occurrences were considered for AST and FST, respectively. In this work, grid cells were considered as independent given the explanatory variables and the probability of presence was modeled for each one of them. The chosen spatial resolution attenuates some of the bias caused by spatial dependence between nearby occurrences because termite reproductives from a mature colony fly only a few hundred meters during their annual dispersal flights (Nutting 1969). Moreover, the available environmental data used in this study were obtained at a 2.5-arcmin (~4 km) resolution and it is appropriate to consider a sampling unit whose size is equal (or close) to it (Elith and Leathwick 2009). Finally, the number of occurrences available after the aforementioned spatial aggregation ensures robustness of the estimates from the statistical models used herein.

Predictor Variables

A set of gridded climatic variables was selected (Table 1) based on both its ability to directly influence the ecophysiology of both AST and FST (Gautam and Henderson 2011), and on suggestions taken in consultation with termite experts. Data for historical climatic conditions were extracted from two sources: (i) the PRISM Climate Group database (Daly et al. 2002) and (ii) the WorldClim (1950-2000) database (Hijmans et al. 2005). General annual trends such as annual total precipitation (prec), average daily mean dew point temperature (dew), and average daily maximum (tmax) and minimum (tmin) temperatures were obtained from the PRISM database, representative of average historical conditions for the years of available occurrence records of both AST and FST. Two bioclimatic variables representing extreme or limiting factors such as maximum temperature of the warmest month (bio5) and minimum temperature of the coldest month (bio6) were chosen from the WorldClim database, representative of 1950-2000 average historical conditions. Both WorldClim and PRISM data were obtained at a 2.5-arcmin (~4 km) resolution and further resampled down using bilinear interpolation to maintain the higher data resolution of the reference spatial grid over the study area. The available time series of historical climate data from PRISM (1895-Present) allowed us to extract those years that matched historical occurrence records for both AST and FST exactly.

In addition to climate variables, the U.S. Geological Survey National Land Cover Database (NLCD) 2006 (Multi-Resolution Land Characteristics Consortium 2012) was also used (see Table 1), which has a native resolution of 30 m. The database comes with 20 land cover classes, which were modified according to the following steps: (1) reduction from 20 to 8 main land cover classes according to the NLCD 2006 product legend; (2) creation of single layers for each land cover class from the previous step; and (3) aggregation of each land cover layer from 30 m to our 2-km reference grid by expressing each cell value as the percentage of land cover contained within.

Finally, centroids of grid cells occupied by termite locations were also used in some of the statistical models (see Tables 2 and 3) in order to account for the geographic proximity between collection sites across the geographic space. Locations were expressed by their projected easting and northing values. All layers, including the 2-km reference grid, were mapped using the NAD83 / Florida GDL Albers projection to minimize distance distortions throughout the study area.

Table 1–caption at the end of file

Most predictor variables in our dataset were highly correlated and their simultaneous presence in statistical models has been proven to cause several problems (e.g. biased parameter estimates or lower efficiency in the estimates) (Farrar and Glauber 1967). Therefore, an *a priori* choice of variables was carried out in order to exclude pairs of highly correlated variables ($r \geq 0.8$). A different set of models was also estimated using principal components obtained from the full set of predictor variables considered herein. Principal component loadings are shown in Supp. Table S1 (available in the online version).

Future Scenarios

The Intergovernmental Panel on Climate Change (IPCC) Fourth Assessment Report (AR4) describes a set of alternative CO_2 emissions scenarios grouped under four main narrative storylines (Intergovernmental Panel on Climate Change 2007). In this study, the A2 emission scenario was used, which forecasts an average increase in global surface temperature of about 3.4°C by 2100. This scenario was preferred over others in order to assess the impact of a larger climate change on both termite species' potential distributions and consider it as a benchmark "worst-case" scenario.

Given the uncertainty associated with the path of future climate change, average projections of annual precipitation and minimum/maximum temperatures for the years 2040-2069 (referred to as 2050s hereafter) were extracted from the Climate Change, Agriculture and Food Security (CCAFS) web portal (Climate Change Agriculture and Food Security 2013). Projections of annual mean dew point temperatures were not available from any data provider, hence this predictor variable could not be considered for future scenarios.

The three following Atmospheric-Oceanic Global Circulation Models (GCMs hereafter) (Diniz-Filho et al. 2009), statistically downscaled using the so-called delta method (Ramirez-Villegas and Jarvis 2010), were selected for the A2 emission scenario and the 2050s time frame: GFDL-CM 2.1, NCAR-CCSM 3.0, UKMO-HadCM3.

A projected population growth scenario in 2060 was obtained from the University of Florida GeoPlan Center (University of Florida Geoplan Center 2013). The dataset assumes no further population growth in areas currently urbanized.

First, we assumed a change in the climatic variables under the A2 emission scenario given by the three selected GCMs, assuming no change in population. Then, we added a population growth scenario together with climate change, resulting in a total of six future scenarios for each species. To create "consensus" maps of projected probabilities in the 2050s, predictions were averaged over the three GCMs. This method has been shown to significantly increase the accuracy of species distribution forecasts (Marmion et al. 2009).

Modeling Approaches

In this study, several models were considered using two statistical approaches for presence-only data (see Tables 2 and 3): (i) maximum entropy (Phillips et al. 2006); and (ii) a Bayesian linear logistic regression adjusted for presence-only data, named Bayesian for presence-only data (BPOD) hereafter (Divino et al. 2011, Divino et al. 2013). The former, was presented in Phillips et al. (2004) and it is widely used for modeling distributions of species (Elith et al. 2011). The BPOD, builds upon the work presented in Ward et al. (2009) while using a Bayesian framework.

Although different in their theoretical backgrounds, both methods use the Bayes' rule as an important point to calculate the probability of presence of the species conditioned on the environment. An outline of the theory, main assumptions, and modeling settings used in both approaches follows.

Maximum Entropy Approach.

Maximum entropy (Maxent hereafter) is a machine-learning method that uses species occurrences and a random sample of background environmental data over a region of interest to predict species distributions. Let us define $Pr(X = x \mid Y = 1)$ to be the probability distribution of covariates, i.e. environmental variables, across locations where the species is observed (Y = 1), and $Pr(X = x \mid Y = 0)$ to be the probability distribution of covariates where the species is absent (Y = 0). The quantity of interest is the probability of presence of a species, $Pr(Y = 1 \mid X = x)$, conditioned on a set of environmental covariates X. Maxent considers the modeling of $Pr(X = x|Y = 1)$ and uses the Bayes' rule to estimate the sought conditional probability distribution:

$$Pr(Y = 1 \mid X = x) = \frac{Pr(x \mid Y = 1)Pr(Y = 1)}{Pr(x)}$$

The core of the Maxent "raw" model output is the estimate of the ratio $Pr(x \mid Y = 1)/P(x)$. This is accomplished by seeking an estimate of $Pr(x \mid Y = 1)$ that is consistent with available occurrence data. Among several possible distributions, one that maximizes the entropy of $Pr(x \mid Y = 1)$ or, in other words, minimizes the relative entropy of $Pr(x \mid Y = 1)$ with respect to $Pr(x)$ (measured using the Kullback-Leibler divergence) is chosen. The distribution of maximum entropy, i.e. closest to the uniform probability distribution or most spread out, is estimated while being subject to a set of constraints imposed by the information available from the environmental conditions where the species occurs.

Environmental variables or functions thereof are known as "features" and are treated as an expanded set of variables to be added as terms in the model specification. A random sample of background locations informs the model about $Pr(x)$. The set of constraints on $Pr(x \mid Y = 1)$ ensures that empirical averages of each feature approximate their averages at sites where the species is present (or a random sample thereof).

The probability distribution of maximum entropy is a Gibbs distribution, which has an exponential form (Della Pietra et al. 1997). Raw exponential values estimated by the model are scale-dependent, e.g. they can be extremely small if the study area is large, and only represent a measure of relative suitability of each site. However, the model can also be transformed from an exponential family model into a logistic model, thus making it more comparable with other machine learning or generalized linear/additive models (Phillips and Dudik 2008).

To calculate the final conditional probability of occurrence $Pr(Y = 1|X = x)$, knowledge of the prevalence of the species $Pr(Y = 1) = \pi$, i.e. the proportion of occupied sites across the study area, is required. However, π is unknown with presence-only data (Ward et al. 2009). In this case, the maximum entropy approach sets this quantity arbitrarily to 0.5.

Bayesian for Presence-only Data (BPOD) Approach.

When dealing with presence-only data, sampling from the reference population of locations cannot be performed under the traditional random sampling design. Specifically, while a random sample of presences is available, a random sample of absences cannot be obtained. Therefore, a random sample of "contaminated controls", i.e. a random sample of locations from the whole reference population (background sample) that can also include some occurrences of the species, is matched with the aforementioned random sample from the available occurrence data (Lancaster and Imbens 1996).

In order to estimate the regression parameters, a two-level scheme is used: (1) a first level describing the probability law that generates the population data; and (2) a second level using a stratified case-control design, modified for presence-only data to select samples from the population. In a traditional logistic regression, the response variable $Y = 0$ marks the absence of an attribute of interest in the population, while $Y = 1$ marks the presence of the same attribute. The key point in the BPOD approach is the introduction of a stratum variable Z, considered as the only observable variable. Specifically, $Z = 0$ means that a location is collected from the whole reference population, while $Z = 1$ indicates that a location is collected from the sub-population of presences. $Z = 1$ implies that $Y = 1$, while $Z = 0$ implies that Y is an unknown value $y \epsilon \{0,1\}$. The introduction of the stratum variable Z allows us to define a linear logistic regression, adjusted for presence-only data. Denoting by $Pr(Z = 1|C = 1, X = x)$ the probability that a location is sampled ($C = 1$) from the set of locations where the species of interest is present ($Z = 1$) and with covariates $X = x$, the linear logistic model for presence-only data can be defined as:

$$\text{logit}Pr(Z = 1|C = 1, X = x) = x\beta + q,$$

where q is a correction term, depending on the number of presences truly observed and the unknown number of presences hidden in the sample of "contaminated" controls. An approximation of q can be derived iteratively within the estimation algorithm. After prior distributions are defined over the parameters of interest (the linear coefficients β and the unobserved responses in the sample of "contaminated" controls), Bayesian inference can be carried out through Markov Chain Monte Carlo (MCMC) techniques (Robert and Casella 2004). In particular, an algorithm including a data-augmentation step (Tanner and Wong 1987) is used to obtain an estimate of the unknown empirical prevalence π of the species of interest, jointly with linear coefficients of the logistic model.

Evaluation of Model Performance.

In this study, model performance is evaluated according to three criteria: (i) prediction accuracy of occurrence data, i.e. model sensitivity expressed by the percentage of correctly predicted occurrences in the sample; (ii) goodness of fit, using both the Akaike Information Criterion (AIC, Akaike 1974) and its corrected version (AICc, Burnham and Anderson 2002); and (iii) ecological realism, i.e. assessing predictions against prior biological knowledge of a species. AIC and AICc for all Maxent models were calculated using the ENMTools (Warren and Seifert 2011) which uses Maxent "raw" suitability scores, i.e. exponential values standardized over the study area. Several other traditional statistical evaluation metrics such as Cohen's Kappa (Cohen 1960) or the area under the receiver operating characteristic curve (AUC, Hanley and Mcneil 1982) are commonly used with presence-absence (or pseudo-absence) data. However, in this study we do not make any assumption of pseudo-absence for background data. While model sensitivity was compared across all models and both statistical approaches, AIC and AICc values were only used to compare the relative quality of each model within the same statistical approach in order to provide a mean for model selection. This is crucial because Maxent’s model structure is different from BPOD, hence values of both AIC and AICc cannot be compared across models considered in both approaches.

Sampling Scheme.

The following background sampling schemes were used with respect to Maxent and BPOD modeling approaches.

Each Maxent model was run 16 times, with the background sample size set to 10,000 randomly selected points. Although there are not set guidelines regarding the ideal number of background points to use in each situation, some recent studies found that predictive accuracy of Maxent was best with about 10,000 points (Barbet-Massin and Jiguet 2012) over areas comparable in size to our study. Moreover, some studies found that predictive accuracy of Maxent was best with about 10,000 points (Barbet-Massin and Jiguet 2012). All other settings in the MaxEnt software have been used with their default values (Phillips et al. 2006).

Each BPOD model was run 500 times, with sample size set according to the presence/background ratio of 1:4, as used by Ward et al. (2009). Specifically, in AST a sample of 65 observed presences was matched with a background sample of 65×4=260 locations (total sample size n=325), while for FST a sample of 160 observed presences was matched with a background sample of 160×4=640 locations (total sample size n=800). The MCMC algorithm with data augmentation used 15,000 iterations (10,000 burn-in) to estimate the unknown model parameters.

The reason for using different sampling schemes between Maxent and BPOD is due to the fact that the two approaches have different requirements for reaching robust parameter estimates.

Specifically, Maxent needs a large background sample, while BPOD needs a large number of model replications. Given these constraints, we chose model settings accordingly and used roughly the same amount of “sampling information” (see Supp. Table S2-S3 available for the online version). In both approaches, parameter estimates were obtained as averages over all model replications.

Results

Several models were run to predict the current potential distribution of both AST and FST. A list of the best performing models is shown in Table 2 for AST and Table 3 for FST, together with their evaluation metrics.

Table 2–caption at the end of file

Table 3–caption at the end of file

For AST, the model that reached the highest overall performance in the maximum entropy approach was M1, while in the BPOD approach it was MPC3, which used the first three principal components as covariates. Fig. 2 (a-b) shows the current potential distributions of AST predicted by the best overall models in both approaches, thus BPOD-MPC3 and Maxent-M1, respectively. Southeastern Florida and the Keys Islands show a much higher suitability compared to other areas, matching the general pattern of recorded occurrences. Low probabilities are also predicted along the east coast up to central Florida and on the west coast around urban areas such as Ft. Myers and Tampa.

Figure2–caption at the end of file

For FST, the model that reached the highest overall performance in the maximum entropy approach was MPC3, while in the BPOD approach it was MPC6, using the first three principal components and all six principal components as covariates, respectively. Fig. 3 (a-b) shows the current potential distributions of FST predicted by the best overall models in both approaches, thus BPOD-MPC6 and Maxent-MPC3, respectively. Highest suitability values are associated with urbanized areas across the entire state. Although no occurrences were recorded in some urban areas, a medium-to-high suitability is predicted for the species in areas such as North-West Florida around Pensacola, along the west coast in Sarasota and Port Charlotte, along the east coast in Melbourne and Palm Coast, and all the Keys islands south-west of Key Largo.

Figure3–caption at the end of file

Future predicted probabilities of presence were derived using a model from the BPOD approach for both AST and FST. Due to data availability (see Materials and Methods), the BPOD-M1 model was chosen to predict their future distributions. Fig. 4 (a-b) shows the contemporary predictions calculated using model BPOD-M1 for AST and FST, respectively. A visual inspection suggests that predictions are not much different from the best models that used principal components as covariates, with the exception of a few areas for FST such as the Keys

Islands or the west coast of Florida where the suitability is slightly lower. Fig. 4 (c-d) shows average "consensus" projected probabilities, i.e. averaged over the three GCMs, for AST and FST, respectively, under climate change conditions for the 2050s time period and given no change in land cover. The climate variables that are projected to the future from M1 are precipitation, bio5, and bio6. Urban areas in southeast Florida seem to have an increased predicted probability of presence for AST, while for FST changes in suitability are less noticeable. The population growth scenario (Fig. 4 e-f) increases the percentage of areal units occupied by developed areas, thus increasing the variable "devel" (refer to Table 1) in our model. The effect of a combined change in climate and developed areas increases the predicted probabilities of presence for both AST and FST. However, the effect is much more noticeable for the latter across the whole study area.

Figure4–caption at the end of file

Discussion

The performance of the BPOD approach on both species was shown to be consistently better than the widely used maximum entropy method, with a few exceptions, in terms of sampling sensitivity (see Table 3). Whenever the model covariates were highly informative on a species geographical distribution (e.g., for AST), the BPOD approach performed consistently better than maximum entropy. In fact, the highest sensitivity reached by any Maxent model was 61%, hence lower than the worst BPOD model (76%). When the model covariates are less informative for predicting distribution, as for the FST, BPOD performs better than MaxEnt in 78% of the cases.

Finally, the best BPOD model gave more realistic predictions from an ecological perspective compared to the best Maxent model for both species. Specifically, for FST maximum entropy tends to over-predict areas across the entire state, far apart from recorded occurrences, and under-predict areas close to them (Fig. 3). Although this phenomenon is less pronounced for AST, areas in the metropolitan southeast Florida are under-predicted nearby recorded occurrences.

The BPOD approach makes a better use of the information from PCA-derived variables compared to maximum entropy, as its predictive power increases until reaching an optimum in terms of sensitivity and both information criteria (see Tables 2-3). However, such models behave in a slightly different manner between the two species. In particular, BPOD models for AST reach an optimum with a smaller number of PCA-derived variables than FST (3 vs. 6 principal components, respectively). This probably means that the original environmental variables enclosed in the first three principal components are sufficient to explain the ecological niche of AST in Florida. FST, tolerating broader climatic and environmental gradients than AST, has attained generic species status in Florida where it occurs in all major human population centers of the State. This result also suggests that some environmental factors influencing the habitat of FST may be missing from these analyses.

Maxent models reported in this paper were estimated by fitting linear responses to relationships between response and predictor variables in order to keep comparability between the two different statistical approaches. Maxent models fitting more complex responses were also tested but had a much lower predictive performance compared to the ones fitting linear features. A major advantage of the BPOD approach over maximum entropy is that the MCMC algorithm does not require the *a priori* knowledge of the population prevalence as it is considered as a further parameter in the model. This overcomes the issue of prevalence specification pointed out by Ward et al. (2009). A Bayesian modeling framework allows flexibility in the treatment of uncertainty while making full inference on the probability of presence possible. However, a more formal comparison between the two statistical approaches based on artificial data is suggested for future studies.

In this paper, statistically downscaled climate projections for the 2050s were preferred over dynamically downscaled projections, such as the CLARReS10 dataset for the Southeast United States (Stefanova et al. 2012). Although the latter are able to incorporate regional-scale processes, their spatial resolution (~10 km) was too coarse to assess the effect of variation in climate and urbanization on the same scale used for contemporary predictions for both termite species. Climate change under the A2 scenario for the 2050s has a moderate effect on both species' geographical distribution. Conversely, a combined effect of climate change with a population growth scenario has a larger impact on their projected probabilities, especially for FST. This suggests that both termite species are influenced by urban development much more than by climate alone.

Two issues not fully addressed in this work are the residual autocorrelation that may still persist among neighboring occurrences and the problem of observer bias (Syfert et al. 2013). In order to reduce spatial autocorrelation, we chose a spatial resolution at which termite occurrences can be assumed independent of each other given the explanatory variables (see Materials and Methods). Spatially explicit models, i.e. models with spatial autoregressive component (Cressie 1993) or latent spatially structured component (Zuur et al. 2009), might be available to refine our final predictions. However, a reasonable way of generating pseudo-absences must be found and these models are computationally intensive to estimate. The issue of observer bias would be hard to address in the models developed herein because the data comes from different sources and involves multiple data collectors (see Materials and Methods). All samples were not collected using road accessibility criteria, hence standard solutions, e.g. adding information on road distance within the models (Phillips and Dudik 2008), could not be implemented in this study. The treatment of such a complex issue is deferred to future work.

Acknowledgements

The authors would like to thank both anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper. We would like to thank David Bucklin and James Watling from the University of Florida for their valuable help and suggestions throughout the development of this work. The authors would also like to thank Lydia Stefanova from the Center for Oceanic-Atmospheric Prediction Studies at Florida State University for her valuable input on downscaled climate projections. The MCMC algorithm used to estimate parameters in the Bayesian logistic regression approach was developed in FORTRAN by Dr. Fabio Divino (co-author, email: fabio.divino@unimol.it).

References

Cohen, J. 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20: 37–46.

Farrar, D. E., and R. R. Glauber. 1967. Multicollinearity in regression analysis: the problem revisited. Rev. Econ. Stat. 49: 92-107.

Nutting, W. L. 1969. Flight and colony foundation. *In* K. Krishna, F. M. Weesner (eds.), Biology of termites. Academic Press, New York, 233-282.

Akaike, H. 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19: 716–723.

Hanley, J. A., and B. J. McNeil. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiol. 143: 29-36.

Tanner, M., and W. Wong. 1987. The calculation of posterior distribution by data augmentation. J. Am. Stat. Assoc. 82: 528-550.

Cressie, N. 1993. Statistics for spatial data. John Wiley & Sons, Inc., New York.

Lancaster, T., and G. Imbens. 1996. Case-control studies with contaminated controls. J. Econ. 71: 145-160.

Della Pietra, S., V. Della Pietra, and J. Lafferty. 1997. Inducing features of random fields. IEEE Trans. Pattern Anal. Mach. Intell. 19: 380-393.

Burnham, K. P., and D. R. Anderson. 2002. Model selection and multimodel inference: a practical information-theoretic approach, 2nd ed. Springer-Verlag.

Daly, C., W. P. Gibson, G. H. Taylor, G. L. Johnson, and P. Pasteris. 2002. A knowledge-based approach to the statistical mapping of climate. Clim. Res. 22: 99-113.

Phillips, S. J., M. Dudík, and R. E. Schapire. 2004. A maximum entropy approach to species distribution modeling. Proc. 21st Int. Conf. Mach. Learn. 655-662.

Robert, C. P., and G. Casella. 2004. Monte Carlo statistical methods, 2nd ed. Springer-Verlag New York, Inc., Secaucus, NJ.

Guisan, A., and W. Thuiller. 2005. Predicting species distribution: offering more than simple habitat models. Ecol. Lett. 8: 993-1009.

Hijmans, R. J., S. E. Cameron, J. L. Parra, P. G. Jones, and A. Jarvis. 2005. Very high resolution interpolated climate surfaces for global land areas. Int. J. Clim. 25: 1965-1978.

Elith, J., C. H. Graham, R. P. Anderson, M. Dudík, S. Ferrier, A. Guisan, R. Hijmans, F. Huettmann, J. R. Leathwick, A. Lehmann, J. Li, L. G. Lohmann, B. A. Loiselle, G. Manion, C. Moritz, M. Nakamura, Y. Nakazawa, J. M. Overton, A. T. Peterson, S. J. Phillips, K. Richardson, R. Schachetti-Pereira, R. E. Schapire, J. Soberon, S. Williams, M. S. Wisz, and N. E. Zimmermann. 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecogr. 29: 129-151.

Phillips, S. J., R. P. Anderson, and R. E. Schapire. 2006. Maximum entropy modeling of species geographic distributions. Ecol. Model. 190: 231-259.

Elith, J., and J. Leathwick. 2007. Predicting species distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines. Divers. Distrib. 13: 265-275.

Intergovernmental Panel on Climate Change. 2007. IPCC fourth assessment report (AR4). Url: http://goo.gl/3C2amF. Accessed: 03/24/2014

Phillips, S. J., and M. Dudik. 2008. Modeling of species distributions with maxent: new extensions and a comprehensive evaluation. Ecogr. 31: 161-175.

Diniz-Filho, J. A., L. M. Bini, T. F. Rangel, R. D. Loyola, C. Hof, D. Nogués-Bravo, and M. B. Araújo. 2009. Partitioning and mapping uncertainties in ensembles of forecasts of species turnover under climate change. Ecogr. 32: 897-906.

Elith, J., and J. R. Leathwick. 2009. Species distribution models: ecological explanation and prediction across space and time. Annu. Rev. Ecol. Evol. Syst. 40: 677-697.

Franklin, J. 2009. Mapping species distributions: spatial inference and prediction. Cambridge University Press, Cambridge, UK.

Li, H.-F., W. Ye, N.-Y. Su, and N. Kanzaki. 2009. Phylogeography of *Coptotermes gestroi* and *Coptotermes formosanus* (Isoptera: Rhinotermitidae) in Taiwan. Ann. Entomol. Soc. Am. 102: 684-693.

Marmion, M., M. Parviainen, M. Luoto, R. K. Heikkinen, and W. Thuiller. 2009. Evaluation of consensus methods in predictive species distribution modelling. Divers. Distrib. 15: 59-69.

Ward, G., T. Hastie, S. C. Barry, J. Elith, and J. R. Leathwick. 2009. Presence-only data and the EM algorithm. Biom. 65: 554-563.

Zuur, A., E. N. Ieno, N. Walker, A. A. Saveliev, and G. M. Smith. 2009. Mixed effect models and extensions in ecology with R. Springer.

Li, H.-F., R. L. Yang, and N. Y. Su. 2010. Interspecific competition and territory defense mechanisms of *Coptotermes formosanus* and *Coptotermes gestroi* (Isoptera: Rhinotermitidae). Env. Entomol. 39: 1601-1607.

Ramirez-Villegas, J., and A. Jarvis. 2010. Downscaling global circulation model outputs: the delta method. Decis. Policy Anal. Working Pap. No. 1.

Warton, D. I., and L. C. Sheperd. 2010. Poisson point process models solve the “pseudo-absence problem” for presence-only data in ecology. Ann. Appl. Stat. 4: 1383-1402.

Chakraborty, A., A. E. Gelfand, A. M. Wilson, A. M. Latimer, and J. A. Silander. 2011. Point pattern modelling for degraded presence-only data over large regions. J. R. Stat. Soc.: Series C (App. Stat.) 5: 757-776.

Divino, F., N. Golini, G. J. Lasinio, and A. Penttinen. 2011. Data augmentation approach in Bayesian modelling of presence-only data. Procedia Environ. Sci. 7: 38-43.

Elith, J., S. J. Phillips, T. Hastie, M. Dudík, Y. E. Chee, and C. J. Yates. 2011. A statistical explanation of maxent for ecologists. Divers. Distrib. 17: 43-57.

Gautam, B. K., and G. Henderson. 2011. Relative humidity preference and survival of starved Formosan subterranean termites (Isoptera: Rhinotermitidae) at various temperature and relative humidity conditions. Env. Entomol. 40: 1232-1238.

Scheffrahn, R. H., and W. Crowe. 2011. Ship-borne termite (Isoptera) border interceptions in Australia and onboard infestations in Florida, 1986–2009. Fla. Entomol. 94: 57-63.

Warren, D. L., and S. Seifert. 2011. Environmental niche modeling in maxent: the importance of model complexity and the performance of model selection criteria. Ecol. Appl. 21: 335-342.

Barbet-Massin, M., and F. Jiguet. 2012. Selecting pseudo-absences for species distribution models: How, where and how many? Methods Ecol. Evol. 3: 327-338.

Multi-Resolution Land Characteristics Consortium, 2012. National land cover database 2006. http://www.mrlc.gov/nlcd2006.php.

Stefanova, L., V. Misra, S. Chan, M. Griffin, J. J. O’Brien, and T. J. I. Smith. 2012. A proxy for high-resolution regional reanalysis for the southeast united states: Assessment of precipitation variability in dynamically downscaled reanalyses. Climate Dynamics 38: 2449-2466.

Climate Change Agriculture and Food Security. 2013. GCM downscaled data portal. Url: http://www.ccafs-climate.org/data/. Accessed: 08/21/2013

Divino, F., N. Golini, G. Jona Lasinio, and A. Penttinen. 2013. Bayesian modeling and MCMC computation in linear logistic regression for presence-only data. (submitted). Cornell University Library. Url: http://arxiv.org/abs/1305.1232. Accessed: 08/21/2013

Evans, T. A., B. T. Forschler, and J. K. Grace. 2013. Biology of invasive termites: a worldwide review. Ann. Rev. Entomol. 58: 455-474.

Li, H.-F., I. Fujisaki, and N.-Y. Su. 2013. Predicting habitat suitability of *Coptotermes gestroi* (Isoptera: Rhinotermitidae) with species distribution models. J. Econ. Entomol. 106: 311-321.

Scheffrahn, R. H. 2013. Overview and current status of non-native termites (Isoptera) in Florida. Fla. Entomol. (in press).

Syfert, M. M., M. J. Smith, and D. A. Coomes. 2013. The effect of sampling bias and model complexity on the predictive performance of maxent species distribution models. PLoS ONE 8.

University of Florida GeoPlan Center. 2013. Florida projected population growth - 2060 Url: http://www.fgdl.org. Accessed: 08/21/2013

Table 1. Climatic and environmental variables used and their data source for both AST and FST.

Database	Variable	Description
PRISM	prec	Annual total precipitation
	dew	Average daily mean dew point temperature
	tmax	Average daily maximum temperature
	tmin	Average daily minimum temperature
WORLDCLIM	bio5	Maximum temperature of the warmest month
	bio6	Minimum temperature of the coldest month
NLCD 2006	water	Open water or permanent ice/snow cover
	devel	High percentage (≥ 30%) of constructed materials (e.g. asphalt, concrete, buildings, etc.).
	barren	Bare rock, gravel, sand, silt, clay, or other earthen material, with little or no "green" vegetation
	forest	Tree cover (> 6 m tall). Tree canopy accounts for 25% to 100% of the cover
	shrub	Natural/semi-natural woody vegetation with aerial stems (< 6 m tall)
	herb	Natural/semi-natural herbaceous vegetation (75% - 100% of the cover)
	cultiv	Herbaceous vegetation that has been planted or is intensively managed for the production of food, feed, or fiber (75% - 100% of the cover)
	wetlands	Soil or substrate is periodically saturated with or covered with water

Table 2. List of main models used for AST, their sampling sensitivity, and information criteria. M1: X (easting), prec, bio5, bio6, all land cover variables. M2: prec, bio5, bio6, all land cover variables. M3: X (easting), Y (northing), prec, bio5, all land cover variables. MPC*x*: *x* stands for the number of principal components used as covariates. The best models are highlighted in bold.

Approach	Model	Sampling Sensitivity	AIC	AICc
	M1	0.96	62.2	38.2
	M2	0.96	63.2	38.4
	M3	0.96	67.7	41.1
	MPC1	0.76	118.3	112.4
BPOD	MPC2	0.90	78	70.2
	MPC3	**0.97**	**48.8**	**39**
	MPC4	0.97	50.1	38.3
	MPC5	0.97	51.4	37.8
	MPC6	0.97	52.5	36.9
	M1	**0.61**	**823.4**	**825.2**
	M2	0.61	828.7	830.6
	M3	0.60	857.6	860
	MPC1	0.40	1066	1066
Maxent	MPC2	0.61	942	942
	MPC3	0.60	828.8	829.2
	MPC4	0.60	829.1	829.8
	MPC5	0.58	831.7	832.5
	MPC6	0.57	830.2	831.1

Table 3. List of main models used for FST, their sampling sensitivity, and information criteria. M1: X (easting), prec, bio5, bio6, all land cover variables. M2: prec, bio5, bio6, all land cover variables. M3: X (easting), Y (northing), prec, bio5, all land cover variables. MPC*x*: *x* stands for the number of principal components used as covariates. The best models are highlighted in bold.

Approach	Model	Sampling Sensitivity	AIC	AICc
BPOD	M1	0.73	391.1	366
	M2	0.73	391.4	367
	M3	0.73	391.5	367.2
	MPC1	0.05	689.2	683.2
	MPC2	0.53	529.7	521.8
	MPC3	0.71	396.3	386.4
	MPC4	0.71	395.1	383.2
	MPC5	0.72	388.6	374.8
	MPC6	**0.74**	**382.8**	**367**
Maxent	M1	0.55	2712.9	2714.2
	M2	0.55	2713.1	2714
	M3	0.57	2712	2713.2
	MPC1	0.66	1177	1177.1
	MPC2	0.59	1047.4	1047.6
	MPC3	**0.66**	**956.8**	**957.2**
	MPC4	0.57	968.5	969.1
	MPC5	0.61	963.6	964.6
	MPC6	0.58	961.1	962.5

Figure Captions:

Fig. 1. Florida occurrences of AST (green) and FST (purple). Available in color online.

Fig. 2. Current predicted probabilities of presence for AST. (a) BPOD-MPC3 model. (b) Maxent-M1 model. Darker red areas correspond to areas with higher probabilities. Available in color online.

Fig. 3. Current predicted probabilities of presence for FST. (a) BPOD-MPC6 model. (b) Maxent-MPC3 model. Darker red areas correspond to areas with higher probabilities. Available in color online.

Fig. 4. Current and average projected probabilities of presence for the 2050s time period. (a) BPOD-M1 contemporary predictions for AST. (b) BPOD-M1 contemporary predictions for FST. (c) BPOD-M1 projected predictions for AST averaged over the GFDL-CM 2.1, NCAR-CCSM 3.0, and UKMO-HadCM3 global circulation models under the A2 emission scenario. (d) BPOD-M1 projected predictions for FST averaged over the GFDL-CM 2.1, NCAR-CCSM 3.0, and UKMO-HadCM3 global circulation models under the A2 emission scenario. (e) BPOD-M1 projected predictions for AST averaged over the GFDL-CM 2.1, NCAR-CCSM 3.0, and UKMO-HadCM3 global circulation models under the A2 emission scenario and population growth. (f) BPOD-M1 projected predictions for FST averaged over the GFDL-CM 2.1, NCAR-CCSM 3.0, and UKMO-HadCM3 global circulation models under the A2 emission scenario and

population growth. Darker red areas correspond to areas with higher probabilities. Available in color online.

FIGURE 1

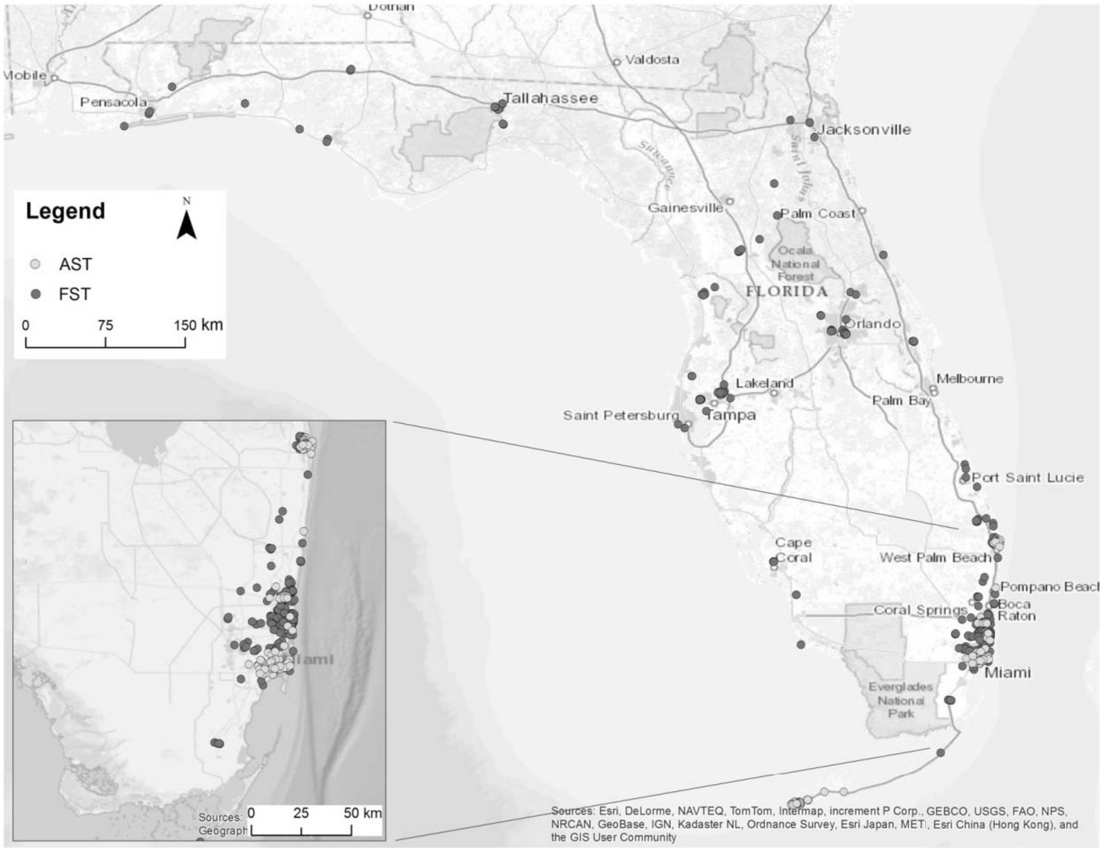
Legend
AST
FST
0 75 150 km
N
Mobile
Pensacola
Dothan
Tallahassee
Valdosta
Jacksonville
Gainesville
Palm Coast
Ocala National Forest
FLORIDA
Orlando
Lakeland
Tampa
Saint Petersburg
Melbourne
Palm Bay
Port Saint Lucie
Cape Coral
West Palm Beach
Pompano Beach
Coral Springs
Boca Raton
Miami
Everglades National Park
0 25 50 km
Sources: Esri, DeLorme, NAVTEQ, TomTom, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), and the GIS User Community

FIGURE 2

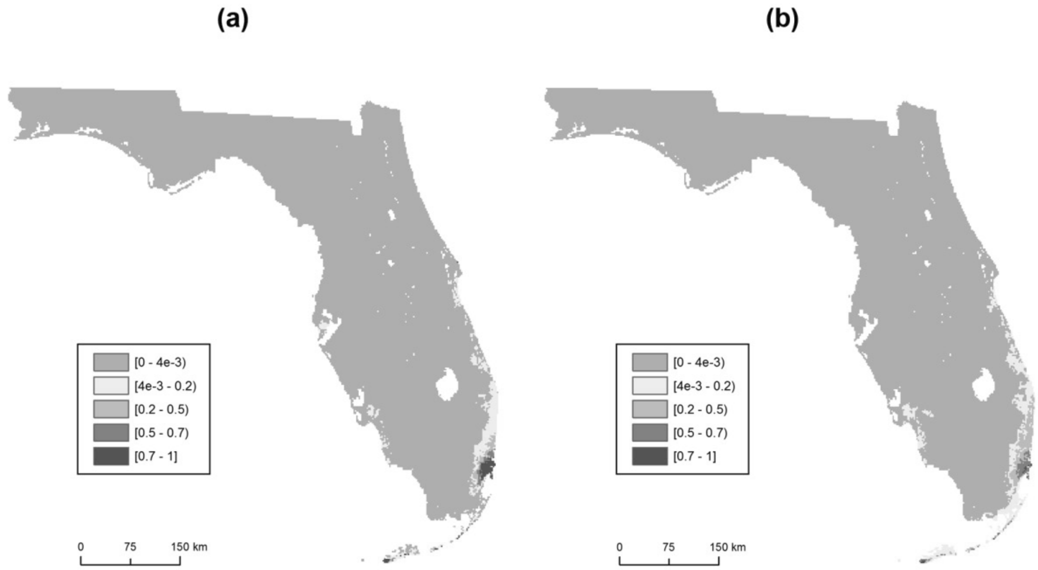

FIGURE 3

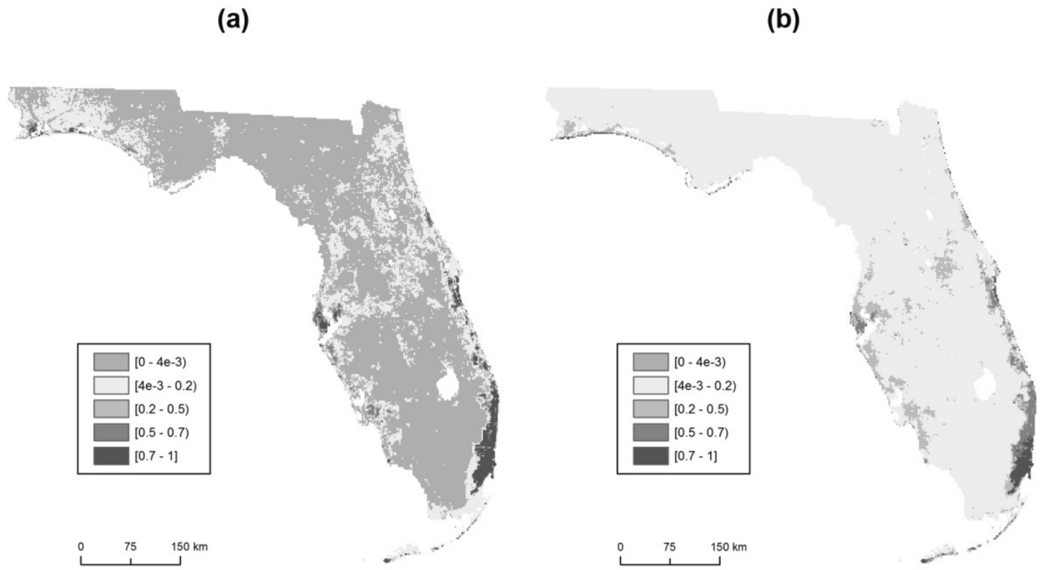

FIGURE 4