Title:

Integrating over uncertainty in spatial scale of response within multispecies occupancy models yields more accurate assessments of community composition

Running title: Spatial-scale selection in occupancy models

Author Details:

Luke Owen Frishkoff[1,*]

D. Luke Mahler[1]

Marie-Josée Fortin[1]

1. Department of Ecology and Evolutionary Biology, University of Toronto, Toronto, ON M5S 1A1

* to whom correspondence should be addressed: frishkol@gmail.com

Keywords: Imperfect detection, habitat selection, landscape, forest fragmentation, habitat conversion, joint community model, focal-site multi-scale

Abstract:

1. Species abundance and community composition are affected not only by the local environment, but also by broader landscape and regional context. Yet determining the spatial scale at which landscapes affect species remains a persistent challenge that hinders ecologists' abilities to understand how environmental gradients influence species presence and shape entire communities, especially in the face of data deficient species and imperfect species detection.

2. Here we present a Bayesian framework that allows uncertainty surrounding the 'true' spatial scale of species' responses (*i.e.,* changes in presence/absence) to be integrated directly into a community hierarchical model.

3. This scale selecting multi-species occupancy model (ssMSOM) estimates the scale of response, and shows high accuracy and correct type I error rates across a broad range of simulation conditions. In contrast, ensembles of single species GLMs frequently fail to detect the correct spatial scale of response, and are often falsely confident in favoring the incorrect spatial scale, especially as species' detection probabilities deviate from perfect.

4. Integrating spatial scale selection directly into hierarchical community models provides a means of formally testing hypotheses regarding spatial scales of response, and more accurately determining the environmental drivers that shape communities.

Introduction

Features of the landscape beyond the local scale often affect the processes that give rise to patterns of community composition (Wiens 1989; Levin 1992; Kneitel & Chase 2004; Dray *et al.* 2012; Fortin *et al.* 2012; McGarigal *et al.* 2016). As a result, ecologists have sought to quantify what landscape features, in what contexts, and at what spatial scales explain the presence and abundance of species. Yet determining how species respond to the landscape has been challenging, in part because the relevant spatial scale(s) at which environmental conditions affect species and communities are rarely known *a priori*. This difficulty has led to uncertainty regarding the conclusions of many landscape level studies (Jackson & Fahrig 2015). The development of statistical methods that more robustly incorporating scales of responses within the statistical analysis of communities (Borcard & Legendre 2002; Jombart *et al.* 2009; Matthiopoulos *et al.* 2011; Dray *et al.* 2012; Warton *et al.* 2015; Ovaskainen *et al.* 2016, 2017), and more accurately convey uncertainty regarding these scales (Chandler & Hepinstall-Cymerman 2016), have the potential to accelerate basic and applied ecological research.

When considering landscape level effects on species presence, abundance, or biomass, two properties of the species are generally of interest. First, at what spatial scale does the species respond to the environment (Desrochers *et al.* 2010), and second, how do they respond (positively or negatively)? The most commonly used approach for determining spatial scale of response (*i.e.,* the spatial context, spatial contingency; Fortin *et al.* 2012) quantifies the average environmental value within buffers of various radii (Holland *et al.* 2004; Weaver *et al.* 2012; Zuckerberg *et al.* 2012; McGarigal *et al.* 2016), and then repeats a statistical analysis using the environmental covariate at each spatial

63 scale (Figure 1). For each species in turn, or for some community-level index like species
64 richness or diversity (*e.g.*, Shannon Index), the most likely spatial scale (as quantified by
65 AICc, correlation coefficient, or slope parameter value) is selected to represent the best
66 match of a species' response to landscape heterogeneity.

67 This multi-scale analysis approach has been successful in elucidating species and
68 community responses (McGarigal *et al.* 2016). By considering landscapes as a whole it
69 has helped quantify the benefits of small forest fragments to biological communities, and
70 related ecosystem services (Karp *et al.* 2013; Mendenhall *et al.* 2016). More generally it
71 has highlighted that species respond to different environmental conditions at different
72 spatial scales, and that species distribution models possess greater predictive power when
73 these multiple scales are directly incorporated (Desrochers *et al.* 2010; Weaver *et al.*
74 2012). However, the current multi-scale approach does present a number of problems
75 related to estimating the spatial scale of response, exacerbating uncertainty by treating
76 species individually rather than the community as an integrated whole, and ignoring
77 issues with species detectability. All of these will inflate error in estimating the true
78 spatial scale of response, and quantifying how species respond to the environment.

79 First, single species model comparison approaches that select a single best model
80 typically neither quantify nor integrate over uncertainty regarding scale selection. This
81 means that other parameters may be biased if the 'most likely' scale is not the true scale.
82 Relatedly, the set of scales analyzed is often quite small (Desrochers *et al.* 2010; Jackson
83 & Fahrig 2015), and as a result is unlikely to even include the true spatial scale. Meta-
84 analysis has shown that the most likely spatial scale is often at one of the extremes of
85 those analyzed—suggesting that the true spatial scale is even more extreme (Jackson &

Fahrig 2015). Recently, Chandler & Hepinstall-Cymerman (2016) proposed a modeling approach that internalizes spatial scale estimation within a single species model by using smoothing kernels to average landscape variables around focal sites. This single-species model addresses these two problems by maximizing likelihood over the spatial scale parameter, and also allows confidence intervals to be calculated around it. Further, because spatial scale is a continuous (albeit bounded) parameter, it eliminates the problem of not including the true spatial scale among the scales assessed (provided an appropriate range of scales is investigated).

The second major problem with the standard approach is that it makes assessments of multiple species in a species-by-species fashion. In the study of entire communities, however, this approach is often inadequate, because it ignores rare species (which are both typically of greatest conservation concern, and may be trophically influential). This problem is especially true in tropical communities, which are particularly under threat from land-use change, and where species rarity is common (MacArthur 1969; Hubbell 2001). Further, a species by species approach is prone to estimation bias and loss of power (Ovaskainen & Soininnen 2011; Banks-Leite *et al.* 2014). Hierarchical joint community models have been proposed to move beyond piecemeal assessments (Ovaskainen & Soininnen 2011; Warton *et al.* 2015; Ovaskainen *et al.* 2017). By assuming that species parameters come from common distributions, overall community error is minimized and rare species can be included in analyses.

Finally, imperfect detection of species is a problem for animal communities generally, especially when species traits, site characteristics, or the time or conditions of observation influence detectability. Multi-species occupancy models (MSOMs) are a

109 commonly adopted solution to account for imperfect and variable detection within
110 communities (Iknayan *et al.* 2013). MSOMs are typically implemented in a Bayesian
111 framework (relying on MCMC to overcome challenges in maximizing likelihoods when
112 numerous random effects exist). Yet, model comparison and selection is still difficult to
113 implement with Bayesian models (*e.g.*, Hooten & Hobbs, 2015; but see Lele *et al.* 2007).
114 Together this means that comparing models across a large range of spatial scales would
115 be both time extremely time consuming (because MCMCs are relatively slow), and non-
116 trivial to implement. Because MSOMs are not easily amenable to adequate testing at
117 multiple spatial scales, the scale of response has generally not been incorporated into
118 community analyses that incorporate imperfect detection. Consequently, their power has
119 not been sufficiently directed to understand how landscape features structure community-
120 level processes.

121 Fortunately scale selection can be integrated directly into MSOMs by establishing
122 a parameter that allows the spatial scale at which species respond to be estimated (*e.g.*,
123 Chandler & Hepinstall-Cymerman, 2016). When extended to the community as a whole,
124 incorporating scale selection into the model should result in the appropriate spatial
125 scale(s) being estimated directly from the data within a single model run, and also ensure
126 that other parameter estimates are not biased because they were analyzed at the wrong
127 spatial scale. One of us (LOF) recently developed the foundations of the approach
128 presented here in an attempt to internalize scale selection within hierarchical multispecies
129 models to overcome uncertainty about what spatial scale to analyze data for two specific
130 empirical studies (Frank et al. In Press; Karp et al, In review). However, this technique
131 was not fully developed in those works, and it remains unclear whether this approach has

132 correct type I error rates, and whether it does indeed increase power and accuracy over
133 more traditional approaches. Here we fully describe, and demonstrate the use of, a multi
134 spatial scale selection multi-species occupancy model (hereafter; ssMSOM), and test its
135 performance in estimating both spatial scale of response, and species' strengths of
136 response to the environment (*i.e.*, a landscape-level covariate). We compare this approach
137 to the standard method of analysis: a series of species-by-species GLMs. While the
138 approach here is demonstrated with a simple single season occupancy model for multiple
139 species, the internalized scale selection is generalizable. For example, it could be directly
140 integrated into abundance models, or combined with flexible hierarchical modeling
141 approaches to query population dynamics through times, or the effects of species traits,
142 phylogenetic relatedness, or interspecific competition on community structure (Yackulic
143 *et al.* 2014; Frishkoff *et al.* 2017).
144
145 **Methods**
146 *Model overview:*
147 The scale selecting multi-species occupancy model (ssMSOM) estimates parameters
148 related to occupancy and detection probabilities in communities containing multiple
149 species (indexed by i), across multiple sites (j), with multiple site visits (k). The observed
150 detection histories ($Y_{i,j,k}$) are assumed to derive from unobserved (latent) occupancy states
151 ($Z_{i,j}$, where $Z_{i,j} = 1$ for presence, and 0 for absence) and detection probabilities determined
152 by species, and site ($P_{i,j}$; visit based variability in detection is ignored for simulations and
153 models, but could be incorporated if desired). Specifically:

$$Y_{i,j,k} \sim Bern(Z_{i,j} * P_{i,j}).$$

154 The occupancy state ($Z_{i,j}$) in turn was assumed to come from some underlying occupancy
155 probability according to:

156 $$Z_{i,j} \sim Bern(\psi_{i,j}).$$

157 The detection process was modeled according to:

$$\text{logit}(\text{P}_{i,j}) = \alpha 0_i + \alpha 1_i * Env_{j,1} .$$

158 And the occupancy probability ($\psi_{i,j}$) according to:

$$\text{logit}(\psi_{i,j}) = \beta 0_i + \beta 1_i * Env_{j,s} + \gamma 0_j .$$

159 Here $Env_{j,s}$ is a site-by-scale matrix of some landscape level environmental variable
160 (centered and scaled within each column). All parameters in the α and β groups are
161 estimated for each species, with species terms drawn from a normal distribution of mean
162 (μ) and variance (σ^2) estimated from the data. γ terms were random intercepts for each
163 site (variance estimated from data around a mean of 0) designed to incorporate consistent
164 differences in occupancy probabilities in all species across sites that are not accounted for
165 by *Env*. The indexing value *s* (representing columns of the $Env_{j,s}$ matrix) spans multiple
166 spatial scales, and the parameter value of *s* that best fits the data is estimated from the
167 model. Because models are implemented using MCMC, this process results in a posterior
168 distribution for values of *s*, which fully integrates over the uncertainly regarding the
169 proper spatial scale, and which further can be used to select the most appropriate spatial
170 scale (*e.g.,* posterior mean or mode) or an interval of spatial scales that well describe the
171 data. This formulation is conceptually similar to generalized linear models that integrate
172 over phylogenetic uncertainty in tree topology (de Villemereuil *et al.* 2012). For
173 simplicity, the environmental effect of a species' detection probability is assumed to
174 come from the environmental conditions at the finest (*i.e.*, most local) spatial scale. This

assumption could be relaxed if there is reason to expect that more distant environmental conditions somehow affect detection probability.

For demonstration and testing purposes, we here assume that a single environmental gradient affects community composition. However, the ssMSOM could be generalized to include multiple environmental conditions (multiple site-by-scale environment matrixes, e.g., $Env1_{j,s}$ $Env2_{j,t}$ $Env3_{j,u}$, etc.), each affecting communities at different spatial scales (with scale parameters *s, t, u*, etc., each independently estimated from the data).

Simulation conditions

In order to test the performance of the ssMSOM, we simulated communities, using an *Env* matrix based on empirical landscape forest cover. Spatial forest cover data for simulation and analysis came from northwestern Costa Rica, used as part of a study of how local and landscape level habitat conversion affects community composition (Karp et. al. In review). In that study, sites were selected to ensure that local forest cover varied independently from landscape level forest cover. To measure surrounding forest cover, all tree cover within 1.5km of sites was classified using high-resolution Google Earth images obtained from 2013-2016. The resulting 5m-resolution tree cover map was verified based on ground-truthed data collected in the field. For analysis, site level forest cover proportion was calculated in radii from 50m to 1500m, in 50m increments, resulting in an *Env* matrix with 30 columns. For the ssMSOM it would be most appealing to use the smallest increments possible to generate the largest number of spatial scales possible, since using few spatial scales makes it likely that the true scale will not be

among the analyzed set. We settled on 50m increments because of a balance of computational efficiency in model runs, and increments that approximate a continuous stretch from our smallest to largest spatial scale.

To test the performance of ssMSOM under a variety of conditions, we simulated 120 communities (four at each of the 30 spatial scales), each with 16 species (N_{sp}), across 50 sites (N_{site}), with three site visits per site (N_{visit}). All species parameters were drawn from normal distributions, generating diversity in species' commonness (simulations included both common and rare species), and species' responses to tree cover (some responded negatively and others responded positively to 'deforestation'). This diversity of overall commonness and responses to the environment mimics patterns observed in many empirical systems. We repeated these simulations under five alternative detection scenarios:

1. Perfect detection, where the probability of detecting a species at a site if the site is occupied is 1.

2. High detection probability: average detectability equals ~ 0.5.

3. Low detection probability: average detectability equals ~ 0.25.

4. Low detection with detection affected by local environment: average detectability equals ~ 0.25, but rising to ~0.5 under high local values of *Env* and dropping to ~0.1 under low local values of *Env*.

5. Low detection with species-specific variation in detectability by local environment: Same as 4, but some species increase in detectability with increasing values of local *Env*, while others decrease in detectability.

For an overview of all simulation parameters see Table 1.

221

222 *Model comparison*

223 To examine ssMSOM performance, versus a typical analysis strategy for this type of
224 data, we compared it to a series of single species models fit using maximum likelihood
225 across all 30 spatial scales. These models are referred to as piecemeal GLMs throughout
226 and are described by standard binomial GLM functions of the form:

$$\text{logit}(\psi_j) = \beta 0 + \beta 1 * Env_j$$

$$Y_j \sim bern(\psi_j)$$

227 where ψ_j is the naïve occupancy probability of the focal species at the focal spatial scale,
228 and Y_j is the naïve occupancy state (*i.e.*, whether a species was detected at a site across all
229 site visits or not). To match the approach taken by empirical analyses species with
230 observations at fewer than 10% of all sites were excluded from analysis (because data
231 would presumably be insufficient for precise parameter estimates; *e.g.*, Desrochers *et al.*
232 2010; Zuckerberg *et al.* 2012). We then used AICc to choose the optimal spatial scale for
233 each species in turn.

234 We focus on two core questions when evaluating the ssMSOM versus standard
235 GLM approaches. First, does using an integrated community analysis provide more
236 accurate estimates of the correct spatial scale (*s*) than a piecemeal approach? For species
237 for which $\beta 1$ is close to 0, the spatial scale of response cannot be evaluated in the
238 piecemeal GLMs because the scale of response is undefined if the species does not
239 respond to the environment. For this set of analyses we therefore additionally excluded
240 all species for which $\beta 1$ was not significantly different from 0 in the most likely GLM, as
241 estimation of true spatial scale should be more accurate for the remaining species.

Second, even when estimating spatial scale is not the primary goal and is therefore considered a nuisance variable, does integrating over uncertainty regarding the correct spatial scale result in more accurate estimates of how species respond to the environment ($\beta 1$) than a piecemeal approach? To answer this question we additionally compared parameter estimates with GLMs using the true spatial scale under which simulations were conducted. While for empirical data the true spatial scale is never known without error, using it here represents the 'best-case scenario' for community level analyses of landscape level responses to the environment.

To quantify the accuracy of the ssMSOM versus a piecemeal GLM approach, we calculated the root mean square error (RMSE) across the entire community of the family of parameter estimates from the true simulated value. With regards to '*s*', we consider the posterior mode in the case of the ssMSOM, where as for the GLMs we consider the spatial scale that minimizes AICc for each species. To ensure the results are comparable in both approaches we calculate RMSE for each species in turn, even though in the case of the ssMSOM the parameter *s* is estimated for all species simultaneously, and is therefore identical for all species.

We also examine coverage probability of the posterior estimates (*i.e.*, the inverse of type I error). If models are behaving as expected, the 95% CIs of the parameter estimates should contain the true value 95% of the time. For ssMSOMs and GLMs the coverage probabilities for $\beta 1$ can be calculated directly from species-specific parameter estimates. Similarly coverage probabilities around '*s*' for the ssMSOMs can be calculated using equal tail Bayesian credible intervals around the posterior of *s*. To calculate a value equivalent to coverage probabilities of the spatial scale in the case of the piecemeal

GLMs we first calculated the AICc weight for all spatial scales for a given species, and then asked whether the true spatial scale was within the top 95% of the cumulative model weights.

Model Fitting

Models were fit using JAGS through the R environment. Simulation code in R and JAGS model code is available in the supplement. For MCMC analyses diffuse priors were used throughout, with a flat prior placed on '*s*'.

Results

Inferring spatial scale

The posterior mode of spatial scales from the ssMSOM tended to accurately estimate the true spatial scale of response, and had relatively low error, typically off by less than 100m under the conditions simulated (Figures 2 and 3). In contrast, piecemeal GLMs failed to consistently recover the true spatial scale for the majority of species in the community, even when detection was perfect. The degree of error was lower when restricting analysis to only those species for which the lowest AICc GLM showed a significant relationship with the environmental gradient, though RMSE across the entire community was still >3X that of the ssMSOM (Figure 3a). Further, when detection itself varied along the environmental gradient at local scales in a species-specific manner, using a standard GLM approach resulted in error in estimating species response scales that is no better (and sometime worse) than guessing a scale at random (Figure 3a). Not surprisingly, error in estimating the scale of response within piecemeal GLMs was

288 greatest for both species that were detected in a small number of sites, as well as species
289 detected in the majority of sites (Figure 3b).

290 The ssMSOM demonstrated correct type I error rates when estimating *s*,
291 regardless of detection regime. In contrast, piecemeal GLMs showed inflated type I error
292 when estimating the true spatial scale, which was exacerbated as detection probability
293 deviated from perfect (Figure 4). This behavior was further accentuated when excluding
294 species that did not have a significant response to the environmental gradient at its most
295 likely spatial scale, such that nearly 20% of all species assessments did not include the
296 true spatial scale model in the top 95% Akaike weighted models under a low-detection
297 regime with specific-specific variation in detectability by environment.
298
299 *Estimating species responses to the environment*
300 Estimates of species responses to the environment ($\beta 1$) were more accurate in the
301 ssMSOM than in piecemeal GLMs, even when GLMs were run using the true spatial
302 scale (Figure 5). These patterns were not strongly affected by detection regime, though in
303 general estimates are more accurate when detection probabilities are high. Type I error
304 does however strongly shift with detection. If detection is perfect, and the true spatial
305 scale in known *a priori*, then a piecemeal GLM approach performs as well as the
306 ssMSOM (Figure 6). However, when the spatial scale must be inferred from the data
307 GLMs generate falsely confident results, with the true values of species responses
308 excluded from the 95% confidence intervals up to 30% of the time under some simulated
309 conditions (*i.e.* 6X the nominal type I error rate). In contrast the ssMSOM possess 95%
310 CIs that behave has expected, regardless of detection regime.

311

312 **Discussion**

313 Here we described and tested the statistical properties of the ssMSOM against the
314 standard method for ascertaining species' and communities' scales of response. We find
315 that internalizing scale selection into the model results in greater community wide
316 accuracy for key parameter estimates, and reduces the probability of making incorrect
317 inferences. The key strength of the ssMSOM is that it does not rely on setting the spatial
318 scale *a priori*. Like the approach of Chandler & Hepinstall-Cymerman (2016), the
319 ssMSOM avoids the problem of researchers selecting only a few scales to investigate,
320 which are too narrow to include the true scale of response (Jackson & Fahrig 2015). This
321 of course requires that researchers first extract landscape data from as broad a range of
322 scales as possible, ideally in the finest increments possible. This allows spatial scale to be
323 treated as nearly continuous, such that 95% CIs can be created, and inferences made as
324 with any other continuous parameter in the model. When taking a flexible scale
325 estimation approach it is essential to use as fine scale environmental data as possible. If
326 environmental data are coarse with respect to the resolution at which species interact with
327 the environment then the estimated spatial scale of response will be strongly upwardly
328 biased and overall model performance will suffer (Mendenhall *et al.* 2011).

329

330 *Examples of empirical use*

331 Two recent studies have demonstrated the power of using the scale selection routine from
332 the ssMSOM (*i.e.*, indexing the *Env* matrix by scale) when analyzing empirical datasets.
333 Frank et al. (in press) used a phylogenetic occupancy model (Frishkoff *et al.* 2017) with

334 the internalized spatial scale selection method presented here, finding that bat responses
335 to deforestation are strongly phylogenetically conserved. Similarly, Karp et al (in review)
336 used Bayesian spatial scale selection embedded within an *N*-mixture model to examine
337 how bird communities responded to habitat conversion while accounting for imperfect
338 detection in order to understand how β-diversity was structured along land-use and
339 climate gradients. In both cases spatial scale selection strongly supported deforestation
340 affecting the communities at fairly small spatial scales. While in both cases scales at over
341 a kilometer away from focal sites were queried, for bats the posterior distribution peaked
342 at 50m, and excluded all scales above 100m, while for birds scales below 300m were
343 favored. Because of the ssMSOM framework, these studies were able to analyze both
344 common and rare species. Had these studies relied on individual GLMs (or species-by-
345 species occupancy models) the uncertainty around the scale of response would likely
346 have been extremely high, and un-estimatable for the majority of rare species. This was
347 particularly important in the case of Neotropical bats for which rare and hard to detect
348 species tended to be found in natural forests. Indeed, if imperfect detection were not
349 taken into account species richness would have appeared to have been unaffected by
350 forest loss, when in fact it declined sharply (Frank *et al. In press*). These early examples
351 of embedding spatial scale selection into hierarchical models highlight the broad
352 applicability of the method. The ssMSOM approach is easily extended to abundance
353 models (i.e., *N*-mixture or recapture models), or indeed any Bayesian implementation of
354 multispecies models with or without detection for which the true spatial scale of response
355 is unknown could benefit from the general approach.
356

Assumptions, limitations, and future directions

Critically, for the simulations presented (and within the ssMSOM itself) there is the assumption that a single, true spatial scale exists at which all species respond to the environment. This may or may not be true for a given assemblage in nature. Empirical studies have shown that different species respond to different spatial scales (*e.g.*, Chambers *et al.* 2016), and theoretical approaches suggest that some species traits may modulate the scale of response (Jackson & Fahrig 2012). However, we show in our simulations that empirical analyses conducted on a species by species basis (as past studies have been) are often unable to recover the true spatial scale at which species respond, and show high heterogeneity in the scale of response even if all species are simulated to respond at the same spatial scale. While many species likely do respond at different scales, this finding casts some doubt on the specific estimates of scales of response presented in past empirical studies. The high degree of inaccuracy inherent in the piecemeal GLM approach may be partially responsible for the lack of correlation between empirically estimated scales of response, and species traits thought to modulate these scales (Jackson & Fahrig 2015).

Future development of the ssMSOM and similar community wide approaches should be able to relax the assumption that all species have the same scale of response, although doing so may diminish the ability to precisely estimate response scales for rare species. One path would be to estimate spatial scale separately for two or more groups of species, delimited based on natural history knowledge, functional guild placement, or other *a priori* expectations (Pacifici *et al.* 2014). An *a priori* grouping based approach, however, is at best an imperfect solution. Ideally individual species' scales of response

would be allowed to vary from one another—using random effect structures within the scale selection component of the model would be one logical way to do so. Allowing random variation among species could additionally allow species' level covariates to affect the scale of response, thereby facilitating testing the hypothesis that some species' traits correlate with scale of response that both maintains high power, and is minimally afflicted by type I error.

In real communities, species' responses to the broader landscape might be predicated on conditions at the local scale (*i.e.*, an interaction between local and landscape scales). For example, a farm-land bird species might benefit from landscape level tree cover when the local habitat is agriculture, but might only exist in forest when there are low amounts of landscape level tree cover because it uses forest edge habitat. Allowing interaction terms between landscape and local effects (*e.g.*, forest cover as estimated within a point count radius) will allow these types of species interactions with the environment to be tested (Matthiopoulos *et al.* 2011; Paton & Matthiopoulos 2016).

Chandler and Hepinstall-Cyberman (2016) pointed out that the step function used to calculate proportion of focal habitat within a given radius has no theoretical basis, and instead favor a Gaussian weighting function. This approach could be easily implemented with the Bayesian framework presented here, by indexing the *Env* matrix based on the output of the weighting function over incremental changes in its key parameter. While, alternatives to the commonly used step functions are certainly appealing on theoretical grounds, at least one study that examined Gaussian weighting versus a step function radius method found that models performed roughly equivalently (Timm *et al.* 2016).

403 *Conclusion*

404 Humans are altering landscapes across the globe, such that the remaining extent of
405 natural habitats are often much diminished and severely fragmented (Haddad *et al.* 2015).
406 Such complex, heterogeneous landscapes challenge ecologists' abilities to discern the
407 underlying environmental drivers of community composition. Yet achieving successful
408 conservation strategies in these landscapes requires simultaneously describing and
409 predicting how these spatially heterogeneous environments affect not just individual
410 species, but entire communities. Internalizing spatial scale selection within community
411 models offers one approach to uncover the environmental drivers behind such community
412 change while accommodating the unavoidable uncertainty in the 'true' scale of species'
413 responses. The ssMSOM possess high accuracy and correct type I error rates when both
414 identifying the spatial scale of response, and the direction and magnitude with which
415 individual species respond to environmental gradients. This approach represents a
416 promising path forward for understanding the ecological drivers of community
417 composition, and the consequences of ongoing environmental change.

418

419 **Acknowledgements**

420 We would like to thank D. S. Karp and A. Echeverri for use of fine-scale forest cover
421 data from northwestern Costa Rica. This study was supported by a University of Toronto
422 Ecology and Evolutionary Biology Postdoctoral Fellowship to LOF and NSERC
423 Discovery Grants to DLM and M-JF.

424 **Data Accessibility**: R scripts for simulating communities and running ssMSOMs are
425 available in the supplemental material.

426 **References**
427
428 Banks-Leite, C., Pardini, R., Boscolo, D., Cassano, C.R., Püttker, T., Barros, C.S. &
429 Barlow, J. (2014). Assessing the utility of statistical adjustments for imperfect
430 detection in tropical conservation science (J. Matthiopoulos, Ed.). *Journal of*
431 *Applied Ecology*, n/a-n/a.
432 Borcard, D. & Legendre, P. (2002). All-scale spatial analysis of ecological data by means
433 of principal coordinates of neighbour matrices. *Ecological Modelling*, **153**, 51–68.
434 Chambers, C.L., Cushman, S.A., Medina-Fitoria, A., Martínez-Fonseca, J. & Chávez-
435 Velásquez, M. (2016). Influences of scale on bat habitat relationships in a forested
436 landscape in Nicaragua. *Landscape Ecology*, **31**, 1299–1318.
437 Chandler, R. & Hepinstall-Cymerman, J. (2016). Estimating the spatial scales of
438 landscape effects on abundance. *Landscape Ecology*, **31**, 1383–1394.
439 Desrochers, A., Renaud, C., Hochachka, W.M. & Cadman, M. (2010). Area-sensitivity
440 by forest songbirds: Theoretical and practical implications of scale-dependency.
441 *Ecography*, **33**, 921–931.
442 Dray, S., Pélissier, R., Couteron, P., Fortin, M.J., Legendre, P., Peres-Neto, P.R., Bellier,
443 E., Bivand, R., Blanchet, F.G., De Cáceres, M., Dufour, A.B., Heegaard, E.,
444 Jombart, T., Munoz, F., Oksanen, J., Thioulouse, J. & Wagner, H.H. (2012).
445 Community ecology in the age of multivariate spatial analysis. *Ecological*
446 *Monographs*, **82**, 257–275.
447 Fortin, M.J., James, P.M.A., MacKenzie, A., Melles, S.J. & Rayfield, B. (2012). Spatial
448 statistics, spatial regression, and graph theory in ecology. *Spatial Statistics*, **1**, 100–
449 109.
450 Frank, H.K., Frishkoff, L.O., Mendenhall, C.D., Daily, G.C. & Hadly, E.A. Phylogeny,
451 traits and biodiversity of a Neotropical bat assemblage: Close relatives show similar
452 responses to local deforestation. *The American naturalist*.
453 Frishkoff, L.O., De Valpine, P. & M'Gonigle, L.K. (2017). Phylogenetic occupancy
454 models integrate imperfect detection and phylogenetic signal to analyze community
455 structure. *Ecology*, **98**, 198–210.
456 Haddad, N.M., Brudvig, L.A., Clobert, J., Davies, K.F., Gonzalez, A., Holt, R.D.,
457 Lovejoy, T.E., Sexton, J.O., Austin, M.P., Collins, C.D., Cook, W.M., Damschen,
458 E.I., Ewers, R.M., Foster, B.L., Jenkins, C.N., King, A.J., Laurance, W.F., Levey,
459 D.J., Margules, C.R., Melbourne, B.A., Nicholls, A.O., Orrock, J.L., Song, D.-X. &
460 Townshend, J.R. (2015). Habitat fragmentation and its lasting impact on Earth's
461 ecosystems. *Science Advances*, **1**, e1500052–e1500052.
462 Holland, J.D., Bert, D.G. & Fahrig, L. (2004). Determining the Spatial Scale of Species'
463 Response to Habitat. *BioScience*, **54**, 227.
464 Hooten, M.B. & Hobbs, N.T. (2015). A Guide to Bayesian Model Selection. *Ecological*
465 *Monographs*, **85**, 3–28.
466 Hubbell, S.P. (2001). *The unified Neutral Theory of Biodiversity and Biogeography*.
467 Princeton University Press, Princeton, NY.
468 Iknayan, K.J., Tingley, M.W., Furnas, B.J. & Beissinger, S.R. (2013). Detecting
469 diversity: emerging methods to estimate species diversity. *Trends in ecology &*
470 *evolution*, **29**, 97–106.
471 Jackson, H.B. & Fahrig, L. (2015). Are ecologists conducting research at the optimal

scale? *Global Ecology and Biogeography*, **24**, 52–63.
Jackson, H.B. & Fahrig, L. (2012). What size is a biologically relevant landscape? *Landscape Ecology*, **27**, 929–941.
Jombart, T., Dray, S. & Dufour, A.B. (2009). Finding essential scales of spatial variation in ecological data: A multivariate approach. *Ecography*, **32**, 161–168.
Karp, D.S., Mendenhall, C.D., Sandí, R.F., Chaumont, N., Ehrlich, P.R., Hadly, E.A. & Daily, G.C. (2013). Forest bolsters bird abundance, pest control and coffee yield (J. Lawler, Ed.). *Ecology Letters*, **16**, 1339–1347.
Kneitel, J.M. & Chase, J.M. (2004). Trade-offs in community ecology: Linking spatial scales and species coexistence. *Ecology Letters*, **7**, 69–80.
Lele, S.R., Dennis, B. & Lutscher, F. (2007). Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods. *Ecology letters*, **10**, 551–63.
Levin, S.A. (1992). The problem of pattern and scale in ecology. *Ecology*, **73**, 1943–1967.
MacArthur, R.H. (1969). Patterns of communities in the tropics. *Biological Journal of the Linnean Society*, **1**, 19–30.
Matthiopoulos, J., Hebblewhite, M., Aarts, G. & Fieberg, J. (2011). Generalized functional responses for species distributions. *Ecology*, **92**, 583–589.
McGarigal, K., Zeller, K.A. & Cushman, S.A. (2016). Multi-scale habitat selection modeling: introduction to the special issue. *Landscape Ecology*, **31**, 1157–1160.
Mendenhall, C.D., Sekercioglu, C.H., Brenes, F.O., Ehrlich, P.R. & Daily, G.C. (2011). Predictive model for sustaining biodiversity in tropical countryside. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 16313–6.
Mendenhall, C.D., Shields-Estrada, A., Krishnaswami, A.J. & Daily, G.C. (2016). Quantifying and sustaining biodiversity in tropical agricultural landscapes. *Proceedings of the National Academy of Sciences*, 201604981.
Ovaskainen, O., Abrego, N., Halme, P. & Dunson, D. (2016). Using latent variable models to identify large networks of species-to-species associations at different spatial scales. *Methods in Ecology and Evolution*, **7**, 549–555.
Ovaskainen, O., Gleb Tikhonov, Norberg, A., Blanchet, F.G., Duan, L., Dunson, D., Roslin, T. & Abrego, N. (2017). How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecology Letters*, **20**, 561–576.
Ovaskainen, O. & Soininnen, J. (2011). Making more out of sparse data: hierarchical modeling of species communities. *Ecology*, **92**, 289–295.
Pacifici, K., Zipkin, E.F., Collazo, J.A., Irizarry, J.I. & Dewan, A. (2014). Guidelines for a priori grouping of species in hierarchical community models. *Ecology and Evolution*, **4**, 877–888.
Paton, R.S. & Matthiopoulos, J. (2016). Defining the scale of habitat availability for models of habitat selection. *Ecology*, **97**, 1113–1122.
Timm, B.C., McGarigal, K., Cushman, S.A. & Ganey, J.L. (2016). Multi-scale Mexican spotted owl (Strix occidentalis lucida) nest/roost habitat selection in Arizona and a comparison with single-scale modeling results. *Landscape Ecology*, **31**, 1209–1225.
de Villemereuil, P., Wells, J. a, Edwards, R.D. & Blomberg, S.P. (2012). Bayesian models for comparative analysis integrating phylogenetic uncertainty. *BMC*

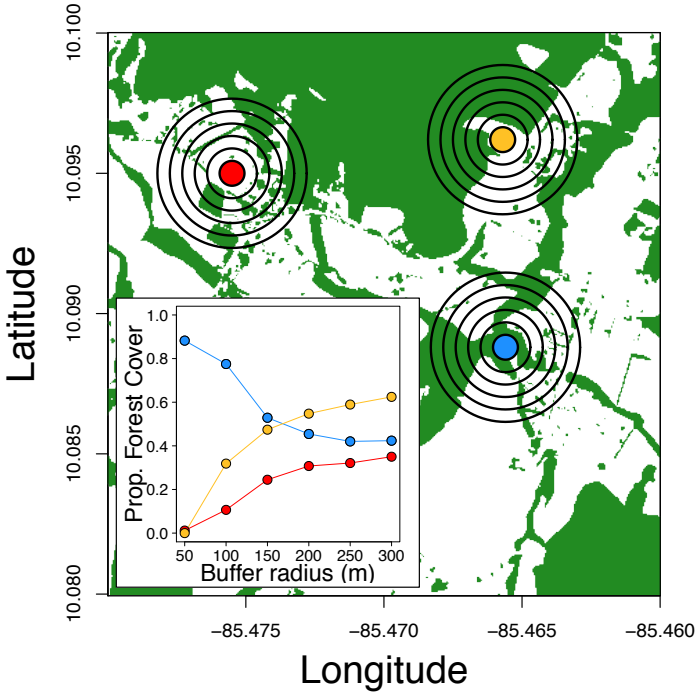
518 *evolutionary biology*, **12**, 102.
519 Warton, D.I., Blanchet, F.G., O'Hara, R.B., Ovaskainen, O., Taskinen, S., Walker, S.C.
520 & Hui, F.K.C. (2015). So Many Variables: Joint Modeling in Community Ecology.
521 *Trends in Ecology and Evolution*, **30**, 766–779.
522 Weaver, J.E., Conway, T.M. & Fortin, M.-J. (2012). An invasive species' relationship
523 with environmental variables changes across multiple spatial scales. *Landscape*
524 *Ecology*, **27**, 1351–1362.
525 Wiens, J.A. (1989). Spatial Scaling in Ecology. *Functional Ecology*, **3**, 385–397.
526 Yackulic, C.B., Reid, J., Nichols, J.D., Hines, J.E., Davis, R. & Forsman, E. (2014). The
527 roles of competition and habitat in the dynamics of populations and species
528 distributions. *Ecology*, **95**, 265–279.
529 Zuckerberg, B., Desrochers, A., Hochachka, W.M., Fink, D., Koenig, W.D. & Dickinson,
530 J.L. (2012). Overlapping landscapes: A persistent, but misdirected concern when
531 collecting and analyzing ecological data. *Journal of Wildlife Management*, **76**,
532 1072–1080.
533
534

Figures and Tables

Table 1: Simulation conditions of communities.

		Detection Regime				
Component	Parameter	Perfect	High	Low	Low-Env	Low-Env Variable
Detection	$\mu_{\alpha 0}$	100	0	-1	-1	-1
	$\sigma_{\alpha 0}$	0	1	1	1	1
	$\mu_{\alpha 1}$	0	0	0	1	1
	$\sigma_{\alpha 1}$	0	0	0	0	1
Occupancy	$\mu_{\beta 0}$	-1	-1	-1	-1	-1
	$\sigma_{\beta 0}$	1	1	1	1	1
	$\mu_{\beta 1}$	0.5	0.5	0.5	0.5	0.5
	$\sigma_{\beta 1}$	1	1	1	1	1
	$\sigma_{\gamma 0}$	0.1	0.1	0.1	0.1	0.1
	s	variable	variable	variable	variable	variable
Sample Size	*N*sp	16	16	16	16	16
	*N*site	50	50	50	50	50
	*N*visit	3	3	3	3	3

542

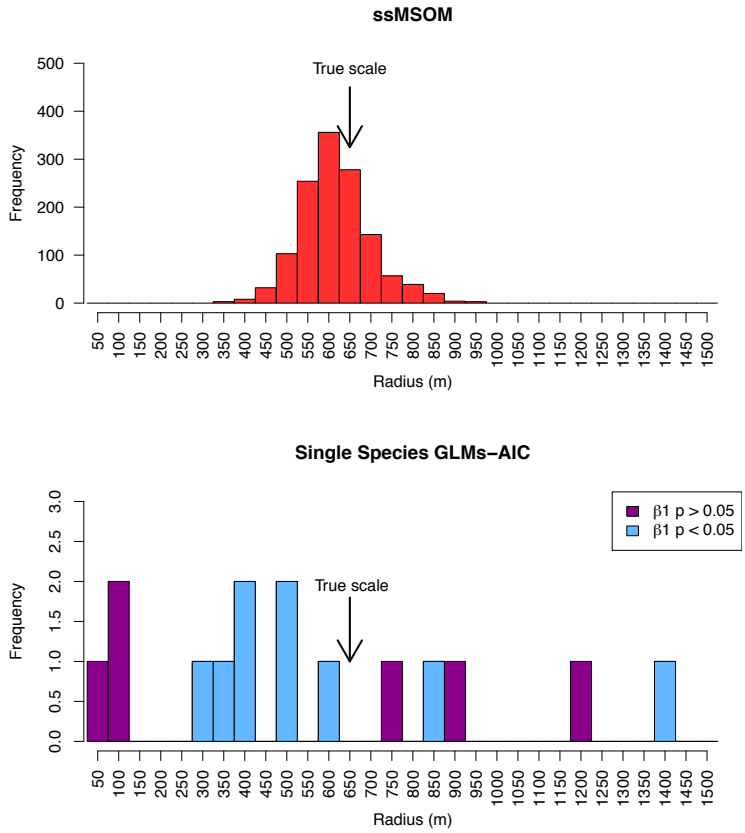


543

544 **Figure 1.** Map of empirical 5m-resolution subsection of forest cover landscape in
545 northwestern Costa Rica with three fictitious sampling points that break the correlation
546 structure between local and broader landscape forest cover. Colored points represent
547 50m-point count radii, and each successive buffer represents an increase in radius of
548 50m. Only 50-300m radii are shown, though for simulation analyses radii extended to
549 1500m.

550

551



552
553 **Figure 2.** Example histograms of single species GLMs, and the ssMSOM. Example
554 comes from low detection case (mean detection = ~0.25). Upper panel depicts posterior
555 distribution of the spatial scale that describes species responses to the environment from
556 the ssMSOM. Lower panel depicts the spatial scale that minimizes AICc for each single-
557 species GLM, after removing one species that was observed in fewer than 10% of sites
558 (*i.e.* 15 species remaining). Purple bars are species for which the response to the
559 environment does not differ significantly from 0, at the 'best' spatial scale, whereas blue
560 species have significantly positive or negative responses to the environment at the 'best'
561 spatial scale.
562

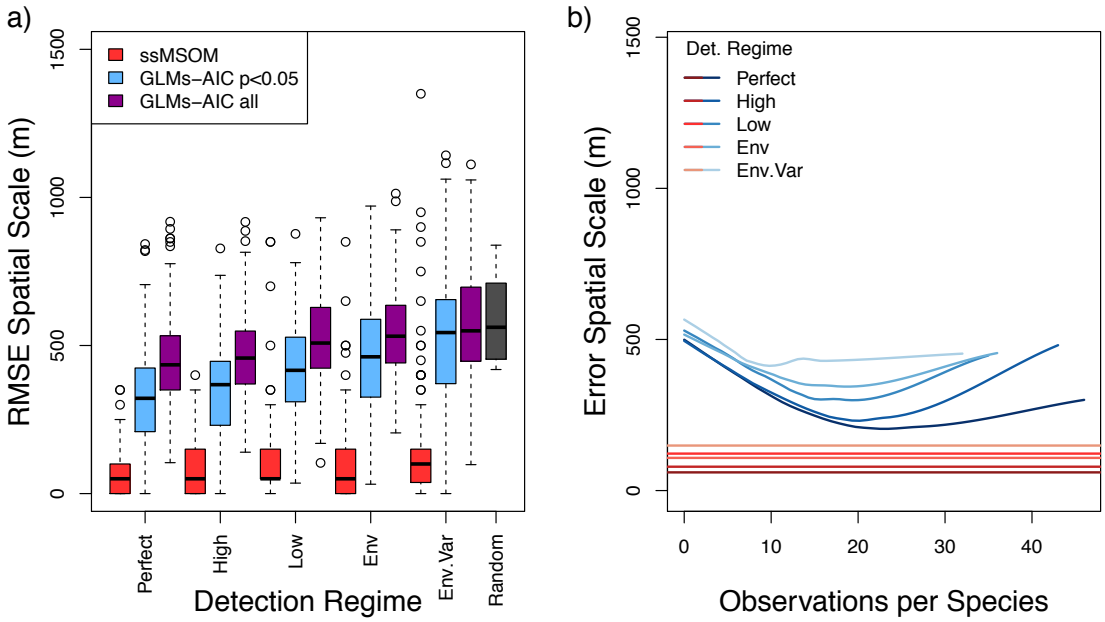


Figure 3. (a) General comparison of root mean square error for spatial scale across 120 simulations per detection regime, comparing posterior mode from ssMSOM, and AICc selected spatial scale for each species in single cases (splitting off only species with statistically significant responses to the environment [in blue], from all species [in purple]). 'Random' indicates the distribution of RMSEs that one would obtain if selecting spatial scales randomly along the uniform range from 50m to 1500m. (b) Blue lines in right panel depict lowess smoothers through all individual species' error in spatial scale estimation from piecemeal GLMs (across all simulated communities), as a function of the number of individuals observed. Red lines show mean error across all ssMSOMs (all species in the community are assumed to respond at the same scale, so species' level error is invariant to number of observations). Ability to estimate a species' scale of response suffers when species are either too rare, or too common, when using piecemeal GLMs.

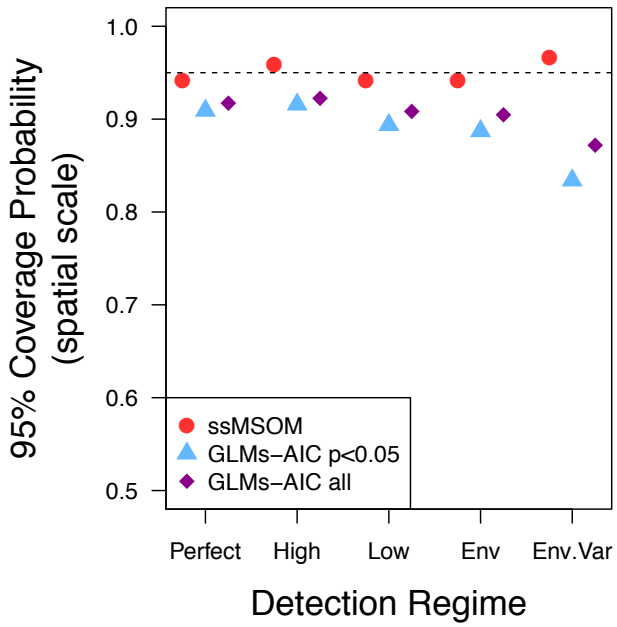


Figure 4. Quality of inference of true spatial scale declines for single species cases declines as detection becomes lower confounded with environment. For ssMSOM coverage probability indicates the proportion of simulations (n = 120 per regime) for which the true spatial scale was within the 95% CIs of the spatial scale posterior. For GLMs coverage probability indicates the proportion of species across simulations for which the true spatial scale was within the top 95% of Akaike weighted models.

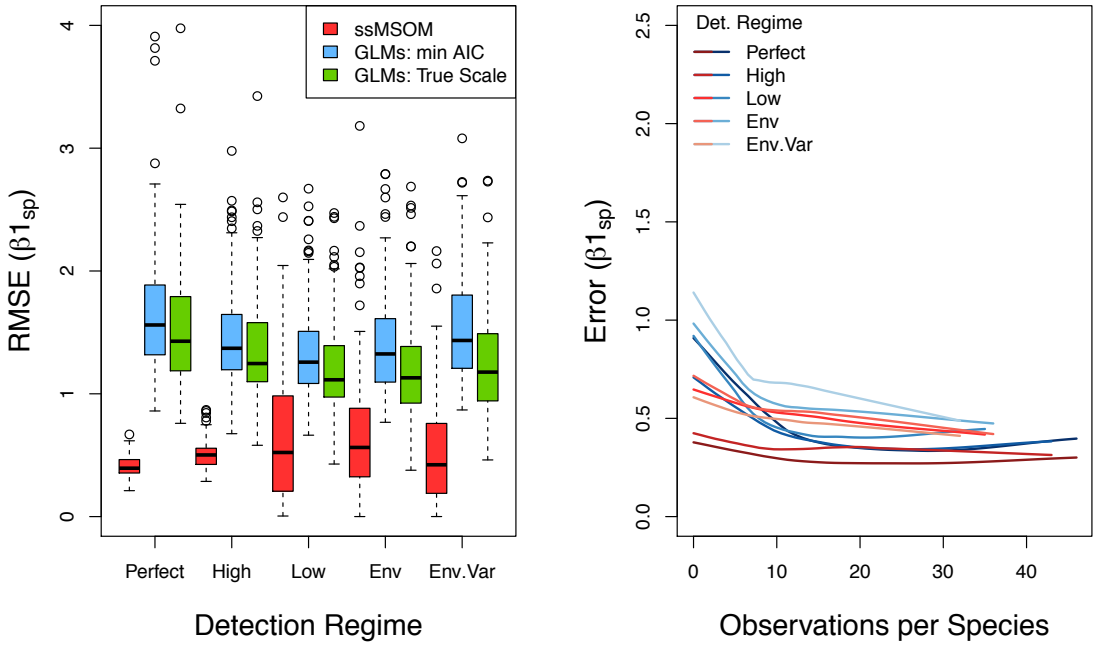


Figure 5. Accuracy of estimates of species responses to the environment ($\beta 1_{sp}$). Left panel depicts boxplots of community level RMSE of $\beta 1_{sp}$ values for each of 120 simulations per detection regime. Species observed in fewer than 10% of sites were removed from GLM estimates in left panel. Right panel depicts lowess smoothers through individual species' error in $\beta 1_{sp}$ (across all simulated communities), as a function of the number of individuals observed. Here blue lines represent estimates from piecemeal GLMs (using min. AICc), whereas red lines represent estimates from ssMSOMs.

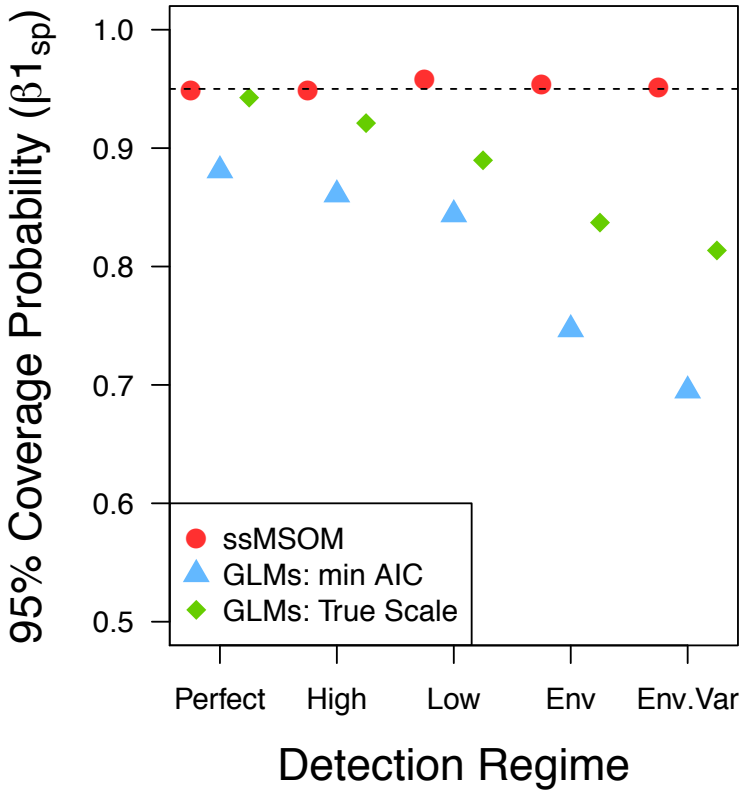


Figure 6. Coverage probabilities around true value of response to the environment ($\beta 1_{sp}$). Each point represents the proportion of species across all simulations for which the true value fell within the 95% CIs of the estimate. For GLMs all species observed in fewer than 10% of sites were excluded from the analysis. GLMs min AICc indicates each species' GLM at the 'best' spatial scale. GLM True scale is the GLM from at the spatial scale that the responses were simulated at, regardless of whether this GLM possessed the lowest AICc.