

CMEE MiniProject:

You are what you eat; Model optimisation for predicting predator size from prey size and associated abiotic factors, in global marine systems.



Katherine Bickerton, *Imperial College, London*

Word Count = [2597]

1 You are what you eat; Model optimisation for predicting predator size from
2 prey size and associated abiotic factors, in global marine systems.

3 **Abstract**

4 Marine systems are one of the most difficult to monitor on the planet and one of the most at risk.
5 Those at higher trophic levels are often the most at risk due to dependence on the stability of the
6 rest of the food web. In this study, I aim to compare linear regression, generalised additive (GAM)
7 and generalised linear mixed effects (GLMM) models for predicting predator size, based upon prey
8 size, feeding interaction, habitat, temperature and depth. The optimal model was a generalised
9 additive model (GAM), with prey mass, temperature and depth as explanatory variables. Feeding
10 interaction was also found to have an effect using GLMM, but violated the assumptions of GAM,
11 therefore could not be included in the optimal model. These findings implies that predator size could
12 be predicted from prey size and abiotic factors, which are less labour intensive to collect and could
13 improve efficiency of conservation research.

14 **Introduction**

15 With ever increasing risks to biodiversity (Kerr and Currie, 1995), and the close proximity of a sixth
16 mass extinction event (Barnosky et al., 2011), understanding the dynamics of ecological
17 systems is increasingly important. One of the main threats to biodiversity is overexploitation, where
18 many populations have been reduced to an unsustainable level. This is especially true of marine
19 systems where, despite regulations, approximately 63% of fish stocks are below a the threshold
20 population size required to persist (Heithaus et al., 2008).

21

22 Population data is particularly difficult to collect in marine systems, especially on pelagic organisms in
23 deep oceans, often restricting data collection to shallow coastal areas. One means of supplementing
24 information in data poor systems is to construct predictive models from existing data and extrapolate
25 with a variety of parameters (Levins, 1966). This method can also aid in targeting studies to test likely

26 models and aid in mitigation against overexploitation.

27

28 One classical model within ecosystem dynamics is that between predators and prey (Freedman and
29 Waltman, 1983). Information regarding predator and prey physiology and consumption is some of the
30 most easily available in marine systems. The most common method is stomach contents analysis of
31 organisms caught as by-catch or stranded (Turesson et al., 2002). This allows understanding of the
32 position of each species within the food chain, and which prey sources are most important to or most
33 threatened by specific predators.

34

35 This study uses an existing dataset of marine predators, both in coastal and pelagic zones to con-
36 struct a model that could be used to predict predator size based upon available prey. The main aim
37 is to investigate the relationship between predator mass and prey mass, on a global scale, across a
38 range of habitats, temperatures and depths. As opposed to defining a typical "null hypothesis", this
39 study is considering several different models for predator mass, based on the above factors, and will
40 select and interpret the optimal model for the system.

41 **Methods**

42 **Data Compilation**

43 The data used for this was complied by Barnes et al, 2008, and comprises 19,625 records of 57
44 different marine predator species, from 27 locations globally, and 18 studies (Barnes et al., 2008).
45 The dataset included predator species, masses and lengths of predators and prey, habitat, location,
46 type of feeding interaction, depth, mean annual air temperature and mean annual precipitation, for
47 each predation record.

48

49 As this study aimed to examine relationship between predator and prey sizes, I chose one dimension,
50 mass, as the measure of size, and excluded length as the two variables would violate the assumption
51 of independence. The mean average of predator and prey masses, temperature and depth for each

52 predator species were calculated. For the categorical variables, habitat and feeding interaction, the
53 most common for each predator species was selected. Habitats were sorted into coastal and pelagic,
54 dependent on the definitions of the habitats mentioned in each study. Feeding interactions were
55 classified into either piscivorous (feeding only on fish), predacious (generalist predators, though diet
56 may include fish), and planktivorous (feeding only on plankton), as defined by Barnes et al, 2008.
57 A new data frame was compiled containing the averages calculated for each species, then used in
58 model fitting.

59 **Model Building and Fitting**

60 I chose to compare three different types of model: a linear regression model, a generalised additive
61 model (GAM) and a generalised linear mixed effects model (GLMM). The response variable, predator
62 mass, did not have a linear distribution, therefore it was log transformed for use in the linear models,
63 and as the response variable must be consistent for models to be comparable, predator mass was
64 also log transformed for the GAM.

65

66 When building the models, I started with the assumption that predator mass would be a function
67 of the average mass of prey species. Additional explanatory variables were also considered and
68 selected for in each model, dependent on whether they improved model fit, described below. The
69 additional variables considered were: feeding interaction, habitat, depth and temperature. For each
70 model, combinations of the explanatory variable were tested and the Akaike Information Criterion
71 (AIC) calculated for each model, then compared to give the best fit to the data. The optimal models
72 from each type of model were then compared, again using AIC, to find the model that overall best
73 described the data. The linear regression model comparison was carried out using the "step" function
74 in R, however no equivalent function was available for GAMs and GLMMs, so comparisons were
75 carried out manually using anova tests and AIC to systematically reject variables which decreased
76 the fit of the model.

77

78 **Model 1: Linear Regression Model** - The linear regression model for predator mass was defined
79 as below, where m = mass in grams and ϵ is the error not explained by the explanatory variable.

$$m_{pred} \sim m_{prey} + \epsilon$$

80

81 Linear regression models assume that the data is linear, therefore predator mass, prey mass and
82 mean depth were log transformed. It also assumes collinearity and independence (Zuur et al., 2009),
83 therefore any explanatory variables that could interact must be removed. Due to this, habitat and
84 feeding interaction were removed from the model as they correlated with depth and prey mass re-
85 spectively. Finally, normally distributed variables are required, which occurred when the data was log
86 transformed.

87

88 **Model 2: GAM** - The general additive model uses a smoothing function on each explanatory vari-
89 able, to map each individuals fit to the response variable, and uses a Gaussian distribution as the
90 variables used are continuous. The equation takes the following form, where m = mass, f denotes the
91 smoothing function, x_n denote the explanatory variable and ϵ is error not explained by the variables.

$$m_{pred} \sim f(x_1) + f(x_2) + \dots + f(x_n) + \epsilon$$

92

93 The assumptions for GAM also require independence therefore habitat and feeding interaction were
94 also excluded from the GAM models. Log transforms were not required for the model, as GAMs do
95 not require linear variables (Zuur et al., 2009), however log transformed models were compared as
96 they were used in the previous model.

97

98 **Model 3: GLMM** - Mixed models give the ability to account for non-independence of variables by
99 adding them as random effects, meaning they effect the data but not predictably (Zuur et al., 2009).

100 The following equation gives the general model for the GLMM, where m = mass, subscripts indicate

101 fixed and random variables, d = mean depth, h = habitat, i = feeding interaction and ϵ accounts for
102 any other error.

$$m_{pred} \sim m_{prey-fixed} + d_{fixed} + h_{random} + i_{random} + \epsilon$$

103

104 As a type of linear model, prey mass and depth were both log transformed to give a linear dis-
105 tribution. Additionally collinearity and independence are also assumptions of GLMMs unless the
106 non-independent factors are accounted for as random effects, allowing the inclusion of habitat and
107 feeding interactions in this model.

108 **Computing Languages**

109 Python version 3.5.2 (Python Core Team, 2018) was used to manipulate the raw data into the data
110 frame used for model building, as the pandas package (McKinney, 2010) is fast and efficient at
111 manipulating and building data frames and csv files.

112

113 R version 3.4.4 (R Core Team, 2018) was used for model fitting, selection and plotting models. R
114 was most appropriate for this due to the wealth of packages available for model fitting and plotting.
115 The package mgcv (Wood, 2011) fits GAM models and allows plots equivalent to the residual plots
116 that can be produced for linear models. The lme4 package (Bates et al., 2015) fits linear mixed
117 effects models and gives information about the significance of each different factor. R also has an
118 inbuilt function to calculate the Akaike Information Criterion which was used to select the optimal
119 model. Finally, the ggplot2 (Wickham, 2016) package was used to generate plots used in this report
120 (specified in figure legend when used).

121

122 Shell scripts in bash were used to compile the \LaTeX document into a pdf format with references from
123 the associated BibTeX file and to run the final project, as bash has inbuilt commands to run R and
124 Python script files.

125 Results

126 Model 1: Linear Regression Model

127 Initially the linear regression model tested contained the three independent explanatory variables:
128 prey mass, depth and temperature. To fulfil the assumption of linearity, prey mass and depth were
129 both log transformed (note that predator mass is log transformed for all models), and a comparison
130 using AIC was made between the fit of the model with and without log transforms ($AIC_{notlogged} =$
131 343.8 , $AIC_{logged} = 236.1$). The smaller AIC value of the log transformed model indicated that it better
132 fitted the data and the large difference between the values shows a significantly different fit.

133
134 Each variable in the model was then tested using the step function in R, which removed the least
135 significant variable in the model then refitted it to the data until the optimal model was found. This
136 model only contained one significant explanatory variable, prey mass (linear regression: $R^2 = 0.90$,
137 $F_{1,55} = 496.3$, $p < 0.001$). The R^2 value for this model was high and positive meaning 90% of
138 the variation in predator mass could be explained by prey mass. Figure 1a indicates the fit of the
139 regression model (shown by a black line) to the data. The coefficients of the regression line were
140 extracted from the model to give an overall equation of:

$$\log(m_{pred}) = 5.41 + 0.761\log(m_{prey})$$

141

142 Where m = mass in grams, therefore when predator mass increases with prey mass.

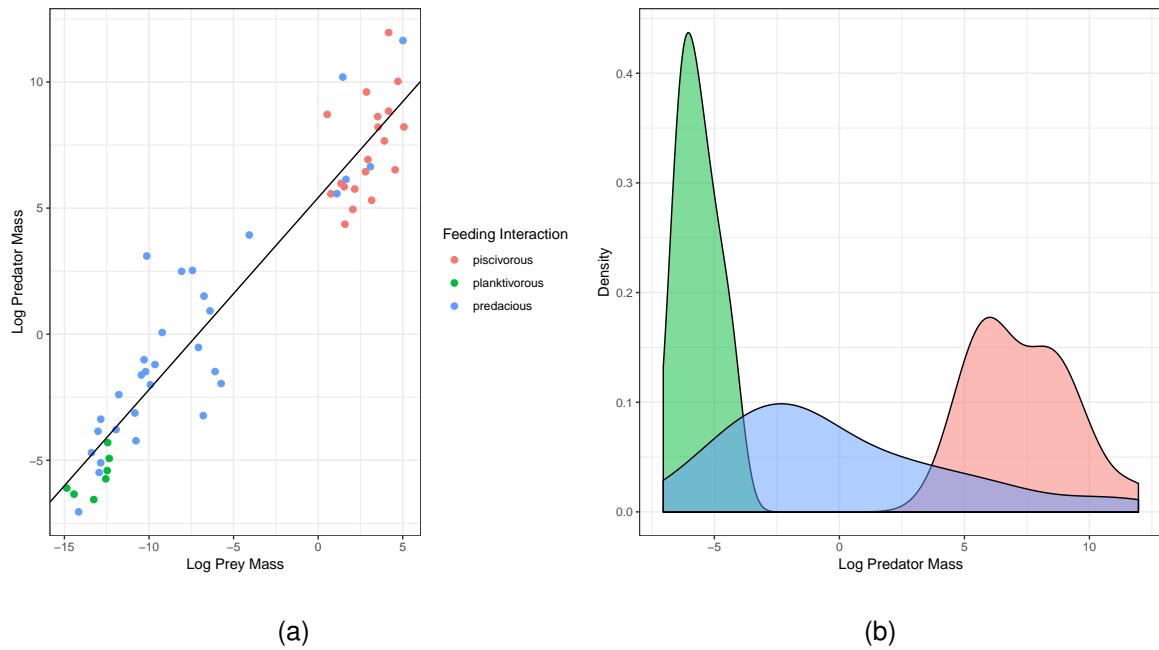


Figure 1: (a) Mean average mass of 57 predator species as a function of mean average mass of prey, both log transformed. The black line represents the linear regression model: $\log(m_{pred}) = 5.41 + 0.761\log(m_{prey})$, colours represent the three feeding interactions, piscivorous (red, $n=19$), predacious (blue, $n=31$) and planktivorous (green, $n=7$). (b) Density plot of log transformed mean average predator masses by feeding interaction, colours same as (a).

Although feeding interaction was not accounted for in the linear regression model, due to possible correlation with prey size, I split Figure 1a by feeding interaction and there appears to be strong grouping, with larger predators tending to be piscivorous and smaller to be planktivorous, with generalised predation more broadly spread. This relationship is further explored in Figure 1b, a density plot of each feeding interaction against predator mass, which reiterates this relationship. Planktivorous species also appear to be confined to a small range of predator masses whereas piscivorous cover a broader range of predator masses. This is explained further in the linear mixed model section.

Model 2: Generalised Additive Model (GAM)

The initial GAM, as with the linear regression model, included all independent explanatory variables, however there is no equivalent of the step function available for GAMs so models fits were tested

154 manually. In each step, the least significant factor was removed, and model refitted, then compared
155 to the previous model. Log transforms were also tested in the same way. The final model with the
156 best fit to the data contained all three explanatory variables, but only prey mass was log transformed:

$$\log(m_{pred}) \sim f(\log(m_{prey})) + f(temp) + f(depth) + \epsilon$$

157

158 Where m = mass in grams, ϵ indicates error not accounted for by the explanatory variables, and
159 all variables are averaged. Figure 2 shows the relationship between each of the smoothed variables
160 and predator mass. Figure2top shows predator mass increased with prey mass, as seen in the linear
161 regression model. Figure 2middle shows a more complex relationship with the smallest predators
162 found at lower temperatures, then medium to larger predators varying within a similar range of tem-
163 peratures. Figure 2bottom shows another linear relationship where larger predators are found at
164 higher depths, whereas small predators tend to stay in shallower areas.

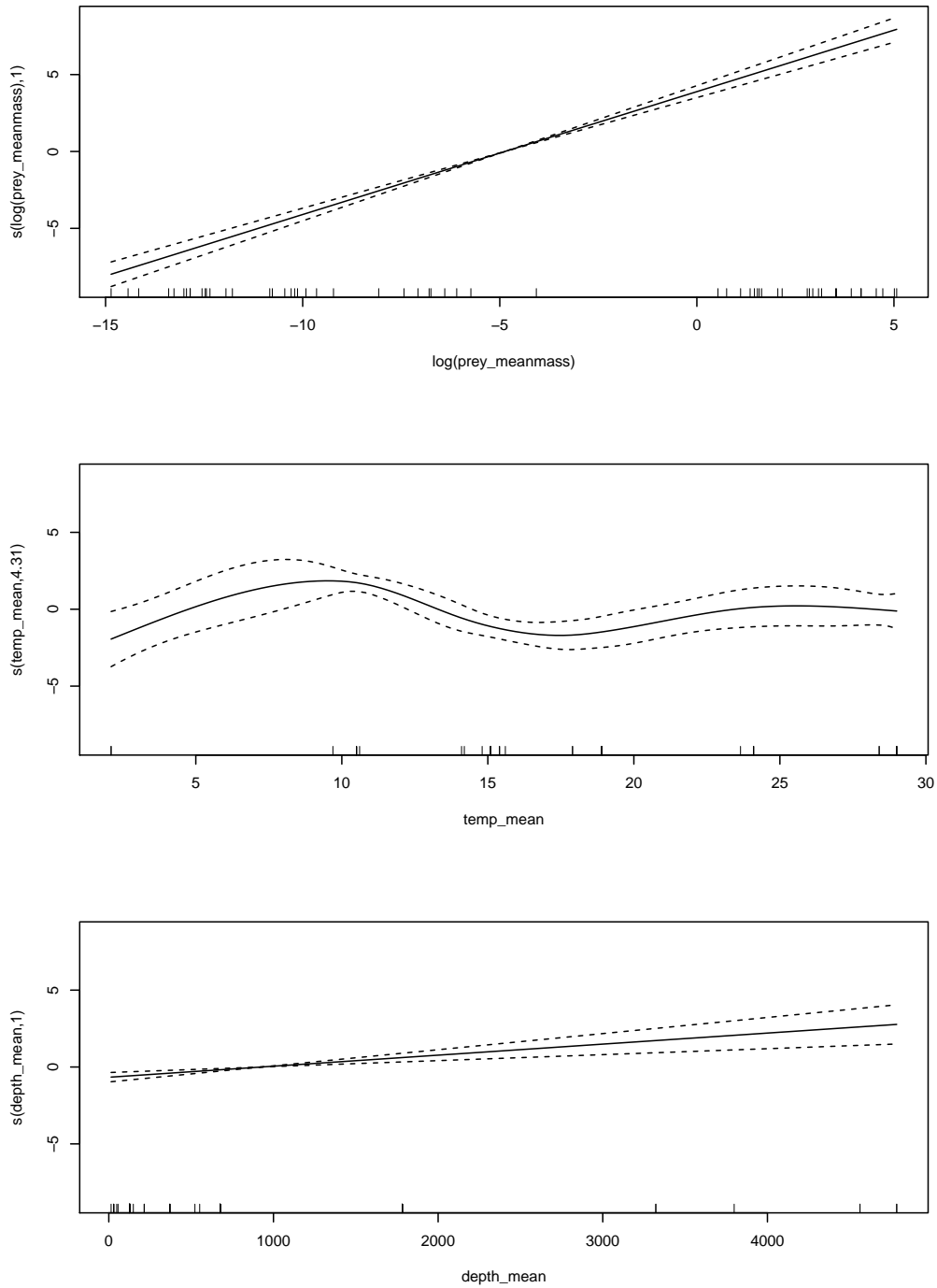


Figure 2: Smoothed functions of the explanatory variables used in the GAM model, where dotted lines give standard error and average predator mass is log transformed, $n=57$. Top: predator mass increases linearly with log transformed prey mass, middle: temperature fluctuates with predator mass change, indicating no clear relationship, bottom: predator mass increase linearly with depth.

165 **Model 3: Linear Mixed Effects Model (GLMM)**

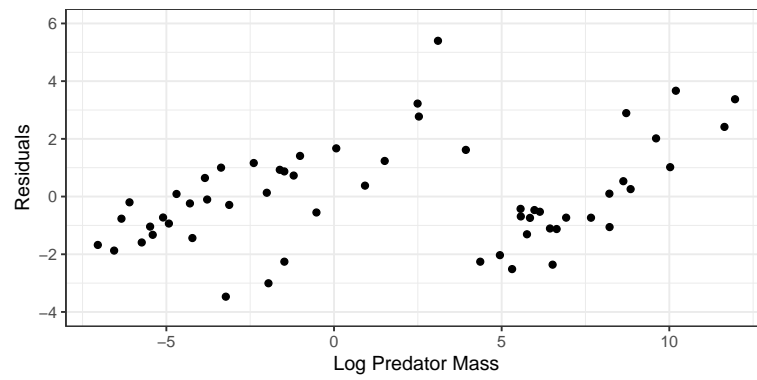
166 As with the first two models, I started the GLMM model with the maximum explanatory variables, in-
167 cluding habitat and feeding interaction as random effects which allows for their lack of independence.
168 I then compared the models manually, excluding the least significant factor and comparing the fit of
169 the new models using analysis of variance (ANOVA) as GLMMs are a type of linear model. The AIC
170 values calculated gave the best fitting model as:

$$\log(m_{pred}) \sim \log(m_{prey}) + feeding_{random} + \epsilon$$

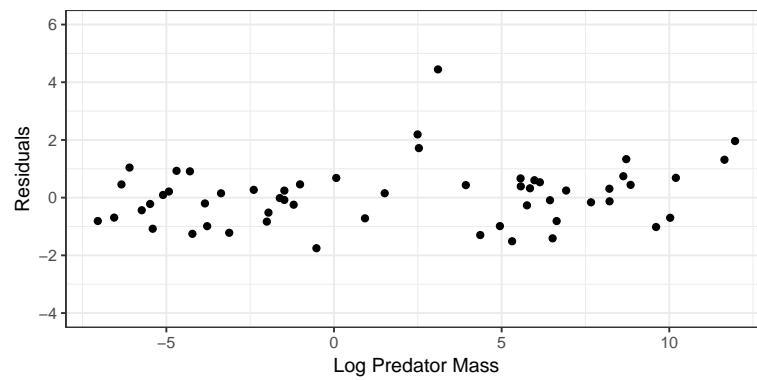
171 Where m = mass in grams, feeding interaction is included as a random effect and epsilon indicates
172 variance not explained by the other variables. This model produced a very similar relationship to the
173 linear regression model, however also explains the grouped feeding interactions observed in Figure
174 1.

175 **Model Selection**

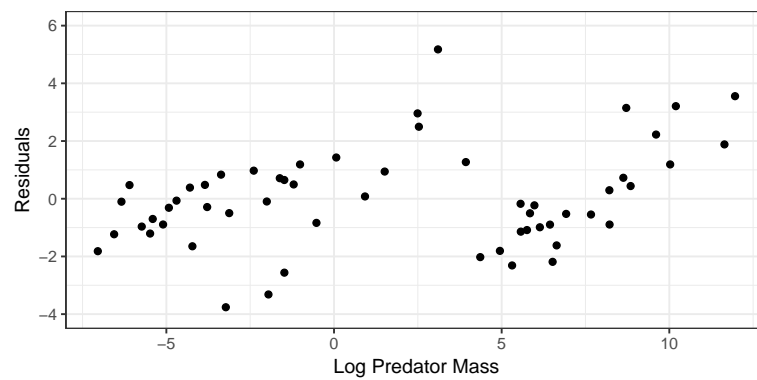
176 Once optimal models were chosen for each type of model, the AIC values were compared, with the
177 linear regression model giving an AIC = 232.5, GAM AIC = 200.4 and GLMM AIC = 239.0, therefore
178 the GAM has the best fit. Additionally, Figure 3 shows the residual plots for each of the final models,
179 and GAM has the narrowest range of residuals and most even spread compared to the other models,
180 further supporting that it is the most appropriate model for the data.



(a)



(b)



(c)

Figure 3

Discussion

This study aimed to find an optimum model with which to predict predator mass, based upon a range of potential factors. Three types of model were tested: linear regression, generalised additive (GAM) and generalised linear mixed effects (GLMM), and the optimal model was found using GAM method. This model indicated that predator mass was a function of average prey mass, annual air

186 temperature at the location and average depth at which the prey species was found. Predator mass
187 (averaged and log transformed) had a positive linear relationship with prey mass, therefore as prey
188 mass increased, so did predator mass. Additionally, predator mass also increased with depth and
189 varied with temperature fluctuations.

190
191 Many marine predators, especially those of larger mass, tend to have large home ranges and en-
192 counter a wide variety of prey (Costa, 2016). This makes them particularly difficult to monitor,
193 whereas smaller species are often more localised, therefore easily monitored. The model generated
194 could be useful in locations where potential prey species are present, as expected size of predator
195 species could be predicted. This could also highlight those areas where predators are absent de-
196 spite prey availability. Furthermore, species at higher trophic levels tend to be more vulnerable to
197 disturbance as their persistence depends on an entire chain of species, as opposed to one producer
198 (Purvis et al., 2000), so having a means to predict their presence, or lack of, could aid monitoring
199 groups globally.

200
201 In the optimal GLMM, the factor of feeding interaction was present, and relatively distinct clusters for
202 piscivorous and planktivorous species were shown in 1. However, this variable was not included in
203 the final model as it was unlikely to be independent of prey mass, and GAMs do not allow for random
204 effects. Despite this, it is likely that feeding interaction does effect predator mass. This could be
205 accounted for by generalised additive mixed effect models (Zuur et al., 2009), which combine the
206 inclusion of random effects with a GAM, and would be the next logical step from this study.

207
208 Another important assumption of this study, was that density of predators or prey had no affect on
209 predator mass or was constant. This may oversimplify the system as, for example, many whale
210 species feed on plankton however the density of plankton allow a much higher body mass and many
211 piscivorous fish species have been observed increasing predation when only smaller prey species
212 are available (Vezina, 1985). This contradicts the findings in this study that indicated planktivorous

213 predators had some of the lowest body masses. Therefore, another area for further study would be
214 to include a factor for density into the model.

215 **Conclusion**

216 Overall, this study showed that the assumptions of models can have a marked effect on which ex-
217 planatory variables are selected during optimisation. For marine predator prey systems, GAM were
218 found to have the best fit to the data, and the function of mixed effects models to include random
219 effects also showed interactions that may have otherwise been missed. To further this study, models
220 combining these two types of model are advised, especially if used in predictions for conservation.

221 **References**

222 Jeremy T Kerr and David J Currie. Effects of Human Activity on Global Extinction Risk. *Conservation*
223 *Biology*, 9(5):1528–1538, 1995.

224 Anthony D. Barnosky, Nicholas Matzke, Susumu Tomiya, Guinevere O. U. Wogan, Brian Swartz,
225 Tiago B. Quental, Charles Marshall, Jenny L. McGuire, Emily L. Lindsey, Kaitlin C. Maguire, Ben
226 Mersey, and Elizabeth A. Ferrer. Has the Earth's sixth mass extinction already arrived? *Nature*,
227 471:51–57, 2011. ISSN 0028-0836. doi: 10.1038/nature09678.

228 Michael R. Heithaus, Alejandro Frid, Aaron J. Wirsing, and Boris Worm. Predicting eco-
229 logical consequences of marine top predator declines. *Trends in Ecology & Evolution*,
230 23(4):202–210, apr 2008. ISSN 01695347. doi: 10.1016/j.tree.2008.01.003. URL
231 <https://linkinghub.elsevier.com/retrieve/pii/S0169534708000578>.

232 Richard Levins. The Strategy of Model Building in Population Biology. *American Sci*, 54(4):421–431,
233 1966. ISSN 00030996. doi: 10.2307/27836590. URL <http://www.jstor.org/stable/27836590>.

234 H I Freedman and Paul Waltman. Mathbio_1984.Pdf, 1983.

235 Hakan Turesson, A. Persson, and C. Bronmark. Prey size selection in piscivorous pikeperch (Sti-

236 zosterion lucioperca) includes active prey choice. *Ecology of Freshwater Fish*, 11(4):223–233,
 237 2002. ISSN 09066691. doi: 10.1034/j.1600-0633.2002.00019.x.

238 C Barnes, DM Bethea, RD Brodeur, J Spitz, V Ridoux, C Pusineri, BC Chase, ME Hunsicker,
 239 F Juanes, A Kellermann, J Lancaster, F Menard, FX Bard, P Munk, JK Pinnegar, FS Scharf,
 240 RA Rountree, KI Stergiou, C Sassa, A Sabates, and S Jennings. Predator and prey body sizes
 241 in marine food webs: ecological archives E089-051. *Ecology*, 89(3):881, 2008. ISSN 1095-9203.
 242 doi: 10.1890/1.

243 Alain F. Zuur, Elena N. Ieno, Anatoly A. Saveliev, Graham M. Smith, and Neil Walker. *Mixed Ef-*
 244 *fects Models and Extensions in Ecology with R*. Springer, New York, 1 edition, 2009. ISBN
 245 9780387874579.

246 Python Core Team. Python: A dynamic, open source programming language., 2018. URL
 247 <https://www.python.org/>.

248 Wes McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod
 249 Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 51–56, 2010.

250 R Core Team. R: A language and environment for statistical computing., 2018. URL
 251 <https://www.r-project.org/>.

252 S.N. Wood. Fast stable restricted maximum likelihood and marginal likelihood estimation of semi-
 253 parametric generalized linear models. *Journal of the Royal Statistical Society (B)*, 73(1):3–36,
 254 2011.

255 Douglas Bates, M Maechler, Benjamin M. Bolker, and Steve Walker. Fitting linear mixed-effects
 256 models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015. doi: 10.18637/jss.v067.i01.

257 H. Wickham. *ggplot2: Elegant graphics for data analysis*. Springer-Verlag, New York, 2016. URL
 258 <http://ggplot2.org>.

259 Gabriel C Costa. Predator Size , Prey Size , and Dietary Niche Breadth Relationships in Marine
 260 Predators. *Ecology*, 90(7):2014–2019, 2016. doi: doi:10.1890/08-1150.1.

261 A. Purvis, J. L. Gittleman, G. Cowlishaw, and G. M. Mace. Predicting extinction risk in declining
262 species. *Proceedings of the Royal Society B: Biological Sciences*, 267(1456):1947–1952, 2000.
263 ISSN 14712970. doi: 10.1098/rspb.2000.1234.

264 AF Vezina. Oecologia Empirical relationships between predator and prey size among terrestrial
265 vertebrate predators. *Oecologia*, pages 555–565, 1985.