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[a,d,e], Claire Saraux[b], and Clara Péron[c,d]

[a]Observatoire PELAGIS UMS 3462, Université de La Rochelle, 5 Allées de l'Océan 17000 La Rochelle, France

[b]Ifremer (Institut Français de Recherche pour l'Exploitation de la Mer), UMR MARBEC, Sète, France

[c]Institute for Marine and Antarctic Studies, University of Tasmania and Australian Antarctic Division, 203 Channel highway, Kingston, Tasmania 7050, Australia

[d]Écologie Spatiale des Populations, Centre d'Écologie Fonctionnelle et Évolutive, 1919 route de Mende, 34293 Montpellier cedex 5, France

[e]authier@gmail.com

February 15, 2016

Contents

List of Figures

List of Tables

Horseshoe Prior

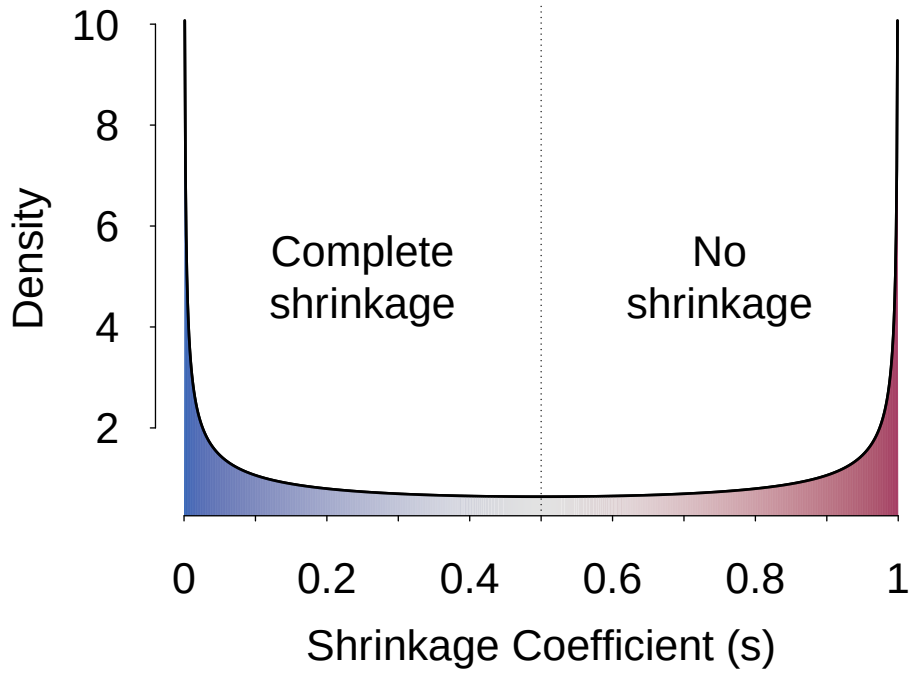

Figure A.1: : Probability density function on the shrinkage coefficient induced with a Horseshoe prior (that is, shrinkage profile of the horseshoe prior). Denoting β_{unshrunk} and β_{shrunk} the unshrunk and shrunk regression coefficient, the shrinkage coefficient s is such that $\beta_{\text{shrunk}} = s \times \beta_{\text{unshrunk}}$. 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.

Data

Boxplot

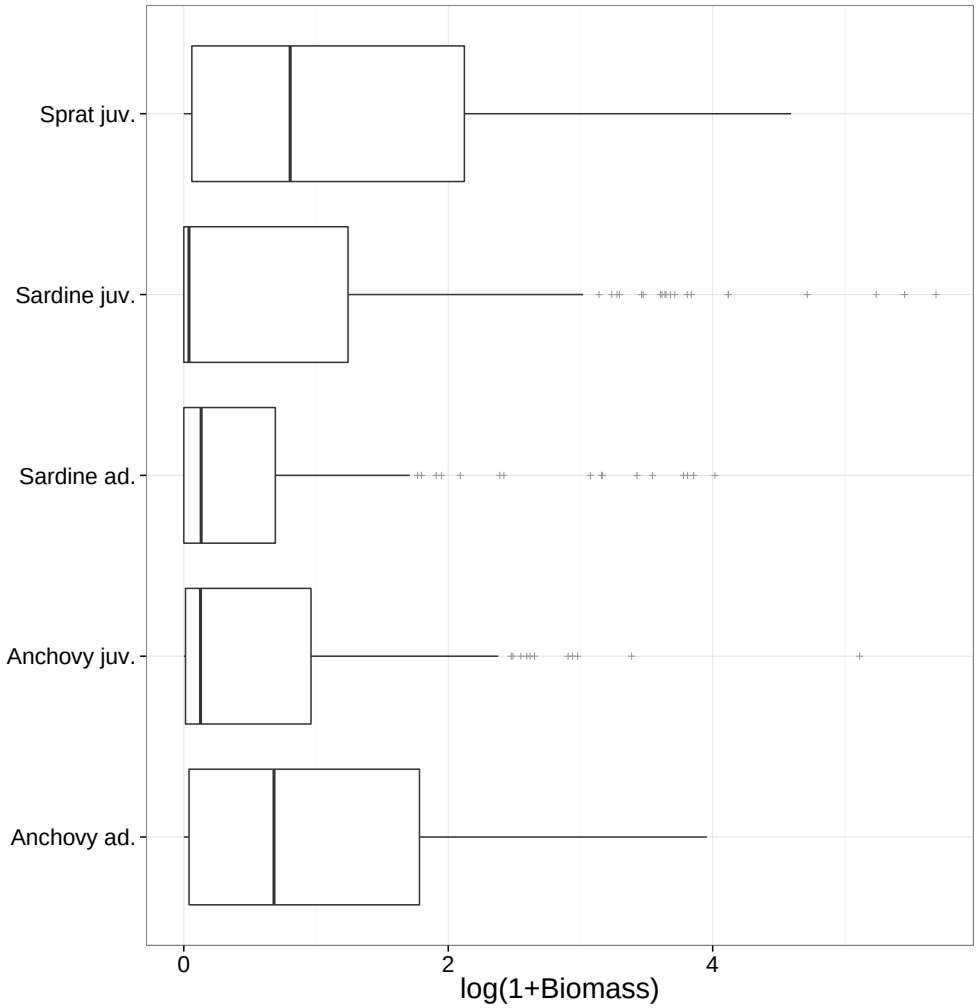

Figure A.2: : Boxplots of the small pelagic fish biomass data during the 2011 PELMED survey. Boxes represent the interquartile range and the vertical line within, the median. Whiskers extend to the lower and higher hinge defined as $1.5\times$ the interquartile range. All data are right-skewed betraying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

Environmental Inputs: Source and Resolution

Table A.1: Data Sources of Environmental Inputs.

Input	Spatial resolution	Temporal	Source	url
Bathymetry	0.01666°		MODIS\Aqua	http://coastwatch.pfel.noaa.gov/coastwatch/CWBrowserWW360.jsp
Sea Surface Temperature	0.05°	weekly	MODIS\Aqua	http://coastwatch.pfel.noaa.gov/coastwatch/CWBrowserWW360.jsp
Chlorophyll a Concentration	0.05°	weekly	MODIS\Aqua	http://coastwatch.pfel.noaa.gov/coastwatch/CWBrowserWW360.jsp

Correlation matrices

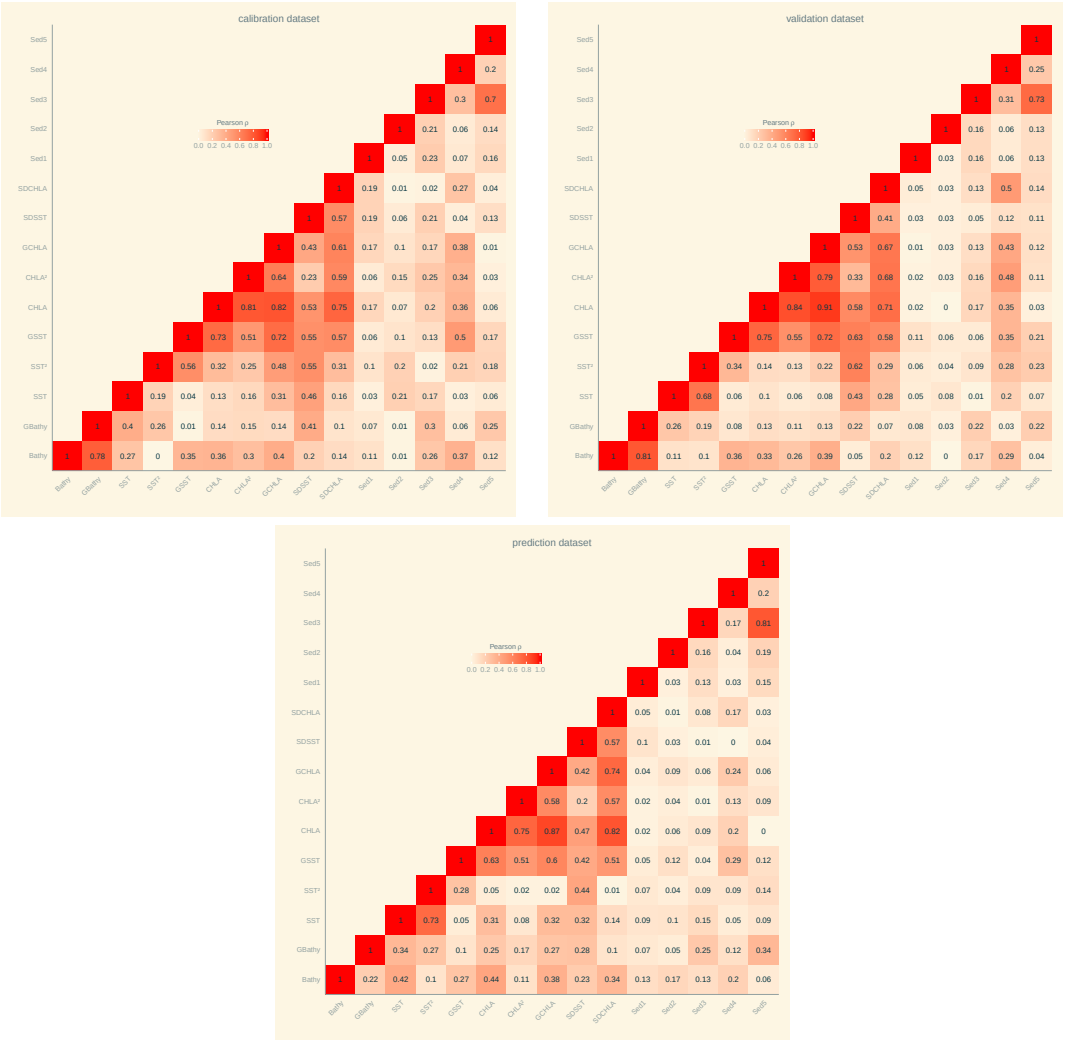

Figure A.3: : Correlation matrices for the different datasets. Empirical pairwise correlations are printed and color-coded (according their absolute magnitude). Calibration and validation data have a similar correlation structure, which is slightly different from that of the prediction data. These graphs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

Maps

Juvenile anchovies

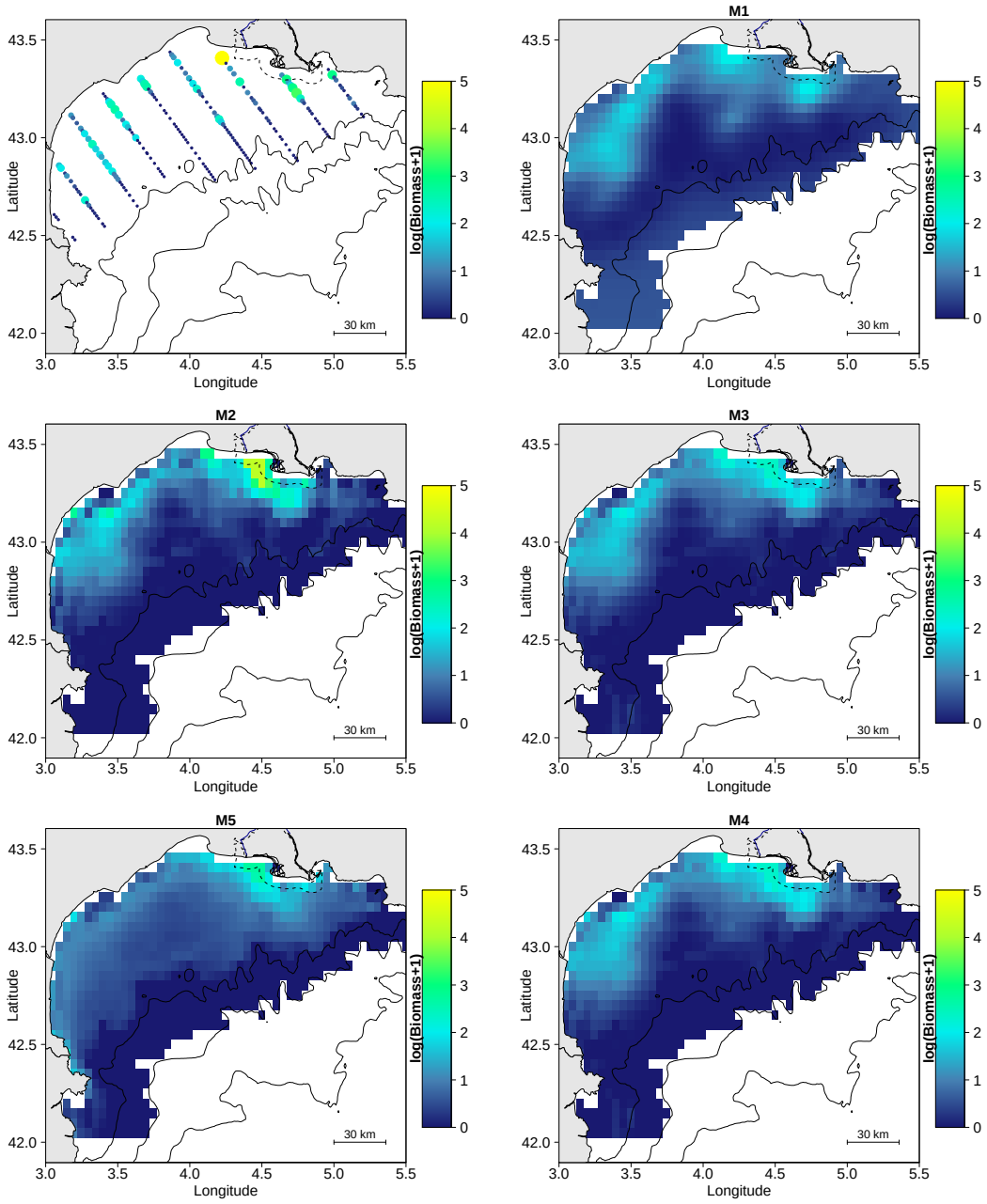

Figure A.4: : Raw data and comparison of model predictions (posterior median) for juvenile European anchovies log-biomasses. The distribution during summer 2011 showed a clear spatial structure. The black dotted line materializes the Carmague Natura 2000 protected area.

Adult anchovies

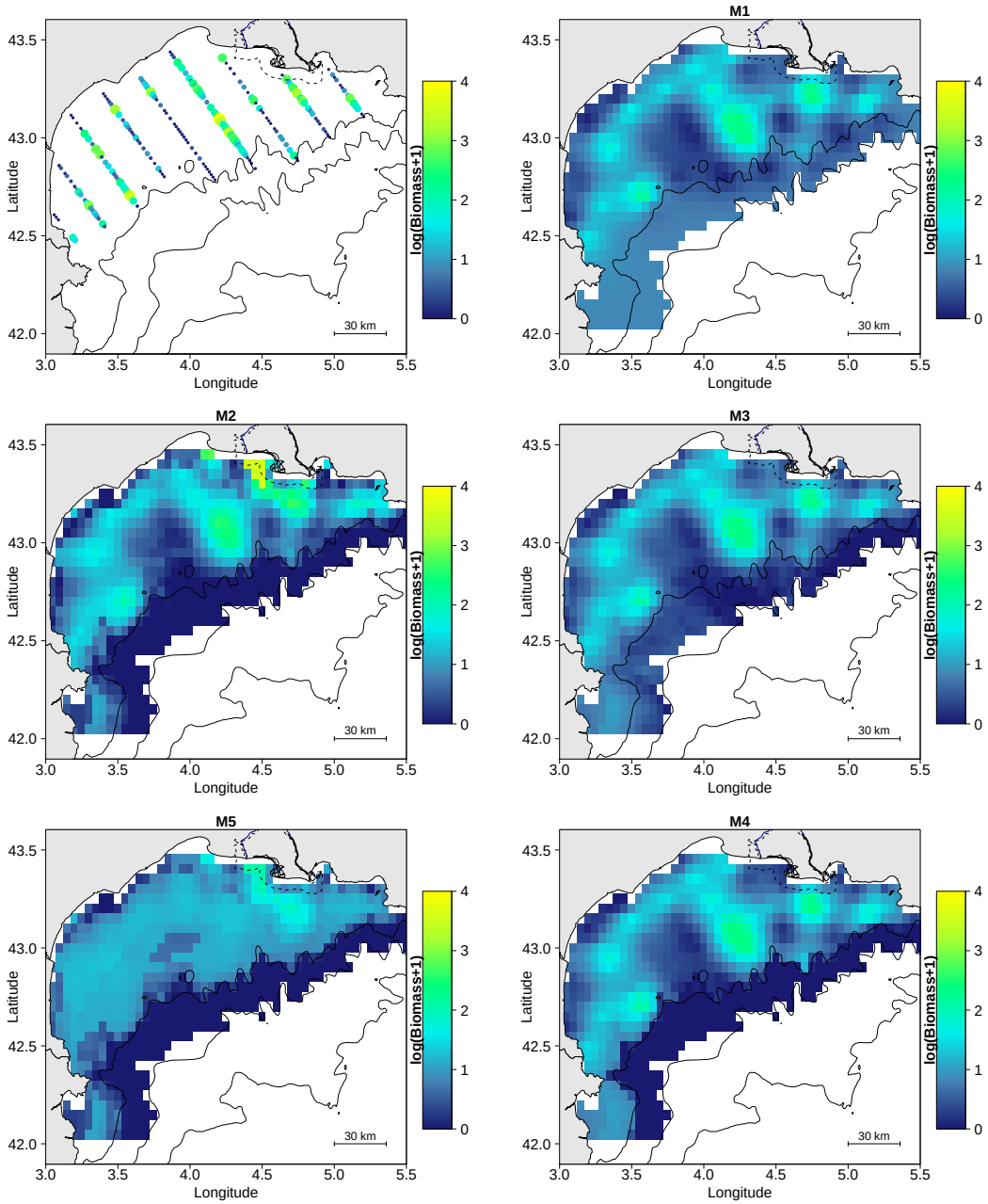

Figure A.5: : Raw data and comparison of model predictions (posterior median) for adult European anchovies log-biomasses. The distribution during summer 2011 showed a clear spatial structure. The black dotted line materializes the Carmague Natura 2000 protected area.

Juvenile sardines

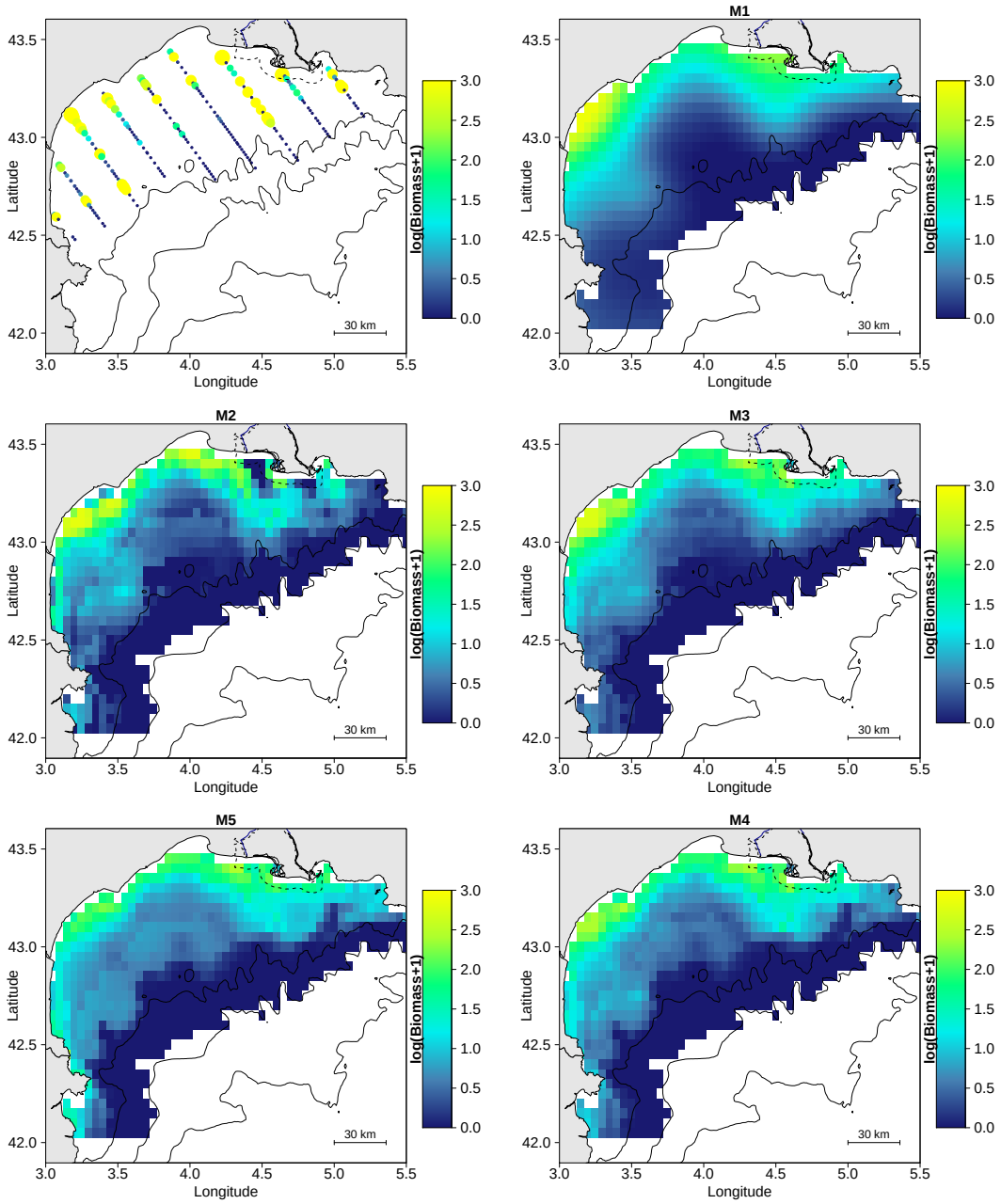

Figure A.6: : Raw data and comparison of model predictions (posterior median) for juvenile European sardine log-biomasses. The distribution during summer 2011 showed a clear pattern linked to depth: juvenile sardines were abundant very close to the coastline of the Gulf of Lion. The black dotted line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

Adult sardines

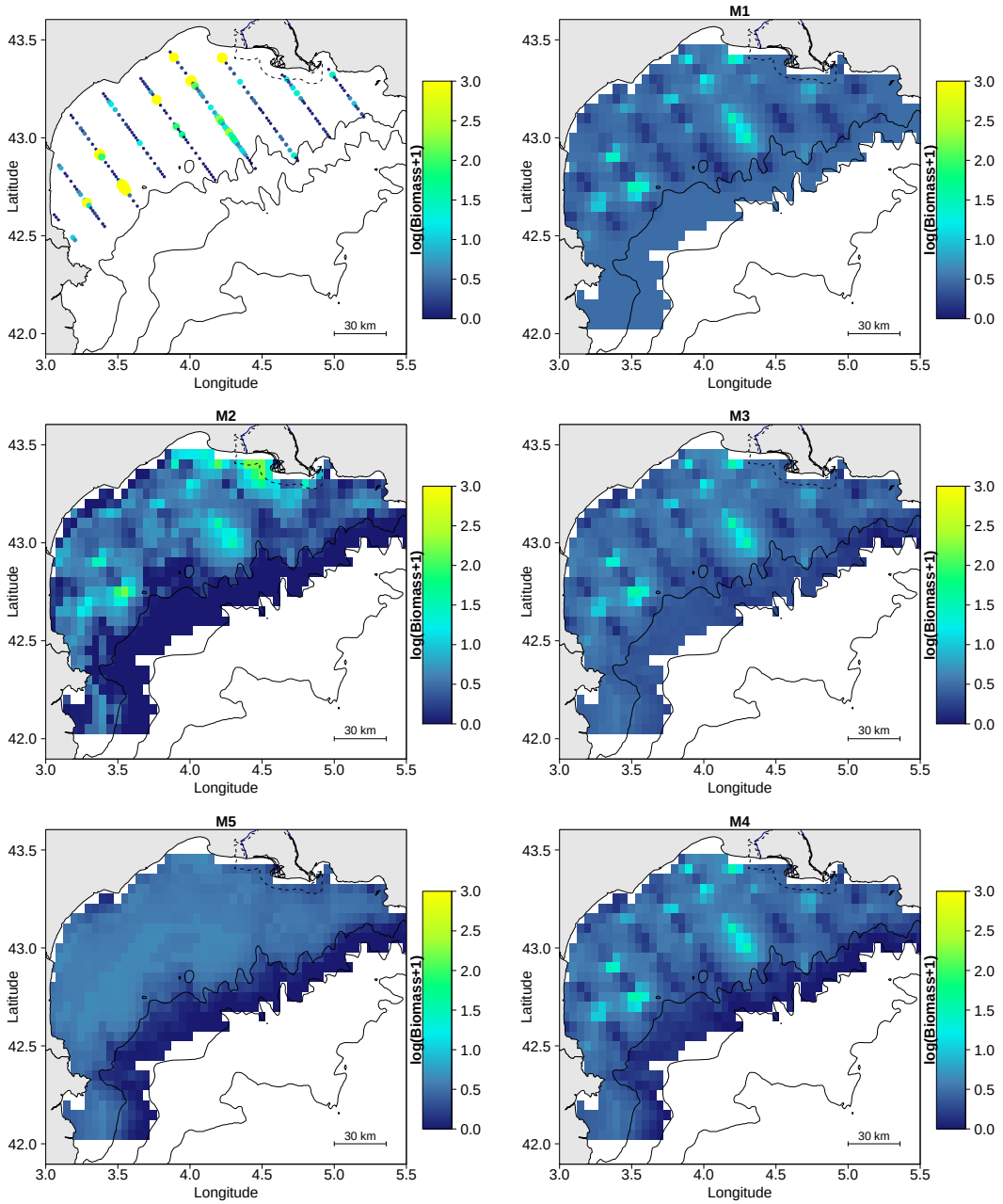

Figure A.7: : Raw data and comparison of model predictions (posterior median) for adult European sardine log-biomasses. The distribution during summer 2011 showed no obvious spatial structure. The black dotted line materializes the Carmague Natura 2000 protected area.

Juvenile sprats

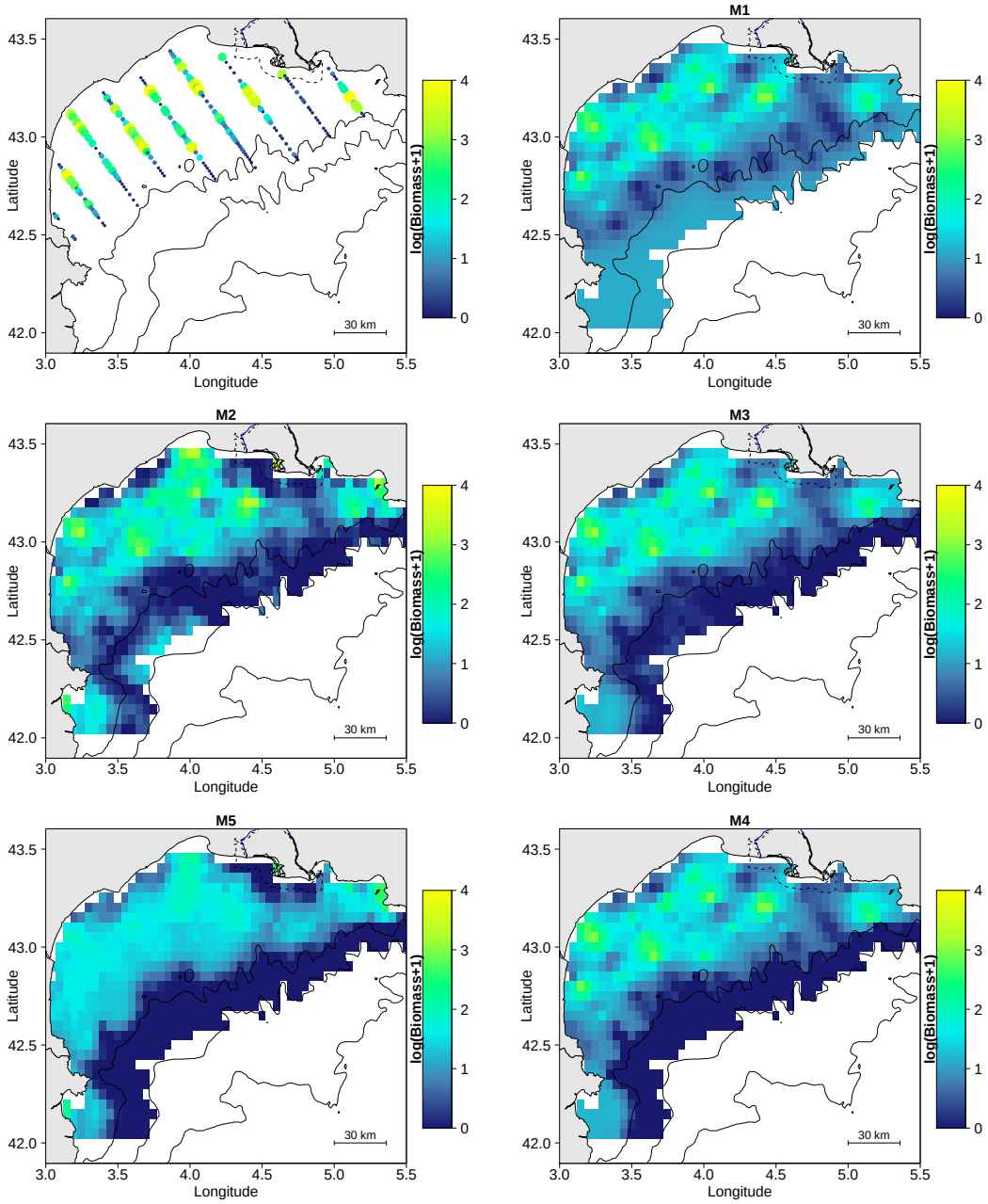

Figure A.8: : Raw data and comparison of model predictions (posterior median) for juvenile sprat log-biomasses. The distribution during summer 2011 showed a clear spatial structure. The black dotted line materializes the Carmague Natura 2000 protected area.

Estimated Regression coefficients

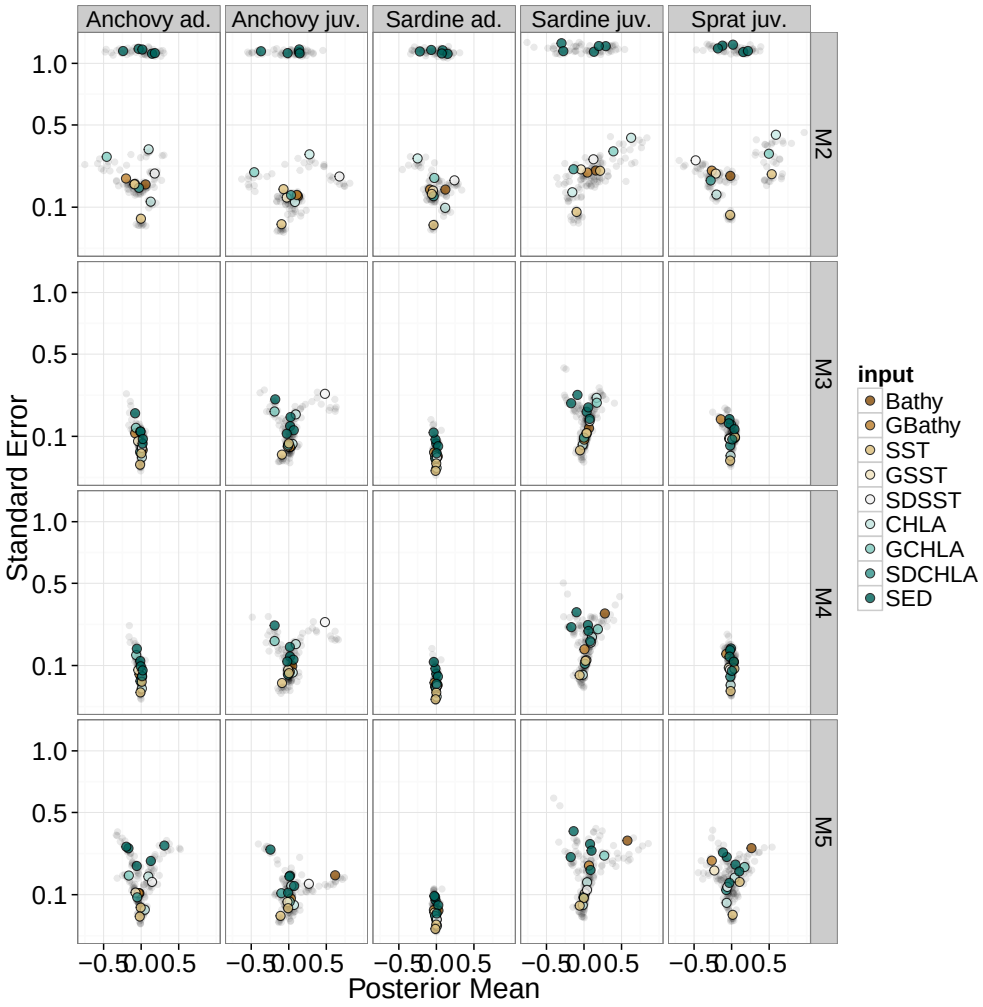

Figure A.9: : Plots of the estimated posterior standard error of the mean against the estimated posterior mean (β_p). Estimates from $\mathcal{M}_2$ were noisy, especially the coefficients linked to sediments. In contrast, the funnel shape of plots from $\mathcal{M}_{3-5}$ illustrates how shrinkage greatly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

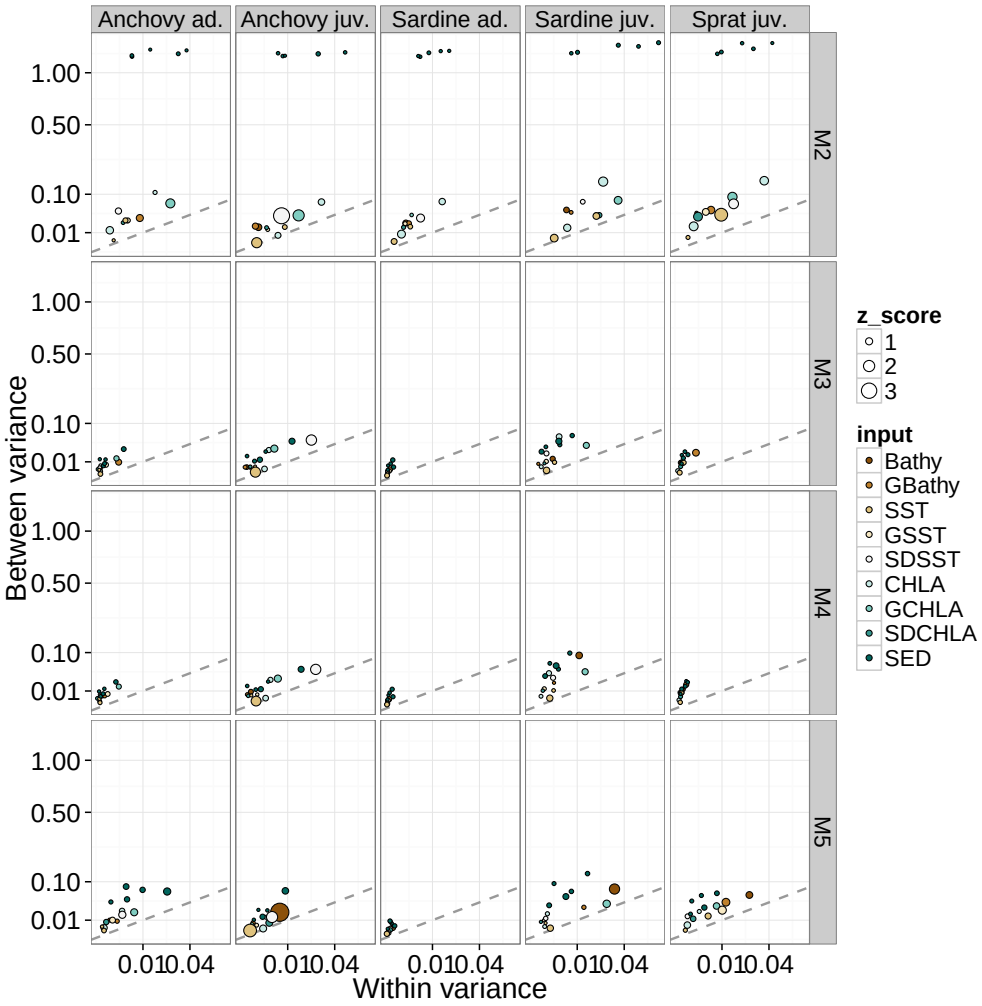

Figure A.10: : Plots of the variance in estimated regression coefficients β_p between cross-validations against the within-variance. Dots are proportional to a z-score (the ratio of estimated posterior mean to its standard error) of the coefficients averaged across the different cross-validation datasets. 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).

STAN code

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

Model $\mathcal{M}_1$

```
data {
 int<lower=1> n_obs; // sample size
 vector<lower=0>[n_obs] BIOMASS; // response variable
 matrix[n_obs,n_obs] DIST; // distance matrix
}

parameters {
 real cst; // intercept
 real<lower=0> sd_spatial; // sill
 real<lower=0.1, upper=200> rho; // range parameter
 real<lower=0> sd_res; // nugget
 vector[n_obs] z; // spatial random effects, Cholesky parametrization
}

model {
 // spatial effects
 matrix[n_obs,n_obs] Sigma; // spatial covariance matrix
 vector[n_obs] spatial; // spatial random effects
 for ( i in 1:(n_obs-1) ) {
  Sigma[i,i] <- square(sd_spatial);
  for ( j in (i+1):n_obs ) {
   // Matern covariance function of order 3/2
   Sigma[i,j] <- (1.0+DIST[i,j]*sqrt(3.0)/rho)*exp(-DIST[i,j]*sqrt(3.0)/rho)*square(sd_spatial);
   Sigma[j,i] <- Sigma[i,j];
  };
 };
 Sigma[n_obs,n_obs] <- square(sd_spatial);
 // 'Matt trick'
 spatial <- cholesky_decompose(Sigma) * z;
 // Priors
 z ~ normal(0.0, 1.0);
 rho ~ uniform(0.1, 200);
 sd_res ~ cauchy(0.0, 1.0);
 sd_spatial ~ cauchy(0 ,1.0);
 cst ~ student_t(7.0, 0.0, 10.0);
 // likelihood
 BIOMASS ~ normal(cst + spatial, sd_res);
}
```

Model $\mathcal{M}_2$

```
data {
 int<lower=1> n_obs;
 int<lower=1> n_pred; // number of predictors
 real<lower=0> BIOMASS[n_obs];
 matrix[n_obs,n_pred] X; // matrix of standardized predictors
 matrix[n_obs,n_obs] DIST;
}

parameters {
 real cst;
 vector[n_pred] beta; // regression coefficients
 real<lower=0> sd_spatial;
 real<lower=0.1, upper=200> rho;
 real<lower=0> sd_res;
 vector[n_obs] z;
}

model {
 // spatial effects
 matrix[n_obs,n_obs] Sigma;
 vector[n_obs] spatial;
 for ( i in 1:(n_obs-1) ) {
  Sigma[i,i] <- square(sd_spatial);
  for ( j in (i+1):n_obs ) {
   Sigma[i,j] <- (1.0+DIST[i,j]*sqrt(3.0)/rho)*exp(-DIST[i,j]*sqrt(3.0)/rho)*square(sd_spatial);
   Sigma[j,i] <- Sigma[i,j];
  };
 };
 Sigma[n_obs,n_obs] <- square(sd_spatial);
 // 'Matt trick'
 spatial <- cholesky_decompose(Sigma) * z;
 // Priors
 z ~ normal(0.0, 1.0);
 rho ~ uniform(0.1, 200);
 sd_res ~ cauchy(0.0, 1.0);
 beta ~ student_t(7.0, 0.0, 2.5); // independent Student-t priors
 sd_spatial ~ cauchy(0.0, 1.0);
 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  }
```

Model $\mathcal{M}_3$

```
data {
 int<lower=1> n_obs;
 int<lower=1> n_pred;
 real<lower=0> BIOMASS[n_obs];
 matrix[n_obs,n_pred] X;
 matrix[n_obs,n_obs] DIST;
}

parameters {
 real cst;
 vector[n_pred] beta;
 real<lower=0> global; // global shrinkage parameter
 vector<lower=0>[n_pred] local; // local shrinkage parameters
 real<lower=0> sd_spatial;
 real<lower=0.1, upper=200> rho;
 real<lower=0> sd_res;
 vector[n_obs] z;
}

model {
 // spatial effects
 matrix[n_obs,n_obs] Sigma;
 vector[n_obs] spatial;
 for ( i in 1:(n_obs-1) ) {
  Sigma[i,i] <- square(sd_spatial);
  for ( j in (i+1):n_obs ) {
   Sigma[i,j] <- (1.0+DIST[i,j]*sqrt(3.0)/rho)*exp(-DIST[i,j]*sqrt(3.0)/rho)*square(sd_spatial);
   Sigma[j,i] <- Sigma[i,j];
  };
 };
 Sigma[n_obs,n_obs] <- square(sd_spatial);
 // 'Matt trick'
 spatial <- cholesky_decompose(Sigma) * z;
 // Priors
 z ~ normal(0.0, 1.0);
 rho ~ uniform(0.1, 200);
 sd_res ~ cauchy(0.0, 1.0);
 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  }
```

Model $\mathcal{M}_4$

```
data {
 int<lower=1> n_obs;
 int<lower=1> n_pred;
 real<lower=0> BIOMASS[n_obs];
 matrix[n_obs,n_pred] X;
 matrix[n_obs,n_obs] DIST;
 // indicator variable, =1 if BIOMASS=0, 0 otherwise
 int<lower=0,upper=1> IS_ZERO[n_obs];
}

parameters {
 real cst_beta;
 real cst_alpha;
 vector[n_pred] beta;
 vector[n_pred] alpha; // coefficients for zero-inflated model
 real<lower=0> global_beta;
 vector<lower=0>[n_pred] local_beta;
 real<lower=0> global_alpha;
 vector<lower=0>[n_pred] local_alpha;
 real<lower=0> sd_spatial;
 real<lower=0.1, upper=200> rho;
 real<lower=0> sd_res;
 vector[n_obs] z;
}

model {
// spatial effects
 matrix[n_obs,n_obs] Sigma;
 vector[n_obs] spatial;
 for ( i in 1:(n_obs-1) ) {
  Sigma[i,i] <- square(sd_spatial);
  for ( j in (i+1):n_obs ) {
   Sigma[i,j] <- (1.0+DIST[i,j]*sqrt(3.0)/rho)*exp(-DIST[i,j]*sqrt(3.0)/rho)*square(sd_spatial);
   Sigma[j,i] <- Sigma[i,j];
  };
 };
 Sigma[n_obs,n_obs] <- square(sd_spatial);
 // '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  }
```

Model $\mathcal{M}_5$

```
data {
 int<lower=1> n_obs;
 int<lower=1> n_pred;
 real<lower=0> BIOMASS[n_obs];
 matrix[n_obs,n_pred] X;
 int<lower=0,upper=1> IS_ZERO[n_obs];
}

parameters {
 real cst_beta;
 real cst_alpha;
 vector[n_pred] beta;
 vector[n_pred] alpha;
 real<lower=0> global_beta;
 vector<lower=0>[n_pred] local_beta;
 real<lower=0> global_alpha;
 vector<lower=0>[n_pred] local_alpha;
 real<lower=0> sd_res;
}

model {
 // Priors
 sd_res ~ cauchy(0.0, 1.0);
 global_beta ~ cauchy(0, sd_res);
 local_beta ~ cauchy(0, global_beta);
 beta ~ normal(0, local_beta);
 global_alpha ~ cauchy(0.0, 1.0);
 local_alpha ~ cauchy(0.0, global_alpha);
 alpha ~ normal(0.0, local_alpha);
 cst_beta ~ student_t(7.0, 0.0, 10.0);
 cst_alpha ~ student_t(7.0, 0.0, 10.0);
 // Likelihood
 for ( i in 1:n_obs ) {
  real mu;
  real prob_zero;
  real u ;
  mu <- cst_beta + dot_product(beta, X[i]);
  // data augmentation: probit model for zero-inflation
```

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

References

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