Final Myxozoan Data Analysis Notes/Results

**Samples:**

1267 samples from 31 sites and 4 years of data (2013-2016)

**Data handling:**

There were several site related issues that were dealt with. They are listed below.

Site related issues:

a.) Sites that had multiple visits in 1 year (only 2 in 2015) – only the largest sample was used (they were collected several weeks apart).

b.) Potomac 2016 sample- did not use because it is a mainstem site (don’t have info on nor did include other mainstem information)

c.) Left Unionville as 2 sites because of differences in NHD land use between site 1 and 2 (~7% different in percent ag)

d.) Harrisburg sampling was all considered 1 site although there may have been slight differences in sampling location up and downstream- Harrisburg includes anything categorized as Harrisburg or Rockville

e.) Low sample sizes were included (total less than 10= 5 sampling events)

f.) Schuylkill River samples- upon review of collection sheets, it was determined that the 2015 sample was a composite from two sites. 15 out of 20 fish were from Port Clinton which has additional years of sampling data and 5 fish were from Berne. Berne and Port Clinton have very different land use. Berne has about 50% ag and Port Clinton has little to no ag. All samples were negative for myxozoans. The samples were split up with 15 and 5 to be accurate- separating out the Schuylkill sample does weaken the ag relationship some, but it is the most accurate for the samples and it has a large amount of uncertainty associated with the estimate

g.) Kettle Creek Samples – upon further review, it appears samples 1-8 were YOY and the second half of the sample were age 1, included 1-8

h.) Samples only included traditional samples collected during July and August (all other early or late samples were omitted)

i.) Conococheague Creek – confusion as to what site (s) samples were from. Appeared to be 2 in 2013, one in 2014, not sure how many in 2016. Talked to MD DNR (Michael Kashiwagi) said the samples were most likely a composite of 5 sites. Got land use info on 5 possible sites. Averaged and included all samples as one site due to management uncertainty (some sites were recorded as a specific site, but too much uncertainty based on management discussion). In general all sites were high ag.

Land Use Predictors: In addition all land use predictor variables were in proportions, logit transformed, and scaled before analysis,

-Note: logit had to remap ag because of 0 in one of the streams, remapped to 0.025 to 0.975

**Model code:**

###Model statement

sink("model6.txt")

cat("

model{

for(i in 1:n){

y[i]~dbin(p.bound[i],1)

p.bound[i]<-max(0, min(1, p[i]))

logit(p[i])<-Xbeta[i]

Xbeta[i]<- b.site[site[i]] + b.year[year[i]]

}

for(j in 1:n.site){

b.site[j]~dnorm(mu.hat[j],tau.site)

mu.hat[j] <- b.0 + b[1] \* agNHD[j] + b[2]\*devNHD[j] + b[3]\*agaccum[j] +b[4]\*devaccum[j]

}

b.0~dnorm(0,0.001)

# Bayesian LASSO - a Laplace (double exponential) prior

for(i in 1:4){

b[i] ~ ddexp(0, lambda1)

}

lambda1 ~ dexp(10)

for(j in 1:n.year){

b.year[j]~dnorm(0, tau.year)

}

tau.site<-pow(sigma.site,-2)

tau.year<-pow(sigma.year,-2)

sigma.site~dunif(0,100)

sigma.year~dunif(0,100)

}

",fill = TRUE)

sink()

#########Let's Run in JAGS

##RATIONALE FOR JAGS - want to be able to give probability along with relationship between myxo prevalance

#and different predictors

## Bundle data

data <- list(y=dat4$m, site=dat4$site,agNHD=agNHD, devNHD=devNHD,agaccum=agaccum,devaccum=devaccum,year=dat4$year.num,

n = nrow(dat4), n.site=length(unique(dat4$stream)),

n.year=length(unique(dat4$year)) )

# Create starting values

inits <- function(){list(b.0=rnorm(1),

sigma.site=runif(1), sigma.year=runif(1)

)}

# Parameters to estimate/track

parameters <- c("b.0","b.site","b","b.year","sigma.year", "sigma.site")

# MCMC setting

n.iter <- 90000

n.thin <- 3

n.burn <- 40000

n.chain <- 3

########## JAGS ANALYSIS FOR AG ###########

outall <- jags(data = data, inits = inits, parameters.to.save = parameters,

model.file = "model6.txt", n.chains = n.chain, n.thin = n.thin, n.iter = n.iter,

n.burnin = n.burn, parallel=T)

**R output:**

> print(outall, digits = 3)

JAGS output for model 'model6.txt', generated by jagsUI.

Estimates based on 3 chains of 90000 iterations,

burn-in = 40000 iterations and thin rate = 3,

yielding 50001 total samples from the joint posterior.

MCMC ran in parallel for 4.901 minutes at time 2018-02-07 20:17:03.

mean sd 2.5% 50% 97.5% overlap0 f Rhat n.eff

b.0 -0.399 0.390 -1.224 -0.393 0.332 TRUE 0.894 1.007 4053

b.site[1] -1.297 0.496 -2.316 -1.287 -0.368 FALSE 0.993 1.003 4489

b.site[2] -1.572 0.508 -2.622 -1.556 -0.636 FALSE 0.996 1.002 4038

b.site[3] 1.368 0.587 0.263 1.351 2.570 FALSE 0.990 1.002 5267

b.site[4] 0.275 0.551 -0.811 0.276 1.352 TRUE 0.705 1.002 3700

b.site[5] 0.318 0.451 -0.596 0.322 1.179 TRUE 0.790 1.004 5128

b.site[6] 1.384 0.475 0.450 1.379 2.318 FALSE 0.994 1.003 6649

b.site[7] 0.363 0.394 -0.453 0.368 1.096 TRUE 0.869 1.007 3156

b.site[8] -0.699 0.463 -1.642 -0.692 0.178 TRUE 0.947 1.004 2761

b.site[9] -0.777 0.758 -2.341 -0.756 0.660 TRUE 0.857 1.001 6535

b.site[10] -2.671 0.677 -4.113 -2.634 -1.456 FALSE 0.999 1.001 8763

b.site[11] -0.452 0.431 -1.311 -0.452 0.377 TRUE 0.884 1.005 3609

b.site[12] -1.837 0.604 -3.103 -1.814 -0.723 FALSE 0.997 1.001 37740

b.site[13] 0.426 0.413 -0.410 0.426 1.213 TRUE 0.886 1.005 3883

b.site[14] 0.326 0.429 -0.530 0.326 1.156 TRUE 0.813 1.005 3673

b.site[15] -0.383 0.408 -1.227 -0.373 0.374 TRUE 0.865 1.005 2394

b.site[16] 0.311 0.413 -0.530 0.315 1.099 TRUE 0.817 1.006 4233

b.site[17] -2.415 0.655 -3.818 -2.378 -1.234 FALSE 0.999 1.001 26012

b.site[18] -1.609 0.981 -3.701 -1.545 0.131 TRUE 0.965 1.000 22202

b.site[19] -0.835 0.564 -1.986 -0.818 0.228 TRUE 0.942 1.002 4549

b.site[20] -0.598 0.555 -1.729 -0.583 0.447 TRUE 0.876 1.002 4355

b.site[21] -0.212 0.398 -1.034 -0.205 0.525 TRUE 0.740 1.006 2771

b.site[22] -0.200 0.651 -1.500 -0.190 1.070 TRUE 0.623 1.001 4658

b.site[23] 0.413 0.492 -0.559 0.413 1.365 TRUE 0.823 1.003 2951

b.site[24] -0.771 0.412 -1.626 -0.764 -0.010 FALSE 0.976 1.005 3462

b.site[25] 0.298 0.458 -0.613 0.297 1.190 TRUE 0.771 1.004 3346

b.site[26] -0.428 0.385 -1.252 -0.418 0.282 TRUE 0.912 1.007 2669

b.site[27] 0.818 0.501 -0.160 0.812 1.805 TRUE 0.957 1.003 5088

b.site[28] -0.627 0.544 -1.737 -0.616 0.401 TRUE 0.893 1.002 3987

b.site[29] -1.542 0.682 -2.967 -1.512 -0.290 FALSE 0.990 1.001 6918

b.site[30] 0.538 0.648 -0.713 0.534 1.829 TRUE 0.805 1.001 13691

b.site[31] -0.291 0.399 -1.122 -0.283 0.453 TRUE 0.815 1.006 4100

b[1] 0.200 0.210 -0.210 0.200 0.615 TRUE 0.837 1.000 17293

b[2] -0.318 0.243 -0.805 -0.315 0.153 TRUE 0.911 1.000 20089

b[3] 0.241 0.258 -0.251 0.236 0.768 TRUE 0.831 1.000 22337

b[4] -0.185 0.297 -0.785 -0.179 0.389 TRUE 0.737 1.000 18702

b.year[1] 0.003 0.345 -0.650 -0.003 0.754 TRUE 0.494 1.011 2426

b.year[2] -0.294 0.343 -0.981 -0.280 0.409 TRUE 0.892 1.012 2186

b.year[3] 0.155 0.343 -0.468 0.135 0.937 TRUE 0.751 1.010 2132

b.year[4] 0.186 0.356 -0.438 0.155 1.002 TRUE 0.763 1.009 1887

sigma.year 0.504 0.523 0.092 0.356 1.838 FALSE 1.000 1.012 4940

sigma.site 1.055 0.204 0.719 1.033 1.516 FALSE 1.000 1.000 50001

deviance 1541.095 8.677 1525.988 1540.483 1559.807 FALSE 1.000 1.000 50001

Successful convergence based on Rhat values (all < 1.1).

Rhat is the potential scale reduction factor (at convergence, Rhat=1).

For each parameter, n.eff is a crude measure of effective sample size.

overlap0 checks if 0 falls in the parameter's 95% credible interval.

f is the proportion of the posterior with the same sign as the mean;

i.e., our confidence that the parameter is positive or negative.

DIC info: (pD = var(deviance)/2)

pD = 37.6 and DIC = 1578.739

DIC is an estimate of expected predictive error (lower is better).

>

> #Probability that the trend with ag posNHD

> pagNHD<- mean(outall$sims.list$b[,1] > 0)

> pagNHD

[1] 0.8369433

>

> pdevNHD <-mean(outall$sims.list$b[,2]<0)

> pdevNHD

[1] 0.9110818

>

> #Probability that the trend with NHD ag is positive

> pagaccum <- mean(outall$sims.list$b[,3] > 0)

> pagaccum

[1] 0.8313834

> #Probability that the trend with NHD development is negative

> pdevaccum <- mean(outall$sims.list$b[,4] < 0)

> pdevaccum

[1] 0.7372253

> ##overall intercept

> int<-plogis(outall$sims.list$b.0)

> mean(int)

[1] 0.404651

> quantile(int,0.025)

2.5%

0.2273194

> quantile(int,0.975)

97.5%

0.5821298

#Note: for summaries of year, intercept was added to account for how the model was coded (b.0 was included in the site level)

**Traceplot given in folder** –traceplot - not great for year components, but ran multiple time and get fairly consistent results (see rerun comp summary for more info) and convergence is indicated based on Rhat values. The model does not always converge, but I am able to get relatively consistent results when it does converge. Also compared to raw data and the trends and estimates seem sensitive enough to the actual data.

**Raw Data comparisons:**

Site comparison (observed and model estimates)

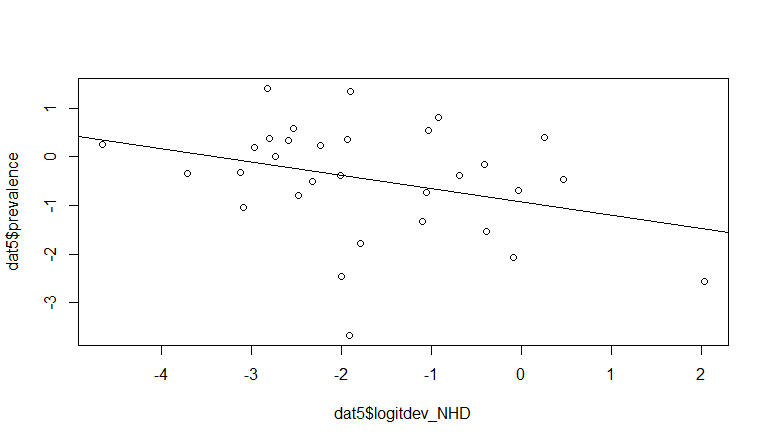
Ones with the largest deviation had characteristics that could be explained. For example, site 31 had low myxo prevalence, low sample size (n=5), and relatively high ag- higher estimate makes sense).

Year comparison (observed and model estimate)

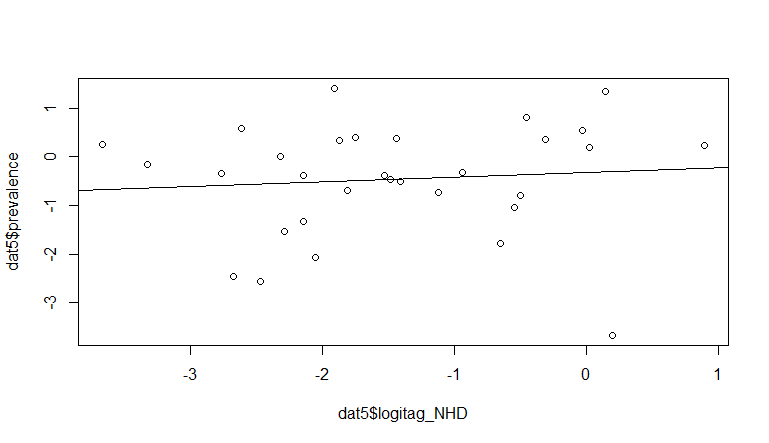
Some things to note: year 4 which has the largest deviation had the lowest sample size, year 3 which has the closest estimate had the largest sample size (although year 2 was also close in sample size), the general trend in the data is in line with what was observed ( higher vs. lower prevalence).

Generally compared ag and development and those trends also were in line with what was observed when using raw data. Logit transformed prevalence and land use variables

Development

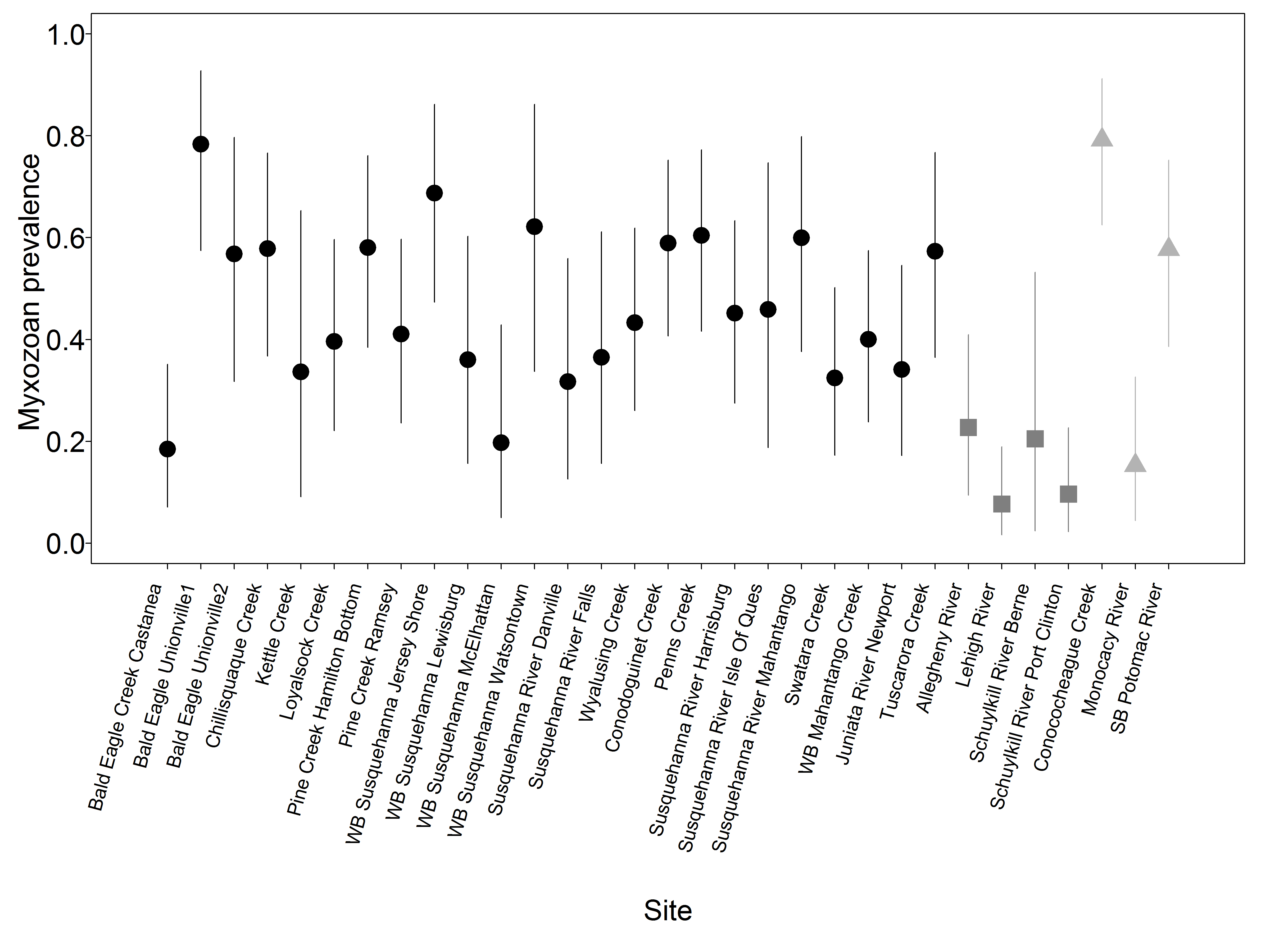


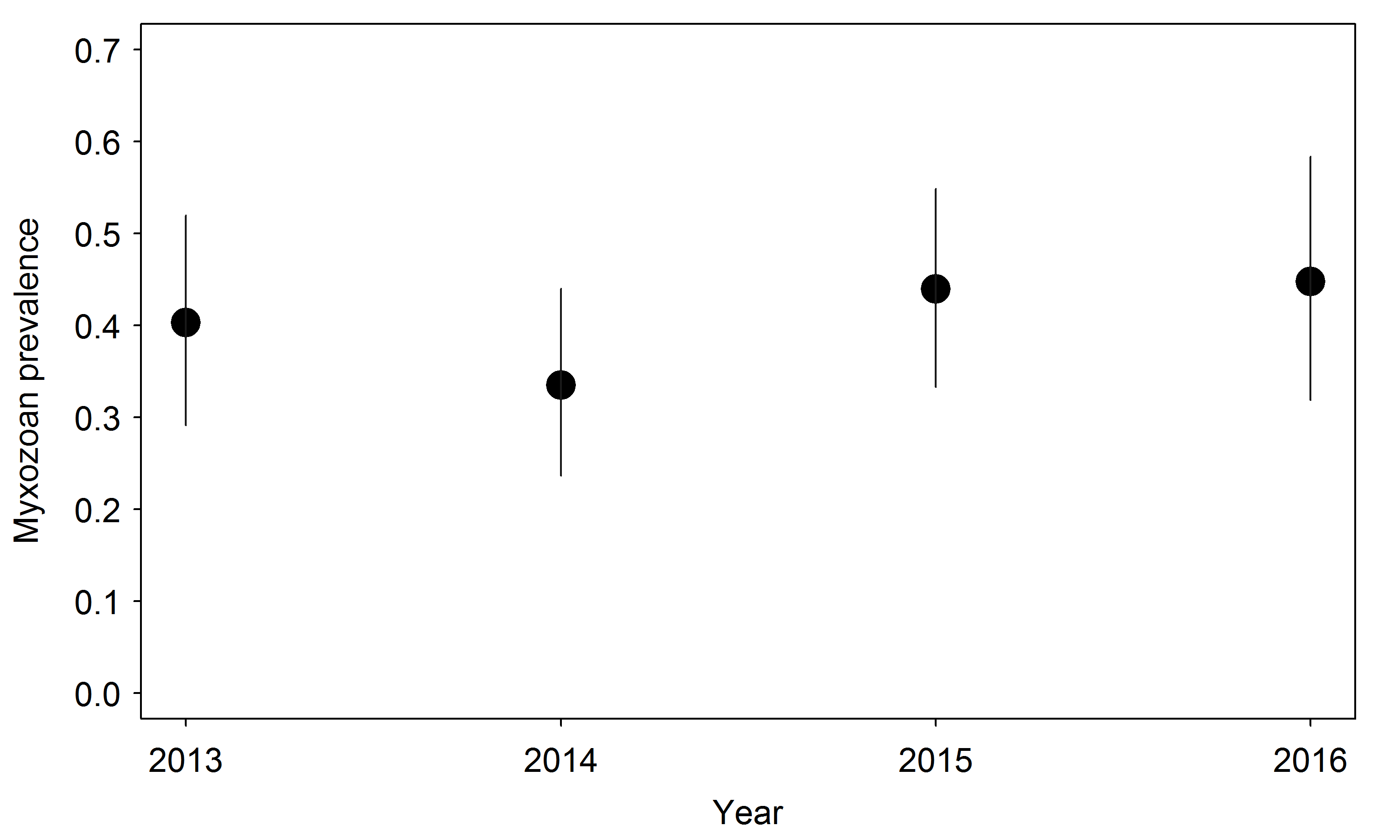
Agriculture

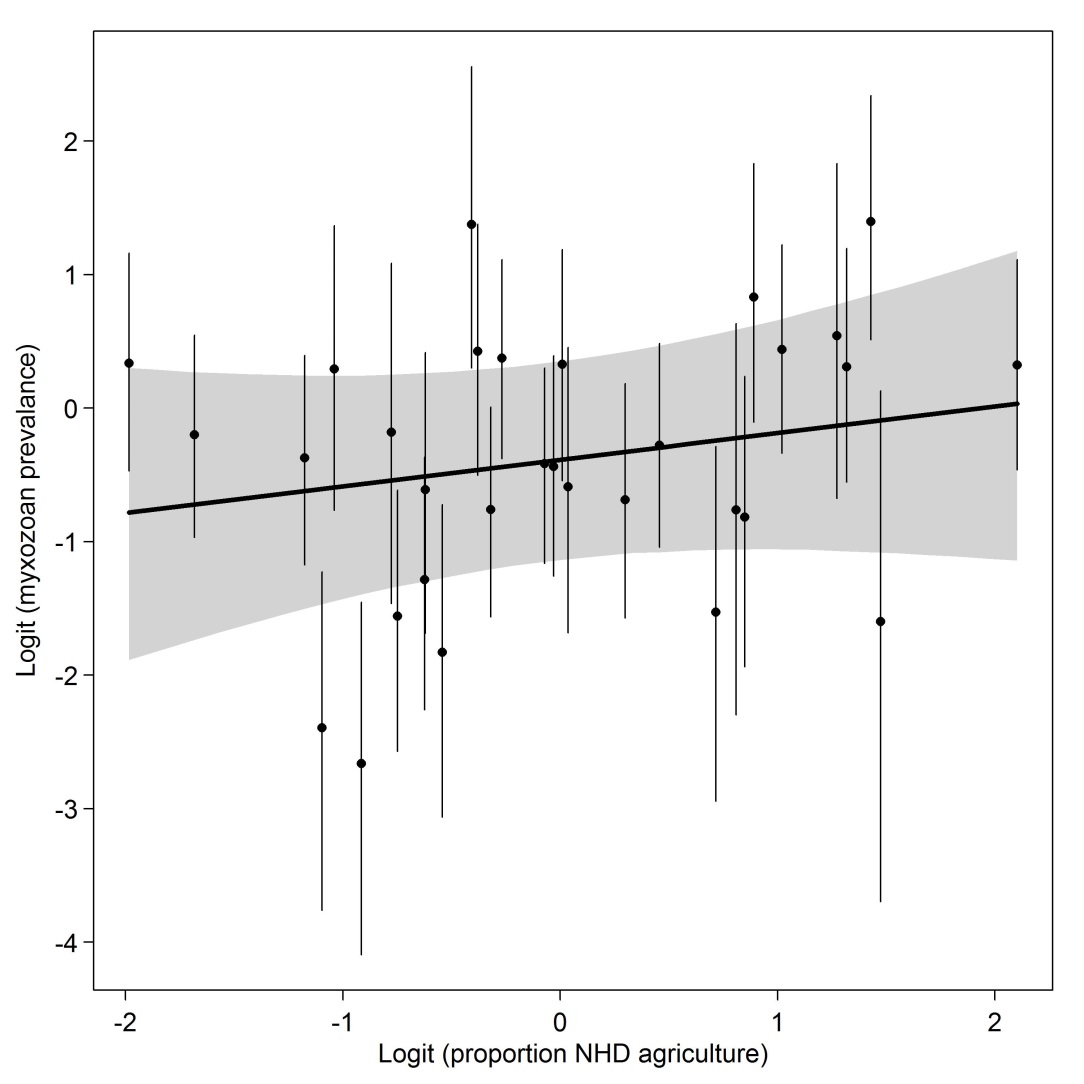


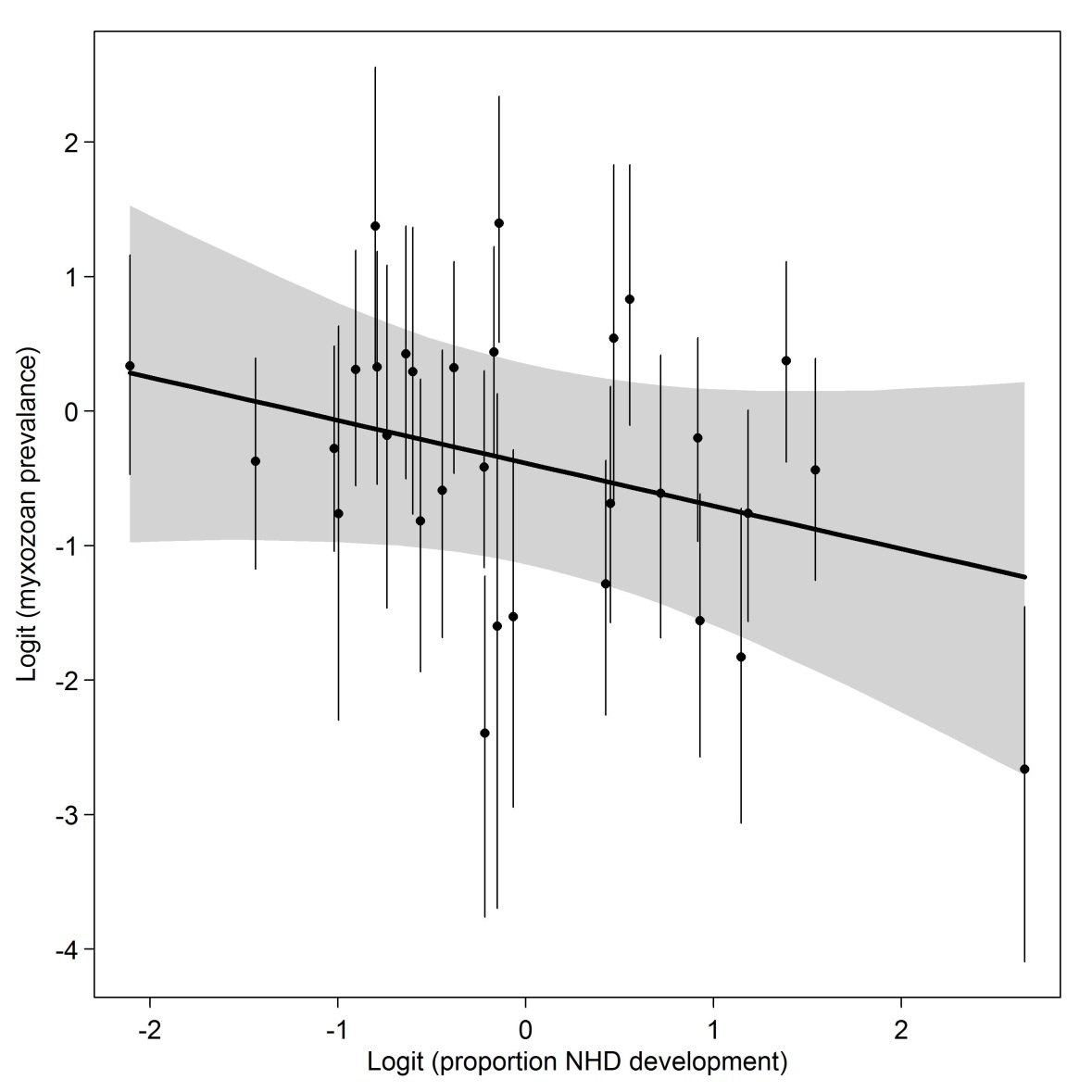
Good thing is the relationships are similar to what the model prediction look like.

**Final Plots:**

****

****

****

****