Ecology, 91(8), 2010, pp. 2446–2454

Unmodeled observation error induces bias when inferring patterns and dynamics of species occurrence via aural detections

Brett T. McClintock,[1,5] Larissa L. Bailey,[2] Kenneth H. Pollock,[3] and Theodore R. Simons[4]

[1]*U.S. Geological Survey, Patuxent Wildlife Research Center, 12100 Beech Forest Road, Laurel, Maryland 20708 USA*
[2]*Department of Fish, Wildlife, and Conservation Biology, Colorado State University, Fort Collins, Colorado 80523 USA*
[3]*Biology, Biomathematics, and Statistics, Campus Box 7617, North Carolina State University, Raleigh, North Carolina 27695 USA*
[4]*U.S. Geological Survey North Carolina Cooperative Fish and Wildlife Research Unit, Department of Biology, Campus Box 7617, North Carolina State University, Raleigh, North Carolina 27695 USA*

Abstract. The recent surge in the development and application of species occurrence models has been associated with an acknowledgment among ecologists that species are detected imperfectly due to observation error. Standard models now allow unbiased estimation of occupancy probability when false negative detections occur, but this is conditional on no false positive detections and sufficient incorporation of explanatory variables for the false negative detection process. These assumptions are likely reasonable in many circumstances, but there is mounting evidence that false positive errors and detection probability heterogeneity may be much more prevalent in studies relying on auditory cues for species detection (e.g., songbird or calling amphibian surveys). We used field survey data from a simulated calling anuran system of known occupancy state to investigate the biases induced by these errors in dynamic models of species occurrence. Despite the participation of expert observers in simplified field conditions, both false positive errors and site detection probability heterogeneity were extensive for most species in the survey. We found that even low levels of false positive errors, constituting as little as 1% of all detections, can cause severe overestimation of site occupancy, colonization, and local extinction probabilities. Further, unmodeled detection probability heterogeneity induced substantial underestimation of occupancy and overestimation of colonization and local extinction probabilities. Completely spurious relationships between species occurrence and explanatory variables were also found. Such misleading inferences would likely have deleterious implications for conservation and management programs. We contend that all forms of observation error, including false positive errors and heterogeneous detection probabilities, must be incorporated into the estimation framework to facilitate reliable inferences about occupancy and its associated vital rate parameters.

Key words: auditory detection; colonization; detection probability; false negative; false positive; imperfect detection; local extinction; measurement error; monitoring; observation error; site occupancy; species occurrence.

Introduction

The use of site occupancy models for inferring patterns and dynamics of species occurrence has surged in recent years (e.g., Wintle et al. 2005, Mazerolle et al. 2007). Many of the recently developed occupancy probability models acknowledge that species are detected imperfectly as a result of observation error. One form of observation error, arising when a species is present but fails to be detected (i.e., false negative error), has been a primary focus of these models (e.g., MacKenzie et al. 2002, 2003). This focus is justifiable because such errors are generally unavoidable, regardless of sampling design efforts. Receiving considerably less attention is another form of observation error that arises when a species is incorrectly detected as present when it is in fact absent (i.e., false positive error), and it remains standard practice to assume false positive errors are of little concern (but see Royle and Link 2006).

The models of MacKenzie et al. (2002, 2003) facilitate unbiased estimation of occupancy probability conditional on there being no false positives and on the sufficient incorporation of explanatory variables for the false negative detection process. To help satisfy the latter condition, variables related to environmental conditions or observer abilities can be identified a priori and measured. However, other factors related to detection that are more difficult to assess, such as site-variable abundance (Royle and Nichols 2003), can result in detection probability heterogeneity between sites that induces bias in standard occupancy estimators. When detections rely on auditory cues (e.g., avian or amphibian vocalizations), another form of heterogeneity can arise due to site-variable calling distances (i.e., the distance

Manuscript received 16 July 2009; revised 2 November 2009; accepted 9 November 2009. Corresponding Editor: J. R. Sauer.
[5] E-mail: brett.mcclintock@gmail.com

between species vocalizations and the observer). Little information is available about the effects of site-variable calling distance on these estimators, but it has been demonstrated in simulated field study experiments that songbird (Simons et al. 2007) and calling anuran (McClintock et al. 2010) observation error increases with distance. Simons et al. (2009) also found substantial measurement error when observers attempted to estimate the distance to a sound source in the absence of visual cues. These findings suggest that it may be difficult to properly account for the potential effects of distance when solely relying on auditory cues for species detection. Failure to properly account for heterogeneity in site detection probabilities, such as that which can be induced by distance, will generally result in underestimation of species occurrence (Royle and Nichols 2003, Royle 2006).

Assuming the false negative detection process can be adequately explained, the standard occupancy modeling assumption of no false positive detections is likely reasonable in many cases. For example, one might contend that studies relying on the physical capture of individual animals for species identification would have negligible false positive error rates. However, studies relying on visual or auditory detections may be more prone to false positive errors because of the additional uncertainty these sampling methods inherently possess. There is mounting evidence that aural detections, routinely relied upon in avian (e.g., Sauer et al. 2003) or amphibian (e.g., Weir and Mossman 2005) monitoring programs, may be particularly prone to false positive errors (Genet and Sargent 2003, Lotz and Allen 2007, Simons et al. 2007; McClintock et al. 2010). Simons et al. (2007) reported an overall 4% misidentification rate when any of six species in their simulated songbird point count surveys was detected, with individual species misidentification rates ranging from 0% for the Hooded Warbler (*Wilsonia citrina*) to 10% for the Northern Parula (*Parula americana*). Using a similar system to simulate calling anuran field surveys, McClintock et al. (2010) found that calls for five species were misidentified as another species in 5% of all detections, ranging from 0% for the spring peeper (*Pseudacris crucifer*) to 11% for the wood frog (*Rana sylvatica*).

Relative to point count surveys of abundance, Nichols et al. (2009) and Simons et al. (2009) suggested that simplified monitoring protocols, such as occupancy approaches, may yield better results in the face of observation error. However, little remains understood about the potential effects of unmodeled observation error on standard estimators of occupancy and its associated vital rate parameters (e.g., colonization and local extinction probabilities). Lotz and Allen (2007) identified some potential problems with standard occupancy models under false positive errors, but no clear or predictable trends were apparent using their methodology because the true occupancy states and identities of the calling anuran species in their study were unknown. Royle and Link (2006) demonstrated via computer simulation that standard occupancy estimators were substantially biased with false positive error probabilities $\geq$5%, resulting in underestimation of detection probability and overestimation of site occupancy. These studies suggest that the false positive error rates observed by Simons et al. (2009) and McClintock et al. (2010) would likely result in unreliable inferences about patterns of occurrence. To our knowledge, there has yet to be an examination of the potential effects these errors may also have on inferences about changes in species occurrence over time.

Simulated field surveys provide a unique opportunity to investigate unmodeled observation error because the true species identities and occupancy states can be known and controlled. We conducted a series of field experiments to assess the potential problems induced by false positive errors and site detection probability heterogeneity on standard dynamic models of site occupancy, colonization, and local extinction probabilities. After describing how we utilized a simulated calling anuran system to produce data suitable for the dynamic occupancy model of MacKenzie et al. (2003), we evaluate the model's performance relative to the true occupancy states. We conclude by discussing the implications of our findings to facilitate more reliable inferences by properly accounting for all forms of observation error when using auditory detections to assess species occurrence and dynamics over time.

Methods

To better understand factors affecting the auditory detection process, we simulated simple anuran call surveys in a series of field experiments in North Carolina. The simulation system was a modification of a bird song system developed by Simons et al. (2007), where user-specified series of MP3 sound clips were transmitted from a laptop computer and broadcast from an arrangement of remote speakers. Speakers were placed along a straight line at 5-m intervals between 10 m and 55 m from observers stationed at a survey point. Individual anuran species call clips, lasting 20 seconds, were broadcast one at a time from each of the 10 speakers. Five species were selected based on their expected relative detectabilities, including the upland chorus frog (*Pseudacris feriarum*), spring peeper (*Pseudacris crucifer*), pickerel frog (*Rana palustris*), southern leopard frog (*Rana sphenocephala*), and wood frog (*Rana sylvatica*). To simulate a period with no species calling, a 20-second call clip of silence (NONE) was also produced and broadcast from the speakers. Three background noise treatments in the form of competing species were applied by broadcasting additional clips from another speaker placed 35m from the observers. These competing species treatments were a continuous spring peeper chorus, a single southern leopard frog calling continuously, and a control treatment containing no competing species vocalizations. With the exception of *P. crucifer*, each call clip (including NONE) was broadcast six different times from each distance for each treatment. The *P. crucifer* call clip was broadcast six times from each distance for the control treatment only. The six replications at each distance were randomly ordered within three distinct sessions, where each session consisted of two of the six replications for all three treatments. Five expert observers, including three North American Amphibian Monitoring Program (NAAMP; Weir and Mossman 2005) participants, were recruited to participate in the experiment and record any calls detected and identified during each of the 20-second call clips. The observers were provided a list of eleven species for potential detection during the experiments, and all of these species were common to the Piedmont region of North Carolina. The additional "phantom" species that never called included the northern cricket frog (*Acris crepitans*), American toad (*Bufo americanus*), Fowler's toad (*Bufo fowleri*), gray treefrog (*Hyla chrysoscelis* or *Hyla versicolor*), bullfrog (*Rana catesbeiana*), and green frog (*Rana clamitans*). Four of the observers participated in the first two sessions, and two of the observers completed the southern leopard frog treatment of the second session and the entire third session. There were therefore two or four replications for each call clip at each distance, totaling $n = 3000$ recorded responses to 960 species call clips played. Empirical results may be found in McClintock, Bailey, Simons, and Pollock (*unpublished manuscript*), but it was generally found that species, distance, time (session), ambient noise, observer effects, and interactions of these were important factors explaining the false negative detection process. Distance and observer effects were the best overall predictors of false positive errors, but ambient noise and competing species treatments also affected these error rates for some species. All of the phantom species except *B. fowleri* were falsely detected at least once, and these errors were most common for *A. crepitans* and *B. americanus*. McClintock et al. (2010) reported only on the empirical results of these experiments, in terms of the false negative and false positive auditory detection process, for specific anuran species of interest. Here, we examine the potential biases such errors can induce when inferring patterns and dynamics of species occurrence.

To investigate the effects of false positive errors and site-variable calling distances on inferences about occupancy and its changes over time, we produced data suitable for the multi-season, single species dynamic site occupancy model of MacKenzie et al. (2003) based on the replication within this experimental design. We first defined a "season" to be a session as described above. A "site" was defined by species, distance, and treatment category, with two clips played (i.e., two sites) for each species at each distance and treatment. A "visit" within each season was defined as an observer replicate, such that there were four visits in the first season, two or four visits in the second season (depending on treatment), and two visits in the third. The first and second call clips at each distance and treatment combination were then combined across all three sessions (seasons) for each species. The detection histories for each site were then modified accordingly for each of seven species (*A. crepitans*, *B. americanus*, *P. crucifer*, *P. feriarum*, *R. palustris*, *R. sphenocephala*, and *R. sylvatica*) with sufficient numbers of detections to estimate species-specific site occupancy (ψ), colonization (γ), and local extinction (ε) probabilities. For example, a site labeled "1_RASY_10m_CONTROL" indicated the first replicate of an *R. sylvatica* call at 10m with no competing species treatment. This site would have a detection history of 1010 1111 01 for the *R. sylvatica* analysis if this species was heard and correctly identified by the first and third observers during the first session, all observers during the second session, and only the second observer during the third session. Supposing the second observer falsely identified the call as being from *R. palustris* during the first season, the detection history for this site would be modified to 0100 0000 00 for the *R. palustris* analysis. If this were the only false positive occurring for this site, then the detection history for all other species analyses would be 0000 0000 00.

This approach created 320 sites, with 120 control sites (no competing species calls) and 100 sites for both the spring peeper and southern leopard frog competing species treatments. There were 20 additional sites for the control treatment because the *P. crucifer* call clip was only played during this treatment. For each of the seven species, only those sites where the species call was played were designated as occupied. By nature of the design, this meant true ψ was constant (i.e., $\gamma = \varepsilon = 0$) across seasons for each species. For example, *R. palustris*, *R. sphenocephala*, and *R. sylvatica* calls were played twice at each of the 10 distances for each of the three treatments, yielding for each species an overall $\psi = 60/320$, with $\psi = 20/120$ for the control treatment sites and $\psi = 20/100$ for the spring peeper chorus and southern leopard frog treatment sites. For *P. crucifer*, only the control and southern leopard frog treatment sites were considered, with overall $\psi = 20/220$, $\psi = 20/120$ for the control treatment sites, and $\psi = 0/100$ for the southern leopard frog treatment sites. Similarly, only the control and spring peeper treatment sites were considered for *R. sphenocephala*, with overall $\psi = 40/220$, $\psi = 20/120$ for the control sites, and $\psi = 20/100$ for the spring peeper chorus sites. As *A. crepitans* and *B. americanus* call clips never actually played, $\psi = 0/320$ for these species. We hypothesized that false positive detections would result in underestimation of detection probabilities and overestimation of ψ, γ, and ε. We also hypothesized that failing to account for distance effects would generally result in underestimation of ψ and overestimation of γ and ε.

We performed separate dynamic occupancy analyses for each of the seven species in Program MARK (White and Burnham 1999). To facilitate numerical convergence, we implemented the ψ_t, γ_t, and ε_t parameterization, where site occupancy probability for the first season (ψ_1) is estimated directly and subsequent occupancy estimates are derived using the recursive equation $\psi_{t+1} = \psi_t(1 - \varepsilon_t) + (1 - \psi_t)\gamma_t$ for $t = 1, 2$ (MacKenzie et al. 2003). Each analysis followed a standardized procedure where a general season- and distance-dependent structure was held constant for ψ_1, γ, and ε while first investigating different parameterizations for detection probability (p). A season-dependent structure was used for ψ_1, γ, and ε when distance effect parameters were inestimable. If season effect parameters were inestimable (e.g., due to a lack of detections for a given species), then intercept-only structures were used for these parameters. Possible covariates on p included effects for season, observer, treatment, true distance category (10–55 m at 5-m increments), a logit-linear trend on distance, ambient noise (dB), and two-way interactions of these variables. We also examined an index score of observer ability obtained through an online exam completed by each of the observers (McClintock et al. 2010) as a potentially more parsimonious alternative to separating observer effects. Akaike's information criterion adjusted for small sample sizes (AIC_c, Burnham and Anderson 2002) was used as evidence of the relative support for the different models. Once the best supported structure for p was identified for each species, we kept this structure fixed to examine different parameterizations for ψ_1, γ, and ε. Covariates examined included season, treatment, true distance category, a logit-linear trend on distance, and a logit-quadratic trend on distance. We investigated distance effects on ψ_1, γ, and ε to simulate the case where a site covariate is hypothesized to influence both detection and true occupancy parameters (e.g., patch size). Models assuming no changes in occupancy over time (i.e., $\gamma_t = \varepsilon_t = 0$) were also investigated. Finally, we combined any additional structures with high support for p with the best supported structures for ψ_1, γ, and ε. To investigate the consequences of unmodeled site detection probability heterogeneity in the presence of false positive errors, we then repeated this process for each of the seven species with all effects related to distance removed.

To help dissect the separate roles false positive errors and heterogeneity may play in biasing occupancy estimators, we performed additional analyses for the five species that were actually present (*P. crucifer*, *P. feriarum*, *R. palustris*, *R. sphenocephala*, and *R. sylvatica*) using corrected detection histories with all false positives removed. We followed the same procedure as above, comparing results with and without distance effects. Here, we primarily focus on the estimation of species occurrence dynamics (ψ_1, γ, and ε), but readers interested in how the false negative and false positive detection processes are affected by distance and other factors are referred to McClintock et al. (2010) and Appendix A. We note that because our objective was to evaluate the potential consequences of false positives and heterogeneity on standard estimators of occupancy (and its dynamics), the misclassification model proposed by Royle and Link (2006) was dismissed on a priori grounds (see *Discussion* and Appendix B).

Results

The raw numbers of false positive detections for *A. crepitans*, *B. americanus*, *P. crucifer*, *P. feriarum*, *R. palustris*, *R. sphenocephala*, and *R. sylvatica* were 19, 5, 3, 9, 15, 37, and 6, respectively. These errors constituted 100%, 100%, 1%, 2%, 3%, 11%, and 2% of all respective positive detections for these species. For *A. crepitans*, *B. americanus*, and *R. palustris*, these errors tended to occur more frequently at greater distances. For species with higher detection probabilities, these errors alone were sufficient to result in naïve estimates of site occupancy (i.e., $p = 1$) that were larger than ψ (see Appendix C). Due to the variable detectability and false positive error rates for species, the best AIC_c supported models tended to differ greatly between species and analyses (Table 1). However, an examination of the season-dependent estimates for ψ_t, γ_t, and ε_t using the species- and analysis-specific minimum-AIC_c model structure for p makes comparisons of the general biases across species and analyses more tractable (Table 2). We summarize our most important findings below, but interested readers are referred to Appendix A for complete results by species.

For the analyses including false positive errors, spurious effects on ψ_1, γ, and ε due to season, competing species treatment, or distance were found for all species except *P. crucifer* (Table 1). Spurious effects were found between seasons for *B. americanus*, *R. palustris*, *R. sylvatica*, and *R. sphenocephala* and between treatments for *P. feriarum* and *R. palustris*. Effects for distance were supported for *A. crepitans*, *R. palustris*, and *R. sphenocephala*. Although the magnitude of these effects varied by species, false positives generally induced positive biases in ψ_1, γ, and ε. Unmodeled distance effects were associated with negative biases in ψ and positive biases in γ and ε (Table 2).

When both false positives and distance effects were included in the analysis, complicated patterns and dynamics for species occurrence were supported. For *A. crepitans* and *R. sphenocephala*, their respective minimum-AIC_c models predicted different site occupancy probabilities for each season, generally decreasing with calling distance after the first season regardless of treatment (Fig. 1). All estimates for *A. crepitans* exhibited positive biases, most notably for ψ_t in the second and third seasons. For *R. sphenocephala*, occupancy estimates at the shortest distances increased with each successive season, particularly between the second and third seasons. At further distances, estimates declined between the first and second seasons and increased less dramatically between the second and third seasons. Positive biases in occupancy were demonstrated for all distances during the first and third seasons, but in the second season site occupancy was

TABLE 1. Akaike's information criterion adjusted for small sample sizes (AIC_c) supported models and AIC_c model weights (w_i) of site occupancy (ψ_1), colonization (γ), and local extinction (ε) probabilities for seven species under four analysis scenarios with and without the incorporation of distance effects on detection probability (p): false positive errors without distance effects (FP), false positive errors with distance effects (FPD), no false positive errors without distance effects (NF), and no false positive errors with distance effects (NFD).

	Model			
Species and analysis	ψ_1	γ	ε	w_i
Acris crepitans				
FP	I	I	I	0.36
FPD	I	L^-	L^-	0.22
Bufo americanus†				
FP	I	S	I	0.65
FPD	I	S	I	0.51
Pseudacris crucifer				
FP	I	I	I	0.17
FPD	I	I	I	0.15
NF	SL^-	(NE)	(NE)	0.88
NFD	SL^-	(NE)	(NE)	0.88
Pseudacris feriarum				
FP	I	SL^-	I	0.34
FPD	I	SL^-	I	0.32
NF	I	(NE)	(NE)	0.53
NFD	I	(NE)	(NE)	0.60
Rana palustris				
FP	CO^-	S, SL^-	I	0.23
FPD	CO^-	S, SL^-, L^+	L^+	0.49
NF	I	I	I	0.19
NFD	I	(NE)	(NE)	0.52
Rana sphenocephala				
FP	I	S	I	0.40
FPD	I	S	L^+	0.30
NF	I	S	S	0.20
NFD	I	(NE)	(NE)	0.62
Rana sylvatica				
FP	I	I	I	0.50
FPD	I	I	S	0.40
NF	I	I	SP^-	0.17
NFD	I	(NE)	(NE)	0.66

Notes: Covariates examined included effects for sites receiving the control (CO), spring peeper chorus (SP), or southern leopard frog (SL) competing species treatments and intercept only (I). Distance linear trend (L) effects were investigated in the FPD and NFD analyses. Season effects (S) and no effects (indicated by NE in parentheses) were also examined for γ and ε. Superscripts indicate whether estimates for selected effects were positive (+) or negative (–). All effects on ψ_1, γ, and ε are spurious except a negative southern leopard frog treatment effect on ψ_1 for *P. crucifer* or an intercept-only effect on ψ_1 for all other species.

† Deduced from the alternative ψ_t and ε_t parameterization with derived γ_t (MacKenzie et al. 2003) because it showed improved numerical convergence.

overestimated at shorter distances and underestimated at further distances. For *R. palustris*, the minimum-AIC_c model predicted different occupancy probabilities for each season, with increases or decreases by site distance depending on the treatment (Fig. 2). Only with false positives removed and distance effects included were the correct models with constant occupancy (i.e., $\gamma = \varepsilon = 0$) selected by AIC_c for all species. When false positive errors were removed and distance effects excluded, spurious effects were still found for *R. palustris*, *R. sphenocephala*, and *R. sylvatica* as a result of inadequate modeling of site detection probability heterogeneity (Table 1). Despite no sites being occupied by *A. crepitans* or *B. americanus*, false positive errors resulted in variable seasonal biases in ψ, γ, and ε for both species (Table 2).

Only three false positive detections occurred for *P. crucifer*, but with $\psi = 0.09$, estimates from the minimum-AIC_c model (15% of all model weight) exhibited considerable positive bias in occupancy for all seasons ($\hat{\psi} = 0.17$, SE = 0.03). There was a large degree of model selection uncertainty in this analysis, and the identical model including an additional southern leopard frog treatment effect on ψ_1 received 14% of the AIC_c weight. Although $\psi_t = 0.17$ for the control sites and $\psi_t = 0.00$ for the southern leopard frog treatment sites, this model predicted $\hat{\psi} = 0.17$ (SE = 0.03) and $\hat{\psi} = 0.08$ (SE = 0.05), respectively. There were nine false positive detections for *P. feriarum*, but biases in occupancy were considerably smaller for this species than for *P. crucifer* (Table 2).

DISCUSSION

Despite the relatively simple and controlled system under examination, false positive errors repeatedly led to misleading inferences about occupancy. The estimators were generally biased, and subject to spurious variability due to season, treatment, or distance effects. As hypothesized, false positive errors tended to induce positive biases in standard dynamic occupancy models, but we had not anticipated the substantial variability in these biases as a result of the false positive detection process itself varying with distance, treatment, or season.

The magnitude and direction of these biases were driven largely by the timing and pattern of all positive detections, and not simply by the number (or proportion) of false positive errors. In their small simulation study, Royle and Link (2006) assumed false positives occurred randomly across sites and found positive biases in occupancy estimators to be reduced when these errors occurred less frequently. However, false positives did not occur randomly across sites in our study. As a result, we found that far fewer false positive detections (<5% of all detections) can induce substantial bias in these models. For example, *P. crucifer* had perfect true positive detection at every distance and the lowest proportion (1%) of false positive errors, but its occupancy estimates exhibited far greater biases when compared to most other species. Because false positives happened to occur on sites receiving the southern leopard frog treatment (of which none were actually occupied), the best supported model predicted lower detection probabilities for southern leopard frog treatment sites (i.e., site heterogeneity) and identical occu-

Table 2. Estimated bias (and standard error) for season-dependent site occupancy ($\hat{\psi}_t$), colonization ($\hat{\gamma}_t$), and local extinction ($\hat{\varepsilon}_t$) probabilities for seven species under four analysis scenarios with and without the incorporation of distance effects on detection probability (p): false positive errors without distance effects (FP), false positive errors with distance effects (FPD), no false positive errors without distance effects (NF), and no false positive errors with distance effects (NFD).

Species and analysis	True ψ_t	Bias (SE)						
		$\hat{\psi}_1$	$\hat{\psi}_2$	$\hat{\psi}_3$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\varepsilon}_1$	$\hat{\varepsilon}_2$
A. crepitans†	0.00							
FP		0.02 (0.01)	0.18 (0.15)	0.32 (0.25)	0.17 (0.15)	0.17 (0.15)	0.00 (0.00)	0.00 (0.00)
FPD		0.03 (0.01)	0.09 (0.03)	0.11 (0.05)	0.08 (0.03)	0.08 (0.03)	0.68 (0.32)	0.68 (0.32)
B. americanus‡	0.00							
FP		0.00 (0.00)	0.02 (0.01)	0.00 (0.00)	0.02 (0.01)	0.00 (0.00)	1.00 (0.00)	1.00 (0.00)
FPD		0.00 (0.00)	0.02 (0.01)	0.00 (0.00)	0.02 (0.01)	0.00 (0.00)	1.00 (0.00)	1.00 (0.00)
P. crucifer	0.09							
FP		0.07 (0.03)	0.08 (0.03)	0.08 (0.03)	0.01 (0.01)	0.00 (0.00)	0.00 (0.00)	0.05 (0.05)
FPD		0.07 (0.03)	0.08 (0.03)	0.08 (0.03)	0.01 (0.01)	0.00 (0.00)	0.00 (0.00)	0.05 (0.05)
NF		0.00 (0.02)	0.00 (0.02)	0.00 (0.02)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
NFD		0.00 (0.02)	0.00 (0.02)	0.00 (0.02)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
P. feriarum	0.19							
FP		0.01 (0.02)	0.01 (0.02)	0.01 (0.02)	0.02 (0.01)	0.01 (0.01)	0.05 (0.03)	0.06 (0.03)
FPD		0.01 (0.02)	0.01 (0.02)	0.01 (0.02)	0.02 (0.01)	0.01 (0.01)	0.05 (0.03)	0.06 (0.03)
NF		0.00 (0.02)	0.00 (0.02)	0.00 (0.02)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
NFD		0.00 (0.02)	0.00 (0.02)	0.00 (0.02)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
R. palustris	0.19							
FP		−0.02 (0.02)	0.02 (0.02)	0.00 (0.02)	0.06 (0.01)	0.02 (0.01)	0.08 (0.04)	0.14 (0.04)
FPD		0.04 (0.02)	0.03 (0.02)	0.01 (0.02)	0.01 (0.01)	0.00 (0.00)	0.06 (0.03)	0.11 (0.04)
NF		−0.04 (0.02)	−0.01 (0.02)	0.00 (0.02)	0.03 (0.01)	0.02 (0.01)	0.00 (0.00)	0.02 (0.02)
NFD		0.00 (0.02)	0.00 (0.02)	0.00 (0.02)	0.01 (0.01)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
R. sphenocephala	0.18							
FP		−0.05 (0.02)	−0.03 (0.02)	0.11 (0.03)	0.06 (0.02)	0.19 (0.03)	0.24 (0.08)	0.12 (0.07)
FPD		0.02 (0.03)	0.01 (0.03)	0.13 (0.04)	0.02 (0.02)	0.17 (0.03)	0.12 (0.07)	0.07 (0.07)
NF		−0.06 (0.02)	−0.07 (0.02)	−0.01 (0.03)	0.02 (0.01)	0.07 (0.02)	0.15 (0.07)	0.00 (0.00)
NFD		−0.01 (0.03)	−0.01 (0.03)	0.00 (0.03)	0.00 (0.00)	0.01 (0.01)	0.00 (0.00)	0.00 (0.00)
R. sylvatica	0.19							
FP		−0.04 (0.02)	−0.03 (0.02)	−0.02 (0.02)	0.04 (0.01)	0.04 (0.01)	0.13 (0.05)	0.18 (0.05)
FPD		0.03 (0.02)	0.01 (0.02)	0.01 (0.02)	0.00 (0.00)	0.00 (0.00)	0.06 (0.03)	0.00 (0.00)
NF		−0.05 (0.02)	−0.04 (0.02)	−0.02 (0.02)	0.03 (0.01)	0.04 (0.01)	0.05 (0.04)	0.14 (0.05)
NFD		0.00 (0.02)	0.00 (0.02)	0.00 (0.02)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)

Notes: True ψ_t varied by species but was constant across seasons. For all species, true $\gamma_t = \varepsilon_t = 0$.

† Model constrained $\gamma_1 = \gamma^2$ and $\varepsilon_1 = \varepsilon_2$ because season-dependent γ_t and ε_t were not uniquely estimable.

‡ Model constrained $\varepsilon_1 = \varepsilon_2$ because season-dependent ψ_t and ε_t were not uniquely estimable.

pancy probabilities for all sites, regardless of treatment. Not only was this a substantial overestimate of *P. crucifer* site occupancy, but it completely obscured the fact that there was a definitive difference in occupancy between treatment sites.

Perhaps even more alarming, spurious patterns and dynamics of species occurrence would be inferred based on detections of species that were not even present in the system. In the case of *A. crepitans*, a complicated story would be required to explain the variability in site occupancy as a function of both season and distance (Fig. 1a). The best evidence for this species suggests an increase in overall site occupancy probabilities over the seasons, with lower occupancy at further distance sites in the second and third seasons. These patterns were simply an artifact of the tendency of false positive errors to occur as a non-random process, these being particularly more likely at greater distances in the second season. Based on the evidence for *B. americanus* (Table 2), one might erroneously conclude it is a very rare or elusive species warranting additional monitoring efforts. We strongly recommend that practitioners monitoring rare or elusive species carefully consider the possibility of false positive errors before making inferences about species occurrence based on a relatively small number of detections.

As hypothesized, unmodeled heterogeneity due to site-variable calling distances tended to induce biases in these estimators. The best supported models for these species were unbiased in all parameters only when we removed false positives and accounted for distance effects. By sheer coincidence, the positive biases induced by false positives and the negative biases induced by unmodeled heterogeneity occasionally resulted in the false positive analysis excluding distance effects performing better, in terms of occupancy estimation, than expected. However, colonization and local extinction probabilities tended to be substantially overestimated,

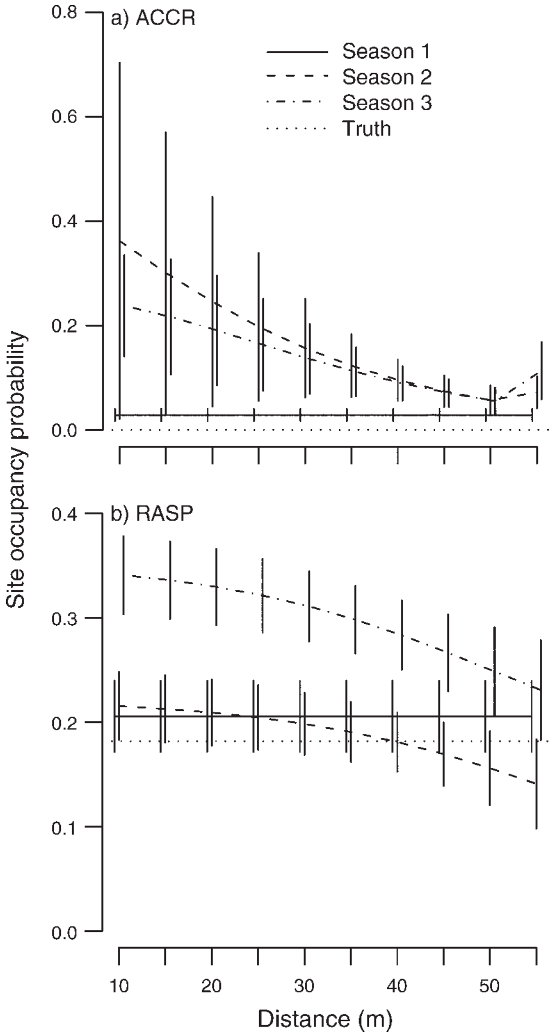

FIG. 1. Site occupancy estimates (±SE) from field calling survey simulations for (a) *Acris crepitans* (ACCR) and (b) *Rana sphenocephala* (RASP) across three seasons by site distance under the minimum-AIC$_c$ model including false positive errors and distance effects. True site occupancy was constant between seasons and site distances.

thereby leading to particularly misleading inferences about the dynamics within the system. As this implies, deliberately ignoring heterogeneity to counter the biases induced by false positive errors does not present a defensible means for addressing the issue.

We acknowledge that the simple system under consideration here was artificial. However, we believe it captures the essence of the potential pitfalls when applying standard occupancy estimators to field data susceptible to false positive errors and heterogeneous detection probabilities. When assembling the data, the site detection histories were composed in an objective, systematic manner to consistently and accurately portray the relative importance of these forms of observation error for each species. Except when individual species data were confounded by the competing species treatment (i.e., *P. crucifer* and *R. sphenocephala*), all available detection and non-detection data were used in each analysis. This allowed the effects of false positives and detection heterogeneity to be assessed relative to the complete detection histories for each species within the system. In other words, we did not choose to include (or omit) sites with false positives or distance effects in an effort to deliberately exaggerate (or understate) the impacts these can have on inferences about species occurrence. As this was a simulated field study with only five (albeit expert) observers over a short period of time, we do not believe our findings are fully representative of the expected level of error in actual field surveys. This will likely vary according to species, local abundances, environmental conditions, and observer abilities. We have instead focused on evaluating the biases of standard dynamic occupancy estimators given the levels of observation error found within this specific system. Although the detection process in this system was likely simpler than that found in actual surveys, the number of species, sites, observers, seasons, and visits were quite representative of the sampling efforts typically afforded in these studies. By conducting simulated surveys in the field with human observers, we believe this has provided a degree of realism that would be very difficult to reproduce in the simplified conditions generally assumed in computer simulation experiments or call recognition tests (e.g., Genet and Sargent 2003, Royle and Link 2006). To our knowledge, this is the first field investigation of the impacts these types of observation error can have on occupancy models when truth was known and controlled.

As with false negative detections, we contend that false positive detections are unlikely to be eliminated through study design alone. These errors must therefore also be incorporated into the estimation framework, based on probability theory, to make reliable inferences about occupancy and its changes over time. The use of ancillary information about false positive error rates from previous studies (e.g., Genet and Sargent 2003) or current online questionnaires (e.g., Weir 2009) may provide a (limited) means for correcting standard occupancy estimators. Although conceptually valuable, the model-based solution proposed by Royle and Link (2006) does not solve the false positive issues identified herein for several reasons. As acknowledged by Royle and Link (2006), in the absence of a priori information the standard occupancy data contain no information to distinguish the false positive and false negative detection processes. In other words, in the presence of misclassification one cannot know that any detection was actually a true positive detection, and thus, one cannot definitively know that any site was actually occupied. Consequently, there are symmetries in the Royle and Link (2006) likelihood such that no unique set of solutions exists for ψ, the probability of a true positive detection (p_{11}), and the probability of a false positive detection (p_{10}). For example, their model cannot be

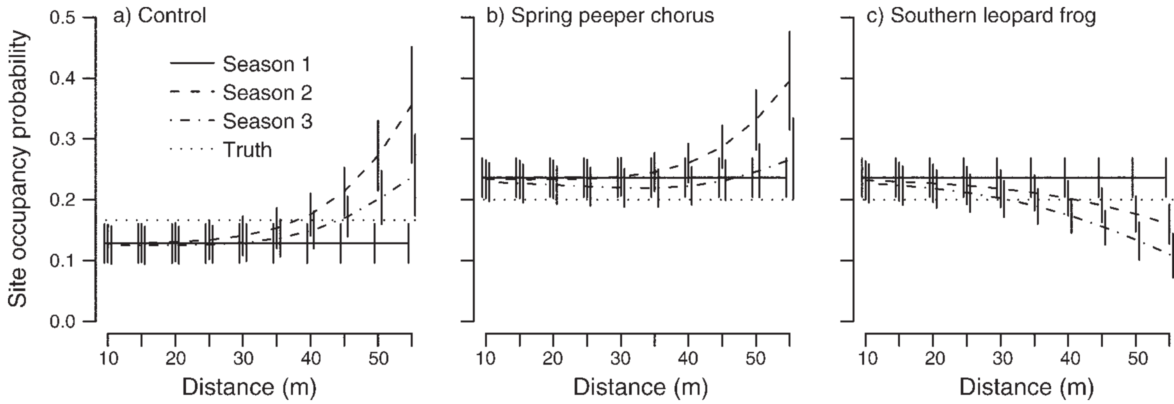

FIG. 2. Site occupancy estimates (±SE) from field calling survey simulations for *Rana palustris* across three seasons by site distance category from the minimum Akaike's information criterion adjusted for small sample sizes (AIC_c) model including false positives and distance effects. Estimates varied between background competing species treatments: (a) control, (b) continuous spring peeper chorus, and (c) continuous southern leopard frog call. True site occupancy was constant between seasons and distance categories but varied slightly by treatment.

used to identify a "phantom" species for which all detections were false positives (e.g., *A. crepitans* and *B. americanus*). If one were to assume a priori that $p_{11} > p_{10}$ (as reasonably suggested by Royle and Link), then one would erroneously conclude that a phantom species were present. Only if it were assumed (perhaps less reasonably) that $p_{10} > p_{11}$ could an unbiased estimate of $\hat{\psi} \approx 0$ potentially be obtained. However, the symmetries also cause estimation to be unstable when p_{11} and p_{10} are similar (e.g., *A. crepitans*, *B. americanus*, and *R. sylvatica*). In this case, there are more than two (and possibly many) distinct sets of solutions for the model parameters that receive identical support. Importantly, this misclassification model cannot distinguish false positive errors from heterogeneity in true positive detection probabilities, even with a large sample consisting mostly (or entirely) of true positive detections (e.g., Fitzpatrick et al. 2009). By design the sites in this experiment exhibited considerable heterogeneity in p_{11}, thus the Royle and Link (2006) model would be inappropriate for occupancy estimation for any of the target species. For example, although true $\psi_t = 0.18$ for *R. sphenocephala*, the model yields $\hat{\psi}_1 = 0.10$ (SE $= 0.02$), $\hat{\psi}_2 = 0.11$ (SE $= 0.02$), and $\hat{\psi}_3 = 0.32$ (SE $= 0.04$) (see Appendix B for additional examples).

To our knowledge, an occupancy modeling framework incorporating both false positive errors and heterogeneity has yet to be formally developed. Although a step in the right direction, the misclassification model proposed by Royle and Link (2006) has seen little use for reasons described above, and because false positive errors are generally considered to be trivial. We wish to help dispel this myth and encourage the development of identifiable occupancy estimation models that utilize direct information about the false positive detection process. This idea has recently begun to appear in the capture-mark-recapture literature (e.g., Lukacs and Burnham 2005, Link et al. 2010). Fortunately, there is much promise for such model development under standard occupancy sampling protocols utilizing auditory detections. For example, the detection data generated from anuran calling surveys often include an index of calling intensity, ranging from a single calling individual to a full chorus (e.g., Weir and Mossman 2005). Assuming well trained observers, it may be reasonable to assume that a detection based on the call of a single individual may more likely be a false positive detection than one based on a continuous chorus of many individuals. If one were to assign probabilities to these indices that reflected the relative certainty that any given detection was a true positive detection, therein would lie information about the false positive detection process. With it becoming increasingly apparent that false positives are an important, but too often overlooked, component of observation error in studies of species occurrence, this would constitute one of the many productive avenues for future research in this area.

Acknowledgments

This project was funded by the U.S. Geological Survey Amphibian Research and Monitoring Initiative. We thank our volunteer observers, A. Braswell, E. Corey, J. Hall, J. Humphries, and L. Weir, for their time and support.

Literature Cited

Burnham, K. P., and D. R. Anderson. 2002. Model selection and multimodel inference: an information-theoretic approach. Second edition. Springer-Verlag, New York, New York, USA.

Fitzpatrick, M. C., E. L. Preisser, A. M. Ellison, and J. S. Elkinton. 2009. Observer bias and the detection of low-density populations. Ecological Applications 19:1673–1679.

Genet, K. S., and L G. Sargent. 2003. Evaluation of methods and data quality from a volunteer-based amphibian call survey. Wildlife Society Bulletin 31:703–714.

Link, W. A., J. Yoshizaki, L. L. Bailey, and K. H. Pollock. 2010. Uncovering a latent multinomial: analysis of mark–recapture data with misidentification. Biometrics. 66:178–185.

Lotz, A., and C. R. Allen. 2007. Observer bias in anuran call surveys. Journal of Wildlife Management 71:675–679.

Lukacs, P. M., and K. P. Burnham. 2005. Estimating population size from DNA-based closed capture-recapture data incorporating genotyping error. Journal of Wildlife Management 69:396–403.

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

MacKenzie, D. I., J. D. Nichols, J. E. Hines, M. G. Knutson, and A. B. Franklin. 2003. Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly. Ecology 84:2200–2207.

Mazerolle, M. J., L. L. Bailey, W. L. Kendall, J. A. Royle, S. J. Converse, and J. D. Nichols. 2007. Making great leaps forward: accounting for detectability in herpetological field studies. Journal of Herpetology 41:672–689.

McClintock, B. T., L. L. Bailey, K. H. Pollock, and T. R. Simons. 2010. Experimental investigation of observation error in anuran call surveys. Journal of Wildlife Management, *in press*.

Nichols, J. D., L. Thomas, and P. B. Conn. 2009. Inferences about landbird abundance from count data: recent advances and future directions. Pages 201–235 *in* D. L. Thomson, E. G. Cooch, and M. J. Conroy, editors. Modeling demographic processes in marked populations. Springer, New York, New York, USA.

Royle, J. A. 2006. Site occupancy models with heterogeneous detection probabilities. Biometrics 62:97–102.

Royle, J. A., and W. A. Link. 2006. Generalized site occupancy models allowing for false positive and false negative errors. Ecology 87:835–841.

Royle, J. A., and J. D. Nichols. 2003. Estimating abundance from repeated presence–absence data or point counts. Ecology 84:777–790.

Sauer, J. R., J. E. Fallon, and R. Johnson. 2003. Use of North American Breeding Bird Survey data to estimate population change for bird conservation regions. Journal of Wildlife Management 67:372–389.

Simons, T. R., M. W. Alldredge, K. H. Pollock, and J. M. Wettroth. 2007. Experimental analysis of the auditory detection process on avian point counts. Auk 124:986–999.

Simons, T. R., K. H. Pollock, J. M. Wettroth, M. W. Alldredge, K. Pacifici, and J. Brewster. 2009. Sources of measurement error, misclassification error, and bias in auditory avian point count data. Pages 237–254 *in* D. L. Thomson, E. G. Cooch, and M. J. Conroy, editors. Modeling demographic processes in marked populations. Springer, New York, New York, USA.

Weir, L. A. 2009. USGS frog quiz. ⟨http://www.pwrc.usgs.gov/frogquiz⟩

Weir, L. A., and M. J. Mossman. 2005. North American amphibian monitoring program (NAAMP). Pages 307–313 *in* M. J. Lannoo, editor. Amphibian declines: conservation status of United States species. University of California Press, Berkeley, California, USA.

White, G. C., and K. P. Burnham. 1999. Program MARK: survival estimation from populations of marked individuals. Bird Study 46:120–139.

Wintle, B. A., R. P. Kavanagh, M. A. McCarthy, and M. A. Burgman. 2005. Estimating and dealing with detectability in occupancy surveys for forest owls and arboreal marsupials. Journal of Wildlife Management 69:905–917.

APPENDIX A

Species-specific occupancy analysis results (*Ecological Archives* E091-176-A1).

APPENDIX B

Royle and Link (2006) misclassification model results (*Ecological Archives* E091-176-A2).

APPENDIX C

Naïve estimates of occupancy with false positives (*Ecological Archives* E091-176-A3).