

¹ Towards causal occupancy models

² Maxwell B. Joseph (maxwell.b.joseph@colorado.edu), Dept. of Ecology and Evolutionary
³ Biology, University of Colorado, Boulder, CO, USA, 80309

⁴ Daniel L. Preston (daniel.preston@colorado.edu), Dept. of Ecology and Evolutionary Biol-
⁵ ogy, University of Colorado, Boulder, CO, USA, 80309

⁶ Pieter T. J. Johnson (pieter.johnson@colorado.edu), Dept. of Ecology and Evolutionary
⁷ Biology, University of Colorado, Boulder, CO, USA, 80309

⁸ Abstract

⁹ Introduction

¹⁰ Focus mostly on need for integrating SEM and occupancy models Describe typical approaches
¹¹ for each End with an overview of how they fit together

¹² Methods

¹³ Data collection

¹⁴ From 2009 to 2013, field crews used standardized methods to assess amphibian site occupancy
¹⁵ in 171 wetlands in the San Francisco Bay Area of California, USA. We surveyed wetlands
¹⁶ twice per summer using a combination of visual encounter, seine, and dipnet surveys (Crump
¹⁷ and Bury 1994). For each survey, we recorded all lifestages of each species of amphibian
¹⁸ that were observed. Sites were considered to be “occupied” if larval amphibians were present
¹⁹ in the site. This approach avoided problems associated with temporary non-breeding vis-
²⁰ itation of other habitats by vagile adults. To quantify grazing intensity, we recorded the
²¹ number of cow paddies within three meters of shoreline, and also recorded the length of the
²² shoreline with a handheld GPS unit. As a second measure of grazing intensity, we made
²³ a qualitative judgement of whether the wetland was disturbed by cattle. Further, at each
²⁴ site we collected water samples that were frozen in 8 oz Nalgene bottles adn later analyzed
²⁵ for ammonium (NH_4^+) and total nitrogen (N) concentrations using standard protocols (see:
²⁶ <http://snobear.colorado.edu/Kiowa/Kiowaref/procedure.html>).
²⁷ Lastly, we estimated the percentage of pond shoreline that was vegetated, and noted whether
²⁸ each wetland was permanent or ephemeral.

²⁹ Conceptual model

³⁰ Drawing upon previous literature on the impacts of cattle grazing (Kauffman et al. 1983,
³¹ Jansen and Healey 2003, Schmutzler et al. 2008, Adams et al. 2009, ???), we developed a
³² conceptual model to represent the different ways grazing could affect amphibian communities
³³ (Figure 1).

34 We hypothesized that livestock would alter wetland ecosystems through physical mechanisms
 35 including trampling and the removal of vegetation due to grazing, and chemical mechanisms,
 36 including inputs of nitrogenous waste products from urine and feces. In turn, these changes
 37 to wetland ecosystems were predicted to affect the occupancy states of pond-breeding am-
 38 phibians. Previous studies indicate that characteristics of shoreline vegetation influence
 39 amphibian breeding, and that water chemistry can alter reproductive success and occupancy
 40 probability (Freda and Dunson 1986, Rowe and Dunson 1995, Rouse et al. 1999, Jansen
 41 and Healey 2003, Brodman et al. 2003, Egan and Paton 2004, Burne and Griffin 2005, Arl
 42 and Hiteman 2009). Separating out the causal pathways underlying the multiple possible
 43 mechanisms through which grazing alters amphibian occupancy is important from a practical
 44 standpoint, as each one might be targeted differently with management interventions such as
 45 vegetation restoration, reductions in nutrient inputs, and partial vs. total cattle exclusion).

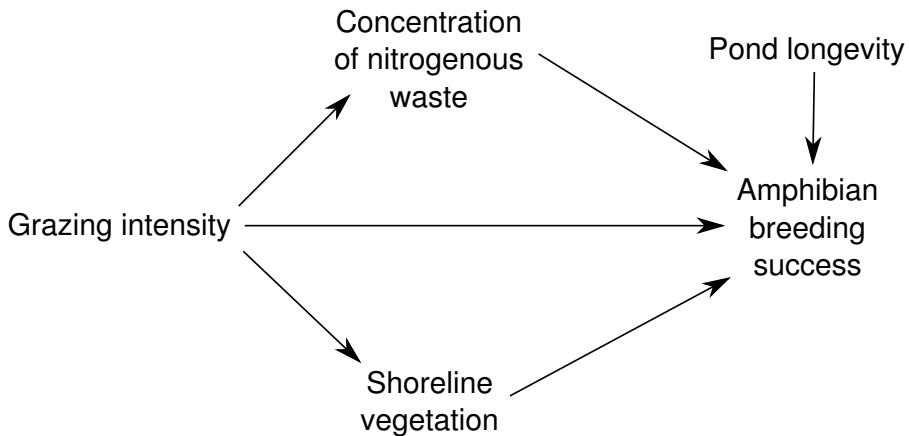


Figure 1: Conceptual model showing multiple causal pathways by which cattle grazing affects amphibians.

46 Model formalization

47 To formalize the conceptual model of Figure 1, we adopt a hierarchical Bayesian structural
 48 equation framework, in which we distinguish between observations, process, and parameter
 49 components of the model (Cressie et al. 2009).

50 The resulting model unites multi-species occupancy models with structural equation models,
 51 as a step towards more mechanistic models of species occurrence (Lee 2007, Dorazio et al.
 52 2010).

53 Data model

54 Traditional structural equation models assume normally distributed observed indicator vari-
 55 ables (Grace 2006). Conveniently, recent methodological advances facilitate the integration

56 of non-normal indicator variables, including those from the exponential family of distributions,
 57 and finite-mixture distributions, broadening the applicability of these methods to
 58 datasets common in ecology (Lee 2007). Here, qualitative observer judgments of whether
 59 sites were disturbed by cattle were modeled as a Bernoulli random variable with latent
 60 grazing intensity as a continuous covariate:

$$61 \quad Y_1[j, k] \sim Bern(p_{Y_1}[j]) \\ logit(p_{Y_1}[j]) = \beta_{Y_1,0} + \beta_{Y_1,1}\xi[j]$$

62 for the j^{th} site and the k^{th} survey. Observed cow paddy counts are treated as a second
 63 indicator (specifically, a multi-method indicator (Grace 2006)) and modeled as a Poisson
 64 random variable with an offset for shoreline perimeter of site j ($\mu_{perim}[j]$) in meters:

$$65 \quad Y_2[j, k] \sim Poisson(\lambda[j, k]) \\ \frac{\lambda_{j,k}}{\mu_{perim}[j, k]} = e^{\beta_{Y_2,0} + \beta_{Y_2,1}\xi[j]}$$

66 Because pond perimeter is measured with error, we model the true perimeter values as
 67 follows, with μ_{perim} representing the true pond perimeter value:

$$68 \quad log(perim[j, k]) \sim N(\mu_{perim}[j], \sigma_w) \\ \mu_{perim}[j] \sim N(\alpha_{perim}, \sigma_a)$$

69 where σ_w represents measurement error and variation within a season, α_{perim} represents the
 70 (log) average perimeter across all wetlands, and σ_a represents the variability in perimeter
 71 among wetlands.

72 Because shoreline vegetation, v , ranged from 0-100% with a non-negligible number of 0%,
 73 100%, and intermediate observations, we use a measurement model with a zero-one inflated
 74 beta response, which is a finite mixture distribution with a Bernoulli component that pro-
 75 duces 0's and 1's, and a beta component that produces values on the interval (0, 1) (Ospina
 76 and Ferrari 2012):

$$P(v; \alpha, \gamma_3, \mu_v, \phi_3) = \begin{cases} \alpha(1 - \mu_v) & v = 0 \\ \alpha\mu_v & v = 1 \\ (1 - \alpha)f(v; \mu_v, \phi) & 0 < v < 1 \end{cases}$$

$$logit(\mu_v[j, k]) = \beta_{v,0} + \beta_{v,1}\eta_1[j]$$

$$logit(\alpha[j, k]) = a_0 + a_1\eta_1[j] + a_2\eta_1[j]^2$$

77 Here, α determines the extent to which the beta or binomial mixture components dominate
 78 the probability density function. The second degree polynomial term with coefficient a_2

79 causes extreme values of the latent true value of shoreline vegetation cover η_1 to increase
 80 the probability of an observer recording either 0% or 100% shoreline vegetation cover. Last,
 81 $f(v; \mu_v, \phi)$ is the probability density function of the beta distribution, parameterized in terms
 82 of its mean and variance to facilitate covariate inclusion (Ospina and Ferrari 2012).
 83 Log-transformed concentrations of ammonium and total N in water were used as multi-
 84 method indicators of latent N concentration. Because both total N and NH_4^+ concentrations
 85 tend to increase over the course of the summer due to inputs and pond drying, we also
 86 include an effect of repeat survey number, coded as an indicator variable for the 2nd survey,
 87 i.e. $k \in 0, 1$.

$$\log(NH_4^+[j, k]) \sim N(\beta_{NH_4^+, 0} + \beta_{NH_4^+, 1}\eta_2[j] + \beta_{NH_4^+, 2}k, \sigma_{NH_4^+})$$

$$\log(N[j, k]) \sim N(\beta_{N, 0} + \beta_{N, 1}\eta_2[j] + \beta_{N, 2}k, \sigma_N)$$

88 We adopt an occupancy modeling approach for the data model describing amphibian detec-
 89 tion and non-detection. Observations of the i^{th} species at the j^{th} site on the k^{th} repeat survey
 90 are represented by $Y[i, j, k]$. We treat these observations as Bernoulli random variables with
 91 probability $p[i, j, k]z[i, j]$, where p is the probability of detection and z is the latent binary
 92 presence/absence state. As usual, z is only partly observed: if species i was seen at site j
 93 on any survey, it was present, but if it was not seen on any survey, it is possible that it was
 94 present but unobserved (MacKenzie et al. 2002):

$$Y[i, j, k] \sim Bernoulli(p[i, j, k]z[i, j])$$

95 We treat fish presence and pond permanence as directly observed quantities, which is sup-
 96 ported by consistency of observations both within and across years in this system.

97 Process model

98 Our process model relates the latent quantities: grazing intensity ξ , shoreline vegetation η_1 ,
 99 N concentration η_2 , the true occupancy states Z , and the probability of detection P . The
 100 relationships among latent quantities and their indicators is shown in figure 2.

101 We treat grazing intensity as an exogenous latent variable, unaffected by the other latent
 102 quantities, with mean 0 and standard deviation 1 as an identifiability constraint. Due to
 103 cattle grazing on and trampling shoreline vegetation, we model a linear effect of cattle grazing
 104 intensity on shoreline vegetation:

$$\eta_1 \sim N(\gamma_1\xi, 1)$$

105 where the standard deviation term, set to 1 for purposes of identifiability, represents the
 106 influence of other, unmodeled factors on shoreline vegetation cover. Nitrogenous inputs
 107 from cattle excretion in and around wetlands are treated similarly:

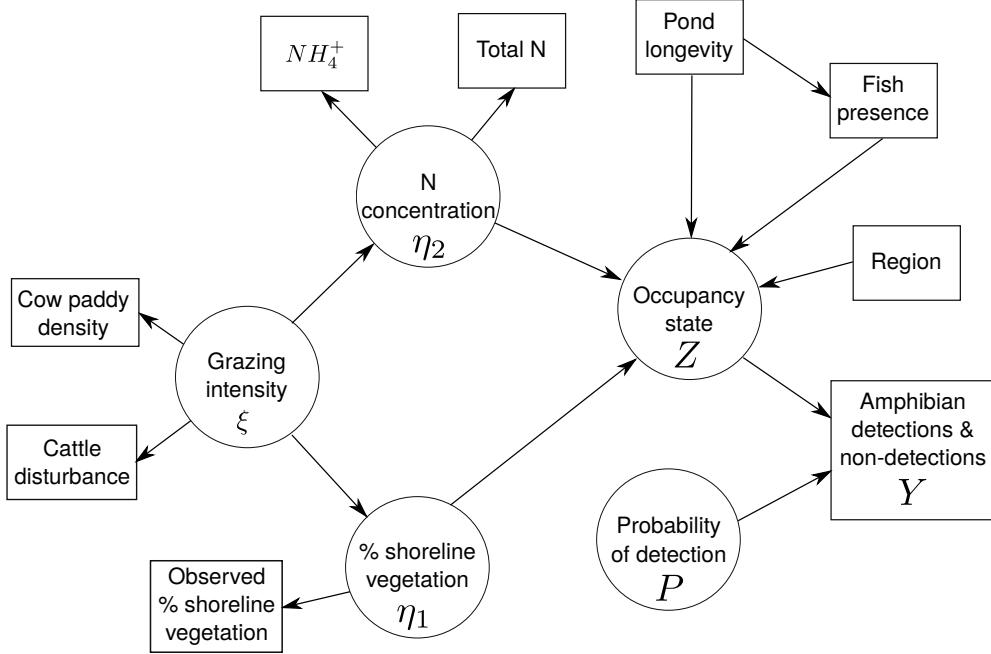


Figure 2: Directed acyclic graph illustrating relationships between unknown quantities (circles) and observed indicator variables (squares).

$$\eta_2 \sim N(\gamma_2 \xi, 1).$$

108 We assume that the true occupancy states are Bernoulli random variables with probability
 109 of occupancy $\psi[i, j]$ for the i^{th} species at the j^{th} site, such that $Z[i, j] \sim Bernoulli(\psi[i, j])$.
 110 We use a logit-link to model the effects of observed and latent covariates on ψ :

$$logit(\psi[i, j]) = \alpha_{species}[i] + \alpha_{refuge}[i, r[j]] + \alpha_{site}[i, j]$$

111 where $\alpha_{species}$ accounts for species specific differences in overall occupancy across the entire
 112 study area, α_{refuge} accounts for regional differences in occupancy rates within each species
 113 (with region r being indexed by site j). Last, α_{site} represents the local effects of fish, pond
 114 permanence, shoreline vegetation and nitrogenous compound concentrations. These terms
 115 can be decomposed as follows:

$$\alpha_{species}[i] \sim N(\mu_{\psi_{species}}, \sigma_{\alpha_{species}})$$

116 Here, $\alpha_{species}$ can be seen to be a variable intercept term, with values normally distributed
 117 around a community mean, and some among species variation in overall occupancy rates.

$$\alpha_{refuge}[i, j] \sim N(0, \sigma_{\alpha_{refuge}}[i])$$

118 This varying intercept term accounts for among-region variation within species in occupancy,
 119 with varying species-specific standard deviations, to account for the fact that the degree of
 120 regional variation in occupancy rates varies among species. Local covariate effects enter the
 121 final term:

$$\alpha_{site}[i, j] = N(\beta_\psi[i, 1]fish[j] + \beta_\psi[i, 2]perm[j] + \beta_\psi[i, 3]\eta_1[j] + \beta_\psi[i, 4]eta_2[j], \sigma_{\alpha_{site}})$$

122 Here, $\alpha_{site}[i, j]$ accounts for the effects of our occupancy covariates, and furthermore accounts
 123 for among-site variation from unmodeled factors, with the term $\sigma_{\alpha_{site}}$. This type of hierar-
 124 chical parameterization facilitates the decomposition of variance in occupancy into sources
 125 at multiple levels (e.g. species-specific, regional, and local).

126 We employ a fairly simple detection model, in which species vary in their detection prob-
 127 abilities, and there is also a difference in detection probabilities between first and second
 128 visits.

$$logit(p[i, j, k]) = \alpha_p[i] + \beta_p k$$

129 where α_p is a species-specific mean, and the last term represents the effect of early vs. late
 130 summer surveys.

131 Parameter model

132 We assumed that species responses to covariates with respect to occupancy and detection
 133 probabilities would be centered around 0 on a logit scale, with some variance that represents
 134 the variability in species responses, such that:

$$\begin{aligned} \beta_\psi &\sim N(0, \sigma_{\beta_\psi}) \\ \beta_p &\sim N(0, \sigma_{\beta_p}) \end{aligned}$$

136 Further, we assumed that for each species, landscape level rarity and mean detection proba-
 137 bilities would be logit-normally distributed around community level means (to be estimated):

$$\begin{aligned} \alpha_\psi &\sim N(\mu_{\alpha_\psi}, \sigma_{\alpha_\psi}) \\ \alpha_p &\sim N(\mu_{\alpha_p}, \sigma_{\alpha_p}) \end{aligned}$$

138 Hierarchical parameters corresponding to community-level variance received semi-
 139 informative half-Cauchy priors that were weighted towards small values to reduce bias
 140 (Gelman 2006). We adopt vague priors for all other parameters, except the loading terms
 141 for indicator variables which were constrained to be positive (e.g. increases in latent grazing
 142 intensity ξ correspond to increases in its indicators Y_1 and Y_2). Lastly, prior information
 143 based on previous work has implicitly entered the model in the form of missing effect
 144 pathways in structural equation component, e.g. we assume that amphibian community
 145 composition does not affect grazing intensity.

147 Parameter estimation

148 We used Just Another Gibbs Sampler ([JAGS](#)) to draw samples from the joint posterior
149 distribution of all parameters, with XX chains running for XX iterations, thinning by XX
150 after a XX iteration burn-in period (Plummer 2003). Convergence was assessed using visual
151 inspections of traceplots and the Gelman-Rubin potential scale reduction factor - chains were
152 considered to have converged when the estimated convergence diagnostics (R values) were
153 < 1.1 (Gelman and Rubin 1992, Brooks and Gelman 1998).

154 Model assessment

155 We assess model performance using the area under the receiver operating characteristic curve
156 (AUC), which estimates out of sample predictive ability (Zipkin et al. 2009). Because we do
157 not have out of sample data, we use K-fold cross validation.

158 We randomly subdivided our dataset into 3 sets (folds), and treated all amphibian
159 detection/non-detection data as missing parameters to be estimated, producing posterior
160 distributions for the probability of occupancy ψ for each species at each site that was in
161 each fold. These estimated occupancy probabilities are then compared to the posterior
162 distribution of true occupancy states generated by fitting a full model with all of the data
163 (Zipkin et al. 2009). This process is repeated for each fold generating three posterior
164 distributions for AUC that were averaged to generate a posterior distribution for the mean
165 AUC [citation]. Using 3-fold cross validation rather than a one-off holdout method is a
166 compromise between reducing bias associated with the random selection of a holdout set,
167 and increasing variance in AUC and computational cost [citation].

168 As this is a new method, we also explored the potential for systematic biases in parameter
169 estimates. We simulated 100 datasets with known structure that matches the structure
170 assumed by our model and attempted to recover the known parameters by fitting the model
171 (Gimenez et al. 2012). Parameter recovery was good, with relatively low levels of bias,
172 increasing our confidence in the method (Figure 3).

173 Results

174 Overall the structural equation component of the model performed well. The grazing sub-
175 model combined information from cow paddy density counts and disturbance classifications
176 to generate values of latent grazing intensity for each site (Figure 4). As we expected based
177 on field observations, many sites experience moderate to high levels of grazing, while fewer
178 experience very low levels of grazing.

179 Similarly, the shoreline vegetation submodel performed well, capturing the fact that some
180 shorelines are either devoid or completely covered by vegetation, with a much variability
181 between these two extremes (Figure 5).

182 Adequate fit was also observed for the latent N variable and observed log-transformed total
183 N and NH_4^+ concentrations, with increasing N concentrations in late summer compared to

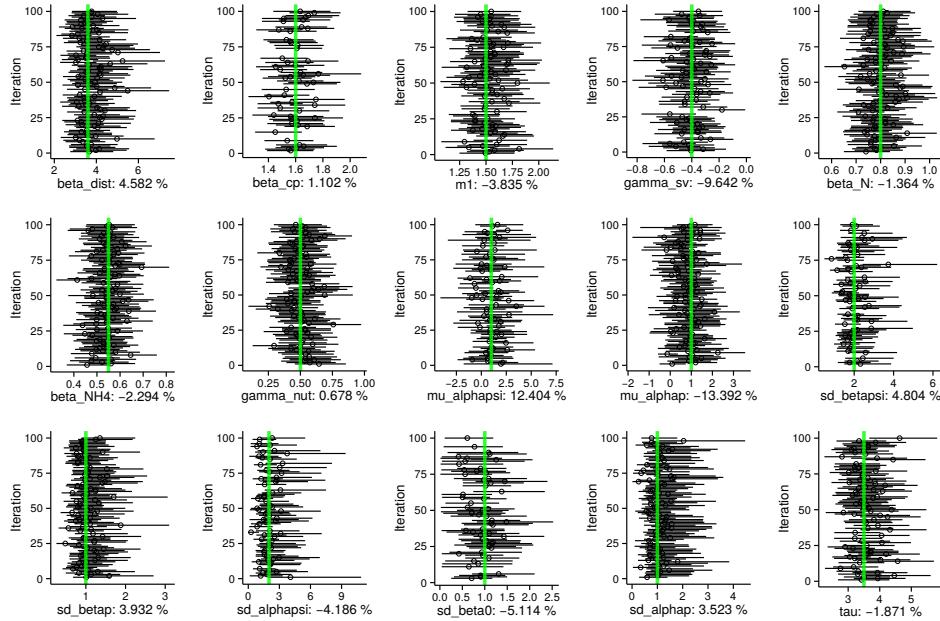


Figure 3: Results from a parameter recovery simulation in which 100 data sets were simulated from the model, and parameters were estimated. Percent bias represents the average deviation between the posterior medians and the true values for each parameter. Results are shown for MCMC results that converged.

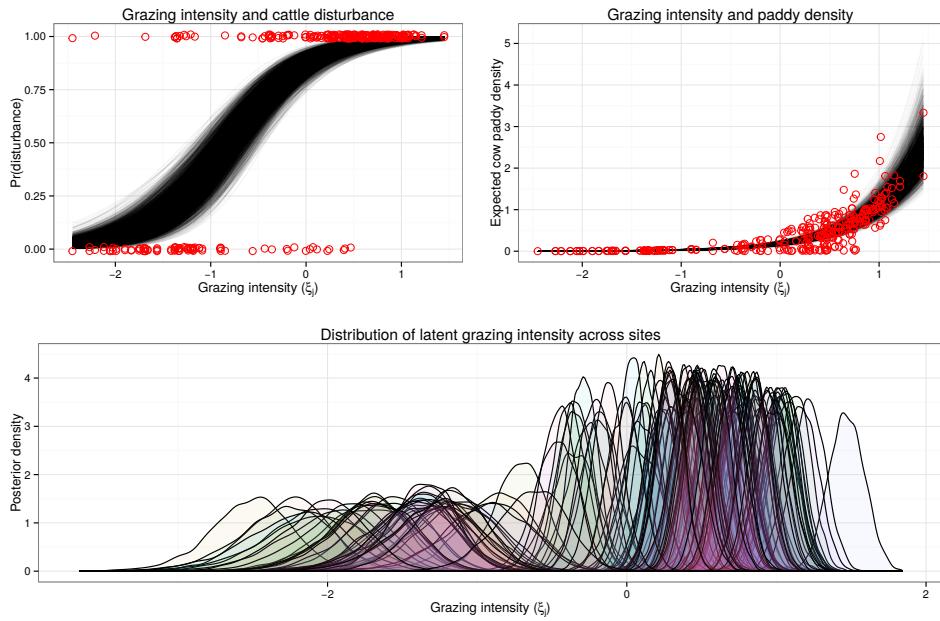


Figure 4: Fit of the grazing submodel to the observed cow paddy and disturbance data. In the upper panels, lines are drawn for each iteration in the posterior simulation, with observed data shown as jittered red points (x-axis values represent posterior modes). The lower panel shows latent grazing intensity posterior density estimates for each of the 172 sites.

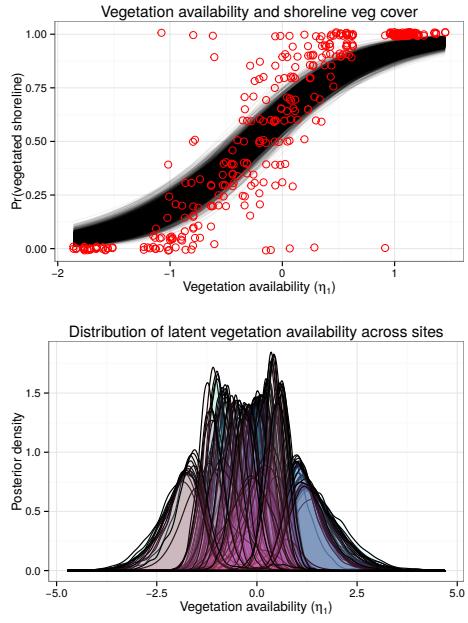


Figure 5: Fit of the shoreline vegetation submodel to observed cover data. A black line representing the expected value is drawn for each posterior draw, with observed data shown as open red points.

184 early summer (Figure 6).

185 Species-specific effects were overall consistent with expectations. Figure 7 shows the direct
 186 effects of grazing, shoreline vegetation, nutrients, and pond permanence, as well as plots
 187 that illustrate the overall effect of cattle grazing (incorporating the effect of grazing on
 188 vegetation and nutrients). Cattle grazing had weak positive direct effects on the occurrence
 189 of multiple native species, including California tiger salamanders (HDPI: -0.075 to 1.063),
 190 California red-legged frogs (HDPI: -0.075 to 0.843), and California newts (HDPI: -0.100
 191 to 0.805). Pond permanence had a strong positive effect on American bullfrog occurrence
 192 (HDPI: 0.128 to 3.926). Shoreline vegetation tended to have weak effects, positively affecting
 193 chorus frog occurrence (HDPI: -0.032 to 1.228). Finally, nutrient concentrations had weak
 194 positive effects on the occurrence of western toads (HDPI: -0.054 to 1.324) and California
 195 newts (-0.109 to 0.968).

196 Accounting for the effect of grazing on shoreline vegetation and nutrients allows us to illus-
 197 trate the total effect of grazing intensity on occupancy probability for each species. Overall,
 198 grazing intensity appears to have weak total effects on occurrence probabilities, with the
 199 exception of California newts, which show a positive response, most likely because they
 200 respond positively to grazing (directly) as well as nutrient concentrations, and these pos-
 201 tive responses are not directly canceled out by their slightly positive response to shoreline
 202 vegetation (Figure 7).

203 Shoreline vegetation had a positive influence on the detectability of chorus frogs (HDPI:

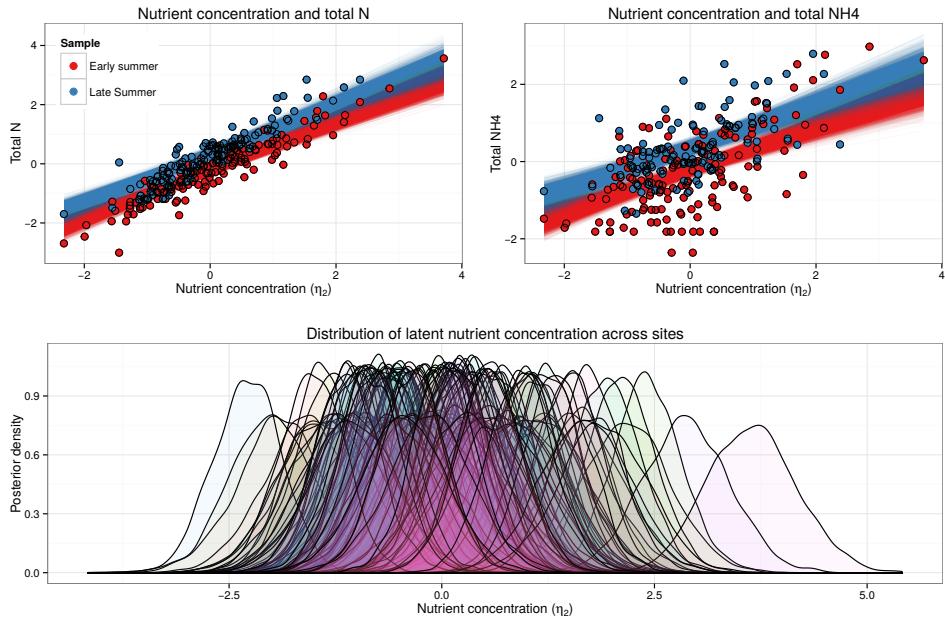


Figure 6: Fit of nitrogenous waste submodel to observed N data. The red lines and dots represent posterior expected values for early summer measurements; blue represents late summer measurements.

204 0.216 to 0.928), and the majority of species were easier to detect on the first, compared to
 205 the second visit in the summer (supplemental figure?).

206 • Species responses to grazing etc. (caterplot of effects + line plot showing increasing
 207 grazing & psi)

208 Expect positive effects of fish on bullfrogs because eggs are unpalatable, larvae are good at
 209 predator avoidance.

210 Discussion

211 Future methods incorporate spatiotemporal dynamics

212 Management implications

₂₁₃ Conclusion

₂₁₄ Acknowledgments

₂₁₅ References

- ₂₁₆ Adams, M. J., C. A. Pearl, B. McCreary, S. K. Galvan, S. J. Wessell, W. H. Wente, W.
₂₁₇ Chauncey, and A. B. Kuehl. 2009. Short-Term Effect of Cattle Exclosures on Columbia
₂₁₈ Spotted Frog (*Rana luteiventris*) Populations and Habitat in Northeastern Oregon Short-
₂₁₉ Term Effect of Cattle Exclosures on Columbia Spotted Frog (*Rana luteiventris*) Populations
₂₂₀ and Habitat in Northeast. *Journal of Herpetology* 43:132–138.
- ₂₂₁ Arl, J. U. E. E., and H. O. H. W. Hiteman. 2009. EFFECTS OF PULSED NITRATE
₂₂₂ EXPOSURE ON AMPHIBIAN DEVELOPMENT 28:1331–1337.
- ₂₂₃ Brodman, R., J. Ogger, T. Bogard, A. J. Long, R. a. Pulver, K. Mancuso, and D. Falk. 2003.
₂₂₄ Multivariate Analyses of the Influences of Water Chemistry and Habitat Parameters on the
₂₂₅ Abundances of Pond-Breeding Amphibians. *Journal of Freshwater Ecology* 18:425–436.
- ₂₂₆ Brooks, S., and A. Gelman. 1998. General methods for monitoring convergence of iterative
₂₂₇ simulations. *Journal of computational and graphical ...* 7:434–455.
- ₂₂₈ Burne, M. R., and C. R. Griffin. 2005. Habitat associations of pool-breeding amphibians in
₂₂₉ eastern Massachusetts, USA. *Wetlands Ecology and Management* 13:247–259.
- ₂₃₀ Cressie, N., C. a Calder, J. S. Clark, J. M. Ver Hoef, and C. K. Wikle. 2009. Accounting
₂₃₁ for uncertainty in ecological analysis: the strengths and limitations of hierarchical statisti-
₂₃₂ cal modeling. *Ecological applications : a publication of the Ecological Society of America*
₂₃₃ 19:553–70.
- ₂₃₄ Crump, M., and R. Bury. 1994. Visual encounter surveys. Pages 84–92 in W. Heyer, M.
₂₃₅ Donnelly, R. McDiarmid, L. Hayek, and M. Foster, editors. *Measuring and monitoring biolog-*
₂₃₆ *ical diversity, standard methods for amphibians.* Smithsonian Institution Press, Washington
₂₃₇ DC.
- ₂₃₈ Dorazio, R. M., M. Kéry, J. A. Royle, and M. Plattner. 2010. Models for inference in
₂₃₉ dynamic metacommunity systems. *Ecology* 91:2466–75.
- ₂₄₀ Egan, R., and P. Paton. 2004. Within-pond parameters affecting oviposition by wood frogs
₂₄₁ and spotted salamanders. *Wetlands* 24:1–13.
- ₂₄₂ Freda, J., and W. Dunson. 1986. Effects of low pH and other chemical variables on the local
₂₄₃ distribution of amphibians. *Copeia* 1986:454–466.
- ₂₄₄ Gelman, A. 2006. Prior distributions for variance parameters in hierarchical models.
₂₄₅ *Bayesian Analysis* 1:515–533.
- ₂₄₆ Gelman, A., and D. B. Rubin. 1992. Inference from iterative simulation using multiple
₂₄₇ sequences. *Statistical science* 7:457–511.

- 248 Gimenez, O., T. Anker-Nilssen, and V. Grosbois. 2012. Exploring causal pathways in
249 demographic parameter variation: path analysis of mark-recapture data. *Methods in Ecology*
250 and *Evolution* 3:427–432.
- 251 Grace, J. B. 2006. Structural Equation Modeling and Natural Systems. Page 378. Cam-
252 bridge Univ Press.
- 253 Jansen, A., and M. Healey. 2003. Frog communities and wetland condition: relationships
254 with grazing by domestic livestock along an Australian floodplain river. *Biological Conser-*
255 *vation* 109:207–219.
- 256 Kauffman, J., W. Krueger, and M. Vavra. 1983. Effects of late season cattle grazing on
257 riparian plant communities. *Journal of Range Management* 36:685–691.
- 258 Lee, S. 2007. Structural Equation Modeling: A Bayesian Approach. Wiley.
- 259 MacKenzie, D., J. Nichols, and G. Lachman. 2002. Estimating site occupancy rates when
260 detection probabilities are less than one. *Ecology* 83:2248–2255.
- 261 Ospina, R., and S. L. Ferrari. 2012. A general class of zero-or-one inflated beta regression
262 models. *Computational Statistics & Data Analysis* 56:1609–1623.
- 263 Plummer, M. 2003. JAGS: A program for analysis of Bayesian graphical models using Gibbs
264 sampling.
- 265 Rouse, J. D., C. a Bishop, and J. Struger. 1999. Nitrogen pollution: an assessment of its
266 threat to amphibian survival. *Environmental health perspectives* 107:799–803.
- 267 Rowe, C., and W. Dunson. 1995. Impacts of hydroperiod on growth and survival of larval
268 amphibians in temporary ponds of central Pennsylvania, USA. *Oecologia* 102:397–403.
- 269 Schmutz, A. C., M. J. Gray, E. C. Burton, and D. L. Miller. 2008. Impacts of cattle on
270 amphibian larvae and the aquatic environment. *Freshwater Biology* 53:2613–2625.
- 271 Zipkin, E. F., A. DeWan, and J. Andrew Royle. 2009. Impacts of forest fragmentation
272 on species richness: a hierarchical approach to community modelling. *Journal of Applied*
273 *Ecology* 46:815–822.