Comparing the performance of three data weighting methods when allowing for time-varying selectivity

Haikun Xu[1*,2] (hkxu@iattc.org), James T. Thorson[3,4] (James.Thorson@noaa.gov), Richard D. Methot[5] (Richard.Methot@noaa.gov)

[1] Inter-American Tropical Tuna Commission, 8901 La Jolla Shores Drive, La Jolla, CA 92037, USA

[2] Previously at School of Aquatic and Fishery Sciences, University of Washington, Box 355020, Seattle, WA 98105, USA

[3] Habitat and Ecosystem Process Research program, Alaska Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, 7600 Sand Point Way NE, Seattle, WA 98115

[4] Previously at Fishery Resource Analysis and Monitoring Division, Northwest Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, 2725 Montlake Blvd. East, Seattle, WA 98112, USA

[5] NOAA Senior Scientist for Stock Assessments, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, 2725 Montlake Blvd. East, Seattle, WA 98112, USA

** Corresponding author address:*

Haikun Xu, Inter-American Tropical Tuna Commission,

8901 La Jolla Shores Drive, La Jolla, CA 92037, USA

Telephone: +1 8583342824; E-Mail: hkxu@iattc.org

Abstract

How to properly weight composition data is an important ongoing research topic for fisheries stock assessments and multiple methods for weighting composition data have been developed. Although several studies indicated that properly accounting for time-varying selectivity can reduce estimation biases in population biomass and management-related quantities, no study to date has compared the performance of widely-used data weighting methods when allowing for time-varying selectivity. Here, we conducted four simulation experiments for this topic, aiming to provide guidance on weighting age-composition data given time-varying selectivity. The first simulation experiment showed that over-weighting should be avoided in general and even when estimating time-varying selectivity. The second simulation experiment compared three data weighting methods (McAllister-Ianelli, Francis, and Dirichlet-multinomial), within which the Dirichlet-multinomial method outperformed the other two methods when selectivity is time-varying. The third and fourth simulation experiments further showed that given time-varying selectivity, the Dirichlet-multinomial method still performed well when age-composition data were over-dispersed and when the level of selectivity variation needed to be simultaneously estimated. Our simulation results support using the Dirichlet-multinomial method when estimating time-varying fishery selectivity. Also, the simulation results suggest that improving stock assessments by accounting for time-varying selectivity requires simultaneously addressing data weighting and time-varying selectivity.

Keywords: Data weighting; time-varying selectivity; Dirichlet-multinomial method; age-composition data

1. Introduction

Fisheries managers use stock assessment models to predict the likely impact of alternative management actions on fishery sustainability. In many jurisdictions worldwide, fisheries managers are recommended or required to manage fishery catches and population abundance in accordance with management targets or limits that are determined from stock assessment models (Methot 2009). Accurate predictions of likely management impacts require stock assessment models to appropriately approximate biological processes including growth, mortality, maturity, and reproduction. To estimate these different processes, modern assessment models typically fit to a wide range of data sources including abundance indices, subsamples of age/length/sex-composition in fishery-independent surveys or fishery operations, and total fishery landings (Maunder and Punt 2013).

Composition data from a fishery are usually not independent between ages and contain a reduced amount of information than they would do if sampled independently (Francis 2011, Maunder 2011, Thorson 2014). Due to, for example, age- or length-specific behaviors such as schooling and aggregating, the age and length of fish from the same set are more similar than from different sets. Namely, composition samples are positively correlated among adjacent age or length bins, contradicting the assumption of random sampling in the widely used multinomial distribution for composition data (Francis 2011). This phenomenon, which is referred to as “over-dispersion”, increases the variance in composition samples and decreases the effective sample size. The weighting of composition data in stock assessment models is positively related to the effective sample size, which is used by stock assessment scientists to accommodate unknown observation error and model mis-specification.

Stock assessment models will often estimate different values for stock status and productivity when fitted to different subsets of available data (Maunder et al. 2017). In particular, inferring trends in population abundance from age- and length-composition sampling depends upon correct specification of many biological processes including mortality, growth, and availability to survey or fishery operations, and mis-specification of these processes will cause information in age- and length-composition data to be biased with respect to true trends in abundance (Minte-Vera et al. 2017). For this reason and others, several papers have suggested that age- and length-composition data should be "down-weighted" relative to abundance index data whenever the two provide conflicting information about abundance trends (Francis 2011, 2014). Widely-used methods for weighting composition data include the methods by McAllister and Ianelli (1997), Francis (2011), and the linear parameterization of the Dirichlet-multinomial (D-M) likelihood (Thorson et al. 2017). These and other methods all have in common that they down-weight age- and length-composition data more when the assessment model predictions and available data are greatly different, and down-weight less (or even up-weight) when predictions and data match well. However, these methods also differ in well-documented respects: the McAllister-Ianelli (M-I) and Francis methods require iteratively fitting a stock assessment model and do not characterize model uncertainty caused by estimating data-weights, while the D-M method can be efficiently estimated as a model parameter with associated measure of uncertainty (Francis 2017, Thorson 2018).

One main reason for down-weighting composition data is that stock assessment models explain process error as observation or sampling error (Maunder and Piner 2017). For example, when a time-varying selectivity is mis-specified to be time-invariant, the stock assessment model attributes the discrepancy between observed and expected compositions solely to the error in the composition sampling process. As an alternative, analysts may instead revise a stock assessment model such that it is better able to predict available data. There are many biological processes that could cause the proportion of an age/length/sex-composition that are selected by a given fishery or survey operation to vary over time, including spatial patterns in fishery effort (Sampson and Scott 2012), environmentally-driven changes in vertical distribution (Kotwicki et al. 2015), or spatial redistribution among well- and poorly-sampled habitats (Thorson et al. 2013a). In general, these processes will cause "model mis-specification", wherein a model cannot match available data even if unlimited or perfect data are available. In these cases, a stock assessment can estimate additional fixed or random effects representing time-varying selectivity, and this will generally increase the match between available data and model predictions (Lowe et al. 2017, Martell and Stewart 2014, Xu et al. 2018). In fact, a number of simulation studies have shown that properly accounting for time-varying selectivity can reduce estimation biases in population biomass and management-related quantities (Stewart and Martell 2014, Stewart and Monnahan 2017, Thorson and Taylor 2014, Xu et al. 2018).

By increasing the match between model predictions and data, estimating time-varying selectivity will clearly impact the degree of data weighting estimated by different methods. However, no study to date has compared the performance of widely-used data weighting methods in the case where the assessment model estimates time-varying selectivity. Under the assumption of constant selectivity, Maunder (2011) showed that estimating the effective sample size of composition data led to an improvement over using the nominal sample size (the number of fish sampled each year) if the corresponding true selectivity varied from year to year. Under the assumption of time-varying selectivity, in comparison, estimating the effective sample size of composition data (namely, weighting composition data) is more problematic because the true level of variation in selectivity is unknown. Using a simulation approach, Stewart and Monnahan (2017) explored the effects of data weighing on the performance of models with or without process error in selectivity. Based on simulation results, they concluded that assessment models should allow for process error in selectivity and should not excessively down-weight composition data.

The main objective of this paper was to compare the performance of three data weighting methods when allowing for time-varying selectivity. Previous studies (e.g., Hulson et al. 2012) has evaluated the performance of several data weighting methods using simulations, but our study is the first to use simulation to compare the performance of data weighting methods in assessment models that estimate time-varying selectivity. We first conducted a simulation experiment to evaluate the sensitivity of model performance to the extent to which fisheries age-composition data are weighted in assessment models both with and without process error in selectivity, given that the true selectivity is time-varying. This experiment aimed to answer the question: what are the consequences of under- or over-weighting age-composition data when process error in selectivity is ignored or estimated? We then conducted three simulation experiments to compare the performance of three (M-I, Francis, and D-M) data weighting methods given that the true selectivity is time-varying, aiming to address the following questions:

1) Which data weighting method performs best when the assessment model estimates time-varying selectivity?
2) How is the performance of the best data weighting method degraded owing to the over-dispersion in age-composition data?
3) Can we simultaneously weight age-composition data and estimate the selectivity variation penalty in stock assessments?

2. Materials and methods

In this paper, we compared three methods for weighting age-composition data based on simulation experiments that were undertaken by modifying an age-structured simulation-estimation package *CCSRA* (Thorson and Cope 2015). We first described the basic structure and hypotheses for the operating model (OM), sampling model (SM), and estimation model (EM) used in our simulation experiments. We then described in detail how the OM, SM, and EM were configured and how model performance was evaluated in each simulation experiment. In each simulation experiment, the OM simulated the true population dynamics from which the SM generated observation data. The EM was then fitted to the generated observation data and model performance was evaluated by comparing the estimates of population attributes that the EM provided with the corresponding true values that the OM simulated.

2.1. Simulation models

2.1.1. Operating model

The OM was an age-structured model (Table 1) that allows fishery selectivity to vary either independently or correlated from a specified parametric functional form. It was used in this study to simulate the true population dynamics for two species, Pacific hake (*Merluccius productus*) and Pacific sardine (*Sardinops sagax*), that correspond to a "periodic" and "opportunistic" type of life history, respectively (Table 2). The level of recruitment variation (σ_R in Eq. T1.2) was specified to be low (0.4) for Pacific hake and to be either low (0.4) or high (0.8) for Pacific sardine, in order to compare the performance of data weighting methods under contrasting levels of recruitment variation. A higher level of recruitment variation caused a larger contrast in each year's age-composition observation. The two types of life history with differing recruitment assumptions were hereafter referred to as hake-low, sardine-low, and sardine-high. The OM included one fishery and the selectivity of which in age a and year t was specified to be a product of a parametric (logistic) form and a random deviation term away from the parametric form:

$$S_{a,t} = \frac{1}{1 + e^{-S_{slope}(a - S_{50})}} \times e^{\varepsilon_{a,t}} \quad (1)$$

Particularly, the non-parametric deviation term ($\varepsilon_{a,t}$), which can be treated as a process error in fishery selectivity, was specified to follow a two-dimensional AR(1) process:

$$\mathrm{vec}(\boldsymbol{\varepsilon}) \sim \mathrm{MVN}\left(\mathbf{0}, \sigma_S^2 \mathbf{R} \otimes \widetilde{\mathbf{R}}\right) \quad (2a)$$

$$\mathbf{R}_{a,\tilde{a}} = \rho_a^{|a - \tilde{a}|} \quad (2b)$$

$$\widetilde{\mathbf{R}}_{t,\tilde{t}} = \rho_t^{|t - \tilde{t}|} \quad (2c)$$

where σ_S (>0) is the standard deviation of selectivity deviations that controls the degree of variation in fishery selectivity and ρ_a (-1<ρ_a<1) and ρ_t (-1<ρ_t<1) are two AR(1) coefficients that control the degree to which selectivity deviations are autocorrelated in age and time, respectively. The deviations of fishery selectivity are identical and independent (IID) when ρ_a and ρ_t are both zeroes because this specification simplifies Eq. 2a to be

$$\varepsilon_{a,t} \sim \mathrm{N}(0, \sigma_S^2) \quad (3)$$

We explored four OMs with differing autocorrelation cases under a moderate level of selectivity variation:

1. OM1 ("*Independent*"). The deviations of fishery selectivity are independent ($\rho_a = 0; \rho_t = 0; \sigma_S = 0.4$);
2. OM2 ("*Time-correlated*"). The deviations of fishery selectivity are highly autocorrelated in time ($\rho_a = 0; \rho_t = 0.8; \sigma_S = 0.4$);

3. OM3 ("*Age-correlated*"). The deviations of fishery selectivity are highly autocorrelated in age ($\rho_a = 0.8; \rho_t = 0; \sigma_S = 0.4$);
4. OM4 ("*Age- and time-correlated*"). The deviations of fishery selectivity are highly autocorrelated in both age and time ($\rho_a = 0.8; \rho_t = 0.8; \sigma_S = 0.4$).

We used the *mvrnorm* function in the MASS R package (version 7.3-50, Venables and Ripley 2002) to simulate the autocorrelated process error in fishery selectivity. Estimating selectivity deviations is usually difficult for the youngest and oldest age groups due to a lack of adequate age-composition samples for those age groups, so in the OM we assumed that $\varepsilon_{a,t} = \varepsilon_{2,t}$ for $a < 2$ and $\varepsilon_{a,t} = \varepsilon_{7,t}$ for $a > 7$, namely, $\text{vec}(\boldsymbol{\varepsilon}) = (\varepsilon_{2,1}, \dots, \varepsilon_{2,T}, \varepsilon_{3,1}, \dots, \varepsilon_{3,T}, \dots \dots, \varepsilon_{7,1}, \dots, \varepsilon_{7,T})'$. Due to this assumption, the simulated time-varying selectivity cannot be dome-shaped (for $a > 7$, $\varepsilon_{a,T} \equiv \varepsilon_{7,T}$). The parametric selectivity profile as well as the associated variability (induced by the random deviation term) for Pacific hake and Pacific sardine were compared in Figure 1. For both species, we set the plus-group (A) and last simulation year (T) to be 15 and 20, respectively. Fishing mortality was simulated according to an effort-dynamics model (T1.6; more details in Thorson et al. (2013b)) that was used to generate contrast in spawning biomass (SB): the fishery (Fig. 2, left column) drove SB down to about 40% (see Table 2 for parameter values) of the unfished level over 20 years (Fig. 2, right column). A detailed description of how the life history parameters were derived for the two types of life histories (Pacific hake and Pacific sardine) also can be found in Thorson and Cope (2015).

2.1.2. Sampling model

The sampling model generated the following observation data from the true population dynamics specified by the OM:

- Fishery total catch in weight, which was assumed to be known without error.
- Fishery index of abundance (I), which was assumed to be log-normally distributed with a coefficient of variation of CV_{abund} and catchability of q : $\ln(I_t)\sim\mathrm{N}(\log(qB_t), \ln(1 + CV_{abund}^2))$, where $B_t = \sum_{a=0}^{A} N_{a,t} w_a S_{a,t}$ is the exploitable biomass in year t.
- Fishery age-composition data $(\mathbf{A})$, which was assumed to be drawn from a multinomial distribution with a sample size of n_{true} : $\boldsymbol{A}_t\sim\mathrm{Multinomial}(\boldsymbol{P}_t, n_{true})$, where $\boldsymbol{P}_t = \boldsymbol{C}_t / \sum_{a=0}^{A} C_{a,t}$ is the true age-composition proportion in year t.

We assumed that both the index of abundance and age-composition data were informative ($CV_{abund} = 0.1$ and $n_{true} = 200$) and were available every year during the simulation period. Therefore, model performance was not limited by low-quality data and should be primarily determined by how properly the fishery age-composition data ($\boldsymbol{A}_t$) were weighted.

2.1.3. Estimation model

The estimation model had the same population dynamics as the operating model except whether and how fishery selectivity varied over age and time. Three EMs with differing selectivity specifications were considered in each simulation experiment:

- EM1 ("*zero deviations*"). Selectivity of the fishery was specified to be constant by fixing $\hat{\sigma}_S$ as zero: $\hat{S}_a = \frac{1}{1+e^{-\hat{S}_{slope}(a-\hat{S}_{50})}}$. This specification for fishery selectivity is still a common practice in stock assessments.
- EM2 ("*IID deviations*"). Selectivity of the fishery was specified to be age- and time-varying and the deviations of which were specified to be identical and independent of age and time: $\hat{S}_{a,t} = \frac{1}{1+e^{-\hat{S}_{slope}(a-\hat{S}_{50})}} \times e^{\hat{\varepsilon}_{a,t}}$, where $\hat{\varepsilon}_{a,t}\sim\mathrm{N}(0, \hat{\sigma}_S^2)$.

- EM3 ("*AR deviations*"). Selectivity of the fishery was specified to be age- and time-varying and the deviations of which were specified to be autocorrelated: $\hat{S}_{a,t} = \frac{1}{1+e^{-\hat{S}_{slope}(a-\hat{S}_{50})}} \times e^{\hat{\varepsilon}_{a,t}}$, where $\mathrm{vec}(\hat{\boldsymbol{\varepsilon}}) \sim \mathrm{MVN}\left(\mathbf{0}, \hat{\sigma}_S^2 \mathbf{R} \otimes \widetilde{\mathbf{R}}\right)$, $\mathbf{R}_{a,\tilde{a}} = \hat{\rho}_a^{|a-\tilde{a}|}$, and $\widetilde{\mathbf{R}}_{t,\tilde{t}} = \hat{\rho}_t^{|t-\tilde{t}|}$.

This study was focused on two data weighting issues in stock assessments: sensitivity of model performance to data weighting and which data weighting method performs better when estimating age- and time-varying selectivity. In some simulation experiments, we assumed that the hyper-parameters for selectivity deviations ($\hat{\sigma}_S$, $\hat{\rho}_a$, and $\hat{\rho}_t$) were known without error; in other simulation experiments, we estimated these hyper-parameters. We included both simulation experiments to determine model performance either in an idealized case (when these hyper-parameters are known) or in a more realistic case (when they must be estimated).

The age-composition data from the fishery were assumed to be drawn from a multinomial distribution with an estimated effective sample size of n_{eff}. It specified the extent to which the fishery age-composition data were weighted:

$$\boldsymbol{A}_t \sim \mathrm{Multinomial}\left(\widehat{\boldsymbol{P}}_t, n_{eff}\right) \quad (4)$$

where $\widehat{\boldsymbol{P}}_t = \widehat{\boldsymbol{C}}_t / \sum_{a=0}^{A} \hat{C}_{a,t}$ is the expected age-composition proportion in year t.

Unless otherwise noted, the three EMs were correctly specified (fixed at the true values) for all model parameters except unfished recruitment (R_0), annual recruitment (R_t), parametric selectivity (S_{slope} and S_{50}), selectivity deviations ($\varepsilon_{\mathrm{a,t}}$ in EMs *IID deviations* and *AR deviations*), and annual fully-selected fishing mortality (F_t). Among those estimated parameters, R_0, S_{slope}, S_{50} and F_t were estimated as fixed effects, and $\varepsilon_{\mathrm{a,t}}$ and R_t were estimated as random effects. We used Template Model Builder (TMB, Kristensen et al. (2016)) to implement mixed-effect parameter estimation.

In TMB, the marginal likelihood of fixed effect parameters was calculated using the Laplace approximation to integrate across random effects (Kristensen et al. 2016) and fixed effect parameters were then estimated via maximizing the marginal likelihood within the R (version 3.4.0) computing environment (R Core Team 2017). We used the *nlminb* function to minimize the negative of the marginal log-likelihood and confirmed model convergence based on the convergence flag the function provided and a positive-definite Hessian.

2.2. Simulation experiments

In this study, we conducted four related simulation experiments. A summary of the factorial design of the OM and EM in each experiment can be found in Table 3.

2.2.1. What is the impact of under, right, or over-weighting on model performance?

The first simulation experiment aimed to evaluate the sensitivity of estimation performance of the three EMs to data weighting, given that the true fishery selectivity had independent or autocorrelated deviations. We compared the performance of each EM in estimating SB under three data weighting scenarios: 1) under-weighting age-composition data by a factor of 10, which was realized by setting $n_{input} = 0.1 \times n_{true} = 20$ in the three EMs; 2) right-weighting age-composition data, which was realized by setting $n_{input} = n_{true} = 200$ in the three EMs; and 3) over-weighting age-composition data by a factor of 10, which was realized by setting $n_{input} = 10 \times n_{true} = 2000$ in the three EMs. In this simulation experiment, four hundred simulation replicates with unique process errors (in recruitment and selectivity) and observation errors (in abundance index and age-composition observations) were generated for every combination of population dynamics and OM case. Each EM (*zero deviations*, *IID deviations*, or *AR deviations*) was then fitted to every generated simulation replicate individually under three data weighting scenarios (under, right, or over-weighting). We evaluated the estimation performance of the three EMs based on the mean absolute relative error (MARE) in the estimate of final year SB: $\text{mean}(|\widehat{SB}_{t=20}/SB_{t=20} - 1|)$. This metric took both accuracy and precision into consideration.

2.2.2. How well can we estimate effective sample size given time-varying selectivity

The second simulation experiment aimed to compare three widely-used data weighting methods in stock assessments:

- McAllister-Ianelli (M-I) method (McAllister and Ianelli 1997). The effective sample size for the multinomial distribution was iteratively estimated through a tuning algorithm. In this study, it was computed as the harmonic mean of annual effective sample sizes

$$n_{eff} = \frac{T}{\sum_{t=1}^{T}\left(\frac{1}{n_{eff_t}}\right)} \quad (5a)$$

$$n_{eff_t} = \frac{\sum_{a=0}^{A}\left(\hat{P}_{a,t}\left(1 - \hat{P}_{a,t}\right)\right)}{\sum_{a=0}^{A}\left(P_{a,t} - \hat{P}_{a,t}\right)^2} \quad (5b)$$

  and iteratively tuned until its relative difference between two iterations was less than 5%.

- Francis method (Francis 2011). The effective sample size for the multinomial distribution was also iteratively estimated through a tuning algorithm. Specifically, it is the inverse of the variance of normalized differences between the observed (P'_t) and expected ($\hat{P}'_t$) mean ages in age-composition

$$n_{eff} = \frac{1}{\text{Var}\left(\frac{P'_t - \hat{P}'_t}{\sqrt{v_t}}\right)} \quad (6a)$$

$$P'_t = \sum_{a=0}^{A} \left(aP_{a,t}\right) \quad (6b)$$

$$\hat{P}'_t = \sum_{a=0}^{A} \left(a\hat{P}_{a,t}\right) \quad (6c)$$

$$v_t = \sum_{a=0}^{A} \left(a^2\hat{P}_{a,t}\right) - {\hat{P}'_t}{}^2 \quad (6d)$$

and iteratively tuned until its relative difference between two iterations was less than 5%.

- Dirichlet-multinomial (D-M) method (Thorson et al. 2017). Different from the two tuning methods above, the D-M method estimated the effective sample size based on maximum likelihood. By assuming that age-composition data followed the linear parameterization of the Dirichlet-multinomial distribution, the effective sample size of the age-composition data was computed as

$$n_{eff} = \frac{1 + \theta n_{input}}{1 + \theta} \quad (7)$$

where n_{input} was the input sample size of the age-composition data. The D-M method estimated the effective sample size by fixing age-composition data ($n_{input}P_{a,t}$) and instead estimating θ as a parameter. The likelihood associated with the age-composition data was

$$L_{comp} \propto \prod_{t=1}^{T} \left(\frac{\Gamma\left(\theta n_{input}\right)}{\Gamma\left(n_{input} + \theta n_{input}\right)} \prod_{a=0}^{A} \frac{\Gamma\left(n_{input}P_{a,t} + \theta n_{input}\hat{P}_{a,t}\right)}{\Gamma\left(\theta n_{input}\hat{P}_{a,t}\right)} \right) \quad (8)$$

The three data weighting methods were compared based on two metrics: the ratio of estimated effective sample size to true sample size (n_{eff}/n_{true}) and the MARE in the estimate of final year SB. In this simulation experiment, four hundred simulation replicates with unique process errors and observation errors were generated for every combination of population dynamics and OM case.

OM case *Independent* approximately matched the simulation scenario explored in Thorson et al. (2017), but other OM cases represented the first attempt to explore the sensitivity of the D-M method to model mis-specification.

2.2.3. How does over-dispersion affect D-M estimates given time-varying selectivity?

In the third simulation experiment, we evaluated the performance of the three data weighting methods in estimating the effective sample size of over-dispersed age-composition data. Over-dispersed age-composition data were simulated by assuming that the extent of over-dispersion is constant across age and time:

$$\tilde{A}_{a,t} = A_{a,t} \times d \quad (9)$$

where d (>1) denotes the extent of over-dispersion in age-composition data. For the two tuning methods, the estimated effective sample size was a function of age-composition proportion ($P_{a,t}$), which, under this assumption, did not change with the extent of over-dispersion in age-composition data ($P_{a,t} = \tilde{A}_{a,t} / \sum_{a=0}^{A} \tilde{A}_{a,t} = A_{a,t} \times d / \sum_{a=0}^{A} (A_{a,t} \times d) = A_{a,t} / \sum_{a=0}^{A} A_{a,t}$). Thus, the estimated effective sample size based on either tuning method should not be affected by the over-dispersion in age-composition data. The focus of this simulation experiment was indeed on the D-M method, for which n_{input} was specified to be the actual sample size (number of fish sampled; $n_{true} \times d$) of the over-dispersed age-composition data. Eq. 9 simulated a type of over-dispersion case that all fish were caught in groups of d individuals with identical age. By this definition, the true sample size would be n_{true}.

We computed the ratio of estimated effective sample size to true sample size (n_{eff}/n_{true}) for evaluating the performance of the D-M method with respect to estimating the effective sample size, given that the age-composition data are over-dispersed. Due to the high computation demand in this simulation experiment, we generated one hundred simulation replicates with unique process errors and observation errors for every combination of population dynamics and OM case.

2.2.4. Can we estimate time-varying selectivity penalty and composition weighting simultaneously?

Lastly, we considered a more realistic situation where the degree of variation in selectivity was estimated rather than known without error. In this simulation experiment, the degree of variation in selectivity was iteratively estimated using a tuning algorithm inspired by Methot and Taylor (2011) and introduced by Xu et al. (2018)

$$\hat{\sigma}_S^2 = \mathrm{SD}(\hat{\boldsymbol{\varepsilon}})^2 + \frac{1}{6T}\sum_{a=2}^{7}\sum_{t=1}^{T}\mathrm{SE}\left(\hat{\varepsilon}_{a,t}\right)^2 \quad (10)$$

To replicate the case of Stock Synthesis (Methot and Wetzel 2013) and other widely-used penalized likelihood models, here $\hat{\sigma}_S$ was estimated via the tuning approach instead of the mixed-effect approach (i.e., EM3 instead of EM4 in Xu et al. 2018). Xu et al. (2018) showed that this tuning algorithm could accurately estimate $\hat{\sigma}_S$ when the effective sample size was known without error. In real-world assessments, however, both n_{eff} and $\hat{\sigma}_S$ are unknown and need to be estimated.

The focus of this simulation experiment was on the combined performance of the D-M method for estimating n_{eff} and the tuning method for estimating $\hat{\sigma}_S$. How the effective sample size and selectivity hyper-parameters were simultaneously estimated in this simulation are described below:

- Step 1: Tune selectivity variability ($\hat{\sigma}_S$) and effective sample size (n_{eff}). $\hat{\sigma}_S$ was iteratively tuned in EM *IID deviations* until matching Eq. 10 within an accuracy of 0.01 while the D-M method was used in every iteration of $\hat{\sigma}_S$ to estimate n_{eff}. $\hat{\sigma}_S$ was then fixed in EM *IID deviation* and the estimated selectivity deviations were extracted.

- Step 2: Estimate selectivity autocorrelations ($\hat{\rho}_a$ and $\hat{\rho}_t$). $\hat{\rho}_a$ and $\hat{\rho}_t$ were estimated based on an "external" estimation method (for more details see the description of EM *AR deviations* in Xu et al. 2018). In brief, the two autocorrelation coefficients were estimated using the maximum likelihood approach by fitting an external stand-alone model to selectivity deviations that EM *AR deviations* estimated in step 1. The external stand-alone model estimated $\hat{\rho}_a$ and $\hat{\rho}_t$ by assuming that selectivity deviations follow the multivariate normal distribution described in Eq. 2 and that both $\hat{\rho}_a$ and $\hat{\rho}_t$ are between 0 and 1 (realized by using the logit transformation).

The combined performance was evaluated according to the ratios of estimated to true values of both n_{eff} and $\hat{\sigma}_S$. In addition, we compared the two estimated autocorrelation coefficients with the corresponding true values to evaluate the performance of the "external" estimation method for selectivity autocorrelations. Due to the high computation demand in this simulation experiment, we generated one hundred simulation replicates with unique process errors and observation errors for every combination of species and OM case.

3. Results

3.1. What is the impact of under, right, or over-weighting on model performance?

Overall, over-weighting age-composition data generally performed worse than under-weighting age-composition data to the same extent. Results of the first simulation showed that over-weighting tended to cause a larger estimation error in the final year SB (Fig. 3) in comparison to under-weighting. Results also showed that over-weighting consistently corresponded to significantly worse estimation performance than under-weighting for EM *AR deviations*, the EM with correctly-specified selectivity (Fig. 3; see Fig. A1 for resampled uncertainty levels).

Whether under-weighting or over-weighting age-composition data performed better varied somewhat among species and OMs. Under OM *Independent*, right-weighting and over-weighting age-composition data performed best and worse, respectively, regardless of how selectivity was specified in the EM (Fig. 3, first column). Over-weighting also performed worse under OM *Time-correlated*, except for EM *IID deviations* for Pacific hake, which performed worst when age-composition data were down-weighted (Fig. 3, second column). It is worth noting that under-weighting could produce the best performing EM (i.e., EM *IID deviations*) in this case. Under OM *Age-correlated*, over-weighting performed better and worse than under-weighting when the variation in selectivity was ignored (EM *zero deviations*) and estimated (EMs *IID deviations* and *AR deviations*), respectively (Fig. 3, third column). Again, right-weighting generally performed best regardless of how selectivity was specified in the EM. Under OM *Age- and time-correlated*, the three data weighting scenarios performed similarly for EMs *zero deviations* and *IID deviations*, at least within the weighting range (0.1×-10×) investigated in this study (Fig. 3, fourth column). For EM *AR deviations*, over-weighting and right-weighting consistently performed worst and best, respectively.

3.2. How well can we estimate the effective sample size given time-varying selectivity?

Among the three data weighting methods (M-I, Francis, and D-M methods), the D-M method provided the most accurate estimated effective sample size regardless of whether the EM allowed for time-varying selectivity. For EM *zero deviations* which mistakenly specified constant selectivity, all three data weighting methods estimated a reduced effective sample size (medians within 0.2×-0.7× of the true sample size) (Fig. 4). This behavior was expected given that these models are mis-specified. For EMs *IID deviations* and *AR deviations*, the effective sample size that the D-M method provided was very accurate while those the M-I method and especially the Francis method estimated were considerably larger than the true sample size (Fig. 4). From a median point of view, the M-I method over-estimated the effective sample size by as much as 4× and 13× for Pacific hake and Pacific sardine, respectively; the Francis method over-estimated the effective sample size by as much as 8× and 20× for Pacific hake and Pacific sardine, respectively. It is worth noting that, by definition (Eq. 7), the effective sample size estimated by the D-M method must be smaller than or approximately the same as (when θ is estimated to be very large) the input sample size, which in this simulation experiment is identical to the true sample size. Therefore, the effective sample sizes estimated by the D-M method were all smaller than the true sample size in this simulation experiment, regardless of whether the EM accounts for time-varying selectivity or not.

Because the D-M method provided the most accurate estimated effective sample size when allowing for time-varying selectivity, both EMs *IID deviations* and *AR deviations* performed best when using the D-M method for data weighting. The first simulation informed us that estimation performance was relatively insensitive to data weighting when age-composition data were under-weighted (Fig. 4). Since all three data weighting methods under-estimated the effective sample size under constant selectivity, the method for data weighting had little effect on the estimation performance of EM *zero deviations* (Fig. 5). For EMs *IID deviations and AR deviations*, in comparison, estimation performance could be significantly affected by the method on which data weighting was based: the D-M and Francis method generally corresponded to smallest and largest errors in the estimate of final year SB, respectively (Fig. 5). That was because, when allowing for time-varying selectivity, the M-I method and especially the Francis method considerably over-estimated the effective sample size, which was, in contrast, slightly under-estimated by the D-M method (Fig. 4). Also, the extent to which the two tuning methods over-estimated the effective sample size when estimating time-varying selectivity was larger for Pacific sardine than Pacific hake. Consequently, the benefit of the D-M method in terms of improving the estimate of final year SB was significant for Pacific sardine, but not for Pacific hake (Fig. 5; see Fig. A2 for resampled uncertainty levels).

Some cases in this simulation suggested that estimating time-varying selectivity (EMs *IID deviations* and *AR deviations*) resulted in worse performance than assuming time-invariant selectivity (EM *zero deviations*) when using either the Francis or M-I method for data weighting (Fig. 5). This is surprising, given that the true selectivity in the OM was simulated to have a moderate level of variation over both age and time (see Eq. 1 and Fig. 1). This pattern indicated that improving stock assessments by accounting for time-varying selectivity requires simultaneously addressing data weighting and the time-varying process.

3.3. How does over-dispersion affect D-M estimates given time-varying selectivity?

The effective sample size of over-dispersed age-composition data was under-estimated by the D-M method for all three EMs, but the degree of under-estimation in all cases remained in a reasonable range (medians larger than 0.2× of the true sample size) (Fig. 6). As expected, the effective sample size was estimated to be considerably below the input sample size (medians within 0.2×-0.6× of the true sample size) for EM *zero deviations*. This result is expected given that this EM is mis-specified. When allowing for time-varying selectivity (EMs *IID deviations and AR deviations*), the median estimated effective sample size that the D-M method provided was above 0.5× of the true sample size under all degrees of over-dispersion (Fig. 6). The bias in the estimated effective sample size became slightly greater as the degree of over-dispersion increased from 2 to

10, which was in accordance with the trend found in a previous study (see Fig. 4 in Thorson et al. (2017). We also noted that although the difference was not dramatic, the D-M method generally performed better for Pacific hake simulations than Pacific sardine simulations.

3.4. Can we estimate time-varying selectivity penalty and composition weighting simultaneously?

Like in the previous simulation experiment where the level of variation in selectivity was assumed known without error, the median estimated effective sample size that the D-M method provided was still above 0.5× of the true sample size (Fig. 7). Moreover, the bias in the estimated effective sample size was still positively related to the degree of over-dispersion in age-composition data. The first simulation suggested that the performance of models that estimated time-varying selectivity was not sensitive to down-weighting age-composition data. As such, it was not surprising to find that MARE was negligibly impacted by the under-estimation of the effective sample size (Fig. A3) within this range of degree of under-estimation (by 10%-50%) (Fig. 7). Namely, model performance was not sensitive to the degree of over-dispersion in age-composition data when using the D-M method for data weighting. Importantly, in combination with the D-M method for data weighting, the tuning method that was developed by Xu et al. (2018) was useful for estimating the level of variation in selectivity. The level of variation in selectivity was only slightly under-estimated (medians within 0.75×-1× of the true level), within which the largest degree of under-estimation occurred under OM *Age- and time-correlated*.

The "external" estimation method for the two autocorrelation coefficients for selectivity deviations ($\hat{\rho}_a$ and $\hat{\rho}_t$) were also useful (Fig. 8). Under OM *Independent* where the true coefficients were both 0, the median estimates that the "external" method provided were between 0 and 0.2. Under OMs *Age-dependent* and *Time-dependent*, where one of the two true coefficients was positive (0.8), the median estimate of that coefficient was only slighted under-estimated (median larger than 0.6). Under OM *Age- and time-dependent* where the true coefficients were both positive (0.8), the median estimates of the two coefficients were mostly above 0.4 In accordance with the finding in Xu et al. (2018), this "external" estimation method generally performed better for $\hat{\rho}_t$ than $\hat{\rho}_a$ (Fig. 8).

4. Discussion

This study aimed to compare the performance of widely-used data weighting methods in the assessment models that allow for time-varying selectivity. We conducted four simulation experiments to evaluate the sensitivity of model estimates to data weighting, and more importantly, to evaluate the performance of M-I, Francis, and D-M methods given time-varying selectivity. For assessment models that estimated time-varying selectivity, over-weighting generally led to larger estimation error in final year SB than did under-weighting to the same extent, suggesting that over-weighting should be avoided even when allowing for time-varying selectivity. Among the three data weighting methods compared in this study, the D-M method out-performed the other two methods when estimating time-varying selectivity. Moreover, the D-M method was still useful even when age-composition data were over-dispersed and the level of variation in selectivity was simultaneously estimated. In conclusion, the D-M method was recommended over the M-I and Francis methods for the assessments that explore time-varying selectivity.

Overall, over-weighting composition data tends to cause larger estimation error in final year SB than does under-weighting composition data to the same extent. The result echoes Francis' (2011) suggestion that age-composition data should not be over-weighted. This suggestion was made based on the idea that while composition data are important to inform selectivity and recruitment variation, the estimated population trend should be primarily driven by the more reliable abundance indices, especially when data conflict exists between abundance indices and composition data. Importantly, this study shows that estimation performance is more degraded by over-weighting than under-weighting regardless of whether selectivity is mis-specified or not. In some cases, data weighting had a larger impact on the estimation performance of assessment models with correctly-specified selectivity than those without, implying that data weighting is important for stock assessments with any selectivity specifications.

For assessment models that estimate time-varying selectivity, the D-M method overall performs better than the M-I and Francis methods with regards to weighting age-composition data. Under the specification of time-varying selectivity, the M-I method and especially the Francis method over-estimate the effective sample size greatly and consequently correspond to large estimation error in final year SB. The fact that the effective sample size is greatly over-estimated by the two tuning methods is likely due to the expected and observed age-compositions tend to match more closely under a more flexible (i.e., time-varying) selectivity specification (Francis 2017, Punt et al. 2014). In contrast, the D-M method under-estimates the effective sample size slightly and consequently corresponds to smaller estimation error in final year SB. When age-composition data are over-dispersed, simulation results show that the D-M method under-estimates the effective sample size to a certain extent. However, the extent of the under-estimation is still smaller in comparison with the extent to which the two tuning methods over-estimate the effective sample size of randomly-sampled age-composition data. Furthermore, assessment models that use the D-M data weighting method consider the uncertainty about data weighting. the D-M method estimates effective sample size as a parameter of the assessment model based on maximum likelihood, so it can propagate the uncertainty about data weighting into the confidence interval of estimated population and management attributes (Thorson et al. 2017). In contrast, both the two tuning methods ignore the uncertainty in estimated effective sample size, leading to under-estimated uncertainty in model estimates (Maunder 2011). However, it should be noted that when using the D-M method for data weighting, the confidence interval of estimated population and management attributes could be over-estimated as the simulations in this study show that the D-M method tends to under-estimate the effective sample size.

For assessment models that estimate time-varying selectivity, the D-M data weighting method is robust for over-dispersed age-composition data, which is a common phenomenon in fisheries. Within the range of over-dispersion we investigated (2×-10×), the D-M method under-estimated the effective sample size by less than 50%, regardless of the OM case and type of life history. The first simulation experiment showed that under-weighting age-composition data by such an extent should only minorly degrade the estimation performance of the assessment model. It should be noted that the D-M method tends to down-weight age-composition data to a much larger extent when selectivity is specified to be constant than time-varying, implying that data weighting based on the D-M method is informed by the goodness-of-fit of age-composition data in the assessment model (Thorson et al. 2017). In addition to the linear parameterization used in the paper, the Dirichlet-multinomial can also be parameterized in another way (i.e., the saturation parameterization). In the simulations in the paper, the input sample size is specified to be identical among years, in which case the two parameterizations result in identical parameter estimates. Future research could compare the two parameterizations when input sample size varies among years.

We are aware that the comparison of the three data weighting methods in the second simulation experiment is tilted towards the D-M method. By construction, the D-M method only allows the effective sample size to be smaller than the input sample size (which is specified to be the true sample size in that simulation), leading to a more restricted parameter space for the effective sample size and a larger probability for the effective sample size to be close to the true sample size. To make a fairer comparison between the three data-weighting methods, we then conducted the third simulation experiment in which the effective sample size can be as large as 10x of the true sample size. Being consistent with the pattern found in Thorson et al. (2017), the performance of the D-M method is negatively related to the ratio of input sample size to true sample size. However, even when the input sample size for the D-M method is specified to be 10x of the true sample size, the D-M method still performs better than the two tuning methods. We recommend ongoing research to accurately estimate the input sample size from field-measurements of age- and length-composition (Stewart and Hamel 2014, Thorson 2014, Thorson and Haltuch 2018) because an accurate starting point for weighting compositional data improves model performance when using the D-M data weighting method.

In terms of estimation performance of an assessment model, correctly specifying the distributional penalty for selectivity deviations is as important as choosing a proper method (i.e., the D-M method) for data weighting. When using the D-M data weighting method, correctly specifies selectivity (EM *AR deviations*) greatly outperformed the other two EMs with mis-specified selectivity (EMs *zero deviations* and *IID deviations*). Several studies have suggested considering data weighting and time-varying selectivity together in stock assessments (Francis 2011, Stewart and Monnahan 2017, Thorson et al. 2017, Wang and Maunder 2017). Results from this study provide another strong support for this suggestion.

According to our simulation study, the D-M method for weighting composition data and the tuning method for penalizing selectivity variation (Xu et al. 2018) are able to provide proper data weighting and selectivity penalizing simultaneously. In real-world stock assessments, both the level of variation in selectivity and the level of over-dispersion in composition data are unknown and need to be estimated. Considering that both methods have been implemented in Stock Synthesis (Methot and Wetzel 2013), a widely used stock assessment package, we recommend users to explore the two methods together in real-world stock assessments. When exploring the two methods simultaneously in stock assessments, we also recommend evaluating the autocorrelations in selectivity deviations using the "external" estimation method, which performs reasonably well in Xu et al. (2018) as well as in this study. It can improve stock assessments if process errors in selectivity are highly autocorrelated.

We note that the performance of the D-M method is likely over-estimated in our idealized simulations. First, our assumption about selectivity deviations in the OM allows the simulated time-varying selectivity to be asymptotic only. However, real fishery selectivity can be dome-shaped (Sampson and Scott 2011, Waterhouse et al. 2014) and the multinomial distribution was found to perform poorly in simulations where selectivity is assumed to be dome-shaped. Second, we only evaluated the impacts of mis-specifying selectivity on the performance of the D-M method in this study. Other biological processes including natural mortality, growth, and maturity were all assumed known without error. In real-world stock assessments, however, these biological processes are likely to vary in complicated ways, such that assessment models are likely misspecified in multiple ways simultaneously. In other words, these biological processes are more or less mis-specified in real-world stock assessments, leading to larger discrepancies between observed and predicted age-composition. Considering that the D-M data weighting method already under-estimates the effective sample size in this study, it may under-estimate the effective sample size to a larger extent in real-world stock assessments.

Third, there is another obvious limitation of this simulation study that can cause over-estimating the performance of the D-M method. In this simulation study, the D-M method was used to weight the age-composition data sampled using a closely-related distribution (i.e., multinomial). Studies (Berg and Nielsen 2016, Berg et al. 2014) have shown that sampling errors in real fishery age-compositions can be positively correlated among ages. The multinomial distribution, however, only allows negative correlations among ages and therefore may not be appropriate for the sampling model that generates age-composition samples (Albertsen et al. 2017, Francis 2014). Indeed, the D-M method may perform worse for length-composition data because the positive correlations among lengths tend to be higher than those among ages. As such, the performance of the D-M method needs to be more closely evaluated in future studies using real fishery age-composition data.

Acknowledgments

We thank Mark Maunder, Jim Ianelli, the associated editor, and two anonymous reviewers for providing valuable inputs to an earlier draft. We also thank Kasper Kristensen for developing Template Model Builder. Haikun Xu was partially funded by NOAA Award NA15OAR4320063 through the University of Washington Joint Institute for the Study of the Atmosphere and Ocean (JISAO).

References

Albertsen, C.M., Nielsen, A., and Thygesen, U.H. 2017. Choosing the observational likelihood in state-space stock assessment models. Canadian Journal of Fisheries and Aquatic Sciences **74**(5): 779-789.

Berg, C.W., and Nielsen, A. 2016. Accounting for correlated observations in an age-based state-space stock assessment model. ICES Journal of Marine Science: Journal du Conseil: fsw046.

Berg, C.W., Nielsen, A., and Kristensen, K. 2014. Evaluation of alternative age-based methods for estimating relative abundance from survey data in relation to assessment models. Fisheries research **151**: 91-99.

Francis, R.I.C.C. 2011. Data weighting in statistical fisheries stock assessment models. Canadian Journal of Fisheries and Aquatic Sciences **68**(6): 1124-1138.

Francis, R.I.C.C. 2014. Replacing the multinomial in stock assessment models: A first step. Fisheries Research **151**: 70-84.

Francis, R.I.C.C. 2017. Revisiting data weighting in fisheries stock assessment models. Fisheries Research **192**: 5-15.

Hulson, P.-J.F., Hanselman, D.H., and Quinn, T.J. 2011. Determining effective sample size in integrated age-structured assessment models. ICES Journal of Marine Science **69**(2): 281-292.

Kotwicki, S., Horne, J.K., Punt, A.E., and Ianelli, J.N. 2015. Factors affecting the availability of walleye pollock to acoustic and bottom trawl survey gear. ICES Journal of Marine Science **72**(5): 1425-1439.

Kristensen, K., Nielsen, A., Berg, C., Skaug, H., and Bell, B. 2016. Template model builder TMB. J. Stat. Softw **70**: 1-21.

Lowe, S.A., Ianelli, J.N., and Palsson, W. 2017. Assessment of the Atka mackerel stock in the Bering Sea and Aleutian Islands [online]. Available from www.afsc.noaa.gov/REFM/Docs/2012/BSAIatka.pdf.

Martell, S., and Stewart, I. 2014. Towards defining good practices for modeling time-varying selectivity. Fisheries Research **158**: 84-95.

Maunder, M.N. 2011. Review and evaluation of likelihood functions for composition data in stock-assessment models: Estimating the effective sample size. Fisheries Research **109**(2): 311-319.

Maunder, M.N., Crone, P.R., Punt, A.E., Valero, J.L., and Semmens, B.X.J.F.R. 2017. Data conflict and weighting, likelihood functions and process error. (192): 1-4.

Maunder, M.N., and Piner, K.R. 2017. Dealing with data conflicts in statistical inference of population assessment models that integrate information from multiple diverse data sets. Fisheries Research **192**: 16-27.

Maunder, M.N., and Punt, A.E. 2013. A review of integrated analysis in fisheries stock assessment. Fisheries Research **142**: 61-74.

McAllister, M.K., and Ianelli, J.N. 1997. Bayesian stock assessment using catch-age data and the sampling-importance resampling algorithm. Canadian Journal of Fisheries and Aquatic Sciences **54**(2): 284-300.

Methot, R.D. 2009. Stock assessment: operational models in support of fisheries management. *In* The future of fisheries science in North America. Springer. pp. 137-165.

Methot, R.D., and Taylor, I.G. 2011. Adjusting for bias due to variability of estimated recruitments in fishery assessment models. Canadian Journal of Fisheries and Aquatic Sciences **68**(10): 1744-1760.

Methot, R.D., and Wetzel, C.R. 2013. Stock synthesis: a biological and statistical framework for fish stock assessment and fishery management. Fisheries Research **142**: 86-99.

Minte-Vera, C.V., Maunder, M.N., Aires-da-Silva, A.M., Satoh, K., and Uosaki, K. 2017. Get the biology right, or use size-composition data at your own risk. Fisheries Research **192**: 114-125.

Punt, A.E., Hurtado-Ferro, F., and Whitten, A.R. 2014. Model selection for selectivity in fisheries stock assessments. Fisheries Research **158**: 124-134.

R Core Team. 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Sampson, D.B., and Scott, R.D. 2011. A spatial model for fishery age-selection at the population level. Canadian Journal of Fisheries and Aquatic Sciences **68**(6): 1077-1086.

Sampson, D.B., and Scott, R.D. 2012. An exploration of the shapes and stability of population–selection curves. Fish and Fisheries **13**(1): 89-104.

Stewart, I.J., and Hamel, O.S. 2014. Bootstrapping of sample sizes for length-or age-composition data used in stock assessments. Canadian journal of fisheries and aquatic sciences **71**(4): 581-588.

Stewart, I.J., and Martell, S.J. 2014. A historical review of selectivity approaches and retrospective patterns in the Pacific halibut stock assessment. Fisheries Research **158**: 40-49.

Stewart, I.J., and Monnahan, C.C. 2017. Implications of process error in selectivity for approaches to weighting compositional data in fisheries stock assessments. Fisheries Research **192**: 126-134.

Thorson, J.T. 2014. Standardizing compositional data for stock assessment. ICES Journal of Marine Science **71**(5): 1117-1128.

Thorson, J.T. 2018. Perspective: Let's simplify stock assessment by replacing tuning algorithms with statistics.

Thorson, J.T., Clarke, M.E., Steward, I.J., and Punt, A. 2013a. The implications of spatially varying catchability on bottom trawl surveys of fish abundance: a proposed solution involving underwater vehicles. **70**(2): 294-306.

Thorson, J.T., and Cope, J.M. 2015. Catch curve stock-reduction analysis: An alternative solution to the catch equations. Fisheries Research **171**: 33-41.

Thorson, J.T., and Haltuch, M.A. 2018. Spatiotemporal analysis of compositional data: increased precision and improved workflow using model-based inputs to stock assessment. Canadian Journal of Fisheries and Aquatic Sciences(999): 1-14.

Thorson, J.T., Johnson, K.F., Methot, R.D., and Taylor, I.G. 2017. Model-based estimates of effective sample size in stock assessment models using the Dirichlet-multinomial distribution. Fisheries Research **192**: 84-93.

Thorson, J.T., Minto, C., Minte-Vera, C.V., Kleisner, K.M., and Longo, C. 2013b. A new role for effort dynamics in the theory of harvested populations and data-poor stock assessment. Canadian Journal of Fisheries and Aquatic Sciences **70**(12): 1829-1844.

Thorson, J.T., and Taylor, I.G. 2014. A comparison of parametric, semi-parametric, and non-parametric approaches to selectivity in age-structured assessment models. Fisheries Research **158**: 74-83.

Venables, W., and Ripley, B. 2002. Modern applied statistics (Fourth S., editor) New York. Springer.

Wang, S.-P., and Maunder, M.N. 2017. Is down-weighting composition data adequate for dealing with model misspecification, or do we need to fix the model? Fisheries Research **192**: 41-51.

Waterhouse, L., Sampson, D.B., Maunder, M., and Semmens, B.X. 2014. Using areas-as-fleets selectivity to model spatial fishing: asymptotic curves are unlikely under equilibrium conditions. Fisheries research **158**: 15-25.

Xu, H., Thorson, J.T., Methot, R.D., and Taylor, I.G. 2018. A new semi-parametric method for autocorrelated age-and time-varying selectivity in age-structured assessment models. Canadian Journal of Fisheries and Aquatic Sciences(999): 1-18.

Figures and Tables

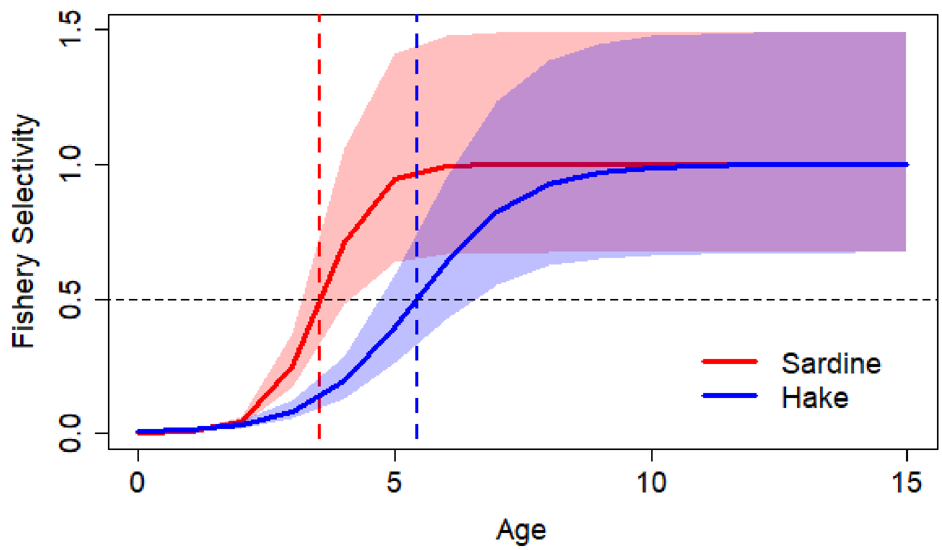

Figure 1. Comparison of the parametric fishery selectivity for the two types of life history (Pacific hake and Pacific sardine) as a function of age. The shaded areas show the ±1 standard deviation range of selectivity variation. The vertical dashed lines mark the age at 50% selection of the fishery.

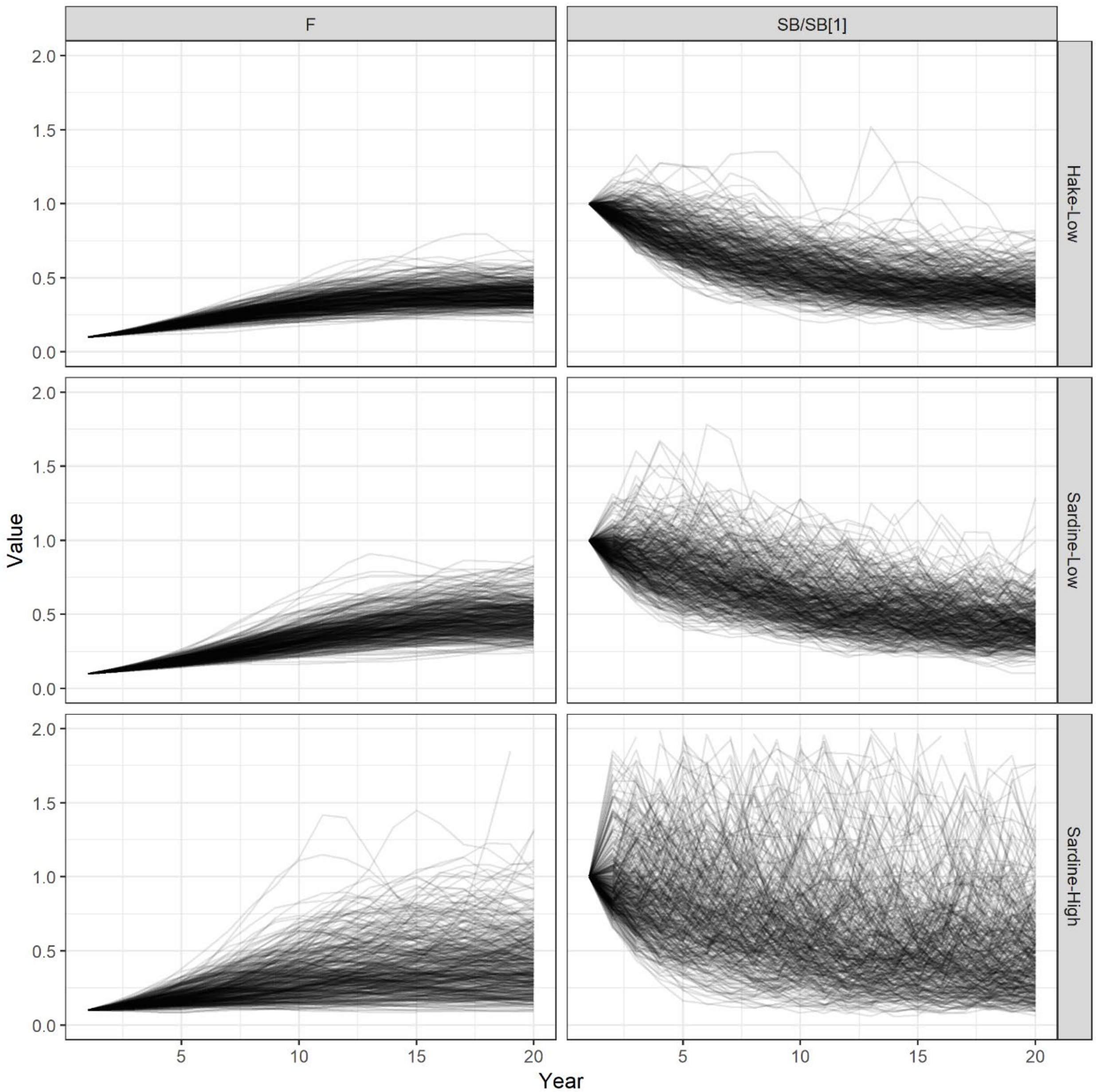

Figure 2. 1st simulation experiment: trajectories of fully-selected fishing mortality (left) and spawning biomass (right) for the four hundred replicates. To facilitate the comparison among replicates, spawning biomass is rescaled to have an initial ($t = 1$) value of 1.

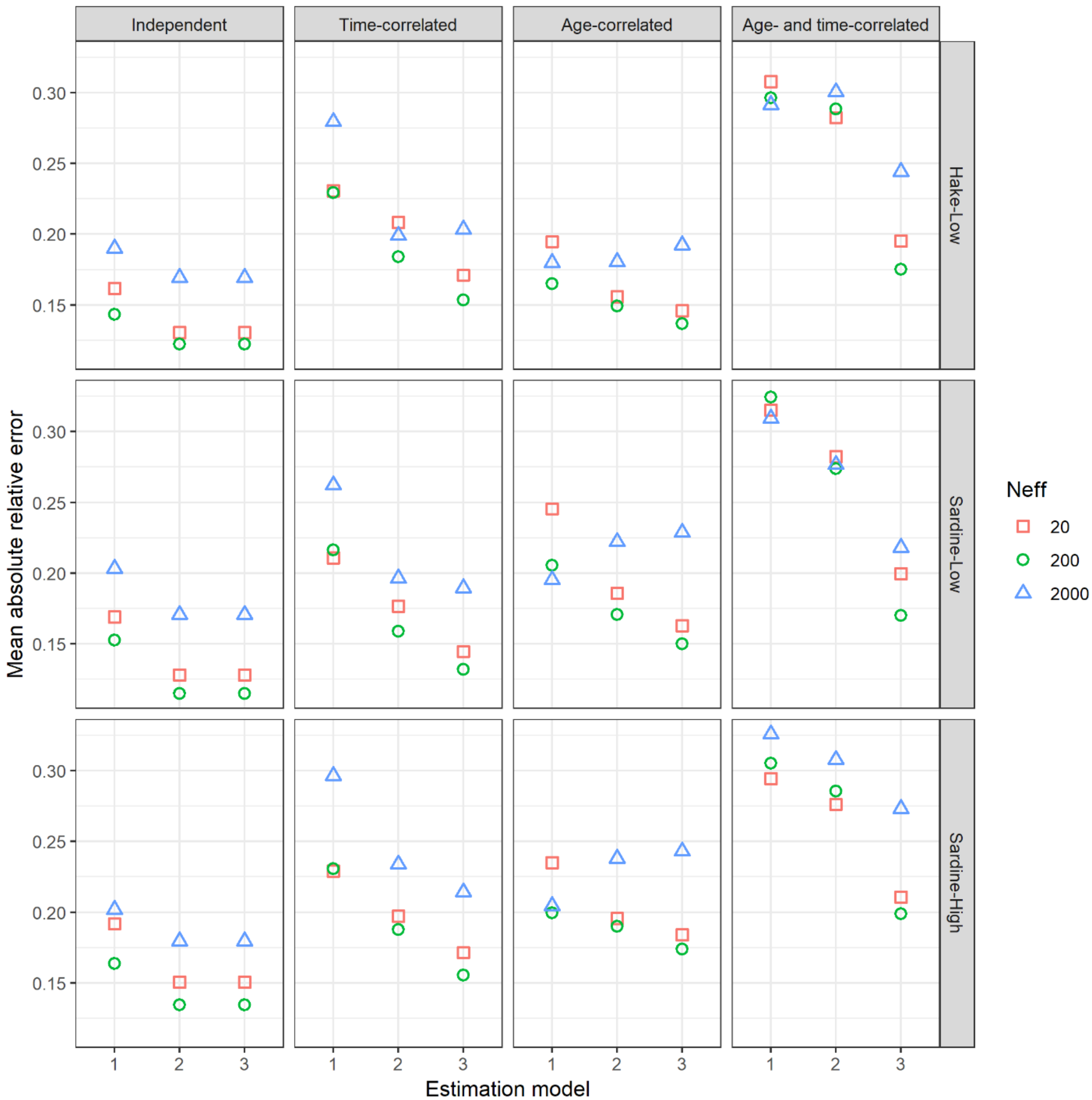

Figure 3. 1st simulation experiment: mean absolute relative error in the estimate of final year spawning biomass under the scenario of under-weighting (red circle), right-weighting (green triangle), or over-weighting (blue square) age-composition data. The four columns correspond to the four autocorrelation cases for simulated selectivity deviations: *Independent*, *Time-correlated*, *Age-correlated*, and *Age- and time-correlated*. EM1-3 have different selectivity specifications: *zero deviations*, *IID deviations*, and *AR deviations*.

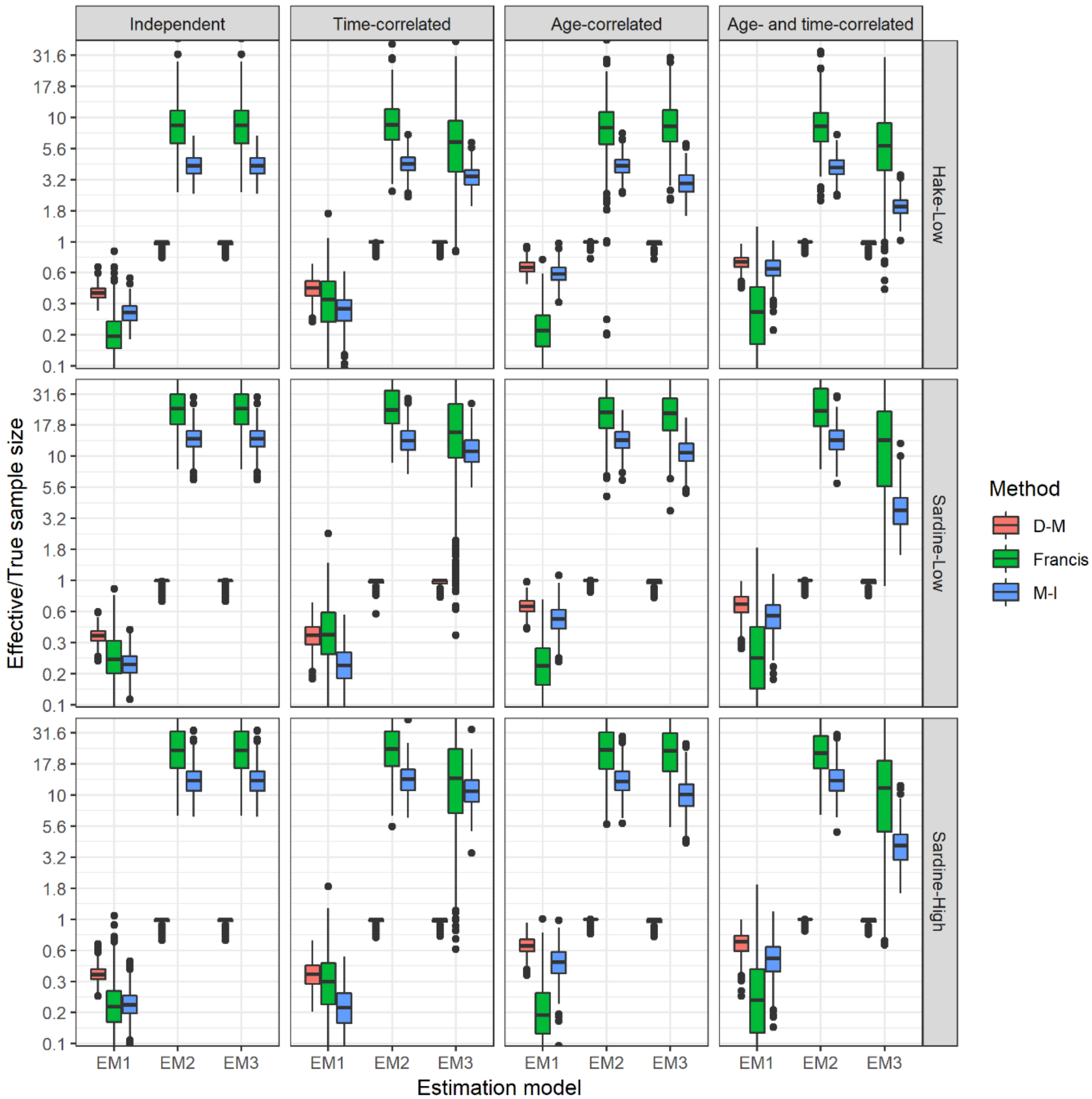

Figure 4. 2nd simulation experiment: boxplot for the ratio of effective sample size to true sample size using the Dirichlet-multinomial (D-M), Francis, and McAllister-Ianelli (M-I) methods. The lower and upper hinges mark the first and third quantiles and the two whiskers extend to the value no further than 1.5 interquartile range from the corresponding hinge. The four columns correspond to the four autocorrelation cases for simulated selectivity deviations: *Independent*, *Time-correlated*, *Age-correlated*, and *Age- and time-correlated*. EM1-3 have different selectivity specifications: *zero deviations*, *IID deviations*, and *AR deviations*.

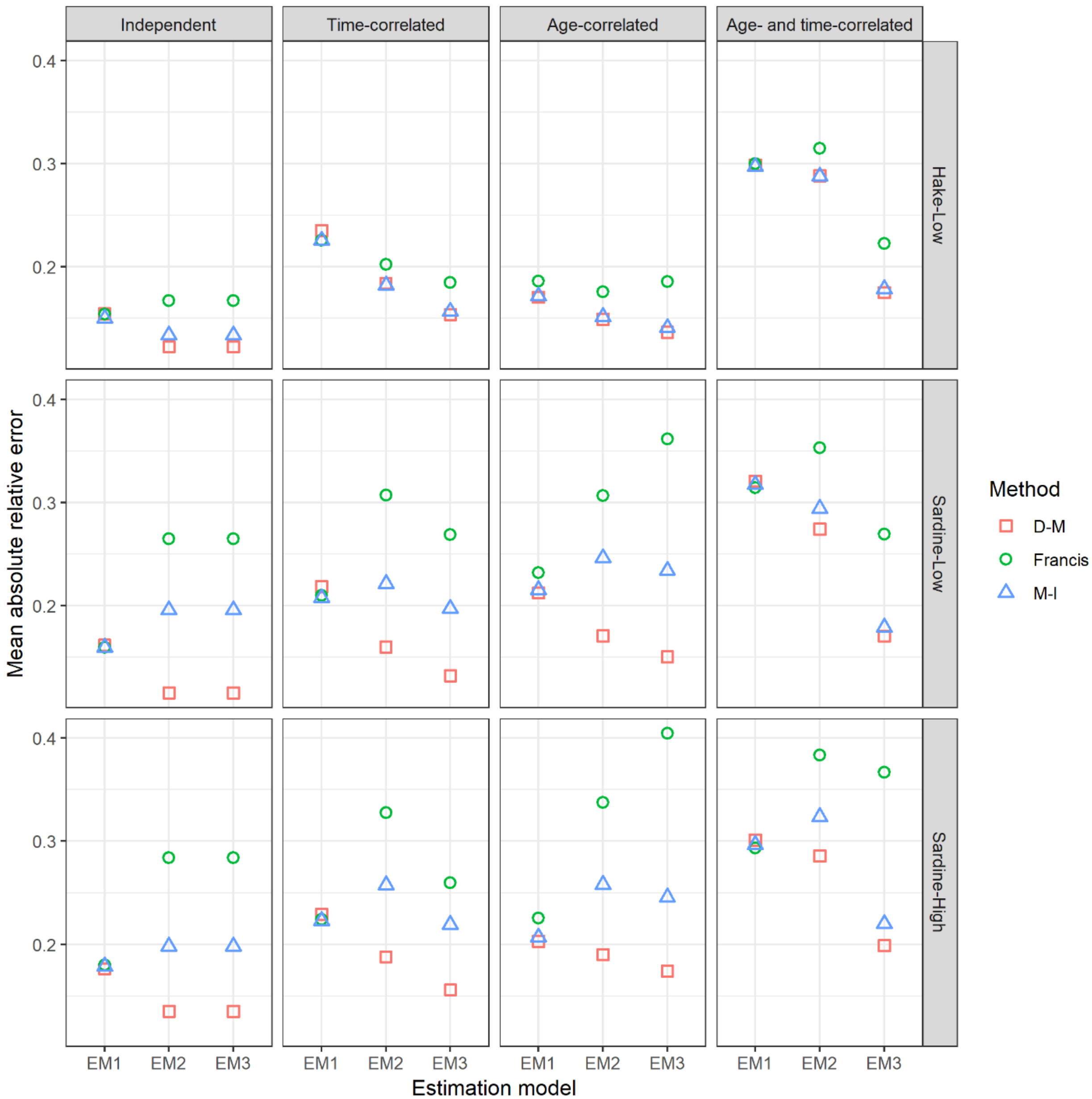

Figure 5. 2nd simulation experiment: mean absolute relative error in the estimate of final year spawning biomass under the Dirichlet-multinomial (D-M), Francis, and McAllister-Ianelli (M-I) methods. The four columns correspond to the four autocorrelation cases for simulated selectivity deviations: *Independent*, *Time-correlated*, *Age-correlated*, and *Age- and time-correlated*. EM1-3 have different selectivity specifications: *zero deviations*, *IID deviations*, and *AR deviations*.

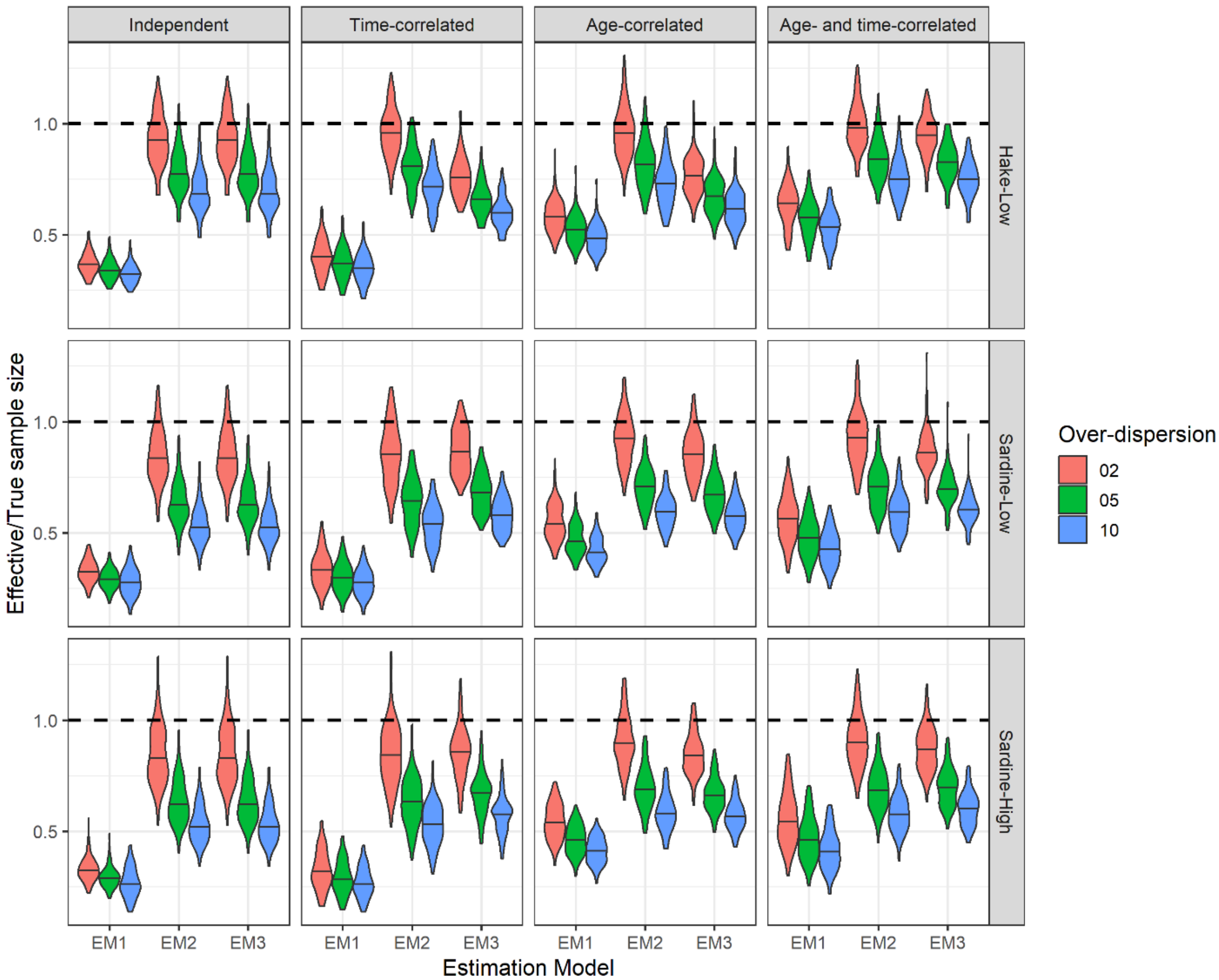

Figure 6. 3rd simulation experiment: violin plots for the ratio of effective to true sample size under three degrees of over-dispersion (d = 2, 5, and 10) in age-composition data. The horizontal line in the violin plot denotes the median. The four columns correspond to the four autocorrelation cases for simulated selectivity deviations: *Independent*, *Time-correlated*, *Age-correlated*, and *Age- and time-correlated*. EM1-3 have different selectivity specifications: *zero deviations*, *IID deviations*, and *AR deviations*.

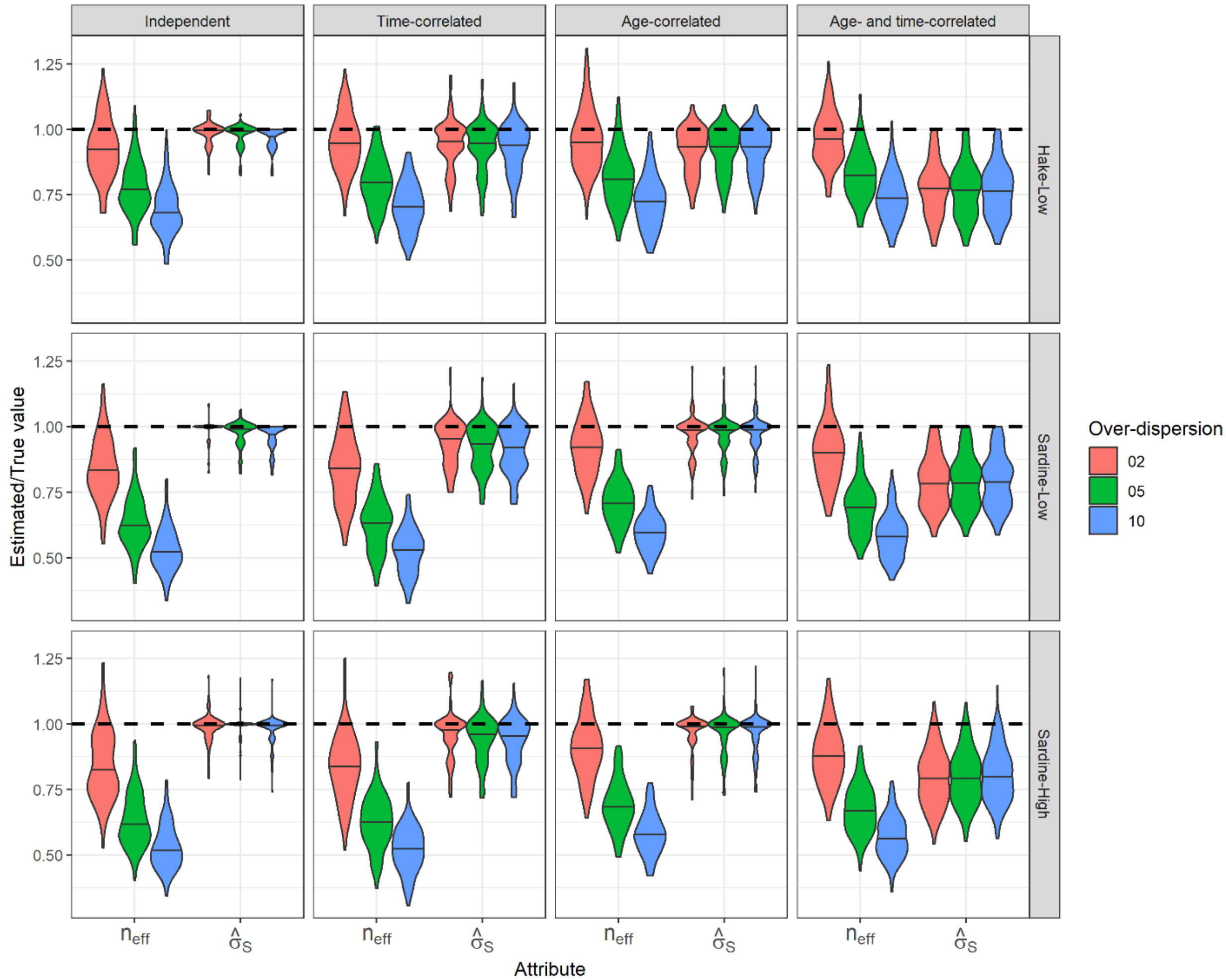

Figure 7. 4th simulation experiment: violin plots for the ratio of effective sample size (n_{eff}) to true sample size and the ratio of estimated ($\hat{\sigma}_S$) to the true level of selectivity variation under three degrees of over-dispersion (d = 2, 5, and 10) in age-composition data. The horizontal line in the violin plot denotes the median. The four columns correspond to the four autocorrelation cases for simulated selectivity deviations: *Independent*, *Time-correlated*, *Age-correlated*, and *Age- and time-correlated*.

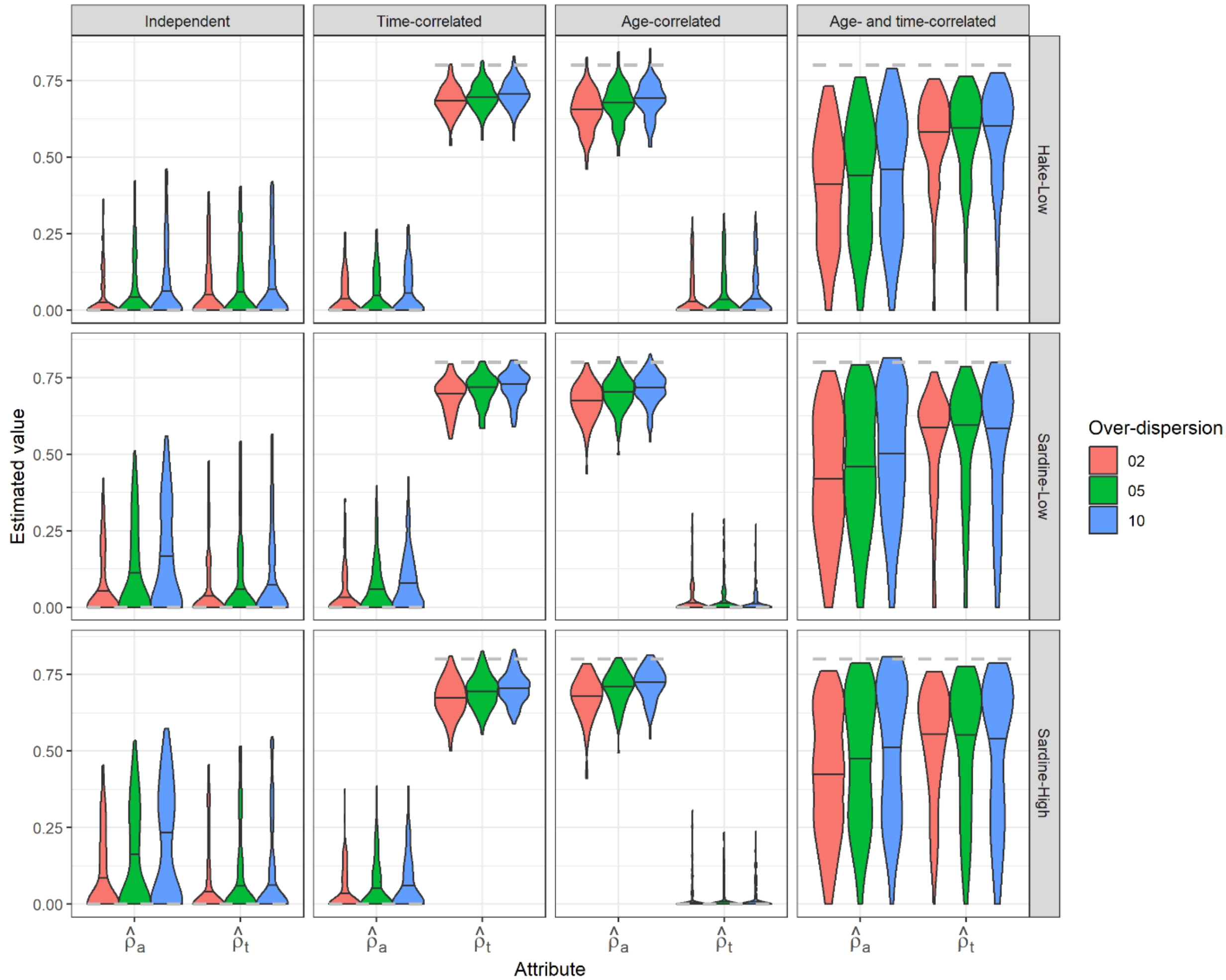

Figure 8. 4th simulation experiment: violin plots for the estimates of selectivity autocorrelations in age ($\hat{\rho}_a$) and time ($\hat{\rho}_t$) under three degrees of over-dispersion (d = 2, 5, and 10) in age-composition data. The horizontal line in the violin plot denotes the median. The four columns correspond to the four autocorrelation cases for simulated selectivity deviations: *Independent*, *Time-correlated*, *Age-correlated*, and *Age- and time-correlated*. Horizontal dashed lines mark the true value for each autocorrelation coefficient in selectivity.

Table 1. Population dynamic equations in the operating model and estimation model.

No.	Equation	Comment
T1.1	$N_{a,t} = \begin{cases} R_t & a=0 \\ N_{a-1,t-1}\exp(-S_{a,t-1}F_{t-1}-M) & 0<a<A \\ N_{A-1,t-1}\exp(-S_{A,t-1}F_{t-1}-M)+N_{A,t-1}\exp(-S_{A,t-1}F_{t-1}-M) & a=A \end{cases}$	Stock equations
T1.2	$\ln(R_t) \sim \mathrm{N}\left(\ln\left(\frac{4hR_0SB_t}{SB_0(1-h)+SB_t(5h-1)}\right)-\frac{\sigma_R^2}{2},\sigma_R^2\right)$	Recruitment
T1.3	$SB_t = \sum_{a=0}^{A} w_a M_a N_{a,t}$	Spawning biomass
T1.4	$C_{a,t} = N_{a,t}\frac{S_{a,t}F_t}{S_{a,t}F_t+M}(1-e^{-S_{a,t}F_t-M})$	Catch-at-age
T1.5	$\ln(N_{a,1}) \sim \mathrm{N}\left(\ln(R_0e^{-aM})-\frac{\sigma_R^2}{2},\sigma_R^2\right)$	Initial conditions
T1.6	$F_t = F_{t-1}\left(\frac{SB_{t-1}}{\gamma SB_0}\right)^{\lambda}\ (F_{t=1}=0.1)$	Fishing mortality

Table 2. Parameter values for the two types of life history investigated in this study.

Parameter Name	Symbol	Pacific hake	Pacific sardine
Natural mortality rate	M	0.386 yr-1	0.552 yr-1
Length at age 0	L_0	1 cm	1 cm
Asymptotic maximum length	$Linf$	90 cm	30 cm
Von Bertalanffy growth coefficient	k	0.20 yr-1	0.30 yr-1
Log-maximum annual spawner per spawner	$LMARR$	2	1
Age at 50% selection in the fishery	S_{50}	5.44	3.55
Rate of change in selectivity at age	S_{slope}	1	2
Age at 50% maturity	a_{mat}	5.44	3.55
Steepness of the Beverton-Holt SR function	h	0.83	0.55
Ratio of equilibrium SB to unfished SB	γ	0.4	0.4
Acceleration rate in fishing mortality	λ	0.2	0.2

Table 3. Summary of the factorial design for each simulation experiment in this study. The columns from left to right represent experiment number, operating models (1-4 represent *Independent*, *Time-correlated*, *Age-correlated*, and *Age- and time-correlated)*, estimation model (1-3 represent *zero deviations*, *IID deviations*, and *AR deviations*), how the effective sample size is estimated in the estimation model, data weighting methods (McAllister-Ianelli (M-I), Francis, and Dirichlet-multinomial (D-M)), the degree of over-dispersion in simulated age-composition data, the input sample size for the D-M method, and how the level of variation in selectivity is specified in the estimation model.

Exp	OM	EM	n_{eff}	Data weighting	Over-dispersion	n_{input} (D-M)	Sel var
1	1-4	1-3	Fixed	0.1x, 1x, 10x n_{true}	-	-	true
2	1-4	1-3	Estimated	M-I, Francis, D-M	-	n_{true}	true
3	1-4	3	Estimated	D-M	2x, 5x, 10x	2x, 5x, 10x n_{true}	true
4	1-4	3	Estimated	D-M	2x, 5x, 10x	2x, 5x, 10x n_{true}	estimated

Appendix

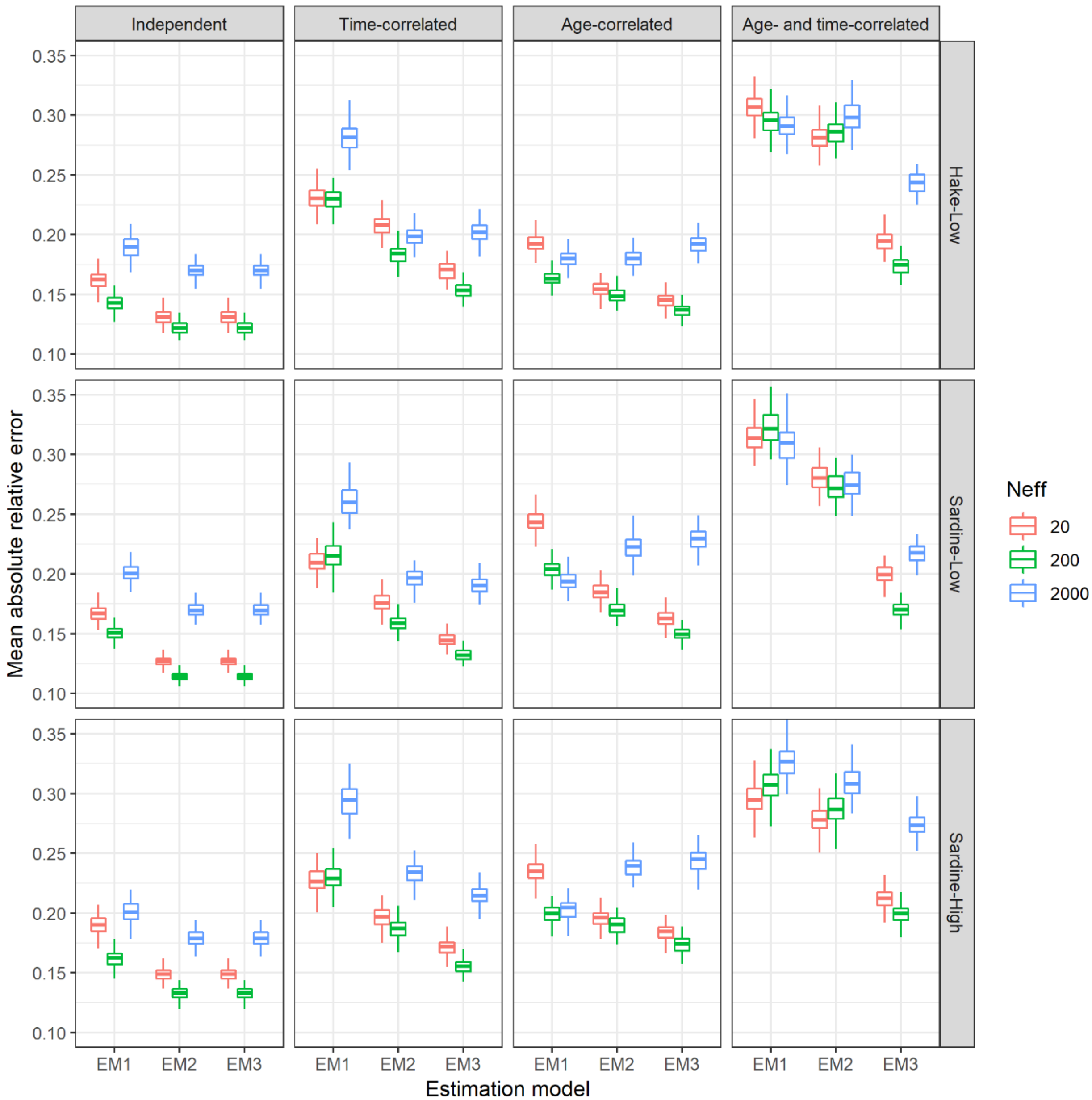

Figure A1. 1st simulation experiment: boxplot for the mean absolute relative error in the estimate of final year spawning biomass showed in Figure 3. To estimate the uncertainty of the mean absolute relative error, the 400 replicates in this simulation experiment were randomly resampled with replacement for 400 times.

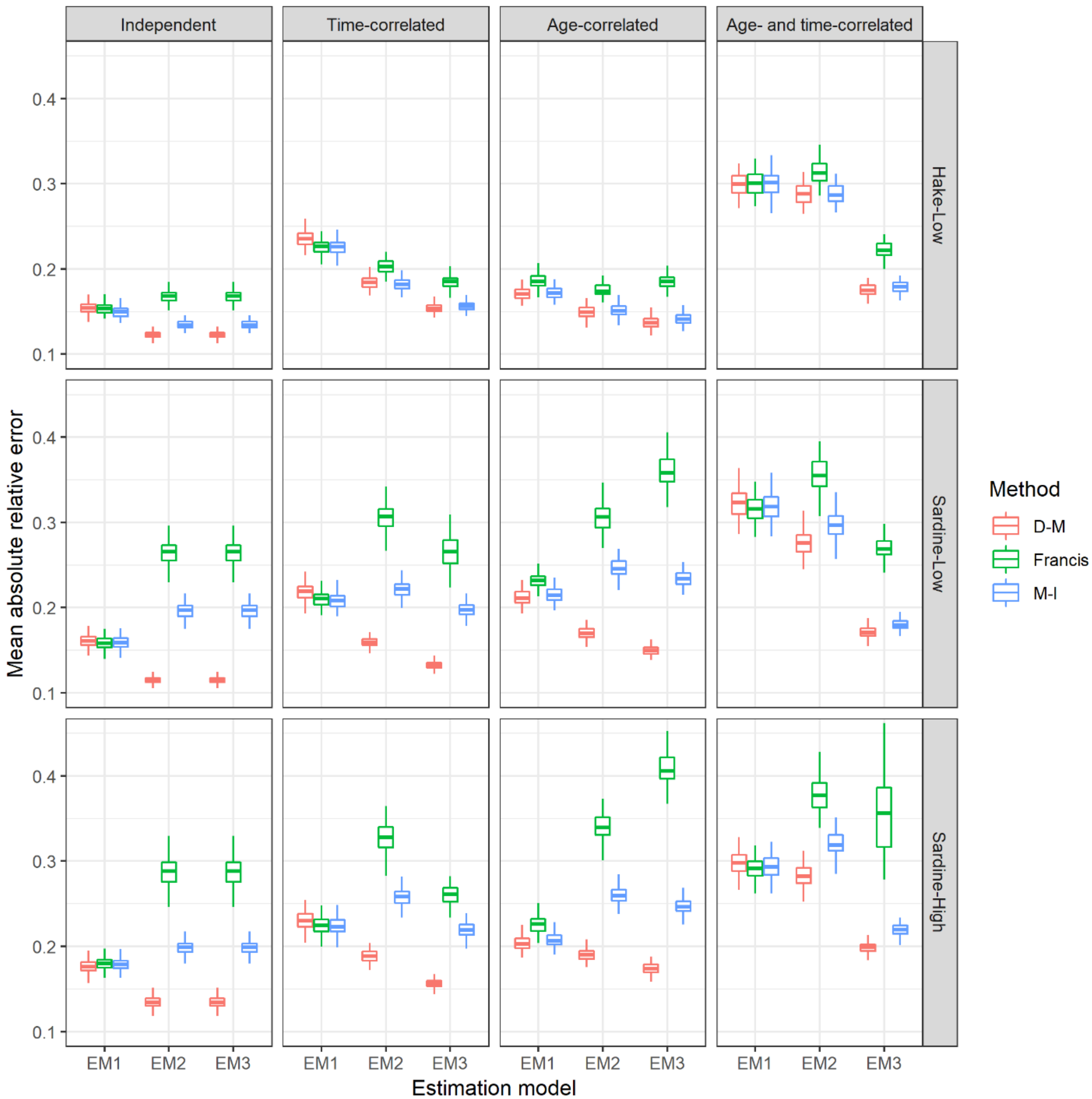

Figure A2. 2^{nd} simulation experiment: boxplot for the mean absolute relative error in the estimate of final year spawning biomass showed in Figure 5. To estimate the uncertainty of the mean absolute relative error, the 400 replicates in this simulation experiment were randomly resampled with replacement for 400 times.

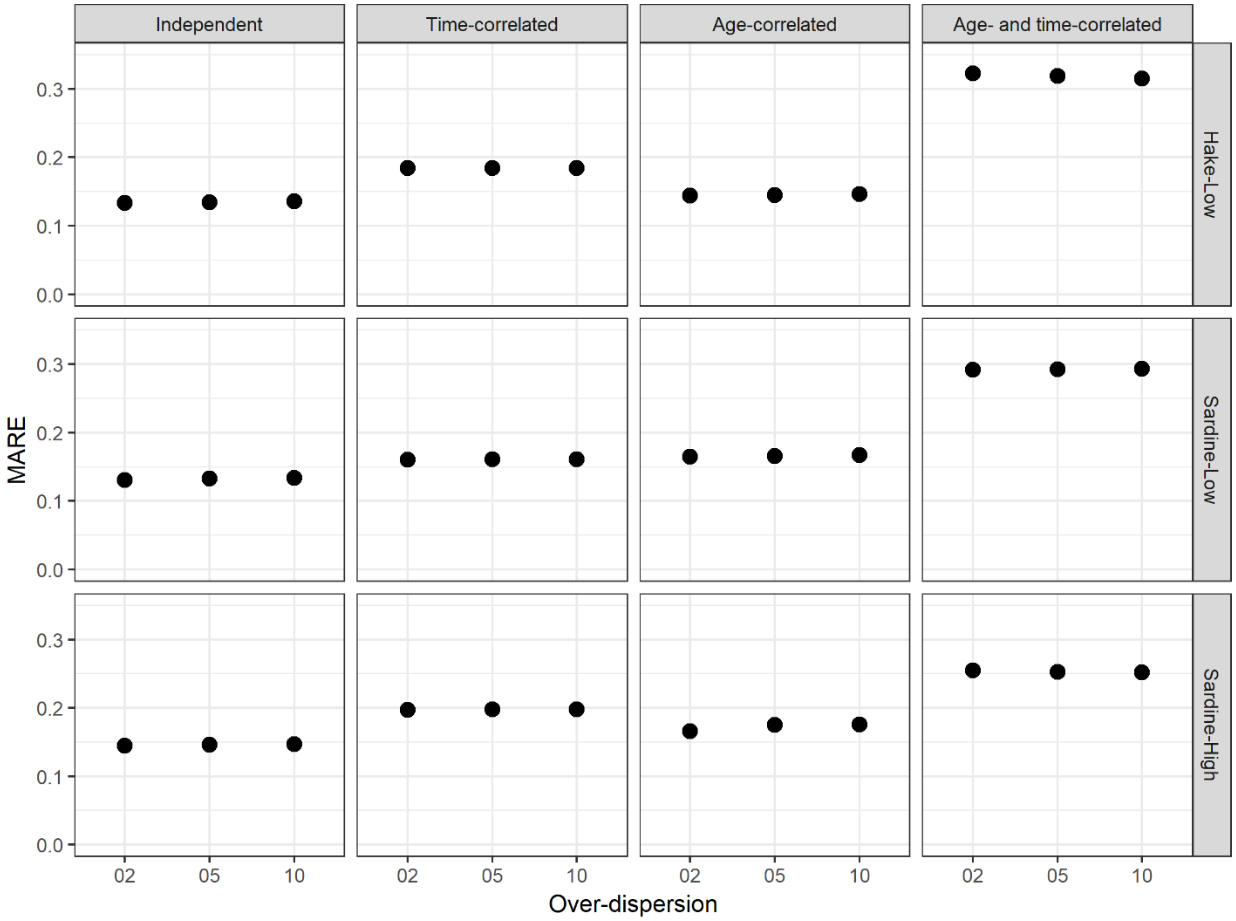

Figure A3. 4th simulation experiment: mean absolute relative error in the estimate of final year spawning biomass three degrees of over-dispersion (d = 2, 5, and 10) in age-composition data. The four columns correspond to the four autocorrelation cases for simulated selectivity deviations: *Independent*, *Time-correlated*, *Age-correlated*, and *Age- and time-correlated*. EM1-3 have different selectivity specifications: *zero deviations*, *IID deviations*, and *AR deviations*.