Importance of Prior Predictive Checks in Bayesian Stock Assessment Models

Kyuhan Kim, Philipp Neubauer, and Kath Large

CAPAM virtual workshop
31 Jan. – 3 Feb. 2022

DRAGONFLY
Data Science

Introduction

- Fisheries management decisions in New Zealand are based on stock assessment models where inferences are typically carried out using a Bayesian approach

- The choice of priors on model parameters tends to be made based on the biological meanings of individual parameters (e.g., log-normal distributions on positive parameters)

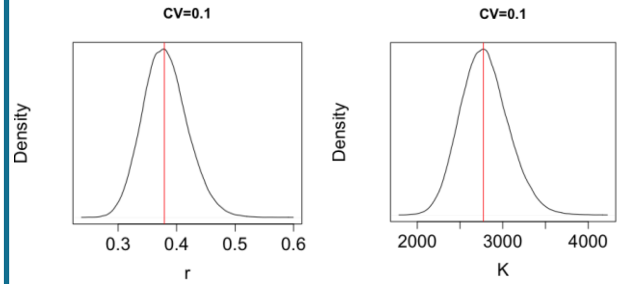
- Thorson and Cope (2017) pointed out a problem of placing a uniform prior on a scale parameter in the context of the model

- Using independent priors for individual parameters can potentially drive stock assessment models to implausible spaces (e.g., negative biomass)

- Walters et al. (2006) and Froese et al. (2017) used Monte Carlo simulations to derive plausible ranges of parameters, based on domain knowledge (e.g., non-negative biomass)

Introduction (simple example for PPCs)

Priors



Draw samples

$r^{(1)}, r^{(2)}, \ldots, r^{(5)}$

Draw samples

$K^{(1)}, K^{(2)}, \ldots, K^{(5)}$

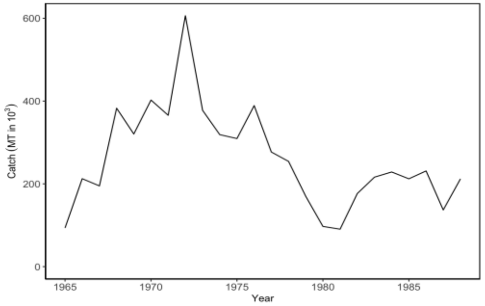
Input

Model

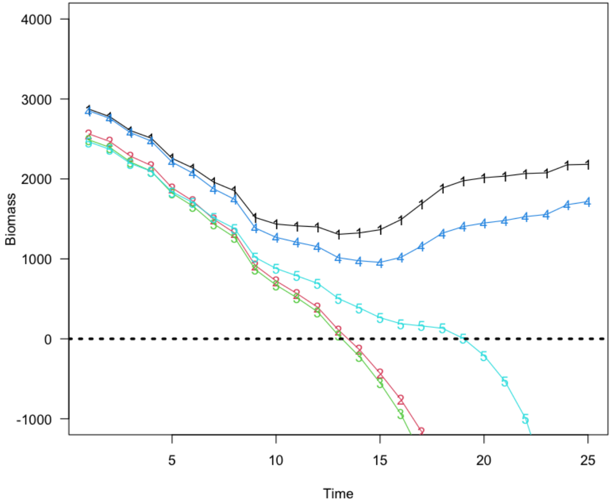
$$\begin{cases} B_1 & = K \\ B_{t+1} & = B_t + r \cdot B_t \cdot \left(1 - \frac{B_t}{K}\right) - C_t \end{cases}$$

Input

Catch



Output



Implausible space

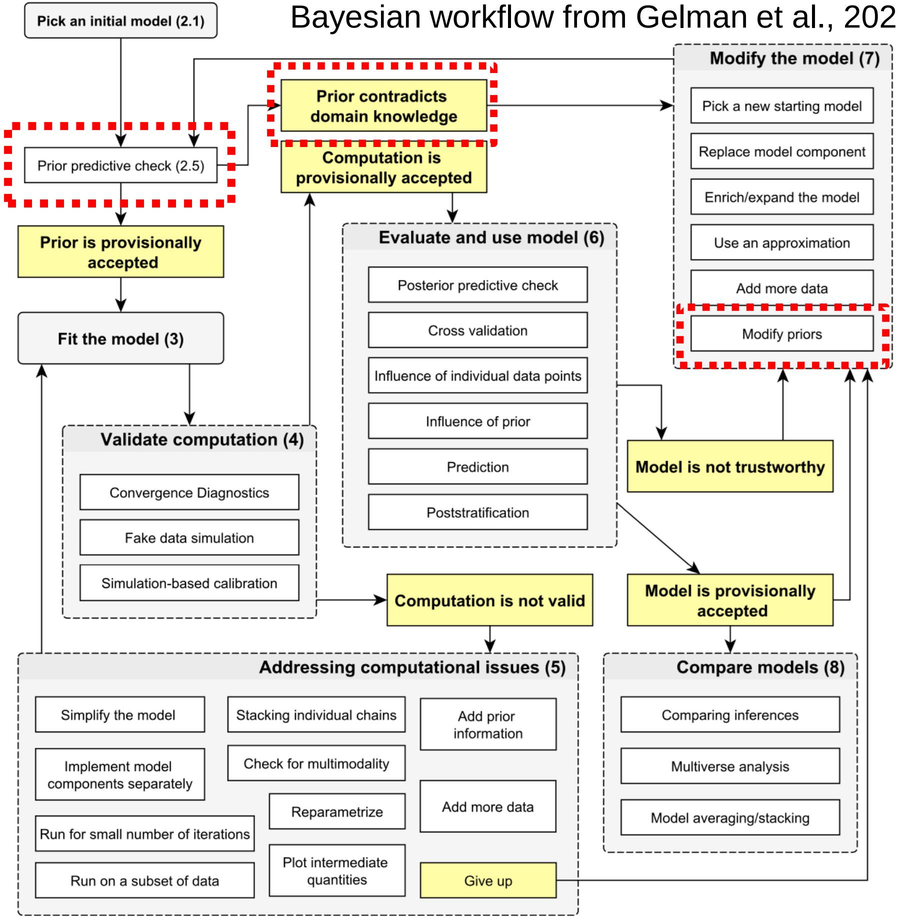
- Doing this process with many prior samples, and check if the predictive biomass trajectories are in the plausible space (i.e., $B_t > 0$).
- This simple process shows that independent priors on individual parameters may have different implications for the model outcome

Introduction

- Prior predictive checks are used to see how credible your assumptions are in terms of model outcomes.

- This simulation-based diagnostics allow you to understand the implications of a prior distribution in the context of a generative model

- Prior predictive checks are the crucial part of the Bayesian workflow (see the flowchart).



Bayesian workflow from Gelman et al., 2020

Pick an initial model (2.1)

Prior predictive check (2.5)

Prior contradicts domain knowledge

Computation is provisionally accepted

Prior is provisionally accepted

Fit the model (3)

Validate computation (4)

Convergence Diagnostics

Fake data simulation

Simulation-based calibration

Computation is not valid

Evaluate and use model (6)

Posterior predictive check

Cross validation

Influence of individual data points

Influence of prior

Prediction

Poststratification

Modify the model (7)

Pick a new starting model

Replace model component

Enrich/expand the model

Use an approximation

Add more data

Modify priors

Model is not trustworthy

Model is provisionally accepted

Addressing computational issues (5)

Simplify the model

Implement model components separately

Run for small number of iterations

Run on a subset of data

Stacking individual chains

Check for multimodality

Reparametrize

Plot intermediate quantities

Add prior information

Add more data

Give up

Compare models (8)

Comparing inferences

Multiverse analysis

Model averaging/stacking

The key points of this talk are...

- To show that Bayesian stock assessment models are no exceptions to the necessity of prior predictive checks.

- To demonstrate that assuming independent priors, where each marginal prior contains true information on individual parameters, can potentially drive stock assessment models to *a priori* implausible spaces, leading to biased inferences for key model parameters.

- To show that considering a single joint prior derived from priors over model inputs and outputs in terms of plausible ranges of model outcomes (e.g., non-negative values for stock biomass) is a necessary step in Bayesian stock assessments in order to eliminate this bias.

We performed simulation experiments to support the points above, using the logistic production and age-structured models.

Logistic Production Model

Logistic Production Model in a Bayesian setting

Logistic production model

$$\begin{cases} B_1 & = K \\ B_{t+1} & = B_t + r \cdot B_t \cdot \left(1 - \dfrac{B_t}{K}\right) - C_t \end{cases}$$

$$I_t = q \cdot B_t \cdot e^{\varepsilon_t}, \quad \text{where} \quad \varepsilon_t \overset{\text{iid}}{\sim} N(0, \sigma_o^2)$$

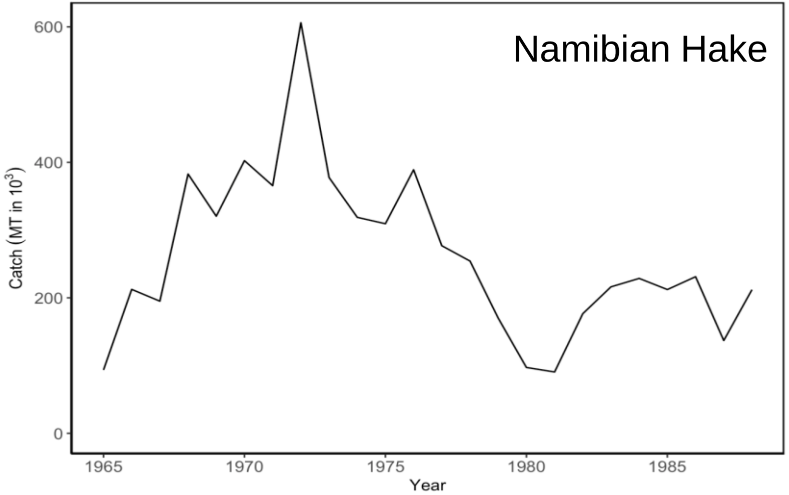
Priors

$$\begin{aligned} r &\sim \text{Log-Normal}(\log(0.379), 0.294^2); & \text{CV} = 0.3 \\ K &\sim \text{Log-Normal}(\log(2772.6), 0.472^2); & \text{CV} = 0.5 \\ q &\sim \text{Log-Normal}(\log(0.0006), 0.833^2); & \text{CV} = 1 \\ \sigma_o^2 &\sim \text{Log-Normal}(\log(0.3^2), 0.833^2); & \text{CV} = 1 \end{aligned}$$

Input values (from Polacheck et al. 1993)

$$r = 0.379; \quad K = 2772.6$$

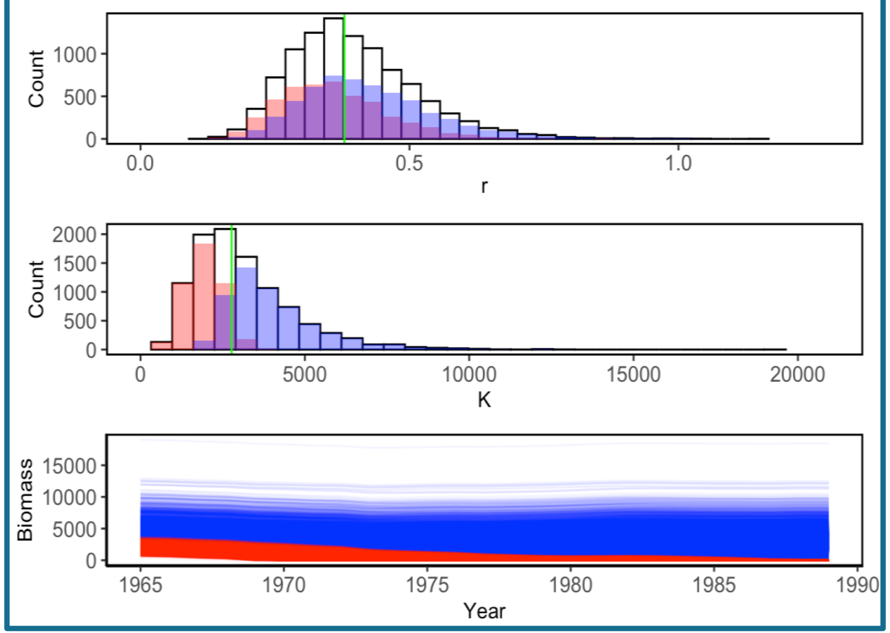
$$q = 0.0006; \quad \sigma_o^2 = 0.3^2$$



Simulation process for prior predictive checks

1. Draw a sample of *r* and *K* from their priors (i.e., the Log-Normal priors in our model)
2. Predict annual biomass B_t, using the samples drawn from 1 as inputs of the model
3. Then, check if B_t outside plausible bounds (e.g., $B_t > 0$)
4. Repeat 1-2 10 000 times to obtain a prior predictive space of B_t

$B_t \leq 0$ (red) vs. $B_t > 0$ (blue)



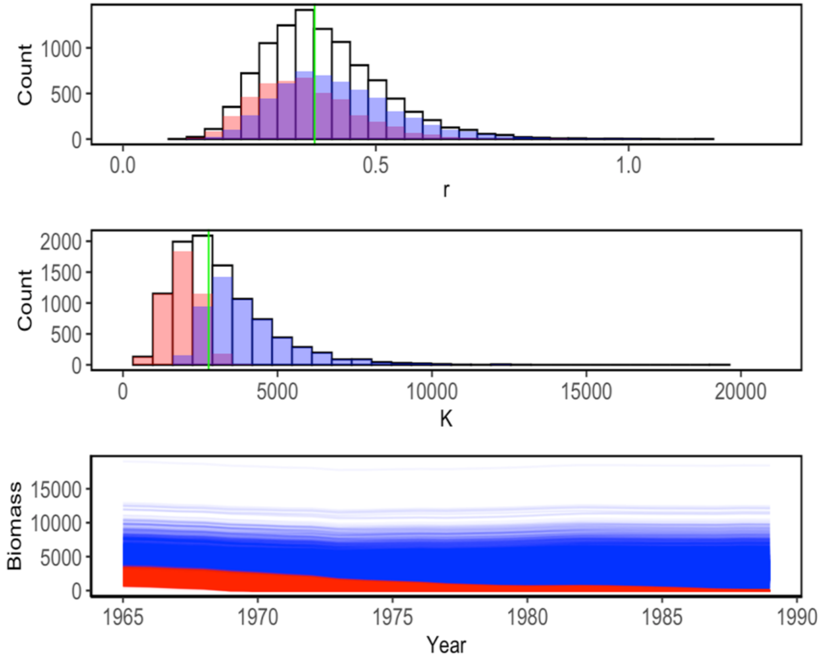
Examining the prior predictive space

- Some pairs of samples predicted the extinction of the population ($B_t \leq 0$), but we know that the population still exists (i.e., domain knowledge).

- Because of those problematic samples (i.e., red in the plot), an effective space of each prior (blue) becomes different from its original form (white).

- To develop a joint prior that predicts the biomass in the plausible space ($B_t > 0$), we need to find a relationship between *r* and *K.*

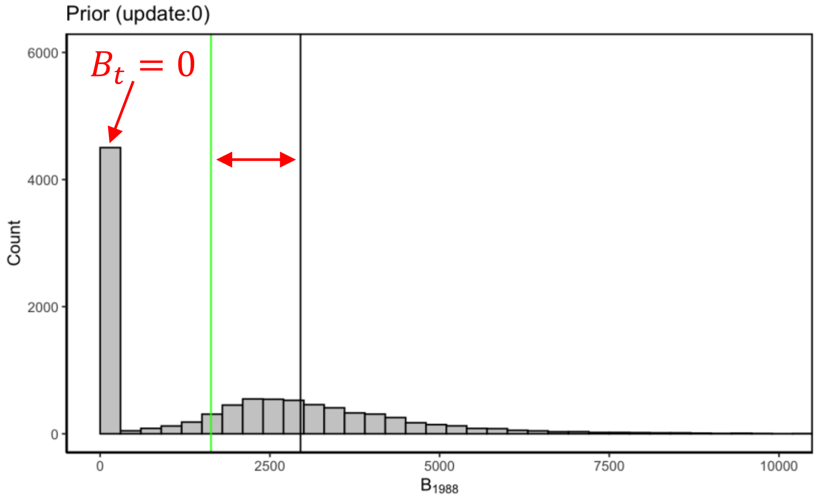
$B_t \leq 0$ (red) vs. $B_t > 0$ (blue)



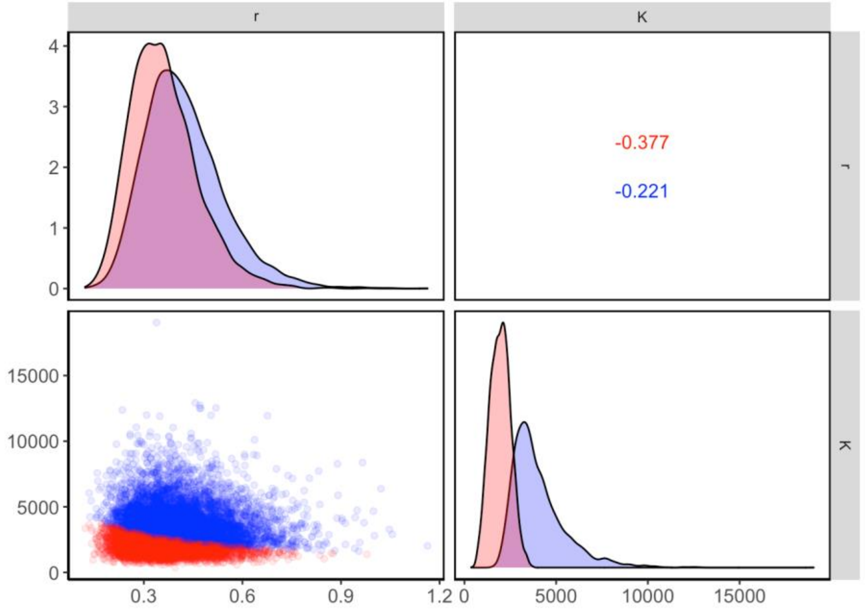
Examining the prior predictive space

Prior predictive space of the last year biomass

- Green vertical line: true biomass
- Black vertical line: the median of the predictive space, excluding the first bar of the histogram



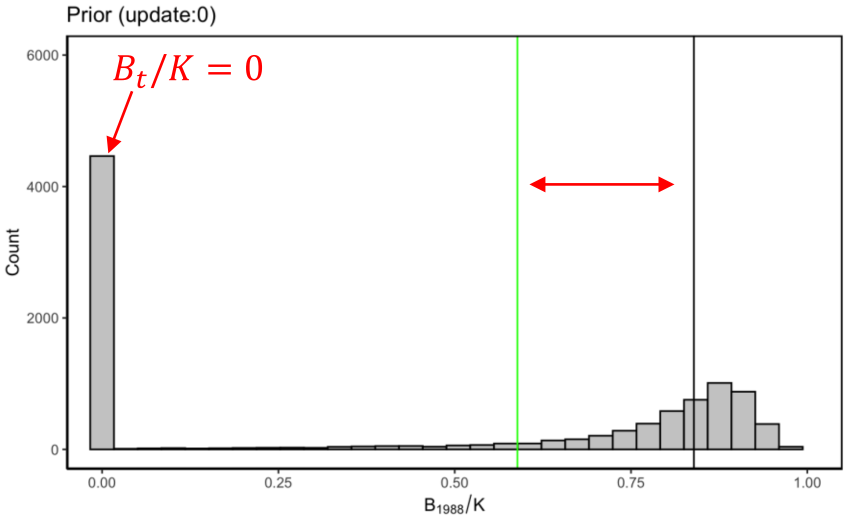
$B_t \leq 0$ (red) vs. $B_t > 0$ (blue)



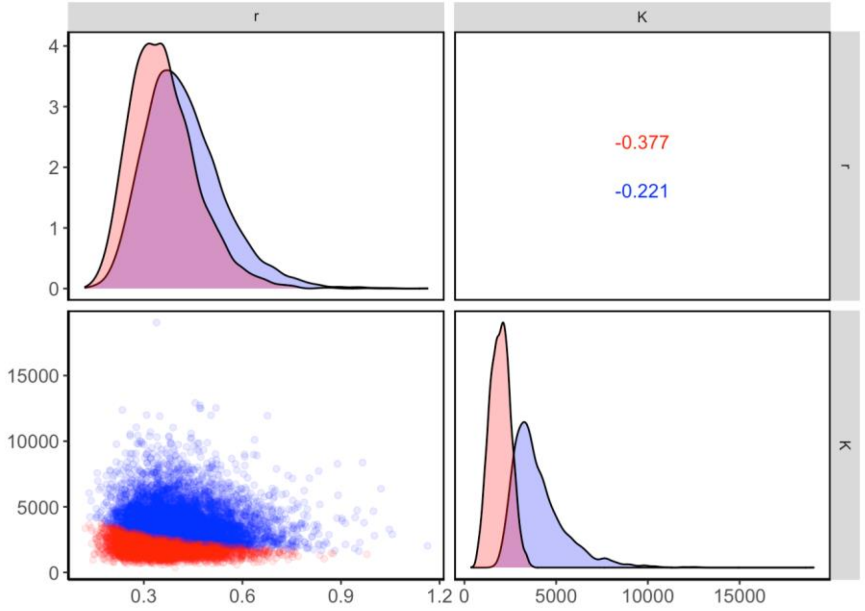
Examining the prior predictive space

Prior predictive space of the last year stock status

- Green vertical line: true biomass
- Black vertical line: the median of the predictive space, excluding the first bar of the histogram



$B_t \leq 0$ (red) vs. $B_t > 0$ (blue)

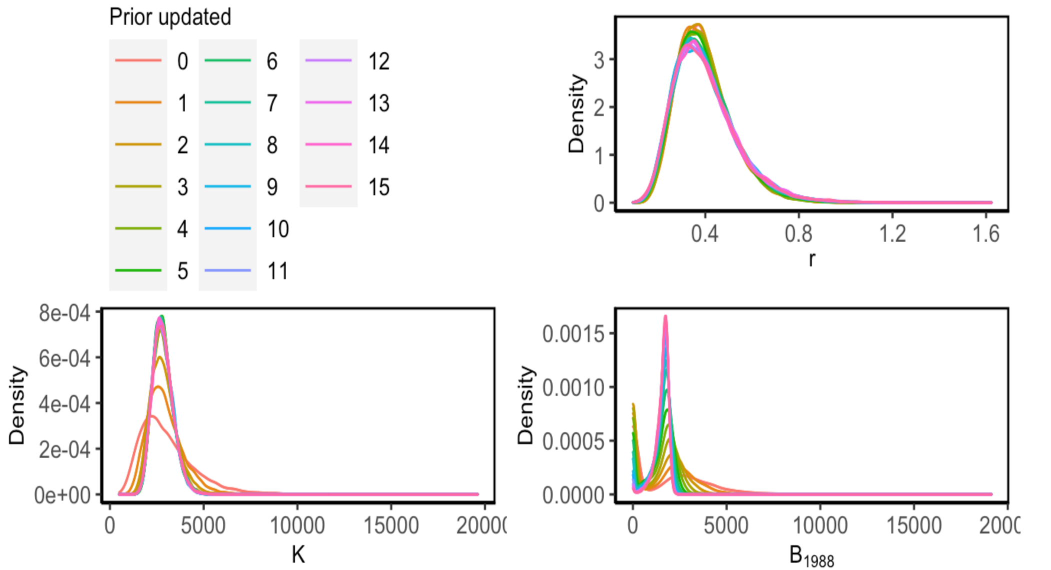


Steps to develop a joint (correlated) prior from a sampling and resampling process

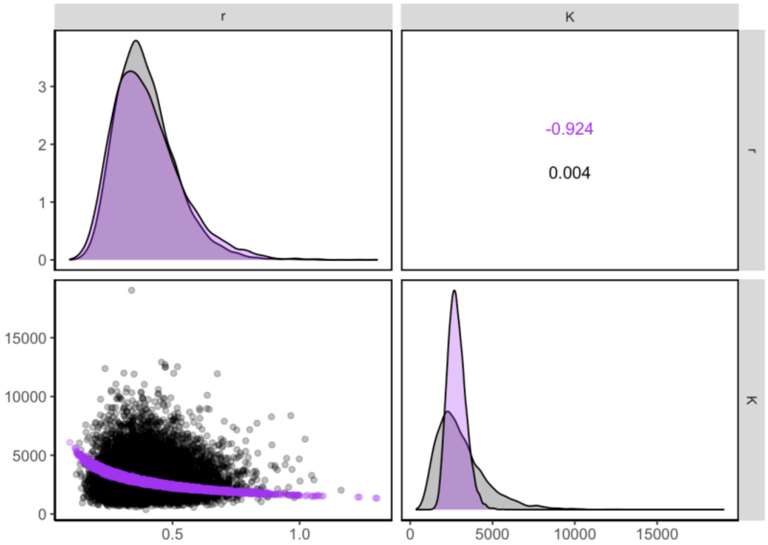
1. Draw 10 000 samples of *r* and *K* from the Log-Normal priors
2. Predict annual biomass, using the samples as inputs of the model
3. Remove a pair of *r* and *K* samples which drive the biomass to extinction $(\text{i.e., } B_t \leq 0)$
4. Calculate a covariance matrix for log(*r*) and log(*K*), using the remaining samples
5. Redraw 10 000 samples of *r* and *K* from a MVN distribution to incorporate the covariance structure
6. Repeat 2-5 until over 99% of samples predict $B_t > 0$

Priors developed from the resampling technique

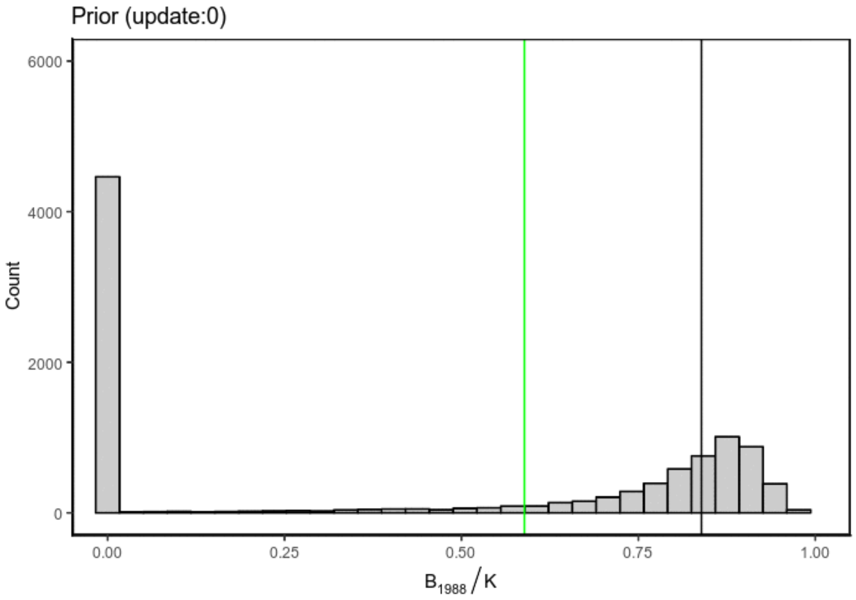
Marginal distributions of all sequentially updated priors

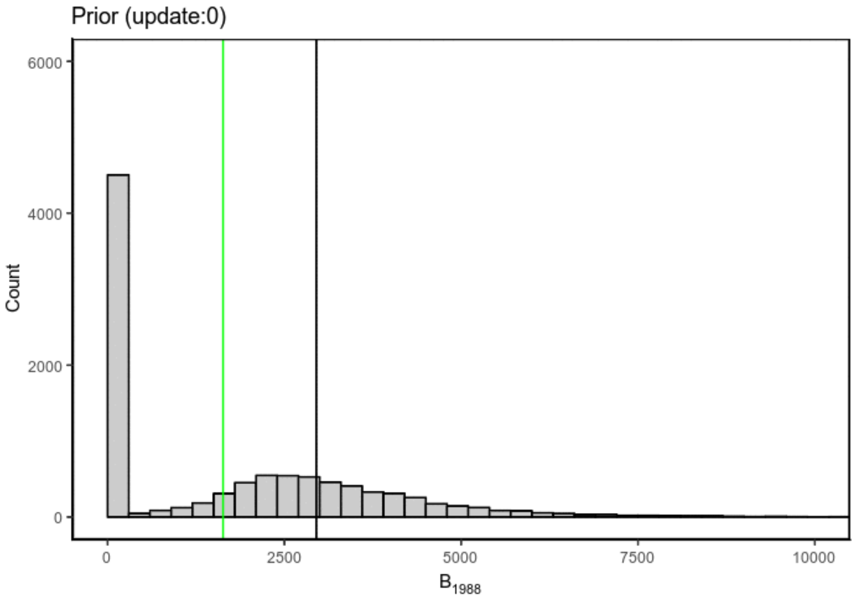


The original independent priors (black) vs. the final correlated prior (purple)

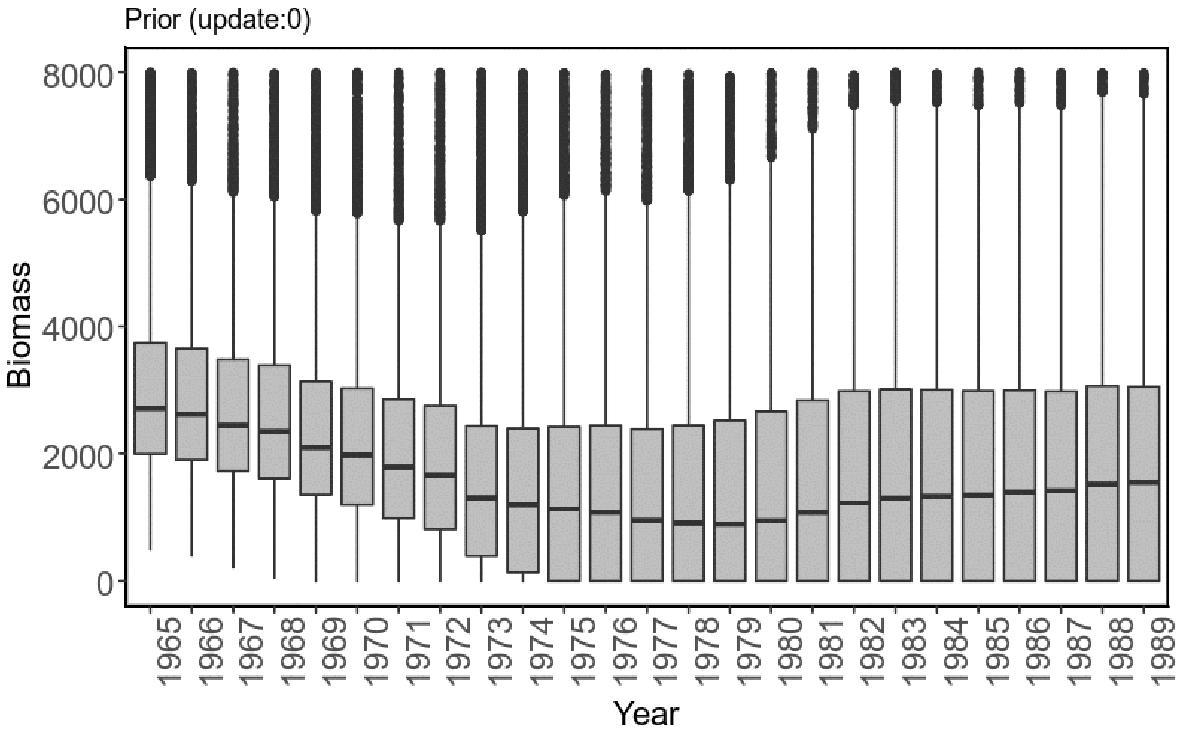


Priors developed from the resampling technique





Priors developed from the resampling technique

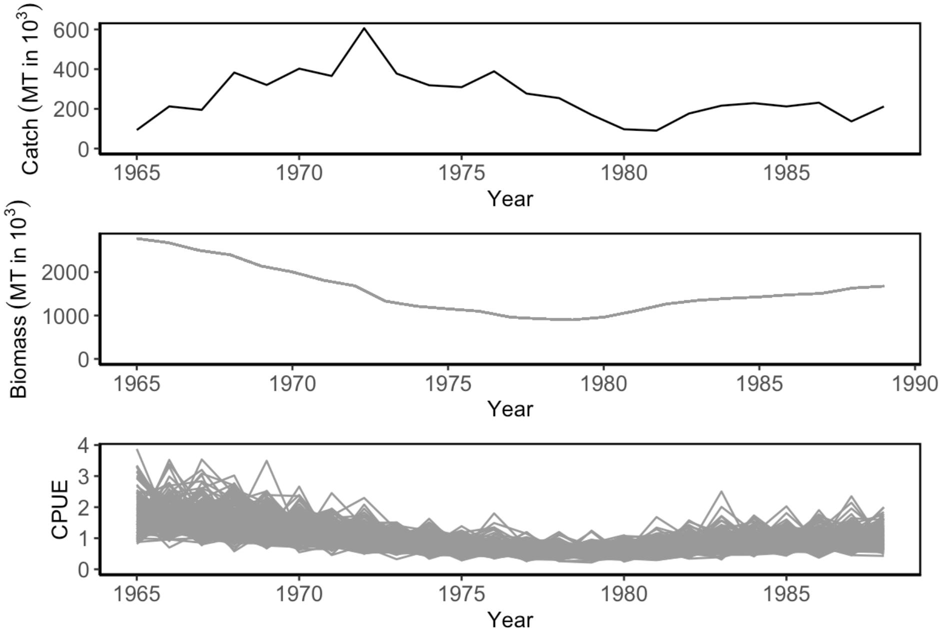


Simulation experiment (procedure)

Parametric bootstrap test

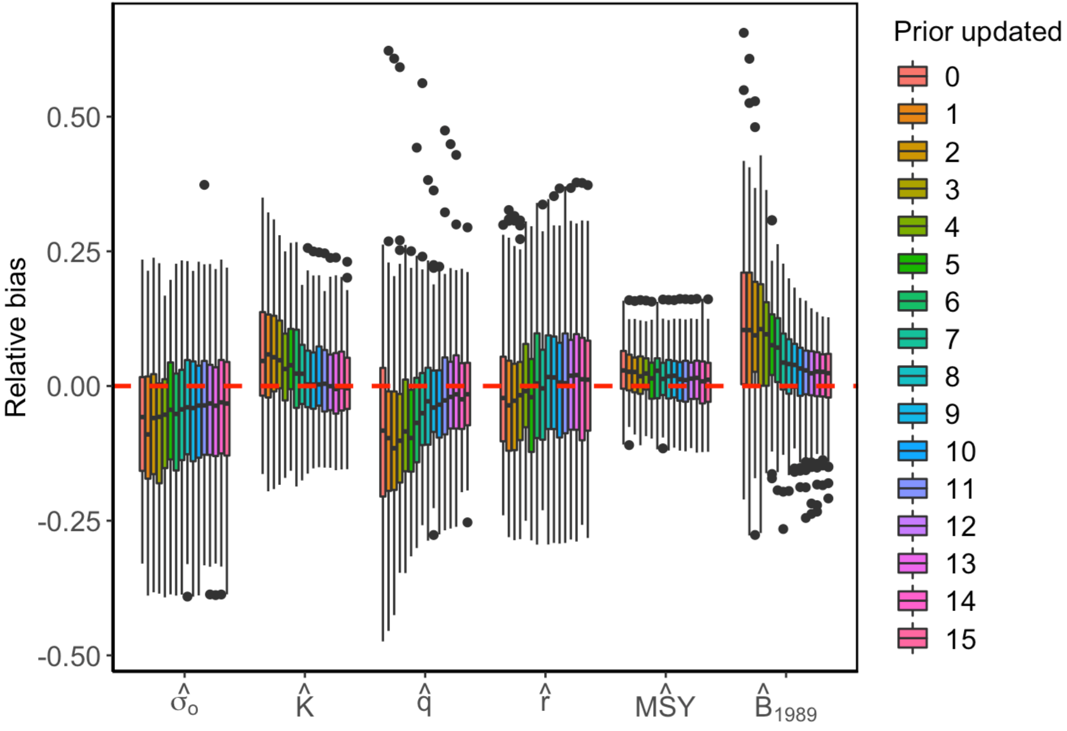
1. Simulate data (i.e., CPUE) given the input values and the model
2. Fit the full Bayesian model to simulated data, using Stan
3. Check model convergence (i.e., no divergent transitions and Rhat <1.05)
4. Calculate a relative bias of the estimates of the parameters (i.e., the median of the posterior)
5. repeat 1-4 200 times

Simulated population and data

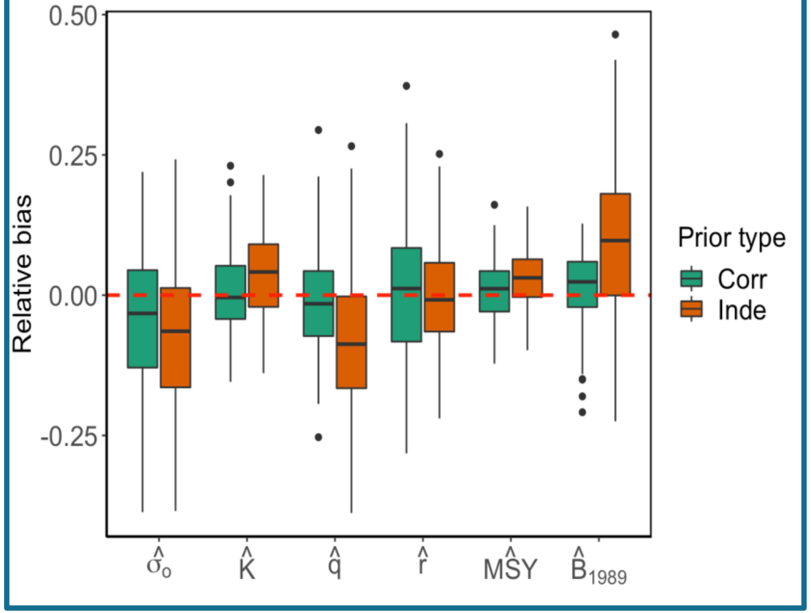


Simulation experiment (results)

Impact of priors on parameter estimation



with correlation (dark green) vs. without correlation (brown)



Age-structured Model

Age-structured Model in a Bayesian setting

Process equation

For $t = 1$,

$$N_{a,1} = \begin{cases} R_0, & \text{for } a = 1 \\ N_{a-1,1} \cdot e^{-M}, & \text{for } 2 \leq a < A \\ \dfrac{N_{A-1,1} \cdot e^{-M}}{1 - e^{-M}}, & \text{for } a = A \end{cases}$$

For $t > 1$,

$$N_{a,t} = \begin{cases} \dfrac{\alpha \cdot \mathrm{Egg}_{t-1}}{1 + \beta \cdot \mathrm{Egg}_{t-1}}, & \text{for } a = 1 \\ N_{a-1,t-1} \cdot e^{-M} \cdot (1 - v_{a-1} \cdot U_{t-1}), & \text{for } 2 \leq a < A \\ N_{a-1,t-1} \cdot e^{-M} \cdot (1 - v_{a-1} \cdot U_{t-1}) \\ \quad + N_{a,t-1} \cdot e^{-M} \cdot (1 - v_a \cdot U_{t-1}), & \text{for } a = A \end{cases}$$

Observation equation (abundance index)

$$I_t = q \cdot B_t \cdot e^{\varepsilon_t}; \quad \varepsilon_t \overset{\text{iid}}{\sim} N(0, \sigma_o^2),$$

$$\text{where } B_t = \sum_{a=1}^{A} w_a \cdot v_a \cdot N_{a,t}$$

Observation equation (age composition; see Thorson et al., 2017)

$$\boldsymbol{n}_t \sim \text{Dirichlet-multinomial}(n_t^{\text{input}}, \boldsymbol{\gamma}_t, \hat{\boldsymbol{P}}_t),$$

$$\text{where } \gamma_{a,t} = \theta \cdot n_t^{\text{input}} \cdot \hat{P}_{a,t};$$

$$\left(n_t^{\text{eff}} = \frac{1}{1+\theta} + n_t^{\text{input}} \cdot \frac{\theta}{1+\theta}\right)$$

Age-structured Model in a Bayesian setting

Priors

$$\alpha \sim \text{Log-Normal}(\log(5.941), 0.472^2); \quad \text{CV} = 0.5$$

$$\beta \sim \text{Log-Normal}(\log(0.00628), 0.472^2); \quad \text{CV} = 0.5$$

$$M \sim \text{Log-Normal}(\log(0.35), 0.198^2); \quad \text{CV} = 0.2$$

$$\sigma_o^2 \sim \text{Log-Normal}(\log(0.3^2), 0.833^2); \quad \text{CV} = 1$$

$$\theta \sim \text{Log-Normal}(\log(0.05), 0.833^2); \quad \text{CV} = 1$$

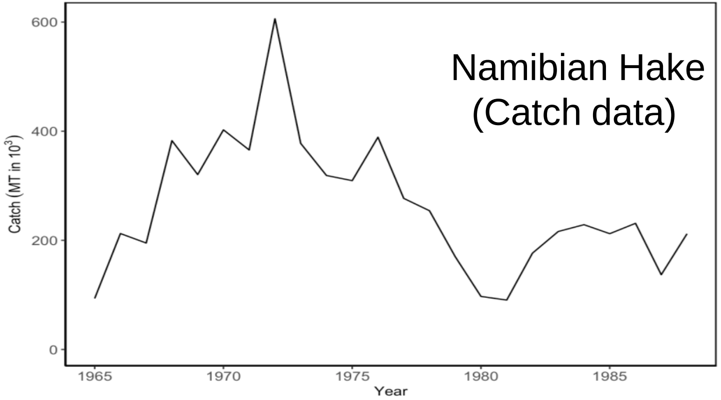
$$q \sim \text{Log-Normal}(\log(0.0006), 0.833^2); \quad \text{CV} = 1$$

Input values (from Forrest et al., 2008)

$$\alpha = 5.941; \quad \beta = 0.00628; \quad M = 0.35$$

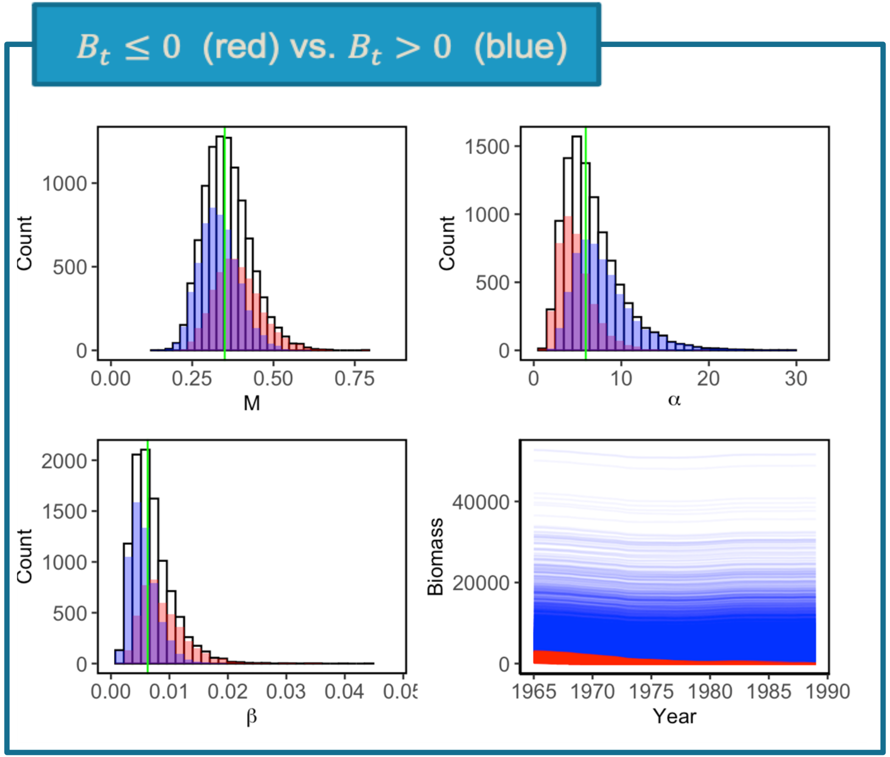
$$\sigma_o^2 = 0.3^2; \quad \theta = 0.05; \quad q = 0.0006$$

And other model parameters (e.g., Maturity, selectivity, size, etc.) are fixed at their true values obtained from Forrest et al., 2008



Simulation process for prior predictive checks

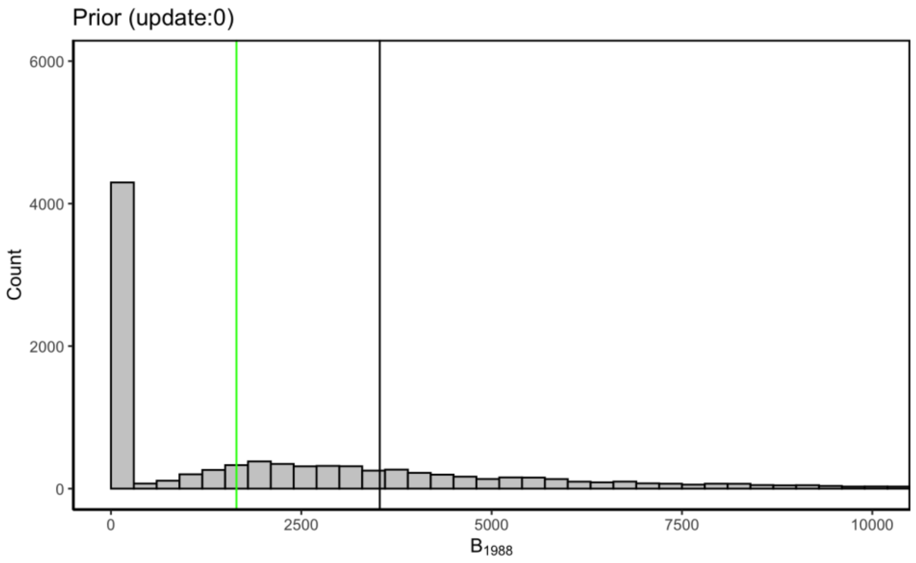
1. Draw a sample of α, β and M from their priors (i.e., the Log-Normal priors in our model)
2. Predict annual biomass B_t, using the samples drawn from 1 as inputs of the model
3. Then, check if B_t outside plausible bounds (e.g., $B_t > 0$)
4. Repeat 1-2 10 000 times to obtain a prior predictive space of B_t



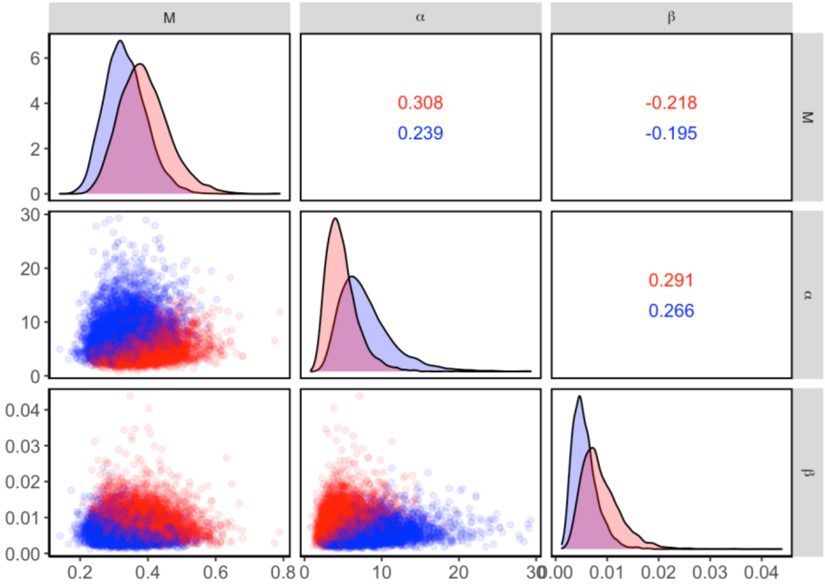
$B_t \leq 0$ (red) vs. $B_t > 0$ (blue)

Examining the prior predictive space

Prior predictive space of the last year biomass

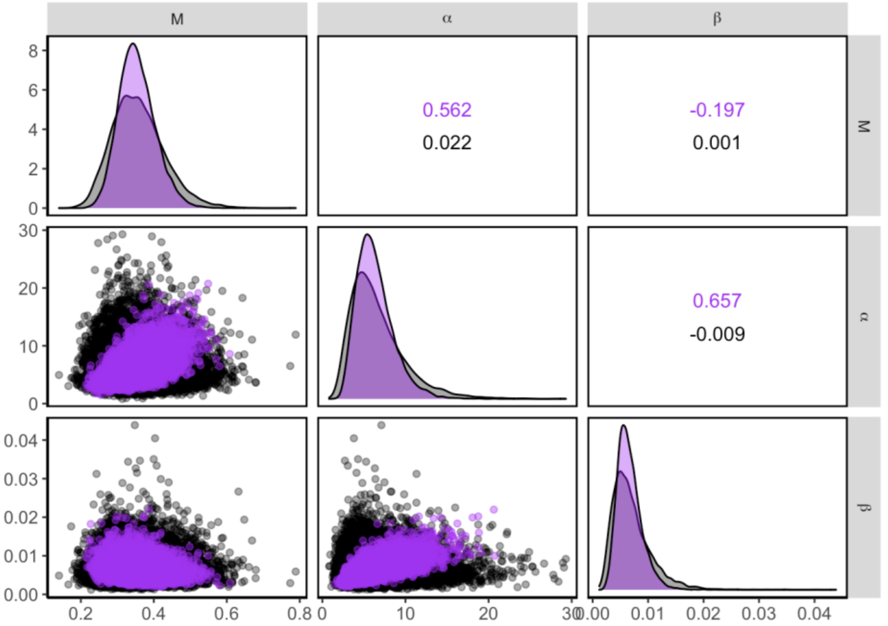


$B_t \le 0$ (red) vs. $B_t > 0$ (blue)

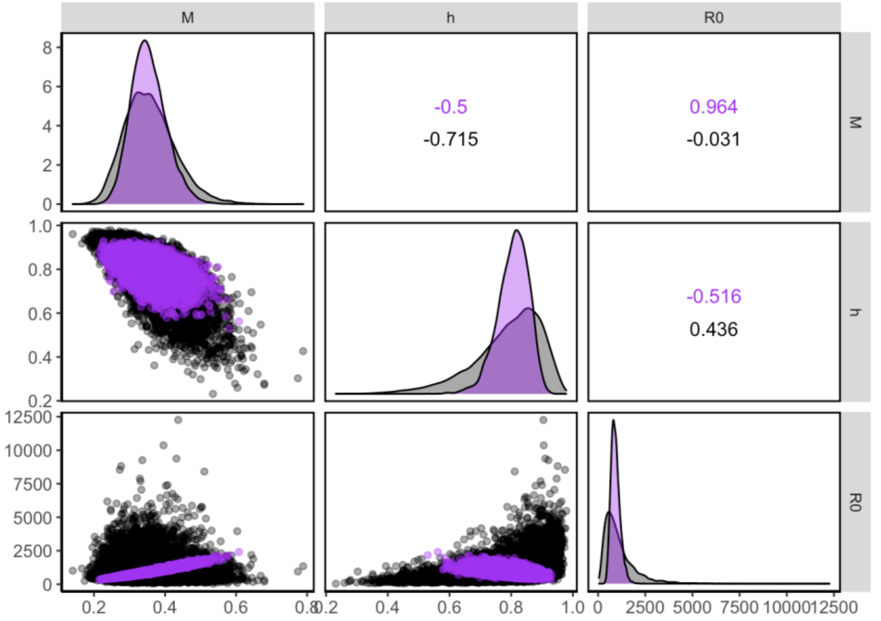


Priors developed from the resampling technique

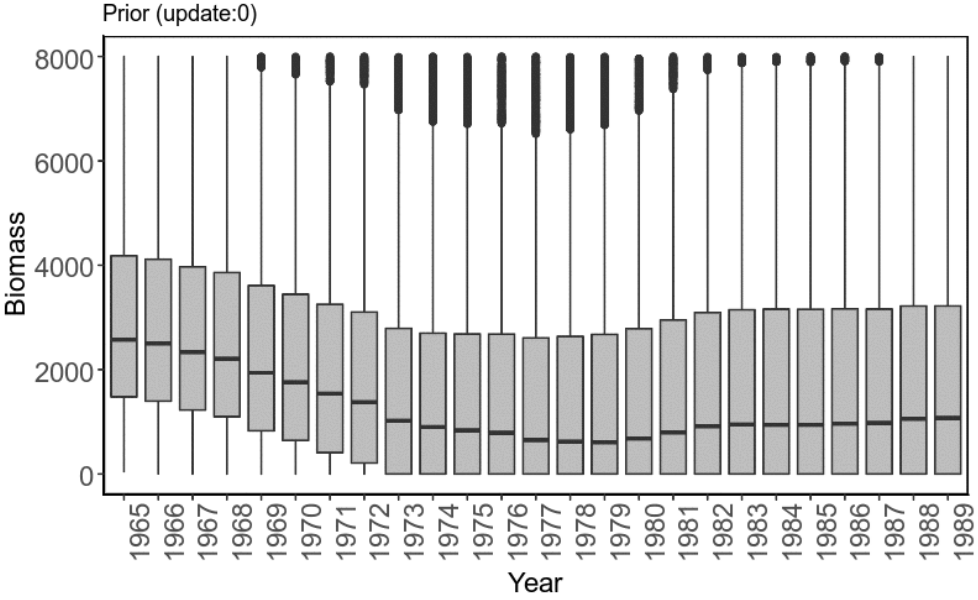
The original independent priors (black) vs. the final correlated prior (purple)

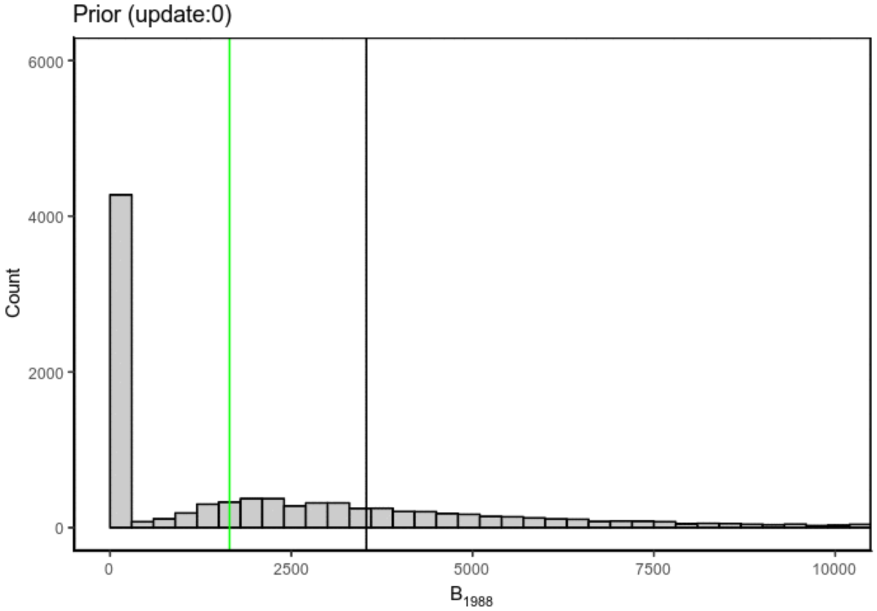


In terms of *h* and *R0*



Priors developed from the resampling technique



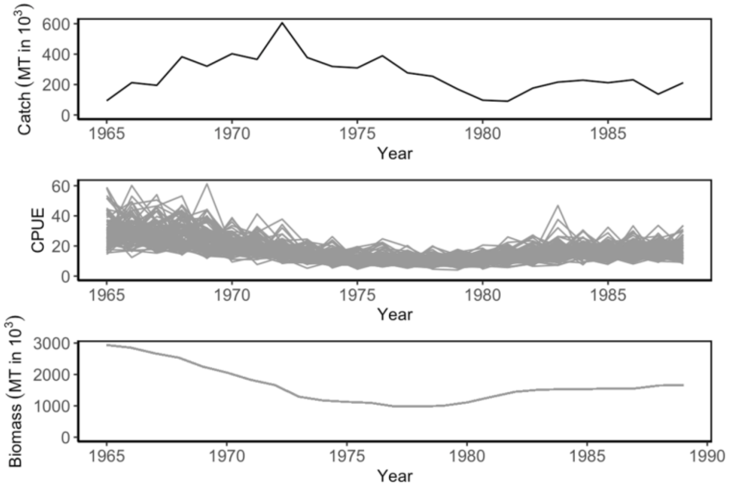


Simulation experiment (procedure)

Parametric bootstrap test

1. Simulate data (i.e., CPUE and age composition) given the input values and the model
2. Fit the full Bayesian model to simulated data (Scenario 1: only CPUE; Scenario 2: both CPUE and age composition), using Stan
3. Check model convergence (i.e., no divergent transitions and Rhat <1.05)
4. Calculate a relative bias of the estimates of the parameters (i.e., the median of the posterior)
5. repeat 1-4 100 times

Simulated population and data

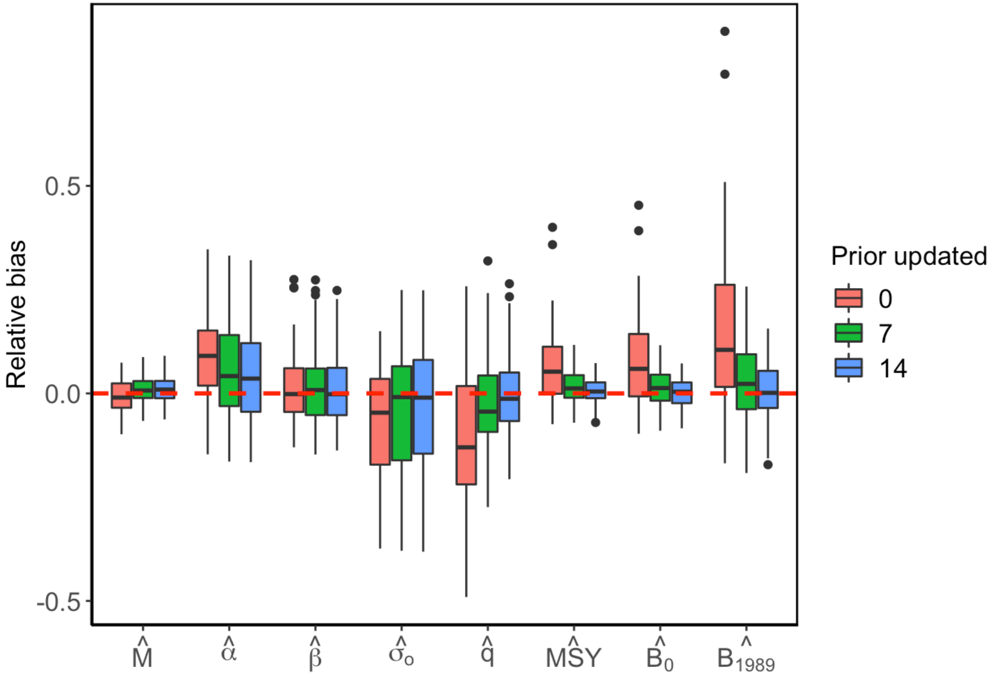


+

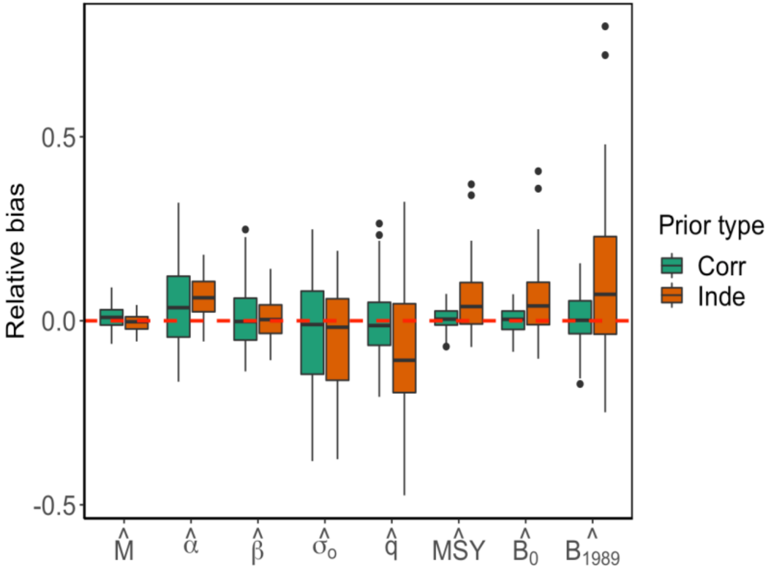
Age composition data

Simulation experiment (results for scenario 1: only CPUE)

Impact of priors on parameter estimation

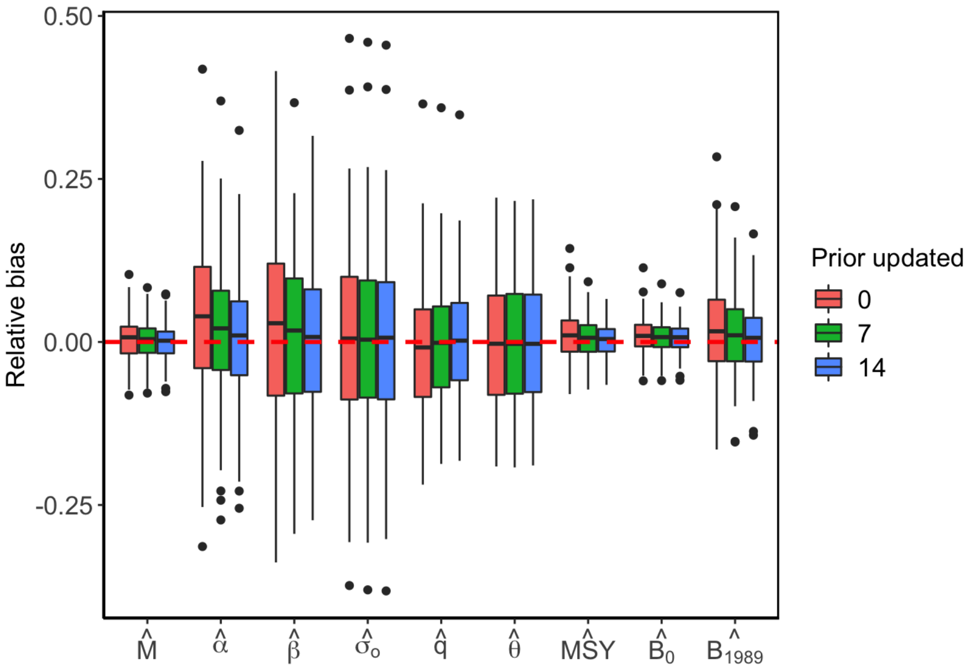


with correlation (dark green) vs. without correlation (brown)

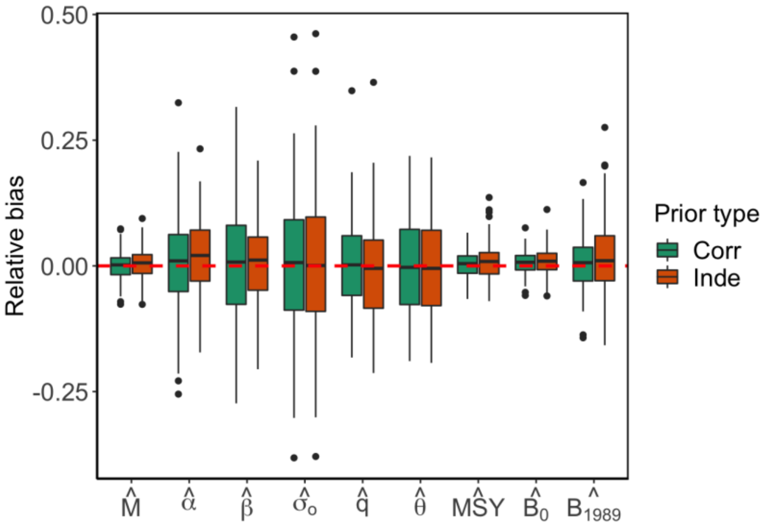


Simulation experiment
(results for scenario 2: both CPUE and age composition)

Impact of priors on parameter estimation



with correlation (dark green) vs. without correlation (brown)



Discussion

- As mentioned in Kennedy et al. (2019), we showed that a poor choice of priors in terms of model outcomes can cause underpowered/biased inferences.

- As Gelman et al. (2017) pointed out, priors should only be interpreted in terms of the likelihood. For example, in the likelihood of the two stock assessment models we used, the biomass was log-transformed (i.e., $\log(B_t) \sim N(\log(qB_t), \sigma_o^2)$); thus, samples, drawn from the independent priors, which predicted $B_t \leq 0$ were not defined in the models.

- As a baseline setting in our simulation studies, we used informative priors, all of which are correctly placed on individual parameters. We note that this simulation setting made the results (relative bias of the estimates) seem less dramatic, but tells us that a prior conflict even occurs in such an ideal situation.

Discussion

- We showed that even using the informative priors, even while centered on the true values of the parameters, can cause biased estimates, especially for the scale parameters (i.e., catchability *q* and the carrying capacity *K*) if the priors are not interpreted in terms of the model outcomes.

- We suggest Bayesian stock assessment models should use prior predictive checks and prior pooling to minimise prior conflict and potential bias in posterior inferences caused by marginal prior choices.

- The iterative resampling procedure allows for pooling of priors over inputs and outputs (e.g., Poole & Raftery 2000), ensuring that Bayesian inference remains well defined, and can be implemented with MCMC assuming a parametric form of the joint prior.

Thank you for your attention

DRAGONFLY
Data Science

Good with data

Additional slides

Logistic Production Model (reparameterised)

Logistic Production Model (reparameterised)

Model structure

$$\begin{cases} B_1 & = K \\ B_{t+1} & = \left[B_t + r \cdot B_t \cdot \left(1 - \frac{B_t}{K} \right) \right] \cdot e^{-q \cdot E_t} \end{cases}$$

$$C_t = \left[B_t + r \cdot B_t \cdot \left(1 - \frac{B_t}{K} \right) \right] \cdot \left(1 - e^{-q \cdot E_t} \right) \cdot e^{\varepsilon_t},$$

$$\text{where} \quad \varepsilon_t \overset{\text{iid}}{\sim} N(0, \sigma_o^2)$$

Input values (from Polacheck et al. 1993)

$$r = 0.379; \quad K = 2772.6$$

$$q = 0.0006; \quad \sigma_o^2 = 0.3^2$$

$$E_t = C_t / I_t$$

Priors

$$r \sim \text{Log-Normal}(\log(0.379), 0.294^2); \quad \text{CV} = 0.3$$

$$K \sim \text{Log-Normal}(\log(2772.6), 0.472^2); \quad \text{CV} = 0.5$$

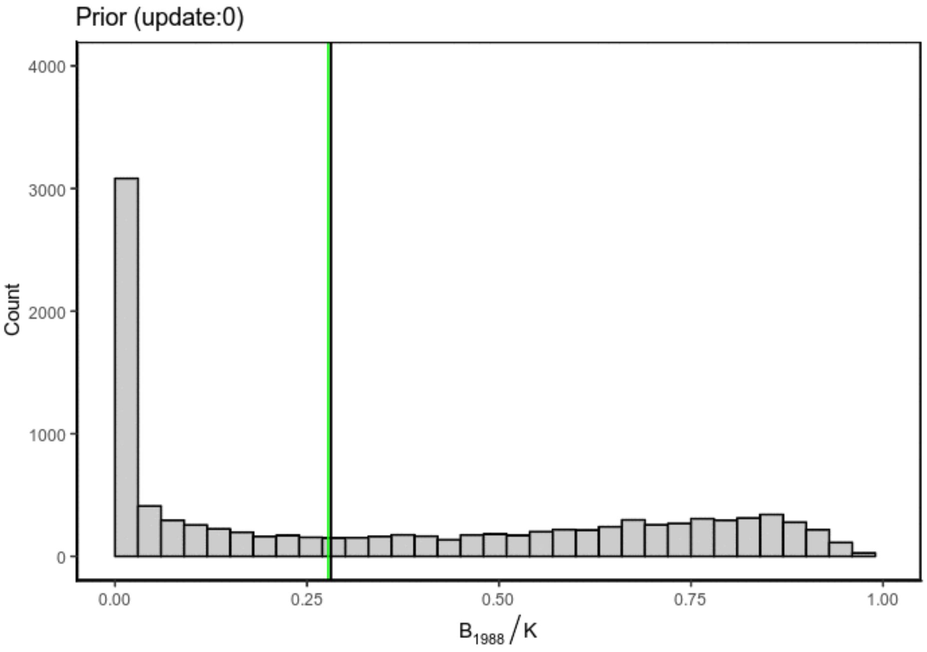
$$q \sim \text{Log-Normal}(\log(0.0006), 0.833^2); \quad \text{CV} = 1$$

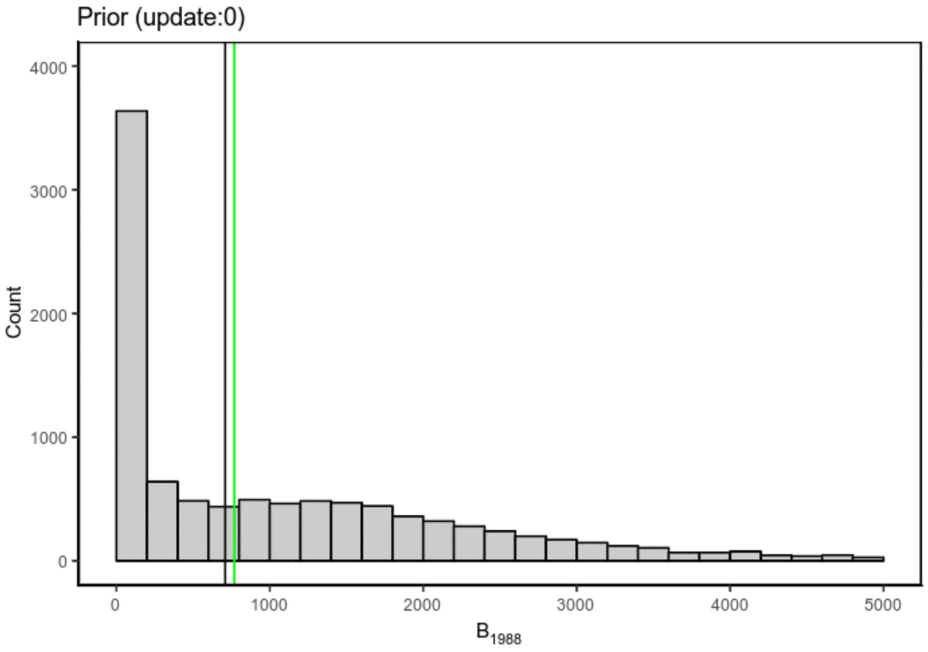
$$\sigma_o^2 \sim \text{Log-Normal}(\log(0.3^2), 0.833^2); \quad \text{CV} = 1$$

Steps to develop a correlated prior from a sampling and resampling process

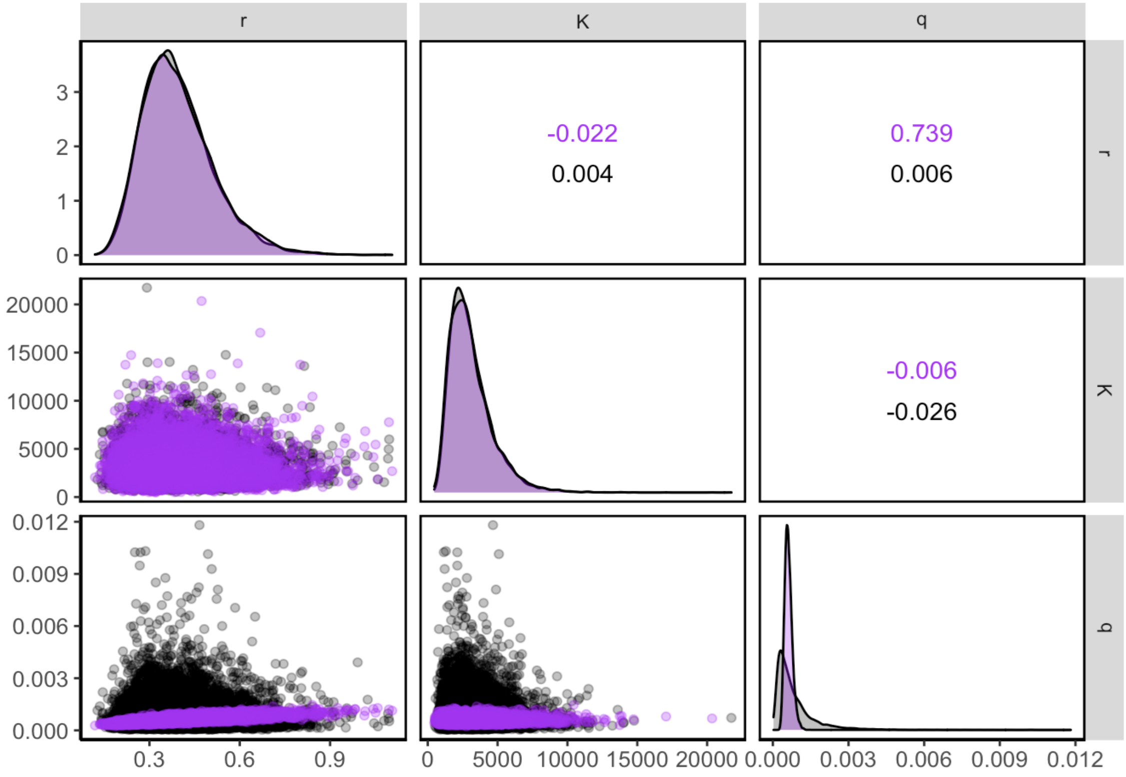
1. Draw 10 000 samples of *r, K,* and *q* from the Log-Normal priors
2. Predict annual biomass, using the samples as inputs of the model
3. Remove a set of *r, K,* and *q* samples which predict the implausibly low stock status in the last year $\text{(i.e., remove those that predict } B_{1988}/K < 0.05)$
4. Calculate a covariance matrix for log(*r*), log(*K*) and log(*q*), using the remaining samples
5. Redraw 10 000 samples of *r, K,* and *q* from a MVN distribution to incorporate the covariance structure
6. Repeat 2-5 until over 99% of samples predict $B_{1988}/K \geq 0.05$

Priors developed from the resampling technique





Priors developed from the resampling technique



Simulation experiment (procedure)

Parametric bootstrap test

1. Simulate data (i.e., Catch) given the input values and the model

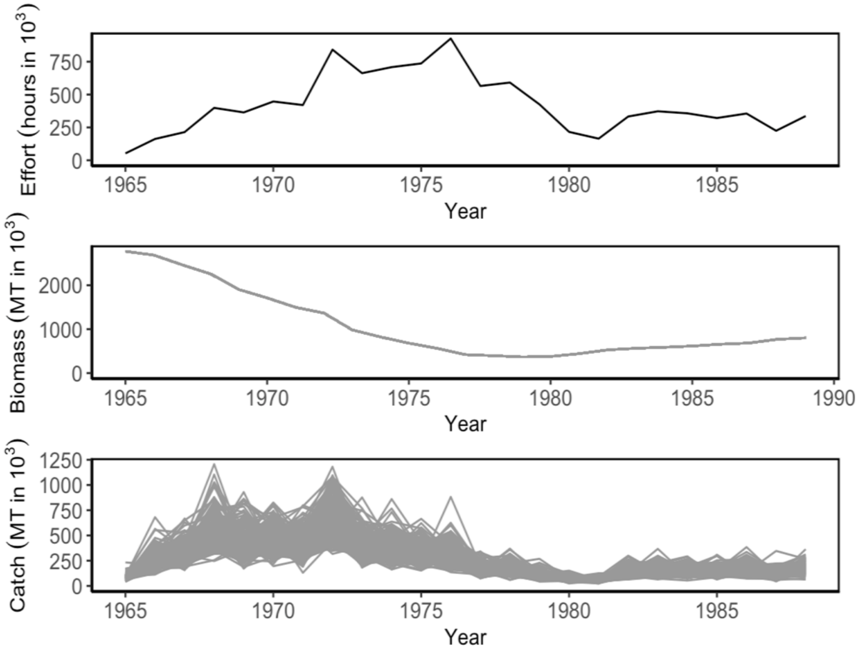
2. Fit the full Bayesian model to simulated data (scenario 1: all data points; scenario 2: last 10 data points), using Stan

3. Check model convergence (i.e., no divergent transitions and Rhat <1.05)

4. Calculate a relative bias of the estimates of the parameters (i.e., the median of the posterior)

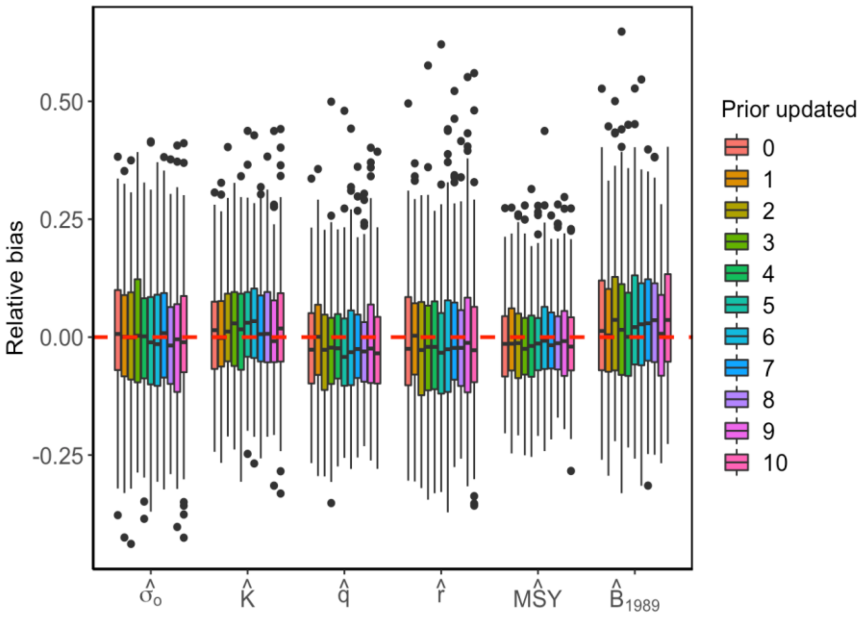
5. repeat 1-4 200 times

Simulated population and data

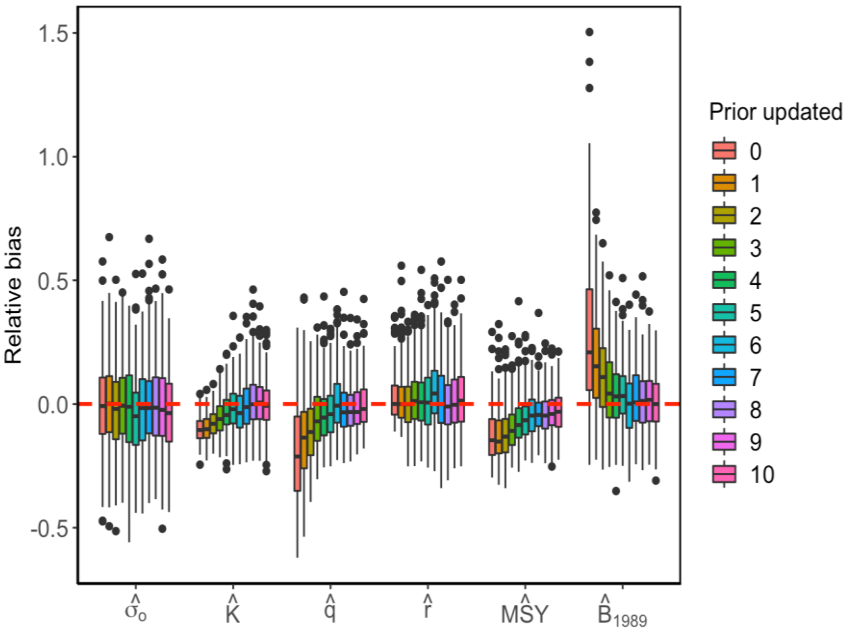


Simulation experiment (results)

Scenario 1: fitted to all data points

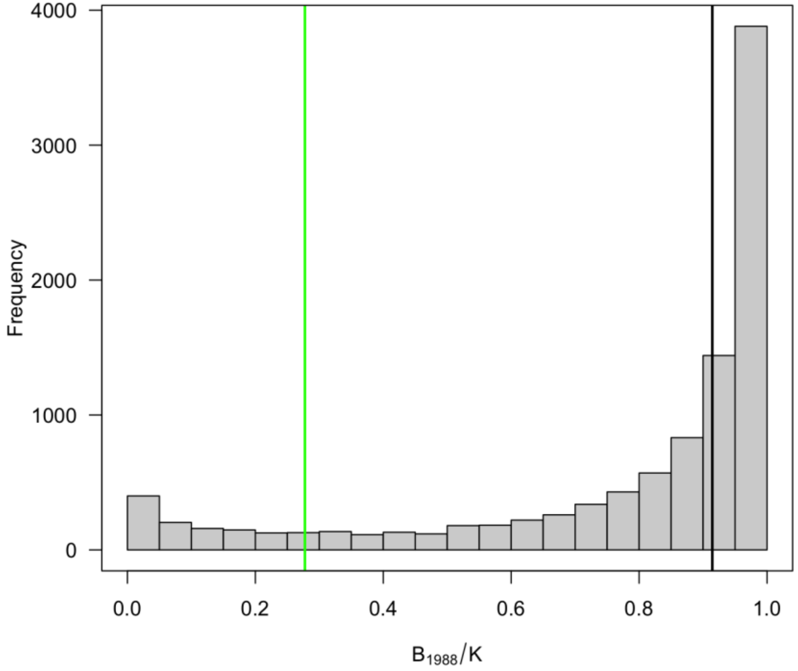


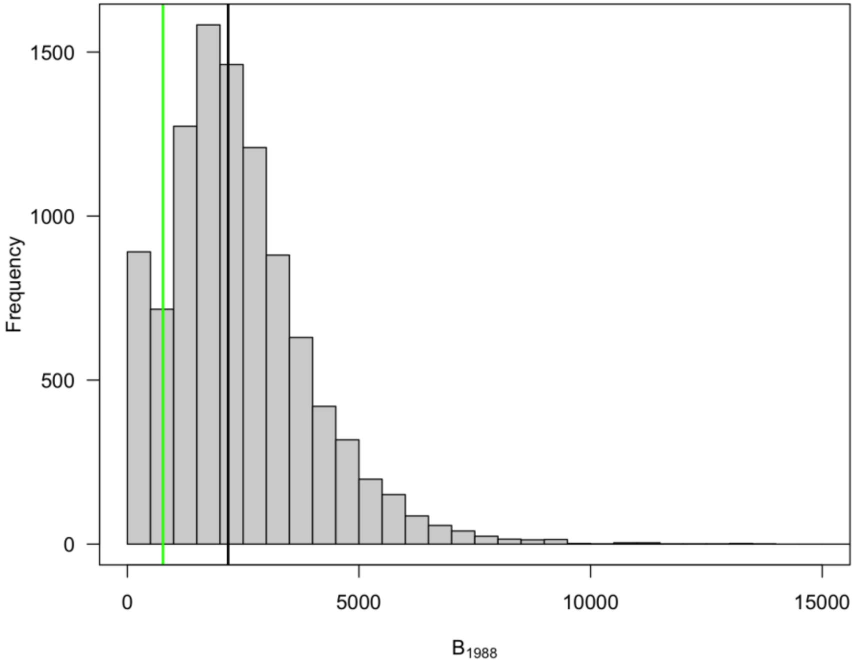
Scenario 2: fitted to the last 10 data points



Uniform prior on log(q) instead of the Log-Normal prior

$$\log(q) \sim \text{Uniform}[\log(10^{-5}), \log(10^{-3})]$$


Frequency
B_{1988}/K


Frequency
B_{1988}