Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2016 doi: 10.1111/2041-210X.12599

Improving distribution data of threatened species by combining acoustic monitoring and occupancy modelling

Marconi Campos-Cerqueira[1]* and T. Mitchell Aide[1,2]

[1]*Department of Biology, University of Puerto Rico-Rio Piedras, San Juan, 00931-3360 Puerto Rico; and* [2]*Sieve Analytics, 7 Gertrudis, San Juan, 00911 Puerto Rico*

Summary

1. Conservation of threatened species relies on predictions about their spatial distribution; however, it is often difficult to detect species in the wild. The combination of acoustic monitoring to improve species detectability and statistical methods to account for false-negative detections can improve species distribution estimates.

2. Here, we combine a novel automated species-specific identification approach with occupancy models that account for imperfect detectability to provide a more accurate species distribution map of the Elfin Woods Warbler *Setophaga angelae*, a rare, elusive and threatened bird species. We also compared three automated species identification/validation approaches to determine which approach provided occupancy estimates similar to manual validation of all recordings. Acoustic data were collected along three elevational gradients (95–1074 m a.s.l) in El Yunque National Forest, Puerto Rico. The detection matrices acquired through automated species-specific identification models and manual validations of all recordings were used to create occupancy models.

3. Although this species has a wider distribution than previously reported, it depends on Palo Colorado forest cover and it mainly occurs between 600 and 900 m a.s.l. Unbiased and precise occupancy models were developed by using automated species identification models and only manually validating 4% of the recordings.

4. Our approach draws on the strength of two active areas of ecological research: acoustic monitoring and occupancy modelling. Our methods provide an effective and efficient way to translate the enormous amount of acoustic information collected with passive acoustic monitoring devices into meaningful ecological data that can be applied to understand and map the distribution of rare, elusive and threatened species.

Key-words: automated species identification models, elusive, passive acoustic monitoring, portable recorders, rare, species distribution

Introduction

Imperfect detection of threatened species remains a challenge for wildlife conservation (MacKenzie *et al.* 2005, 2006). This is because many threatened species have small population sizes and occur in few sites (Manne & Pimm 2001) and they are often elusive, resulting in low number of observations regardless of their abundance and distribution (Chadès *et al.* 2008). Consequently, the population status of many threatened species is unknown, making it essential for conservation biologists to develop new approaches to improve detectability of threatened species.

One solution is passive acoustic monitoring (PAM), which has been successfully used to monitor threatened species. For instance, PAM devices were used to search for the presumably extinct Ivory-Billed Woodpecker (Fitzpatrick *et al.* 2005) and to monitor populations of threatened and rare species such as the Little Spotted Kiwi (Digby *et al.* 2013) and endangered Blue Whales (Miller *et al.* 2015). PAM devices have also been used to increase the number of observations of bats (Bader *et al.* 2015), nocturnal birds (Sberze, Cohn-Haft & Ferraz 2010) and marine cetaceans (Moore *et al.* 2006) that are particularly difficult to detect with traditional sampling techniques. Other advantages of passive acoustic techniques include monitoring populations 24 h a day, for years, at multiple locations simultaneously, and all acoustic information can be permanently stored (Brandes 2008; Celis-murillo, Deppe & Ward 2012). Furthermore, recordings can function as 'museum specimens' providing a permanent record and permitting future analyses. Nevertheless, PAM can require substantial expert effort to extract useful information from the recordings, and this is why many researchers have developed algorithms to automate species identification (Aide *et al.* 2013; Kalan *et al.* 2015). For example, automated species identification methods have been used to identify vocalizations of insects (Chesmore 2004), birds (Briggs *et al.* 2012), bats (Walters *et al.* 2012), primates (Kalan *et al.* 2015), whales (Murray, Mercado & Roitblat 1998) and amphibians (Ospina *et al.* 2013).

*Correspondence author: E-mail: marconi.campos.cerqueira@gmail.com

Along with technological advances to better sample animal populations in the wild, there has been a dramatic increase in development and application of occupancy models that take into account species detectability (Bailey, MacKenzie & Nichols 2014). These models provide useful tools to assess the population status of threatened species because they assign a probability that a species occurs in a sample unit taking into account that a species may be present in a site even though it was not detected. This is important, because ignoring imperfect detectability can underestimate occupancy estimates and bias inferences on the relationship between species occurrence and habitat variables (MacKenzie *et al*. 2006). Occupancy models have been successfully used to monitor population dynamics of the threatened Northern Spotted Owls *Strix occidentalis caurina* (Olson *et al*. 2005), to assess impact of cattle grazing on occupancy of the cryptic California Black Rail *Laterallus jamaicensis coturniculus* (Richmond, Tecklin & Beissinger 2012) and to identify and monitor Tiger *Panthera tigris* populations at the landscape level in India (Karanth *et al*. 2011).

Despite advances in automated species identifications and occupancy models, few studies have combine these techniques (Yates & Muzika 2006; Kalan *et al*. 2015), probably because automated species identification models often include high levels of false-positive detections (Waddle, Thigpen & Glorioso 2009; Zwart *et al*. 2014), which violates a major assumption of most occupancy models (MacKenzie *et al*. 2006). Alternatively, all recordings can be manually validated, but this would be extremely time-consuming. To resolve this problem, there are two potential solutions: (i) implement more complex occupancy models that take into account both false-negative and false-positive errors (Miller *et al*. 2011); or (ii) use automated species identification models to reduce the size of the data set, and then validate all positive identifications to eliminate any false-positive detections.

The objectives of this study were to (i) determine the spatial distribution of Elfin Woods Warbler *Setophaga angelae* a rare, elusive and threatened bird species across 60 sites along an elevation gradient in one of its few remaining populations; and (ii) compare three automated species identification/validation approaches to determine which approach provides occupancy estimates similar to the manual validation of all recordings. In this study, we show that the use of automated species-specific identification models can greatly reduce the amount of recordings that must be validated to develop unbiased and precise occupancy models.

Materials and methods

STUDY SITE AND SPECIES DESCRIPTION

The study was conducted in El Yunque National Forest (EYNF) in north-eastern Puerto Rico (Fig. 1). The EYNF is the largest protected area (115 km^2) of primary forest in Puerto Rico (Lugo 1994) and comprises a series of mountain chains rising to an elevation of 1074 m a.s.l. This elevation gradient has a strong effect on temperature, rain, humidity and the distribution of plants and animals (Garcia-martino *et al*. 1996; Wang *et al*. 2003; González *et al*. 2007; Gould *et al*. 2008; Willig *et al*. 2011; Brokaw *et al*. 2012). There are four main forest types along the elevational gradient in EYNF: Tabonuco forest which is dominated by *Dacryodes excelsa* and occurs between 150 and 600 m a.s.l., Palo Colorado forest which is dominated by *Cyrilla racemiflora*, and occurs between 600 and 950 m a.s.l, Elfin forest which is dominated by *Tabebuia rigida* and *Eugenia boriquensis* and occurs above 950 m a.s.l., and Sierra Palm forest, which is dominated by *Prestoea montana* and can occur anywhere along the elevational gradient. In addition to the four major forest types, EYNF has a considerable area in old secondary forest (>40 years) that occurs mostly at low elevations near the border of the reserve.

Setophaga angelae is a small passerine bird, endemic to the main island of Puerto Rico (Kepler & Parkes 1972). Currently, its distribution is restricted to two protected areas separated by 150 km: EYNF and the Maricao Commonwealth Forest (MCF). The estimated population size is 1800 mature individuals according to IUCN Red List (BirdLife International 2012). Besides having a small population size and a restricted geographical distribution, *S. angelae* is described as rare and cryptic, which could explain its late discovery (Kepler & Parkes 1972). At the time of its description, *S. angelae* was assumed to be restricted to high elevation areas within the Elfin forest (above 950 m a.s.l), although individuals could be found as low as 250 m a.s.l., and in a variety of habitats including Palo Colorado forest, *Podocarpus coriaceus* forest, secondary forest, coffee plantation and pasturelands (González 2008). Banding studies suggest that *S. angelae* is monogamous and territorial throughout the year (Delannoy-Juliá 2009). The territory size was estimated to be approximately one hectare per pair (Kepler & Parkes 1972). Vocalizations include the territorial song (common song), an alarm call and a duet song (https://arbimon.sieve-analytics.com/project/elevation).

SAMPLING DESIGN AND AUTONOMOUS RECORDINGS

Because elevation is a well-known proxy for habitat type, temperature and animal and plant communities (Brokaw *et al*. 2012; Kéry, Gardner & Monnerat 2010), we collected acoustic data in 60 sites in EYNF along three elevational transects (95–1074 m a.s.l; ~20 sampling sites per elevational transect) between March 27 and May 6, 2015. The elevational transects took advantage of roads and trails, but all recorders were placed more than 200 m from any road. Along each elevational transect, two recorders, separated by 200 m, were deployed at 100-m elevation interval (from 95 to 1074 m a.s.l). Recorders collected data at each site within a transect for approximately 1 week and were then moved to another elevation transect. The study occurred during the breeding season when song rate is highest (Arroyo-Vasquez 1992). Due to the small home range of *S. angelae* (~1 ha, Kepler & Parkes 1972), we believe that it is unlikely that birds from one territory would be recorded by more than one recorder.

Recorders consist of one LG smartphone enclosed in a waterproof case with an external connector linked to a Monoprice microphone. The ARBIMON Touch application (https://play.google.com/store/apps/details?id=touch.arbimon.com.arbimontouch&hl=en) was used to schedule recording events. Recorders were placed on trees at a height of 1·5 m and programmed to record 1 min of audio every 10 min for a total of 144 – 1-min recordings per day. We performed field tests in our study area and we have found that *S. angelae* vocalizations can be detected by our recorders up to ~50 m. Therefore, a site is defined here as a three-dimensional hemisphere space with a radius of approximately 50 m around the recorder.

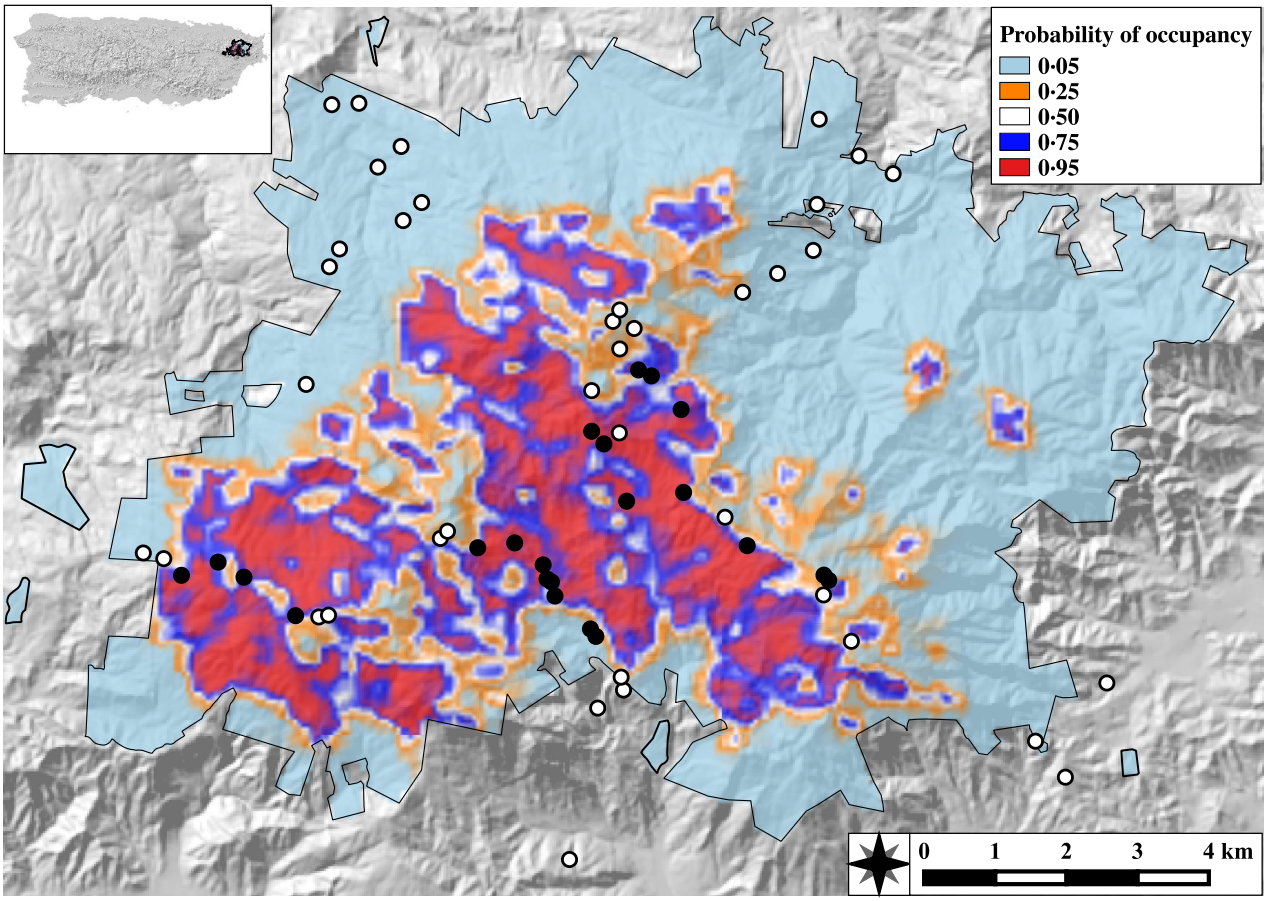

Fig. 1. Map of El Yunque National Forest and its location in NE Puerto Rico. The black circles represent sites where *Setophaga angelae* was detected and the white circles represent sites where the species was not detected. Different colors represent different probabilities of occupancy for *S. angelae* according the top occupancy model from the Full data set.

BIOACOUSTICS DATA PROCESSING AND MANAGEMENT

The spectrograms of all recordings (n = 38 255) were visually inspected, and if the species appear to be present, we listened to the recordings to make the final decision. This resulted in a detection/non-detection matrix that was then used to fit occupancy models that accounted for imperfect detectability (Fig. 2). The results of these analyses were used as the 'gold standard' for comparing results based on three different approaches that used a species identification model created in the ARBIMON analytical platform (https://arbimon.sieve-analytics.com). Below, we summarize the six steps used in creating a species identification model:

1. Create a template of the vocalization and validate a set of recordings: For the model, we used the territorial song because it is the most distinct and most common vocalization. Fifteen examples of the territorial song were selected to create the template, and 208 recordings were used for the validation data set (i.e. recording were the song was present or absent).

2. Create a correlation vector between the song template and the spectrogram. The song template was applied to each of the validated recordings. In this step, the template traverses each spectrogram and produces a vector of similarities for each recording (i.e. correlations between the template and sections of the spectrogram). The correlation was generated by the OpenCV function MatchTemplate (Bradski & Kaehler 2008).

3. Extract features of the vectors from the 208 validated recordings. In this step, 12 features of the correlation vector are extracted: mean, median, minimum, maximum, standard deviation, maximum–minimum, skewness, kurtosis, hyper-skewness, hyper-kurtosis, histogram and cumulative frequency histogram.

4. Create a RandomForest (RF) classifier: the features of the validated recordings (i.e. present or absent) are input into a RandomForest

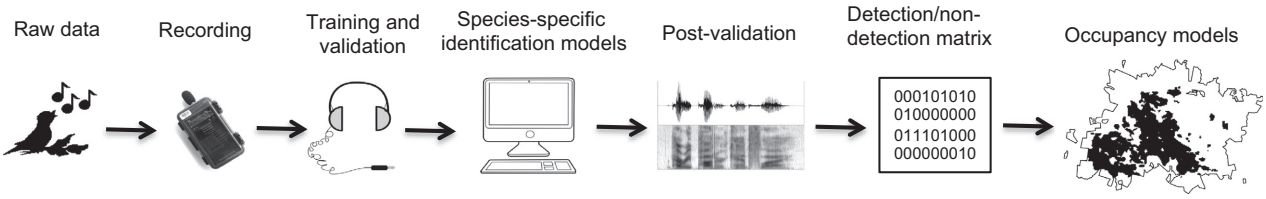

Fig. 2. Workflow of the bioacoustics data processing and analyses.

classifier (Breiman 2001). The goal was to train the RF model for a binary decision of presence or absence of the territorial song in a recording based on the feature vectors. A confusion matrix is provided (Table S1). The model was adjusted to reduce false positives.

5. Apply a Threshold approach: this is an alternative approach that is based on manually setting the maximum similarity correlation level of the vectors necessary to assign a recording as having a positive detection. A confusion matrix is provided (Table S1). The model was adjusted to reduce false positives.

6 Classify all recordings: the RF model and Threshold model were applied to all recordings. This resulted in a data set with a classification of presence or absence based on the RF model and Threshold model for each of the 38 255 recordings.

We then compared the results of the manual validation process with the results from the RF and Threshold approaches. This procedure resulted in four data sets: the manual validation, Threshold, RandomForest and Combined (Table 1). The Threshold, RandomForest and Combined data sets were constructed by manually verifying all the positive detections from the automated species identification models and converting any false-positive detections to true negatives. False-negative detections were assumed to be true-negative detections. We chose not to change false-negative detections because occupancy models can account for this type of error. The Combined data set only included recordings with positive detection in both the RandomForest and Threshold models. Although it is possible to confuse the vocalizations of the Bananaquit *Coereba flaveola* and Elfin Woods Warbler in the field, we are confident that we do not have any false positives in our data sets because the spectrogram analyses allowed us to visualize and compare the vocalizations, making it easy to distinguish the species.

The analyses were based on recordings between 05:00 and 19:00, but to simplify the detection matrix, we summarize detections in two-hour intervals. This simplification resulted in seven sampling occasions per day, where each sampling occasion included 12 recordings in each two-hour interval. Therefore, our most basic sample unit is defined here as one interval with 12 1-min recordings.

OCCUPANCY MODELLING

We used the detection/non-detection matrix generated after the validation of the classified data to fit single-season occupancy models using the package Unmarked in R (Fiske & Chandler 2011). The occupancy probability of each sampling site was estimated taking into account imperfect detection, following a standard maximum-likelihood hierarchical approach (MacKenzie *et al.* 2002). Our models include a sampling level describing the probability of detection conditioned on occupancy (p), and a biological level describing the probability (ψ) that a site is occupied. Both p and ψ are allowed to vary according habitats characteristics. Because both elevation and forest type are expected to influence *S. angelae* occurrence (Kepler & Parkes 1972; Anadón-Irizarry 2006; Arendt, Qian & Mineard 2013), we chose to include these variables in our occupancy models. We included three continuous and standardized variables representing the effect of elevation on both occupancy and detection parameters: 'Elevation', 'Elevation^2' and 'Elevation^3', which provides a first-, second- and third-order polynomial function of the elevation data, respectively (Kéry *et al.* 2010). Additionally, we included the effect of per cent cover of five forest types (Tabonuco forest, Secondary forest, Palo Colorado forest, Sierra Palm forest, Elfin forest and Riparian forest) and forest cover in the occupancy and detection parameters. The per cent cover of each forest type was estimated within a buffer with a radius of 100 m centred on the location of each recorder. Forest type classification was based on vegetation classification maps developed by USDA Forest Service (Gould *et al.* 2008). Lastly, we included a variable 'Hour', coded as 1–7 for each of the 7 2-h sampling periods. This variable was included in the detection parameter, because it is a good predictor of bird vocal activity (Catchpole & Slater 2003). We also included a second ('Hour^2')- and third-order ('Hour^3') polynomial function of the hour data.

The most parsimonious model was chosen by calculating the Akaike Information Criterion (AIC, Burnham & Anderson 2002). We chose a 2-step approach to select the most parsimonious model as proposed by MacKenzie (2006): first, we kept the occurrence part of the model constant (intercept-only) and tested the effect of each variable (e.g. elevation, forest type, hour) separately in the detection parameter. Having identified the most parsimonious model structure for the detection part of the model (Hour + Hour^2), we kept this constant and tested the effect of each variable separately in the occupancy parameter. Because the four major forest types in EYNF are associated with an elevation gradient (Gould *et al.* 2008), we did not include these two type of variables (e.g. elevation and forest type) in the same model. As a result, with the exception of the null model, all other models included at least three variables: one or two variables in the occupancy parameter (e.g. elevation or forest type), and always two variables in the detection parameter (Hour + Hour^2). The same model selection approach was applied to the four different data sets. We used the parametric bootstrap procedure of MacKenzie & Bailey (2004) for assessing goodness-of-fit of our best model. We found no indication of lack of fit for our best model ($P > 0{\cdot}05$).

To create a distribution map for the species in EYNF, we added a grid of 4032 – 3·1 ha hexagons polygons over a map of EYNF and extracted the per cent of vegetation cover of each forest type. We used the function 'predict' from the Unmarked package to estimate the probability of occupancy from each hexagon polygon. We used QGIS (QGIS Development Team 2015) to graph the expected probability of occupancy across EYNF.

Table 1. The four data sets used in the occupancy models

Data set	Recordings	Classification presence	Manually confirmed presence
Full	38 255	–	888
RandomForest	38 255	1603	194
Threshold	38 255	437	62
Combined	38 255	67	51

All 38 255 recordings were manually inspected for the Full data set. For the RandomForest and Threshold data sets, all recordings were classified using the species model and the recordings that were classified as present were manually inspected. The Combined data set only included recordings where both the RandomForest and Threshold models agreed, and these recordings were also manually inspected.

Results

AUTOMATED SPECIES-SPECIFIC IDENTIFICATION MODELS

The confusion matrix of the RandomForest (RF) model had 1% of false positives and 2% of false negatives, while the Threshold model (TH) had 0·4% of false positives and 39·4% of false negatives. These models were then applied to all 38 255 1-min recordings. The RF model classified 1603 recordings with positive detections, the TH model classified 437 recordings with positive detections, and the two models agreed in 67

recordings with positive detections. Following the manual inspection of all positive detections to eliminate any false positives, we were left with 194 recordings from 20 sites with the RF model, 67 recordings from 12 sites with the TH model and 51 recordings from 14 sites RF/TH combined approach (Table 1). False-positive detections were mainly associated with the vocalization of Bananaquit *Coereba flaveola*, the most common and widespread bird species in the study area (Wunderle & Arendt 2011).

OCCUPANCY MODELS

Contrary to our expectations, the per cent of Elfin forest cover was not a positive predictor of *S. angelae* occurrence in EYNF in any of the four data sets. The best-fitting occupancy model for each data set presented the same structure and included the effect of Palo Colorado forest cover in the occupancy parameter and the quadratic effect of hour in the detection parameter (Table S2). The best-fitting occupancy model for all data sets had high Akaike weights (AIC weight > 0·76), suggesting that the per cent of Palo Colorado forest cover is a good predictor of *S. angelae* distribution. Models that included a second-order polynomial function of elevation performed better than models that did not include this covariate (**ΔAIC** < 10) and were always the second-ranked model for all data sets (Table 2).

As the per cent of Palo Colorado forest cover increases, the probability of occupancy of *S. angelae* increases in all data sets ($\beta > 0{\cdot}04$, SE < 0·10). Nevertheless, the predicted relationship between per cent of Palo Colorado forest cover and probability of occupancy for *S. angelae* in EYNF varied across data sets (Fig. 3). The RandomForest and Full data sets had a similar relationship between Palo Colorado forest and species occurrence, and gave more precise occupancy probabilities (i.e. narrow SEs) than the other two data sets. Sites with 50% cover of Palo Colorado forest had 0·53 (SE ± 0·10) and 0·60 (SE ± 0·11) probabilities of occupancy in the top model from RandomForest and Full data sets, respectively.

Although the best-fitting model for all data sets included a quadratic effect of hour in the detection parameter, we found different relationships between time of day and detection probabilities across data sets (Fig. 4). In the Full data set, detection probabilities were higher early in the morning (5–8 am) and decrease during the day. In contrast, in the Threshold and Combined data sets detection probabilities were higher between 9 am and 1 pm, while detection probabilities from RandomForest data set were higher between 11 am and 4 pm.

Even though our best-fitting model did not include the effect of elevation on occupancy, we chose to visually represent this scenario using the second-ranked model for the Full data set, given the strong correlation between elevation and the distribution of Palo Colorado forest (Weaver & Gould 2013; Fig. S1). Not surprisingly, the second-ranked model shows a higher probability of occupancy at intermediate elevations (between 600 and 900 m a.s.l), which coincides with the Palo Colorado forest zone (Fig. 5).

Table 2. The first three ranked model according to AIC for each data set

Data set	#Parameters	AIC	ΔAIC	AIC weight
Threshold				
Ψ (palo colorado) p (hour^2)	5	399·98	0·00	0·91
Ψ (elevation^2) p (hour^2)	6	405·5	5·53	0·06
Ψ (elevation^3) p (hour^2)	7	406·77	6·79	0·03
Combined				
Ψ (palo colorado) p (hour^2)	5	394·4	0·00	0·76
Ψ (elevation^2) p (hour^2)	6	397·99	3·59	0·13
Ψ (elavtion^3) p (hour^2)	7	398·17	3·77	0·12
RandomForest				
Ψ (palo colorado) p (hour^2)	5	768·46	0·00	0·98
Ψ (elevation^2) p (hour^2)	6	776·94	8·48	0·01
Ψ (elevation^3) p (hour^2)	7	777·88	9·42	0·01
Full				
Ψ (palo colorado) p (hour^2)	5	1403·8	0·00	0·99
Ψ (elevation^2) p (hour^2)	6	1413·63	9·83	0·01
Ψ (elevation^3) p (hour^2)	7	1414·04	10·24	0·01

The best model for all data set included the effect of Palo Colorado cover on occupancy parameter (Ψ) and a quadratic effect of hour on detection parameter (p).

Discussion

SETOPHAGA ANGELAE DISTRIBUTION

Contrary to our expectations, Palo Colorado forest and not Elfin forest was the best predictor of *S. angelae* distribution in the EYNF. The most parsimonious model, for all data sets, included a positive effect of Palo Colorado forest on warbler occupancy. In contrast, there was no support for the model that included an effect of Elfin forest cover on warbler occupancy. Although in the species description suggests a strong association with Elfin forest (Kepler & Parkes 1972), our results corroborate more recent studies that documented a positive relationship between Palo Colorado forest and the distribution of *S. angelae* in EYNF (Anadón-Irizarry 2006; Arendt, Qian & Mineard 2013). The occupancy estimates from the Full data set suggest that *S. angelae* may occur in 17% of EYNF, which is much larger than previously thought (Fig. 1). The differences between historical and modern studies may represent a real habitat/elevational shift in response to habitat disturbance (e.g. climate change, hurricanes, habitat loss) or are due to sampling artefact related with imperfect detection (e.g. different detectability by different sampling techniques).

The occupancy models also showed that *S. angelae* has higher probability of occupancy (Ψ > 0·50) in sites between 600 and 900 m a.s.l, which coincides with the Palo Colorado forest zone. Similarly, in the only other known population in the Maricao Commonwealth Forest, the number of detections

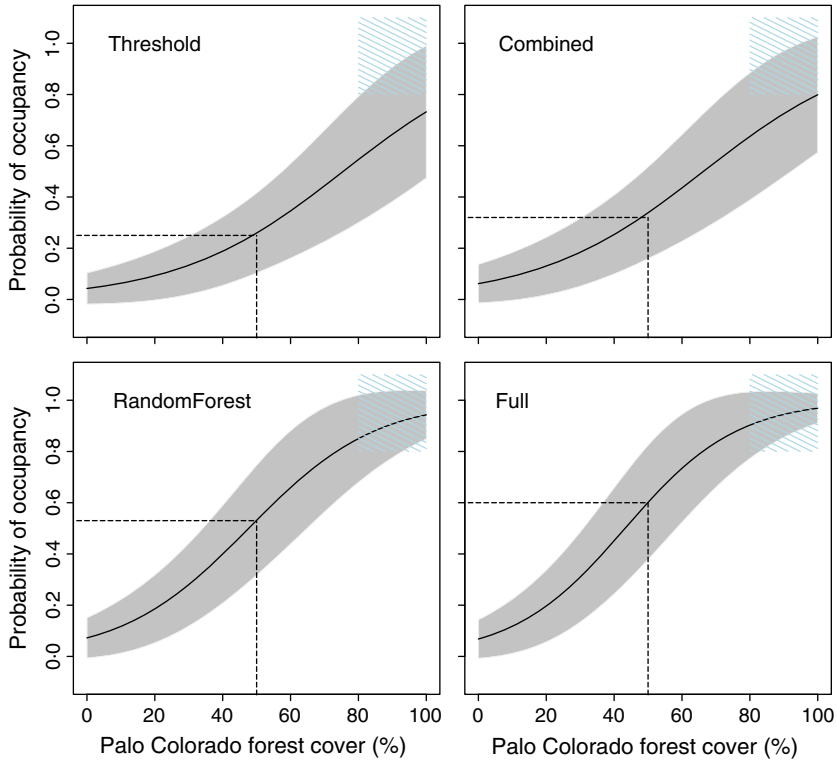

Fig. 3. Predicted relationship between per cent cover of Palo Colorado forest and the probability of occupancy for *Setophaga angelae* in EYNF using the most parsimonious model for each of the four data sets. The grey shaded area represents 95% confidence interval. The dashed blue area highlights the precision of occupancy estimates among the four data sets. Dashed lines indicate the probability of occupancy for 50% cover of Palo Colorado.

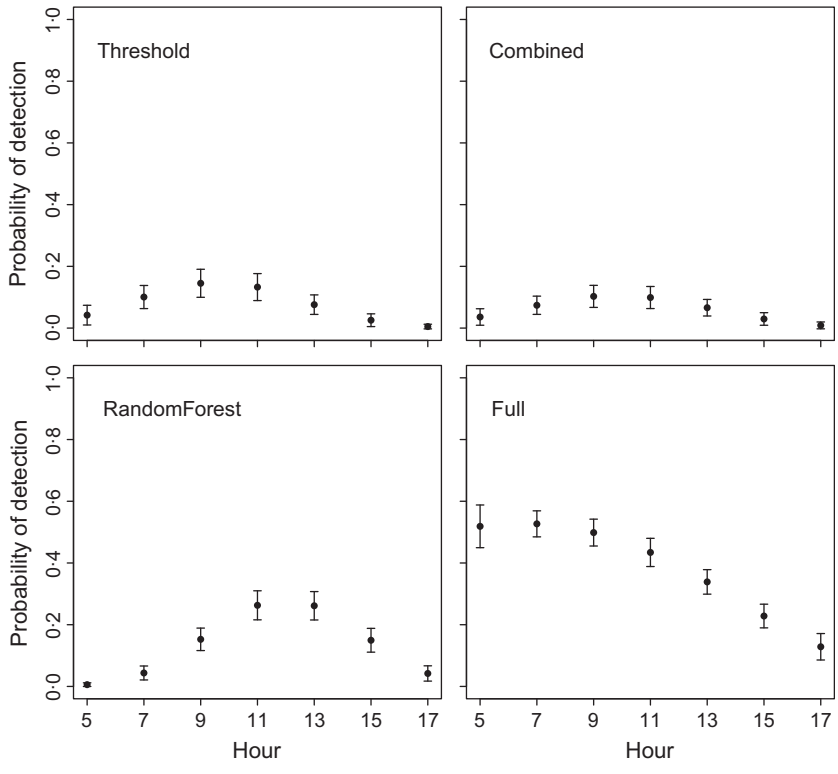

Fig. 4. Predicted relationship between hour and detection probability for *Setophaga angelae*. Variances are represented by 95% confidence intervals. The estimates came from the most parsimonious models.

and the density of *S. angelae* were highest between 600 and 880 m a.s.l, which corresponds with the *Podocarpus coriaceus* forest zone. The restricted elevational distribution of *S. angelae* in two distant and isolated areas may indicate that other variables related to elevation (e.g. temperature, humidity, prey availability, or forest structure) are the main drivers of the species distribution.

AUTOMATED SPECIES IDENTIFICATION AND OCCUPANCY MODELS

Regardless of the amount of recordings that were inspected after the classification process, the distribution of *S. angelae* was always described as a function of the per cent cover of Palo Colorado forest. Furthermore, the

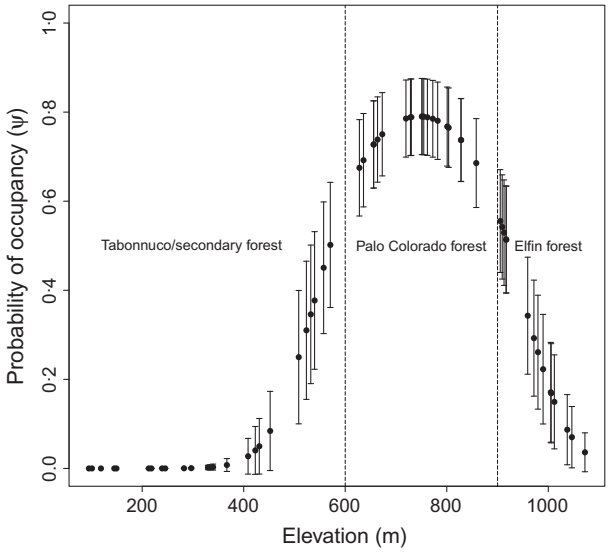

Fig. 5. Predicted relationship between elevation and probability of occupancy for *Setophaga angelae*. Variances are represented by 95% confidence intervals.

results show that automated species identification models can provide reliable data sets once false-positive detections were removed.

Although an increase in per cent of Palo Colorado forest cover always led to an increase in *S. angelae* occurrence, the occupancy response curves varied among the four data sets. Nevertheless, the RandomForest approach yielded occupancy estimates and uncertainty measurements (e.g. standard error) more similar with those obtained through the complete validation of all recordings, and this was achieved by only inspecting ~4% of the entire data set (1603 recordings). Given that we were able to manually inspect ~3000 recordings per day, the combination of the RandomForest model and post-classification inspection can save a substantial amount of time without sacrificing the quality of the results. The Threshold (437 recordings) and Combined (67 recordings) approaches could also reduce time, but the occupancy estimates and uncertainty measurements were not as precise as the RandomForest approach given that they were based on a much smaller data set, which included fewer sites.

While the results from the species-specific identification model can provide a good assessment for modelling species occurrence, it is still essential to conduct a post-validation and eliminate false-positive detections because if they were to remain in the data set the models would overestimate species occurrence and may make erroneous inferences of habitat usage. Although false-positive detections are still a ubiquitous characteristics of automated species-specific identification models, our post-validation approach with the removal of false-positive detections eliminated the need to use more complex models that adjust for misclassification (Miller *et al.* 2011).

The best-fitting model from all data sets included a quadratic effect of hour on the detection parameter, in which the Full data set indicates higher detection probabilities early in the morning, while the Threshold, RandomForest and Combined data sets had higher detection probabilities from 9 am to 4 pm. All recorders were identical, and there should be no systematic variability in their ability to capture the species' vocalization. Consequently, the detection probability from the full data means that birds are singing more and are more available for detection early in the morning. In contrast, the variation in detectability from the Threshold, RandomForest and Combined data sets is associated with the performance of the automated species identifications, which appear to have higher levels of detection during the middle of day when there is less vocal activity from other bird species.

Difficulties in detecting *S. angelae* are not new, and this is believed to have been the reason for the late discovery of this species (Kepler & Parkes 1972). Low vocal activity and vocal similarities with the co-occurring and highly abundant Bananaquit are probably the main factors leading to low acoustic detection of *S. angelae*. Nevertheless, the use of autonomous recorders greatly improved the detection probability of *S. angelae*. In 34 sampling days, we confirmed 888 true-positive detections, while a 17-year study in the same area was based on 1442 detections (Arendt, Qian & Mineard 2013). Another advantage of the autonomous recorders is the ability to sample many sites simultaneously. Furthermore, by increasing both the number of sampling sites and number of observation within a site, one can greatly improve the accuracy and precision of occupancy models (MacKenzie *et al.* 2002).

CONSERVATION AND MANAGEMENT IMPLICATIONS

There are three major conservation implications for the high occupancy of *S. angelae* in the Palo Colorado forest. First, Palo Colorado forest covers a much larger area (3441 ha) than the Elfin Woods forest (368 ha), which means there is almost 10× more potential habitat for this species. This suggests that this species may be more widely distributed than previous suspected. The second implication is that our occupancy estimates were based on sampling that was conducted during the species reproductive season, demonstrating the importance of Palo Colorado forest for the species reproductive ecology. Indeed, one of the three nests described for the species was constructed in a Palo Colorado tree in the Marico Commonwealth Forest (Rodriguez-Mojica 2004). The third implication, based on future scenarios of climate change for EYNF (Scatena 1998), is that the area of Palo Colorado forest will be reduced as Tabonuco forests expands its distribution up the mountain, which could negatively impact populations of *S. angelae*.

Although *S. angelae* may have a wider distribution than previously thought, our study shows that it still occurs in a limited number of habitats and it has a restricted elevational range. In addition, this species is only known from two localities. This combination has been described as an important predictor of extinctions risk in birds (White & Bennett 2015). We believe that these conditions, along with a documented population decline in EYNF (Arendt, Qian & Mineard 2013), provide enough evidence to include this species as vulnerable under the Endangered Species Act (ESA).

Previous surveys of *S. angelae* in others localities along the Cordillera Central in Puerto Rico have failed to detect the species (Anadón-Irizarry 2006). Since this species has low detectability, we highly recommend sampling in historical and new sites, particularly in areas of quality habitat between 600 and 900 m a.s.l using autonomous portable recorders combined with occupancy models. In this study, we demonstrate that the combination of acoustic monitoring and occupancy models can be a valuable tool to predict the distribution of a threatened species. Our approach shows that species-specific identification models combined with post-validation provided occupancy estimates that were comparable with the manually validation of all recordings. From a practical management perspective, this means much time and effort can be saved when using autonomous species identification models to predict species occurrence.

Acknowledgements

We are especially thankful to Paul Furumo, Serge Aucoin, Felipe Cano and Pedro Ríos (USDA Forest Service), Miguel A. Acevedo, Orlando Acevedo-Charry, Nora Alvarez-Berríos, Maria José Andrade, Joseph M. Wunderle, Wayne Arendt and Andres Hernandez-Serna. In addition, we thank Tim Lucas (UCL) and three anonymous reviewers for their comments. MCC was supported by the fellowship 'Science without borders' from the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) at Brazil (8933/13-8).

Data accessibility

The *S. angelae* data including recordings, sampling sites coordinates, elevation data, validation data, automated species identification model and classification data are permanently stored and available at https://arbimon.sieve-analytics.com/project/elevation/dashboard.

References

Aide, T.M., Corrada-Bravo, C., Campos-Cerqueira, M., Milan, C., Vega, G. & Alvarez, R. (2013) Real-time bioacoustics monitoring and automated species identification. *PeerJ*, **1**, e103.

Anadón-Irizarry, V. (2006) *Distribution, habitat occupancy and population density of the elfin-woods warbler (Dendroica Angelae) in Puerto Rico*. Master thesis, University of Puerto Rico, Puerto Rico.

Arendt, W.J., Qian, S.S. & Mineard, K.A. (2013) Population decline of the Elfin-woods Warbler *Setophaga angelae* in eastern Puerto Rico. *Bird Conservation International*, **23**, 136–146.

Arroyo-Vasquez, B. (1992) Observations of the breeding biology of the Elfin Woods Warbler. *The Wilson Bulletin*, **104**, 362–365.

Bader, E., Jung, K., Kalko, E.K., Page, R.A., Rodriguez, R. & Sattler, T. (2015) Mobility explains the response of aerial insectivorous bats to anthropogenic habitat change in the Neotropics. *Biological Conservation*, **186**, 97–106.

Bailey, L.L., MacKenzie, D.I. & Nichols, J.D. (2014) Advances and applications of occupancy models. *Methods in Ecology and Evolution*, **5**, 1269–1279.

BirdLife International (2012) *Dendroica angelae*. *The IUCN red list of Threatened Species*. URL http://dx.doi.org/10.2305/IUCN.UK.2012-1.RLTS.T22721749A39857971.en [accessed 20 November 2015]

Bradski, G. & Kaehler, A. (2008) *Learning OpenCV: Computer Vision With the OpenCV Library*. O'Reilly Media, Inc, Sebastopol, California, USA.

Brandes, T.S. (2008) Automated sound recording and analysis techniques for bird surveys and conservation. *Bird Conservation International*, **18**, S163–S173.

Breiman, L. (2001) Random forests. *Machine Learning*, **45**, 5–32.

Briggs, F., Lakshminarayana, B., Neal, L., Fern, X.Z., Raich, R., Hadley, S.J., Hadley, A.S. & Betts, M.G. (2012) Acoustic classification of multiple simultaneous bird species: a multi-instance multi-label approach. *The Journal of the Acoustical Society of America*, **131**, 4640–4650.

Brokaw, N., Crowl, T., Lugo, A., McDowell, W., Scatena, F., Waide, R. & Willig, M. (2012) *A Caribbean Forest Tapestry: The Multidimensional Nature of Disturbance and Response* (ed. N. Brokaw). Oxford University Press, Oxford.

Burnham, K.P. & Anderson, D.R. (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer-Verlag, New York, New York, USA.

Catchpole, C.K. & Slater, P.J. (2003) *Bird Song: Biological Themes and Variations*. Cambridge University Press, New York, New York, USA.

Celis-murillo, A., Deppe, J.L. & Ward, M.P. (2012) Effectiveness and utility of acoustic recordings for surveying tropical birds. *Journal of Field Ornithology*, **83**, 166–179.

Chadès, I., McDonald-Madden, E., McCarthy, M.A., Wintle, B., Linkie, M. & Possingham, H.P. (2008) When to stop managing or surveying cryptic threatened species. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 13936–13940.

Chesmore, D. (2004) Automated bioacustic identification of species. *Anais da Academia Brasileira de Ciências*, **76**, 436–440.

Delannoy-Juliá, C. (2009) Elfin-woods Warbler (*Setophaga angelae*), Neotropical Birds Online (T.S. Schulenberg, Editor). Cornell Lab of Ornithology, Ithaca. URL http://neotropical.birds.cornell.edu/portal/species/overview?p_p_spp=569996 [accessed 20 November 2015]

Digby, A., Towsey, M., Bell, B.D. & Teal, P.D. (2013) A practical comparison of manual and autonomous methods for acoustic monitoring. *Methods in Ecology and Evolution*, **4**, 675–683.

Fiske, I. & Chandler, R. (2011) Unmarked: an R package for fitting hierarchical models of wildlife occurrence and abundance. *Journal of Statistical Software*, **43**, 1–23.

Fitzpatrick, J.W., Lammertink, M., Luneau, M.D., Gallagher, T.W., Harrison, B.R., Sparling, G.M. *et al.* (2005) Ivory-billed Woodpecker (*Campephilus principalis*) persists in continental North America. *Science*, **308**, 1460–1462.

Garcia-martino, A.R., Warner, G.S., Scatena, F.R. & Civco, D.L. (1996) Rainfall, runoff and elevation relationships in the Luquillo Mountains of Puerto Rico. *Caribbean Journal of Science*, **32**, 413–424.

González, G.M. (2008) Distribución y abundancia de la reinita de bosque enano (*Dendroica angelae*) en el Bosque de Maricao y en áreas adyacentes. Master thesis, University of Puerto Rico, Mayagüez Campus, Mayagüez, Puerto Rico.

González, G., García, E., Cruz, V., Borges, S., Zalamea, M. & Rivera, M.M. (2007) Earthworm communities along an elevation gradient in Northeastern Puerto Rico. *European Journal of Soil Biology*, **43**, S24–S32.

Gould, W.A., Alarcon, C., Fevold, B., Jimenez, M.E., Martinuzzi, S., Potts, G., Quiñones, M., Solórzano, M. & Ventosa, E. (2008) *The Puerto Rico Gap Analysis Project Volume 1: Land Cover, Vertebrate Species Distributions, and Land Stewardship*. US Forest Service IITF-GTR-39, Río Piedras, Puerto Rico.

Kalan, A.K., Wagner, O.J.J., Heinicke, S., Mundry, R., Boesch, C. & Kuehl, H.S. (2015) Towards the automated detection of primates using passive acoustic monitoring. *Ecological Indicators*, **54**, 217–226.

Karanth, K.U., Gopalaswamy, A.M., Kumar, N.S., Vaidyanathan, S., Nichols, J.D. & MacKenzie, D.I. (2011) Monitoring carnivore populations at the landscape scale: occupancy modelling of tigers from sign surveys. *Journal of Applied Ecology*, **48**, 1048–1056.

Kepler, C.B. & Parkes, K.C. (1972) A new species of warbler (Parulidae) from Puerto Rico. *The Auk*, **89**, 1–18.

Kéry, M., Gardner, B. & Monnerat, C. (2010) Predicting species distributions from checklist data using site-occupancy models. *Journal of Biogeography*, **37**, 1851–1862.

Lugo, A.E. (1994) Preservation of primary forests in the Luquillo Mountains, Puerto Rico. *Conservation Biology*, **8**, 1122–1131.

MacKenzie, D.I. (2006) Modeling the probability of resource use: the effect of, and dealing with, detecting a species imperfectly. *The Journal of Wildlife Management*, **70**, 367–374.

MacKenzie, D.I. & Bailey, L.L. (2004) Assessing the fit of site-occupancy models. *Journal of Agricultural, Biological, and Environmental Statistics*, **9**, 300–318.

MacKenzie, D.I., Nichols, J.D., Lachman, G.B., Droege, S., Royle, J.A. & Langtimm, C.A. (2002) Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, **83**, 2248–2255.

MacKenzie, D.I., Nichols, J.D., Sutton, N., Kawanishi, K. & Bailey, L.L. (2005) Improving inferences in populations studies of rare species that are detected imperfectly. *Ecology*, **86**, 1101–1113.

MacKenzie, D.I., Nichols, J.D., Royle, J.A., Pollock, K.H., Bailey, L.L. & Hines, J.E. (2006) *Occupancy Estimation and Modeling: Inferring Patterns and Dynamics of Species Occurrence*. Academic Press, New York, New York, USA.

Manne, L.L. & Pimm, S.L. (2001) Beyond eight forms of rarity: which species are threatened and which will be next? *Animal Conservation*, **4**, 221–229.

Miller, D.A., Nichols, J.D., McClintock, B.T., Grant, E.H.C., Bailey, L.L. & Weir, L.A. (2011) Improving occupancy estimation when two types of

observational error occur: non-detection and species misidentification. *Ecology*, **92**, 1422–1428.
Miller, B.S., Barlow, J., Calderan, S., Collins, K., Leaper, R., Olson, P. *et al.* (2015) Validating the reliability of passive acoustic localisation: a novel method for encountering rare and remote Antarctic blue whales. *Endangred Species Research*, **26**, 257–269.
Moore, S.E., Stafford, K.M., Mellinger, D.K. & Hilderbrand, J.A. (2006) Listening for large whales in the offshore waters of Alaska. *BioScience*, **56**, 49–55.
Murray, S.O., Mercado, E. & Roitblat, H.L. (1998) The neural network classification of false killer whale (*Pseudorca crassidens*) vocalizations. *The Journal of the Acoustical Society of America*, **104**, 3626–3633.
Olson, G.S., Anthony, R.G., Forsman, E.D., Ackers, S.H., Loshl, P.J., Dugger, K.M., Glenn, E.M. & Ripple, W.J. (2005) Modeling of site occupancy dynamics for northern spotted owls, with emphasis on the effects of barred owls. *Journal of Wildlife Management*, **69**, 918–932.
Ospina, O.E., Villanueva-Rivera, L.J., Corrada-Bravo, C.J. & Aide, T.M. (2013) Variable response of anuran calling activity to daily precipitation and temperature: implications for climate change. *Ecosphere*, **4**, 1–12.
QGIS Development Team (2015) *QGIS Geographic Information System*. Open Source Geospatial Foundation Project. URL http://qgis.osgeo.org [accessed 18 November 2015]
Richmond, O.M.W., Tecklin, J. & Beissinger, S.R. (2012) Impact of cattle grazing on the occupancy of a cryptic, threatened rail. *Ecological Applications*, **22**, 1655–1664.
Rodriguez-Mojica, R. (2004) First report of cavity-nesting in elfin-woods warbler *Dendroica angelae* at Maricao State Forest, Puerto Rico. *Cotinga*, **22**, 21–23.
Sberze, M., Cohn-Haft, M. & Ferraz, G. (2010) Old growth and secondary forest site occupancy by nocturnal birds in a neotropical landscape. *Animal Conservation*, **13**, 3–11.
Scatena, F.N. (1998) An assessment of climate change in the Luquillo Mountains of Puerto Rico. *Third International Symposium on Water Resources, San Juan, Puerto Rico*. pp. 193–198. American Water Resources Association, Washington, DC, USA.
Waddle, J., Thigpen, T. & Glorioso, B. (2009) Efficacy of automatic vocalization recognition software for anuran monitoring. *Herpetological Conservation and Biology*, **4**, 384–388.
Walters, C.L., Freeman, R., Collen, A., Dietz, C., Brock, F.M., Jones, G. *et al.* (2012) A continental-scale tool for acoustic identification of European bats. *Journal of Applied Ecology*, **49**, 1064–1074.
Wang, H., Hall, C.A., Scatena, F.N., Fetcher, N. & Wu, W. (2003) Modeling the spatial and temporal variability in climate and primary productivity across the Luquillo Mountains, Puerto Rico. *Forest Ecology and Management*, **179**, 69–94.
Weaver, P. & Gould, W. (2013) Forest vegetation along environmental gradients in northeastern Puerto Rico. *Ecological Bulletins*, **54**, 43–65.
White, R.L. & Bennett, P.M. (2015) Elevational distribution and extinction risk in birds. *PLoS ONE*, **10**, e0121849.
Willig, M.R., Presley, S.J., Bloch, C.P., Castro-Arellano, I., Cisneros, L.M., Higgins, C.L. *et al.* (2011) A complex metacommunity structure for gastropods along an elevational gradient. *Oikos*, **120**, 480–488.
Wunderle, J.M. & Arendt, W.J. (2011) Avian studies and research opportunities in the Luquillo Experimental Forest: a tropical rain forest in Puerto Rico. *Forest Ecology and Management*, **262**, 33–48.
Yates, M. & Muzika, R. (2006) Effect of forest structure and fragmentation on site occupancy of bat species in Missouri Ozark forests. *Journal of Wildlife Management*, **70**, 1238–1248.
Zwart, M.C., Baker, A., McGowan, P.J. & Whittingham, M.J. (2014) The use of automated bioacoustic recorders to replace human wildlife surveys: an example using nightjars. *PLoS ONE*, **9**, e102770.

Received 17 March 2016; accepted 25 May 2016
Handling Editor: Kate Jones

Supporting Information

Additional Supporting Information may be found online in the supporting information tab for this article:

Fig. S1. Relationship between Percent of Palo Colorado cover and elevation in EYNF.

Table S1. Confusion matrix of the species-specific identification model.

Table S2. Model selection results for *S. angelae* occupancy analyses using the four datasets.