Towards causal occupancy models

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

Daniel L. Preston (daniel.preston@colorado.edu), Dept. of Ecology and Evolutionary Biology, University of Colorado, Boulder, CO, USA, 80309

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

Abstract

Introduction

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

Methods

Data collection

From 2009 to 2013, field crews used standardized methods to assess amphibian site occupancy in 171 wetlands in the San Francisco Bay Area of California, USA. We surveyed wetlands twice per summer using a combination of visual encounter, seine, and dipnet surveys (Crump and Bury 1994). For each survey, we recorded all lifestages of each species of amphibian that were observed. Sites were considered to be "occupied" if larval amphibians were present in the site. This approach avoided problems associated with temporary non-breeding visitation of other habitats by vagile adults. To quantify grazing intensity, we recorded the number of cow paddies within three meters of shoreline, and also recorded the length of the shoreline with a handheld GPS unit. As a second measure of grazing intensity, we made a qualitative judgement of whether the wetland was disturbed by cattle. Further, at each site we collected water samples that were frozen in 8 ox Nalgene bottles adn later analyzed for ammonium ($NH4^+$) and total nitrogen (N) concentrations using standard protocols (see: http://snobear.colorado.edu/Kiowa/Kiowaref/procedure.html).
Lastly, we estimated the percentage of pond shoreline that was vegetated, and noted whether each wetland was permanent or ephemeral.

Conceptual model

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

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

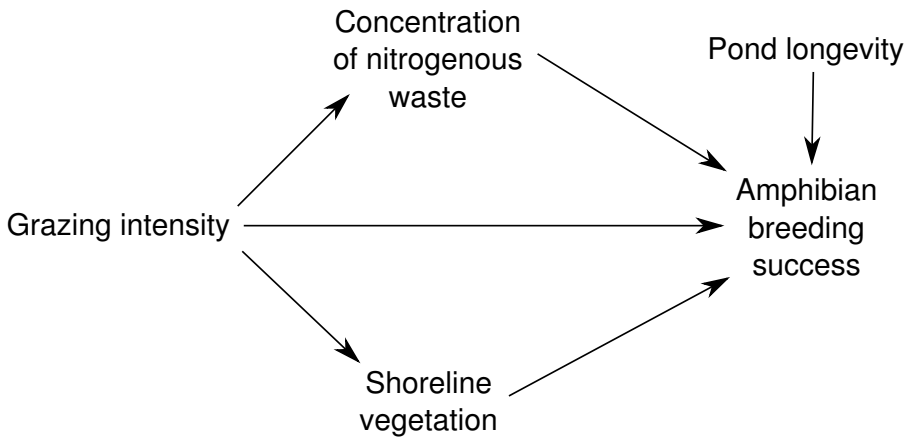

Figure 1: Conceptual model showing multiple causal pathways by which cattle grazing affects amphibians.

Model formalization

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

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

Data model

Traditional structural equation models assume normally distributed observed indicator variables (Grace 2006). Conveniently, recent methodological advances facilitate the integration

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

$$\begin{aligned} Y_1[j,k] &\sim Bern(p_{Y_1}[j]) \\ logit(p_{Y_1}[j]) &= \beta_{Y_1,0} + \beta_{Y_1,1}\xi[j] \end{aligned}$$

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

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

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

$$\begin{aligned} log(perim[j,k]) &\sim N(\mu_{perim}[j], \sigma_w) \\ \mu_{perim}[j] &\sim N(\alpha_{perim}, \sigma_a) \end{aligned}$$

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

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

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

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

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

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

causes extreme values of the latent true value of shoreline vegetation cover η_1 to increase the probability of an observer recording either 0% or 100% shoreline vegetation cover. Last, $f(v; \mu_v, \phi)$ is the probability density function of the beta distribution, parameterized in terms of its mean and variance to facilitate covariate inclusion (Ospina and Ferrari 2012).

Log-transformed concentrations of ammonium and total N in water were used as multi-method indicators of latent N concentration. Because both total N and NH_4^+ concentrations tend to increase over the course of the summer due to inputs and pond drying, we also include an effect of repeat survey number, coded as an indicator variable for the 2nd survey, i.e. $k \in 0, 1$.

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

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

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

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

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

Process model

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

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

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

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

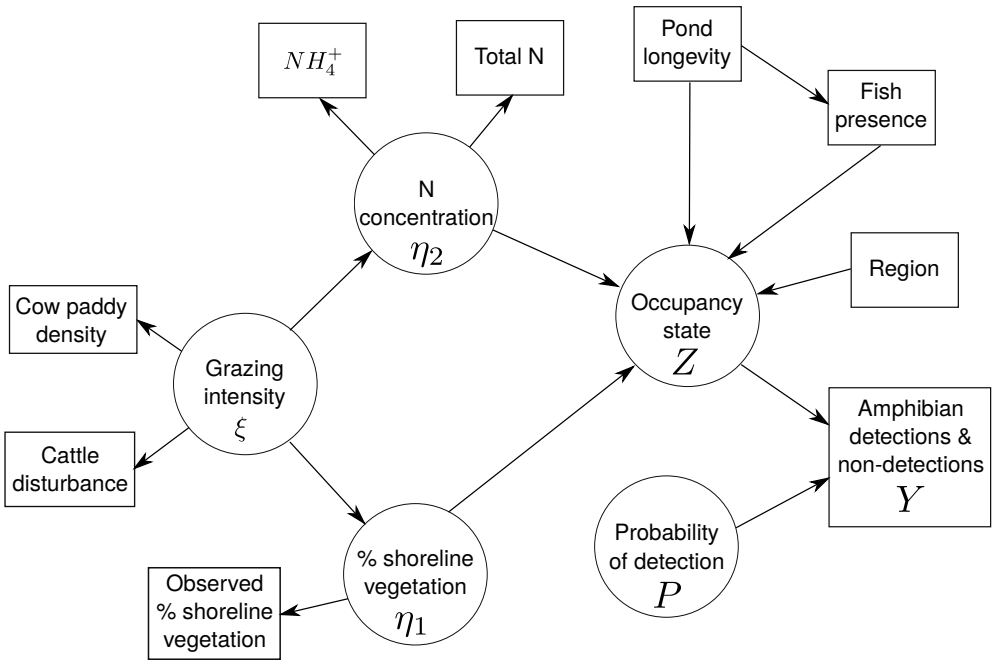

Figure 2: Directed acyclic graph illustrating relationships between unknown quantities (circles) and observed indicator variables (squares).

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

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

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

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

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

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

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

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

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

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

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

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

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

Parameter model

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

$$\beta_{\psi,} \sim N(0, \sigma_{\beta_{\psi}})$$
$$\beta_p \sim N(0, \sigma_{\beta_p})$$

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

$$\alpha_{\psi} \sim N(\mu_{\alpha_{\psi}}, \sigma_{\alpha_{\psi}})$$
$$\alpha_p \sim N(\mu_{\alpha_p}, \sigma_{\alpha_p})$$

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

Parameter estimation

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

Model assessment

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

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

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

Results

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

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

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

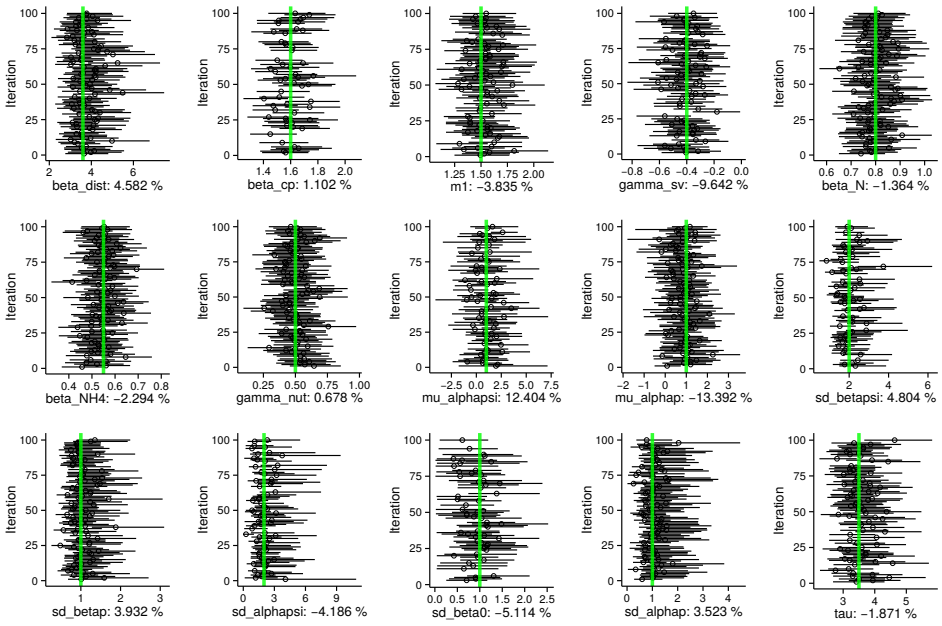

Figure 3: Results from a parameter recovery simulation in which 100 data sets were simulated from the model, and parameters were estimated. Percent bias represents the average deviation between the posterior medians and the true values for each parameter. Results are shown for MCMC results that converged.

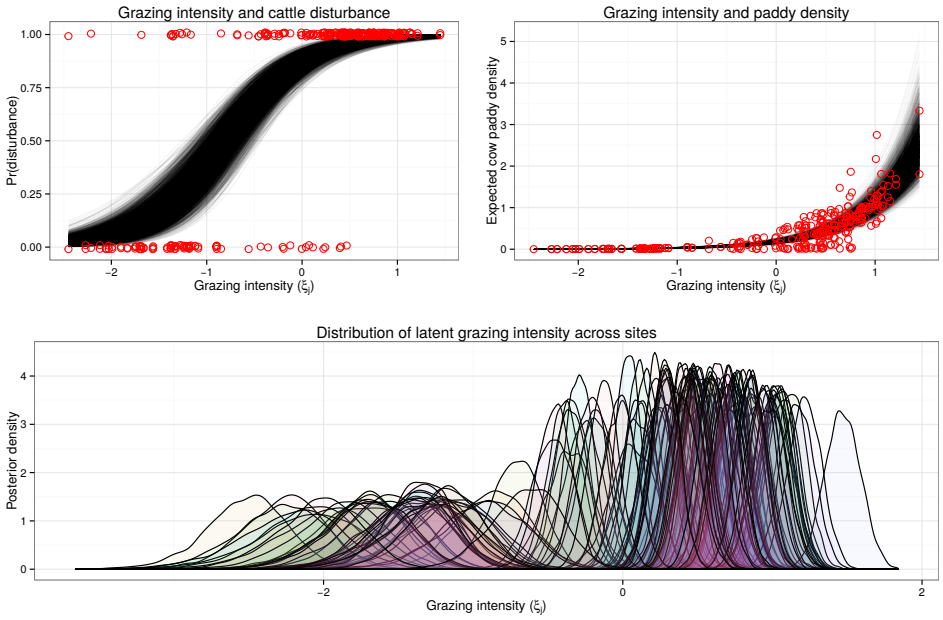

Figure 4: Fit of the grazing submodel to the observed cow paddy and disturbance data. In the upper panels, lines are drawn for each iteration in the posterior simulation, with observed data shown as jittered red points (x-axis values represent posterior modes). The lower panel shows latent grazing intensity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

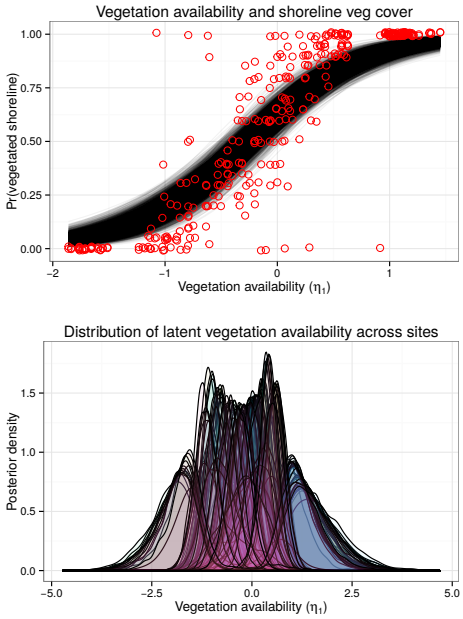

Figure 5: Fit of the shoreline vegetation submodel to observed cover data. A black line representing the expected value is drawn for each posterior draw, with observed data shown as open red points.

early summer (Figure 6).

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

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

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

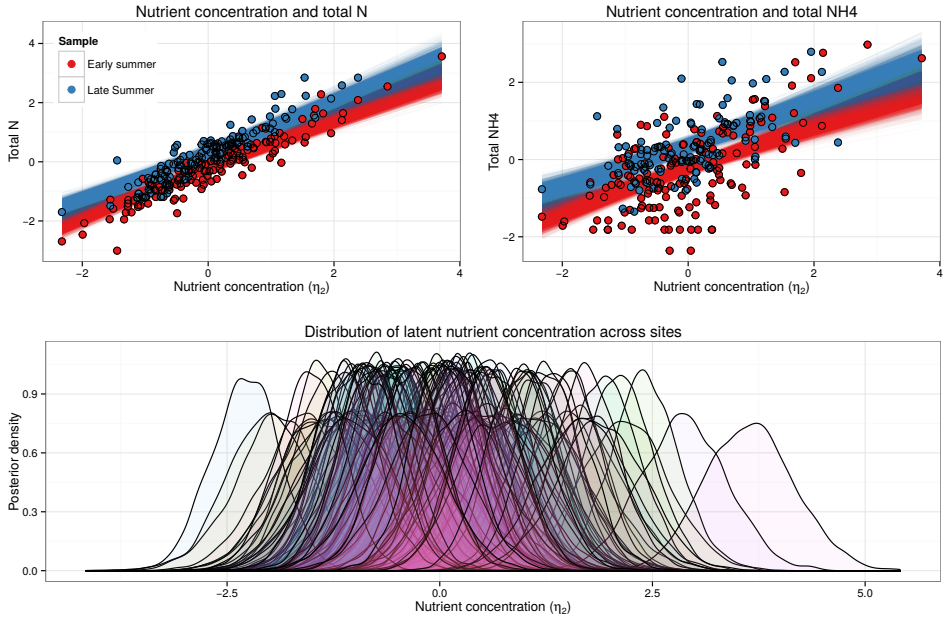

Figure 6: Fit of nitrogenous waste submodel to observed N data. The red lines and dots represent posterior expected values for early summer measurements; blue represents late summer measurements.

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

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

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

Discussion

Future methods incorporate spatiotemporal dynamics

Management implications

Conclusion

Acknowledgments

References

Adams, M. J., C. A. Pearl, B. Mccreary, S. K. Galvan, S. J. Wessell, W. H. Wente, W. Chauncey, and A. B. Kuehl. 2009. Short-Term Effect of Cattle Exclosures on Columbia Spotted Frog (Rana luteiventris) Populations and Habitat in Northeastern Oregon Short-Term Effect of Cattle Exclosures on Columbia Spotted Frog (Rana luteiventris) Populations and Habitat in Northeast. Journal of Herpetology 43:132–1338.

Arl, J. U. E. E., and H. O. H. W. Hiteman. 2009. EFFECTS OF PULSED NITRATE EXPOSURE ON AMPHIBIAN DEVELOPMENT 28:1331–1337.

Brodman, R., J. Ogger, T. Bogard, A. J. Long, R. a. Pulver, K. Mancuso, and D. Falk. 2003. Multivariate Analyses of the Influences of Water Chemistry and Habitat Parameters on the Abundances of Pond-Breeding Amphibians. Journal of Freshwater Ecology 18:425–436.

Brooks, S., and A. Gelman. 1998. General methods for monitoring convergence of iterative simulations. Journal of computational and graphical … 7:434–455.

Burne, M. R., and C. R. Griffin. 2005. Habitat associations of pool-breeding amphibians in eastern Massachusetts, USA. Wetlands Ecology and Management 13:247–259.

Cressie, N., C. a Calder, J. S. Clark, J. M. Ver Hoef, and C. K. Wikle. 2009. Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. Ecological applications : a publication of the Ecological Society of America 19:553–70.

Crump, M., and R. Bury. 1994. Visual encounter surveys. Pages 84–92 in W. Heyer, M. Donnely, R. McDiarmid, L. Hayek, and M. Foster, editors. Measuring and monitoring biological diversity, standard methods for amphibians. Smithsonian Institution Press, Washington DC.

Dorazio, R. M., M. Kéry, J. A. Royle, and M. Plattner. 2010. Models for inference in dynamic metacommunity systems. Ecology 91:2466–75.

Egan, R., and P. Paton. 2004. Within-pond parameters affecting oviposition by wood frogs and spotted salamanders. Wetlands 24:1–13.

Freda, J., and W. Dunson. 1986. Effects of low pH and other chemical variables on the local distribution of amphibians. Copeia 1986:454–466.

Gelman, A. 2006. Prior distributions for variance parameters in hierarchical models. Bayesian Analysis 1:515–533.

Gelman, A., and D. B. Rubin. 1992. Inference from iterative simulation using multiple sequences. Statistical science 7:457–511.

Gimenez, O., T. Anker-Nilssen, and V. Grosbois. 2012. Exploring causal pathways in demographic parameter variation: path analysis of mark-recapture data. Methods in Ecology and Evolution 3:427–432.

Grace, J. B. 2006. Structural Equation Modeling and Natural Systems. Page 378. Cambridge Univ Press.

Jansen, A., and M. Healey. 2003. Frog communities and wetland condition: relationships with grazing by domestic livestock along an Australian floodplain river. Biological Conservation 109:207–219.

Kauffman, J., W. Krueger, and M. Vavra. 1983. Effects of late season cattle grazing on riparian plant communities. Journal of Range Management 36:685–691.

Lee, S. 2007. Structural Equation Modeling: A Bayesian Approach. Wiley.

MacKenzie, D., J. Nichols, and G. Lachman. 2002. Estimating site occupancy rates when detection probabilities are less than one. Ecology 83:2248–2255.

Ospina, R., and S. L. Ferrari. 2012. A general class of zero-or-one inflated beta regression models. Computational Statistics & Data Analysis 56:1609–1623.

Plummer, M. 2003. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling.

Rouse, J. D., C. a Bishop, and J. Struger. 1999. Nitrogen pollution: an assessment of its threat to amphibian survival. Environmental health perspectives 107:799–803.

Rowe, C., and W. Dunson. 1995. Impacts of hydroperiod on growth and survival of larval amphibians in temporary ponds of central Pennsylvania, USA. Oecologia 102:397–403.

Schmutzer, A. C., M. J. Gray, E. C. Burton, and D. L. Miller. 2008. Impacts of cattle on amphibian larvae and the aquatic environment. Freshwater Biology 53:2613–2625.

Zipkin, E. F., A. DeWan, and J. Andrew Royle. 2009. Impacts of forest fragmentation on species richness: a hierarchical approach to community modelling. Journal of Applied Ecology 46:815–822.