

Ecography

**ECOG-01633**

Authier, M., Saraux, C. and Péron, C. 2016. Variable selection and accurate predictions in habitat modelling: a shrinkage approach. – *Ecography* doi: 10.1111/ecog.01633

**Supplementary material**

**Appendix 1**

# Variable Selection and Accurate Predictions in Habitat Modelling: a Shrinkage Approach - Appendix

Matthieu Authier<sup>a,d,e</sup>, Claire Saraux<sup>b</sup>, and Clara Péron<sup>c,d</sup>

<sup>a</sup>Observatoire PELAGIS UMS 3462, Université de La Rochelle, 5 Allées de l’Océan 17000 La Rochelle, France

<sup>b</sup>Ifremer (Institut Français de Recherche pour l’Exploitation de la Mer), UMR MARBEC, Sète, France

<sup>c</sup>Institute for Marine and Antarctic Studies, University of Tasmania and Australian Antarctic Division, 203 Channel highway, Kingston, Tasmania 7050, Australia

<sup>d</sup>Écologie Spatiale des Populations, Centre d’Écologie Fonctionnelle et Évolutive, 1919 route de Mende, 34293 Montpellier cedex 5, France

<sup>e</sup>authier@gmail.com

February 15, 2016

## Contents

|                 |   |
|-----------------|---|
| List of Figures | 1 |
| List of Tables  | 2 |

## List of Figures

|   |   |
|---|---|
| A.1 : Probability density function on the shrinkage coefficient induced with a Horseshoe prior (that is, shrinkage profile of the horseshoe prior). Denoting $\beta_{\text{unshrunk}}$ and $\beta_{\text{shrunk}}$ the unshrunk and shrunk regression coefficient, the shrinkage coefficient $s$ is such that $\beta_{\text{shrunk}} = s \times \beta_{\text{unshrunk}}$ . If this coefficient is 0, there is complete shrinkage and $\beta_{\text{shrunk}} = 0$ . If this coefficient is 1, there is no shrinkage and $\beta_{\text{shrunk}} = \beta_{\text{unshrunk}}$ . The horseshoe prior favours either complete or no shrinkage. . . . . | 4 |
|---|---|

|    |      |  |    |
|----|------|--|----|
| 23 | A.2  | : Boxplots of the small pelagic fish biomass data during the 2011 PELMED survey. Boxes represent the   |    |
| 24 |      | interquartile range and the vertical line within, the median. Whiskers extend to the lower and higher hinge  |    |
| 25 |      | defined as $1.5 \times$ the interquartile range. All data are right-skewed betraying a large proportion of zeros   |    |
| 26 |      | and a few extreme values. The x-axis is on a logarithmic scale. . . . .  | 5  |
| 27 | A.3  | : Correlation matrices for the different datasets. Empirical pairwise correlations are printed and color-  |    |
| 28 |      | coded (according their absolute magnitude). Calibration and validation data have a similar correlation   |    |
| 29 |      | structure, which is slightly different from that of the prediction data. These graphs were done thanks to  |    |
| 30 |      | code provided by Peter Haschke (R code lifted from <a href="http://www.peterhaschke.com/r/2013/04/23/CorrelationMatrix.html">http://www.peterhaschke.com/r/2013/04/23/</a> |    |
| 31 |      | <code>CorrelationMatrix.html</code> ). . . . .   | 7  |
| 32 | A.4  | : Raw data and comparison of model predictions (posterior median) for juvenile European anchovies log-   |    |
| 33 |      | biomasses. The distribution during summer 2011 showed a clear spatial structure. The black dotted line   |    |
| 34 |      | materializes the Carmague Natura 2000 protected area. . . . .  | 8  |
| 35 | A.5  | : Raw data and comparison of model predictions (posterior median) for adult European anchovies log-  |    |
| 36 |      | biomasses. The distribution during summer 2011 showed a clear spatial structure. The black dotted line   |    |
| 37 |      | materializes the Carmague Natura 2000 protected area. . . . .  | 9  |
| 38 | A.6  | : Raw data and comparison of model predictions (posterior median) for juvenile European sardine log-   |    |
| 39 |      | biomasses. The distribution during summer 2011 showed a clear pattern linked to depth: juvenile sardines   |    |
| 40 |      | were abundant very close to the coastline of the Gulf of Lion. The black dotted line materializes the  |    |
| 41 |      | Carmague Natura 2000 protected area. . . . .   | 10 |
| 42 | A.7  | : Raw data and comparison of model predictions (posterior median) for adult European sardine log-  |    |
| 43 |      | biomasses. The distribution during summer 2011 showed no obvious spatial structure. The black dotted   |    |
| 44 |      | line materializes the Carmague Natura 2000 protected area. . . . .   | 11 |
| 45 | A.8  | : Raw data and comparison of model predictions (posterior median) for juvenile sprat log-biomasses. The  |    |
| 46 |      | distribution during summer 2011 showed a clear spatial structure. The black dotted line materializes the   |    |
| 47 |      | Carmague Natura 2000 protected area. . . . .   | 12 |
| 48 | A.9  | : Plots of the estimated posterior standard error of the mean against the estimated posterior mean ( $\beta_p$ ).  |    |
| 49 |      | Estimates from $\mathcal{M}_2$ were noisy, especially the coefficients linked to sediments. In contrast, the funnel  |    |
| 50 |      | shape of plots from $\mathcal{M}_{3-5}$ illustrates how shrinkage greatly reduced both the estimated posterior mean  |    |
| 51 |      | and standard error of the mean. . . . .  | 13 |
| 52 | A.10 | : Plots of the variance in estimated regression coefficients $\beta_p$ between cross-validations against the within-   |    |
| 53 |      | variance. Dots are proportional to a z-score (the ratio of estimated posterior mean to its standard error) of  |    |
| 54 |      | the coefficients averaged across the different cross-validation datasets. The between-variance was greatest  |    |
| 55 |      | for $\mathcal{M}_2$ illustrate instability in estimation. In contrast, this between-variance was greatly reduced with  |    |
| 56 |      | $\mathcal{M}_{3-5}$ and comparable to the within-variance. The grey dashed line shows the identity line (between-  |    |
| 57 |      | variance = within-variance). . . . .   | 14 |

58 **List of Tables**

59 A.1 Data Sources of Environmental Inputs. . . . . 6

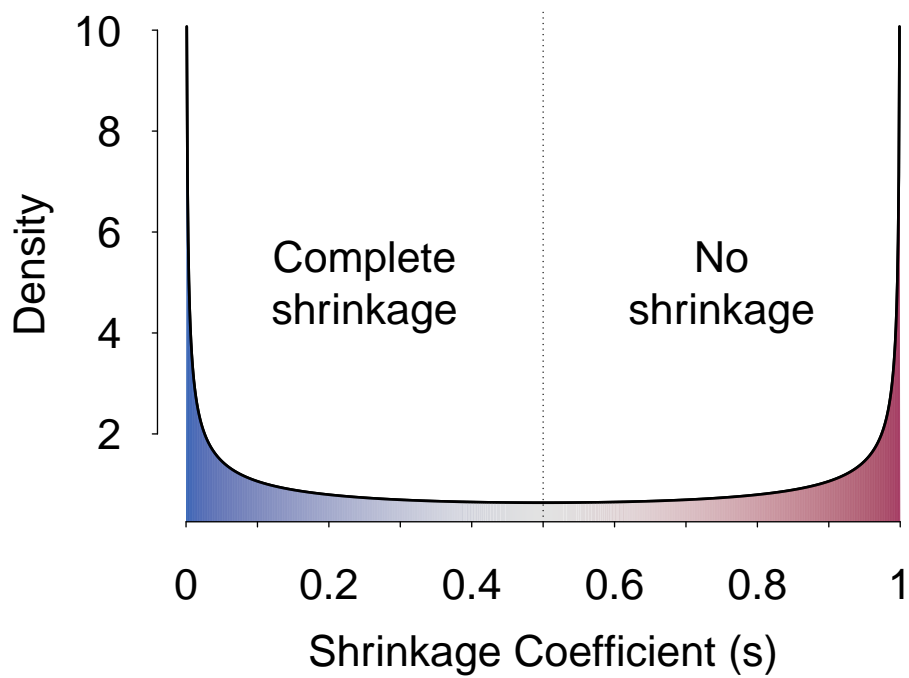


Figure A.1: : Probability density function on the shrinkage coefficient induced with a Horseshoe prior (that is, shrinkage profile of the horseshoe prior). Denoting  $\beta_{\text{unshrunk}}$  and  $\beta_{\text{shrunk}}$  the unshrunk and shrunk regression coefficient, the shrinkage coefficient  $s$  is such that  $\beta_{\text{shrunk}} = s \times \beta_{\text{unshrunk}}$ . If this coefficient is 0, there is complete shrinkage and  $\beta_{\text{shrunk}} = 0$ . If this coefficient is 1, there is no shrinkage and  $\beta_{\text{shrunk}} = \beta_{\text{unshrunk}}$ . The horseshoe prior favours either complete or no shrinkage.

61 **Data**

62 **Boxplot**

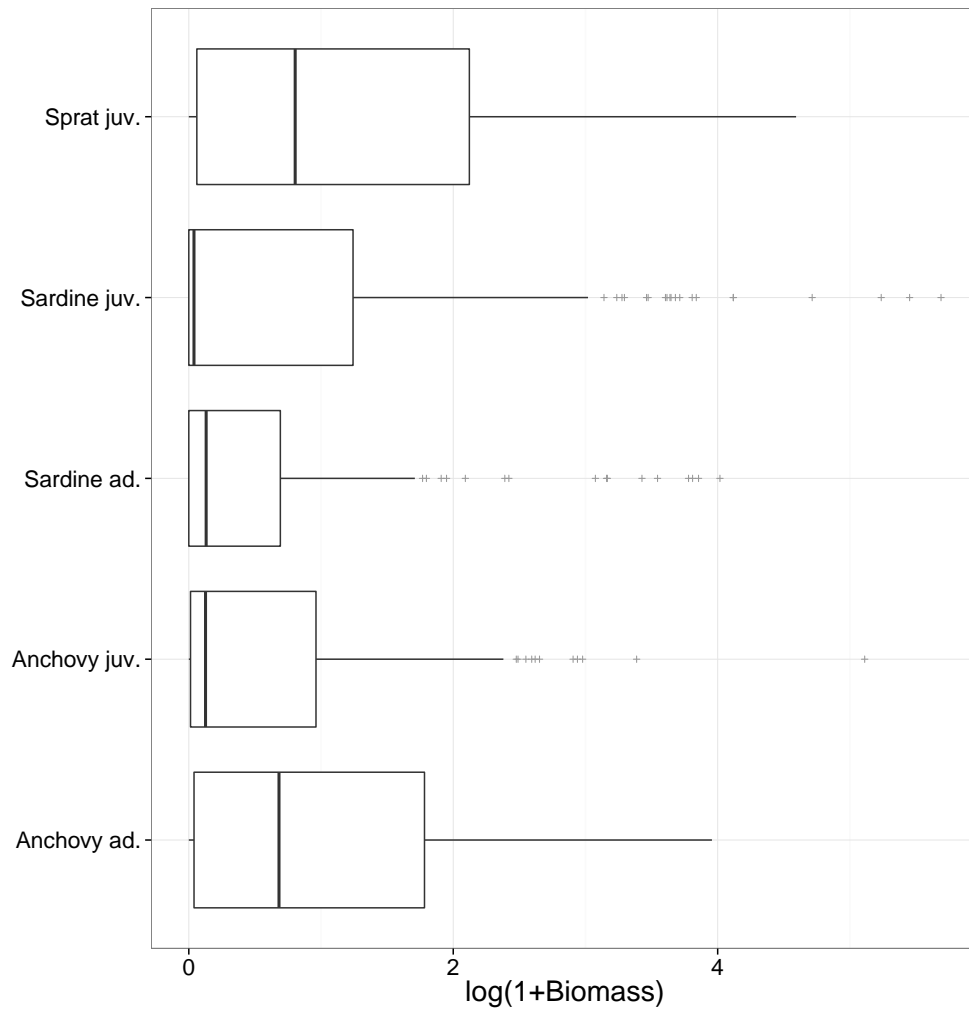


Figure A.2: : Boxplots of the small pelagic fish biomass data during the 2011 PELMED survey. Boxes represent the interquartile range and the vertical line within, the median. Whiskers extend to the lower and higher hinge defined as  $1.5 \times$  the interquartile range. All data are right-skewed betraying a large proportion of zeros and a few extreme values. The x-axis is on a logarithmic scale.

63 **Environmental Inputs: Source and Resolution**

Table A.1: Data Sources of Environmental Inputs.

| Input                       | Spatial resolution | Temporal | Source     | url   |
|-----------------------------|--------------------|----------|------------|---|
| Bathymetry                  | 0.01666°           |          | MODIS\Aqua | <a href="http://coastwatch.pfel.noaa.gov/coastwatch/CWBrowseIWW360.jsp">http://coastwatch.pfel.noaa.gov/coastwatch/CWBrowseIWW360.jsp</a> |
| Sea Surface Temperature     | 0.05°              | weekly   | MODIS\Aqua | <a href="http://coastwatch.pfel.noaa.gov/coastwatch/CWBrowseIWW360.jsp">http://coastwatch.pfel.noaa.gov/coastwatch/CWBrowseIWW360.jsp</a> |
| Chlorophyll a Concentration | 0.05°              | weekly   | MODIS\Aqua | <a href="http://coastwatch.pfel.noaa.gov/coastwatch/CWBrowseIWW360.jsp">http://coastwatch.pfel.noaa.gov/coastwatch/CWBrowseIWW360.jsp</a> |

64 Correlation matrices

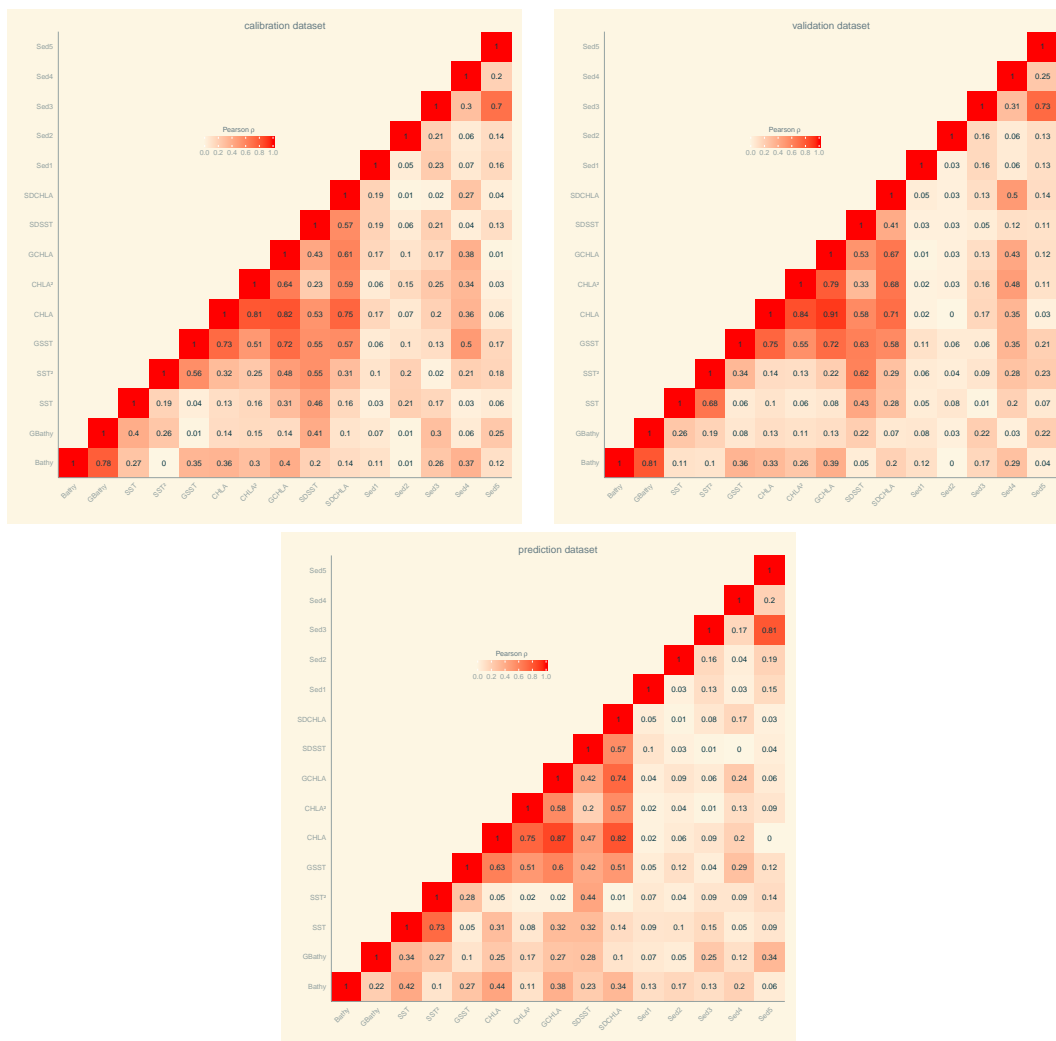


Figure A.3: : Correlation matrices for the different datasets. Empirical pairwise correlations are printed and color-coded (according their absolute magnitude). Calibration and validation data have a similar correlation structure, which is slightly different from that of the prediction data. These graphs were done thanks to code provided by Peter Haschke (R code lifted from <http://www.peterhaschke.com/r/2013/04/23/CorrelationMatrix.html>).



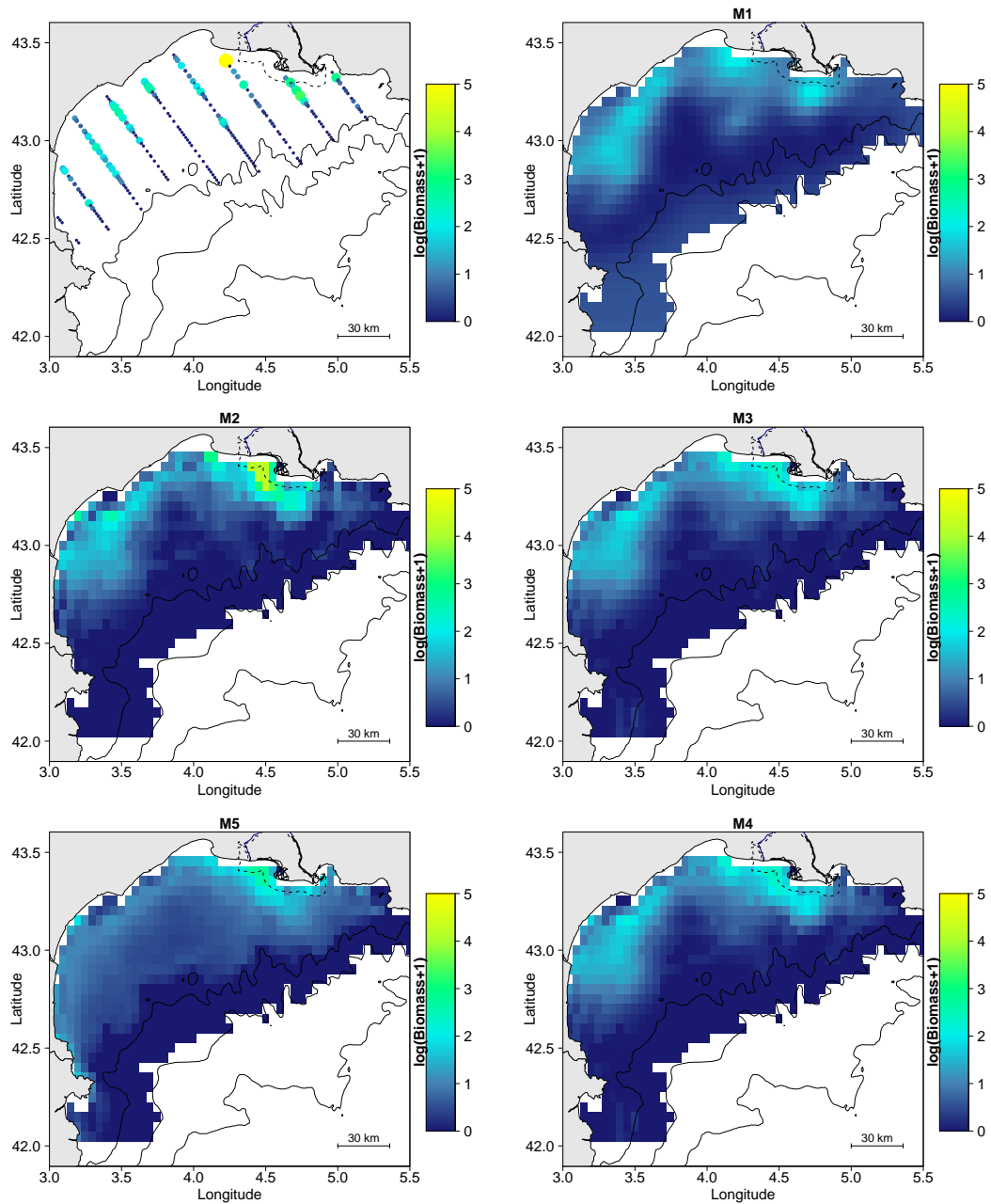


Figure A.4: : Raw data and comparison of model predictions (posterior median) for juvenile European anchovies log-biomasses. The distribution during summer 2011 showed a clear spatial structure. The black dotted line materializes the Carmagne Natura 2000 protected area.

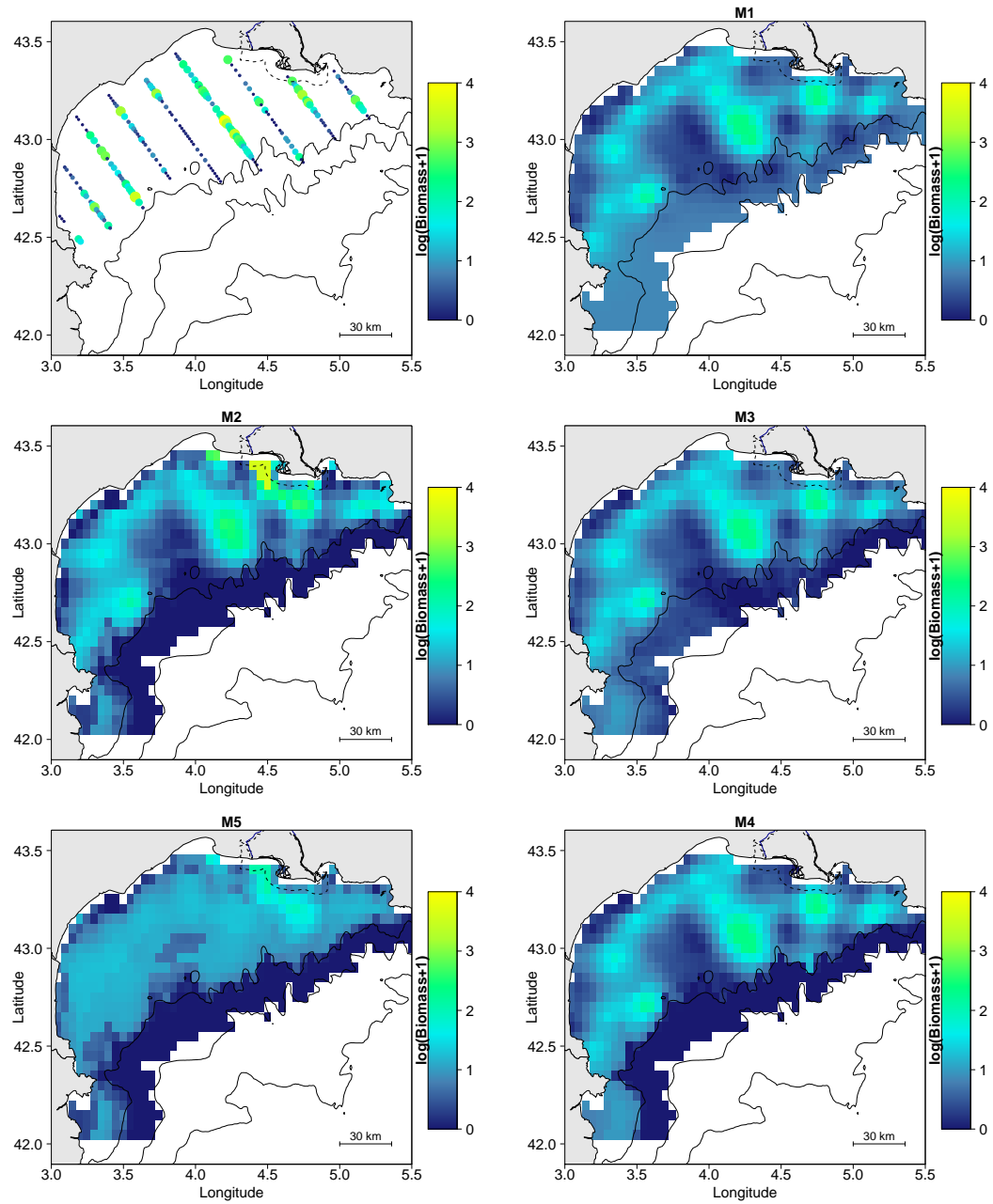


Figure A.5: : Raw data and comparison of model predictions (posterior median) for adult European anchovies log-biomasses. The distribution during summer 2011 showed a clear spatial structure. The black dotted line materializes the Carmague Natura 2000 protected area.

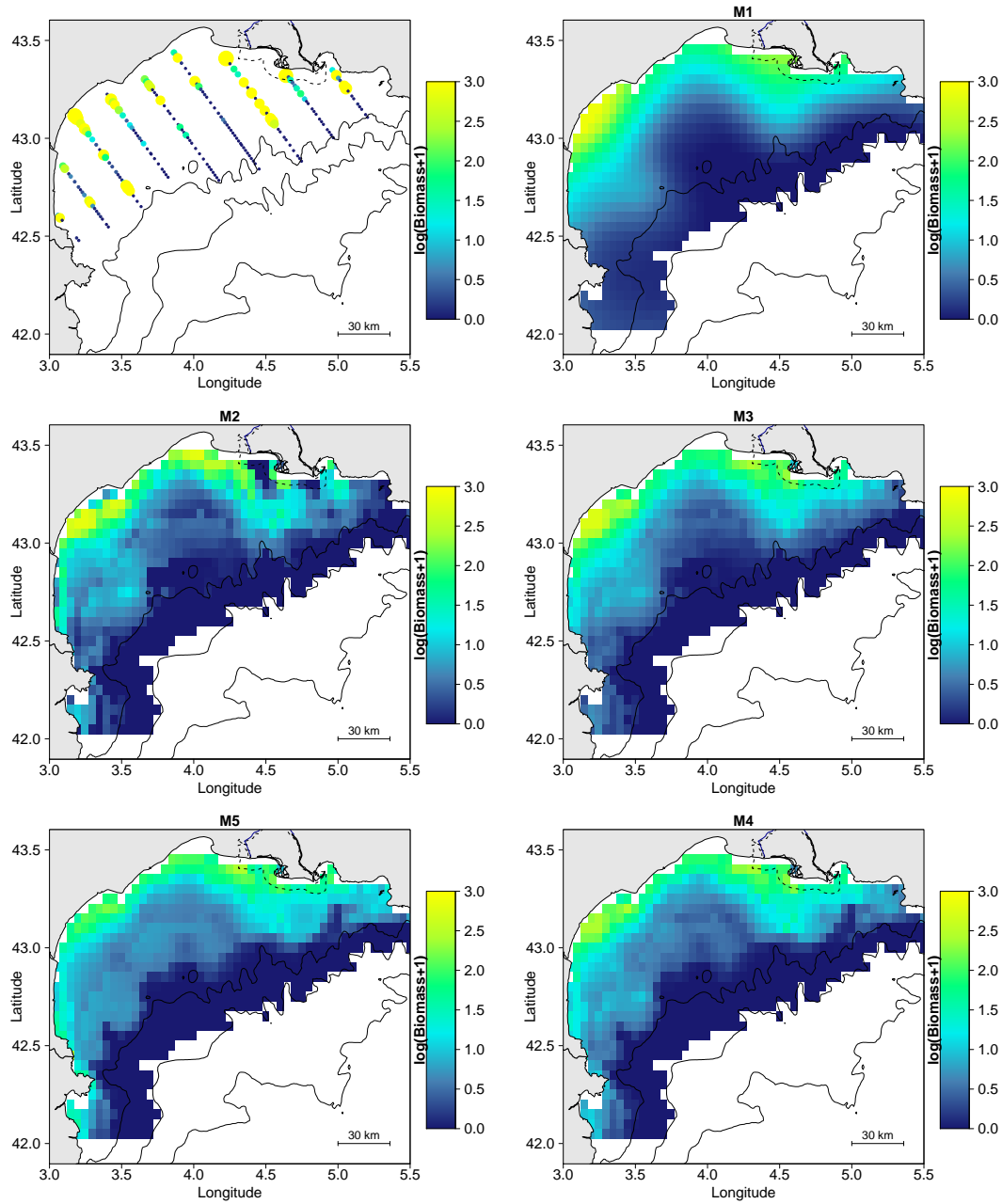


Figure A.6: : Raw data and comparison of model predictions (posterior median) for juvenile European sardine log-biomasses. The distribution during summer 2011 showed a clear pattern linked to depth: juvenile sardines were abundant very close to the coastline of the Gulf of Lion. The black dotted line materializes the Carmague Natura 2000 protected area.

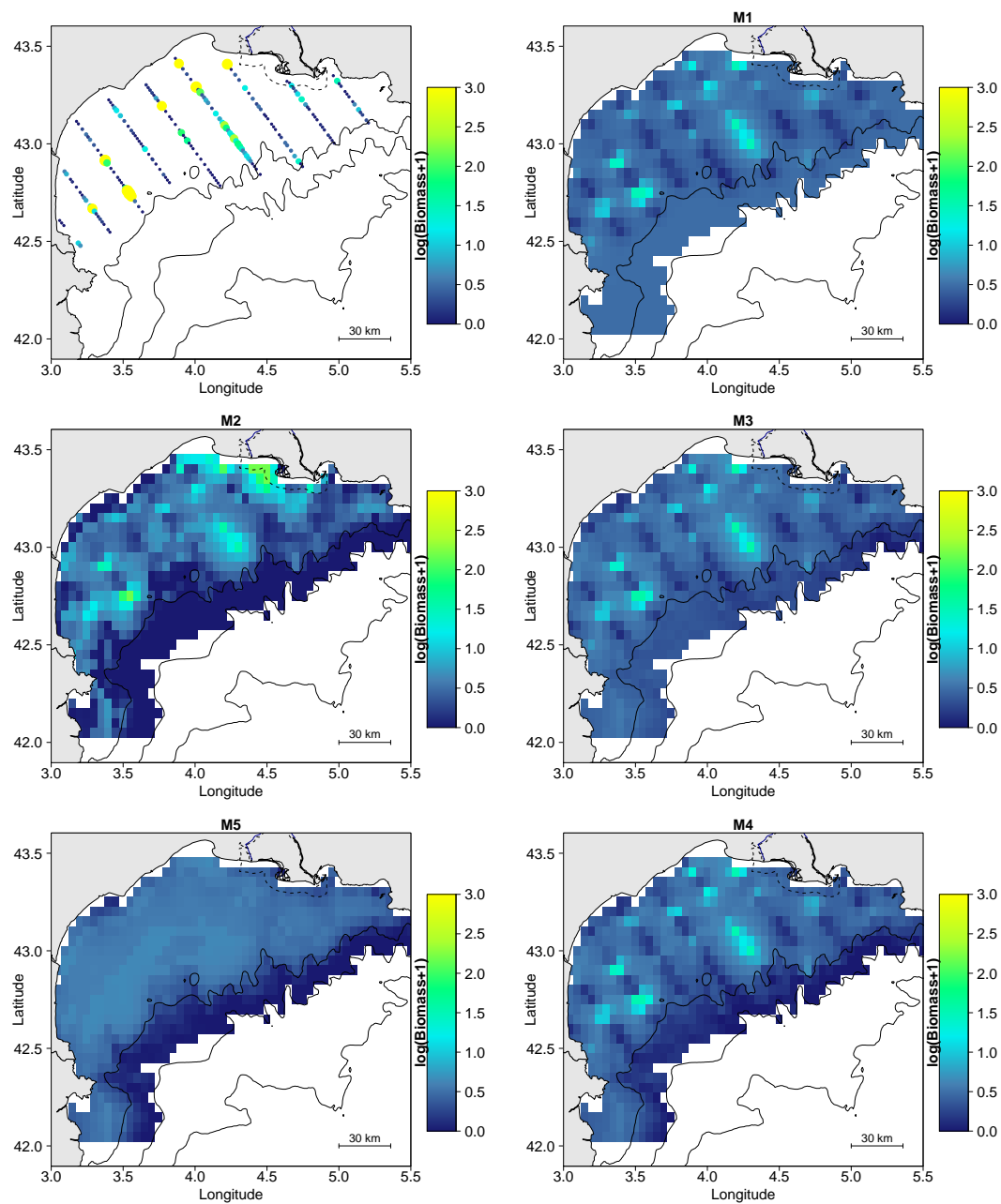


Figure A.7 : Raw data and comparison of model predictions (posterior median) for adult European sardine log-biomasses. The distribution during summer 2011 showed no obvious spatial structure. The black dotted line materializes the Carmague Natura 2000 protected area.

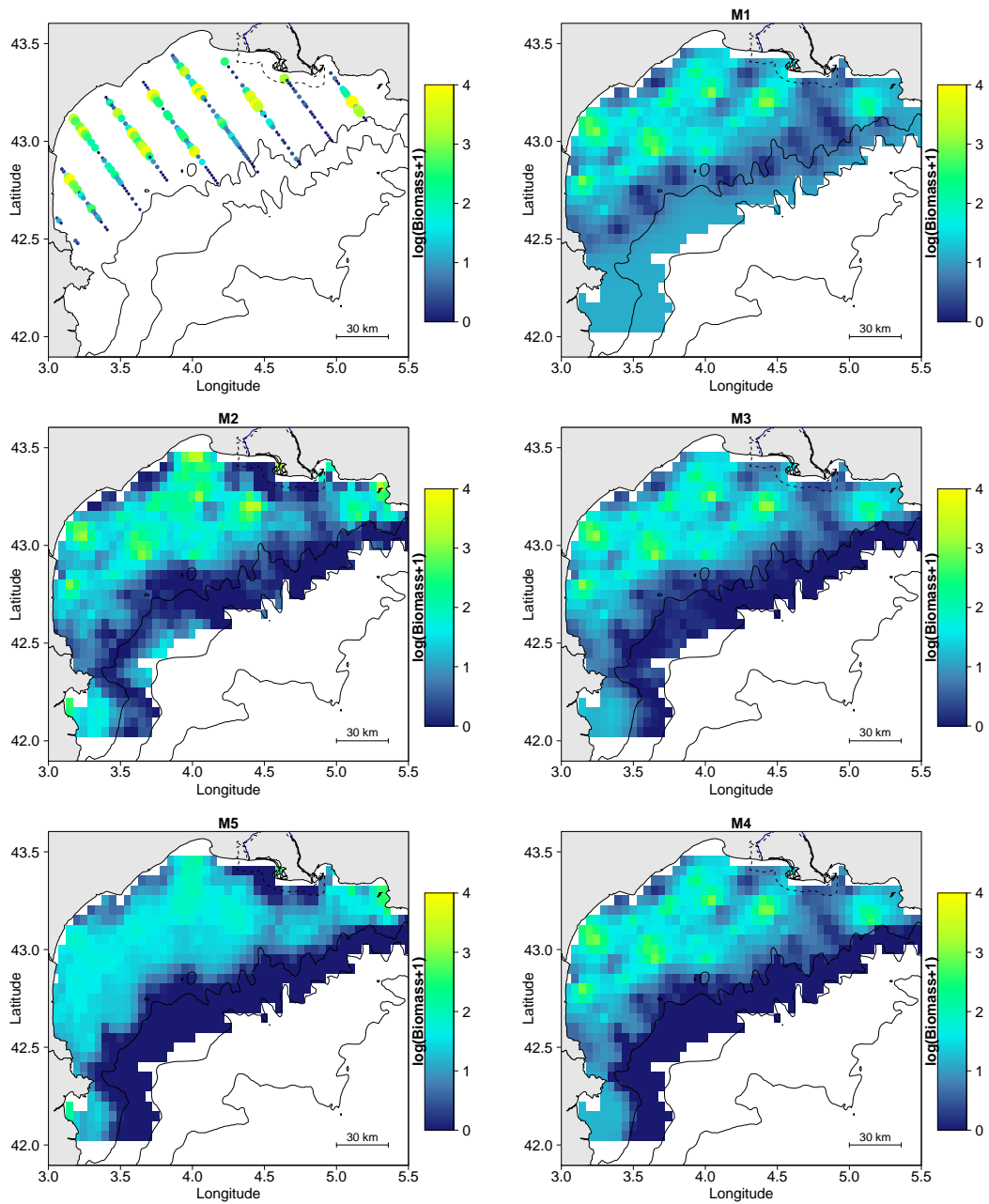


Figure A.8: : Raw data and comparison of model predictions (posterior median) for juvenile sprat log-biomasses. The distribution during summer 2011 showed a clear spatial structure. The black dotted line materializes the Carmague Natura 2000 protected area.

71 **Estimated Regression coefficients**

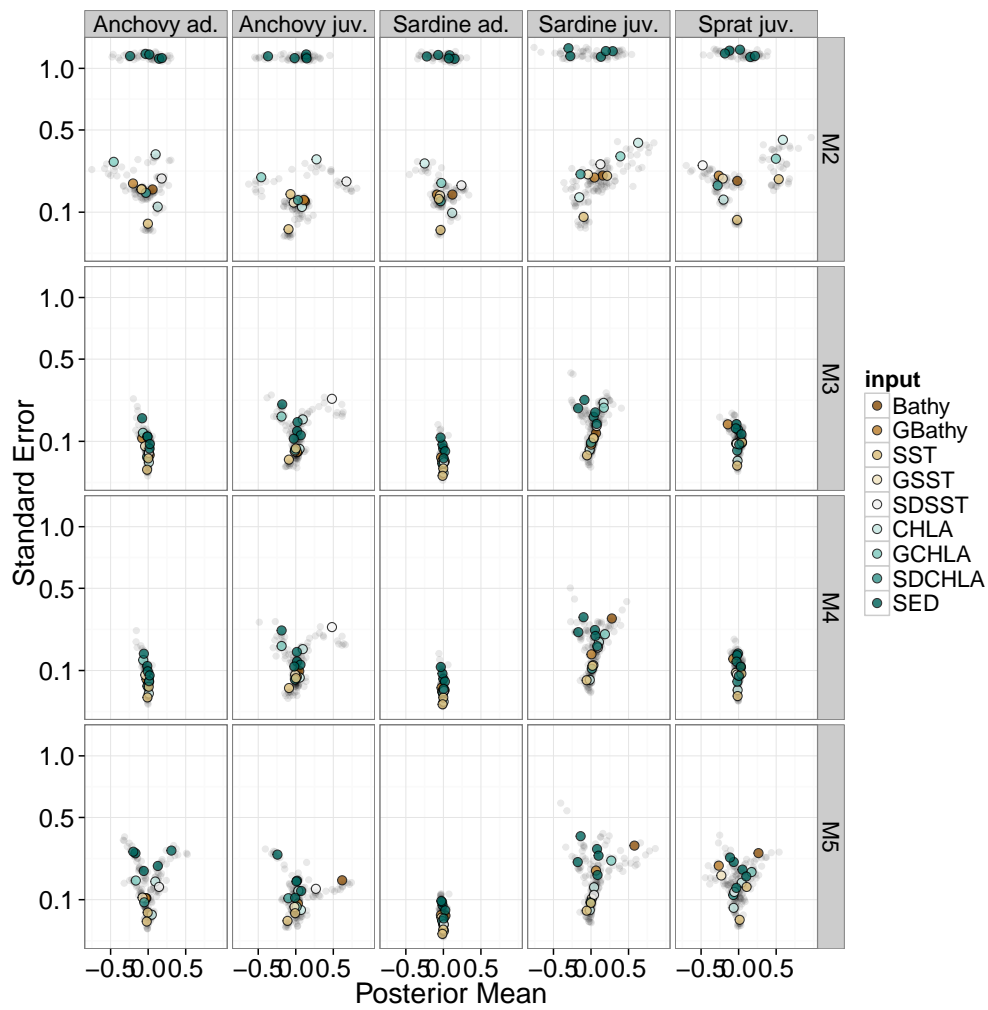


Figure A.9: : Plots of the estimated posterior standard error of the mean against the estimated posterior mean ( $\beta_p$ ). Estimates from  $\mathcal{M}_2$  were noisy, especially the coefficients linked to sediments. In contrast, the funnel shape of plots from  $\mathcal{M}_{3-5}$  illustrates how shrinkage greatly reduced both the estimated posterior mean and standard error of the mean.

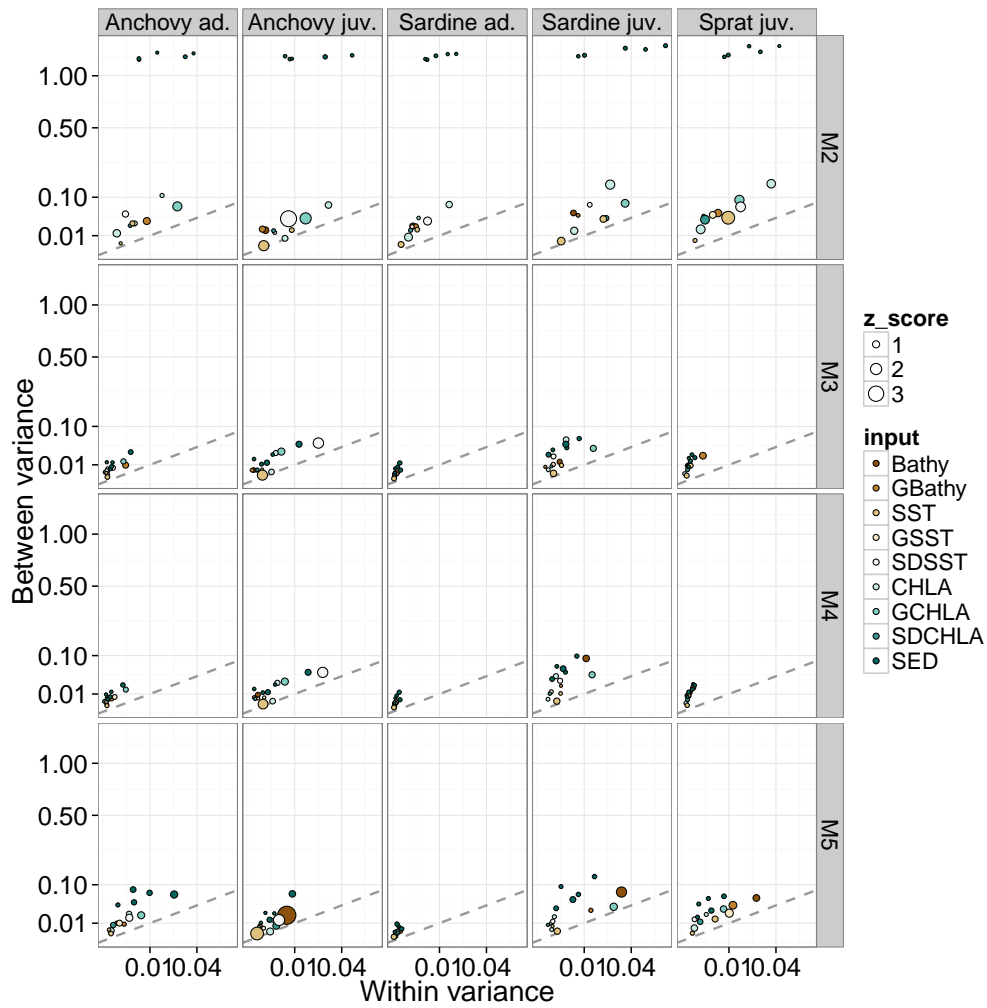


Figure A.10: : Plots of the variance in estimated regression coefficients  $\beta_p$  between cross-validations against the within-variance. Dots are proportional to a z-score (the ratio of estimated posterior mean to its standard error) of the coefficients averaged across the different cross-validation datasets. The between-variance was greatest for  $\mathcal{M}_2$  illustrate instability in estimation. In contrast, this between-variance was greatly reduced with  $\mathcal{M}_{3-5}$  and comparable to the within-variance. The grey dashed line shows the identity line (between-variance = within-variance).

## 72 **STAN code**

73 All model were fitted with `CmdStan v.2.8.0`, which is a command line interface to the Stan probabilistic modelling lan-  
74 guage (Stan Development Team, 2015). Each model was compiled as an executable; *e.g.* model  $\mathcal{M}_1$  was written down into  
75 a text file called `M1.stan`, and then compiled into the executable `M1.exe`. Cross-validation was performed by multiple  
76 calls to the different executables.



## 77 **Model $\mathcal{M}_1$**

```
78 data {
79   int<lower=1> n_obs; // sample size
80   vector<lower=0>[n_obs] BIOMASS; // response variable
81   matrix[n_obs,n_obs] DIST; // distance matrix
82 }
83
84 parameters {
85   real cst; // intercept
86   real<lower=0> sd_spatial; // sill
87   real<lower=0.1, upper=200> rho; // range parameter
88   real<lower=0> sd_res; // nugget
89   vector[n_obs] z; // spatial random effects, Cholesky parametrization
90 }
91
92 model {
93   // spatial effects
94   matrix[n_obs,n_obs] Sigma; // spatial covariance matrix
95   vector[n_obs] spatial; // spatial random effects
96   for ( i in 1:(n_obs-1) ) {
97     Sigma[i,i] <- square(sd_spatial);
98     for ( j in (i+1):n_obs ) {
99       // Matern covariance function of order 3/2
100      Sigma[i,j] <- (1.0+DIST[i,j]*sqrt(3.0)/rho)*exp(-DIST[i,j]*sqrt(3.0)/rho)*square(sd_spatial);
101      Sigma[j,i] <- Sigma[i,j];
102    };
103  };
104  Sigma[n_obs,n_obs] <- square(sd_spatial);
105  // 'Matt trick'
106  spatial <- cholesky_decompose(Sigma) * z;
107  // Priors
108  z ~ normal(0.0, 1.0);
109  rho ~ uniform(0.1, 200);
110  sd_res ~ cauchy(0.0, 1.0);
111  sd_spatial ~ cauchy(0, 1.0);
112  cst ~ student_t(7.0, 0.0, 10.0);
113  // likelihood
114  BIOMASS ~ normal(cst + spatial, sd_res);
115 }
```

## 116 **Model $\mathcal{M}_2$**

```
117 data {
118   int<lower=1> n_obs;
119   int<lower=1> n_pred; // number of predictors
120   real<lower=0> BIOMASS[n_obs];
121   matrix[n_obs,n_pred] X; // matrix of standardized predictors
122   matrix[n_obs,n_obs] DIST;
123 }
124
125 parameters {
126   real cst;
127   vector[n_pred] beta; // regression coefficients
128   real<lower=0> sd_spatial;
129   real<lower=0.1, upper=200> rho;
130   real<lower=0> sd_res;
131   vector[n_obs] z;
132 }
133
134 model {
135   // spatial effects
136   matrix[n_obs,n_obs] Sigma;
137   vector[n_obs] spatial;
138   for ( i in 1:(n_obs-1) ) {
139     Sigma[i,i] <- square(sd_spatial);
140     for ( j in (i+1):n_obs ) {
141       Sigma[i,j] <- (1.0+DIST[i,j]*sqrt(3.0)/rho)*exp(-DIST[i,j]*sqrt(3.0)/rho)*square(sd_spatial);
142       Sigma[j,i] <- Sigma[i,j];
143     };
144   };
145   Sigma[n_obs,n_obs] <- square(sd_spatial);
146   // 'Matt trick'
147   spatial <- cholesky_decompose(Sigma) * z;
148   // Priors
149   z ~ normal(0.0, 1.0);
150   rho ~ uniform(0.1, 200);
151   sd_res ~ cauchy(0.0, 1.0);
152   beta ~ student_t(7.0, 0.0, 2.5); // independent Student-t priors
153   sd_spatial ~ cauchy(0.0, 1.0);
154   cst ~ student_t(7.0, 0.0, 10.0);
```

```
155 // Likelihood
156 for ( i in 1:n_obs ) {
157   BIOMASS[i] ~ normal(cst + dot_product(beta, X[i]) + spatial[i], sd_res);
158 };
159 }
```

## 160 **Model $\mathcal{M}_3$**

```
161 data {
162   int<lower=1> n_obs;
163   int<lower=1> n_pred;
164   real<lower=0> BIOMASS[n_obs];
165   matrix[n_obs,n_pred] X;
166   matrix[n_obs,n_obs] DIST;
167 }
168
169 parameters {
170   real cst;
171   vector[n_pred] beta;
172   real<lower=0> global; // global shrinkage parameter
173   vector<lower=0>[n_pred] local; // local shrinkage parameters
174   real<lower=0> sd_spatial;
175   real<lower=0.1, upper=200> rho;
176   real<lower=0> sd_res;
177   vector[n_obs] z;
178 }
179
180 model {
181   // spatial effects
182   matrix[n_obs,n_obs] Sigma;
183   vector[n_obs] spatial;
184   for ( i in 1:(n_obs-1) ) {
185     Sigma[i,i] <- square(sd_spatial);
186     for ( j in (i+1):n_obs ) {
187       Sigma[i,j] <- (1.0+DIST[i,j]*sqrt(3.0)/rho)*exp(-DIST[i,j]*sqrt(3.0)/rho)*square(sd_spatial);
188       Sigma[j,i] <- Sigma[i,j];
189     };
190   };
191   Sigma[n_obs,n_obs] <- square(sd_spatial);
192   // 'Matt trick'
193   spatial <- cholesky_decompose(Sigma) * z;
194   // Priors
195   z ~ normal(0.0, 1.0);
196   rho ~ uniform(0.1, 200);
197   sd_res ~ cauchy(0.0, 1.0);
198   global ~ cauchy(0.0, sd_res);
```

```
199 local ~ cauchy(0.0, global);
200 beta ~ normal(0.0, local); // this is the horseshoe prior
201 sd_spatial ~ cauchy(0.0, 1.0);
202 cst ~ student_t(7.0, 0.0, 10.0);
203 // Likelihood
204 for ( i in 1:n_obs ) {
205   BIOMASS[i] ~ normal(cst + dot_product(beta, X[i]) + spatial[i], sd_res);
206 };
207 }
```

## 208 **Model $\mathcal{M}_4$**

```
209 data {
210   int<lower=1> n_obs;
211   int<lower=1> n_pred;
212   real<lower=0> BIOMASS[n_obs];
213   matrix[n_obs,n_pred] X;
214   matrix[n_obs,n_obs] DIST;
215   // indicator variable, =1 if BIOMASS=0, 0 otherwise
216   int<lower=0,upper=1> IS_ZERO[n_obs];
217 }
218
219 parameters {
220   real cst_beta;
221   real cst_alpha;
222   vector[n_pred] beta;
223   vector[n_pred] alpha; // coefficients for zero-inflated model
224   real<lower=0> global_beta;
225   vector<lower=0>[n_pred] local_beta;
226   real<lower=0> global_alpha;
227   vector<lower=0>[n_pred] local_alpha;
228   real<lower=0> sd_spatial;
229   real<lower=0.1, upper=200> rho;
230   real<lower=0> sd_res;
231   vector[n_obs] z;
232 }
233
234 model {
235   // spatial effects
236   matrix[n_obs,n_obs] Sigma;
237   vector[n_obs] spatial;
238   for ( i in 1:(n_obs-1) ) {
239     Sigma[i,i] <- square(sd_spatial);
240     for ( j in (i+1):n_obs ) {
241       Sigma[i,j] <- (1.0+DIST[i,j]*sqrt(3.0)/rho)*exp(-DIST[i,j]*sqrt(3.0)/rho)*square(sd_spatial);
242       Sigma[j,i] <- Sigma[i,j];
243     };
244   };
245   Sigma[n_obs,n_obs] <- square(sd_spatial);
246   // 'Matt trick'
```

```

247 spatial <- cholesky_decompose(Sigma) * z;
248 // Priors
249 z ~ normal(0.0, 1.0);
250 rho ~ uniform(0.1, 200);
251 sd_res ~ cauchy(0.0, 1.0);
252 global_beta ~ cauchy(0, sd_res);
253 local_beta ~ cauchy(0, global_beta);
254 beta ~ normal(0, local_beta);
255 global_alpha ~ cauchy(0.0, 1.0);
256 local_alpha ~ cauchy(0.0, global_alpha);
257 alpha ~ normal(0.0, local_alpha);
258 sd_spatial ~ cauchy(0.0, 1.0);
259 cst_beta ~ student_t(7.0, 0.0, 10.0);
260 cst_alpha ~ student_t(7.0, 0.0, 10.0);
261 // Likelihood
262 for ( i in 1:n_obs ) {
263   real mu;
264   real prob_zero;
265   real u ;
266   mu <- cst_beta + dot_product(beta, X[i]) + spatial[i];
267   // data augmentation: probit model for zero-inflation
268   prob_zero <- Phi(cst_alpha + dot_product(alpha, X[i]));
269   // this is the likelihood of a zero-inflated normal model
270   u <- if_else( IS_ZERO[i],
271     log( prob_zero + (1-prob_zero)*exp(normal_log(BIOMASS[i], mu, sd_res)) ),
272     log1m(prob_zero) + normal_log(BIOMASS[i], mu, sd_res) );
273   increment_log_prob(u);
274 };
275 }

```

## 276 **Model $\mathcal{M}_5$**

```
277 data {
278   int<lower=1> n_obs;
279   int<lower=1> n_pred;
280   real<lower=0> BIOMASS[n_obs];
281   matrix[n_obs,n_pred] X;
282   int<lower=0,upper=1> IS_ZERO[n_obs];
283 }
284
285 parameters {
286   real cst_beta;
287   real cst_alpha;
288   vector[n_pred] beta;
289   vector[n_pred] alpha;
290   real<lower=0> global_beta;
291   vector<lower=0>[n_pred] local_beta;
292   real<lower=0> global_alpha;
293   vector<lower=0>[n_pred] local_alpha;
294   real<lower=0> sd_res;
295 }
296
297 model {
298   // Priors
299   sd_res ~ cauchy(0.0, 1.0);
300   global_beta ~ cauchy(0, sd_res);
301   local_beta ~ cauchy(0, global_beta);
302   beta ~ normal(0, local_beta);
303   global_alpha ~ cauchy(0.0, 1.0);
304   local_alpha ~ cauchy(0.0, global_alpha);
305   alpha ~ normal(0.0, local_alpha);
306   cst_beta ~ student_t(7.0, 0.0, 10.0);
307   cst_alpha ~ student_t(7.0, 0.0, 10.0);
308   // Likelihood
309   for ( i in 1:n_obs ) {
310     real mu;
311     real prob_zero;
312     real u ;
313     mu <- cst_beta + dot_product(beta, X[i]);
314     // data augmentation: probit model for zero-inflation
```



```
315 prob_zero <- Phi(cst_alpha + dot_product(alpha, X[i]));
316 // this is the likelihood of a zero-inflated normal model
317 u <- if_else( IS_ZERO[i],
318 log( prob_zero + (1-prob_zero)*exp(normal_log(BIOMASS[i], mu, sd_res)) ),
319 log1m(prob_zero) + normal_log(BIOMASS[i], mu, sd_res) );
320 increment_log_prob(u);
321 };
322 }
```

## 323 **References**

324 Stan Development Team 2015. Stan Modeling Language Users Guide and Reference Manual, Version 2.7.0.