

1 Predicting biodiversity dynamics in response to environmental
2 change

3 Can we do it? A report from assess.sim.basic.R

4 Ryan Batt

5 2015-08-23

6 **Abstract**

7 “Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore
8 et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut
9 aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum
10 dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia
11 deserunt mollit anim id est laborum.”

¹² Contents

¹³	Introduction	3
¹⁴	Overview	3
¹⁵	The Simulation	3
¹⁶	Multispecies Occupancy Models (MSOMs)	4
¹⁷	Conventions and Settings	6
¹⁸	Dimension Conventions	6
¹⁹	Settings	7
²⁰	Species Richness	10
²¹	Definition of species richness	10
²²	Regional Richness	10
²³	Site Specific Richness (N_{site})	12
²⁴	Occupancy Probability, ψ	14
²⁵	Definition of ψ	14
²⁶	Scatter Plot of Aggregated ψ	14
²⁷	Scatter Plot of $\hat{\psi}$ vs ψ_{true} , split by year and replicate	14
²⁸	Occupancy Response Curves	16
²⁹	Probability of Detection, p	19
³⁰	Definition of p	19
³¹	Demo: Effect of MSOM Hierarchy on p	19
³²	Scatter Plot of \hat{p} vs p_{true}	21
³³	Scatter Plot of \hat{p} vs p_{true} , split by year and replicate	21
³⁴	Assessment with Mixed Effects Models	23
³⁵	E.g. LME for ψ Evaluation	23

37 **Introduction**

38 **Overview**

39 As water temperatures change, species may shift the size and location of their geographical ranges, bearing
40 consequences for the food webs and economies linked to those species. However, species don't always respond
41 similarly to shifting temperatures (different thermal tolerances, e.g.), which means that changing temperature
42 may remix the composition and diversity of ecological communities.

43 The biological, spatial, and temporal scale of community diversity shifting in response to climate is massive.
44 A functional definition of a community may consist of 100's or 1000's of species, each of which may be
45 shifting its range at a scale of decades and 100's kilometers. As a result, we need statistical methods for
46 estimating biodiversity that don't rely on heavy replication and that make efficient use of available data.
47 Enter the superstars: on the data side the trawl data set has amazing spatiotemporal and taxonomic extent
48 and resolution; on the statistical side multispecies occupancy models (MSOM) are hierarchical state space
49 models that are designed to estimate species richness and don't require consistent or extensive "replication".
50 Although they're superstars, even these data and models have their limitations and pitfalls.

51 Can we estimate the dynamics of species richness from trawl data using an MSOM? It's a hard question to
52 answer because we can never know the "truth" for sure, but we can get an idea of how reliable our analysis
53 is by simulating fake data, for which we know true values because we created them. The trawl data set is
54 generated by two distinct processes: Nature's data generating process (NDGP), and the process by which
55 humans observe the result of NDGP. So we ask: to what extent is the accuracy of estimates from an MSOM
56 dependent on characteristics of NDGP, and in particular, the way in which we observe the result of NDGP?
57 The strategy for answering this question is to simulate fake data where we approximate Nature but gain
58 knowledge of "truth", "observe" the results of the true process, then try to recover the true species richness
59 from these simulated data.

60 **The Simulation**

61 The goal of this simulation was to use a very basic process to generate presences and absences of species in
62 space and time. In this version of the simulation, there is no explicit connection between years (they are
63 independent). There is a modest spatial connection, because in the simulation an environmental variable
64 determines habitat suitability. I think of this environmental variable as temperature, and I filled a grid with
65 temperatures that ranged from the coldest at the top of the grid (north) and the warmest at the bottom
66 (south) and added random variation among columns in the same row (among longitudes at the same latitude).

67 One level of the simulation mimics NDGP. In this level, NDGP is best characterized by ψ , which is the product
68 of a temperature and species' response curves. I.e., temperatures were used to determine the suitability of
69 each grid cell to each simulated species. This suitability is known as ψ throughout this document.

70 A second level of the simulation mimics human observation of NDGP — what we do when we collect data.
71 This process was simulated by assigning each species has a unique probability of being observed or "detected"
72 (this variable is p). The observation process gets several attempts at observing a given species in a given grid
73 cell; think of this as subdividing each site into subsites, and when you visit each subsite you have probability

74 p of observing a particular species (each species has its own p). Depending on the settings used in the analysis
75 that this document summarizes, the maximum number of subsites can vary, as can the number of subsites
76 per site (OK, fine; the maximum number of subsites in this version is 25, the number of subsites per site
77 varied between 4 and 20, and overall 50% of total possible subsites were sampled).

78 As previously mentioned, the simulation included “time”. In this basic version, not much changes between
79 the “years” for the true process (temperature doesn’t change, nor do the response curves), but the mean of p
80 does change. In a given year, the entire community has an overall mean probability of being detected, and
81 each species randomly deviates from that mean.

82 The simulation also has replicates. To understand the replicates, it needs to be clear that even when a
83 parameter in the simulation does not change, the outcome can change. The replicates hold the realization
84 of the simulated NDGP constant, and draw new realizations of the observation process. I.e., both ψ and
85 p are constant among replicates, and the binary *outcome* of ψ is also held constant, but the outcome for p
86 can change. Furthermore, although each replicate has same values of p (both the mean p and each species’
87 individualized random draw from that distribution), each replicate switches which year is associated with
88 which p ’s. In this way we can observe each outcome of Nature’s data generating process under a series of
89 settings for the human observation process.

90 Multispecies Occupancy Models (MSOMs)

91 Multispecies occupancy models are Bayesian statespace hierarchical models. They distinguish between truth
92 and observation of the truth, and many parameters share a common “parent” distribution. They are very
93 flexible models, and can be adapted to include new types of processes. The MSOM being used here is a
94 relatively simple version of these models. It predicts the probability of each species existing in a grid cell from
95 a logistic regression equation that uses a second-order polynomial of the environmental variable as a covariate.
96 The parameters in this level of the model are hierarchical, with species having their own paramter values,
97 but these individual parameters are not wholly independent in the sense that they share a common parent
98 distribution, which sort of acts to both limit how different they can be and to inform one another. The model
99 also has an observation level, which only has a hierarchical intercept (just a mean) as a predictor variable.

100 The MSOM makes guesses of the true state of the system (whether a species is actually present or not). It
101 then makes guesses at how the observation of that true state might turn out, which is effectively a prediction
102 of what our data will be. The Bayesian model fitting process then uses this comparison of the observed data
103 to the estimate of the observation to tweak the parameters in the MSOM. This process is repeated until the
104 choice of paramters boils down to what is essentially the posterior distribution of the estimated parameters.

105 Right now the MSOM model is fit separately to each year and each replicate. So the model never gets to see
106 multiple years or multiple replicates at the same time. Furthermore, when referring to a parameter value
107 fitted in the MSOM, it is implied that it can be subscripted with time or replicate (because all years and
108 replicates are fit independently).

109 The parameters in the logistic regression that predicts the value of ψ vary among species, although ψ itself
110 varies among species and space, because the regression parameters (subscripted by species) are multiplied by
111 the environmental variable (subscripted by space). More or less, it can be said that, for a given species, ψ

₁₁₂ varies among space because of the environmental variable, and in a given location it varies among species
₁₁₃ because of the regression parameters.

₁₁₄

115 Conventions and Settings

116 In this section I outline the subscripting and notation used in the MSOM analysis and for the simulation. I
117 also outline various settings (number of species simulated, replicates, etc.). Most of the numbers you see
118 (and some of the text) is dynamically generated based on the code that produced the statistics and figures.
119 Therefore, you can refer back to these sections to see what settings may have changed since the last version
120 of this document.

121 **Note:** *I've often found myself having to get creative with subscripts and superscripts. I've tried to be clear an*
122 *consistent, but small inconsistencies likely exist, so don't be confused by them. For example, if you see $\max(Z)$*
123 *and Z_{\max} in two different sections, they are probably referring to the same thing. If you see something*
124 *confusing, let me know (preferably by (creating an issue on GitHub)[<https://github.com/rBatt/trawl/issues>]),*
125 *and I'll fix it.*

126 Dimension Conventions

127 Summary

128 1. Site ($j = 1, 2, \dots, j_{\max} = 12 \times 12 = 144$)

- 129 • Sites are unique combinations of latitude and longitude
- 130
- 131 • The spatial arrangement of sites is *not* arbitrary, although importance depends on settings (see
- 132 *dynamic* below)
- 133
- 134 • The environmental variable X varies among sites (and years, below)

135 2. Sub-sites ($k = 1, 2, \dots$)

- 136 • Sub-sites are only relevant to the “observation” process
- 137 • Each site has the same number of possible sub-sites, but the number of sub-sites observed can vary
- 138 • In this simulation, $k_{\max} = 25$, $k_{\min}^{\text{observed}} = 4$, and $k_{\max}^{\text{observed}} = 20$
- 139
- 140 • Substrata are primarily useful for determining p , the **detection probability**

142 3. Species ($i = 1, 2, \dots i_{\max} = R = 40$)

- 143 • Does not include “augmented” species
- 144 • For this MSOM analysis, the species array was padded with 0’s

145 4. Time ($t = 1, 2, \dots 4$)

- 146 • Time is primarily used to vary the parameters controlling the “true” process
- 147 • When those parameters don’t change, time provides independent*realizations of the same “true”
- 148 process

149 — *Note: only when *dynamic=FALSE* in *sim.spp.proc*

150 5. Replicates ($r = 8$)

- 151 • Replicates are *simulated* repeated human observations of the same *realization* of the “true” process
152 at Time $_t$
- 153 • Replicates are used to vary the parameters that control the “observation” process
- 154 • When those parameters don’t change, each replicate provides an independent* realization of the
155 same “observation” process

156 **In Code**

157 The MSOM analyzes each year $_t$ -replicate $_r$ combination independently. Parameters subscripted by these
158 dimensions are derived from separate analyses.

159 In my code, I’ve tried to be consistent in my use of these indices to describe arrays, matrices, and rasters.
160 Rows are dimension 1, columns dimension 2, etc. The precedent of subscripted dimensions follows the numeric
161 ordering of the list above. E.g., in the matrix $X_{i,t}$ each row will refer to a different species, and each column
162 a different year (note that site $_j$ is skipped, so species $_i$ is “promoted” to dimension 1, the row.). By default, R
163 fills matrices and arrays by column, whereas the **raster** package fills them by row. In most cases where an R
164 object needs to split sites into the lat/ lot components, I make use of the **raster** package. Therefore, the
165 numbering of the sites proceeds row-wise, where each site is numbered according to the order in which it is
166 filled, as in this 2×3 matrix: $J = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}$

167 Note that even though this matrix is numbered row-wise, it is still indexed as $J_{row,column}$, such that $J_{1,2} = 2$.
168 As mentioned previously, this information is primarily important for understanding the code involved with
169 this project, and in most cases it is not crucial to be explicitly aware of the spatial arrangement of sites.

170

171 **Settings**

172 **Simulation Settings**

173 I created a class called "spp", which has methods for **print()**. The **Dimensions** are the number of sites, the
174 number of species, then the number of years.

175 Also **printed** are some richness summary statistics. **All cells** refers to the collective richness over all j
176 taken together. The meaning of **One cell** differs slightly between the true and observed printouts: in the
177 true printout the richness is of a particular site (j), and in the observed printout it is of a particular sub-site
178 (k).

179 **## Dimensions:** 144, 40, 4
180 **## grid.h** = 12
181 **## grid.w** = 12
182 **## grid.t** = 4
183 **##**
184 **## Number Species Possible (ns):**

```

185 ## 40
186 ## Total Species Richness:
187 ## 40
188 ## Total Observed Species Richness:
189 ## 40
190 ##
191 ## Annual Species Richness:
192 ##      Min. 1st Qu. Median   Mean 3rd Qu. Max.
193 ## All cells 39       39       39 39.250 39.25    40
194 ## One cell  0        3        5  5.781   8.00    16
195 ##
196 ##
197 ## Observed Annual Species Richness:
198 ##      Min. 1st Qu. Median   Mean 3rd Qu. Max.
199 ## All cells 35       38       39 38.000 39       39
200 ## One cell  0        0        0  1.868   3       14

```

201 In the MSOM, detectability (p_i) is determined in the form of a logistic regression, which currently only
202 has an intercept (v_0) as predictor (so just a mean). That intercept varies among species (i.e., $v_{0,i}$), and
203 that variation is generated by drawing each individual species's intercept ($v_{0,i}$) from a parent distribution:
204 $v_{0,i} \sim \mathcal{N}(\mu_{v_0}, \sigma_{v_0}^2)$. See [section about \$p\$](#) for more info.

year	mu.v0	sigma.v0
1	-2	2
2	0	2
3	2	2
4	4	2

205

206 Settings for JAGS & MSOM

nChains	nIter	n0s	nSamples
3	50000	10	500

207 In the table above, `nChains`, `nIter`, and `nSamples` are all variables that are strictly pertinent to the Bayesian
208 analysis carried out in JAGS. The `n0s` value refers to the the degree of “data augmentation”. In this process,
209 you add extra species to the data set, and say that they were never observed. For our purposes, this is
210 employed for purely technical reasons, although it can be used to extra further inferences about species
211 richness.
212 The posterior samples from JAGS consist of 500 samples (above). However, to save memory and hard drive

213 space, I have often only saved measures of central tendency for each of these. In this assessment, I have
214 performed all calculations on the **centralT=median** of the posterior samples.

215

²¹⁶ **Species Richness**

²¹⁷ **Definition of species richness**

²¹⁸ Species richness is the number of different species, or more generically, unique taxa. The point is moot in the
²¹⁹ simulation study, and in the empirical trawl data it refers to species.

²²⁰ Estimates of richness (R) can be made spatially or temporally explicit (or neither, or both). In the following
²²¹ figures, different levels of aggregation are performed – for most figures R is split by year (this is true for all
²²² figures but Figure 1). Figure 2 emphasizes temporal dynamics and keeps replicates separated, but aggregates
²²³ over space (the j sites). Figure 3 doesn't aggregate over space or time, but it does aggregate over “replicate”
²²⁴ observations; importantly, while the figure does present any spatial aggregation, it does not retain the spatial
²²⁵ relationship (you can't tell which sites are next to others). The final two figures of the section (Figure 4 and
²²⁶ Figure 5) are similar to the previous figure, except that spatial relationship among points is retained via a
²²⁷ heatmap representation.

²²⁸ None of these estimates of richness include the 10 species that were part of the “data augmented”/ “adding
²²⁹ 0's” process. Richness values can either be true (true simulated NDGP; R^{true}), observed (true simulated
²³⁰ human observation of NDGP; R^{obs}), or MSOM estimates of one of those two (\hat{R}^{true} or \hat{R}^{obs}).

²³¹

²³² **Regional Richness**

²³³ These estimates of species richness only distinguish between replicates and years. They do not contain any
²³⁴ site-specific information.

²³⁵ **Richness Boxplots**

²³⁶ With the boxplots we're mostly looking to see if the estimates of richness vary with the mean probability of
²³⁷ detection, p . In the empirical data, we know that taxonomic identification changed over time (it improved;
²³⁸ generally, more species were ID'd in later years). We also suspect that gear might change, which affects
²³⁹ the probability of observing a species. The “Fraction Capable of Being ID'd” category in the boxplots is
²⁴⁰ essentially the cross-species average of p .

²⁴¹

²⁴² **Richness Time Series**

²⁴³ Text explanation goes here

²⁴⁴ Need explanations for how each panel was calculated.

²⁴⁵ 1. R^{true} is straightforward

²⁴⁶

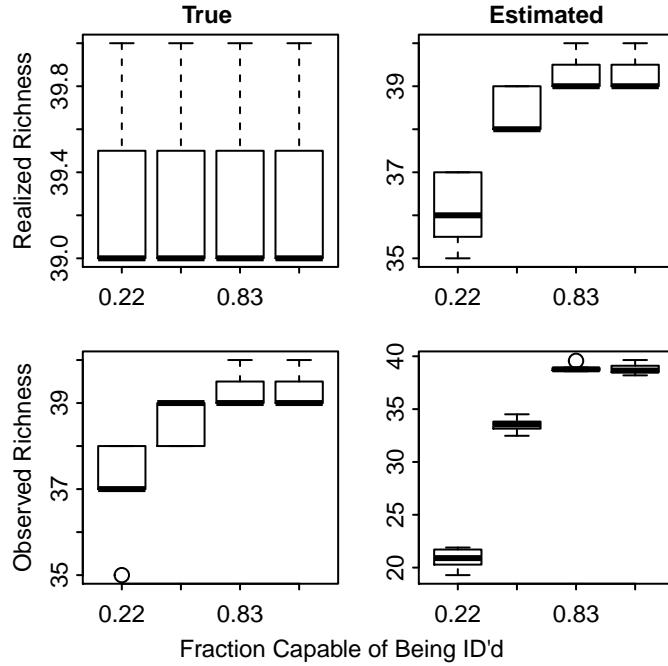


Figure 1: **Figure 1.** Boxplots of species richness. Numeric groupings indicate the average value of p across species during a given year-replicate combination. The panels in the left column are the true simulated values, and the panels on the right are the corresponding MSOM estimates. The top row indicates the latent realized species richness or MSOM estimates of the richness. The bottom row's panels are the simulated observed values of richness (the response variable in the MSOM) and the MSOM estimates of the observed values.

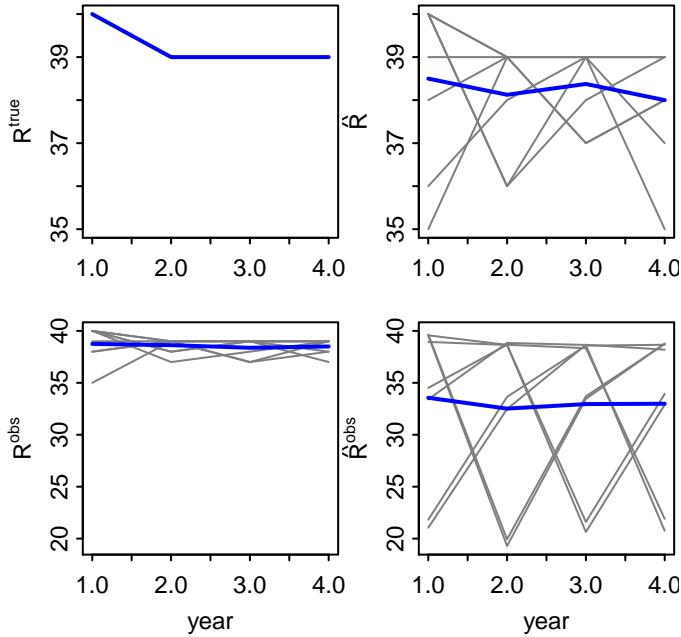


Figure 2: **Figure 2.** Time series of richness. Gray lines are individual replicates. Blue lines are averages. Note that detection probabilities ($p_{i,t,r}$, see [simulation settings above](#), as well as [definition of \$p\$ below](#)) change over time, and their temporal ordering differs among replicates.

247 2. \hat{R} is from $\sum_{i=1}^{R^{true}} \max(\hat{Z}_{i,\nabla j \in J})$; and to be clear, \hat{R} does not include the “unobserved” species introduced
 248 to the MSOM occurrence matrix (Y)

249

250 **Site Specific Richness (Nsite)**

251 **Scatter Plots of Nsite Split by Year**

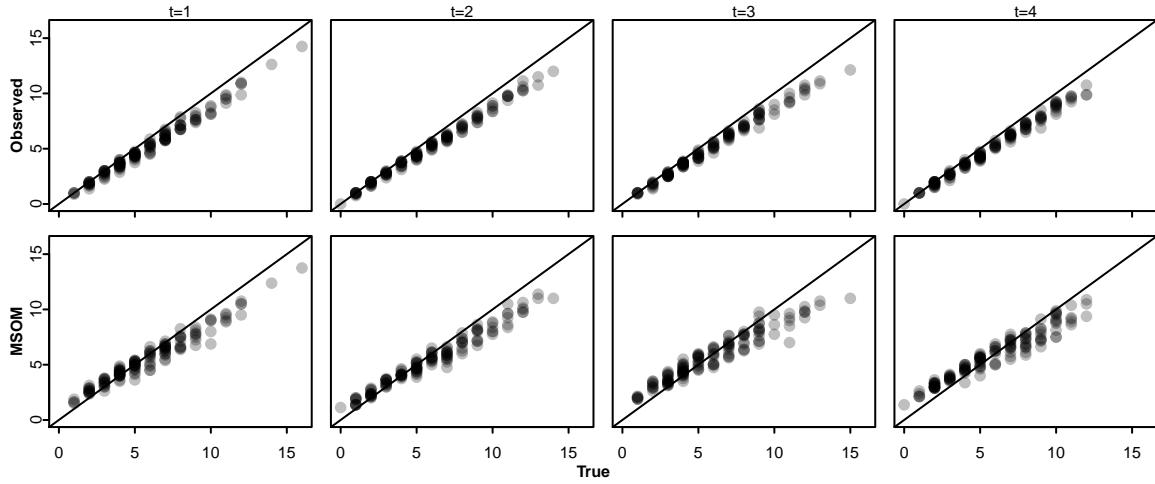


Figure 3: **Figure 3.** Site-specific richness (N_{site}, N_j) from simulated observations (vertical axis, top row; N_j^{obs}) and from MSOM estimates (vertical axis, bottom row, \hat{N}_j) vs true site-specific richness (horizontal axis; N_j^*). The panel columns delineate the years of the simulation. Each point is site-specific ($j = 'rgrid.h' \times 'rgrid.w' = 'rgrid.h * grid.w'$) species richness that has been averaged over the simulated replicate observations ($r = 'rn.obs.reps'$).

252

253 **Maps of Richness (space and time)**

254

255 Text explanation goes here

256

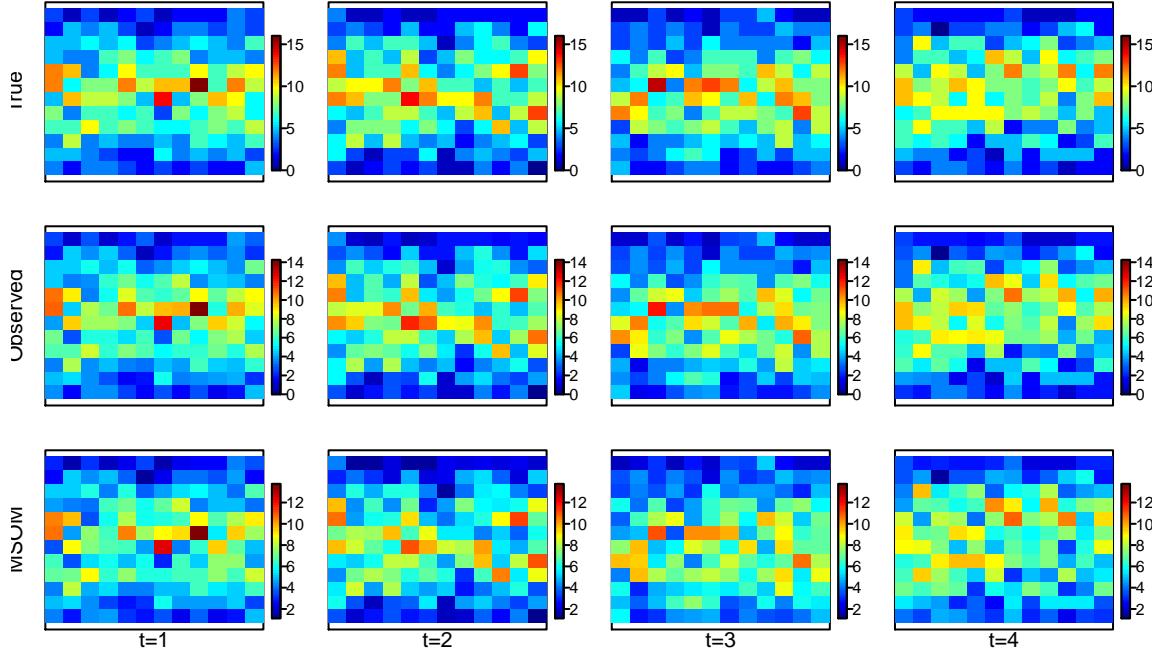


Figure 4. **Figure 4.** Maps of site- and year-specific species richness (N_{site}) from the simulation of the True process (top row), simulation of the Observed process (middle row), and the MSOM estimates (bottom row). X-axis and Y-axis indicate position in 2 dimensional space; it is important to note that the environmental variable changes linearly across the y-axis, and randomly (and much less) across the x-axis. The different columns represent separate years. The environmental variable changes linearly among years (the rate of change is the same for all x-y locations). Colors indicate species richness (warm colors are higher richness than cool colors), averaged over the simulated replicate observations ($r = 'rn.obs.reps'$). Horizontal and vertical axes Each row of panels is scaled independently, columns within a row are scaled equally.

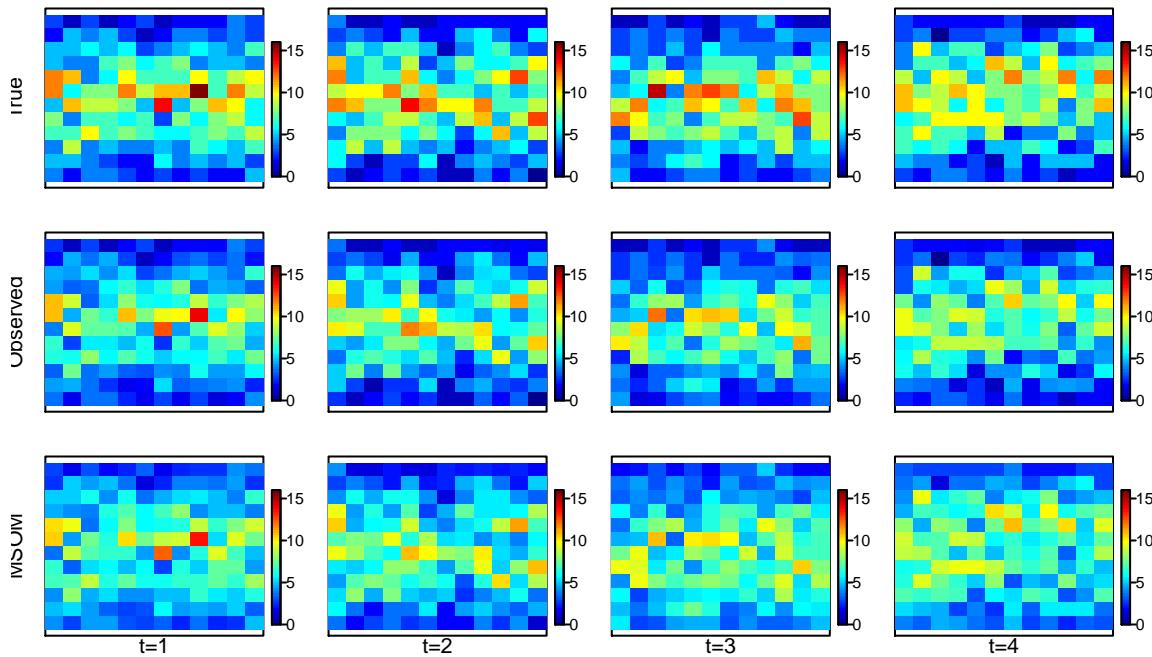


Figure 5: **Figure 5.** Same as previous figure, but all panels are on the same scale.

257 **Occupancy Probability, ψ**

258 **Definition of ψ**

- 259 Definition description goes here
- 260 Probably need to describe how it's generated in the simulation
- 261 As well as how it's estimated in the MSOM
- 262 In particular, important to point out that they may or may not match

263

264 **Scatter Plot of Aggregated ψ**

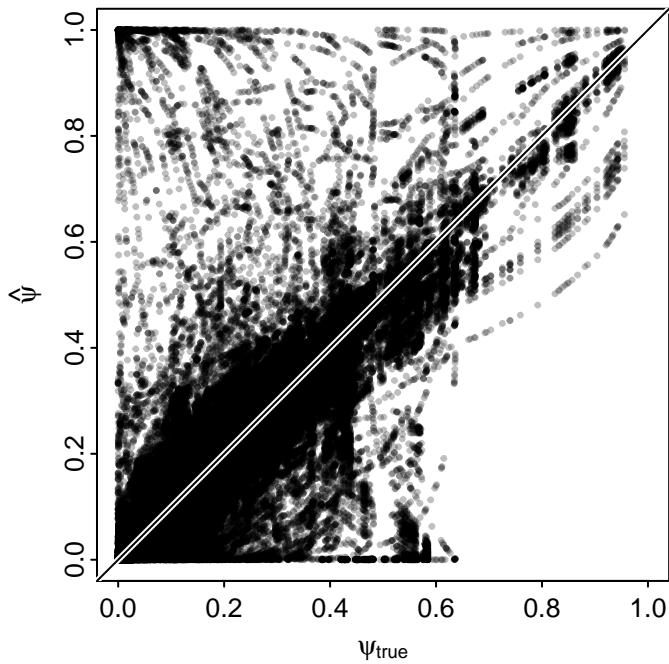


Figure 6: **Figure 6.** MSOM estimates of ψ ($\hat{\psi}$) vs. true values of ψ (ψ_{true}). Each point is a ψ value for a particular site-species-year, averaged across $r = 'rn.obs.reps'$ simulated replicate observations (i.e., the 'true' value is the same, but each simulated replicate has a different outcome of how the same true process was observed). The white and black line is the 1:1 line.

265

266 **Scatter Plot of $\hat{\psi}$ vs ψ_{true} , split by year and replicate**

- 267 Note: what I refer to as p here is really just the probability that a species will be detected if an occupied site is
- 268 sampled. In this simulation, 50% of substrata were sampled, which doesn't influence p , but can add noise to
- 269 its estimates.

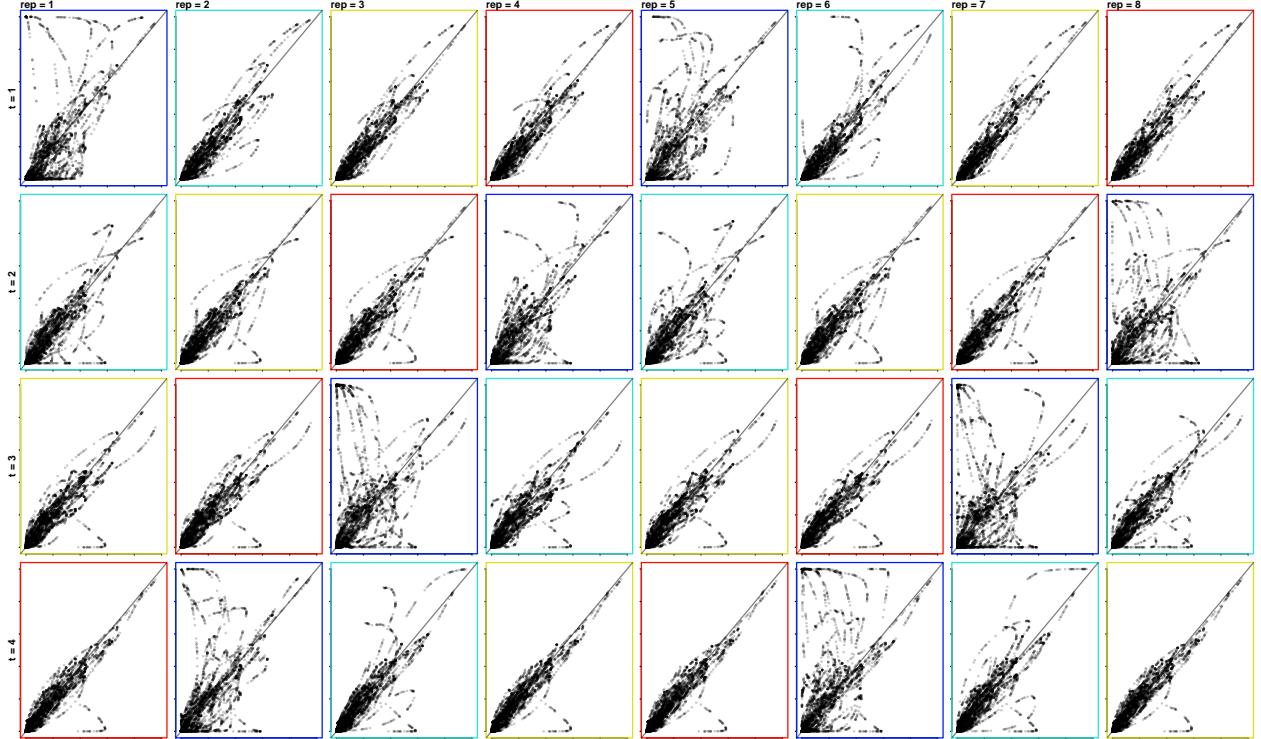


Figure 7: Figure 7. True (horizontal axes) and MSOM estimates (vertical axes) of occupancy probabilities ($\psi_{j,i,t,r}$) of species i occupying a location j in year t . In our simulation, ψ is a function of individual species characteristics (niche) and the environment, the latter of which changes among years. The simulated (true) outcome of each year was subject to r replicate observations of the true process. Each simulated observation (r) was an independent realization, but the r replicates also differed in the probability that a species would be detected (p): the color of the boxes around each panel refer to the among-species average of the probability of detection; warm colors indicate that the mean detection probability is high (red; $p_{max} = \text{round}(\max(\text{taxChance}), 2)$), whereas cool colors indicate that p was low (blue; $p_{min} = \text{round}(\min(\text{taxChance}), 2)$). The year t of the simulated true process changes across the rows of panels, and the simulated replicate observation r changes across columns.

271 Occupancy Response Curves

Occupancy response curves are calculated as $\text{logit}(\psi_i) = \mathbf{X} \times \mathbf{a}_i$, where

$$\mathbf{X} = \begin{pmatrix} 1 & X_{min} & X_{min}^2 \\ \vdots & \vdots & \vdots \\ 1 & X_{max} & X_{max}^2 \end{pmatrix}; \mathbf{a}_i = \begin{pmatrix} a_{0,i} \\ a_{3,i} \\ a_{4,i} \end{pmatrix}$$

272 Therefore, these curves are tantamount to values of ψ , except that ψ generally pertains to a simulated,
 273 observed, or true occupancy probability, whereas the occupancy probability in the response curves is calculated
 274 over hypothetical conditions (i.e., over hypothetical values of the environmental gradient X).

275 True Occupancy Response Curves

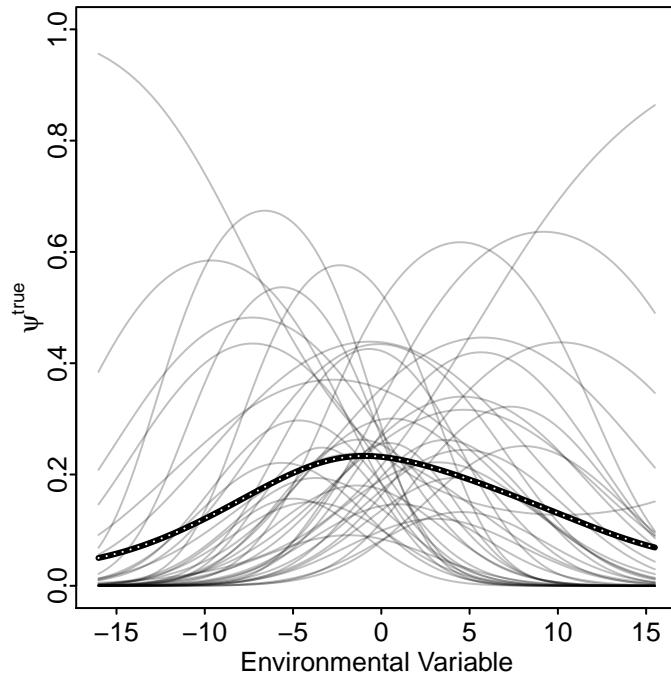


Figure 8: **Figure 8.** True simulated response curves. Vertical axis is the value of ψ^{true} , horizontal axis is the value of the environmental variable that, along with species-specific regression parameters, determines ψ^{true} . The thick line is the among-species mean value of ψ^{true} at a given value of the environmental variable.

276 In the response curve, the values of the environmental variable are an arbitrary gradient, and do not necessarily
 277 correspond to what was observed in the simulated environment (although they are intended to cover the
 278 same range).

279 Estimated Occupancy Response Curves

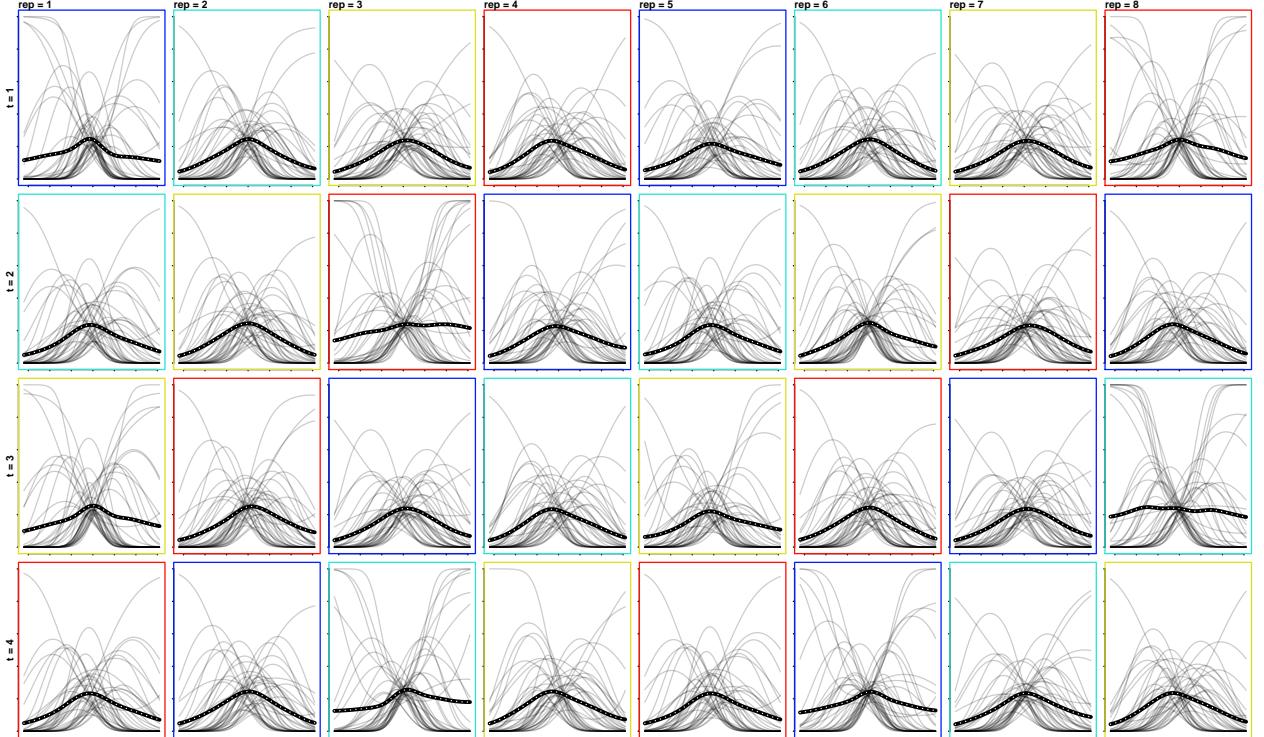


Figure 9: **Figure 9.** Response curves of species' probability of occupancy (ψ_i , vertical axis) across the full range of temperatures in the simulation ($\min(X) = \text{'rround(range.X[1], 1)'}$, and $\max(X) = \text{'rround(range.X[2], 1)'}$). The color of the boxes around each panel refer to the among-species average of the probability of detection; warm colors indicate that the mean detection probability is high (red; $p_{\max} = \text{r round(max(taxChance), 2)}$), whereas cool colors indicate that p was low (blue; $p_{\min} = \text{r round(min(taxChance), 2)}$). The year t of the simulated true process changes across the rows of panels, and the simulated replicate observation r changes across columns.

281 **Probability of Detection, p**

282 **Definition of p**

283 The probability of detection (p), is a species specific parameter in the MSOM model. The MSOM analyzes
284 all years (t) and replicates (r) separately, so I am going to leave those subscripts out of this description. In
285 the simulation, the probability of observing a species is a function of two independent factors:

- 286 1. The probability that site j is occupied by species i ; this is $\psi_{j,i}$
- 287 • $\psi_{j,i}$ is a function of species-specific niche and an environmental variable that changes over space
288 and time
- 289
- 290 • $Z_{j,i}$ is the species- and site-specific richness, which is a function of ψ (given that we're only talking
291 about species that are in the pool of possible species, determined by w_i)
- 292 2. A species-specific (i) chance of being identified (`taxChance`), given that it is present in a location that
293 was sampled (i.e., detectability does not reflect the probability of sampling a place); this detectability
294 parameter is p_i
- 295 • Detectability changed between years.
- 296 • In a given year, $\text{logit}(p_i) \sim \mathcal{N}(\sqrt{\mu}, \sigma^\epsilon)$. μ changed between years (taking on values of -2, 0, 2,
297 and 4), $\sigma^2 = 2$ in all years.
- 298 • The value of p only changes between species (and years), but the observation process occurs at the
299 substratum (k) level. Thus, the parameter is really $p_{j,k,i}$, but for a given i , all $p_{j,k}$ are constant. I
300 represent this probability as p_i with the understanding that this value is repeated over space.
- 301 • $Y_{j,i}$ is the observed version of $Z_{j,i}$.
- 302 • $Y_{j,i} \sim \text{Bern}(p_i \times Z_{j,i})$.
- 303 – Note: Because p is actually subscripted to k , the Y are also actually subscripted to k . Maybe
304 leaving these subscripts out is making things more confusing. I've only excluded them to
305 emphasize how parameters are estimated.
- 306 • Our data about species presence/ absence correspond to $Y_{j,i}$. So it might be useful to think of the
307 MSOM as estimating $\hat{Y}_{j,i}$, which is compared to the observed data $Y_{j,i}^{obs}$.

308

309 **Demo: Effect of MSOM Hierarchy on p**

310 The above plot is interesting because it shows that exceptionally high or exceptionally low chances to be
311 observed only occur for the species that were observed at least once; i.e., this says that if you didn't observe
312 it, it just takes on the mean. That's reasonable, I guess; but I also would think that the things that were
313 never observed could also be things that had a low chance of observability; but they could also have just a
314 low chance of actually being present. So I suppose in the end it just doesn't get informed, and reverts to the
315 mean?

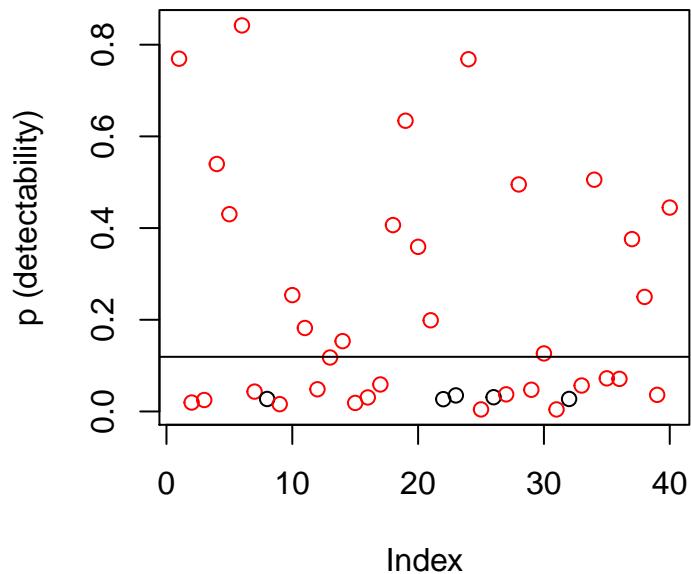


Figure 10: **Figure 10.** Probability of being detected, p . Horizontal line is mean probability. Figure only shows results for the first year of the simulation/ observation, and only 1 replicate. Different points are different species. Probability of being detected is a species-specific parameter (does not vary among sites, e.g.). Red points are species that were observed, black points are species that were never observed.

316 Scatter Plot of \hat{p} vs p_{true}

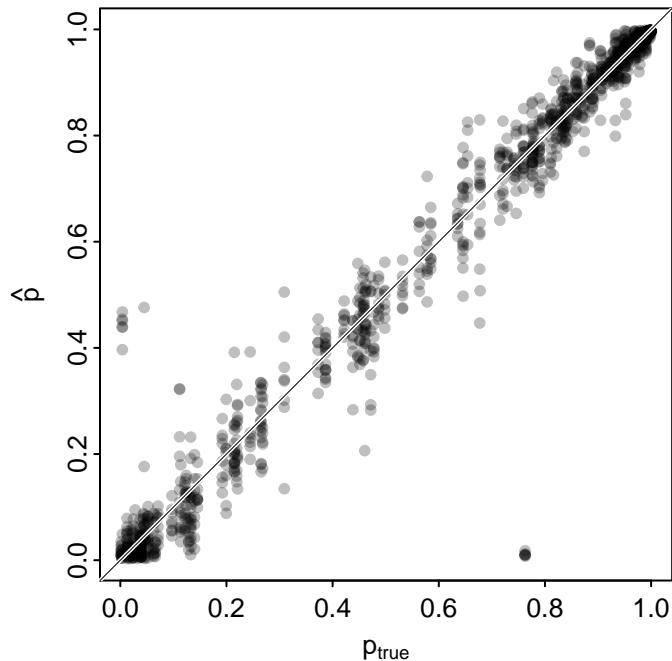


Figure 11: **Figure 11.** MSOM estimates (vertical axis) and true values of p_i , the species-specific (i) detection probability. Each point is subscripted by species ir ifelse(agg.p, 'and', ',') year t , r ifelse(agg.p, paste0('but are averaged among the \$r=', n.obs.reps, '\$ observation replicates'), 'and observation replicate \$r\$').

317

318 Scatter Plot of \hat{p} vs p_{true} , split by year and replicate

319 **Figure.** Caption goes here.

320 Text explanation goes here

321

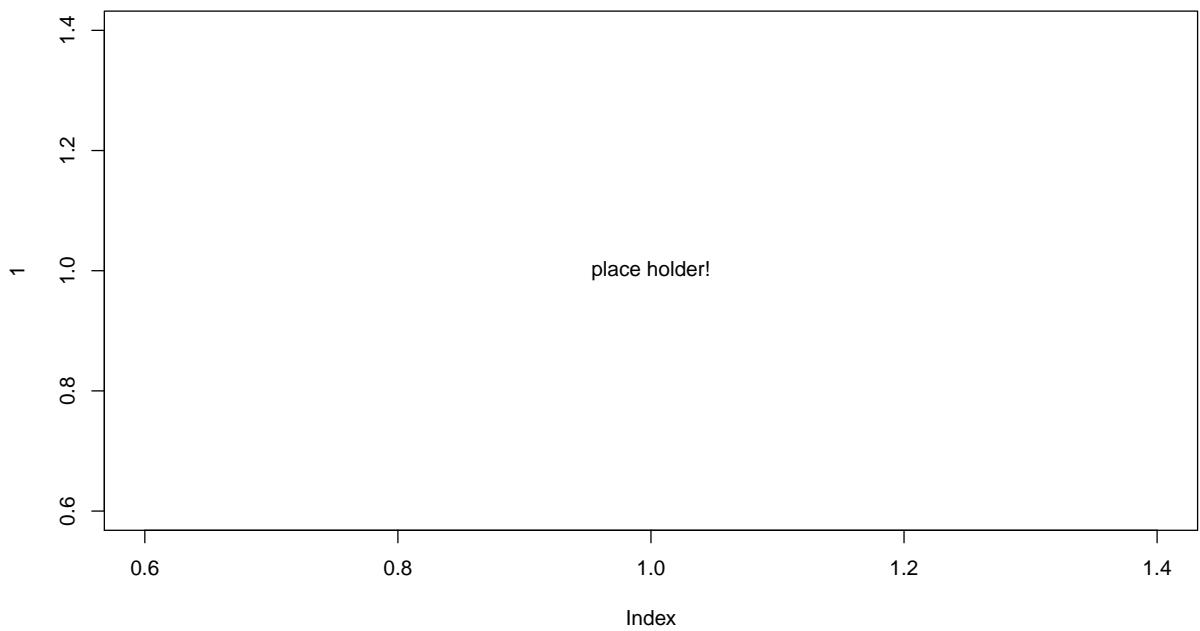


Figure 12: **Figure 12.** Caption goes here.

322 Assessment with Mixed Effects Models

323 E.g. LME for ψ Evaluation

- 324 **Motivation:** MSOM skill might differ across dimensions, trying to figure out what patterns I should expect
325 to pick out (spatial patterns in richness, temporal?) E.g., Is the correlation between MSOM and True the
326 same comparing across sites as comparing across years? Species, reps, also.
- 327 **Motivation:** What factors influence MSOM skill in a given dimension? E.g., Skill in finding differences in ψ
328 across species may depend on p , the chance of being identified. If p changes among years, might also explain
329 Read more about [specifying mixed effects models using lmer in R here](#)
- 330 This example is looking at ψ , probability of an individual species being present

```
# =====
# = LME Model on Psi =
# =====
# Just exploration/ starting point
library(car)
library(lme4)

blah <- reshape2:::melt.array(psi.true, varnames=c("site","spp","time","rep"), value.name="true", as.is=TRUE)
blah.hat <- reshape2:::melt.array(psi.hat, varnames=c("site","spp","time","rep"), value.name="hat", as.is=TRUE)
blah <- cbind(blah, hat=blah.hat[, "hat"])

blah$site <- as.factor(blah$site)
blah$spp <- as.factor(blah$spp)
blah$time <- as.factor(blah$time)
blah$rep <- as.factor(blah$rep)

(blah.mod <- lmer(hat~true+(1|spp)+(1|time), data=blah))

## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl =
## control$checkConv, : Model failed to converge with max|grad| = 0.00329892
## (tol = 0.002)

## Linear mixed model fit by REML ['lmerMod']
## Formula: hat ~ true + (1 | spp) + (1 | time)
## Data: blah
## REML criterion at convergence: -278996.7
## Random effects:
## Groups   Name        Std.Dev.
## spp      (Intercept) 0.028588
## time     (Intercept) 0.005547
```

```

342 ## Residual           0.113441
343 ## Number of obs: 184320, groups: spp, 40; time, 4
344 ## Fixed Effects:
345 ## (Intercept)      true
346 ##          0.01984   0.87989

```

Anova(blah.mod)

```

347 ## Analysis of Deviance Table (Type II Wald chisquare tests)
348 ##
349 ## Response: hat
350 ##          Chisq Df Pr(>Chisq)
351 ## true 246243  1 < 2.2e-16 ***
352 ## ---
353 ## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

354

```

355 ## Warning: Figure(s) 6789101112 with label(s)
356 ## 'psiAggFigpsiPlot.splitScatterresponseCurve.trueresponseCurve.msomegHierarchpPlot.fullScatter.cappPl
357 ## are present in the document but are never referred to in the text.

358 ## Last compiled on: 2015-08-24

359 ## R version 3.1.2 (2014-10-31)
360 ## Platform: x86_64-apple-darwin13.4.0 (64-bit)
361 ##
362 ## locale:
363 ## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
364 ##
365 ## attached base packages:
366 ## [1] parallel    grid       stats      graphics   grDevices utils     datasets
367 ## [8] methods     base
368 ##
369 ## other attached packages:
370 ## [1] kfigr_1.2      lme4_1.1-6      Rcpp_0.11.6      Matrix_1.1-4
371 ## [5] car_2.0-24     rbLib_0.0.2      xtable_1.7-4      rmarkdown_0.7
372 ## [9] knitr_1.10.5    doParallel_1.0.8 iterators_1.0.7    foreach_1.4.2
373 ## [13] R2jags_0.5-6    rjags_3-15      coda_0.16-1      lattice_0.20-29
374 ## [17] igraph_0.7.1    fields_6.9.1     maps_2.3-6       spam_0.41-0
375 ## [21] data.table_1.9.4 raster_2.3-24    sp_1.0-17
376 ##
377 ## loaded via a namespace (and not attached):

```

```
378 ## [1] abind_1.4-0          boot_1.3-17        chron_2.3-45
379 ## [4] codetools_0.2-9      digest_0.6.8       evaluate_0.7
380 ## [7] formatR_1.2         highr_0.5          htmltools_0.2.6
381 ## [10] MASS_7.3-35        mgcv_1.8-3         minqa_1.2.3
382 ## [13] nlme_3.1-118       nnet_7.3-8         numbers_0.5-6
383 ## [16] pbkrtest_0.4-2     plyr_1.8.1         quantreg_5.11
384 ## [19] R2WinBUGS_2.1-19    RcppEigen_0.3.2.1.1 reshape2_1.4.1
385 ## [22] SparseM_1.6        splines_3.1.2      ssh.utils_1.0
386 ## [25] stringr_0.6.2      tools_3.1.2        yaml_2.1.13
```