**Beyond age length keys: ordinal models for predicting age given length**

New Zealand Fisheries Assessment Report 2022/xx

D.N. Webber

A. Dunn

S. Mormede

ISSN xxxx-xxxx (online)

ISBN yyyy-yyyy (online)

month 2022

Requests for further copies should be directed to:

Publications Logistics Officer

Ministry for Primary Industries

PO Box 2526

WELLINGTON 6140

Email: [brand@mpi.govt.nz](mailto:brand@mpi.govt.nz)

Telephone: 0800 00 83 33

Facsimile: 04-894 0300

This publication is also available on the Ministry for Primary Industries websites at:

<http://www.mpi.govt.nz/news-and-resources/publications>

[http://fs.fish.govt.nz](http://fs.fish.govt.nz/) go to Document library/Research reports

**© Crown Copyright – Fisheries New Zealand**

**TABLE OF CONTENTS**

[Executive summary 1](#_Toc98493324)

[1. Introduction 2](#_Toc98493325)

[2. Methods 2](#_Toc98493326)

[2.1 Age length keys 2](#_Toc98493327)

[2.2 Ordinal models 3](#_Toc98493328)

[3. Results 4](#_Toc98493329)

[4. Discussion 1](#_Toc98493330)

[4.1 Recommendations 1](#_Toc98493331)

[4.2 Future research 1](#_Toc98493332)

[5. Acknowledgements 1](#_Toc98493333)

[6. References 1](#_Toc98493334)

[7. Tables 2](#_Toc98493335)

[8. Figures 3](#_Toc98493336)

[9. Appendix I: input for catch at age software 4](#_Toc98493337)

[10. Appendix II: diagnostics plots 5](#_Toc98493338)

# Executive summary

**Webber, D.N.[[1]](#footnote-1); Dunn, A.; Mormede, S. (2022). Beyond age length keys: ordinal models for predicting age given length.**

***New Zealand Fisheries Assessment Report 2022/xx. Yy* p*.***

This document describes.

# Introduction

Age structured stock assessment models require commercial catch or survey age composition data.

Length data is collected for some species. Generally, otoliths are collected and aged for only a subset of these data providing paired age-length data. So there is a need to convert the length composition data for which there are no associated ages to age composition data. This is usually done empirically using an age length key (ALK). Often a small set of otoliths are sampled for ageing each year. This is done by observers onboard fishing vessels or by collecting market samples. These aged fish are used to derive age-length relationships and age-length keys.

Although otoliths are collected at random within tows, the tows are not random with respect to location or time. Hence ages from selected otoliths may contain spatial or temporal bias (i.e., ignore spatial and temporal variability). Sampling may not collect otoliths from critical areas in the fishery in some years. ALKs are non-continuous w.r.t. length (i.e., length data must be discretised) although usually this does not matter as fish are measured to the nearest cm. Often much data is discarded wasting time and money (e.g., entire years worth of aged fish are discarded if it is deemed that there is not enough data). Analysis methods are statistically inefficient, hence estimates of uncertainty may be poor.

A better approach would be model based so that it is able to deal with spatial and temporal variability; model length as a continuous explanatory variable; and enable inference when there is little or no data. Could age fewer fish (saving $$$) for the same level of uncertainty. Ordinal regression.

In this document we use ordinal regression to predict age compositions given length compositions. Ordinal regression has been used for this purpose in the past (Babyn et al. 2021, Berg & Kristensen 2012, Rindorf & Lewy 2001). But this document does so using Bayesian inference and a New Zealand context. It may also use a better model than that used by others (cumulative rather than cratio), not sure yet, see more detail below.

This document covers the following analyses:

1. development of categorial models that predict length composition and associated uncertainty from samples of length and covariates;
2. scaling of the length composition derived in step 1 by the catch;
3. development of ordinals model that can predict age given length from samples of paired age-length and covariates; and
4. conversion of the scaled length composition derived in step 2 to an age composition using the model described in step 3

## Age length keys

An age-length key (ALK) is an joint distribution matrix where each row corresponds to an age and each column to a length-class . An age-length key specifies the relative proportion of fish of length that are age and can be written

where is the conditional probability that a randomly selected fish is of age given it is in length-class . This conditional probability can be determined using an appropriate growth model and is equivalent to which is the normal probability density that a fish of length is age , where is the normal Z-score for a fish of age and length , calculated as

where is the mean length (cm) of a fish of age , and is the standard deviation of the age-length relationship for . The normal Z-score is then converted to a cumulative normal distribution for each age to give the probability that a fish of length will exceed age , and finally this is converted to the probability that a fish of length is age or . The ALK is applied to length-frequency distributions to derive proportions in the catch at age observations

is then compared to the proportions at age of the catch in stock assessment models.

# Methods

## Sampling length from the catch

For a given time (), in a location (), and tow () counts of the number of fish measured by length bin () are sampled from all fish caught in that tow, this can be represented as

where is the number of sampled fish, is an error distribution describing each sample from a tow, and is the true number of fish in that tow. The true number of fish in a tow is drawn from the available population

where is an error distribution describing each tow from the available population and is the true number of fish by length bin () within the available population at any given time () and location ().

By rolling these two error structures together, these counts can be fitted to using a multinomial distribution

|  |  |
| --- | --- |
| Equation 1: |  |

where the expected value of each length bin () can be derived using linear regression

where is the intercept for each length, are coefficients accounting for time, are coefficients accounting for space, and represents any other explanatory variables (e.g., vessel characteristics, different sampling programmes).

## Scaling by the catch

For a given time

## Ordinal models

We denote values of the response variable as , a density function as , and use to refer to the main model parameter, which is usually the mean of the response distribution or some closely related quantity. In a regression framework, is not estimated directly but computed as , where is a predictor term and is the response function (i.e., inverse of the link function).

For ordinal models, is one of the categories . The intercepts of ordinal models are called thresholds and are denoted as , with , whereas does not contain a fixed effects intercept. Note that the applied link functions are technically distribution functions . The density of the **cumulative** family (implementing the most basic ordinal model) is given by

The densities of the stopping ratio (**sratio**) and continuation ratio (**cratio**) families are given by

and

respectively. Both families are equivalent for symmetric link functions such as logit or probit. The cumulative and sratio models use , whereas cratio uses . This is done to ensure that larger values of increase the probability of higher response categories. The linear predictor can be generalised to also depend on the category for a subset of predictors. This leads to category specific effects.

It is not clear what others have used (Babyn et al. 2021, Berg & Kristensen 2012, Rindorf & Lewy 2001), but one may not want to use cratio because each age is not linked due to different year classes / recruitment pulses. Cumulative may be the best choice.

The following explanatory variables were considered:

* length – the length (cm) of each individual fish for which an otolith was collected. This is the continuous length (e.g., not necessarily rounded to the nearest integer)
* fyear – the fishing year that the otolith was collected
* cyear – a continuous version of fyear
* month – the month that the otolith was collected
* origin – the origin of the otolith (SOP, etc)
* latitude – the latitude at which the otolith was collected
* longitude – the longitude at which the otolith was collected

The model that would be most similar to the ALK would be

s(length, by = fyear)

which includes…

The R package brms was used for developing models (Bürkner 2017). The brms package uses Stan (Stan Development Team 2017). Ordinal regression using brms is discussed in Bürkner & Vuorre (no date). Model selection was done using the LOO IC (Vehtari et al. 2017).

* See this page for how to set up in brms <https://kevinstadler.github.io/notes/bayesian-ordinal-regression-with-random-effects-using-brms/>
* This page has the families <https://cran.r-project.org/web/packages/brms/vignettes/brms_families.html>

## Combing the models

We denote values of the response.

# Results

Notes

* The data were ages (ordered integers) and the model treats each age as an ordered categorical variable.
* The most important explanatory variable was length. Other potential explanatory variables included the fishing year that the sampling occurred, month, origin, latitude, and longitude.
* The spline s(length, k = 3) tended to be better than s(length). I think that this is because brms uses k = 10 by default. The form of these two splines looked very similar. Could try fitting a gam to see what k it suggests, but gam may not work for these models.
* The linear term length did not work.
* According to LOO IC, the cumulative(“logit”) distribution was better than cratio or sratio.
* cumulative(“probit”) didn’t work
* estimating different variances ran with some initial complaints about starting values
* the model t2(length, cyear) where cyear is a continuous version of year was good
* adding the terms month and t2(lat, long) worked also

Table 1: LOO IC. Run time is HH:MM:SS.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | ∆ELPD | |  | ELPD LOO | |  | p\_loo | |  | LOO IC | |  |  |
| Model | Family | Mean | SE |  | Mean | SE |  | Mean | SE |  | Mean | SE | Divergent | Run time |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| t2(length, cyear, k = c(10, 26)) + s(lat, long) | cumulative | 0 | 0 |  | -27244.4 | 140.4 |  | 157.7 | 4.5 |  | 54488.8 | 280.9 | 0 | 13:48:20 |
| t2(length, cyear, k = c(3, 26)) + origin + month + s(lat, long) | cumulative | -471.9 | 36.3 |  | -27716.4 | 138.9 |  | 119.7 | 2.7 |  | 55432.7 | 277.9 | 1 | 06:32:15 |
| t2(length, cyear, k = c(3, 26)) + origin + month + t2(lat, long) | cumulative | -472.1 | 36.7 |  | -27716.5 | 139.0 |  | 111.7 | 2.7 |  | 55433.1 | 278.0 | 0 | 06:58:18 |
| t2(length, cyear, k = c(3, 26)) + month + t2(lat, long) | cumulative | -474.6 | 36.5 |  | -27719.0 | 138.9 |  | 108.8 | 2.6 |  | 55438.0 | 277.9 | 2 | 07:05:17 |
| t2(length, cyear, k = c(3, 26)) + s(lat, long) | cumulative | -493.9 | 35.7 |  | -27738.3 | 139.2 |  | 116.4 | 2.7 |  | 55476.6 | 278.4 | 0 | 06:51:13 |
| t2(length, cyear, k = c(3, 26)) + s(lat, long, bs = "gp") | cumulative | -494.7 | 35.8 |  | -27739.1 | 139.3 |  | 112.1 | 2.7 |  | 55478.2 | 278.6 | 0 | 07:45:10 |
| t2(length, cyear, k = c(3, 26)) + t2(lat, long) | cumulative | -498.4 | 36.0 |  | -27742.8 | 139.3 |  | 107.7 | 2.6 |  | 55485.6 | 278.6 | 0 | 06:37:49 |
| t2(length, cyear, k = c(3, 26)) | cumulative | -551.7 | 37.2 |  | -27796.1 | 140.0 |  | 99.2 | 2.5 |  | 55592.1 | 280.0 | 0 | 05:39:48 |
| s(length, k = 3, by = fyear) + month + s(lat, long) | cumulative | -1518.1 | 57.5 |  | -28762.5 | 134.9 |  | 112.3 | 2.9 |  | 57525.0 | 269.9 | 0 | 06:45:42 |
| s(length, cyear, k = c(3, 26)) + origin + month + s(lat, long) | cumulative | -1522.4 | 58.6 |  | -28766.8 | 131.4 |  | 54.3 | 1.6 |  | 57533.6 | 262.8 | 1 | 02:49:16 |
| s(length, cyear, k = c(3, 26)) + origin + month + t2(lat, long) | cumulative | -1538.2 | 59.0 |  | -28782.6 | 131.3 |  | 41.7 | 1.4 |  | 57565.2 | 262.7 | 2 | 03:08:56 |
| s(length, k = 3, by = fyear) + month + t2(lat, long) | cumulative | -1544.1 | 58.2 |  | -28788.5 | 135.1 |  | 110.4 | 3.1 |  | 57577.0 | 270.1 | 126 | 13:16:37 |
| s(length, cyear, k = c(3, 26)) | cumulative | -1616.3 | 60.2 |  | -28860.7 | 131.4 |  | 29.6 | 1.4 |  | 57721.3 | 262.9 | 0 | 02:00:04 |
| s(length, k = 3, by = fyear) | cumulative | -1619.3 | 59.6 |  | -28863.7 | 135.4 |  | 88.5 | 2.7 |  | 57727.4 | 270.8 | 0 | 05:13:06 |
| s(length) | cumulative | -1873.9 | 64.4 |  | -29118.3 | 130.6 |  | 34.1 | 1.4 |  | 58236.6 | 261.1 | 0 | 02:57:54 |
| s(length, bs = "cr") | cumulative | -1875.2 | 64.4 |  | -29119.6 | 130.6 |  | 34.4 | 1.4 |  | 58239.2 | 261.2 | 0 | 02:01:46 |
| t2(length) | cumulative | -1880.2 | 64.3 |  | -29124.7 | 130.6 |  | 30.6 | 1.4 |  | 58249.3 | 261.2 | 0 | 03:27:50 |
| t2(length, bs = "ts") | cumulative | -1882.7 | 64.6 |  | -29127.2 | 130.8 |  | 30.8 | 1.4 |  | 58254.3 | 261.7 | 0 | 03:08:37 |
| s(length, k = 3), disc ~ length | cumulative | -1990.5 | 65.8 |  | -29234.9 | 130.2 |  | 29.1 | 1.4 |  | 58469.8 | 260.5 | 0 | 02:35:36 |
| s(length, k = 3) | cumulative | -2036.0 | 65.8 |  | -29280.4 | 130.9 |  | 28.4 | 1.4 |  | 58560.9 | 261.7 | 0 | 02:46:49 |
| s(length, k = 3) | sratio | -3106.2 | 82.7 |  | -30350.6 | 138.9 |  | 31.3 | 1.6 |  | 60701.2 | 277.9 | 1 | 12:36:58 |
| s(length, k = 3) | cratio | -3106.6 | 82.7 |  | -30351.0 | 138.9 |  | 31.7 | 1.6 |  | 60702.0 | 277.9 | 0 | 10:35:04 |
| length | cumulative | -3434.0 | 74.5 |  | -30678.4 | 132.8 |  | 27.7 | 1.3 |  | 61356.8 | 265.5 | 0 | 00:21:23 |

# Discussion

* This analysis will be repeated just before the CRA 2 stock assessment (or rapid update) and reported in the data FAR for that assessment.

## Recommendations

Recommendations include:

* It appears that when the data is moved from the fine scale.

## Future research

Future research considerations should include:

* Different treatments of time (e.g., Julian date rather than fishing year and month)
* Further exploration of spatial effects

# Acknowledgements

We thank Fisheries New Zealand who awarded the contract SEA2022-28 to Ocean Environmental Ltd. We thank members of the Rock Lobster Working Group for their peer review and helpful discussion throughout.

# References

Babyn, J.; Varkey, D.; Regular, P.; Ings, D.; Mills Flemming, J. (2021). A gaussian field approach to generating spatial age length keys. *Fisheries Research* 240: 105956.

Berg, C.W.; Kristensen, K. (2012). Spatial age-length key modelling using continuation ratio logits. *Fisheries Research* 129–130: 119–126.

Bürkner, P. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software* 80(1): 1–28.

Bürkner, P.; Vuorre, M. (XXXX). Ordinal regression models in psychology: a tutorial. *Unpublished*.

Rindorf, A.; Lewy, P. (2001). Analyses of length and age distributions using continuation-ratio logits. *Canadian Journal of Fisheries and Aquatic Science* 58: 1141–1152.

Stan Development Team (2017). *Stan Modelling Language: User’s Guide and Reference Manual*. Version 2.16.0.

Vehtari, A.; Gelman, A.; Gabry, J (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing* 27: 1413–1432.

# Tables

Table 1: Minimum legal size (MLS) limits for males and females over time in CRA 2. Note that the MLS before 1987 was expressed in terms of tail-length and has been converted to tail-width using the procedure described by Breen et al. (1988).

|  |  |  |
| --- | --- | --- |
|  | MLS (tail-width, mm) | |
| Period | Males | Females |
| 1945–1949 | None | None |
| 1950–1951 | 47 | 49 |
| 1952–1958 | 51 | 53 |
| 1959–1987 | 53 | 58 |
| 1988–1991 | 54 | 58 |
| 1992–2020 | 54 | 60 |

# Figures

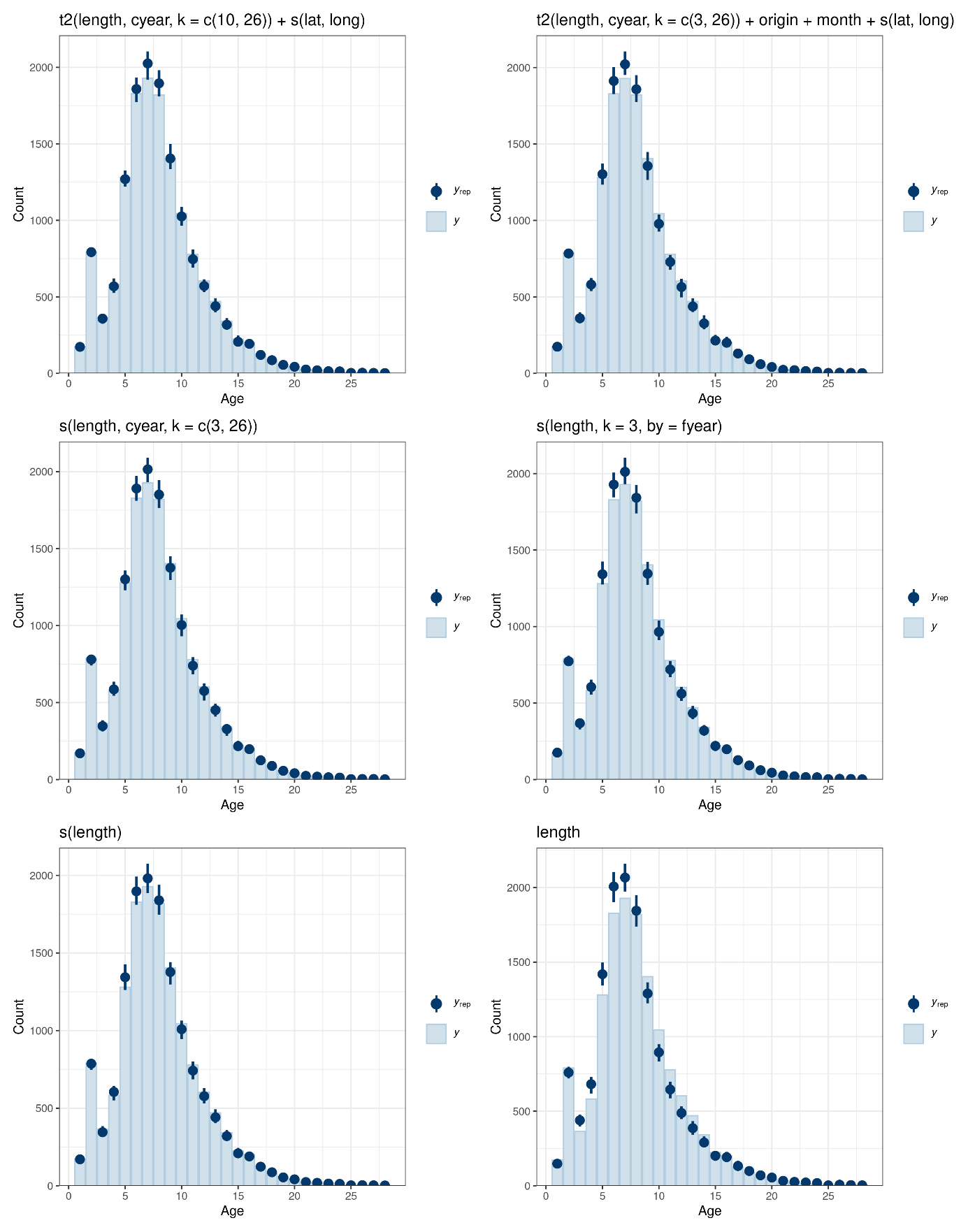


Figure 1: posterior predictive density plots for six different models, including the best model according to LOO IC (top-left). In each plot, *y* is the density of the data (light blue bar) and *yrep* is the density of 100 draws of data from the posterior predictive distribution (dark blue point and whiskers).

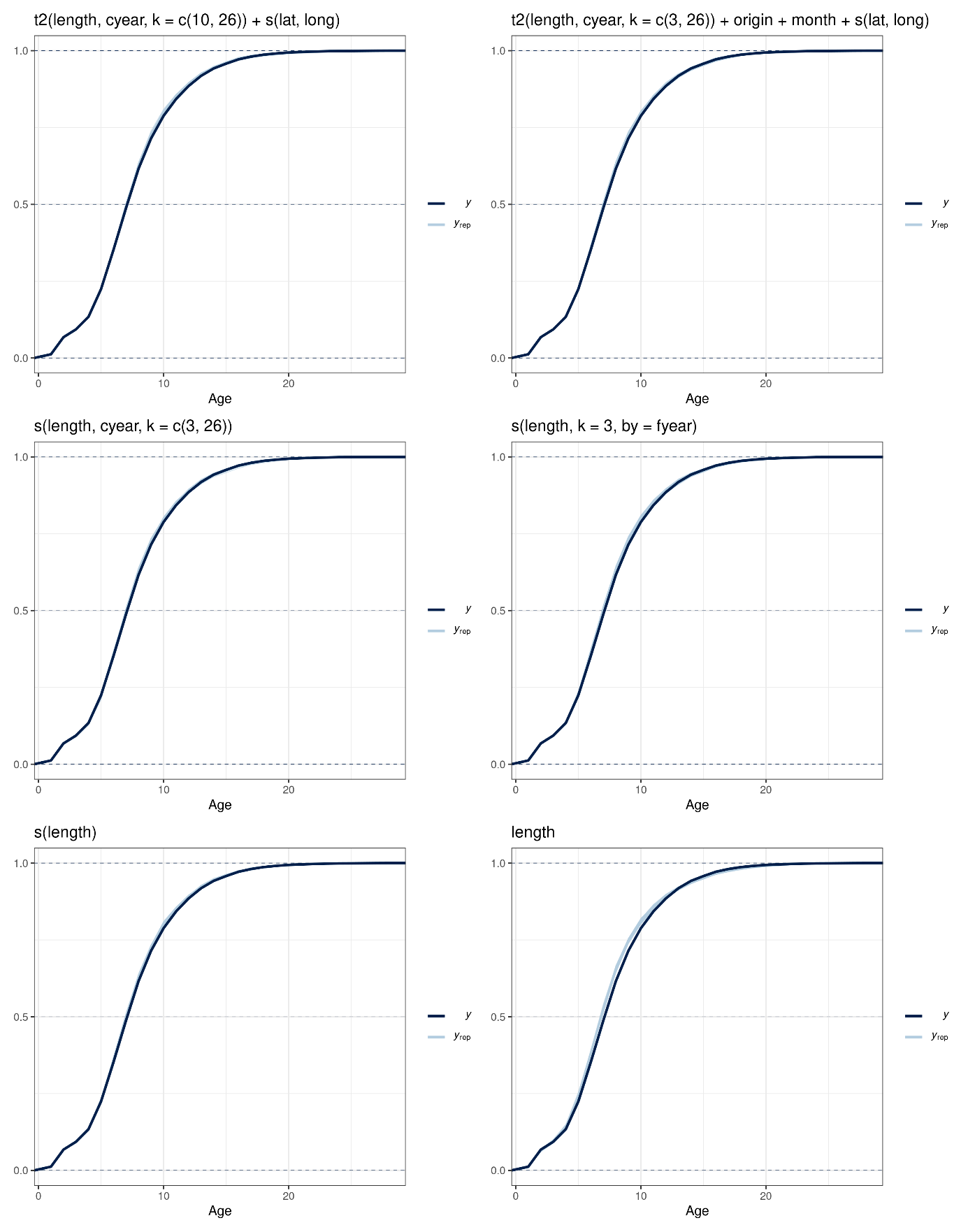


Figure 1: Posterior predictive check plot.

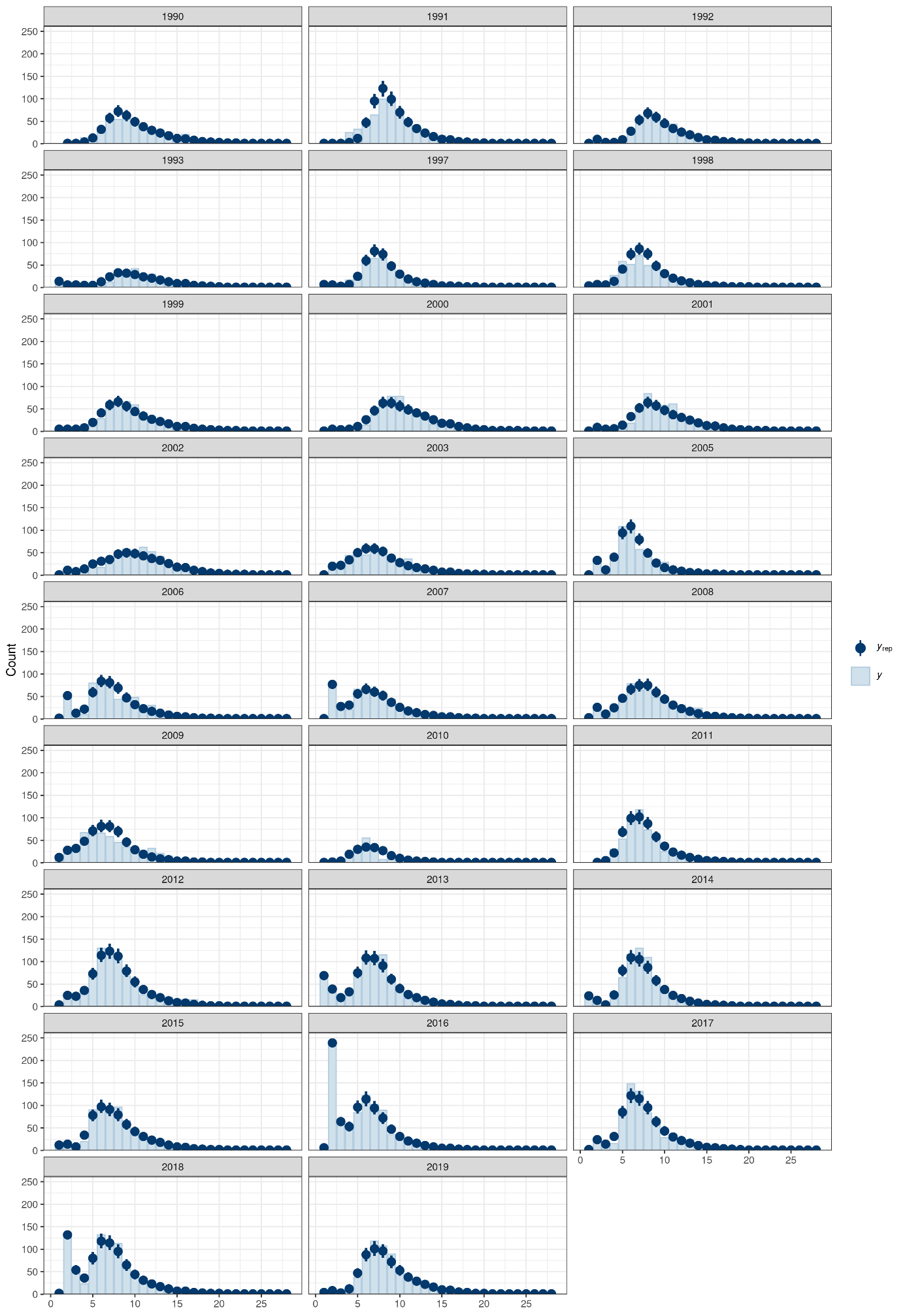


Figure 1: Posterior predictive check plot for the best model.

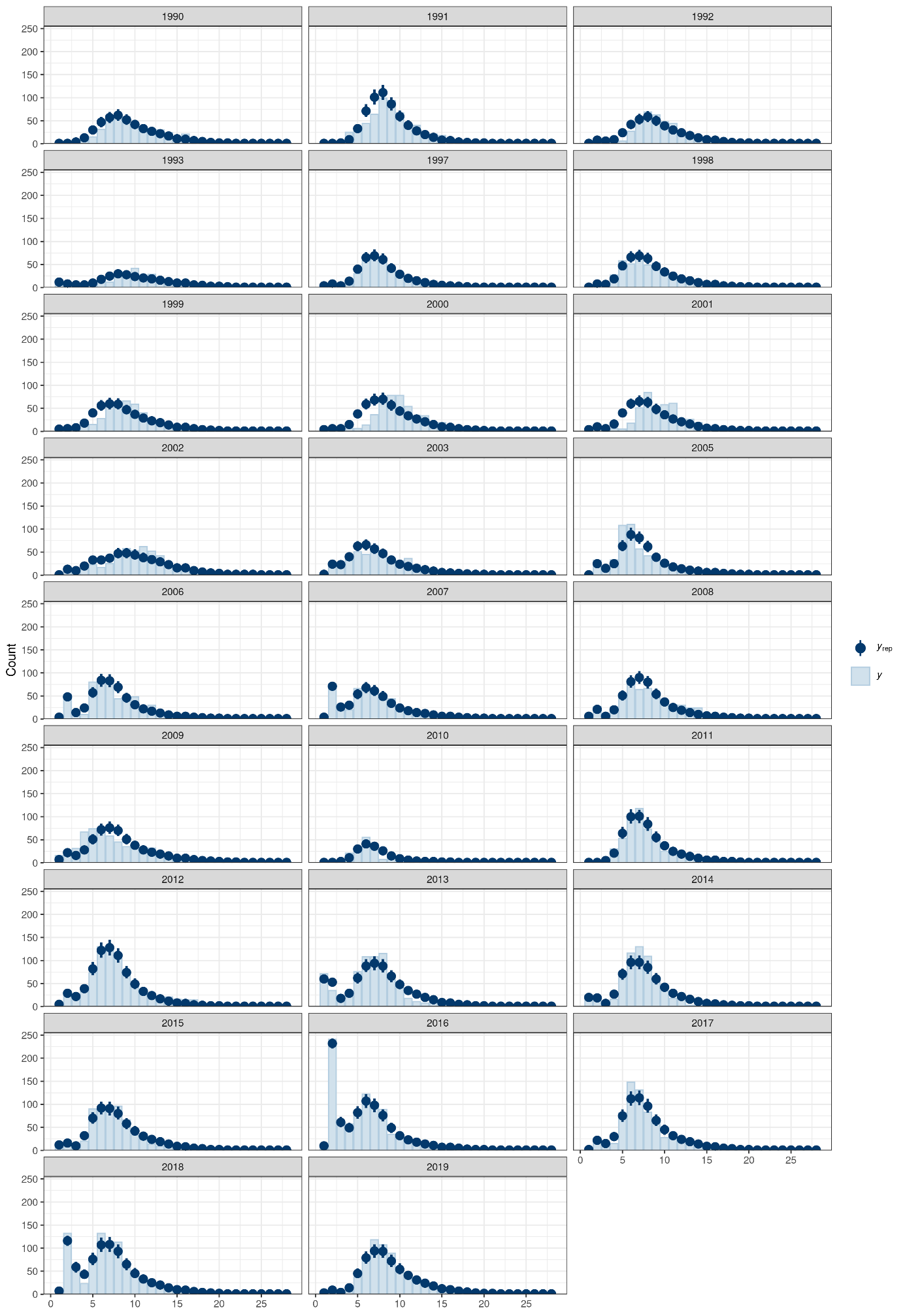


Figure 1: Posterior predictive check plot for the best model with independent years.

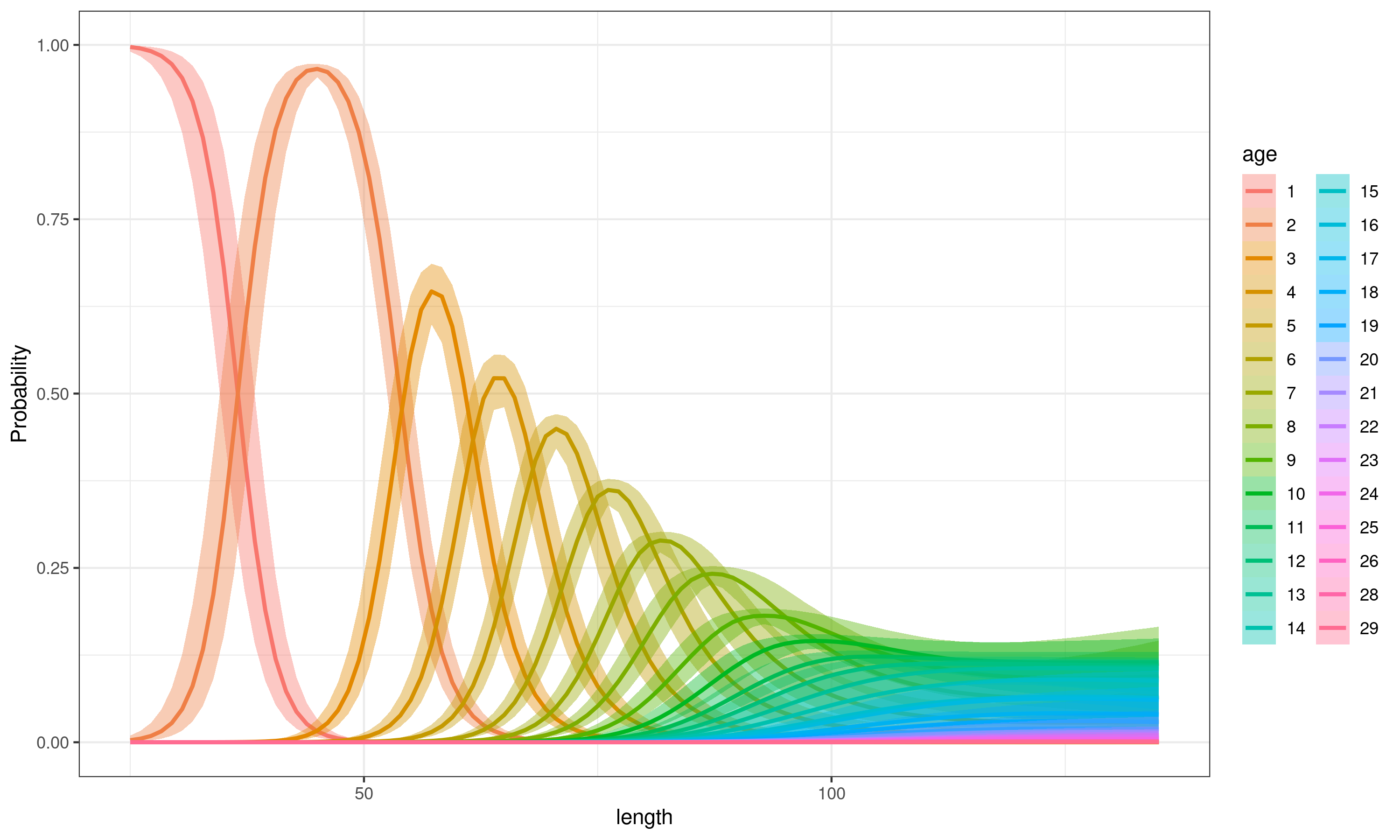


Figure 1: Conditional effect of length in the best model.

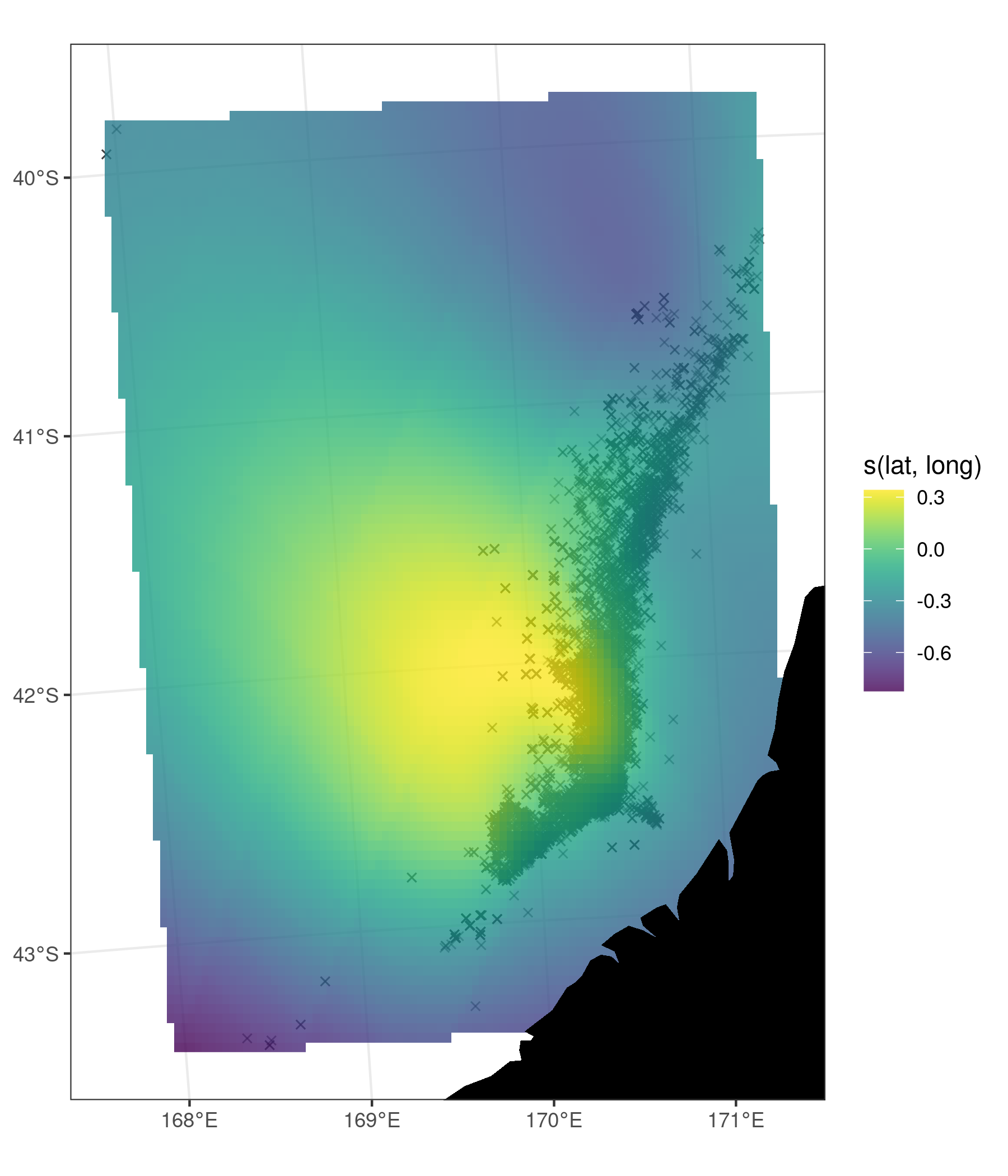


Figure 1: Conditional smooth from the best model and the data points (black crosses).

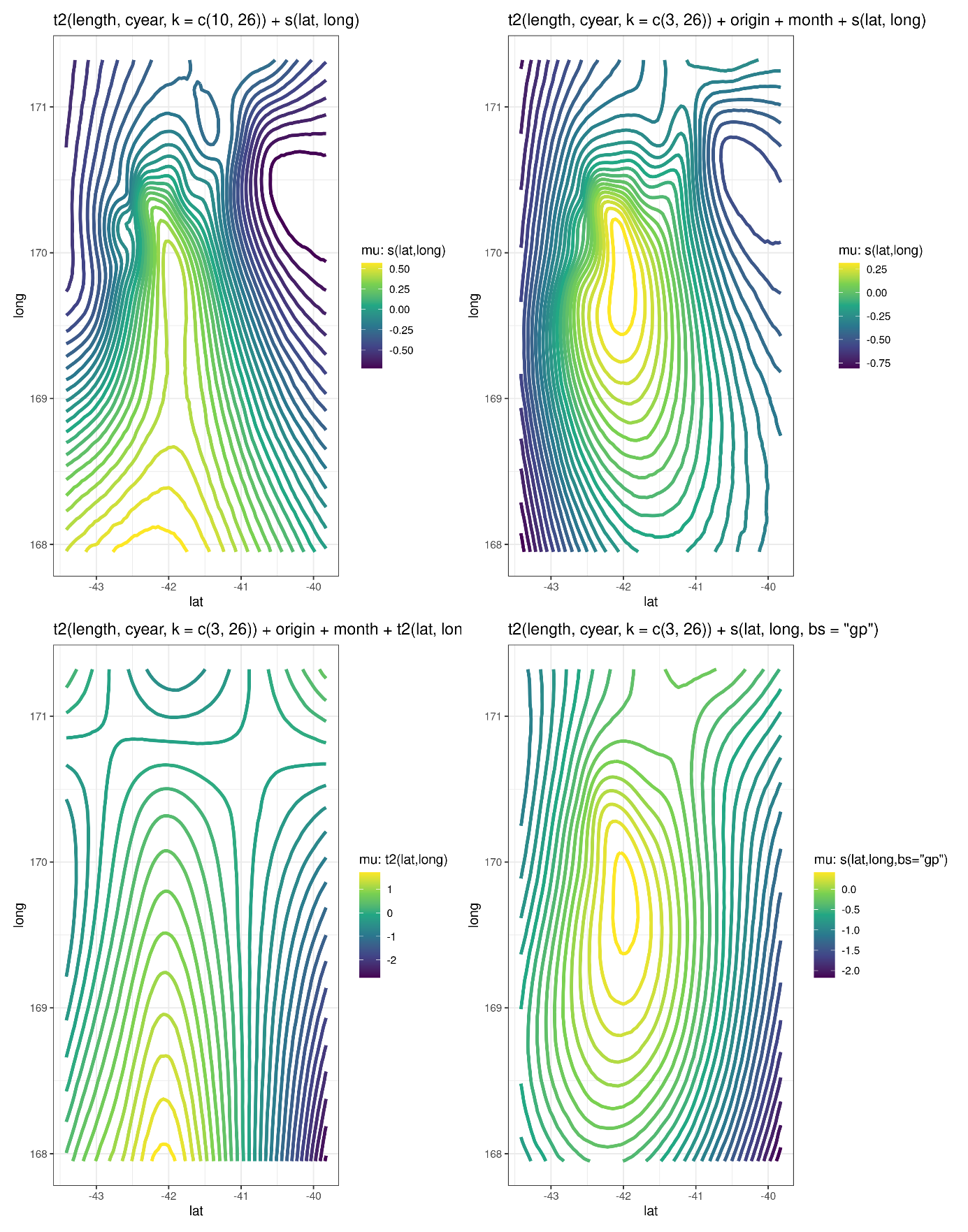


Figure 1: Conditional smooth from the best model and the data points (black crosses).

# Appendix I: input for catch at age software

SOP1990\_wcsi (t.age & t.read data) 1990

163

66 1 6

66 1 6

66 1 8

69 1 5

70 1 4

70 1 6

70 1 7

70 1 12

70 1 13

# Appendix II: diagnostics plots

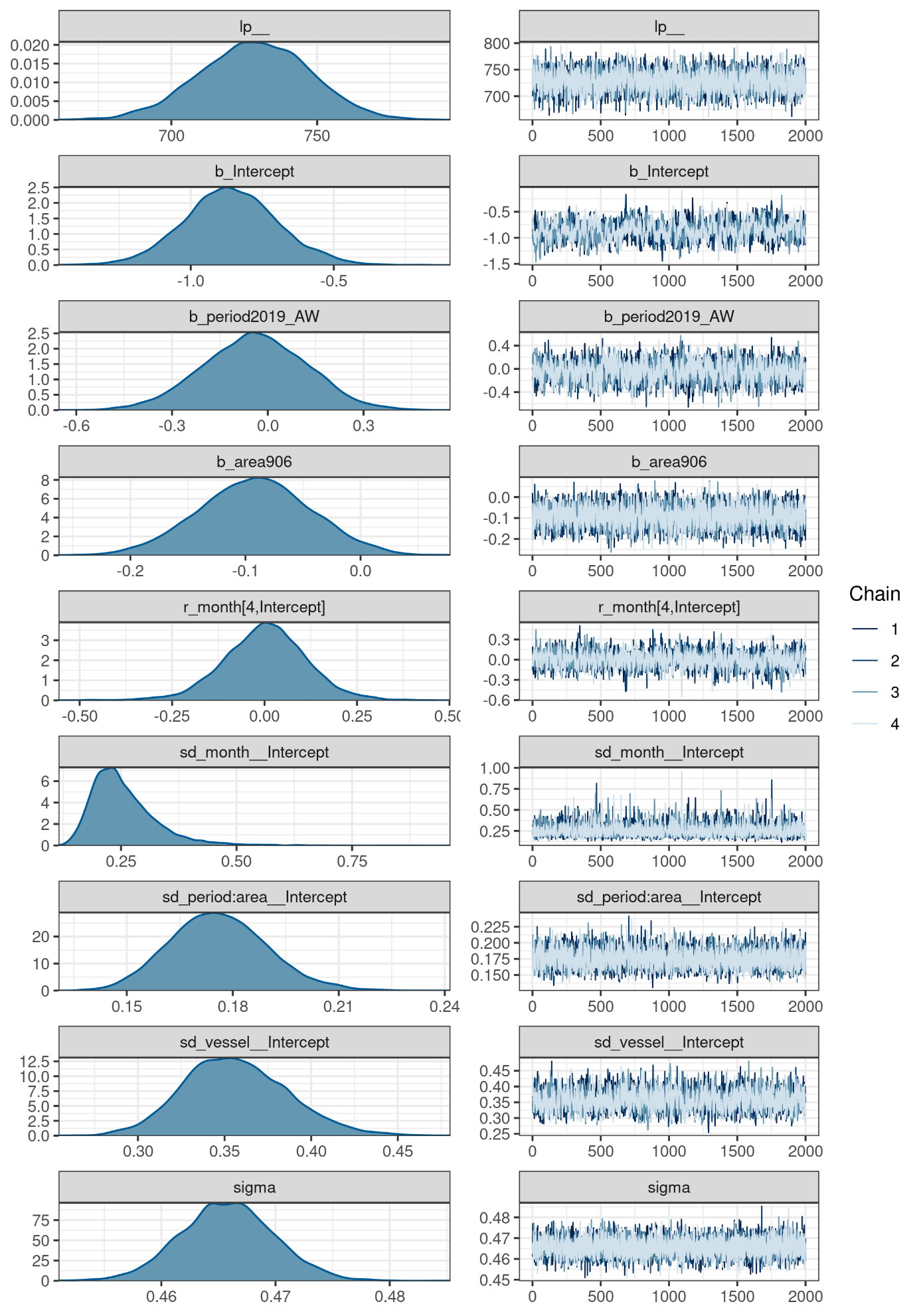


Figure 25: CELR: MCMC trace plots for a selection of parameters.

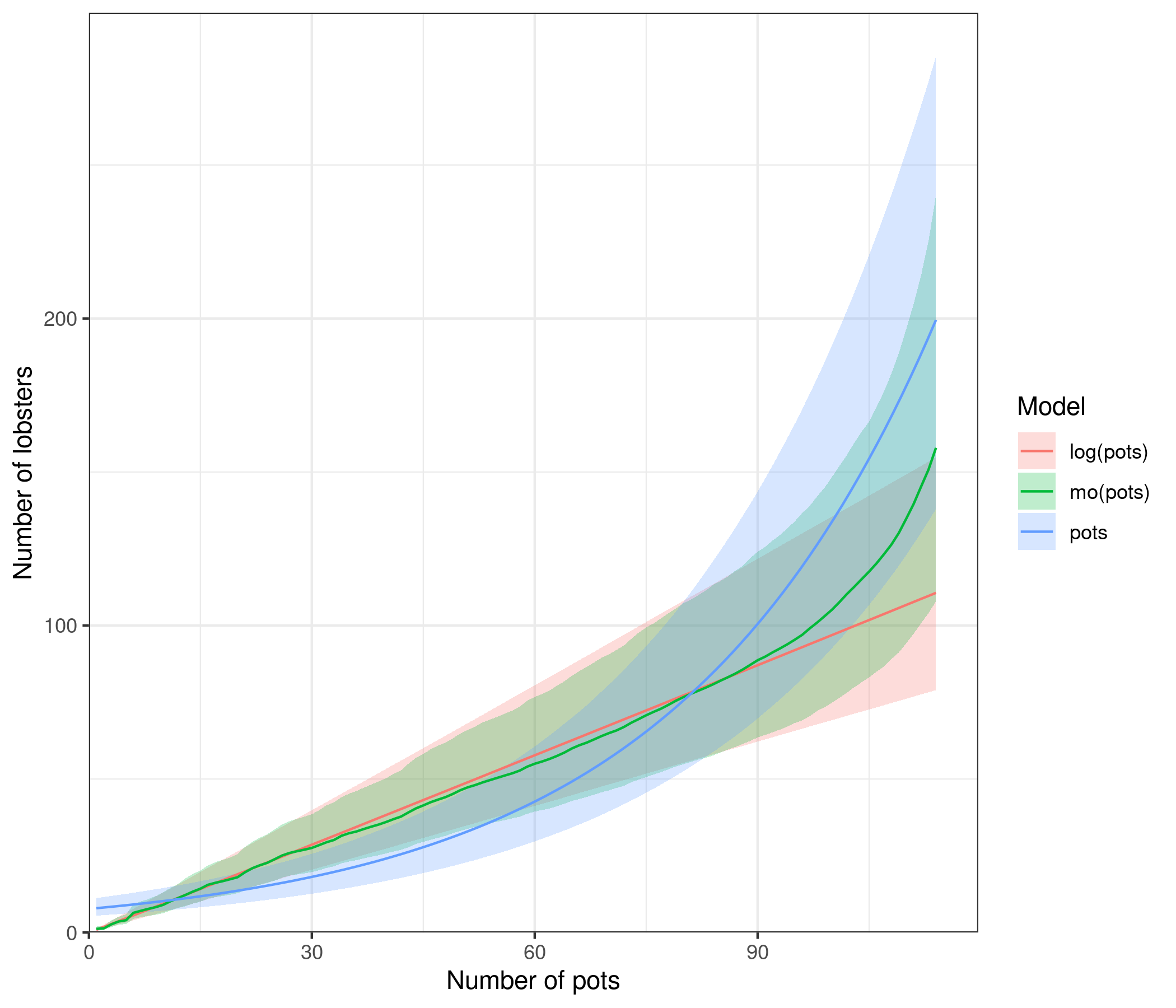


Figure 27: CELR: conditional effect of the number of pots lifted on the number of lobsters caught using the data aggregated by month for the models f(pots) + period + (1 | month) + (1 | vessel) + area + (1 | period × area) where f(pots) is either pots, log(pots), or mo(pots).

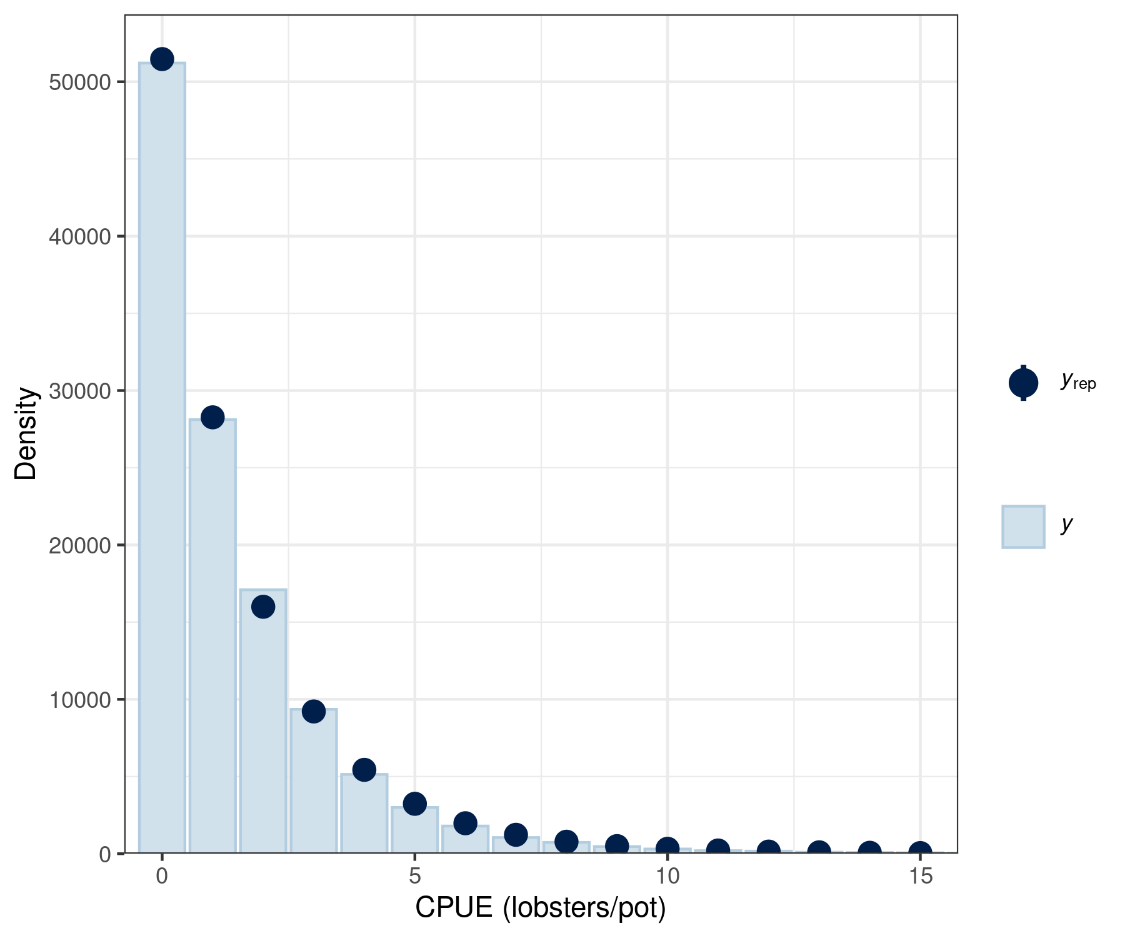


Figure 33: Logbook: posterior predictive density plots for the model ‘period + area + (1 | vessel × area) + (1 | vessel) + (1 | month)’ assuming the Poisson [top] or negative binomial [bottom] distribution. In each plot, *y* is the density of the data (solid blue line) and *yrep* is the density of 100 draws of data from the posterior predictive distribution.

1. Quantifish Ltd, New Zealand [↑](#footnote-ref-1)