Evaluating the impact of log-normal bias-correction on a state-space stock assessment model

Chengxue Li[1,2,*] Jonathan J. Deroba[1] Timothy J. Miller[1]

Christopher M. Legault[1] Charles T. Perretti[1]

[1] Northeast Fisheries Science Center, Woods Hole Laboratory, 166 Water Street, Woods Hole, MA 02543 USA

[2] Saltwater Inc., 733 N Street, Anchorage, AK 99501 USA

[*]Corresponding author: chengxue.li@noaa.gov

Abstract

In state-space stock assessment models, recruitment and numbers-at-age are typically modeled as log-normal random variables, with bias correction applied to ensure that their mean matches the expected mean of the random variable. However, it remains unclear whether estimation error in variance parameters, which influence bias correction, propagates to estimates of population quantities. We conducted simulation-estimation experiments to evaluate the effects of bias correction for log-normal random variables and observations. We found that applying bias correction on observations had minimal impact on estimated population quantities, whereas applying bias correction on the process had a significant effect, because estimation error in variance parameters created bias in population estimates. Specifically, when both recruitment deviations and numbers-at-age transitions were treated as random effects, substantial bias in estimated annual recruitments and SSB was found when bias correction was excluded in the operating model but applied in the estimation model. In contrast, not using bias correction had limited negative effects. Thus, we recommend avoiding bias correction for log-normal random variables in state-space models, especially when multiple random-effects processes are modeled simultaneously. (word count: 175)

Keywords: state-space models, random effects, bias correction, recruitment, numbers-at-age transitions

1 Introduction

State-space population models include random and fixed effects, where random effects represent random processes that are separable from observation error. Random effects now have been widely used to model a variety of process errors in state-space stock assessments (Nielsen and Berg, 2014; Cadigan, 2015; Stock and Miller, 2021). Perhaps the most common random effects used in the state-space assessment model are deviations on recruitment and numbers-at-ages 2+ (NAA). Recruitment and NAA random effects are typically assumed to be log-normally distributed (Stock and Miller, 2021).

Error modeled as normally distributed in log-space (i.e., log-normally distributed), implies that error is multiplicative in natural space. Log-normal error will increase the expected value of the population process in natural space, where that increase is related to the variance of the log-normal distribution. In order to ensure that this increase does not occur, one can adjust the mean of the log-normal distribution, known as "bias-correction" (Methot and Taylor, 2011). Although there is not universal agreement on whether bias-correction should be applied, an important open question is the extent to which bias-correction affects the accuracy of important assessment outputs such as recruitment and spawning stock biomass (SSB). Here, we aim to address that question.

Bias in derived population quantities can be exacerbated by the nonlinear transformation (e.g., exponentiation) of a random variable (Thorson and Kristensen, 2016). Whether applying a bias correction term is sufficient to accurately recapture the true population quantities remains an open question (Deroba and Miller, 2016). Methot and Taylor (2011) claimed that population abundance is informed by observations, which are never perfectly accurate and often exhibit inter-annual variability in both quantity and quality. Ignoring this source of variability can induce bias in the estimation of recruitment variability, mean recruitment, and hence management quantities (Methot and Taylor, 2011; Thorson and Kristensen, 2016). An additional plug-in "multiplier" was proposed in maximum likelihood estimation to provide more accurate recruitment estimates (Methot and Taylor, 2011; Thorson and Kristensen, 2016). However, their approach is not appropriate for state-space models. In their simulation experiments, recruitment was treated as a penalized fixed effect and was not integrated out of the likelihood for estimation. In addition, they fixed the recruitment standard deviation (σ_{Rec}) to avoid potential estimation error. In state-space models, however, σ_{Rec} is estimated using the marginal maximum likelihood, which can influence the utility of the log-normal adjustment and derived population quantities.

In addition, evidence has indicated that when multiple processes are treated as random effects in a state-space model, the process variation may not be reliably partitioned for each process due to processes being confounded with each other (Trijoulet et al., 2020; Li et al., 2024; Liljestrand et al., 2024). Improperly estimated process variance can induce inaccurate adjustment and subsequently bias population quantities. Moreover, when bias correction is applied to multiple random processes (e.g., recruitment and NAA), an interaction among the parameters associated with these random processes is introduced in the marginal maximum likelihood estimation. The impacts of this interaction on derived quantities are not fully understood.

To understand the caveats of applying bias correction to log-normal random variables, as well as observations, we designed a simulation-estimation experiment based on three stocks [Georges Bank (GB) yellowtail flounder: *Limanda ferruginea*, Gulf of Maine (GoM) haddock: *Melanogrammus aeglefinus*, and Atlantic mackerel: *Scomber scombrus*]. We explored scenarios where either recruitment only or both recruitment and NAA were treated as random effects, with different autocorrelation structures (see Section 2.4 for more details). Overall, the goal of this study is to provide guidance on bias-correction of log-normal random effects and observations in state-space assessment models.

2 Methods

2.1 Overview

The Woods Hole Assessment Model (WHAM) is a state-space assessment model (https://timjmiller.github.io/wham) (Stock and Miller, 2021). WHAM can incorporate varying population and fishery processes, including recruitment, NAA, natural mortality, fishing selectivity, and survey catchability (Stock and Miller, 2021). WHAM is currently used to manage various stocks in the US northeast region. Below, we describe the population processes and observations where a log-normal distribution is assumed.

2.2 Population numbers-at-age

The transitions between numbers-at-age are described as:

$$\log(N_{a,y}) = \begin{cases} \mu_{Rec} + \epsilon_{1,y} & \text{when } a = 1 \\ \log(N_{a-1,y-1}e^{-Z_{a-1,y-1}}) + \epsilon_{a,y} & \text{when } 1 < a < A \\ \log(N_{A-1,y-1}e^{-Z_{A-1,y-1}} + N_{A,y-1}e^{-Z_{A,y-1}}) + \epsilon_{A,y} & \text{when } a = A \end{cases} \tag{1}$$

where $N_{a,y}$ is the numbers-at-age a in year y, μ_{Rec} is the mean recruitment in log scale, $Z_{a,y}$ is the total mortality rate for age a in year y [i.e., sum of fishing mortality ($F_{a,y}$) and natural mortality ($M_{a,y}$)], f represents the stock-recruitment function, A defines the plus-group, and ϵ is the error term that represents recruitment ($\epsilon_{1,y}$) and and NAA ($\epsilon_{a,y}$) random effects.

Recruitment and NAA random effects are assumed to be log-normally distributed with bias correction, given as:

$$\epsilon_{a,y} \sim \begin{cases} \mathcal{N}\left(-\frac{\sigma^2_{Rec}}{2}, \sigma^2_{Rec}\right), & \text{if } a = 1 \\ \mathcal{N}\left(-\frac{\sigma^2_{NAA}}{2}, \sigma^2_{NAA}\right), & \text{if } a > 1 \end{cases} \tag{2}$$

where σ_{Rec} represents the variance for recruitment and σ_{NAA} represents the shared variance for all other ages. $\sigma^2/2$ is the bias correction term. If bias correction is not used, the mean of random effects in log space becomes zero instead of $-\sigma^2/2$. Note that when random effects are autocorrelated across years, the bias correction term becomes $-\sigma^2/[2 \cdot (1-\rho_y^2)]$ where ρ_y indicates the first-order autocorrelation across years. For more details please see Stock et al (2021).

For example, assuming that recruitment is random about some mean value, when bias correction is not applied:

$$R_y = \bar{R}_y \cdot e^{\epsilon_y}, \text{ with } \epsilon_y \sim \mathcal{N}(0, \sigma^2_{Rec}) \tag{3}$$

where R_y is recruitment in year y, $\bar{R}_y$ is the mean recruitment estimated in the model, and ϵ_y is the inter-annual deviations from the mean recruitment in log space. The expectation of the $\bar{R}_y \cdot e^{\epsilon_y}$ in Eq. 3 is:

$$E[\bar{R}_y \cdot e^{\epsilon_y}] = E[\bar{R}_y] \cdot E[e^{\epsilon_y}] = \bar{R}_y \cdot e^{\frac{\sigma^2_{Rec}}{2}} \tag{4}$$

Then, because $e^{\frac{\sigma^2_{Rec}}{2}} > 1$

$$E[R_y] \neq \bar{R}_y \tag{5}$$

Note that the median of e^{ϵ_y} is 1 here, therefore $\bar{R}_y$ is “median unbiased”, but $\bar{R}_y$ is always less than $E[R_y]$ and the difference becomes larger as the variance σ^2_{Rec} increases.

Therefore, a bias correction term can be applied here to ensure $E[R_y] = \bar{R}_y$:

$$R_y = \bar{R}_y \cdot e^{\left(\epsilon_y - \frac{\sigma^2_{Rec}}{2}\right)}, \text{ with } \epsilon_y \sim \mathcal{N}(0, \sigma^2_{Rec}) \tag{6}$$

2.3 Aggregate catch and indices

Observed, annual, aggregate fishery catch is also assumed to be log-normally distributed:

$$\log(\widehat{C}_y) \sim \mathcal{N}\left(\log(C_y) - \frac{\sigma^2_C}{2}, \sigma^2_C\right) \tag{7}$$

where $\widehat{C}_y$ is the observed fleet catch in year y, C_y is the unobserved true catch, $\sigma^2_{C_y}$ is an input variance for catch observation (fixed across years in our study), and $-\sigma^2_{C_y}/2$ is the bias correction term. Note that $-\sigma^2_{C_y}/2$ is omitted from the Eq. 7 when bias correction is not applied. With bias correction, the observation model is specified so that the mean equals the true catch (mean-unbiased); without bias correction, the observation model is specified so that the median equals the true catch (median-unbiased).

Observations of annual aggregate indices of abundance are handled identically to the aggregate catch in Eq. 7.

2.4 Operating model

For each stock, operating models (OMs) were constructed by fitting to real fishery and survey data and conditioning on parameter values informed by recent stock assessments (e.g., Northeast Fisheries Science Center, 2019, 2021; NEFMC Scientific and Statistical Committee, 2023). Within each random-effects structure (Tables S1–S3), we developed four OM variants corresponding to different bias correction treatments: (1) bias correction applied to both the random-effects process and observations, (2) bias correction applied to observations only, (3) bias correction applied to the random-effects process only, and (4) no bias correction. Parameters for the random-effects processes in each OM are provided in Tables S1–S3, and fishery and survey configurations are summarized in Table 1.

2.5 Simulation–estimation experiment

The simulation experiment followed a full factorial design (Table 2). For each OM variant, fixed-effect parameters (including variance parameters for random effects) were used to generate 100 realizations of recruitment and population dynamics. Observation error was applied to produce 100 pseudo-datasets per OM. These pseudo-datasets were then fitted with all four estimation model (EM) variants differing in bias correction treatment, producing all OM–EM bias correction combinations within the same random-effects structure. This design produced both self-tests (EM bias correction treatment matches OM bias correction treatment) and cross-tests (EM bias correction treatment differs from OM bias correction treatment) (Deroba et al., 2015). A model was considered converged if two criteria were met: (1) the optimizer had successfully converged, and (2) the Hessian matrix was invertible. Simulations in which any of the four EMs failed to converge were discarded, and additional iterations were run with new random seeds until 100 complete and converged simulation replicates were obtained for every scenario.

2.6 Performance metrics

Model performance was evaluated by calculating the median relative error of recruitment, NAA, and SSB over the model years. The relative error was calculated as:

$$\text{Relative Error}_i = \text{Median}\left(\frac{\widehat{\theta}_{i,y}}{\theta_{i,y}} - 1\right) \tag{8}$$

where $\theta_{i,y}$ represents the true value for year y from the simulated pseudo-dataset i, and $\widehat{\theta}_{i,y}$ is the estimated value from the EM fitting to the pseudo-dataset. Then, the median, 25th quantile, and 75th quantile of these psuedo-dataset medians were calculated.

To provide a performance metric that incorporates both bias and variance simultaneously, the root mean squared relative error (RMSR) was also calculated:

$$\text{RMSR}_i = \sqrt{\frac{1}{T}\sum_{y=1}^{T}\left(\frac{\hat{\theta}_{i,y}-\theta_{i,y}}{\theta_{i,y}}\right)^2} \tag{9}$$

where T is the total number of years in the simulation. The median, 25th quantile, and 75th quantile of these psuedo-dataset medians were calculated.

Considering the asymmetric nature of relative error, which ranges from -1 (100% underestimation) to infinity (∞ overestimation), we also calculated the symmetric signed percentage bias (SSPB) and median log-quotient (MdLQ) to ensure that underestimation and overestimation are penalized equally (Morley et al., 2018):

$$\text{SSPB}_i = 100 \times \text{sign}(\text{MdLQ}_i) \times \left(e^{|\text{MdLQ}_i|} - 1\right) \tag{10}$$

where

$$\text{MdLQ}_i = \text{median}\left(\log\left(\frac{\hat{\theta}_{i,y}}{\theta_{i,y}}\right)\right) \tag{11}$$

Note that a value of zero for either SSPB or MdLQ indicates the model is median-unbiased, while negative and positive values indicate underestimation and overestimation, respectively.

Relative errors of mean recruitment (μ_{Rec}), recruitment standard deviation (σ_{Rec}), NAA standard deviation (σ_{NAA}), and AR(1)-year autocorrelation (ρ_y) (hereafter referred to as random-effects parameters) were also calculated for each pseudo-dataset i:

$$\text{Relative Error}_i = \frac{\hat{\theta}_i}{\theta_i} - 1 \tag{12}$$

In addition, we used the Akaike Information Criterion (AIC) to evaluate the best-fitting EM for each realization i. To compare the relative performance of the EM j, we calculated the delta AIC (dAIC):

$$\Delta\text{AIC}_{i,j} = \text{AIC}_{i,j} - \min_j(\text{AIC}_{i,j}) \tag{13}$$

The the proportions of EMs selected by AIC were also calculated to evaluate AIC's ability to identify the correctly specified model.

3 Results

For OMs with only recruitment random effects, the patterns of self-tests and cross-tests were similar, regardless of whether the autocorrelation structure was IID or AR(1)-year. Similarly, in OMs with both recruitment and NAA random effects, the performance differences between self-tests and cross-tests were consistent across IID and AR(1)-year autocorrelation structures (Figure S1 and Figure S2). Given these consistent patterns, the results from the IID and AR(1)-year OMs were combined for simplicity, resulting in 200 simulation replicates for each random effects structure. The effect of applying the bias correction to the catch and survey observations was trivial relative to the bias correction effect of the process errors. Therefore, to simplify the analysis, we restricted both the OMs and EMs to two primary configurations: one where bias correction was applied to both processes and observations (hereafter 'BC-ON'), and another where it was turned off for both (hereafter 'BC-OFF').

We found that convergence failures were often realization-specific rather than model-specific. In most cases, if one EM failed to converge for a given realization, all EMs failed for that same realization, suggesting that convergence issues were driven by factors other than the bias correction setting. Therefore, to better isolate the effect of bias correction, we calculated a conditional convergence rate: the proportion of simulations that converged for the misspecified model, given that the correctly specified self-test had already converged. This provides a clearer evaluation of how the bias correction mismatch impacts model convergence, with results summarized across all stocks (Table S4).

All performance metrics considered (relative error, RMSR, MdLQ, and SSPB) led to similar conclusions, for brevity, only the relative error results are presented in the main text. The results for the other metrics can be found in the supplementary materials (Figures S4–S12).

3.1 Relative error of recruitment

For OMs with only recruitment random effects, the relative error in recruitment estimates was small across all test scenarios. However, when the OM included both recruitment and NAA random effects, model performance became highly sensitive to mismatches in how bias correction was applied between the OM and EM.

Specifically, applying bias correction in the EM but not in the OM led to substantial overestimation of recruitment (7-37%), with the most severe bias observed for GB yellowtail flounder (37%), followed by Atlantic mackerel (14%) and GoM haddock (7%). Conversely, applying correction in the OM but not the EM resulted in only a slight underestimation for Atlantic Mackerel (8%), while estimates for the other two stocks were relatively unbiased (Figure 1). The poor performance in the former case was corroborated by a greater magnitude of RMSR (Figure S4), MdLQ (Figure S7), and SSPB (Figure S10).

3.2 Relative error of SSB and NAA

Estimates of SSB were accurate when the OM included only recruitment random effects. However, performance degraded in scenarios with both recruitment and NAA random effects, particularly when there was a mismatch in how bias correction was applied.

Specifically, applying bias correction in the EM but not the OM led to a consistent overestimation of SSB (11% for GB yellowtail flounder, 9% for Atlantic mackerel, and 5% for GoM haddock). Conversely, applying correction in the OM but not the EM resulted in a smaller underestimation, which was negligible for GB flounder and GoM haddock (close to 0%) and 7% for Atlantic mackerel (Figure 2). This large positive bias in the former case also translated into a greater overall model error, such as RMSR (Figure S5), MdLQ (Figure S8), and SSPB (Figure S11). The same pattern of bias and overall error was mirrored in the estimates of NAA (Figure S3, Figure S6, Figure S9, Figure S12)

3.3 Relative error of random-effects parameters

Recruitment standard deviation was accurately estimated, or slightly underestimated (Figure 3). When both recruitment and NAA were treated as random effects in the OM, a systematic underestimation of NAA standard deviation was found across self-tests and cross-tests, with the magnitude of underestimation ranging between -30% and -10% (Figure 4). Such consistent underestimation was also found for the AR(1)-year autocorrelation parameter (Figure S13).

3.4 AIC

Although the correctly specified EM was generally preferred based on AIC, the difference in AIC between the correct model and the misspecified model was usually less than two units (Figure 5). Therefore, using the standard rule of thumb that dAIC > 2 represents a significant difference in model performance, AIC would most often not be useful for determining whether bias correction should be applied or not.

4 Discussion

4.1 Log-normal random effects

Our results suggest that bias correction has minimal impact on the estimation of recruitment and SSB in state-space models when only recruitment random effects were present, as both quantities were accurately estimated in self-tests and cross-tests. However, in models with both recruitment and NAA random effects, mismatches in bias correction between the EM and OM (e.g., EM with bias correction while OM without, or vice versa) led to biases in both recruitment and SSB estimates. Recruitment estimates were particularly sensitive, with a maximum median error of 37% overestimation when the EM included bias correction but the OM did not, compared to a maximum median error of 8% underestimation in the opposite scenario. For SSB, biases were generally smaller but still notable, where excluding bias correction in the EM led to less bias in cross-tests, compared to applying it. The lower magnitude of relative error in SSB compared to recruitment in cross-tests is likely due to the relatively smaller process variance in NAA. This reduced variance minimizes the contribution of bias correction to NAA estimates when transformed back to the natural scale, and subsequently to SSB. In contrast, the high variability in recruitment amplifies even small estimation biases in process variance, leading to exponentially larger biases in recruitment estimates when transformed back to the natural scale.

When bias correction is applied to a log-normal random variable, accurately estimating the variance parameters associated with random effects is crucial to ensure that the derived log-normal quantities are correctly transformed back to values on the natural scale. Our study demonstrated that, in most cases, recruitment standard deviation (σ_{Rec}) was well estimated in EMs with bias correction, resulting in accurate population quantity estimates when the OM included only recruitment random effects. In general, with bias correction, both the mean (μ_{Rec}) and standard deviation (σ_{Rec}) of recruitment jointly influence annual recruitment estimates. When only recruitment deviations are treated as random effects, if σ_{Rec} is slightly underestimated, μ_{Rec} may also be underestimated to compensate and maintain a desired solution. This interaction partially explains why recruitment estimates in our study remained relatively unbiased when bias correction was applied in the EM but not in the OM, or vice versa. Additionally, when only recruitment deviations are treated as random effects in the OM, the system becomes simplified by confining the process error to a single source of uncertainty. This allows process variation to be restrictively controlled, even when there is a mismatch between the data generation process (OM) and the fitting process (EM) regarding bias correction. However, when random effects are applied to both recruitment and NAA, their interaction introduces additional complexity to the model, raising estimation challenges in disentangling their individual contributions to the data. This interaction effect, coupled with the mismatch between the data generation process and the fitting process with respect to bias correction, likely contributes to discrepancies in the estimation of population quantities.

The variability of NAA process error (i.e., σ_{NAA}), which contributes to bias correction, can be underestimated in state-space models for several reasons. Simulation studies have shown that estimation bias in variance parameters associated with random effects often arises from multiple confounding processes interacting with one another, regardless of the magnitude of process variation (Li et al., 2024; Liljestrand et al., 2024). Furthermore, variances may not be properly apportioned among different random-effects processes when one process exhibits high variability. For instance, Liljestrand et al. (2024) found that when recruitment and selectivity displayed low variability but survival (i.e., NAA) had high variability, some of the survival variation in the estimation model was misallocated to recruitment. Additionally, underestimation of process variance is common in maximum likelihood estimation, where variance estimates tend to shrink when the sample size is insufficient to fully capture the variability of the process. We found significant correlations between σ_{NAA} and σ_{Rec} in cross-tests with BC-ON in the EM (Figures S15-S16). Given that estimation error of key outputs appears to be related to the level of σ_{NAA} (Figures S17-S18), and that the species with the highest σ_{NAA} also had the largest estimation error in cross-tests (GB yellowtail flounder, Figures 1-2), there appears to be a link between σ_{NAA} and estimation error of key outputs. However, further research is needed to better understand how exactly this error in σ_{NAA} and other parameters propagates to error in recruitment and SSB.

4.2 Log-normal observations

In our study, the bias correction for log-normally distributed observations had a limited impact on derived population quantities, in contrast to the correction for process random effects. This is likely because the observation error variance for fleet catch was fixed at known, low values (CVs ≤ 0.2), following the configuration used in recent assessments for these stocks. Fixing the variance prevents it from being confounded with the process variance during estimation but may also limit the potential influence of the observation-level bias correction.

Our findings appear to contrast with those of Aldrin et al. (2020), who found that applying bias correction to catch data could improve model performance. A key difference lies in the treatment of observation error: in their state-space assessment model (SAM), the catch observation error variance was freely estimated, a flexibility that can make the bias correction more influential, whereas in our study, it was fixed. Furthermore, because the observation error for fleet catch in our study was low and fixed (and generally smaller than the error for the survey indices), the model likely weighted the fleet catch data heavily. As noted by Aldrin et al. (2020), survey index observation error is often less influential as it can be partially absorbed by the estimation of catchability (q). The high confidence placed on the catch data in our models may therefore limit the practical benefit of applying an observation-level bias correction. In addition, our exploratory work indicated that higher observation variance can be misattributed to process variance, a finding consistent with previous reports of variance confounding by Fisch et al. (2023).

We therefore suggest that future simulation studies investigate how the utility of log-normal bias correction for observations is affected by key factors, including: (1) the relative magnitudes of observation and process error, and (2) whether the observation error variance is fixed or freely estimated within the model.

4.3 Implications and future research recommendations

Restricted maximum likelihood (REML) has been proposed as an improvement over marginal maximum likelihood estimation, as it provides an unbiased estimator for the variance of random effects. Unlike marginal maximum likelihood, REML calculates the variance of random effects by integrating the likelihood over both random effects and non-variance fixed effects and has been successfully implemented within Stock Synthesis (Thorson et al., 2015). The application of REML is sparking growing interest in state-space modeling due to its potential to improve variance estimation for random effects and enhance the accuracy of management quantity estimates (Maunder and Thorson, 2019; Thorson, 2019). However, little attention has been given to REML estimation of process variance when multiple confounding random-effects processes occur simultaneously, warranting further exploration in the future.

Our preliminary results suggest that the magnitude of estimation bias in population quantities in cross-tests was not related to the level of recruitment variability but was influenced by the level of the variability of NAA in the OM. For instance, misspecified EMs with bias correction tended to overestimate recruitment, with the degree of overestimation increasing exponentially as σ_{NAA} increased from 0.1 to 0.6 (Figure S17). In contrast, misspecified EMs without bias correction tended to underestimate recruitment as σ_{NAA} increased, though to a lesser extent (Figure S17). Similar patterns were also observed for SSB (Figure S18). Further investigation is needed to better understand the underlying mechanisms driving these patterns.

Studies have demonstrated that ignoring data availability can introduce bias in the log-normal adjustment term and result in inaccurate estimates of log-normal random variables, such as recruitment deviations (Methot and Taylor, 2011; Thorson and Kristensen, 2016). This is because individual recruitment estimates ($\hat{R}_y$) are directly informed by the data, and variations in data quantity and quality across years can introduce additional uncertainty to the estimate of σ_R. Our preliminary analysis of estimates from the intermediate period (with improved data quantity for estimating recruitment and NAA) showed only marginal improvement (Figures S19-S20). This suggests that data quantity and quality are less influential than the estimation of random-effects parameters in adjusting log-normal random variables and deriving management quantities. Future research could explore how accounting for variability in data availability in state-space assessment models might improve estimates of recruitment and other derived quantities.

Overall, bias correction in state-space models should be applied with caution, as its benefits are uncertain when the extent of bias in parameters associated with random effects and their propagation into derived population quantities cannot be reliably quantified. In the absence of strong evidence in support of bias correction, we recommend excluding it, as it appears to have less downside risk in cases where supporting evidence is ambiguous.

5 Acknowledgements

We acknowledge the support provided by Saltwater Inc. for facilitating Chengxue Li's contribution to this project.

6 Competing interests statement

One co-author, Timothy J. Miller, serves as a Guest Editor for CJFAS for this special issue.

7 CRediT authorship contribution statement

Chengxue Li: Conceptualization, Methodology, Software, Writing - original draft, Formal analysis, Visualization.
Jonathan J. Deroba: Conceptualization, Funding acquisition, Supervision, Writing - review & editing.
Timothy J. Miller: Conceptualization, Software, Writing - review & editing.

Christopher M. Legault: Conceptualization, Writing - review & editing.
Charles T. Perretti: Conceptualization, Writing - review & editing.

8 Funding statement

This work was funded by NOAA Fisheries Northeast Fisheries Science Center.

9 Data availability statement

The data underlying this article are available on Github: https://github.com/lichengxue/Bias_Correction_Project.

10 Tables

Table 1. Model configuration for GB Yellowtail Flounder, GoM Haddock, and Atlantic Mackerel. Note: The age composition likelihoods are defined as follows. 'Dirichlet-miss0' uses the Dirichlet distribution where zero observations are treated as missing. 'Logistic-normal-miss0' uses the logistic-normal distribution, also treating zeros as missing. 'Logistic-normal-ar1-miss0' extends the logistic-normal model with a first-order autoregressive (AR1) error structure to account for temporal correlation.

Parameter	**Flounder**	**Haddock**	**Mackerel**
Fleet Catch			
Period	1973-2022	1977-2018	1968-2019
Selectivity form	Logistic	Age-specific	Age-specific
Age comp. likelihood	Dirichlet-miss0	Logistic-normal-miss0	Logistic-normal-ar1-miss0
Survey Indices			
Period	1. 1973-2022	1. 1977-2018	1. 1979-2019
	2. 1973-2022	2. 1977-2018	2. 2009-2019
	3. 1987-2022		3. 1974-2008
Selectivity form	Logistic	Age-specific	Age-specific
Age comp. likelihood	Dirichlet-miss0	Logistic-normal-miss0	Logistic-normal-ar1-miss0

Table 2. Summary of operating models (OMs) and estimation models (EMs) with different random-effects structures and bias correction scenarios. Each OM includes four bias correction scenarios (ON or OFF for process and observation, respectively). Note that a shared AR(1)-year autocorrelation parameter (ρ_y) is used for both recruitment and NAA random effects.

Random Effects Structure	**Parameters**	**Bias-Correct (Proc.)**	**Bias-Correct (Obs.)**
Rec (IID)	σ_{Rec}	ON	ON
		OFF	ON
		ON	OFF
		OFF	OFF
Rec (AR1$_y$)	σ_{Rec}, ρ_y	ON	ON
		OFF	ON
		ON	OFF
		OFF	OFF
Rec+NAA (IID)	$\sigma_{Rec}, \sigma_{NAA}$	ON	ON
		OFF	ON
		ON	OFF
		OFF	OFF
Rec+NAA (AR1$_y$)	$\sigma_{Rec}, \sigma_{NAA}, \rho_y$	ON	ON
		OFF	ON
		ON	OFF
		OFF	OFF

11 Figures

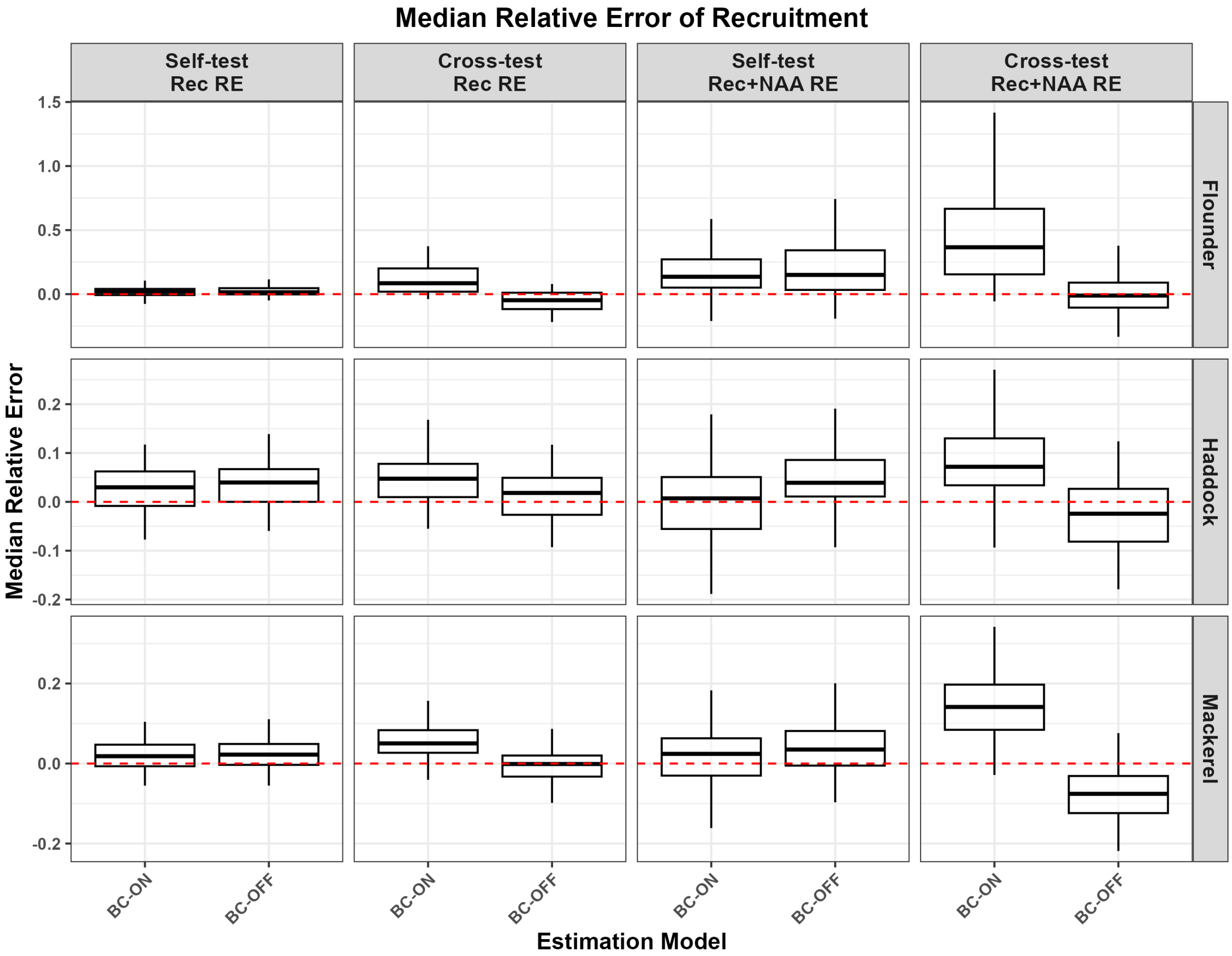

Figure. 1. Median relative error of reccruitment calculated for self-tests and cross-tests. "Rec RE" and "Rec+NAA RE" in the top facet indicate operating models (OMs) with only recruitment random effects and both recruitment and NAA random effects, respectively.

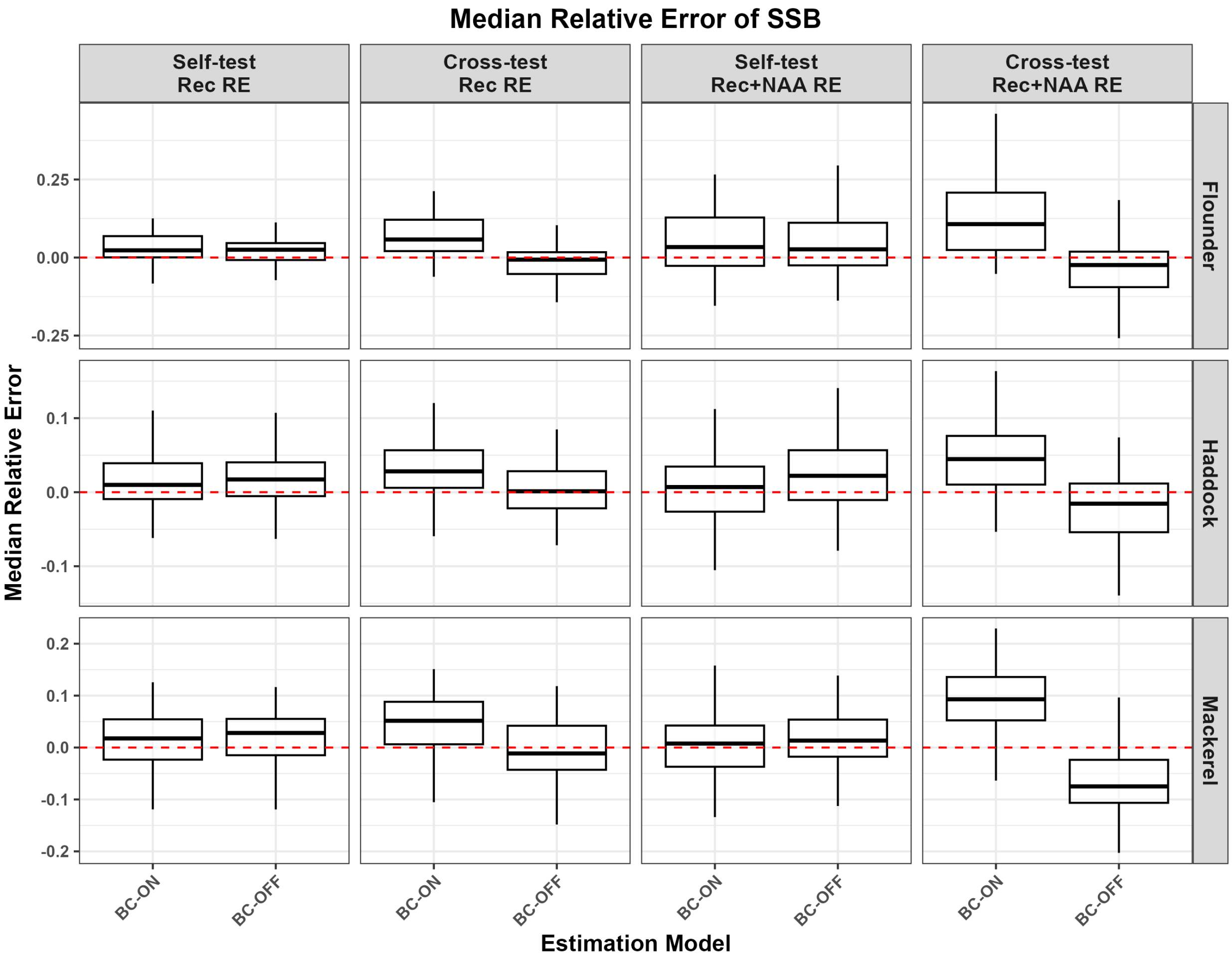

Figure. 2. Median relative error of SSB calculated for self-tests and cross-tests. “Rec RE” and “Rec+NAA RE” in the top facet indicate operating models (OMs) with only recruitment random effects and both recruitment and NAA random effects, respectively.

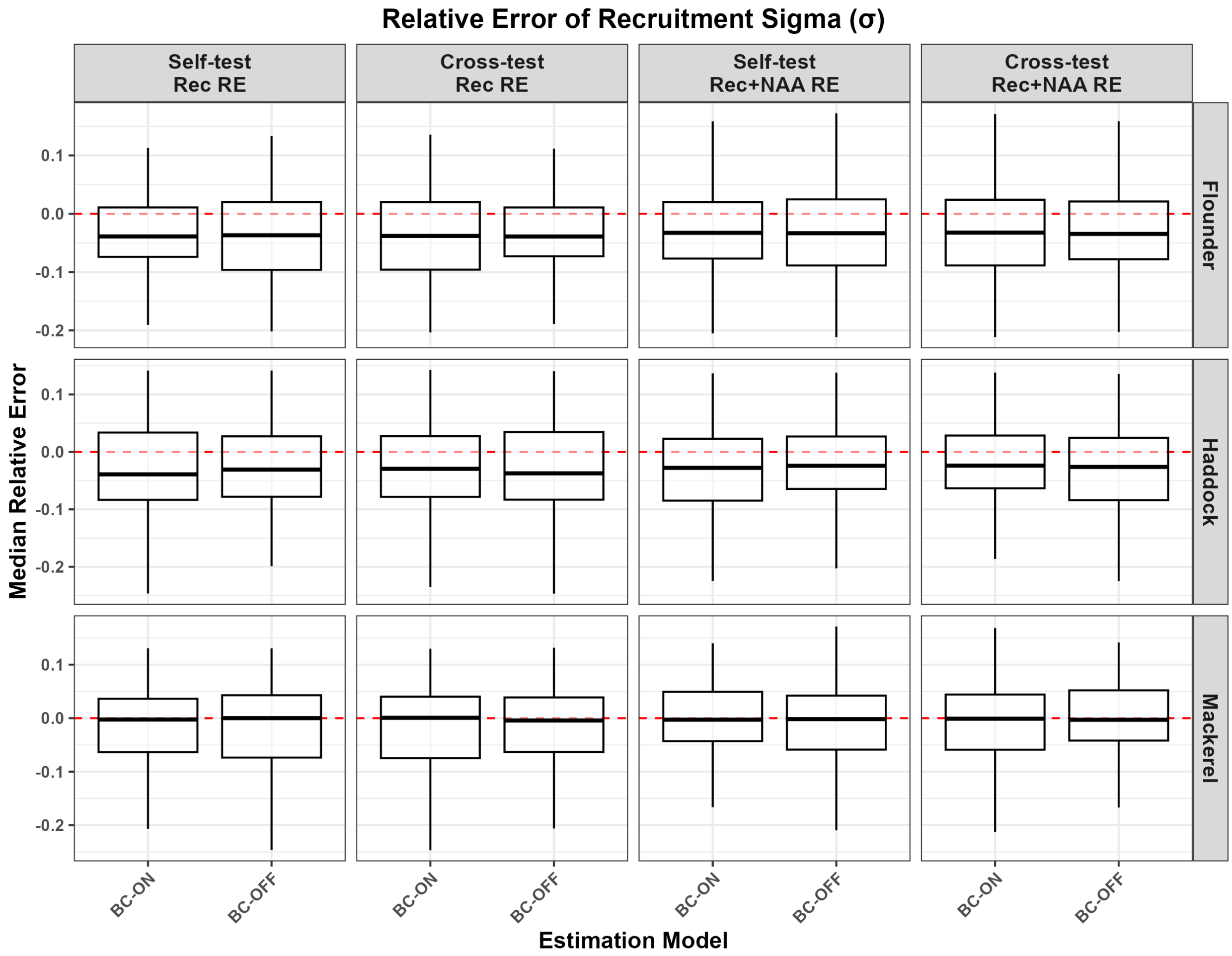

Figure. 3. Relative error of recruitment variance calculated for self-tests and cross-tests. "Rec RE" and "Rec+NAA RE" in the top facet indicate operating models (OMs) with only recruitment random effects and both recruitment and NAA random effects, respectively.

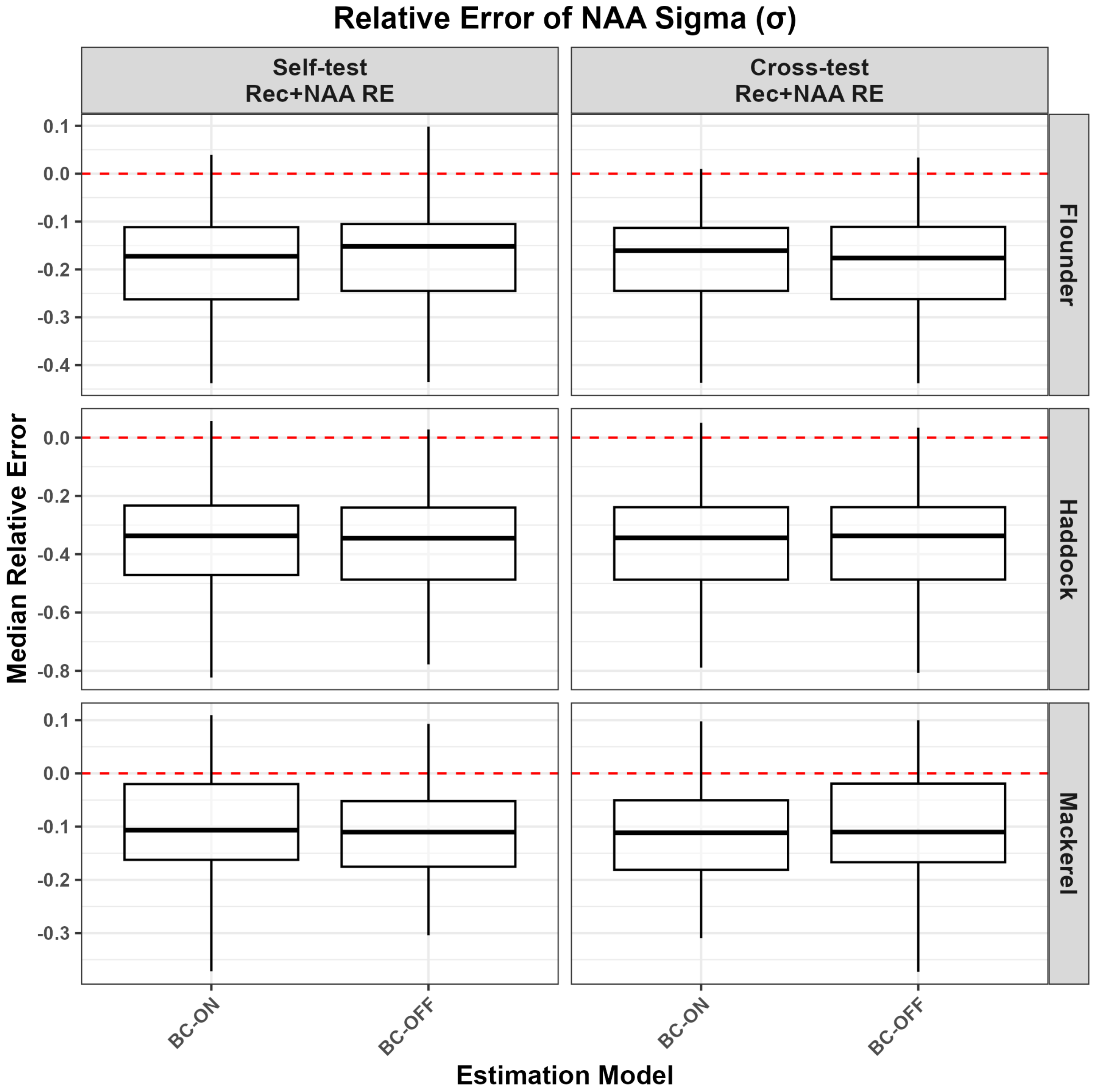

Figure. 4. Relative error of NAA variance calculated for self-tests and cross-tests. "Rec RE" and "Rec+NAA RE" in the top facet indicate operating models (OMs) with only recruitment random effects and both recruitment and NAA random effects, respectively.

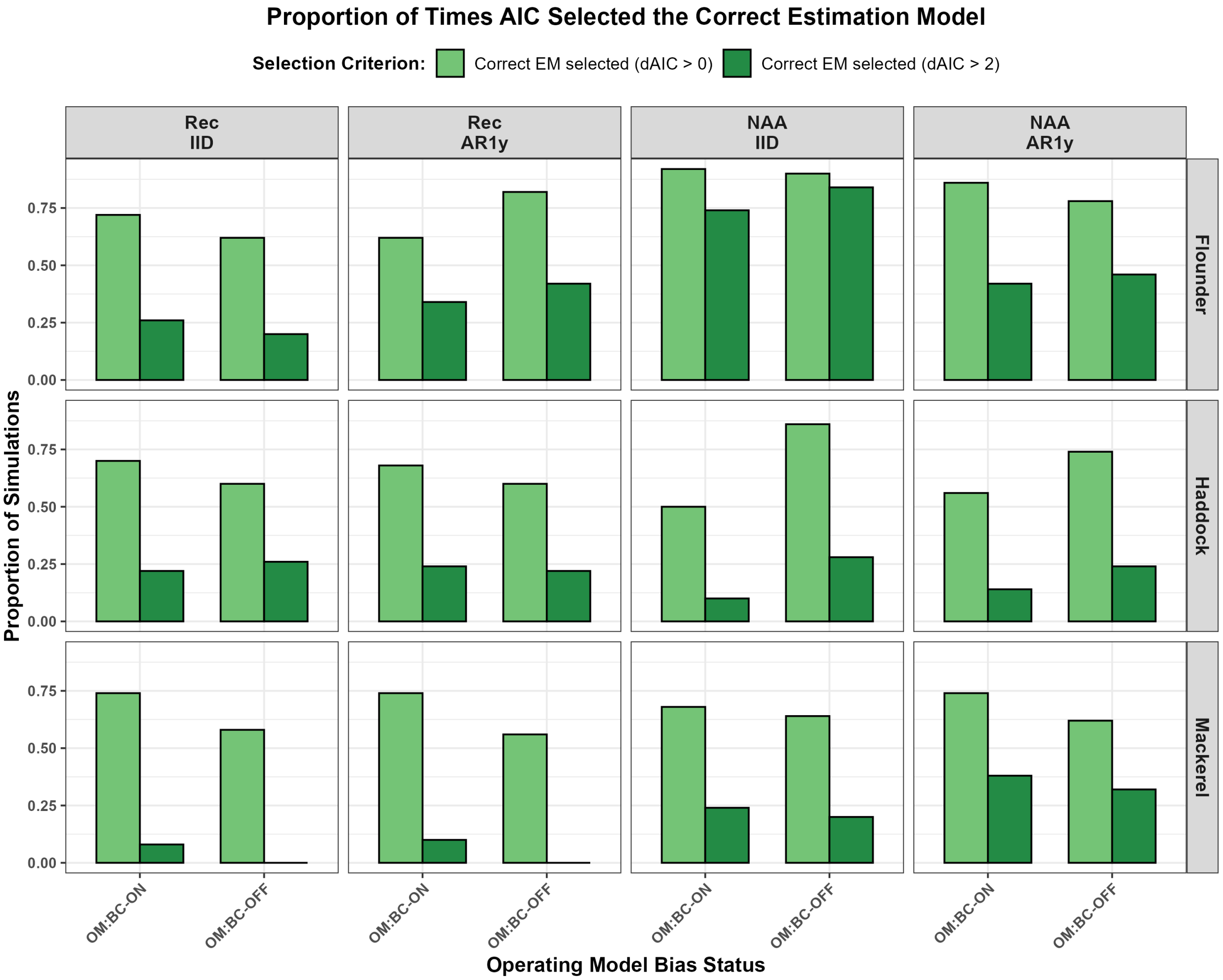

Figure. 5. Proportion of AIC selecting the correct estimation model (EM). The green color represents the proportion of the correct EM selected based on the lowest AIC, while the red color represents the proportion of the correct EM selected when the difference in AIC (dAIC) is greater than 2. The top facet displays operating models (OMs) with different forms of random-effects processes.

12 Supplementary files

Table S1. Parameters associated with random effects processes used for Georges Bank (GB) yellowtail flounder.

Random Effects Structure	Proc./Obs.	Rec sigma	NAA sigma	Rho (AR1_y)
Rec (iid)	ON & ON	1.07	NA	NA
Rec (iid)	OFF & ON	1.07	NA	NA
Rec (iid)	ON & OFF	1.08	NA	NA
Rec (iid)	OFF & OFF	1.08	NA	NA
Rec (ar1_y)	ON & ON	0.37	NA	0.96
Rec (ar1_y)	OFF & ON	0.37	NA	0.96
Rec (ar1_y)	ON & OFF	0.37	NA	0.96
Rec (ar1_y)	OFF & OFF	0.37	NA	0.96
Rec+NAA (iid)	ON & ON	1.23	0.55	NA
Rec+NAA (iid)	OFF & ON	1.23	0.56	NA
Rec+NAA (iid)	ON & OFF	1.24	0.55	NA
Rec+NAA (iid)	OFF & OFF	1.24	0.56	NA
Rec+NAA (ar1_y)	ON & ON	0.55	0.21	0.94
Rec+NAA (ar1_y)	OFF & ON	0.55	0.21	0.94
Rec+NAA (ar1_y)	ON & OFF	0.55	0.21	0.94
Rec+NAA (ar1_y)	OFF & OFF	0.55	0.21	0.94

Table S2. Parameters associated with random effects processes used for Gulf of Maine (GoM) haddock.

Random Effects Structure	Proc./Obs.	Rec sigma	NAA sigma	Rho (AR1_y)
Rec (iid)	ON & ON	1.57	NA	NA
Rec (iid)	OFF & ON	1.57	NA	NA
Rec (iid)	ON & OFF	1.59	NA	NA
Rec (iid)	OFF & OFF	1.59	NA	NA
Rec (ar1_y)	ON & ON	1.16	NA	0.7
Rec (ar1_y)	OFF & ON	1.16	NA	0.7
Rec (ar1_y)	ON & OFF	1.17	NA	0.71
Rec (ar1_y)	OFF & OFF	1.17	NA	0.71
Rec+NAA (iid)	ON & ON	1.60	0.2	NA
Rec+NAA (iid)	OFF & ON	1.60	0.2	NA
Rec+NAA (iid)	ON & OFF	1.62	0.2	NA
Rec+NAA (iid)	OFF & OFF	1.62	0.2	NA
Rec+NAA (ar1_y)	ON & ON	1.18	0.16	0.6
Rec+NAA (ar1_y)	OFF & ON	1.18	0.16	0.6
Rec+NAA (ar1_y)	ON & OFF	1.18	0.17	0.61
Rec+NAA (ar1_y)	OFF & OFF	1.18	0.16	0.61

Table S3. Parameters associated with random effects processes used for Atlantic mackerel.

Random Effects Structure	Proc./Obs.	Rec sigma	NAA sigma	Rho (AR1_y)
Rec (iid)	ON & ON	1.11	NA	NA
Rec (iid)	OFF & ON	1.11	NA	NA
Rec (iid)	ON & OFF	1.11	NA	NA
Rec (iid)	OFF & OFF	1.11	NA	NA
Rec (ar1_y)	ON & ON	1.00	NA	0.46
Rec (ar1_y)	OFF & ON	1.00	NA	0.46
Rec (ar1_y)	ON & OFF	1.01	NA	0.46
Rec (ar1_y)	OFF & OFF	1.01	NA	0.46
Rec+NAA (iid)	ON & ON	1.02	0.28	NA
Rec+NAA (iid)	OFF & ON	1.02	0.28	NA
Rec+NAA (iid)	ON & OFF	1.02	0.28	NA
Rec+NAA (iid)	OFF & OFF	1.02	0.28	NA
Rec+NAA (ar1_y)	ON & ON	0.89	0.32	0.49
Rec+NAA (ar1_y)	OFF & ON	0.89	0.32	0.49
Rec+NAA (ar1_y)	ON & OFF	0.90	0.32	0.48
Rec+NAA (ar1_y)	OFF & OFF	0.90	0.32	0.48

Table S4. Convergence rates of misspecified models across different stocks.

OM	EM	Convergence Rate	Stock	Random Effects Structure
BC-OFF	BC-ON	1.00	Flounder	Rec (IID)
BC-ON	BC-OFF	1.00	Flounder	Rec (IID)
BC-ON	BC-OFF	1.00	Flounder	Rec (AR1y)
BC-OFF	BC-ON	1.00	Flounder	Rec (AR1y)
BC-OFF	BC-ON	1.00	Flounder	Rec+NAA (IID)
BC-ON	BC-OFF	1.00	Flounder	Rec+NAA (IID)
BC-OFF	BC-ON	1.00	Flounder	Rec+NAA (AR1y)
BC-ON	BC-OFF	0.98	Flounder	Rec+NAA (AR1y)
BC-OFF	BC-ON	1.00	Haddock	Rec (IID)
BC-ON	BC-OFF	1.00	Haddock	Rec (IID)
BC-OFF	BC-ON	1.00	Haddock	Rec (AR1y)
BC-ON	BC-OFF	1.00	Haddock	Rec (AR1y)
BC-OFF	BC-ON	0.98	Haddock	Rec+NAA (IID)
BC-ON	BC-OFF	0.96	Haddock	Rec+NAA (IID)
BC-OFF	BC-ON	0.98	Haddock	Rec+NAA (AR1y)
BC-ON	BC-OFF	0.90	Haddock	Rec+NAA (AR1y)
BC-OFF	BC-ON	1.00	Mackerel	Rec (IID)
BC-ON	BC-OFF	1.00	Mackerel	Rec (IID)
BC-OFF	BC-ON	0.98	Mackerel	Rec (AR1y)
BC-ON	BC-OFF	1.00	Mackerel	Rec (AR1y)
BC-OFF	BC-ON	1.00	Mackerel	Rec+NAA (IID)
BC-ON	BC-OFF	1.00	Mackerel	Rec+NAA (IID)
BC-OFF	BC-ON	1.00	Mackerel	Rec+NAA (AR1y)
BC-ON	BC-OFF	0.96	Mackerel	Rec+NAA (AR1y)

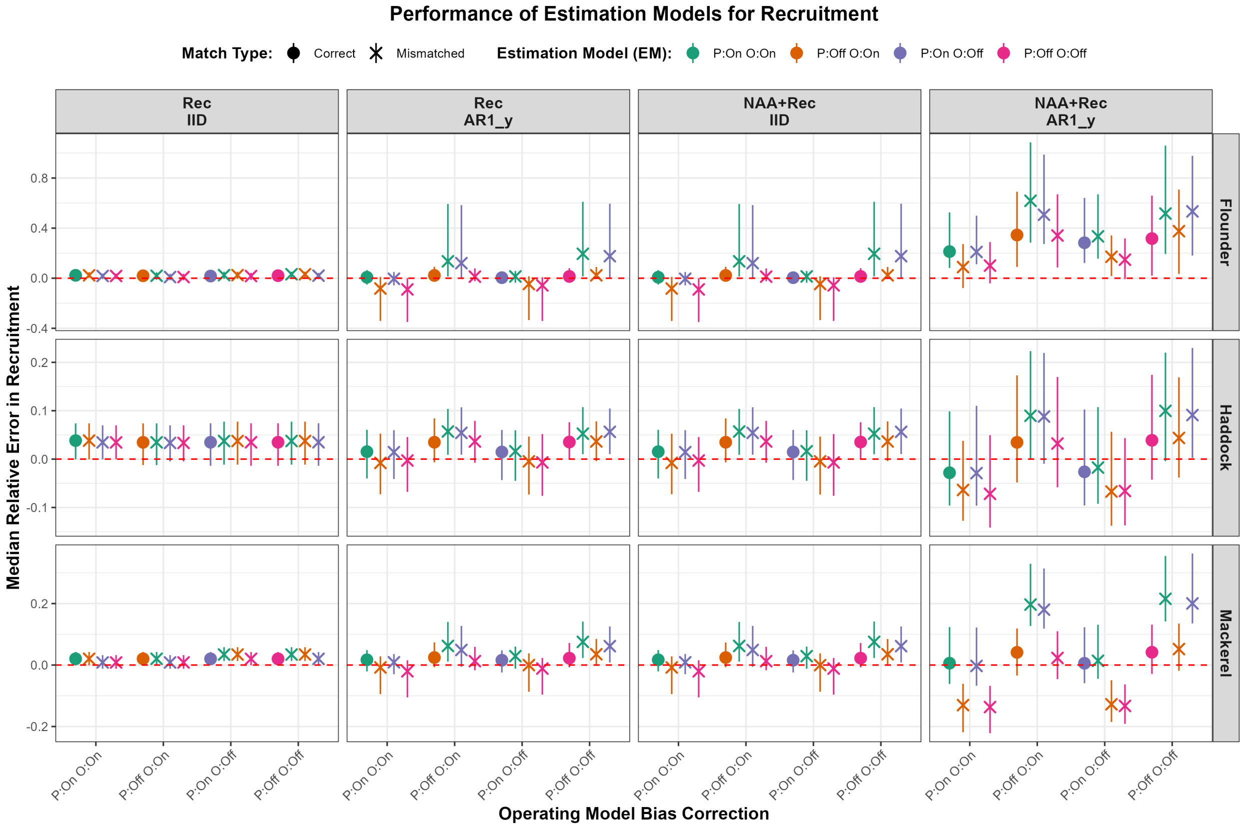

Figure. S1. Median relative error of reccruitment calculated for self-tests and cross-tests. "Rec RE" and "Rec+NAA RE" in the top facet indicate operating models (OMs) with only recruitment random effects and both recruitment and NAA random effects, respectively.

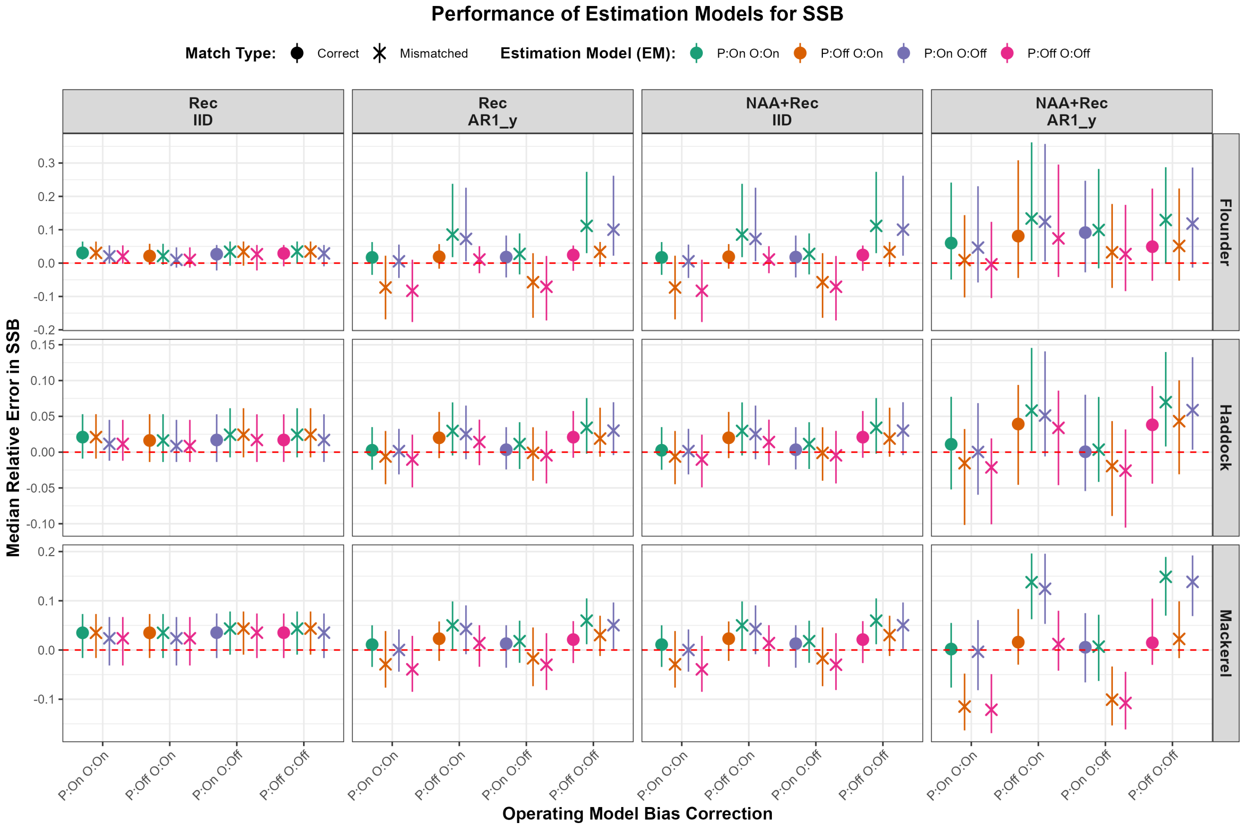

Figure. S2. Median relative error of SSB calculated for self-tests and cross-tests. "Rec RE" and "Rec+NAA RE" in the top facet indicate operating models (OMs) with only recruitment random effects and both recruitment and NAA random effects, respectively.

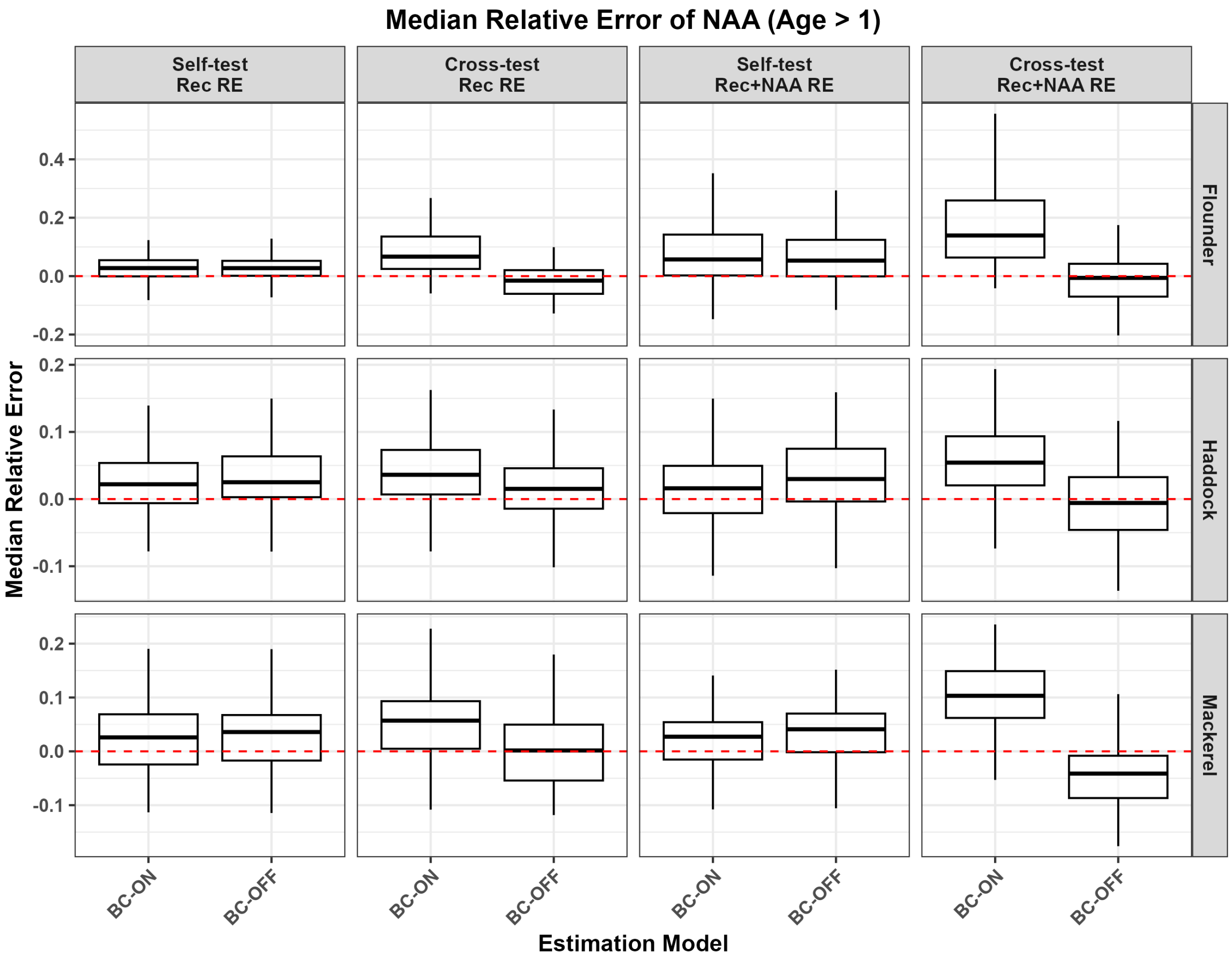

Figure. S3. Median relative error of NAA calculated for self-tests and cross-tests. “Rec RE” and “Rec+NAA RE” in the top facet indicate operating models (OMs) with only recruitment random effects and both recruitment and NAA random effects, respectively.

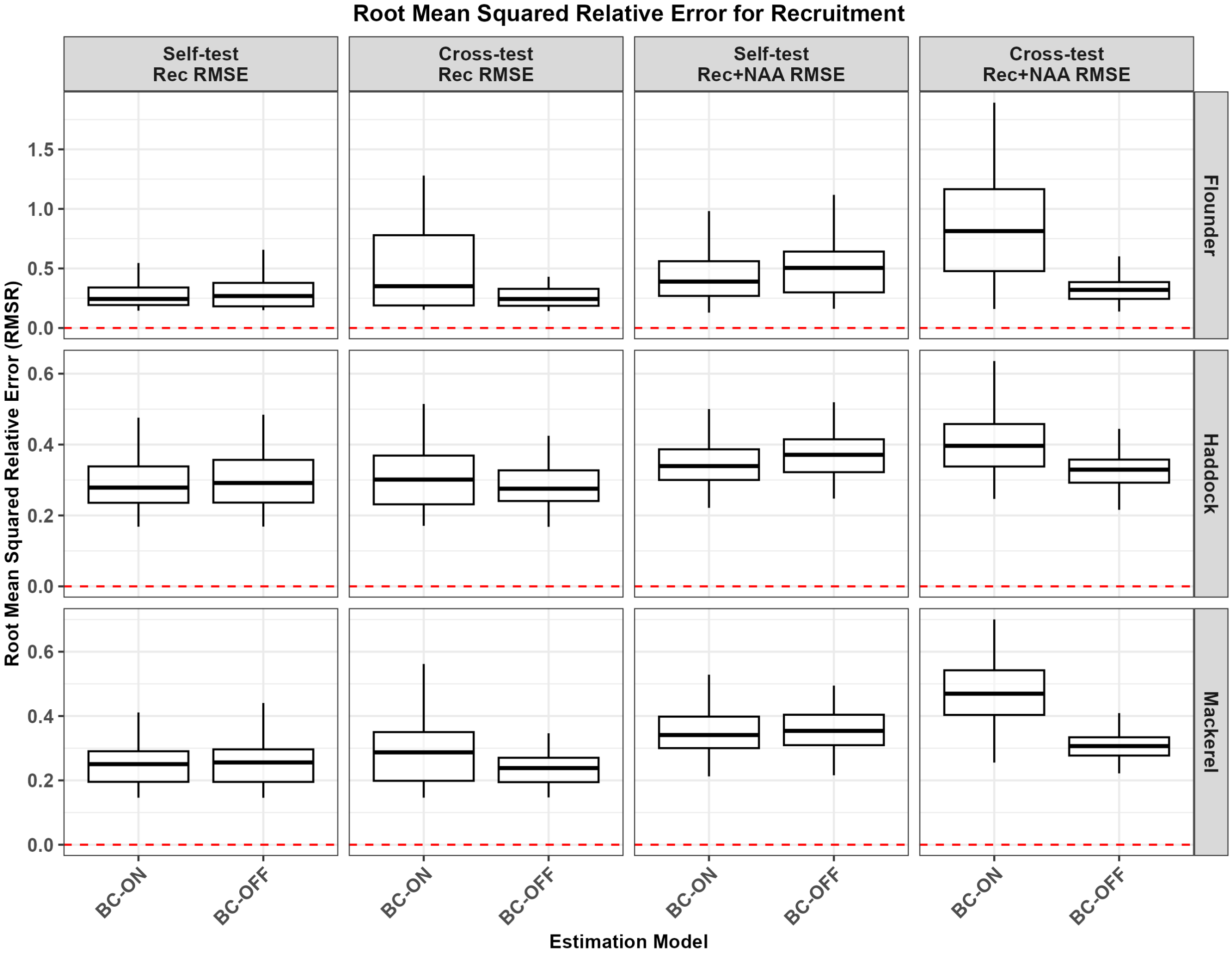

Figure. S4. Root mean squared relative error (RMSR) of recruitment calculated for self-tests and cross-tests. “Rec RE” and “Rec+NAA RE” in the top facet indicate operating models (OMs) with only recruitment random effects and both recruitment and NAA random effects, respectively.

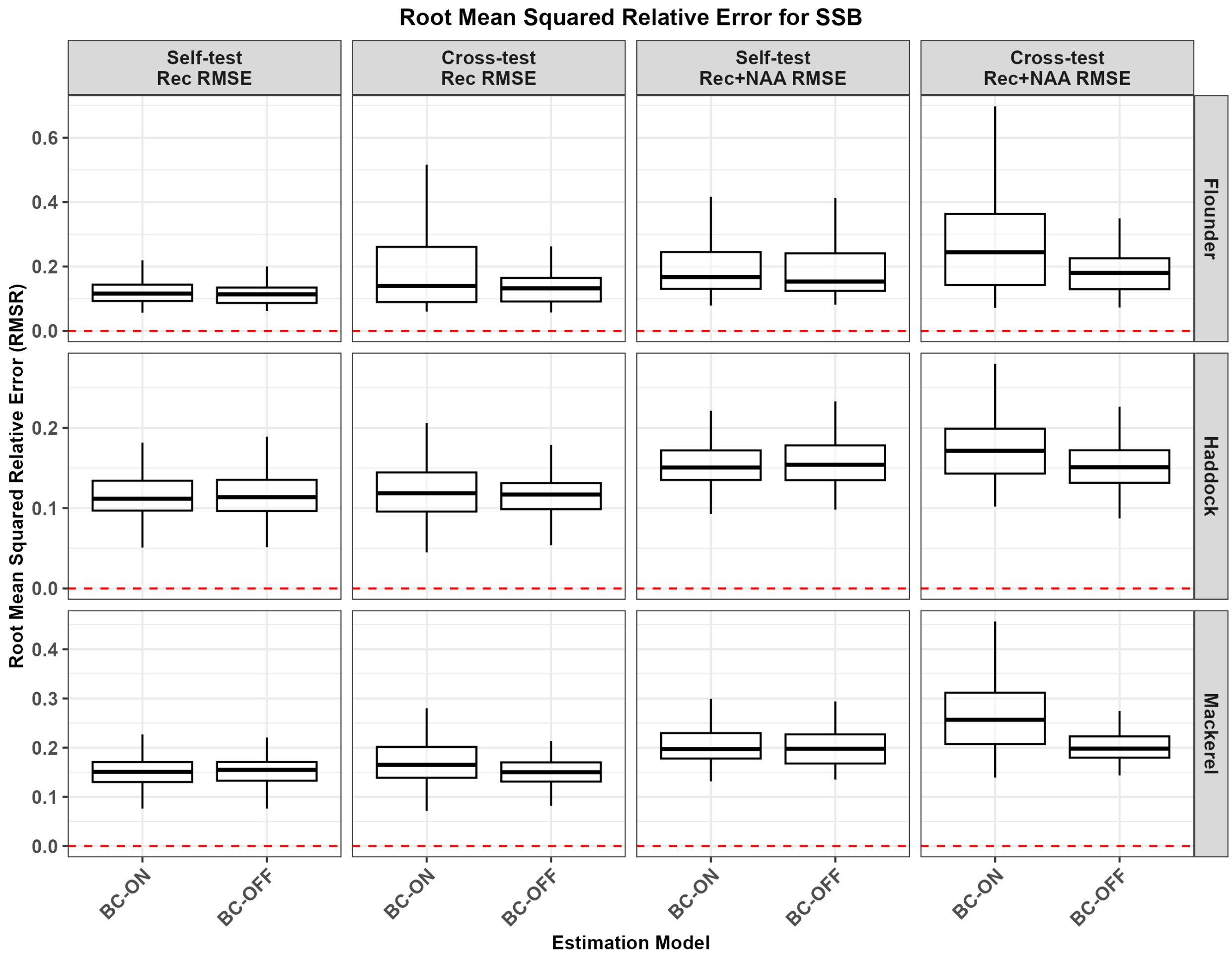

Figure. S5. Root mean squared relative error (RMSR) of SSB calculated for self-tests and cross-tests. "Rec RE" and "Rec+NAA RE" in the top facet indicate operating models (OMs) with only recruitment random effects and both recruitment and NAA random effects, respectively.

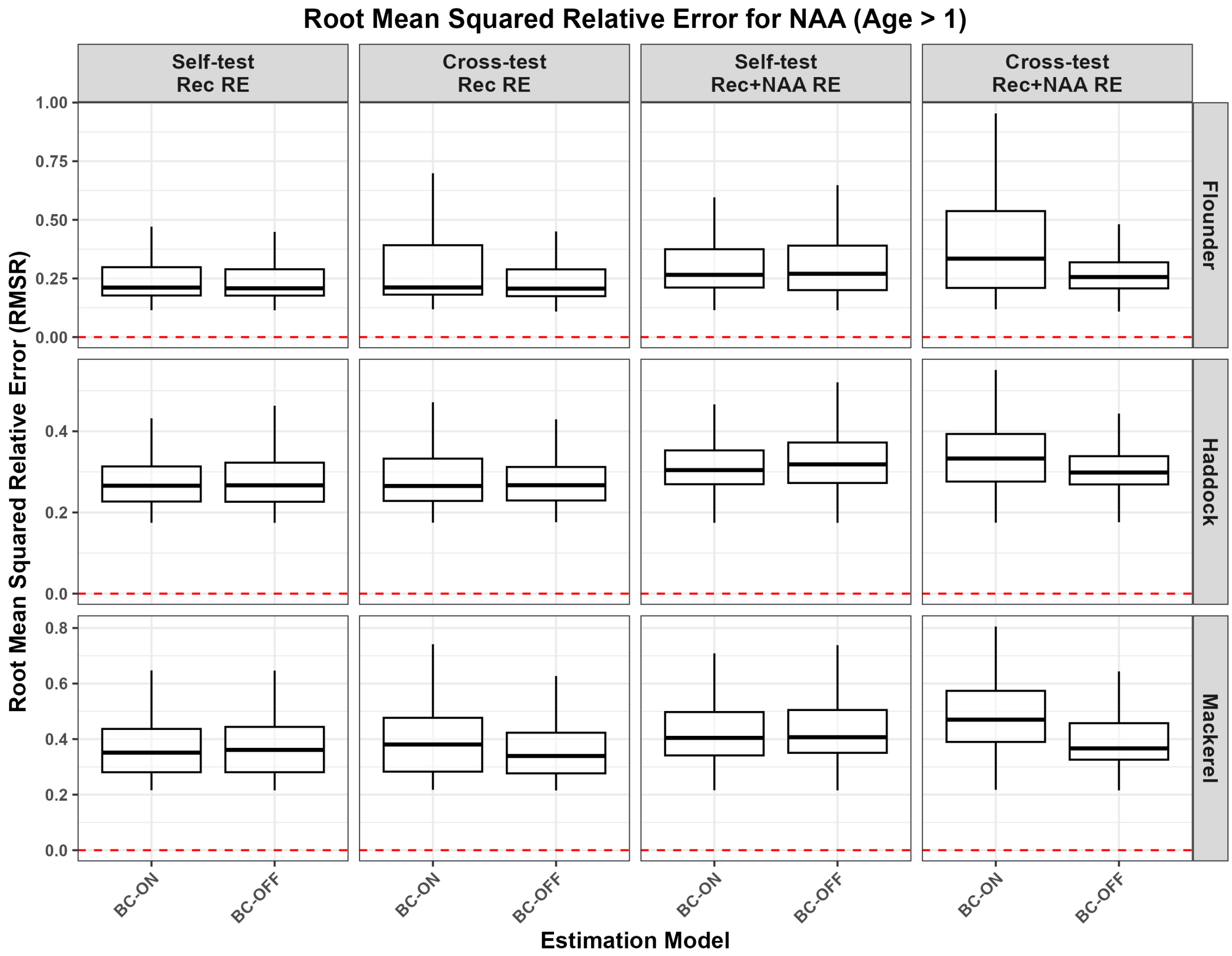

Figure. S6. Root mean squared relative error (RMSR) of NAA calculated for self-tests and cross-tests. “Rec RE” and “Rec+NAA RE” in the top facet indicate operating models (OMs) with only recruitment random effects and both recruitment and NAA random effects, respectively.

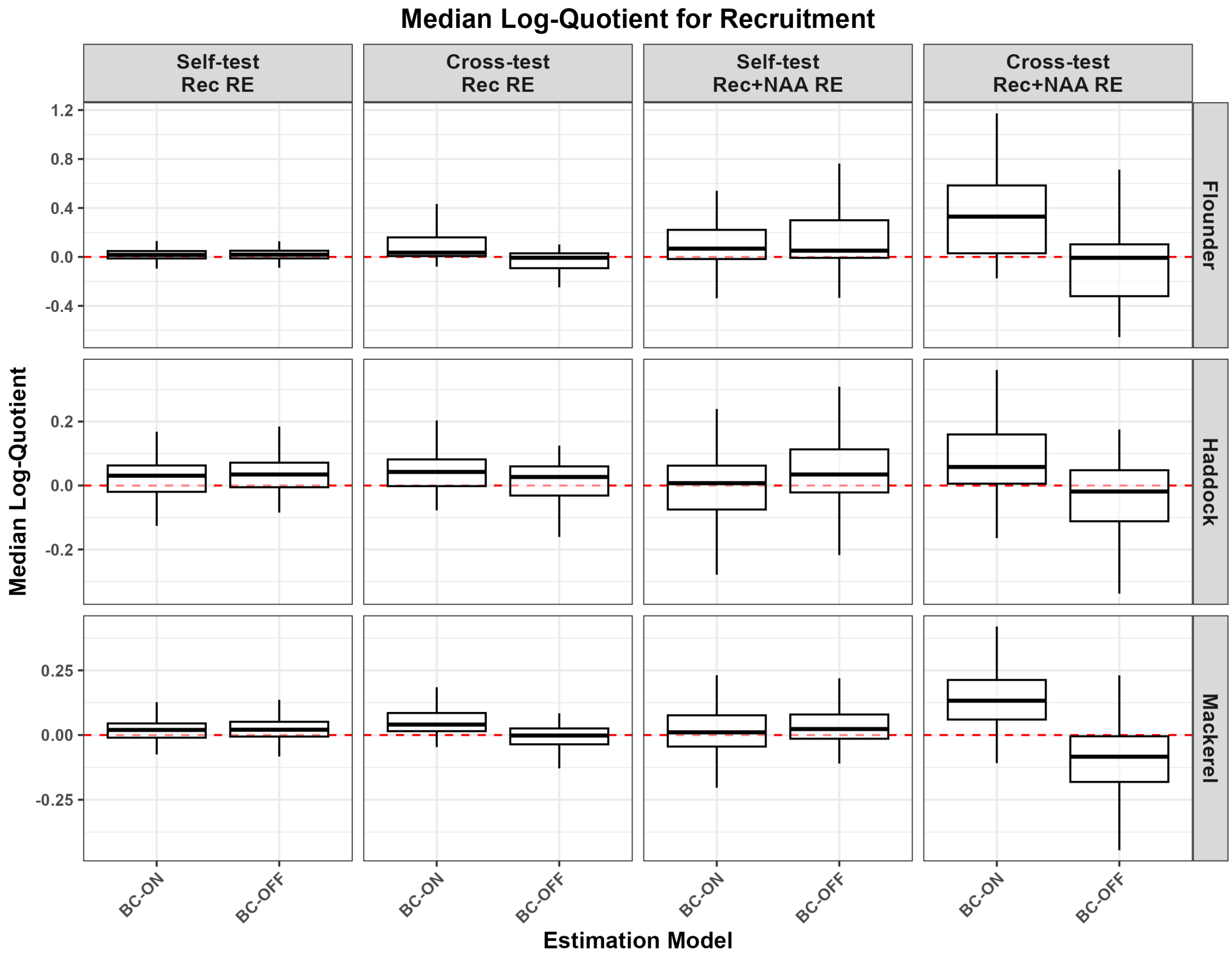

Figure. S7. Median log-quotient (MdLQ) of recruitment calculated for self-tests and cross-tests. "Rec RE" and "Rec+NAA RE" in the top facet indicate operating models (OMs) with only recruitment random effects and both recruitment and NAA random effects, respectively.

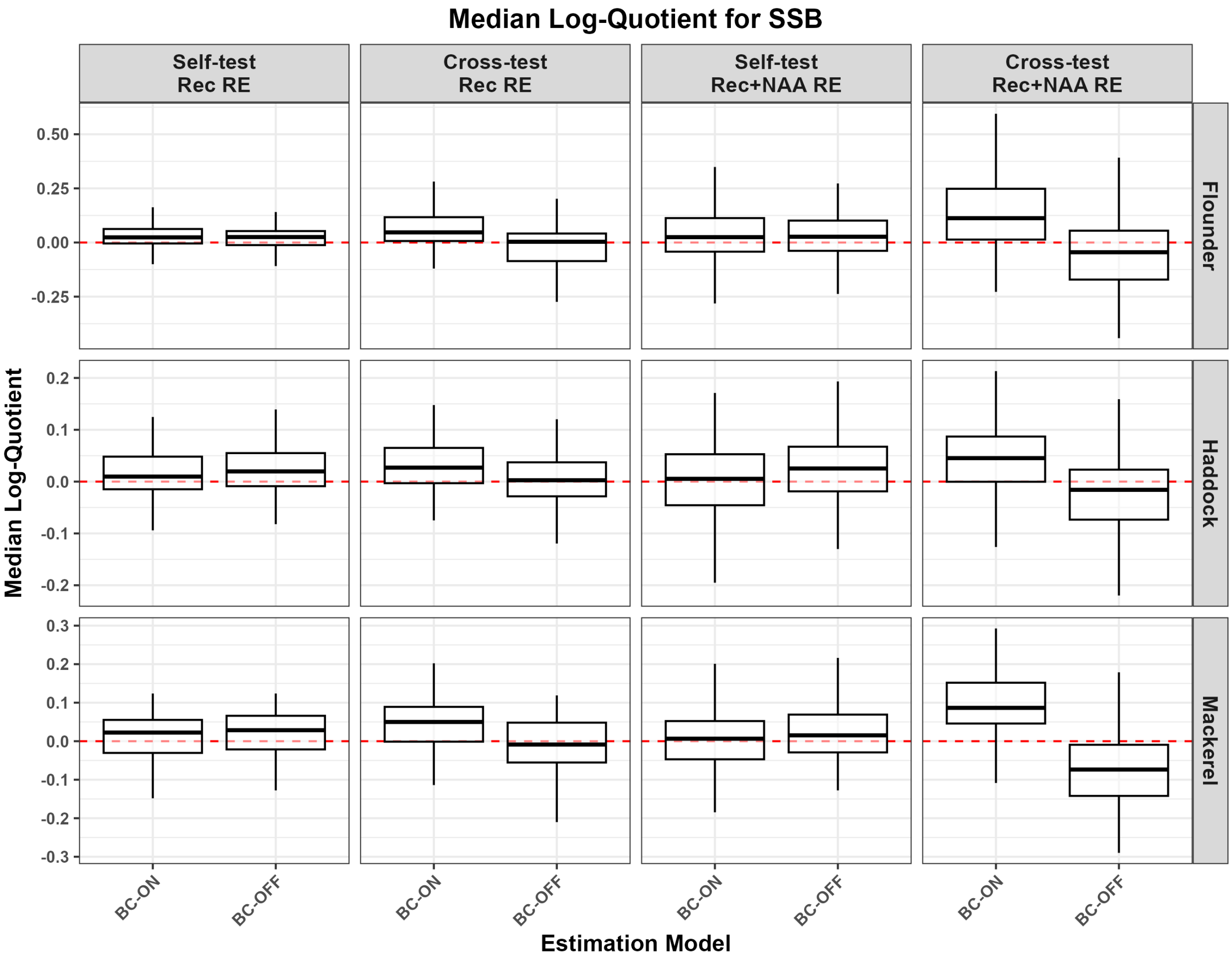

Figure. S8. Median log-quotient (MdLQ) of SSB calculated for self-tests and cross-tests. “Rec RE” and “Rec+NAA RE” in the top facet indicate operating models (OMs) with only recruitment random effects and both recruitment and NAA random effects, respectively.

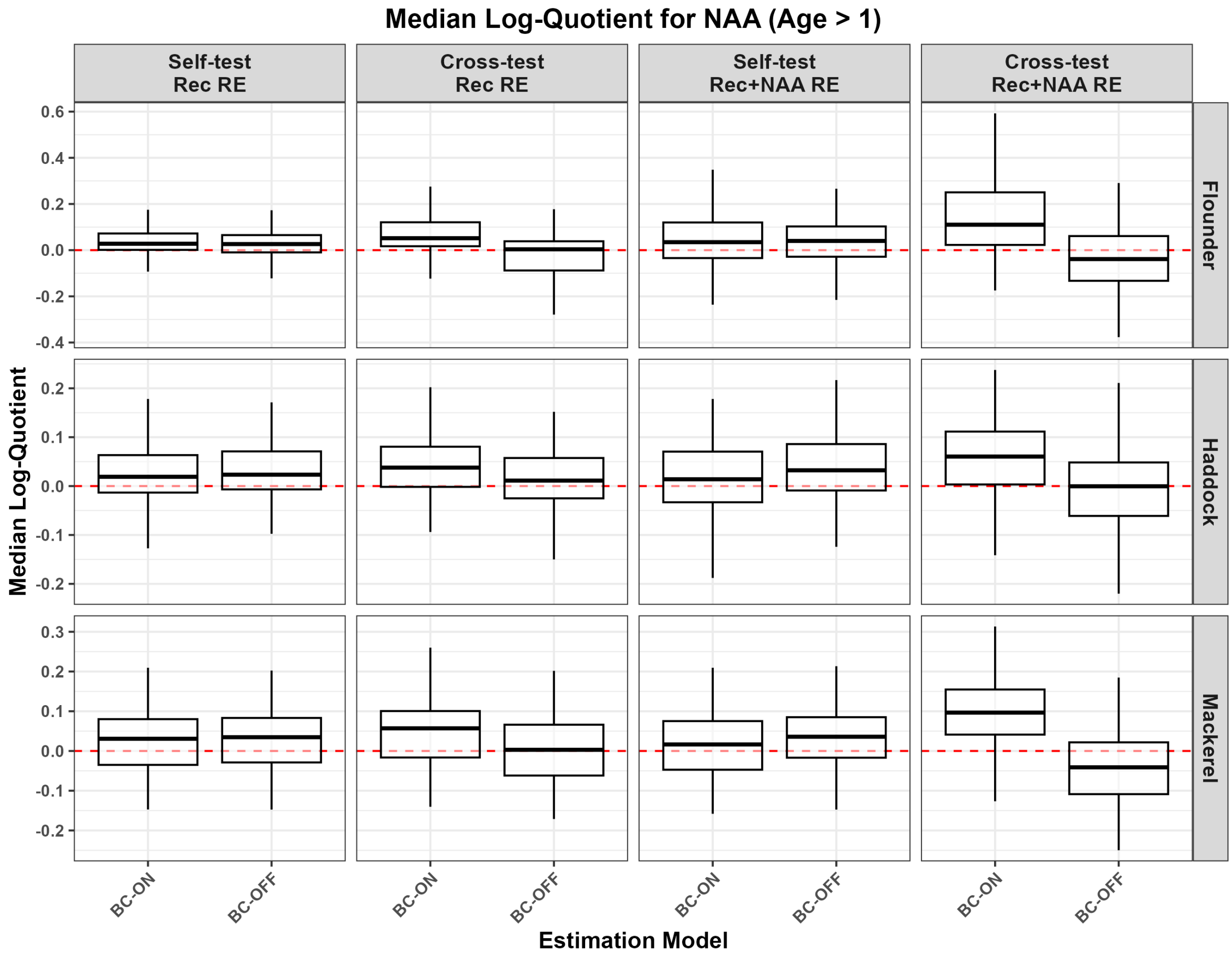

Figure. S9. Median log-quotient (MdLQ) of NAA calculated for self-tests and cross-tests. "Rec RE" and "Rec+NAA RE" in the top facet indicate operating models (OMs) with only recruitment random effects and both recruitment and NAA random effects, respectively.

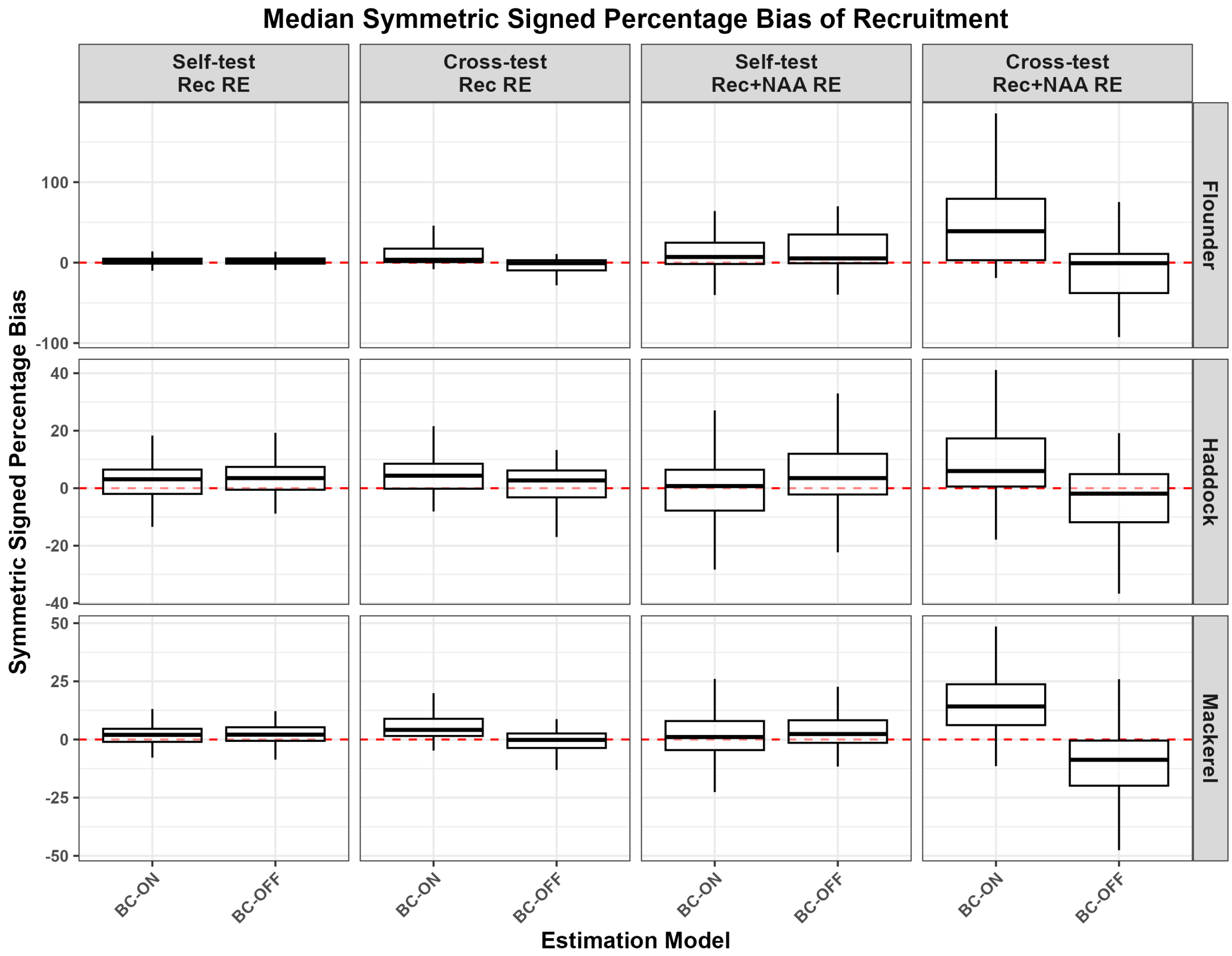

Figure. S10. Median symmetric signed percentage bias (SSPB) of recruitment calculated for self-tests and cross-tests. “Rec RE” and “Rec+NAA RE” in the top facet indicate operating models (OMs) with only recruitment random effects and both recruitment and NAA random effects, respectively.

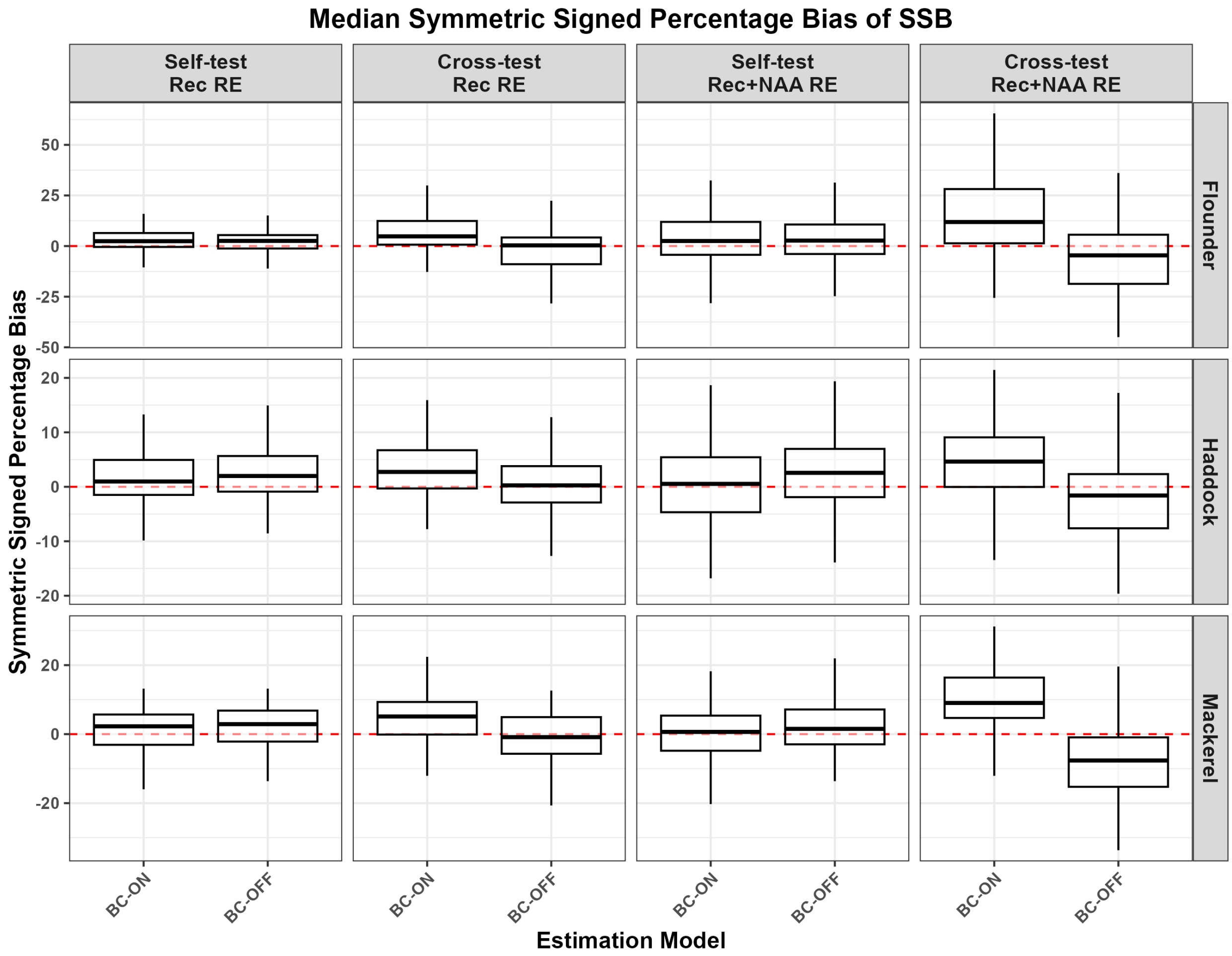

Figure. S11. Median symmetric signed percentage bias (SSPB) of SSB calculated for self-tests and cross-tests. “Rec RE” and “Rec+NAA RE” in the top facet indicate operating models (OMs) with only recruitment random effects and both recruitment and NAA random effects, respectively.

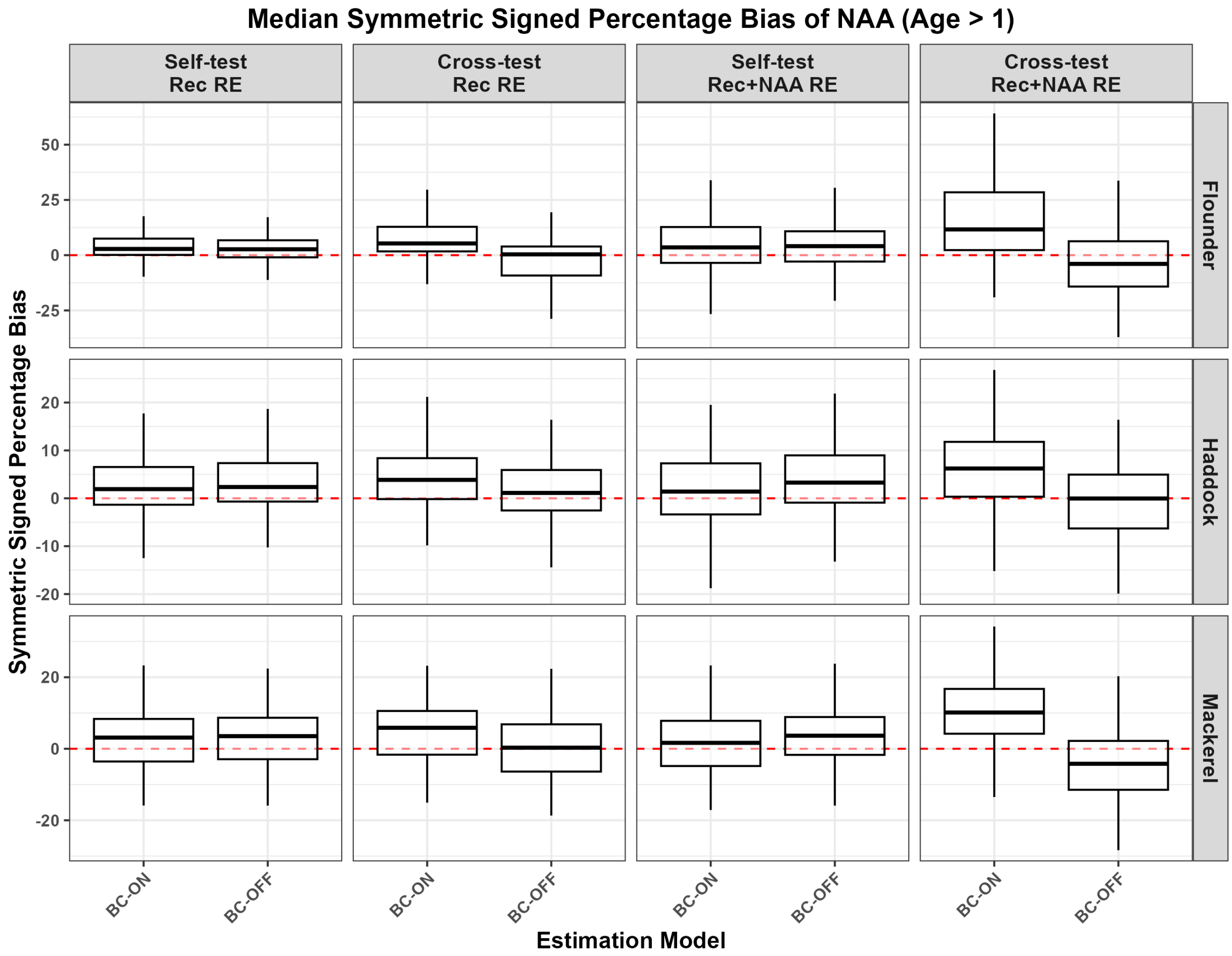

Figure. S12. Median symmetric signed percentage bias (SSPB) of NAA calculated for self-tests and cross-tests. “Rec RE” and “Rec+NAA RE” in the top facet indicate operating models (OMs) with only recruitment random effects and both recruitment and NAA random effects, respectively.

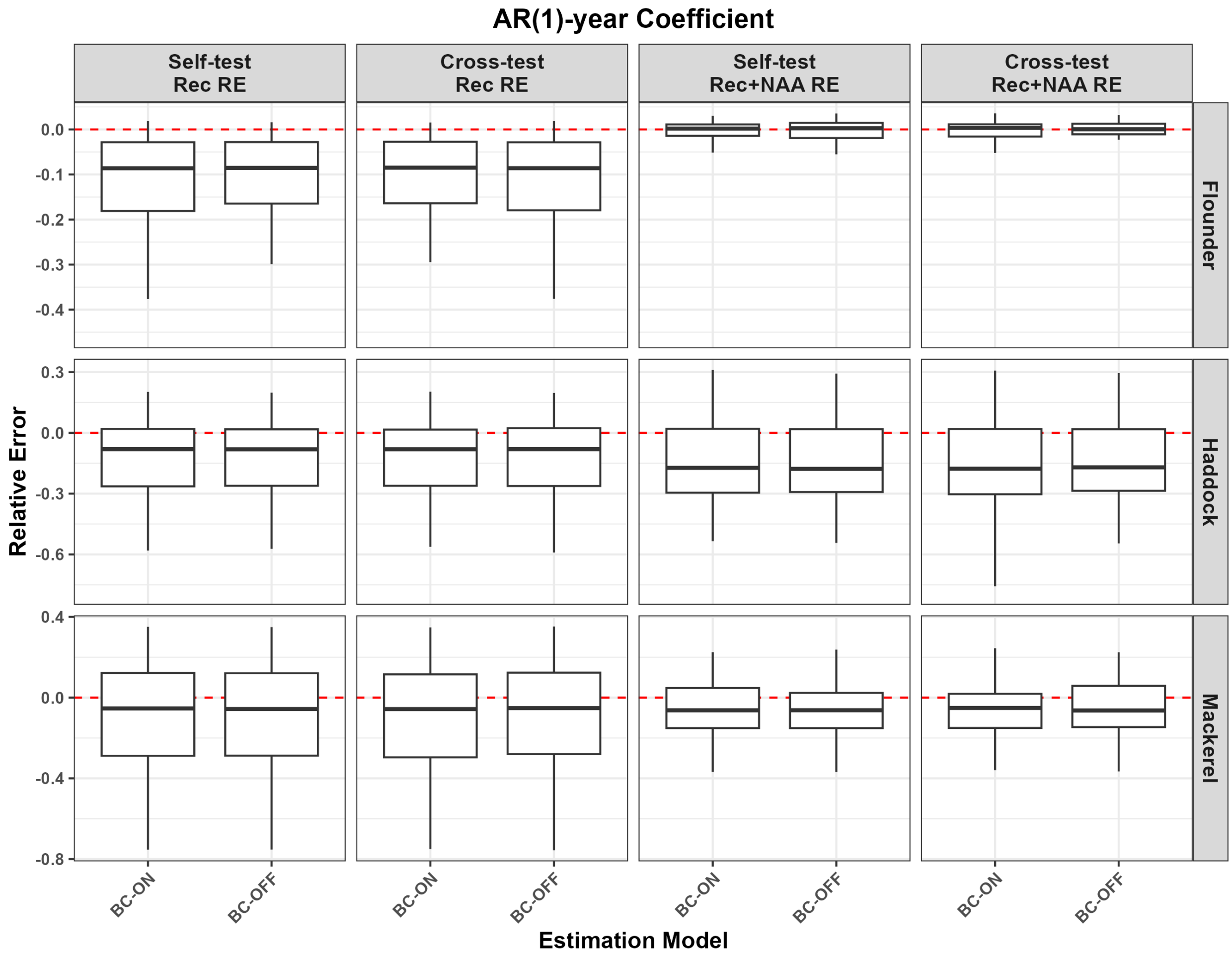

Figure. S13. Relative error of AR(1)-year coefficient calculated for self-tests and cross-tests. "Rec RE" and "Rec+NAA RE" in the top facet indicate operating models (OMs) with only recruitment random effects and both recruitment and NAA random effects, respectively.

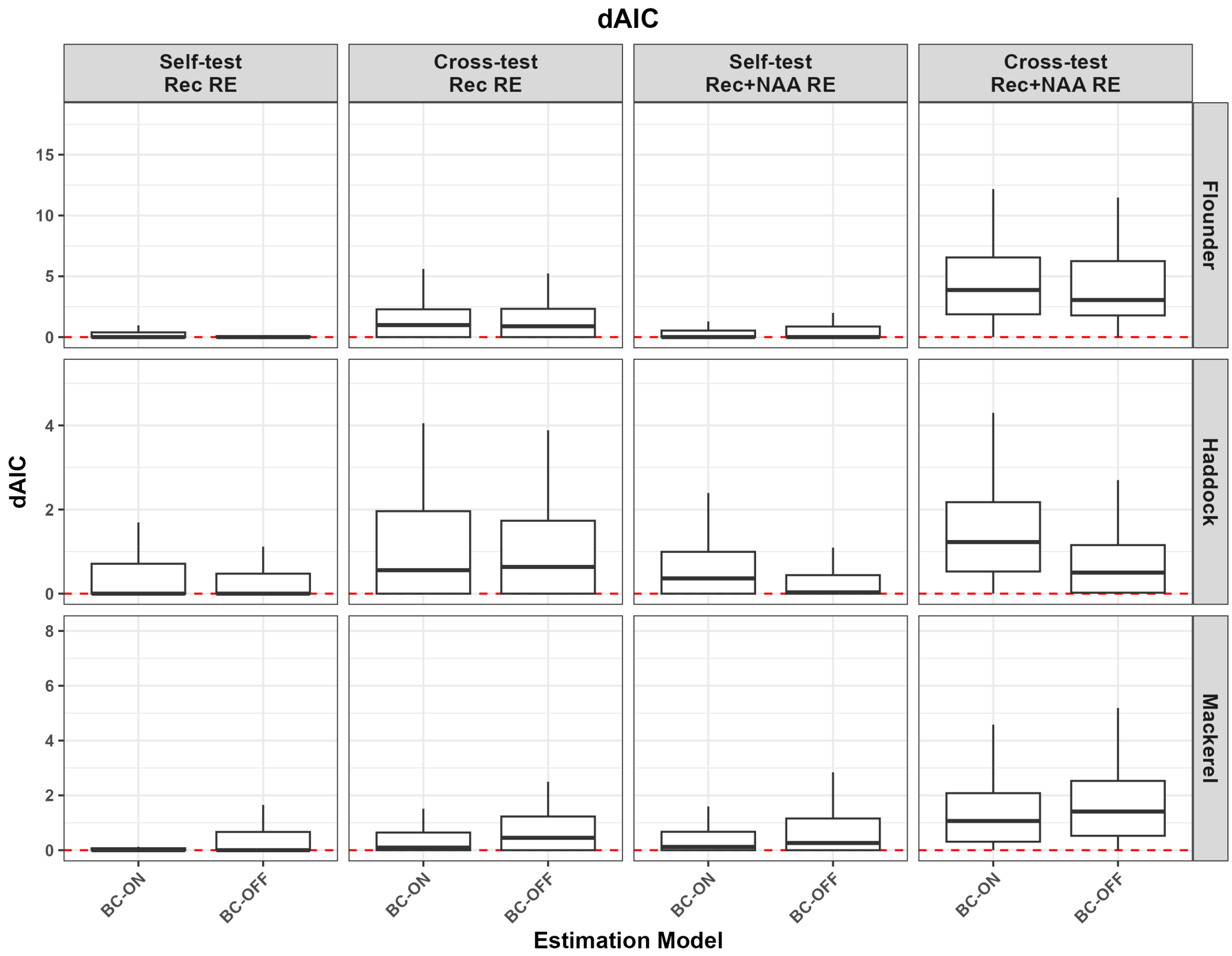

Figure. S14. dAIC calculated for self-tests and cross-tests. “Rec RE” and “Rec+NAA RE” in the top facet indicate operating models (OMs) with only recruitment random effects and both recruitment and NAA random effects, respectively.

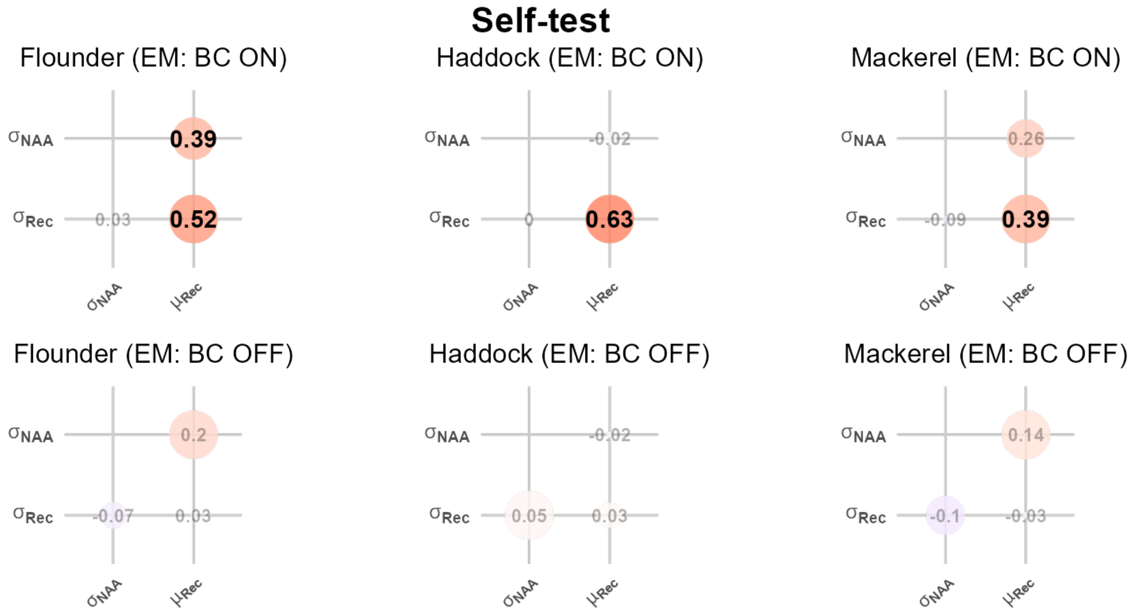

Figure. S15. Correlation plot for the OM with both recruitment and NAA treated as IID random effects. The correlations were calculated from self-tests, where the EM had the same bias correction as the operating model (OM). Correlations in **bold** indicate statistically significant values (p-value $<$ 0.05).

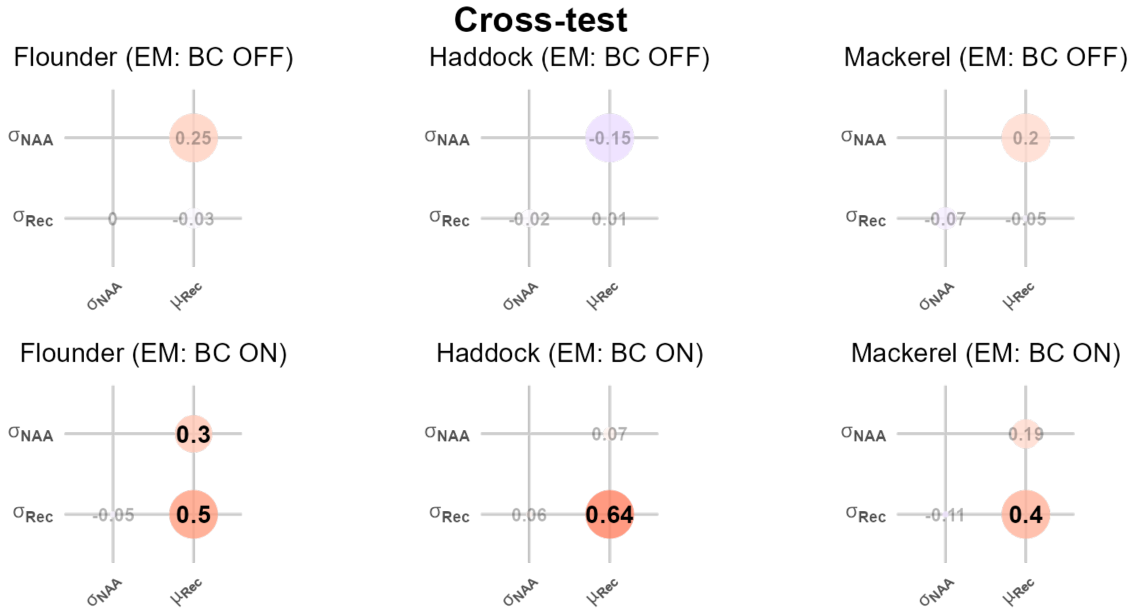

Figure. S16. Correlation plot for the OM with both recruitment and NAA treated as IID random effects. The correlations were calculated from cross-tests, where the EM had a different bias correction than the operating model (OM). Correlations in **bold** indicate statistically significant values (p-value $<$ 0.05).

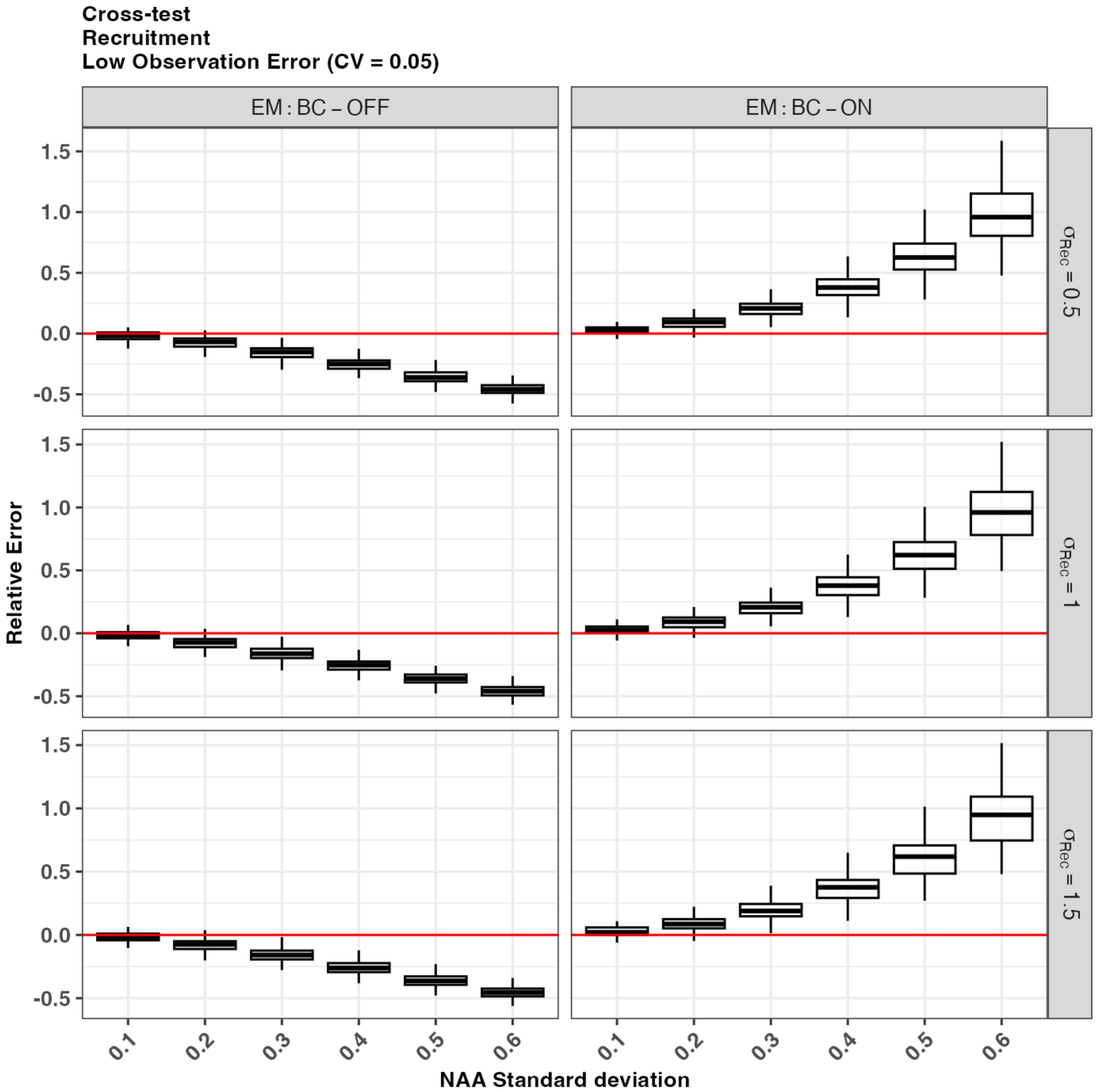

Figure. S17. Relative errors of recruitment estimates summarized from 50 realizations for each scenario. Two operating models (OMs) (with bias correction applied or omitted for both processes and observations) for Gulf of Maine (GoM) haddock with both recruitment and NAA IID random effects (see Table S2) were used to conduct simulation-estimation experiments. The study evaluated the effects of recruitment variability (σ_{Rec} = 0.5, 1, 1.5) and NAA variability (σ_{NAA} = 0.1, 0.2, ... 0.6) in a factorial design through self-tests and cross-tests. To isolate the impact of observation error, the coefficient of variation (CV) for observations was set to 0.05.

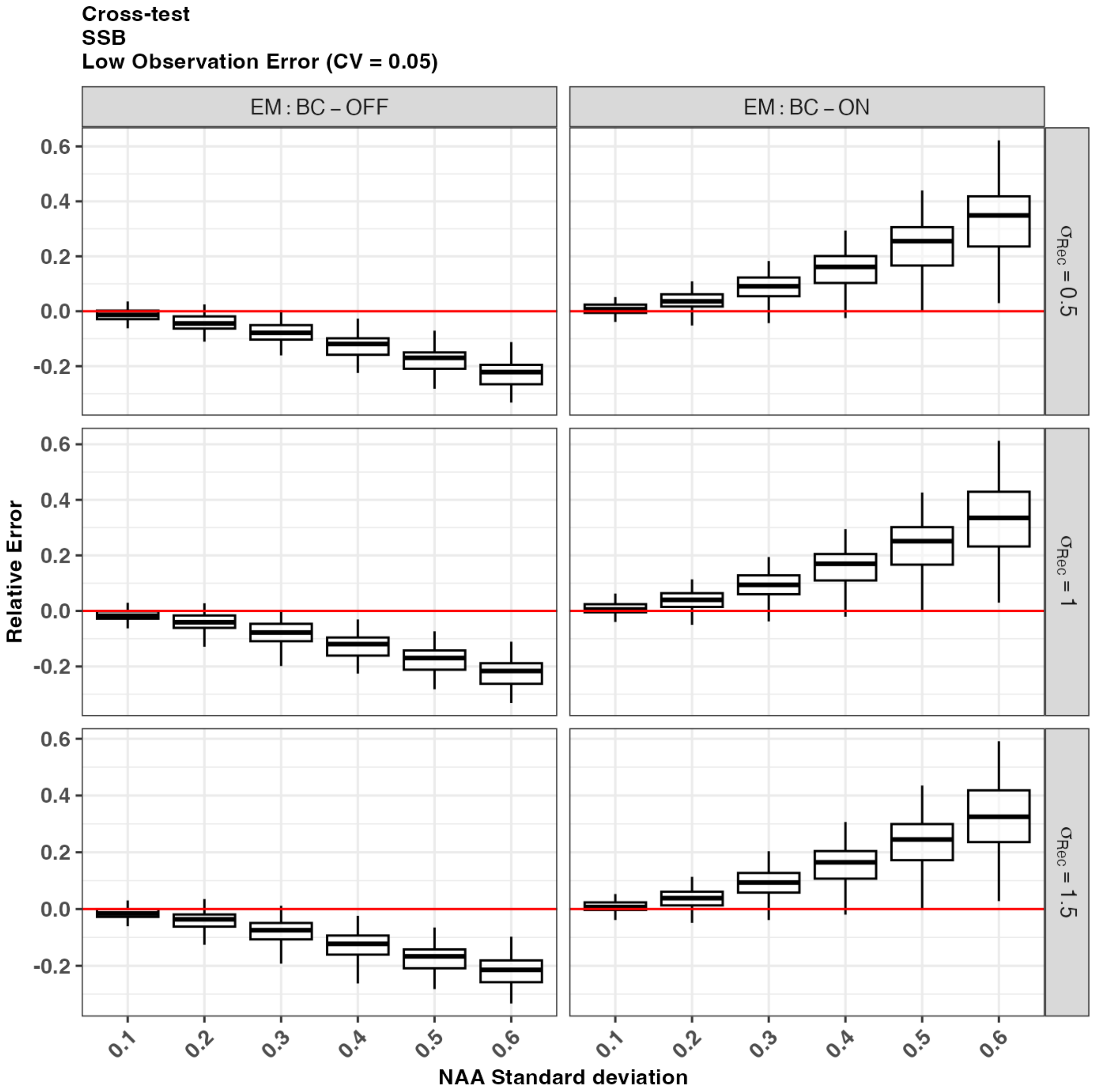

Figure. S18. Relative errors of recruitment estimates summarized from 50 realizations for each scenario. Two operating models (OMs) (with bias correction applied or omitted for both processes and observations) for Gulf of Maine (GoM) haddock with both recruitment and NAA IID random effects (see Table S2) were used to conduct simulation-estimation experiments. The study evaluated the effects of recruitment variability (σ_{Rec} = 0.5, 1, 1.5) and NAA variability (σ_{NAA} = 0.1, 0.2, ... 0.6) in a factorial design through self-tests and cross-tests. To isolate the impact of observation error, the coefficient of variation (CV) for observations was set to 0.05.

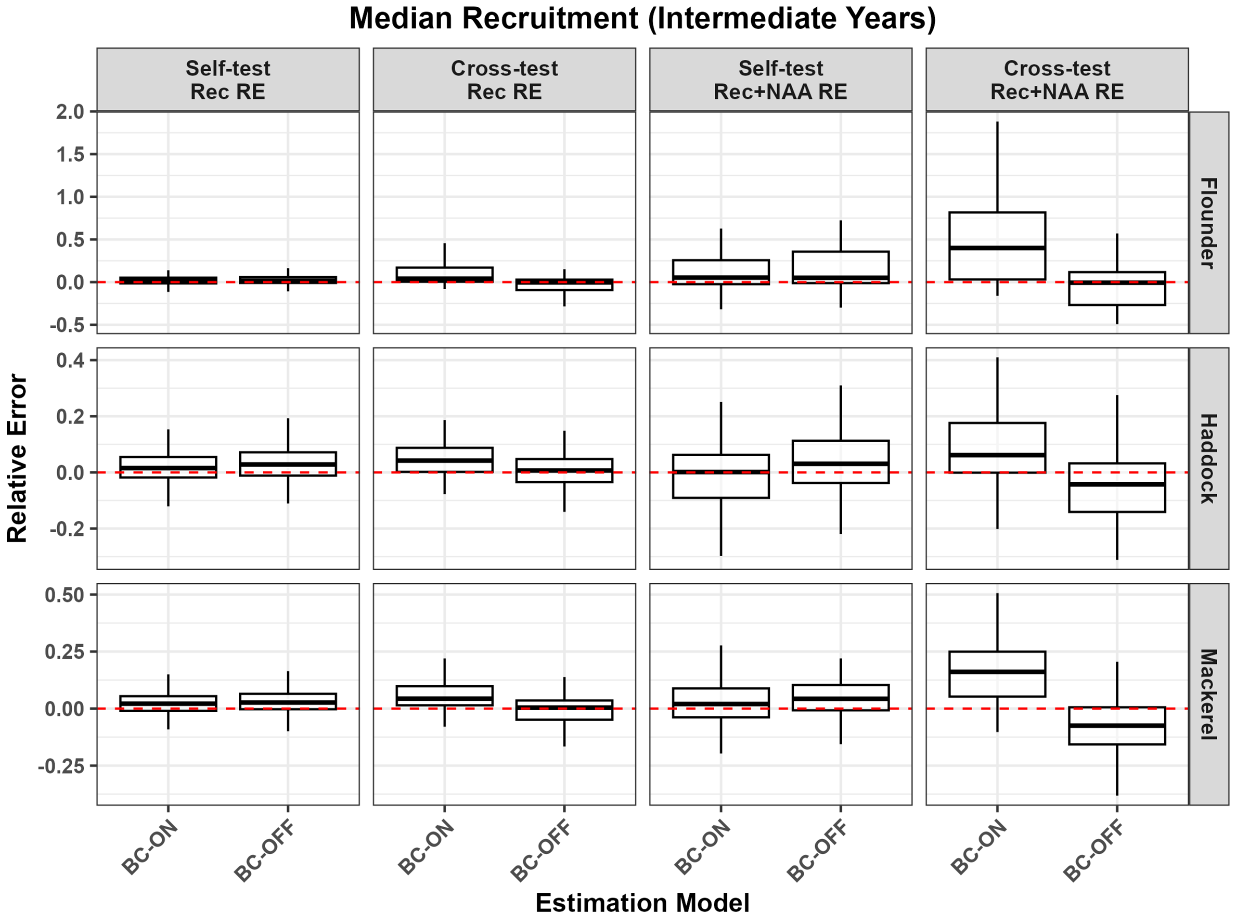

Figure. S19. Median relative eror of recruitment in the intermediate period (with first and last 10 years of estimates removed).

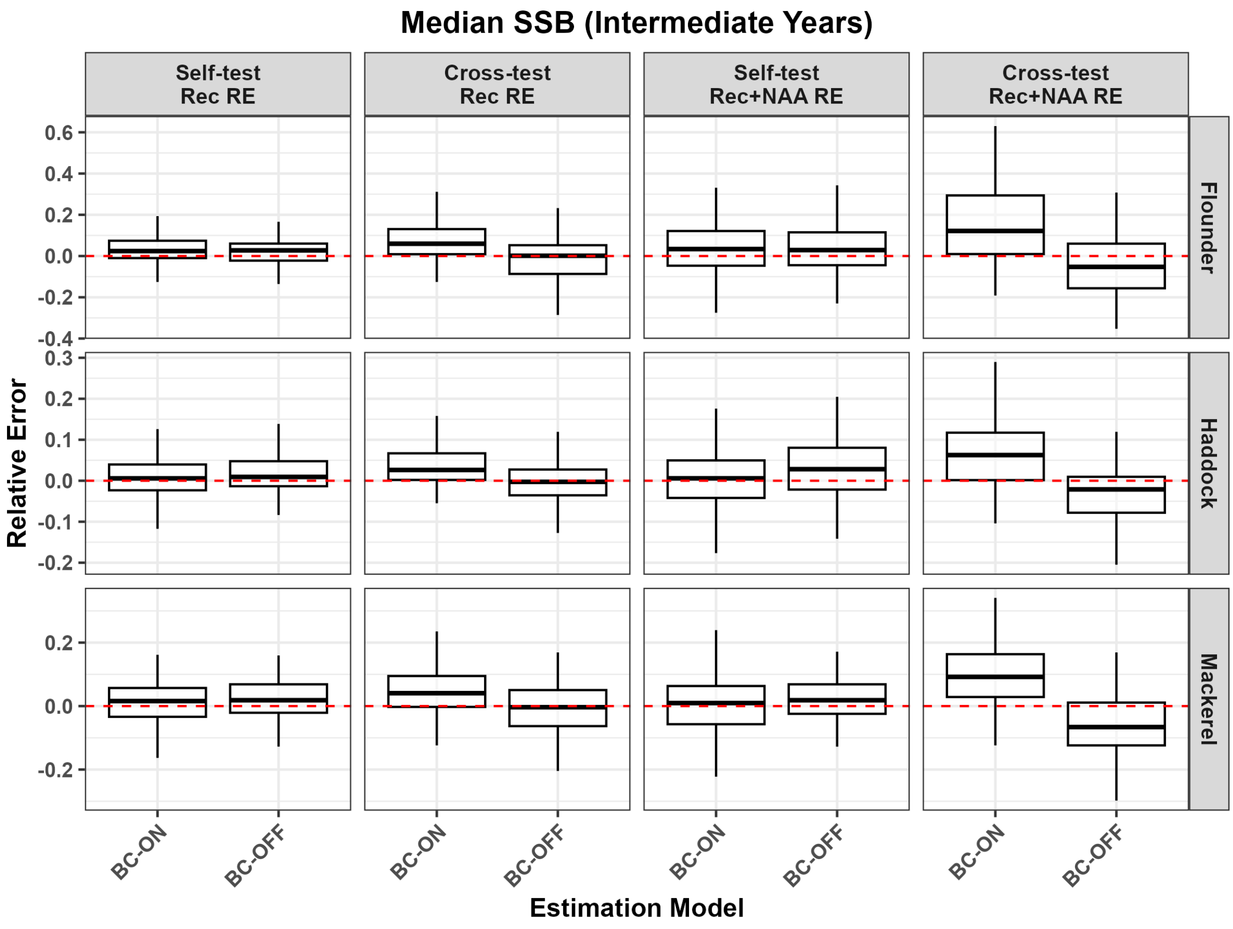

Figure. S20. Median relative eror of SSB in the intermediate period (with first and last 10 years of estiamtes removed).

References

Aldrin, M., Tvete, I. F., Aanes, S., and Subbey, S. 2020. The specification of the data model part in the sam model matters. Fisheries Research **229**: 105585.

Cadigan, N. G. 2015. A state-space stock assessment model for northern cod, including under-reported catches and variable natural mortality rates. Canadian Journal of Fisheries and Aquatic Sciences **73**(2): 296–308.

Deroba, J., Butterworth, D. S., Methot Jr, R., De Oliveira, J., Fernandez, C., Nielsen, A., Cadrin, S., Dickey-Collas, M., Legault, C., Ianelli, J., et al. 2015. Simulation testing the robustness of stock assessment models to error: some results from the ices strategic initiative on stock assessment methods. ICES Journal of Marine Science **72**(1): 19–30.

Deroba, J. J. and Miller, T. J. 2016. Correct in theory but wrong in practice: Bias caused by using a lognormal distribution to penalize annual recruitments in fish stock assessment models. Fisheries Research **176**: 86–93.

Fisch, N., Shertzer, K., Camp, E., Maunder, M., and Ahrens, R. 2023. Process and sampling variance within fisheries stock assessment models: estimability, likelihood choice, and the consequences of incorrect specification. ICES Journal of Marine Science **80**(8): 2125–2149.

Li, C., Deroba, J. J., Miller, T. J., Legault, C. M., and Perretti, C. T. 2024. An evaluation of common stock assessment diagnostic tools for choosing among state-space models with multiple random effects processes. Fisheries Research **273**: 106968.

Liljestrand, E. M., Bence, J. R., and Deroba, J. J. 2024. The effect of process variability and data quality on performance of a state-space stock assessment model. Fisheries Research **275**: 107023.

Maunder, M. N. and Thorson, J. T. 2019. Modeling temporal variation in recruitment in fisheries stock assessment: a review of theory and practice. Fisheries Research **217**: 71–86.

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.

Morley, S. K., Brito, T. V., and Welling, D. T. 2018. Measures of model performance based on the log accuracy ratio. Space Weather **16**(1): 69–88.

NEFMC Scientific and Statistical Committee. 2023. Transboundary resources assessment committee (trac) assessments for eastern georges bank cod and haddock, and georges bank yellowtail flounder. Technical report, New England Fishery Management Council, Newburyport, MA, meeting Report, September 12-14, 2023.

Nielsen, A. and Berg, C. W. 2014. Estimation of time-varying selectivity in stock assessments using state-space models. Fisheries Research **158**: 96–101.

Northeast Fisheries Science Center. 2019. 65th northeast regional stock assessment workshop (saw 65) assessment report. NEFSC Reference Document 19-01, Northeast Fisheries Science Center, Woods Hole, MA.

Northeast Fisheries Science Center. 2021. 2021 management track assessments: Atlantic mackerel. Technical report, Mid-Atlantic Fishery Management Council, Scientific and Statistical Committee, contained within the Report of the 2021 Management Track Assessment Peer Review Panel.

Stock, B. C. and Miller, T. J. 2021. The woods hole assessment model (wham): a general state-space assessment framework that incorporates time-and age-varying processes via random effects and links to environmental covariates. Fisheries Research **240**: 105967.

Thorson, J. T. 2019. Perspective: Let's simplify stock assessment by replacing tuning algorithms with statistics. Fisheries Research **217**: 133–139.

Thorson, J. T., Hicks, A. C., and Methot, R. D. 2015. Random effect estimation of time-varying factors in stock synthesis. ICES Journal of Marine Science **72**(1): 178–185.

Thorson, J. T. and Kristensen, K. 2016. Implementing a generic method for bias correction in statistical models using random effects, with spatial and population dynamics examples. Fisheries Research **175**: 66–74.

Trijoulet, V., Fay, G., and Miller, T. J. 2020. Performance of a state-space multispecies model: What are the consequences of ignoring predation and process errors in stock assessments? Journal of Applied Ecology **57**(1): 121–135.