

# Factors affecting reliability of state-space age-structured assessment models

Timothy J. Miller<sup>1,2</sup>      Greg Britten<sup>3</sup>      Elizabeth N. Brooks<sup>2</sup>

Gavin Fay<sup>4</sup>      Alex Hansell<sup>2</sup>      Christopher M. Legault<sup>2</sup>

Brandon Muffley<sup>5</sup>      Brian C. Stock<sup>6</sup>      John Wiedenmann<sup>7</sup>

12 February, 2025

<sup>1</sup>corresponding author: [timothy.j.miller@noaa.gov](mailto:timothy.j.miller@noaa.gov)

<sup>2</sup>Northeast Fisheries Science Center, Woods Hole Laboratory, 166 Water Street, Woods  
Hole, MA 02543 USA

<sup>3</sup>Woods Hole Oceanographic Institution

<sup>4</sup>SMAST

<sup>5</sup>Institute of Marine Research

<sup>6</sup>Mid-Atlantic Fisheries Management Council

<sup>7</sup>Rutgers University

## Abstract

Random effects can be included in state-space assessment models on many processes and in many ways, and guidance is needed on statistical reliability and model selection criteria. We simulated 72 operating models with varying fishing history, observation error uncertainty, and process error magnitude, correlation, and source (recruitment, survival, fishery selectivity, catchability, and natural mortality). We fit estimating models with different assumptions on the process error source, whether median natural mortality was estimated, and whether a stock-recruit relationship was estimated.

Models that assumed the correct process error source had high convergence and low bias. Bias was also low under most process error assumptions when there was contrast in fishing pressure. Stock-recruitment parameters were only reliably estimated in ideal situations. Marginal AIC most accurately distinguished process errors on recruitment, survival, and selectivity, as well as larger magnitude process errors of other types. Retrospective patterns were generally weak except for recruitment when observation error was high, even with the correct process error assumptions. When models did exhibit some retrospective pattern, estimating median natural mortality often removed it.

## Introduction

Application of state-space models in fisheries stock assessment and management has expanded dramatically within ICES, Canada, and the Northeast US (Nielsen and Berg 2014; Cadigan 2016; Pedersen and Berg 2017; Stock and Miller 2021). State-space approaches that use random effects to parameterize process errors is considered best practice and a requirement for the next generation of stock assessment models (Hoyle et al. 2022; Punt 2023).

Much is known about the reliability of state-space models that are linear or Gaussian (Aeberhard et al. 2018), but applications in fisheries management are nonlinear and typically include multiple types of observations with varying distributional assumptions. Furthermore, there is a wide range of potential random effects structures and model parameters that can be treated as random effects in assessment models. We know relatively little about the factors affecting statistical reliability of such models or the ability of information criteria to distinguish among such alternative structures.

Li et al. (2024) investigated some aspects of inferences for operating models with multiple sources of process error, but there are differences for this paper.

Here we conduct a simulation study with operating models (OMs) varying by degree of observation error uncertainty, source and variability of process error, and fishing history. The simulations from these OMs are fitted with estimation models (EMs) that make alternative assumptions for sources of process error, whether a stock-recruit model was estimated, and whether  $M$  is estimated. We evaluate whether AIC can correctly determine the correct source of process error and stock effects on recruitment. We also evaluate the degree of bias in the outputs of the assessment model that are important for management.

## Methods

We used the Woods Hole Assessment Model (WHAM) to configure operating and EMs in our simulation study (Miller and Stock 2020; Stock and Miller 2021). WHAM is an R package freely available as a github repository. For this study we used version 1.0.6.9000, commit 77bbd94. This package has also been used to configure operating and EMs for closed loop simulations evaluating index-based assessment methods (Legault et al. 2023) and is used for management of haddock, butterfish, American plaice, bluefish, Atlantic cod, black sea bass, and yellowtail flounder in the Northeast US.

We completed a simulation study with a number of OM s that can be categorized based on where process error random effects are assumed: abundance at age (R, R+S), natural mortality (R+M), fleet selectivity (R+Sel), or index catchability (R+q). For each OM assumption about variance of process errors and observations are required and the values we used were based on a review of the range of estimates from applications of WHAM in management of stocks of haddock, butterfish, and American plaice in the NE US.

In total, we configured 72 OM s with alternative assumptions about the source and variability of process errors, level of observation error in indices and age composition data, and contrast in fishing pressure over time. We fitted 20 EM s to observations from each of 100 simulations where process errors were also simulated. For R+M, R+Sel, and R+q OM s, we used unique seeds for each simulation, but we inadvertently used the same 100 seeds for all R and R+S OM s. Each of the EM s made alternative assumptions about the source of process errors and whether natural mortality (or the median for models with process error in natural mortality) was estimated and whether a Beverton-Holt stock recruit relationship was estimated within the EM. Details of each of the operating and EM s are described below.

We did not use the log-normal bias-correction feature for process errors or observations described by (Stock and Miller 2021) for operating and EM s (Li et al. In review). Simulations and model fitting were all carried out on the University of Massachusetts Green

81 High-Performance Computing Cluster. All code we used to perform the simulation study  
 82 and summarize results can be found at [https://github.com/timjmiller/SSRTWG/tree/main/](https://github.com/timjmiller/SSRTWG/tree/main/Project_0/code)  
 83 [Project\\_0/code](https://github.com/timjmiller/SSRTWG/tree/main/Project_0/code).

## 84 **Operating models**

### 85 **Population**

86 The population consists of 10 age classes: ages 1 to 10+ and we assume spawning occurs  
 87 each year 1/4 of the way through the year. The maturity at age was a logistic curve with  
 88  $a_{50} = 2.89$  and slope = 0.88 (Figure S1, top left).

89 Weight at age was generated with a von Bertalanffy growth function

$$L_a = L_\infty \left(1 - e^{-k(a-t_0)}\right)$$

90 where  $t_0 = 0$ ,  $L_\infty = 85$ , and  $k = 0.3$ , and a L-W relationship such that

$$W_a = \theta_1 L_a^{\theta_2}$$

91 where  $\theta_1 = e^{-12.1}$  and  $\theta_2 = 3.2$  (Figure S1, top right).

92 We assumed a Beverton-Holt stock recruit function with constant pre-recruit mortality pa-  
 93 rameters for all OMs. All post-recruit productivity components are constant in the NAA  
 94 and survey catchability process error OMs. Therefore steepness and unfished recruitment  
 95 are also constant over the time period for those OMs (Miller and Brooks 2021). We specified  
 96 unfished recruitment =  $R_0 = e^{10}$  and  $F_{\text{MSY}} = F_{40\%} = 0.348$  equated to a steepness of 0.69  
 97 and  $a = 0.60$  and  $b = 2.4 \times 10^{-5}$  for the

$$N_{1,y} = \frac{a\text{SSB}_{y-1}}{1 + b\text{SSB}_{y-1}}$$

98 Beverton-Holt parameterization (Figure S1, bottom right). For OMs without process errors  
 99 on natural mortality we assumed the rate was assumed 0.2. For OMs with process errors on  
 100 natural mortality the median natural mortality rate was 0.2.

101 We used two fishing scenarios for OMs. In the first scenario, the stock experiences overfishing  
 102 at  $2.5F_{\text{MSY}}$  for the first 20 years and fishing at  $F_{\text{MSY}}$  for the last 20 years (denoted  $2.5F_{\text{MSY}} \rightarrow$   
 103  $F_{\text{MSY}}$ ). In the second scenario, the stock is fished at  $F_{\text{MSY}}$  for the entire time period. The  
 104 magnitude of the overfishing assumptions is based on average estimates of overfishing for NE  
 105 groundfish stocks from (Wiedenmann et al. 2019). Legault et al. (2023) also used similar  
 106 approaches to defining fishing mortality histories for OMs.

107 We specified initial population abundance at age at the equilibrium distribution fishing at  
 108 either  $F = 2.5 \times F_{\text{MSY}}$  or  $F = F_{\text{MSY}}$  for the two alternative fishing histories. This implies  
 109 that, for a deterministic model, the abundance at age would not change from the first year  
 110 to the next.

111 For OMs with time-varying random effects for M, steepness is not constant, but we used  
 112 the same alpha and beta parameters as other OMs this equates to a steepness and R0 at  
 113 the median of the time series process for M. For OMs with time-varying random effects for  
 114 fishery selectivity,  $F_{\text{MSY}}$  is also not constant however we use the same F history as other  
 115 OMs which corresponds to  $F_{\text{msy}}$  at the mean selectivity parameters.

## 116 **Fleets**

117 We assumed a single fleet operating year round for catch observations with logistic selectivity  
 118 for the fleet with  $a_{50} = 5$  and slope = 1 (Figure S1, bottom left). This selectivity is was  
 119 used to define  $F_{\text{MSY}}$  for the Beverton-Holt stock recruitment parameters above. We assumed  
 120 a logistic-normal distribution with no correlation on the multivariate normal scale for the  
 121 age-composition observations for the fleet.

## Indices

Two time series of surveys are assumed and observed in numbers rather than biomass for the entire 40 year period with one occurring in the spring (0.25 of each year) and one in the fall (0.75 of each year). Catchability of both surveys are assumed to be 0.1. Like the fishing fleet, we assumed logistic selectivity for both indices with  $a_{50} = 5$  and slope = 1 and a logistic-normal distribution with no correlation on the multivariate normal scale for the age-composition observations.

## Observation Uncertainty

Standard deviation for log-aggregate catch was 0.1. There were two levels of observation error variance for indices and age composition for both indices and fleet catch. A low uncertainty specification assumed standard deviation of both series of log-aggregate index observations was 0.1 and the standard deviation of the logistic-normal for age composition observations was 0.3 In the high uncertainty specification the standard deviation for log-aggregate indices was 0.4 and that for the age composition observations was 1.5. For all EMs, standard deviation for log-aggregate observations was assumed known whereas that for the logistic-normal age composition observations was estimated.

## Operating models with random effects on numbers at age

For operating models with random effects on recruitment and(or) survival (R, R+S) we assumed marginal standard deviations for recruitment of  $\sigma_R \in \{0.5, 1.5\}$  and marginal standard deviations for older age classes of  $\sigma_{2+} \in \{0, 0.25, 0.5\}$ . The full factorial combination of these process error assumptions and fishing history (2 levels) and observation error (2 levels) scenarios described above results in 24 different R ( $\sigma_{2+} = 0$ ) and R+S operating models (Table S1).

## Operating models with random effects on natural mortality

All R+M OMs treat natural mortality constant across age, but with annually varying random effects. WHAM treats natural mortality as a log-transformed parameter

$$\log M_{y,a} = \mu_M + \epsilon_{M,y}$$

that is a linear combination of a mean log-natural mortality parameter that is constant across ages  $\mu_M = \log(0.2)$  and any annual random effects marginally distributed as  $\epsilon_{M,y} \sim N(0, \sigma_M^2)$ . Uncorrelated random effects were also included on recruitment with  $\sigma_R = 0.5$  (hence, R+M). The marginal standard deviations we assumed for log natural mortality random effects were  $\sigma_M \in \{0.1, 0.5\}$  and AR1 autocorrelation parameters of  $\rho_M \in \{0, 0.9\}$ . The full factorial combination of these process error assumptions and fishing history (2 levels) and observation error (2 levels) scenarios described above results in 16 different R+M OMs (Table S2).

## Operating models with random effects on fleet selectivity

MORE SPECIFICS about correlation of random effects? Both selectivity pars? just correlated by year? WHAM treats selectivity parameter  $s$  as a logit-transformed parameter

$$\log \left( \frac{p_{s,y} - l_s}{u_s - p_{s,y}} \right) = \mu_s + \epsilon_{s,y}$$

that is a linear combination of a mean  $\mu_s$  and any annual random effects marginally distributed as  $\epsilon_{s,y} \sim N(0, \sigma_s^2)$  where the lower and upper bounds of the parameter ( $l_s$  and  $u_s$ ) can be specified by the user. All selectivity parameters are either  $a_50$  or slope parameters and we assume bounds of 0 and 10 for all selectivity parameters for all operating and EMs. The marginal standard deviations we assumed for logit scale random effects were  $\sigma_s \in \{0.1, 0.5\}$  and AR1 autocorrelation parameters of  $\rho_s \in \{0, 0.9\}$ . Like R+M OMs, the



full factorial combination of these process error assumptions and fishing history (2 levels) and observation error (2 levels) scenarios described above results in 16 different R+Sel OMs (Table S3).

## Operating models with random effects on index catchability

Like selectivity parameters, WHAM treats catchability for an index  $i$  as a logit-transformed parameter

$$\log\left(\frac{q_{i,y} - l_i}{u_i - q_{i,y}}\right) = \mu_i + \epsilon_{i,y}$$

that is a linear combination of a mean  $\mu_i$  and any annual random effects marginally distributed as  $\epsilon_{i,y} \sim N(0, \sigma_i^2)$  where the lower and upper bounds of the catchability ( $l_i$  and  $u_i$ ) can be specified by the user. Here we assume bounds of 0 and 1000 for all operating and EMs. For operating and EMs with process errors on catchability, the temporal variation is only assumed for the first index. The marginal standard deviations we assumed for logit scale random effects were  $\sigma_i \in \{0.1, 0.5\}$  and AR1 autocorrelation parameters of  $\rho_i \in \{0, 0.9\}$ . Like R+M and R+Sel OMs, the full factorial combination of these process error assumptions and fishing history (2 levels) and observation error (2 levels) scenarios described above results in 16 different R+q OMs (Table S4).

## Estimation models

For each data set simulated from an OM 20 EMs were fit. A total of 32 different EMs were fit across OMs where the subset of 20 depended on the source of process error in the OM (Table S5). The EMs make different assumptions about the source of process error (R+S, R+M, R+Sel, R+q) and whether there is temporal autocorrelation or not, and whether a Beverton-Holt stock recruit function is estimated and the natural mortality rate ( $\mu_M$ , the constant or mean on log scale for R+M EMs) is estimated. For simplicity we refer to the derived estimate  $e^{\mu_M}$  as the median natural mortality rate regardless of whether natural

mortality random effects are estimated in the EM. The first 20 EMs in Table S5 were fit to simulate data sets from R and R+S OM. EMs 5 to 24 in Table S5 were fit to simulate data sets from R+M OM. EMs 5 to 20 and 25-28 in Table S5 were fit to simulate data sets from R+Sel OM. Finally, EMs 5 to 20 and 29-32 in Table S5 were fit to simulate data sets from R+q OM. The maturity at age, weight at age for catch and SSB, and observation error variance of aggregate catch and indices were all assumed known at the true values.

## Measures of reliability

The first measure of reliability we investigated was frequency of convergence when fitting each EM to the simulated data sets. There are various ways to assess convergence of the fit (e.g., Carvalho et al. 2021), but given the importance of estimates of uncertainty when using assessment models in management, we estimated probability of convergence as measured by occurrence of a positive-definite hessian matrix at the optimized negative log-likelihood that could be inverted. However, we also provided results in the Supplementary Materials for the maximum of the absolute values among all gradients for all fits of a given EM to all simulated data sets from a given OM that produced hessian-based standard errors for all estimated fixed effects. This provides an indication of how poor the calculated gradients can be, but still presumably converged adequately enough for parameter inferences.

## AIC for model selection

We estimated the probability of selection of each process error model structure (R, R+S, R+M, R+Sel, R+q) using marginal AIC. For a given operating model, we compared AIC for EMs that all made the same assumptions about median natural mortality (known or estimated) and stock recruitment model (Beverton-Holt or none).

We also estimated the probability of correctly selecting between models with Beverton-Holt stock recruit function assumed and models without the stock-recruit function (null model).

We made these comparisons between models that otherwise assume the same process error structure as the operating model and both of the compared models either estimate median natural mortality or assume it is known. Contrast in fishing pressure and time series with recruitment at low stock size has been shown to improve estimability of stock-recruitment parameters (Magnusson and Hilborn 2007; Conn et al. 2010). Our preliminary inspections of the proportions of simulations where the correct recruitment model was chosen for a given set of OM factors (including contrast in fishing pressure) indicated generally poor performance of AIC. Therefore, we fit logistic regression models to the indicator of Beverton-Holt models having lower AIC as a function of the log-standard deviation of the true  $\log(\text{SSB})$  (similar to the log of the coefficient of variation for SSB) since simulations with realized SSB producing low and high recruitments would have larger variation in realized SSB.

All results only condition on whether all of the compared estimating models completed the optimization process without failure. We did not condition on convergence as defined by a gradient threshold or invertibility of the hessian because optimization can correctly determine the the correct likelihood that would indicate poor convergence because variance parameters may be at the lower bound of zero correctly for models that assume the incorrect process error structure.

## Bias

For a given model attribute we calculated the relative error

$$\text{RE}(\theta_i) = \frac{\hat{\theta}_i - \theta_i}{\theta_i}$$

from fitting a given estimating model to simulated data set  $i$  configured for a given OM where  $\hat{\theta}_i$  and  $\theta_i$  are the estimated and true values for simulation  $i$ . We estimated bias as the median of the relative errors across all simulations for a given OM and EM combination. We constructed 95% confidence intervals for the median relative bias using the binomial distribu-

tion approach as in Miller and Hyun (2018) and Stock and Miller (2021). We present results for terminal years SSB and recruitment, Beverton-Holt stock recruit parameters ( $a$  and  $b$ ), and median natural mortality rate. Results for terminal year fishing mortality are strongly negatively correlated with those for SSB and provided in the supplementary materials.

Similar to the AIC results, bias results only condition on whether the estimating model completed the optimization process without failure. We did not condition on convergence as defined by a gradient threshold or invertibility of the hessian because the optimized model can provide reliable estimation of SSB,  $F$ ,  $M$ , and stock recruit parameters whether or not the model was able to estimate non-zero random effects. In practice, the model would be reconfigured to remove unnecessary process errors and produce otherwise equivalent parameter estimates.

## Mohn's $\rho$

We estimated Mohn's  $\rho$  for SSB, fully-selected fishing mortality, and recruitment for each EM (Mohn 1999). We estimated 7 peels for each EM. We calculated median 95% confidence intervals for Mohn's  $\rho$  using the same methods as that for relative bias. Similar to the other results, retrospective results only condition on whether all of the peels of a given estimating model completed the optimization process without failure. We did not condition on convergence as defined by a gradient threshold or invertibility of the hessian because the optimized model can provide reliable estimation of SSB, fishing mortality rates, and recruitment whether or not the model was able to estimate non-zero random effects.

## Results

### Convergence performance

For R and R+S OMs, convergence generally declines for most EMs when the median natural mortality rate is estimated and/or the Beverton-Holt stock recruit relationship is estimated even when the process error assumptions of the EMs and OMs match (Figure 1, A). When there was high observation error and constant fishing pressure (change from  $2.5F_{\text{MSY}}$  to  $F_{\text{MSY}}$ ), convergence of all EM process error configurations other than R EMs was poor for R OMs ( $\sigma_{2+} = 0$ ) regardless of whether median natural mortality and stock-recruit relationships were estimated. Convergence of R EMs was high for all R and R+S OMs except when there was high observation error and constant fishing pressure and median natural mortality and stock-recruit relationships were estimated. R+S EMs fit to R OMs exhibited poor convergence regardless of whether natural mortality or a stock-recruit relationship was estimated. R+S EMs fit to R+S OMs had highest convergence rates when there was contrast in fishing pressure and low observation error. Convergence rates were high for all EMs when fit to data from R+S OMs with lower observation uncertainty except those where median natural mortality and/or stock-recruit relationships were estimated.

Convergence of all EMs fitted to R+M OMs was best when the OMs had higher natural mortality process error variability, low observation error, and contrast in fishing pressure (Figure 1, B). R+M EMs that estimated autocorrelation of process errors had poor convergence for R+M OMs when there was low natural mortality process error variability whether or not there was autocorrelation of the simulated process errors. R+S EMs fitted to data generated from R+M OMs always converged poorly whether median natural mortality and the Beverton-Holt stock-recruit relationship were estimated.

R+S EMs, in particular, converged poorly when fit to data generated from R+Sel OMs with lower selectivity process error variability or higher observation uncertainty (Figure 1, C).

R+Sel EMs generally converged better than other EMs for R+Sel OMs with higher process error variability, lower observation uncertainty, and contrast in fishing pressure regardless of whether median natural mortality or a stock recruit relationship was estimated.

For R+q OMs, convergence of R+q EMs is generally better than that of other EMs when there is contrast in fishing history (Figure (1, D)). Convergence of R+S EMs is generally worse than that of all other EMs across all OMs whether or not median natural mortality or a stock recruit relationship is estimated. Again, convergence probability generally declines for all EMs when median natural mortality or a stock recruit relationship is estimated.

We found a wide range of maximum absolute values of gradients for models that converged as defined here (Figure S2). The largest value observed for a given EM and OM combination was typically  $< 10^{-3}$ , but many converged models had values greater than 1. For many OMs, EMs that assumed the correct process error type and did not estimate median natural mortality or the Beverton-Holt stock-recruit relationship produced the lowest gradient values.

## **AIC performance for process error structure**

Marginal AIC accurately determines the correct process error assumptions in EMs when data are generated from R and R+S OMs, regardless of whether median natural mortality or a stock recruit relationship is estimated (Figure 2, A). Adding estimation of median natural mortality or a stock recruit relationship separately has a negligible effect on the accuracy of determining the correct process error assumption. When both are estimated, there is a noticeable reduction in accuracy when OMs have a constant fishing history, observation uncertainty is low and larger variability in recruitment process errors.

For R+M OMs, marginal AIC only accurately determined the correct process error model and correlation structure when observation error was low and variability in natural mortality process errors was high (Figures 2, B). Estimating the median natural mortality rate reduced the accuracy of AIC for OMs that assumed natural mortality process errors were independent.

For OMs with poor accuracy, AIC most frequently selected EMs with process errors in catchability (R+q) or selectivity (R+Sel). Selection of R+S EMs was typically unlikely.

Marginal AIC most accurately determined the correct source of process error and correlation structure for R+Sel OMs with low observation uncertainty (Figures 2, C). When there was low variability in selectivity process errors and high observation uncertainty, R+q or R+S EMs were more likely to have the best AIC. Whether median natural morality or stock recruit relationships were estimated appeared to have little effect on the performance of AIC.

Marginal AIC most accurately determined the correct source of process error and correlation structure for R+q OMs with high variability in catchability process errors (Figures 2,D). The worst accuracy occurred for OMs with low variability in catchability process errors and high observation uncertainty. However, in these OMs, the marginal AIC accurately determined the correct source of process error (but not correlation structure) except when EMs estimated both median natural morality and the stock recruit relationships and OMs assumed a constant fishing pressure.

## **AIC performance for the stock-recruit relationship**

Our comparisons of model performance condition on assuming the true process error configuration is known (EM and OM process error types match). Broadly, we found generally poor accuracy of AIC in selecting models assuming a Beverton-Holt stock recruit function over the null model without an assumed stock-recruit relationship for all OMs. However, we also found increased accuracy of AIC in determining the Beverton-Holt stock-recruit relationship with greater variation in spawning biomass generated in the simulated populations for nearly every OM (Figure 3).

With R and R+S process error assumptions, probability of AIC accuracy for the B-H stock-recruit relationship as a function of SSB variability were greatest for OMs where there was

greater process variability in survival, lower process variability in recruitment and contrast in fishing pressure (Figure 3, A). However, the largest variation in SSB (Figure 3, rug on x-axis) occurred in OMs with larger recruitment variability ( $\sigma_R = 1.5$ ; Figure 3, A, right column group). The weakest overall accuracy of AIC for the Beverton-Holt stock-recruit function was when there was constant fishing pressure (Figure 3, A, second and fourth rows). The largest effect of simultaneously estimating median natural mortality was when there was OMs had high observation uncertainty and constant fishing pressure (Figure 3, A, second row).

For R+M OMs, probability of AIC accuracy for the Beverton-Holt stock-recruit relationship was greatest for OMs where there was contrast in fishing pressure (Figure 3, B). The largest variation in SSB occurred in OMs with larger natural mortality variability and temporal AR1 correlation of the annual random effects ( $\sigma_M = 1.5$ ,  $\rho_M = 0.9$ ; Figure 3, B, right-most column) where AIC accuracy was also high even when there was constant fishing pressure. Interestingly, there seemed to be little difference in AIC accuracy whether there was lower or higher observation uncertainty or whether the EM simultaneously estimated median natural mortality.

For R+Sel OMs, contrast in fishing pressure over time is the primary mechanism creating variation in SSB and these are the OMs where AIC accuracy for the Beverton-Holt stock-recruit relationship was greatest (Figure 3, C, first and third rows). There appears to be little effect of variability or correlation of selectivity process errors or whether median natural mortality was estimated on AIC accuracy.

Like the R+Sel OMs, the greatest accuracy for AIC in selecting the Beverton-Holt stock-recruit relationship occurred for R+q OMs where there was contrast in fishing pressure over time which is also where there was the greatest variation in SSB (Figure 3, D, first and third rows). There were some differences in effects of SSB variation on AIC accuracy when OMs had high observation uncertainty and constant fishing pressure. Across all of the OMs in



Figure 3, the only estimated decline in AIC accuracy with SSB variation occurred when R+q OMs had high observation uncertainty and constant fishing pressure and greater variation in catchability process errors that also had AR1 autocorrelation.

## Bias

### Spawning stock biomass and recruitment

For R OMs ( $\sigma_{2+} = 0$ ), there was no indication of bias (95% confidence intervals included 0) in terminal year SSB for any of the estimating models regardless of process error assumptions, except when no SR assumption was made, recruitment variability was low, and there was contrast in fishing mortality and high observation uncertainty (Figure 4, A). However, errors in terminal SSB estimates were highly variable when median natural mortality was estimated and there was constant fishing pressure and high observation uncertainty (Figure 4, A, second row).

For R+S OMs, the EMs with matching process error assumption generally produced unbiased estimation of terminal SSB except when median natural mortality was estimated and there was high observation uncertainty. In R+S OMs with low observation uncertainty, EMs with incorrect process error assumptions typically provided biased estimation of terminal year SSB. Estimating the Beverton-Holt stock-recruit relationship had little discernible effect on bias of terminal year SSB estimation whereas estimating median M tended to produce more variability in errors in terminal SSB estimation similar to R OMs.

For R+ M OMs with low variability in natural mortality process errors, low observation uncertainty and contrast in fishing mortality over time all EMs produced low variability in SSB estimation error that indicated unbiasedness (Figure 4, B, third row). However, larger variability in natural mortality process errors increased bias of EMs without the correct process error type. Estimating median natural mortality increased variability of SSB estimation error particularly for OMs with high observation uncertainty and constant fishing pressure

381 over time and increased bias in SSB estimation for many R+M OMs. Like R and R+S OMs,  
382 estimating a stock-recruit relationship had little discernible effect on SSB bias.

383 For R+Sel OMs, there was no evidence of bias for any EMs when variability in selectivity  
384 process errors was low, observation uncertainty low and contrast in fishing mortality (Fig-  
385 ure 4, C). The largest bias for all EMs that estimated median natural mortality when the  
386 OMs had high observation uncertainty, constant fishing pressure and greater variability in  
387 selectivity process errors ( $\sigma_{\text{Sel}} = 0.5$ ) or low selectivity process errors ( $\sigma_{\text{Sel}} = 0.1$ ) and low  
388 observation uncertainty. However, there was no evidence of bias of matching R+Sel EMs  
389 when observation uncertainty was low and variation in selectivity process errors was larger,  
390 whether median natural mortality was estimated or not (Figure 4, C, third row). We only  
391 observed an effect of estimating the Beverton-Holt Stock recruit relationship for R+Sel OMs  
392 that had high observation uncertainty and contrast in fishing pressure where estimating the  
393 relationship produced less biased SSB estimation for many EMs (Figure 4, C, top row).

394 All EMs fit to data from R+q OMs with low observation uncertainty and contrast in fishing  
395 pressure exhibited little evidence of bias in terminal SSB estimation except for R+M EMs  
396 when there was no AR1 correlation in catchability process errors (Figure 4, D). Many EMs  
397 also performed well in R+q OMs with low observation uncertainty, but no contrast in fishing  
398 pressure. For R+q OMs with high observation uncertainty and contrast in fishing pressure,  
399 EMs that estimated the Beverton-Holt stock recruit function exhibited less SSB bias than  
400 those that did not. Estimating median natural mortality in the EMs only resulted in much  
401 more variable SSB estimation errors when there was no contrast in fishing pressure (Figure  
402 4, D, first and third rows).

403 For all OM process error types, relative errors in terminal year recruitment were generally  
404 more variable than SSB, but effects of R and R+S OM and EM attributes on bias (i.e,  
405 negative or positive or none) were similar (Figure S5, A). Furthermore, for EM configurations  
406 where bias in terminal bias in SSB was evident, median relative errors in recruitment often

indicated stronger bias in recruitment of the same sign.

## **Beverton-Holt parameters**

In R and R+S OMs, EMs with the correct assumptions about process errors, provided the least biased estimation of Beverton-Holt stock-recruit relationship parameters when there was low observation error, a change in fishing pressure over time, and lower variability of recruitment process errors, and higher variability survival process errors, and there was little effect of estimating natural mortality (Figure S6, A). For other R and R+S OMs, estimating natural mortality often resulted in less biased estimation of stock-recruit parameters. There was generally large variability in relative errors of the stock-recruit parameter estimates, but the lowest variability occurred with low variability in recruitment and little or no variability in survival process errors ( $\sigma_{2+} \in \{0, 0.25\}$ ), and contrast in fishing pressure. Across all R and R+S OMs, relative errors for the  $a$  parameter were often less variable than those for  $b$ .

In R+M OMs, the most accurate estimation of stock-recruit parameters for all EM process error assumptions occurred when there was a change in fishing pressure combined with either low variability in natural mortality process errors and high observation error or vice versa (Figure S6, B). Relative to the R, and R+S OMs, there was even less effect of estimating median natural mortality on estimation bias for the stock-recruit relationship parameters, but similarly there was generally large variability of the parameter estimates.

Bias for stock-recruit parameters was very strong and variability in relative errors was greatest for most EMs fit to R+Sel OMs with constant fishing pressure (Figure S6, C). Less bias in parameter estimation occurred for OMs with a change in fishing pressure over time and of those OMs and the most accurate estimation occurred for EMs with process error types that matched the OMs and the OMs had low observation uncertainty, contrast in fishing mortality and more variable and uncorrelated selectivity process errors. There was little effect of estimating natural mortality on relative errors for stock recruit parameters.

Like R+Sel OMs, relative errors in stock-recruit parameters were less variable for most EM process error types when OMs had contrast in fishing pressure and those OMs that also had lower observation uncertainty generally had the least variability (Figure S6, D). The worst accuracy of stock-recruit parameter estimation regardless of EM type occurred when R+q OMs had low observation uncertainty and constant fishing pressure (Figure S6, D, fourth row).

### Median natural mortality rate

Median natural mortality rate is estimated accurately by all EM process error types for all R OMs except those where observation uncertainty is high and fishing pressure is constant where variability in relative errors is high (Figure S7, A,  $\sigma_{2+} = 0$ ). For R+S OMs estimation of median natural mortality rate is most accurate when observation uncertainty is low and there is contrast in fishing pressure and the EM process error type is correct. There is little effect of estimating stock-recruit relationships on the bias for median natural mortality.

For R+M OMs, median natural mortality was estimated most accurately, regardless of EM process error type, when OMs had a change in fishing pressure and low observation uncertainty (Figure S7, B). However, those R+M OMs that also had greatest variability in AR1 correlated natural mortality process errors only had unbiased estimation when the EM process error type was correct. Like R, and R+S OMs, there was little effect of estimating stock-recruit relationships on the bias for median natural mortality.

All EM process error types produced accurate estimation of median natural mortality rate for R+Sel OMs that had contrast in fishing pressure, low observation uncertainty, and low selectivity process error variability (Figure S7, C). When selectivity process error variability increased the incorrect EM process errors produce more biased estimation of median natural mortality rate. The least accurate estimation occurred for all EM process error types when observation uncertainty was high and there was constant fishing pressure. Again, like other

OM process error types, there was little effect of estimating stock-recruit relationships on the bias for median natural mortality.

Like R+Sel OMs, all EM process error types produced accurate estimation of median natural mortality rate when fit to R+q OMs with contrast in fishing pressure, low observation uncertainty and low catchability process error variability (Figure S7, D). Most EM process error types produced biased estimation of median natural mortality when R+q OMs had high observation uncertainty and constant fishing pressure. Again, there was no discernible effect of estimating stock-recruit relationships on bias.

### **Mohn's $\rho$**

Mohn's  $\rho$  for SSB was small in absolute value for all R and R+S OMs, regardless of EM process error types, and whether median natural mortality rate or stock-recruit relationships were estimated (Figure 5, A). The strongest retrospective patterns occurred in OMs with the largest survival process error variability, high observation uncertainty, and contrast in fishing pressure, but only for EMs without the correct process error type and where median natural mortality rate was assumed known. For R+M, R+Sel, and R+q OMs, Mohn's  $\rho$  was also small in absolute value, but median values were all closer to 0 than the largest values in the R and R+S OMs (Figure 5,B-D). For these OMs, there was no noticeable effect of estimation of median natural mortality rate or stock recruit relationship on Mohn's  $\rho$  for any EM process error types.

Mohn's  $\rho$  for recruitment was small in absolute value for all R OMs with low variability in recruitment process errors, regardless of EM process error type, and whether median natural mortality rate or stock-recruit relationships were estimated (Figure S9, A). However, R and R+S OMs with greater recruitment process variability and higher observation uncertainty had median Mohn's  $\rho$  for recruitment greater than zero for most EMs even when the EM process error type was correct. In R+S OMs with lower observation uncertainty, EMs

with the correct process error type exhibited better median Mohn's  $\rho$  close to 0 than EMs with the incorrect process error type. For R+M, R+Sel, and R+q OM, results for Mohn's  $\rho$  for recruitment are similar to those for SSB, but the range in median values and variation in Mohn's  $\rho$  values for a given OM are generally larger for recruitment (Figure S9, B-D).

## Discussion

### Convergence

Convergence results can be useful for understanding how lack of convergence in applications to real data might direct the practitioner to which alternative random effects configurations is more appropriate. Therefore, the type of convergence that we might use as a diagnostic is important. It is common during assessment model fitting to check that the maximum absolute gradient component is less than some threshold, but there is no standard. However, checking the Hessian for positive definiteness is sometimes done conditional on the convergence criterion. As such, some models may be excluded that in fact have an invertible hessian. We found the hessian at the optimized log-likelihood can often be invertible when the gradient approach with a sensible criterion (e.g.,  $10^{-3}$ ) would indicate lack of convergence.

Another factor affecting the convergence criteria is numerical accuracy. With the process errors modeled as random effects we are optimizing an (Laplace) approximation of the marginal log-likelihood and therefore Hessians and gradients are also with regard to the approximation. Furthermore, numerical methods are used to calculate the hessian (Optimhess?), and also to invert it for variance estimation. All of these approximations are affecting the gradients and invertibility of the hessian (although likely worse for some models than others), which, along with our results, suggests at least investigating whether the hessian is positive definite when the calculated gradients are not terribly large.

Often high gradients and/or lack of positive definite hessian result when selectivity param-

eter estimates are at their bounds (0,1) or random effects variance parameter estimates are essentially 0. Changing the parameter configuration to avoid the bounds or removing process errors with 0 variance will often lead to satisfactory convergence.

Perhaps a greater gradient threshold would be more consistent with hessian-based convergence, but the gradient and hessian-based standard errors should be inspected.

Parameters that are not well estimated can lead the assessment scientist toward a more appropriate and often more parsimonious model configuration. For example, variance parameters that are estimated on log scale as low negative number (variance is close to 0) can happen for random effects estimated with little or no variability. These may sometimes result in a hessian that is not positive definite or due to numerical issues may be invertible, but provide very large standard errors. Generally, we can expect lower probability of convergence when estimating natural mortality or stock-recruit relationships because of the difficulty distinguishing these parameters from others being estimated in assessment model with data typically available.

## AIC

Among the process error configurations we used in OMs, we found AIC to be accurate for process errors on recruitment and survival (R and R+S). Fitting models to other OMs rarely preferred R+S EMs and R and R+S EMs were nearly always selected for the matching OMs. For other sources of process error, accuracy of AIC was improved to useful levels when there was larger variability in the process errors and/or better data quality.

AIC generally performs weakly in determining stock recruit relationships unless there is large contrast in SSB and low variation in recruitment process errors. Although we did not compare models with alternative stock-recruit relationships, we do not expect AIC to perform any better distinguishing these relationships. Our results comport well with those of de Valpine and Hastings (2002) where state-space stock-recruit models were fit just to

SSB and recruitment “observations” and Britten et al. which found AIC could not reliably distinguish alternative environmental effects on stock-recruit parameters. Miller et al. (2016) found AIC to prefer a stock-recruit relationship with environmental effects for SNEMA yellowtail flounder, but there was a large change in stock size estimated over time and flatfish are well-observed by the NEFSC bottom trawl survey used for an index in the assessment model.

## Bias

As we might expect bias in all parameters and assessment output was generally better with lower observation error. Reliable estimation of stock-recruit relationship parameters only appears possible in ideal situations with lower observation errors in age composition and indices, lower variability in recruitment process errors and large contrast in spawning biomass over time. The latter can occur with larger variability in survival or natural mortality process errors, or variation in fishing pressure.

Estimation of median natural mortality appears more feasible than SR parameters. We found low or no evidence of bias for many OM process error assumptions when there was contrast in fishing pressure even when there was greater observation error although it can lead to more variable estimates of natural mortality. For OMs where there was bias in natural mortality due to high observation error (R+Sel, R+q), estimating the stock-recruit relationship seemed to remove the bias.

However, estimation of natural mortality can cause large differences between the true and estimated SSB (that may be unbiased on average) when there is less contrast in fishing pressure over time and higher observation error.

Estimating B-H reduced SSB bias in some situations



## Retrospective patterns

Higher observation error with variation in fishing pressure can produce retrospective patterns even when the the process error assumption is correct, but this retrospective pattern can be reduced if  $M$  or the correct stock recruit relationship is estimated. Retrospective patterns tended to be more variable for recruitment and can be large even when the EM is correct. Therefore, we recommend emphasis on inspection of retrospective patterns primarily for SSB and  $F$ , but further research on retrospective patterns in other assessment model parameters and derived output may be beneficial.

## Conclusions

All together what do results mean? E.g., when do you get good convergence and retrospective patterns and high/none bias? When do you get good convergence and no retros and no/big bias?

When  $R+S$  is preferred it probably means its the correct process error OF THE TYPES CONSIDERED HERE.

## Summary

We conducted a simulation study where we simulated process errors and observations for 72 operating models with alternative assumptions about fishing history, degree of uncertainty in index and age composition observations, type (recruitment, survival, fishery selectivity, catchability and natural mortality), degree of variation, and correlation of process errors. We fit 20 different estimation models to each of 100 simulated set of observations with alternative assumptions type and correlation structure of process errors, median natural mortality was known or estimated, and a B-H stock recruit relationship was assumed or not.

Across simulations we summarized probability of convergence of fitted models, accuracy of

marginal AIC in determining the correct process error assumption and its ability to determine the Beverton-Holt stock recruit relationship, bias in annual spawning stock biomass, in estimation of natural mortality, and in stock-recruit relationship parameters, and severity of retrospective patterns for estimation models.

Alternative measures of convergence performed differently. Invertible Hessians and resulting standard error estimation was possible when criteria based on the gradient of the optimized log-likelihood with respect to the fixed effects parameters. Using Hessian-based convergence, probability of convergence was best for models that assumed the correct source of process error, assumed  $M$  was known, and did not assume stock-recruit relationships.

Using marginal AIC provided most accurate inferences about the process errors on recruitment and survival and selectivity, in that AIC preferred EMs with assumed process errors that matched OMs, and EMs with those assumed process errors were not preferred when alternative process errors were assumed in the OMs. However, when the true process errors were more variable, AIC accuracy increased to a useful level. We found AIC more accurately determined a B-H stock recruit relationship rather than the null model without a S-R relationship when there was low variability in recruitment, low variability in survival random effects, and higher variation in spawning biomass over the time series. When the median natural mortality rate was estimated, we found large bias and uncertainty much more likely in model output such as spawning stock biomass.

Bias in spawning stock biomass estimation was generally low for estimation models that assumed the correct source of process error when there was lower observation error. Reliable estimation of stock-recruit relationship parameters only appears possible in ideal situations with lower observation errors in age composition and indices, lower variability in recruitment process errors and large contrast in spawning biomass over time. We found little evidence of bias for many OM process error assumptions when there was contrast in fishing pressure even when there was greater observation error although it can lead to more variable estimates of

natural mortality. For OMs where there was bias in natural mortality due to high observation error, estimation the stock-recruit relationship seemed to remove the bias. However, estimation of natural mortality can cause large differences between the true and estimated SSB (that may be unbiased on average) when there is less contrast in fishing pressure over time and higher observation error.

Retrospective patterns were generally weak for all estimation models regardless of the true source of process error, but they can be expected for recruitment even for the correct process error assumptions when observation error is high. When models did exhibit some retrospective pattern, estimation the median natural mortality rate tended to remove it.

## Acknowledgements

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

## References

- Aeberhard, W.H., Flemming, J.M., and Nielsen, A. 2018. Review of State-Space Models for Fisheries Science. *Annual Review of Statistics and Its Application* **5**(1): 215–235. doi:10.1146/annurev-statistics-031017-100427.
- Britten, G., Brooks, E.N., and Miller, T.J. Reliability of stock-recruitment function estimation in state space assessment models.
- Cadigan, N.G. 2016. 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. doi:10.1139/cjfas-2015-0047.
- Carvalho, F., Winker, H., Courtney, D., Kapur, M., Kell, L., Cardinale, M., Schirripa, M., Kitakado, T., Yemane, D., Piner, K.R., Maunder, M.N., Taylor, I., Wetzel, C.R., Doering, K., Johnson, K.F., and Methot, R.D. 2021. A cookbook for using model diagnostics in integrated stock assessments. *Fisheries Research* **240**: 105959. doi:https://doi.org/10.1016/j.fishres.2021.105959.
- Conn, P.B., Williams, E.H., and Shertzer, K.W. 2010. When can we reliably estimate the productivity of fish stocks? *Canadian Journal of Fisheries and Aquatic Sciences* **67**(3): 511–523. doi:10.1139/F09-194.
- de Valpine, P., and Hastings, A. 2002. Fitting population models incorporating process noise and observation error. *Ecological Monographs* **72**(1): 57–76.
- Hoyle, S.D., Maunder, M.N., Punt, A.E., Mace, P.M., Devine, J.A., and A’mar, Z.T. 2022. Preface: Developing the next generation of stock assessment software. *Fisheries Research* **246**: 106176.
- Legault, C.M., Wiedenmann, J., Deroba, J.J., Fay, G., Miller, T.J., Brooks, E.N., Bell, R.J., Langan, J.A., Cournane, J.M., Jones, A.W., and Muffley, B. 2023. Data-rich but model-resistant: An evaluation of data-limited methods to manage fisheries with failed age-based stock assessments. *Canadian Journal of Fisheries and Aquatic Sciences* **80**(1): 27–42. doi:10.1139/cjfas-2022-0045.

- Li, C., Deroba, J.J., Miller, T.J., Legault, C.M., and Perretti, C. In review. Guidance on bias-correction of log-normal random effects and observations in state-space assessment models. For Canadian Journal of Fisheries and Aquatic Sciences, State-space special collection.
- 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. doi:10.1016/j.fishres.2024.106968.
- Magnusson, A., and Hilborn, R. 2007. What makes fisheries data informative? *Fish and Fisheries* **8**(4): 337–358. doi:<https://doi.org/10.1111/j.1467-2979.2007.00258.x>.
- Miller, T.J., and Brooks, E.N. 2021. Steepness is a slippery slope. *Fish and Fisheries* **22**(3): 634–645. doi:10.1111/faf.12534.
- Miller, T.J., Hare, J.A., and Alade, L. 2016. A state-space approach to incorporating environmental effects on recruitment in an age-structured assessment model with an application to Southern New England yellowtail flounder. *Canadian Journal of Fisheries and Aquatic Sciences* **73**(8): 1261–1270.
- Miller, T.J., and Hyun, S.-Y. 2018. Evaluating evidence for alternative natural mortality and process error assumptions using a state-space, age-structured assessment model.
- Miller, T.J., and Stock, B.C. 2020. The Woods Hole Assessment Model (WHAM). Available from <https://timjmiller.github.io/wham/>.
- Mohn, R. 1999. The retrospective problem in sequential population analysis: An investigation using cod fishery and simulated data. *ICES Journal of Marine Science* **56**(4): 473–488. doi:10.1006/jmsc.1999.0481.
- Nielsen, A., and Berg, C.W. 2014. Estimation of time-varying selectivity in stock assessments using state-space models. *Fisheries Research* **158**: 96–101. doi:10.1016/j.fishres.2014.01.014.
- Pedersen, M.W., and Berg, C.W. 2017. A stochastic surplus production model in continuous time. *Fish and Fisheries* **18**(2): 226–243. doi:10.1111/faf.12174.

- Punt, A.E. 2023. Those who fail to learn from history are condemned to repeat it: A perspective on current stock assessment good practices and the consequences of not following them. *Fisheries Research* **261**: 106642.
- 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. doi:<https://doi.org/10.1016/j.fishres.2021.105967>.
- Wiedenmann, J., Free, C.M., and Jensen, O.P. 2019. Evaluating the performance of data-limited methods for setting catch targets through application to data-rich stocks: A case study using northeast U.S. Fish stocks. *Fisheries Research* **209**(1): 129–142. doi:<https://doi.org/10.1016/j.fishres.2018.09.018>.

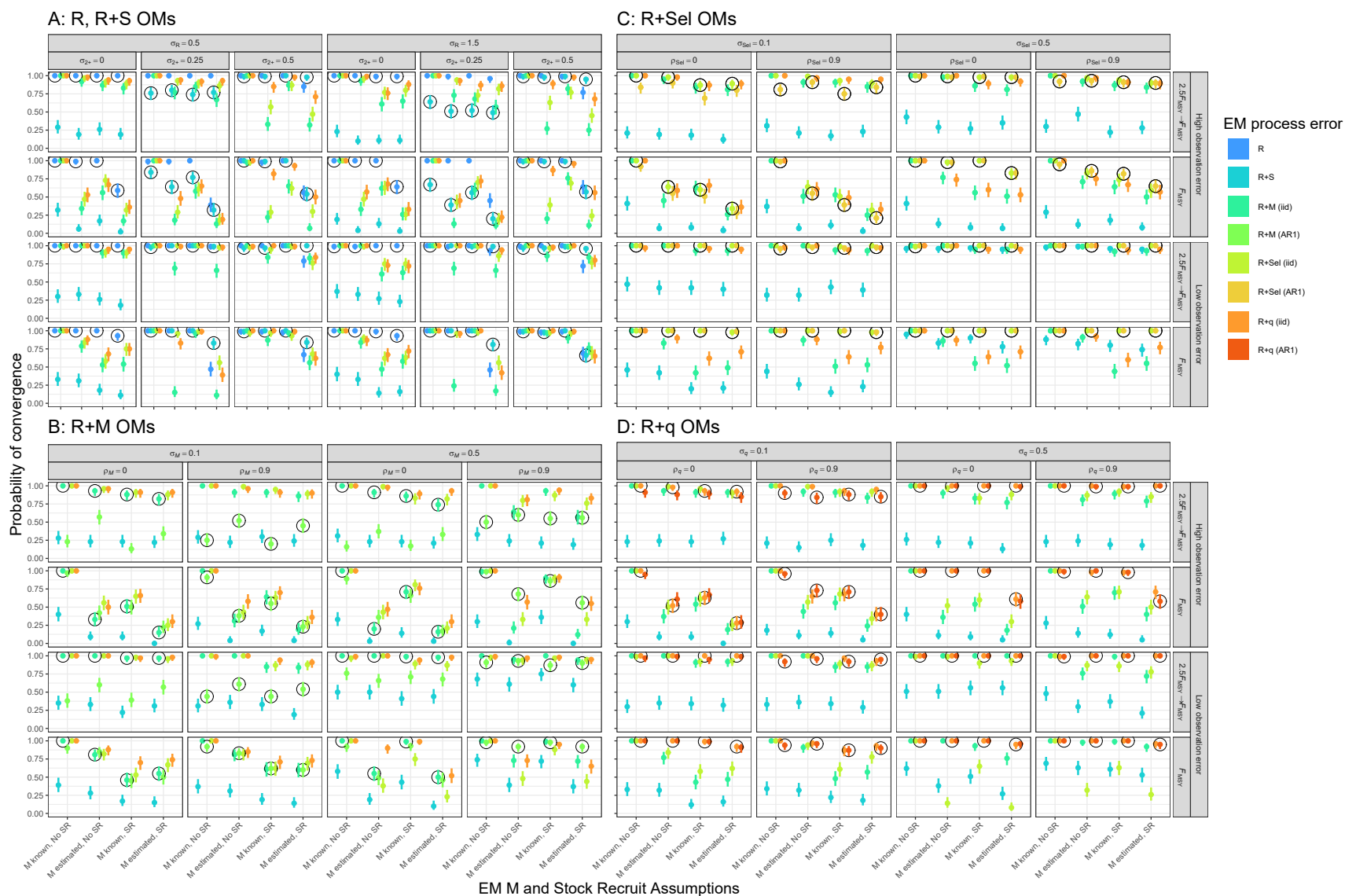


Fig. 1. Estimated probability of fits providing hessian-based standard errors for EMs assuming alternative process error (colored points and lines), and median natural mortality (estimated or known) and Beverton-Holt stock recruit functions (estimated or not; along x-axis) when fitted to operating models that have R and R+S (A), R+Sel (B), R+M (C), or R+q (D) process error structures. Circled values indicate results where the EM process error structure matches that of the operating model and vertical lines represent 95% confidence intervals.

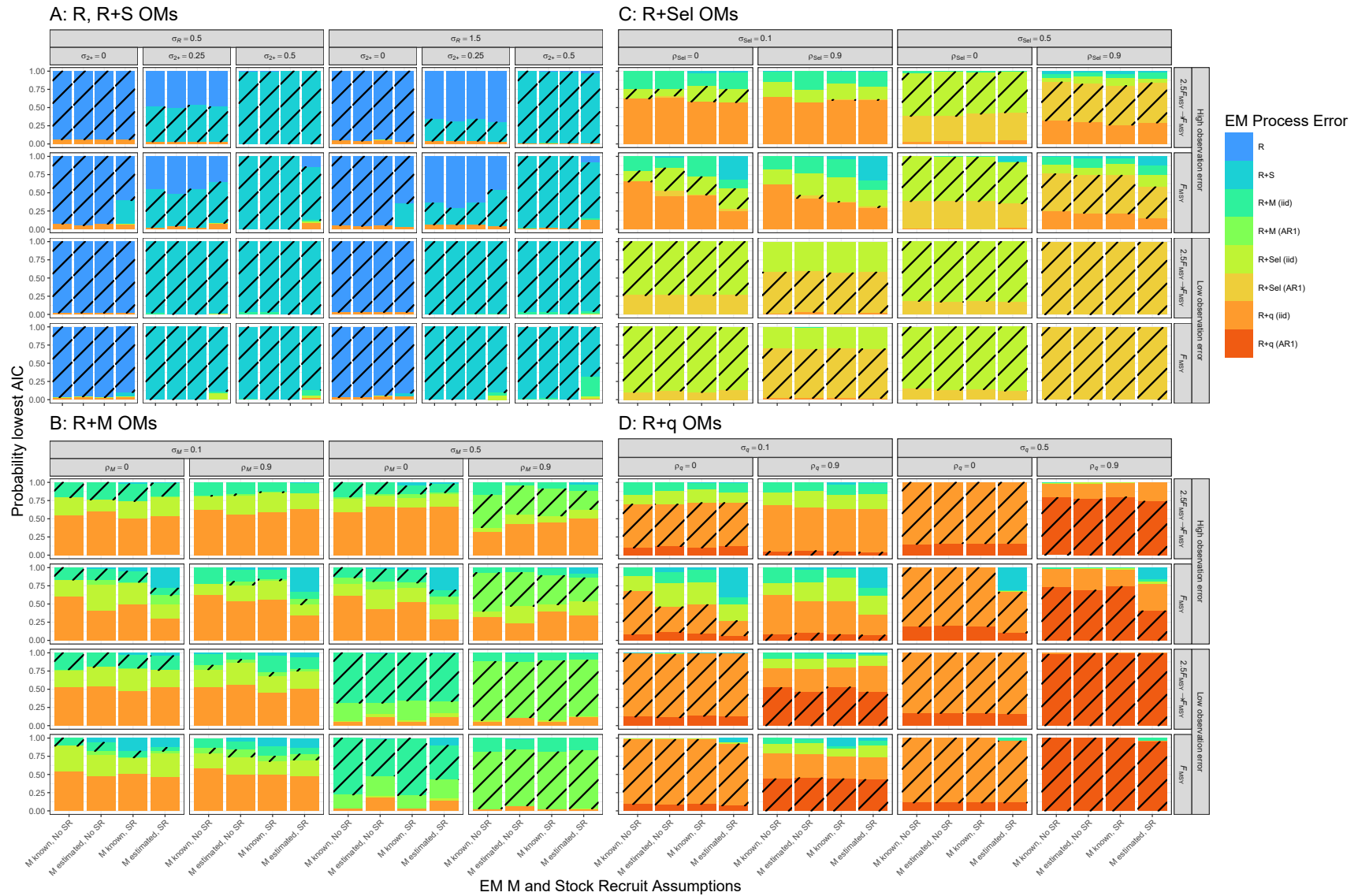


Fig. 2. Estimated probability of lowest AIC for EMs assuming alternative process error structures (colored bars) conditional on alternative assumptions for median natural mortality (estimated or known) and Beverton-Holt stock recruit functions (estimated or not; along x-axis) when fitted to operating models that have R and R+S (A), R+Sel (B), R+M (C), or R+q (D) process error structures. Striped bars indicate results where the EM process error structure matches that of the operating model.



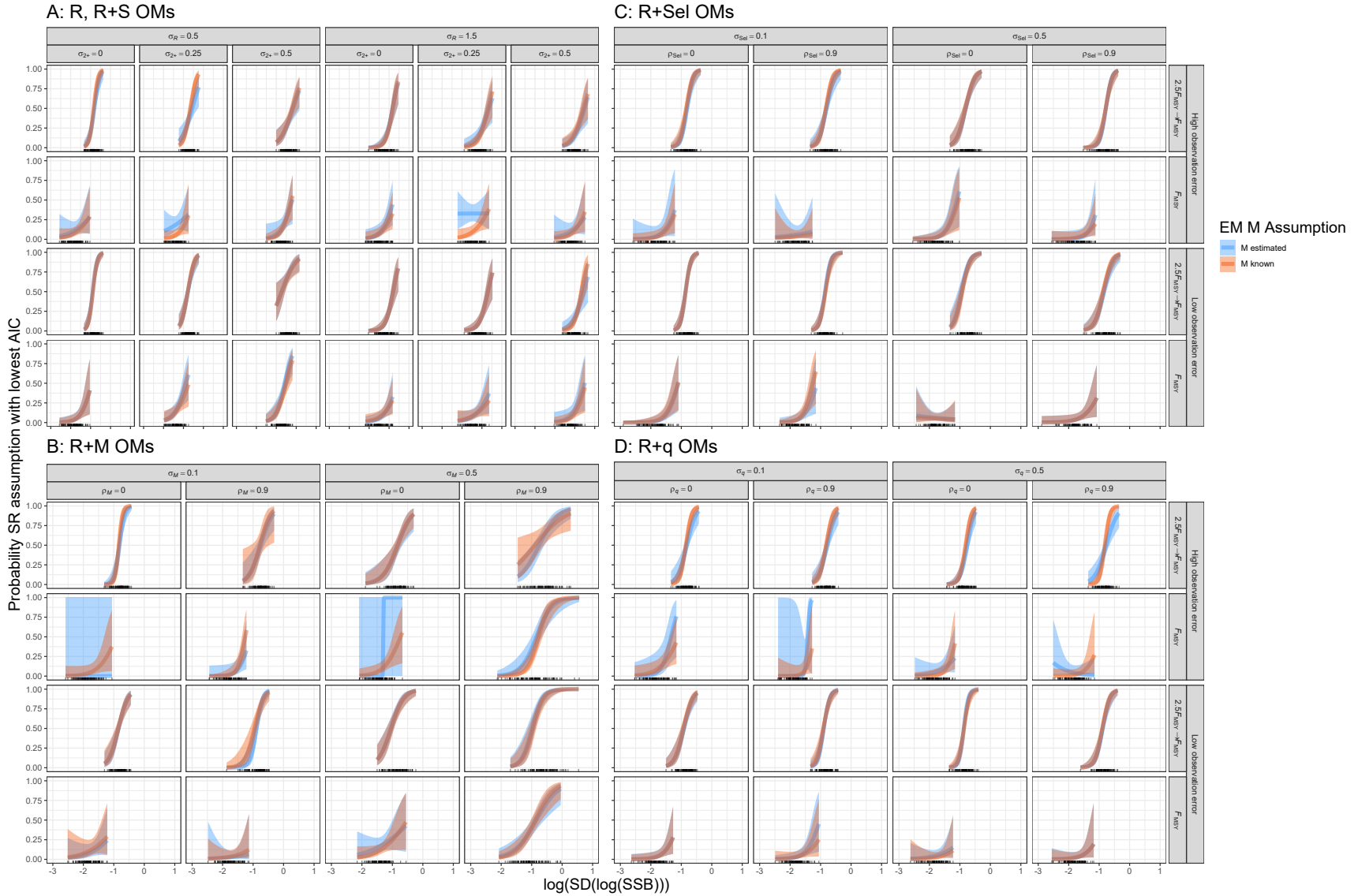


Fig. 3. Estimated probability of lowest AIC from logistic regression on the log-standard deviation of the true log(SSB) in each simulation for estimating model with Beverton-Holt stock recruit functions, rather than the otherwise equivalent EM without the stock recruit function. Results are conditional on alternative assumptions for median natural mortality (estimated or known) and on EMs having the correct process error structure: R and R+S (A), R+Sel (B), R+M (C), or R+q (D). Rug along x-axis denotes  $SD(\log(SSB))$  values for each simulation and polygons represent 95% confidence intervals.

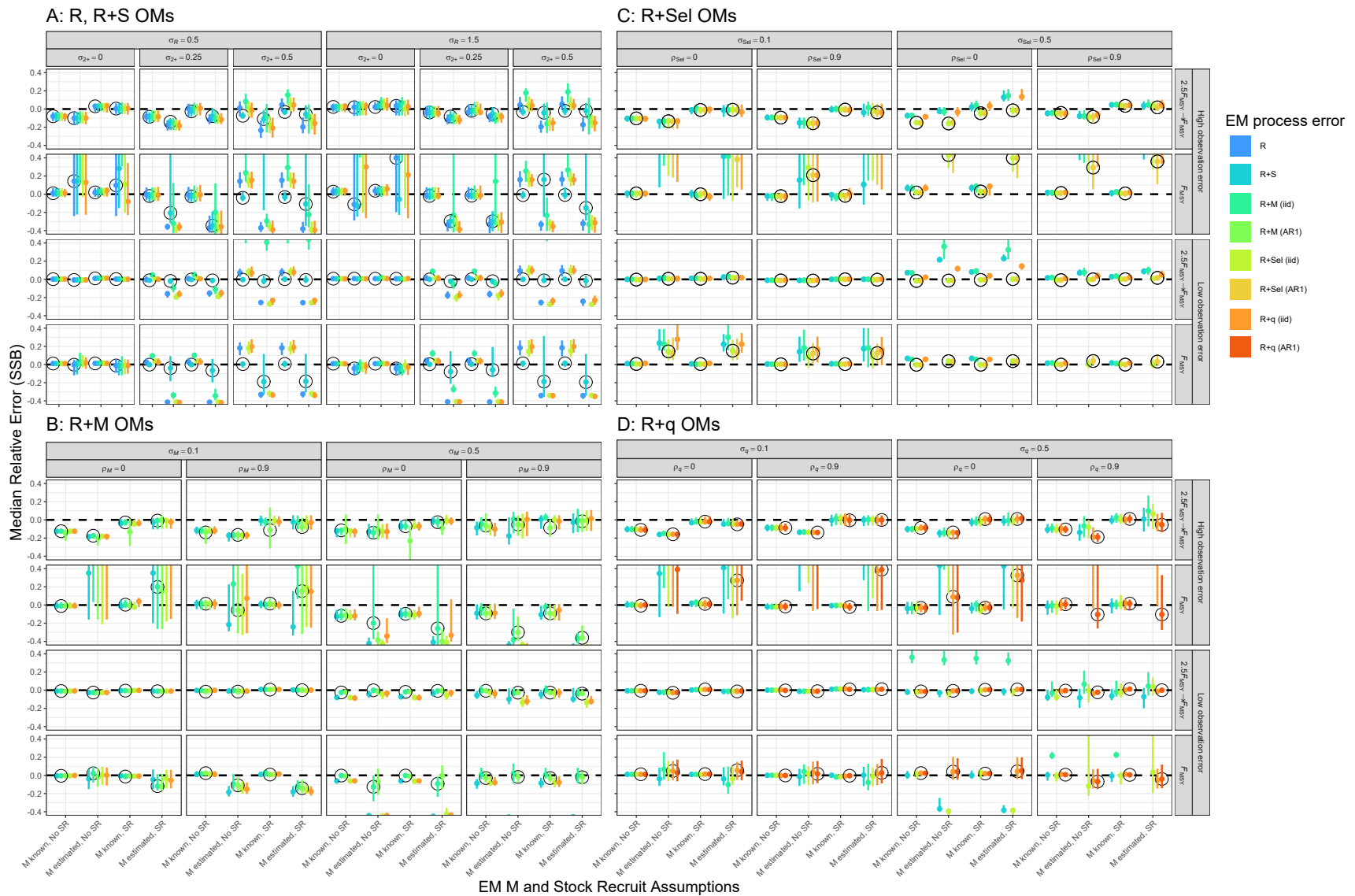


Fig. 4. Median relative error of terminal year SSB for estimating models fitted to data sets simulated with alternative process error structures: R and R+S (A), R+Sel (B), R+M (C), or R+q (D). Diamond shaped points denote results where the EM process error assumption matches that of the operating model. Circled values indicate results where the EM process error structure matches that of the operating model and vertical lines represent 95% confidence intervals.

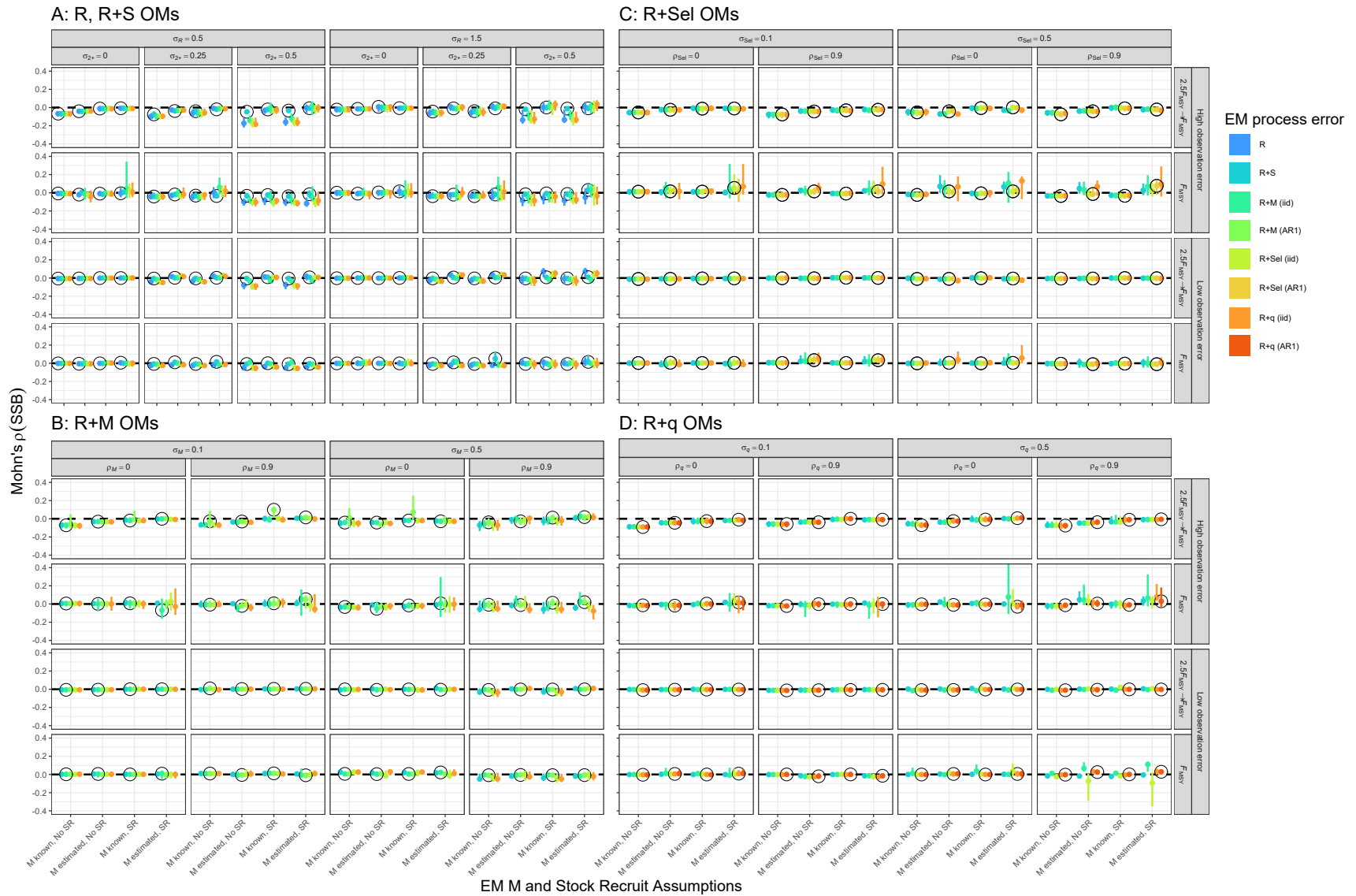


Fig. 5. Median Mohn's rho for SSB for estimating models fitted to data sets simulated with alternative process error structures: R and R+S (A), R+Sel (B), R+M (C), or R+q (D). Circled values indicate results where the EM process error structure matches that of the operating model and vertical lines represent 95% confidence intervals.

## 679 Supplementary Materials

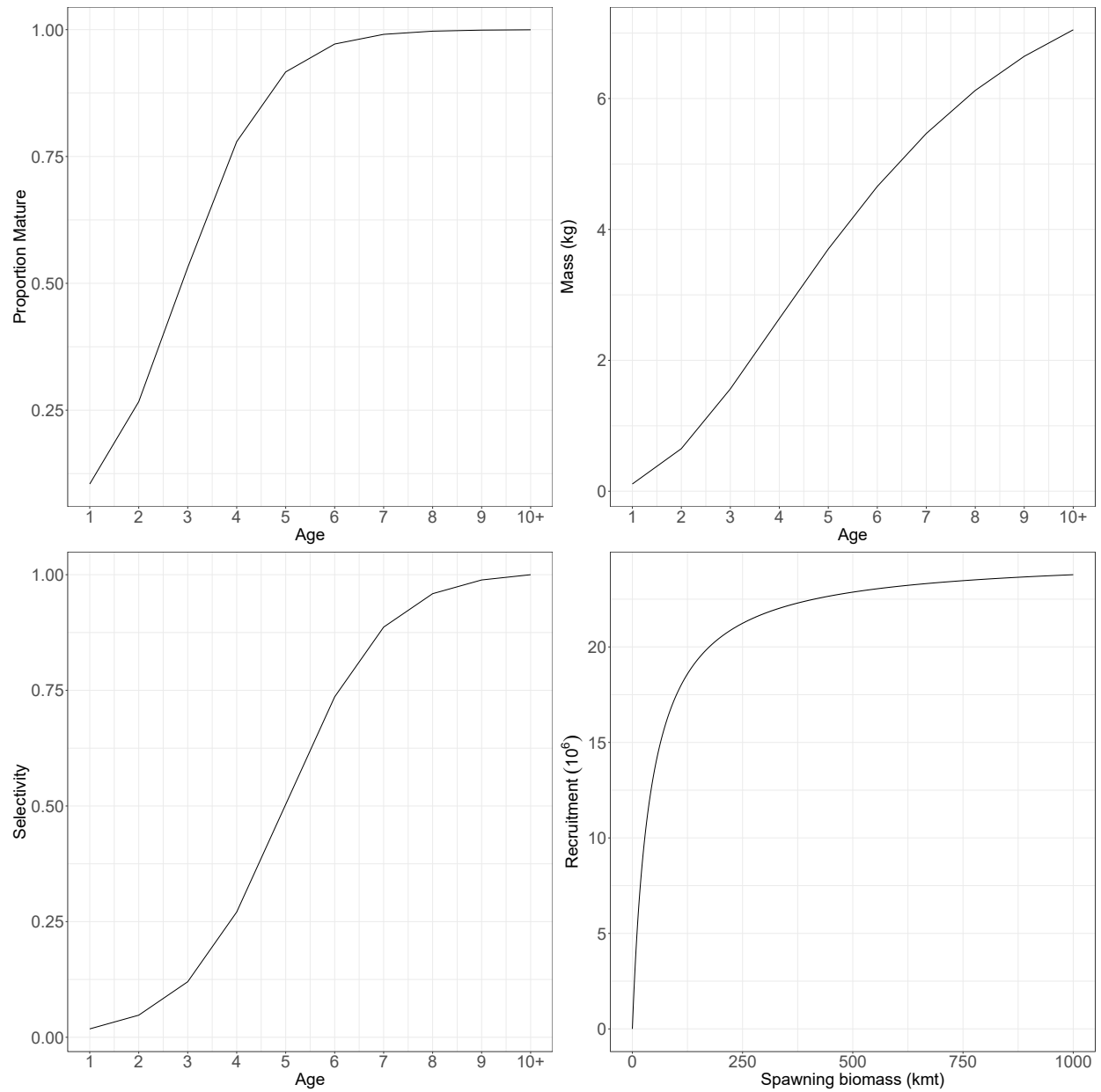


Fig. S1. The proportion mature at age, weight at age, fleet and index selectivity at age, and Beverton-Holt stock recruit relationship assumed for the population in all operating models. For operating models with random effects on fleet selectivity, this represents the selectivity at the mean of the random effects.

Table S1. Distinguishing characteristics of the operating models with random effects on survival. Standard deviations (SD) are for log-normal distributed indices and logistic normal distributed age composition observations (fleet and indices). Fishing mortality changes after year 20 (of 40) for fishing histories where fishing mortality is not constant.

Model	$\sigma_R$	$\sigma_{2+}$	Fishing History	Observation Uncertainty
NAA <sub>1</sub>	0.5		$2.5F_{\text{MSY}} \rightarrow F_{\text{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
NAA <sub>2</sub>	1.5		$2.5F_{\text{MSY}} \rightarrow F_{\text{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
NAA <sub>3</sub>	0.5	0.25	$2.5F_{\text{MSY}} \rightarrow F_{\text{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
NAA <sub>4</sub>	1.5	0.25	$2.5F_{\text{MSY}} \rightarrow F_{\text{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
NAA <sub>5</sub>	0.5	0.50	$2.5F_{\text{MSY}} \rightarrow F_{\text{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
NAA <sub>6</sub>	1.5	0.50	$2.5F_{\text{MSY}} \rightarrow F_{\text{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
NAA <sub>7</sub>	0.5		$F_{\text{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
NAA <sub>8</sub>	1.5		$F_{\text{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
NAA <sub>9</sub>	0.5	0.25	$F_{\text{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
NAA <sub>10</sub>	1.5	0.25	$F_{\text{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
NAA <sub>11</sub>	0.5	0.50	$F_{\text{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
NAA <sub>12</sub>	1.5	0.50	$F_{\text{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
NAA <sub>13</sub>	0.5		$2.5F_{\text{MSY}} \rightarrow F_{\text{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
NAA <sub>14</sub>	1.5		$2.5F_{\text{MSY}} \rightarrow F_{\text{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
NAA <sub>15</sub>	0.5	0.25	$2.5F_{\text{MSY}} \rightarrow F_{\text{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
NAA <sub>16</sub>	1.5	0.25	$2.5F_{\text{MSY}} \rightarrow F_{\text{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
NAA <sub>17</sub>	0.5	0.50	$2.5F_{\text{MSY}} \rightarrow F_{\text{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
NAA <sub>18</sub>	1.5	0.50	$2.5F_{\text{MSY}} \rightarrow F_{\text{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
NAA <sub>19</sub>	0.5		$F_{\text{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
NAA <sub>20</sub>	1.5		$F_{\text{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
NAA <sub>21</sub>	0.5	0.25	$F_{\text{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
NAA <sub>22</sub>	1.5	0.25	$F_{\text{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
NAA <sub>23</sub>	0.5	0.50	$F_{\text{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
NAA <sub>24</sub>	1.5	0.50	$F_{\text{MSY}}$	Index SD = 0.4, Age composition SD = 1.5

Table S2. Distinguishing characteristics of the operating models with random effects on natural mortality. Standard deviations (SD) are for log-normal distributed indices and logistic normal distributed age composition observations (fleet and indices). Fishing mortality changes after year 20 (of 40) for fishing histories where fishing mortality is not constant. For AR1 process errors,  $\sigma$  is defined for the marginal distribution of the processes.

Model	$\sigma_R$	$\sigma_M$	$\rho_M$	Fishing History	Observation Uncertainty
$M_1$	0.5	0.1	0.0	$2.5F_{\text{MSY}} \rightarrow F_{\text{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
$M_2$	0.5	0.5	0.0	$2.5F_{\text{MSY}} \rightarrow F_{\text{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
$M_3$	0.5	0.1	0.9	$2.5F_{\text{MSY}} \rightarrow F_{\text{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
$M_4$	0.5	0.5	0.9	$2.5F_{\text{MSY}} \rightarrow F_{\text{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
$M_5$	0.5	0.1	0.0	$F_{\text{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
$M_6$	0.5	0.5	0.0	$F_{\text{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
$M_7$	0.5	0.1	0.9	$F_{\text{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
$M_8$	0.5	0.5	0.9	$F_{\text{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
$M_9$	0.5	0.1	0.0	$2.5F_{\text{MSY}} \rightarrow F_{\text{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
$M_{10}$	0.5	0.5	0.0	$2.5F_{\text{MSY}} \rightarrow F_{\text{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
$M_{11}$	0.5	0.1	0.9	$2.5F_{\text{MSY}} \rightarrow F_{\text{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
$M_{12}$	0.5	0.5	0.9	$2.5F_{\text{MSY}} \rightarrow F_{\text{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
$M_{13}$	0.5	0.1	0.0	$F_{\text{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
$M_{14}$	0.5	0.5	0.0	$F_{\text{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
$M_{15}$	0.5	0.1	0.9	$F_{\text{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
$M_{16}$	0.5	0.5	0.9	$F_{\text{MSY}}$	Index SD = 0.4, Age composition SD = 1.5

Table S3. Distinguishing characteristics of the operating models with random effects on selectivity. Standard deviations (SD) are for log-normal distributed indices and logistic normal distributed age composition observations (fleet and indices). Fishing mortality changes after year 20 (of 40) for fishing histories where fishing mortality is not constant. For AR1 process errors,  $\sigma$  is defined for the marginal distribution of the processes.

Model	$\sigma_R$	$\sigma_{\text{Sel}}$	$\rho_{\text{Sel}}$	Fishing History	Observation Uncertainty
Sel <sub>1</sub>	0.5	0.1	0.0	$2.5F_{\text{MSY}} \rightarrow F_{\text{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
Sel <sub>2</sub>	0.5	0.5	0.0	$2.5F_{\text{MSY}} \rightarrow F_{\text{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
Sel <sub>3</sub>	0.5	0.1	0.9	$2.5F_{\text{MSY}} \rightarrow F_{\text{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
Sel <sub>4</sub>	0.5	0.5	0.9	$2.5F_{\text{MSY}} \rightarrow F_{\text{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
Sel <sub>5</sub>	0.5	0.1	0.0	$F_{\text{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
Sel <sub>6</sub>	0.5	0.5	0.0	$F_{\text{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
Sel <sub>7</sub>	0.5	0.1	0.9	$F_{\text{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
Sel <sub>8</sub>	0.5	0.5	0.9	$F_{\text{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
Sel <sub>9</sub>	0.5	0.1	0.0	$2.5F_{\text{MSY}} \rightarrow F_{\text{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
Sel <sub>10</sub>	0.5	0.5	0.0	$2.5F_{\text{MSY}} \rightarrow F_{\text{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
Sel <sub>11</sub>	0.5	0.1	0.9	$2.5F_{\text{MSY}} \rightarrow F_{\text{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
Sel <sub>12</sub>	0.5	0.5	0.9	$2.5F_{\text{MSY}} \rightarrow F_{\text{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
Sel <sub>13</sub>	0.5	0.1	0.0	$F_{\text{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
Sel <sub>14</sub>	0.5	0.5	0.0	$F_{\text{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
Sel <sub>15</sub>	0.5	0.1	0.9	$F_{\text{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
Sel <sub>16</sub>	0.5	0.5	0.9	$F_{\text{MSY}}$	Index SD = 0.4, Age composition SD = 1.5

Table S4. Distinguishing characteristics of the operating models with random effects on catchability. Standard deviations (SD) are for log-normal distributed indices and logistic normal distributed age composition observations (fleet and indices). Fishing mortality changes after year 20 (of 40) for fishing histories where fishing mortality is not constant. For AR1 process errors,  $\sigma$  is defined for the marginal distribution of the processes.

Model	$\sigma_R$	$\sigma_q$	$\rho_q$	Fishing History	Observation Uncertainty
$q_1$	0.5	0.1	0.0	$2.5F_{\text{MSY}} \rightarrow F_{\text{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
$q_2$	0.5	0.5	0.0	$2.5F_{\text{MSY}} \rightarrow F_{\text{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
$q_3$	0.5	0.1	0.9	$2.5F_{\text{MSY}} \rightarrow F_{\text{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
$q_4$	0.5	0.5	0.9	$2.5F_{\text{MSY}} \rightarrow F_{\text{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
$q_5$	0.5	0.1	0.0	$F_{\text{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
$q_6$	0.5	0.5	0.0	$F_{\text{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
$q_7$	0.5	0.1	0.9	$F_{\text{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
$q_8$	0.5	0.5	0.9	$F_{\text{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
$q_9$	0.5	0.1	0.0	$2.5F_{\text{MSY}} \rightarrow F_{\text{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
$q_{10}$	0.5	0.5	0.0	$2.5F_{\text{MSY}} \rightarrow F_{\text{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
$q_{11}$	0.5	0.1	0.9	$2.5F_{\text{MSY}} \rightarrow F_{\text{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
$q_{12}$	0.5	0.5	0.9	$2.5F_{\text{MSY}} \rightarrow F_{\text{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
$q_{13}$	0.5	0.1	0.0	$F_{\text{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
$q_{14}$	0.5	0.5	0.0	$F_{\text{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
$q_{15}$	0.5	0.1	0.9	$F_{\text{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
$q_{16}$	0.5	0.5	0.9	$F_{\text{MSY}}$	Index SD = 0.4, Age composition SD = 1.5



Table S5. Distinguishing characteristics of the estimating models.

Model	Recruitment model	Mean $M$	Process error assumption
EM <sub>1</sub>	Mean recruitment	0.2	Recruitment ( $\sigma_R$ estimated)
EM <sub>2</sub>	Beverton-Holt	0.2	Recruitment ( $\sigma_R$ estimated)
EM <sub>3</sub>	Mean recruitment	Estimated	Recruitment ( $\sigma_R$ estimated)
EM <sub>4</sub>	Beverton-Holt	Estimated	Recruitment ( $\sigma_R$ estimated)
EM <sub>5</sub>	Mean recruitment	0.2	Recruitment and survival ( $\sigma_R, \sigma_{2+}$ estimated)
EM <sub>6</sub>	Beverton-Holt	0.2	Recruitment and survival ( $\sigma_R, \sigma_{2+}$ estimated)
EM <sub>7</sub>	Mean recruitment	Estimated	Recruitment and survival ( $\sigma_R, \sigma_{2+}$ estimated)
EM <sub>8</sub>	Beverton-Holt	Estimated	Recruitment and survival ( $\sigma_R, \sigma_{2+}$ estimated)
EM <sub>9</sub>	Mean recruitment	0.2	Recruitment and uncorrelated natural mortality ( $\sigma_R, \sigma_M$ estimated, $\rho_M = 0$ )
EM <sub>10</sub>	Beverton-Holt	0.2	Recruitment and uncorrelated natural mortality ( $\sigma_R, \sigma_M$ estimated, $\rho_M = 0$ )
EM <sub>11</sub>	Mean recruitment	Estimated	Recruitment and uncorrelated natural mortality ( $\sigma_R, \sigma_M$ estimated, $\rho_M = 0$ )
EM <sub>12</sub>	Beverton-Holt	Estimated	Recruitment and uncorrelated natural mortality ( $\sigma_R, \sigma_M$ estimated, $\rho_M = 0$ )
EM <sub>13</sub>	Mean recruitment	0.2	Recruitment and uncorrelated fleet selectivity ( $\sigma_R, \sigma_{Sel}$ estimated, $\rho_{Sel} = 0$ )
EM <sub>14</sub>	Beverton-Holt	0.2	Recruitment and uncorrelated fleet selectivity ( $\sigma_R, \sigma_{Sel}$ estimated, $\rho_{Sel} = 0$ )
EM <sub>15</sub>	Mean recruitment	Estimated	Recruitment and uncorrelated fleet selectivity ( $\sigma_R, \sigma_{Sel}$ estimated, $\rho_{Sel} = 0$ )
EM <sub>16</sub>	Beverton-Holt	Estimated	Recruitment and uncorrelated fleet selectivity ( $\sigma_R, \sigma_{Sel}$ estimated, $\rho_{Sel} = 0$ )
EM <sub>17</sub>	Mean recruitment	0.2	Recruitment and uncorrelated catchability (spring index) ( $\sigma_R, \sigma_q$ estimated, $\rho_q = 0$ )
EM <sub>18</sub>	Beverton-Holt	0.2	Recruitment and uncorrelated catchability (spring index) ( $\sigma_R, \sigma_q$ estimated, $\rho_q = 0$ )
EM <sub>19</sub>	Mean recruitment	Estimated	Recruitment and uncorrelated catchability (spring index) ( $\sigma_R, \sigma_q$ estimated, $\rho_q = 0$ )
EM <sub>20</sub>	Beverton-Holt	Estimated	Recruitment and uncorrelated catchability (spring index) ( $\sigma_R, \sigma_q$ estimated, $\rho_q = 0$ )
EM <sub>21</sub>	Mean recruitment	0.2	Recruitment and AR1 natural mortality ( $\sigma_R, \sigma_M, \rho_M$ estimated)
EM <sub>22</sub>	Beverton-Holt	0.2	Recruitment and AR1 natural mortality ( $\sigma_R, \sigma_M, \rho_M$ estimated)
EM <sub>23</sub>	Mean recruitment	Estimated	Recruitment and AR1 natural mortality ( $\sigma_R, \sigma_M, \rho_M$ estimated)
EM <sub>24</sub>	Beverton-Holt	Estimated	Recruitment and AR1 natural mortality ( $\sigma_R, \sigma_M, \rho_M$ estimated)
EM <sub>25</sub>	Mean recruitment	0.2	Recruitment and AR1 selectivity ( $\sigma_R, \sigma_{Sel}, \rho_{Sel}$ estimated)
EM <sub>26</sub>	Beverton-Holt	0.2	Recruitment and AR1 selectivity ( $\sigma_R, \sigma_{Sel}, \rho_{Sel}$ estimated)
EM <sub>27</sub>	Mean recruitment	Estimated	Recruitment and AR1 selectivity ( $\sigma_R, \sigma_{Sel}, \rho_{Sel}$ estimated)
EM <sub>28</sub>	Beverton-Holt	Estimated	Recruitment and AR1 selectivity ( $\sigma_R, \sigma_{Sel}, \rho_{Sel}$ estimated)
EM <sub>29</sub>	Mean recruitment	0.2	Recruitment and AR1 catchability (spring index) ( $\sigma_R, \sigma_q, \rho_q$ estimated)
EM <sub>30</sub>	Beverton-Holt	0.2	Recruitment and AR1 catchability (spring index) ( $\sigma_R, \sigma_q, \rho_q$ estimated)
EM <sub>31</sub>	Mean recruitment	Estimated	Recruitment and AR1 catchability (spring index) ( $\sigma_R, \sigma_q, \rho_q$ estimated)
EM <sub>32</sub>	Beverton-Holt	Estimated	Recruitment and AR1 catchability (spring index) ( $\sigma_R, \sigma_q, \rho_q$ estimated)

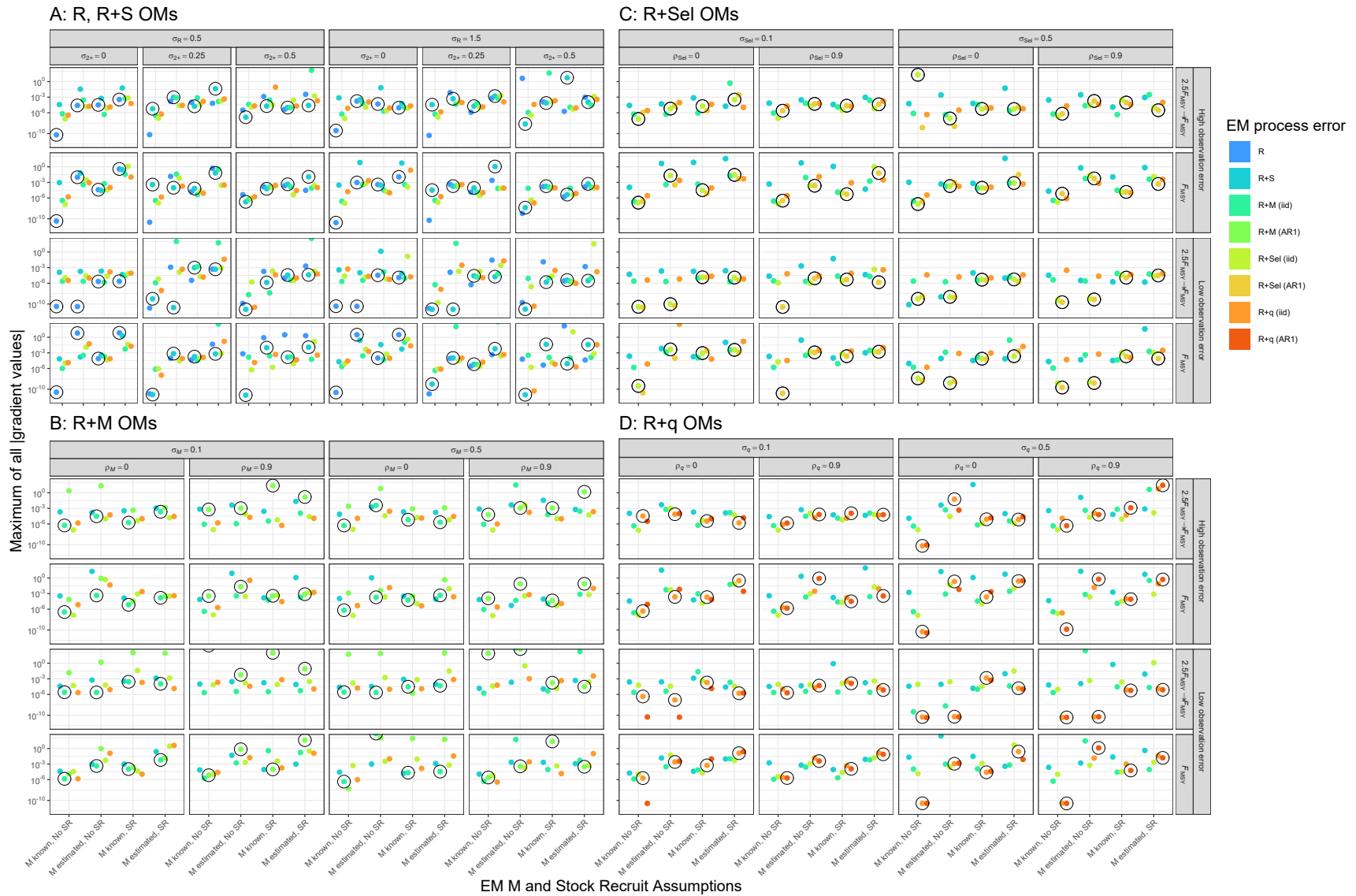


Fig. S2. The maximum of the absolute values of all gradient values for all fits that provided hessian-based standard errors across all simulated data sets of a given OM configuration (A: R and R+S, B: R+M, C: R+Sel, or D: R+q). Results are conditional on EM fits with alternative process error type (colored points and lines), median natural mortality (estimated or known) and recruitment assumptions (Beverton-Holt stock recruit or not). Circled values indicate results where the EM process error structure matches that of the operating model and vertical lines represent 95% confidence intervals.

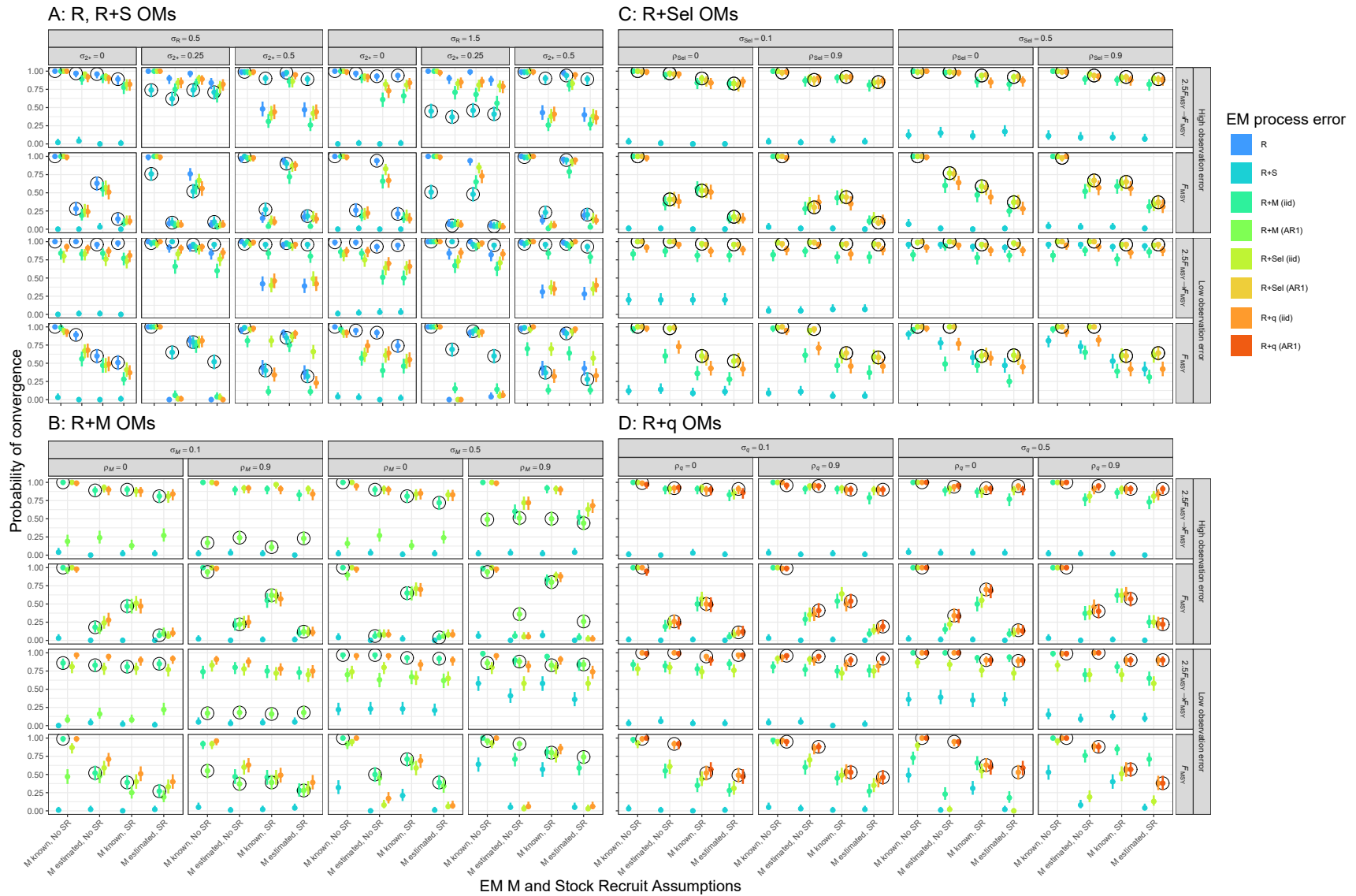


Fig. S3. Probability of estimating models providing maximum absolute values of gradients less than  $10^{-6}$  assuming alternative process error (colored points and lines), and median natural mortality (estimated or known) and Beverton-Holt stock recruit functions (estimated or not; along x-axis) when fitted to operating models that have R and R+S (A), R+Sel (B), R+M (C), or R+q (D) process error structures. Circled values indicate results where the EM process error structure matches that of the operating model and vertical lines represent 95% confidence intervals.

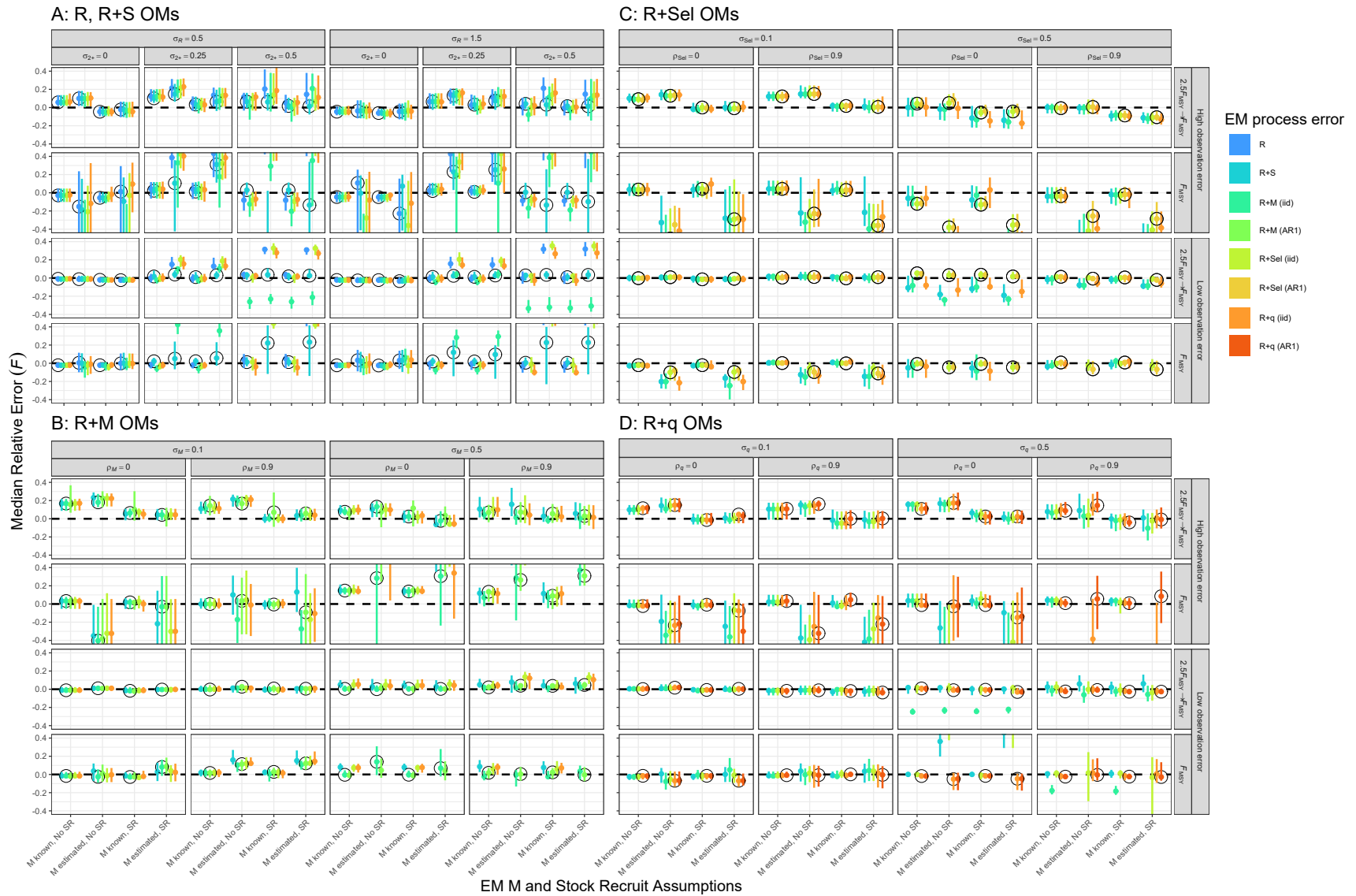


Fig. S4. Median relative error of terminal year fully-selected fishing mortality for estimating models fitted to data sets simulated with alternative process error structures: R and R+S (A), R+Sel (B), R+M (C), or R+q (D). Circled values indicate results where the EM process error structure matches that of the operating model and vertical lines represent 95% confidence intervals.

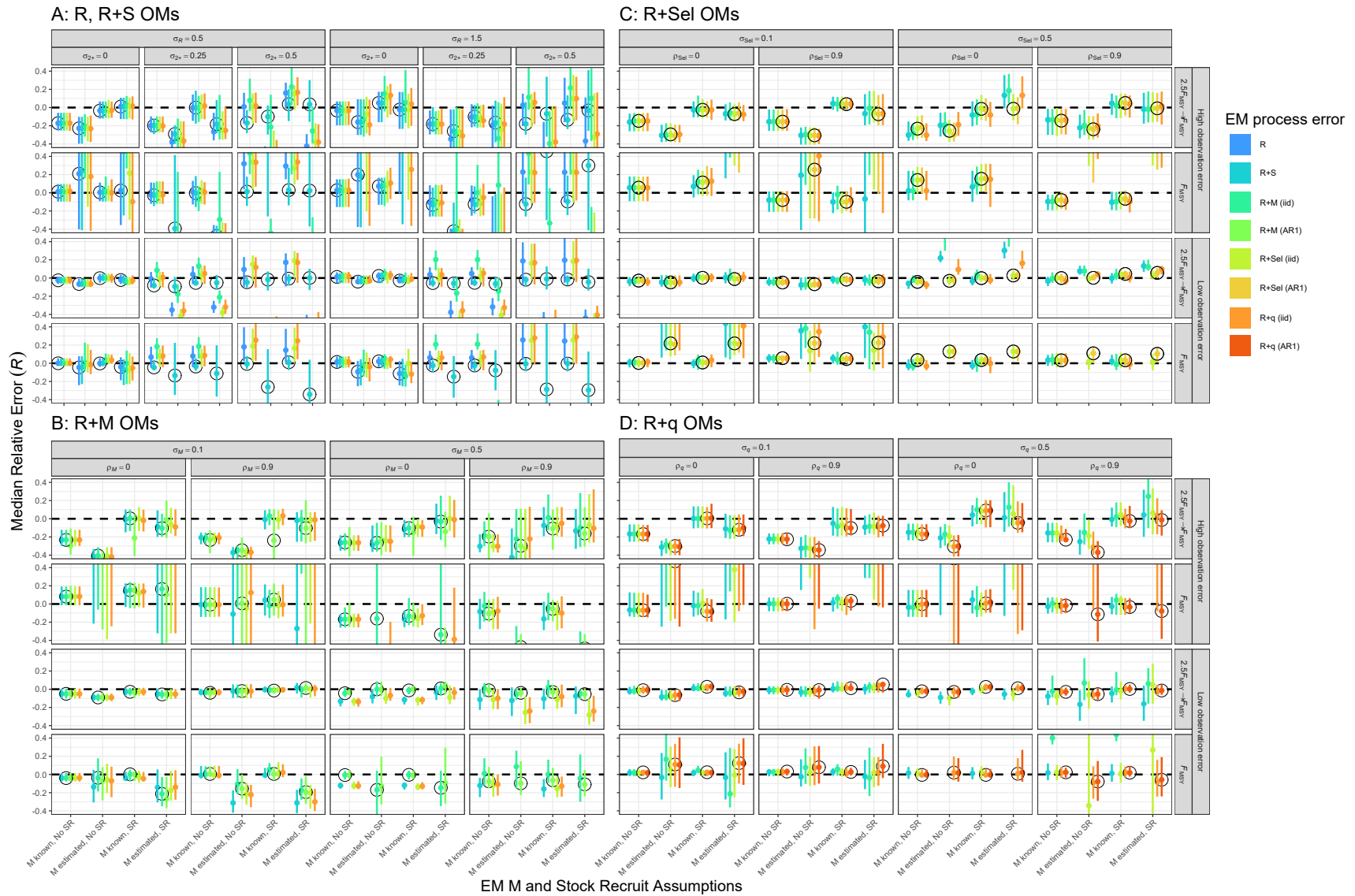


Fig. S5. Median relative error of terminal year recruitment for estimating models fitted to data sets simulated with alternative process error structures: R and R+S (A), R+Sel (B), R+M (C), or R+q (D). Circled values indicate results where the EM process error structure matches that of the operating model and vertical lines represent 95% confidence intervals.

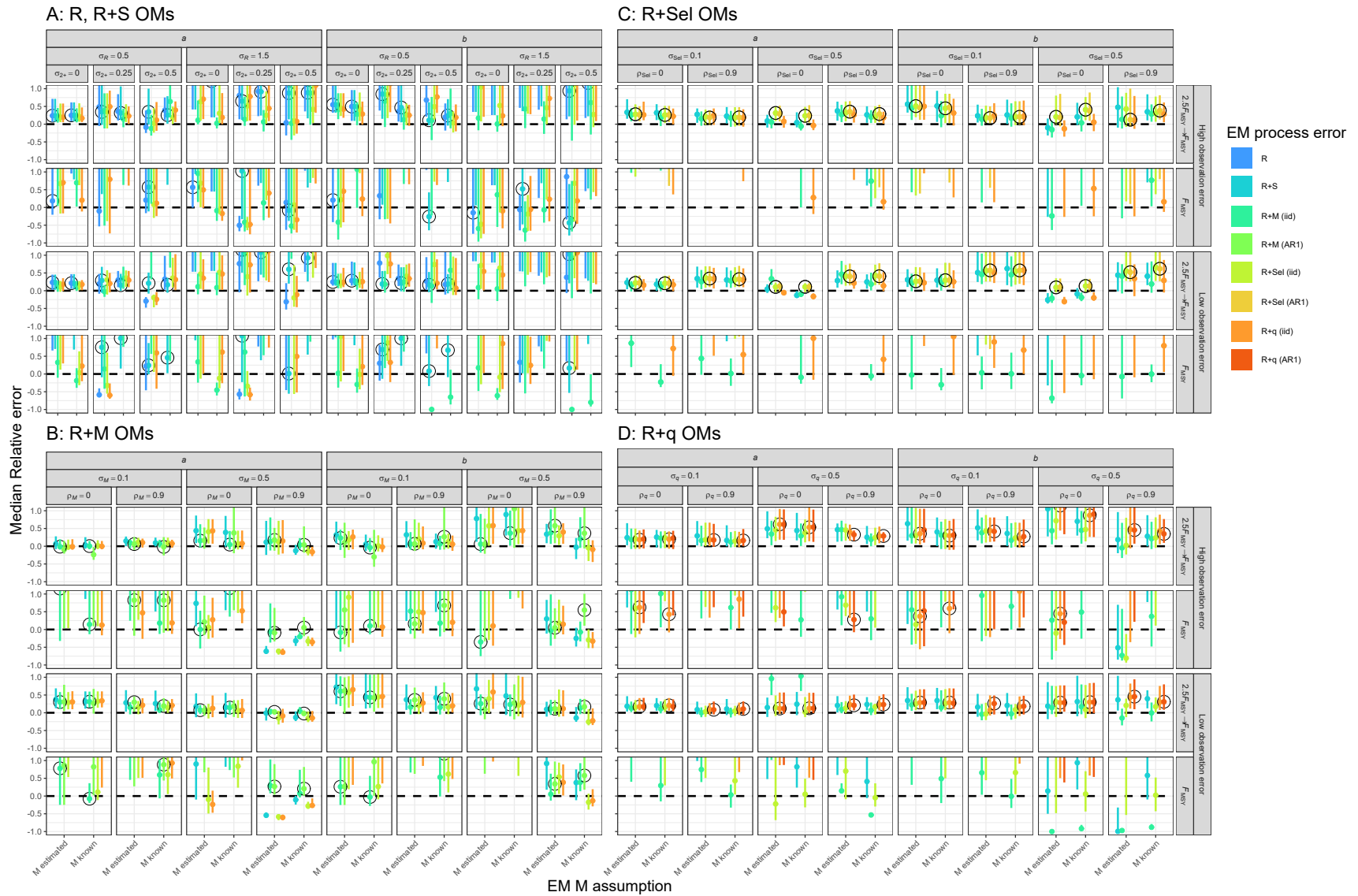


Fig. S6. Median relative error of Beverton-Holt stock recruitment parameters ( $a$  and  $b$ ) for estimating models fitted to data sets simulated with alternative process error structures: R and R+S (A), R+Sel (B), R+M (C), or R+q (D). Circled values indicate results where the EM process error structure matches that of the operating model and vertical lines represent 95% confidence intervals.

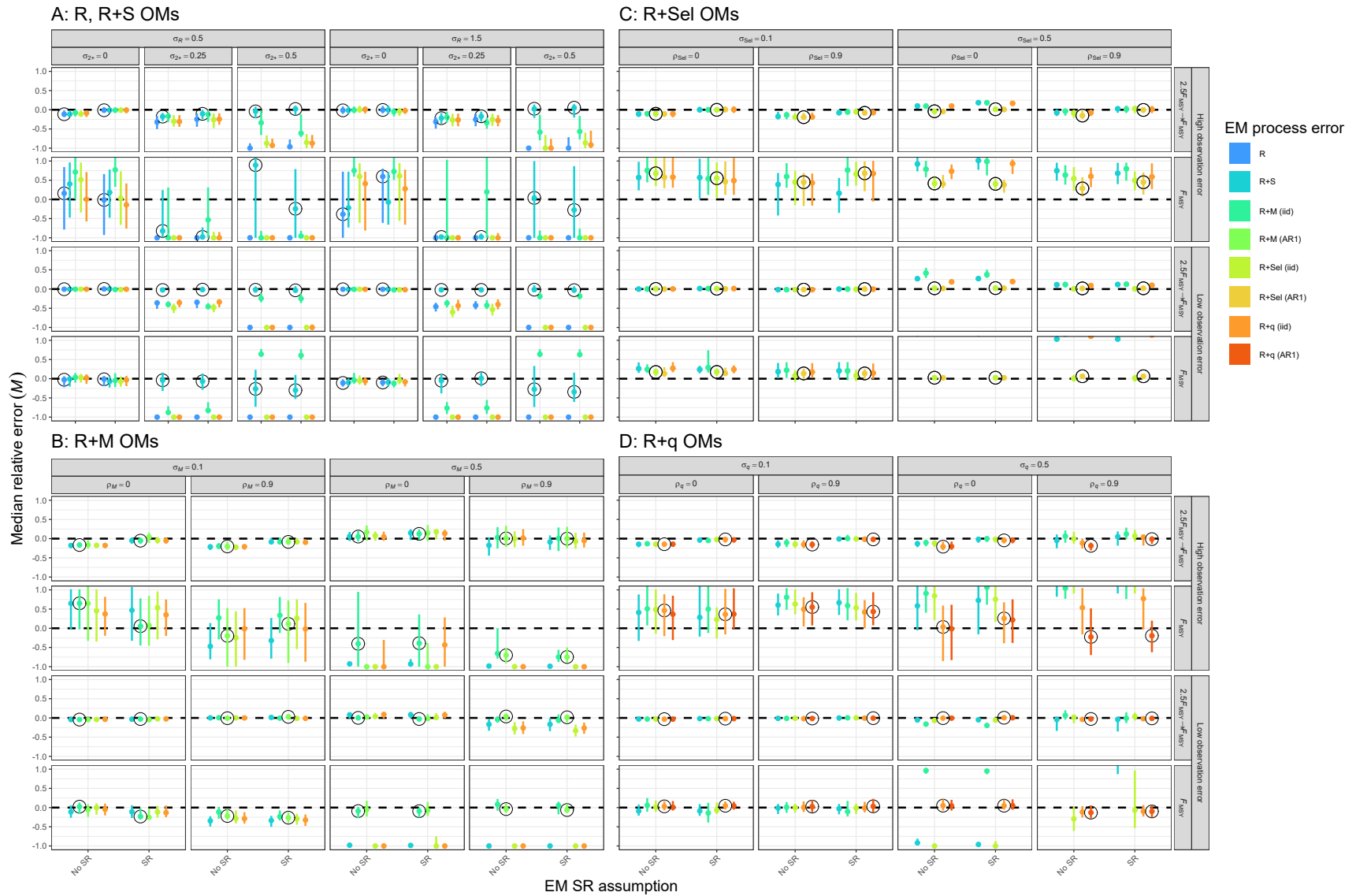


Fig. S7. Median relative error of median natural mortality for estimating models fitted to data sets simulated with alternative process error structures: R and R+S (A), R+Sel (B), R+M (C), or R+q (D). Circled values indicate results where the EM process error structure matches that of the operating model and vertical lines represent 95% confidence intervals.



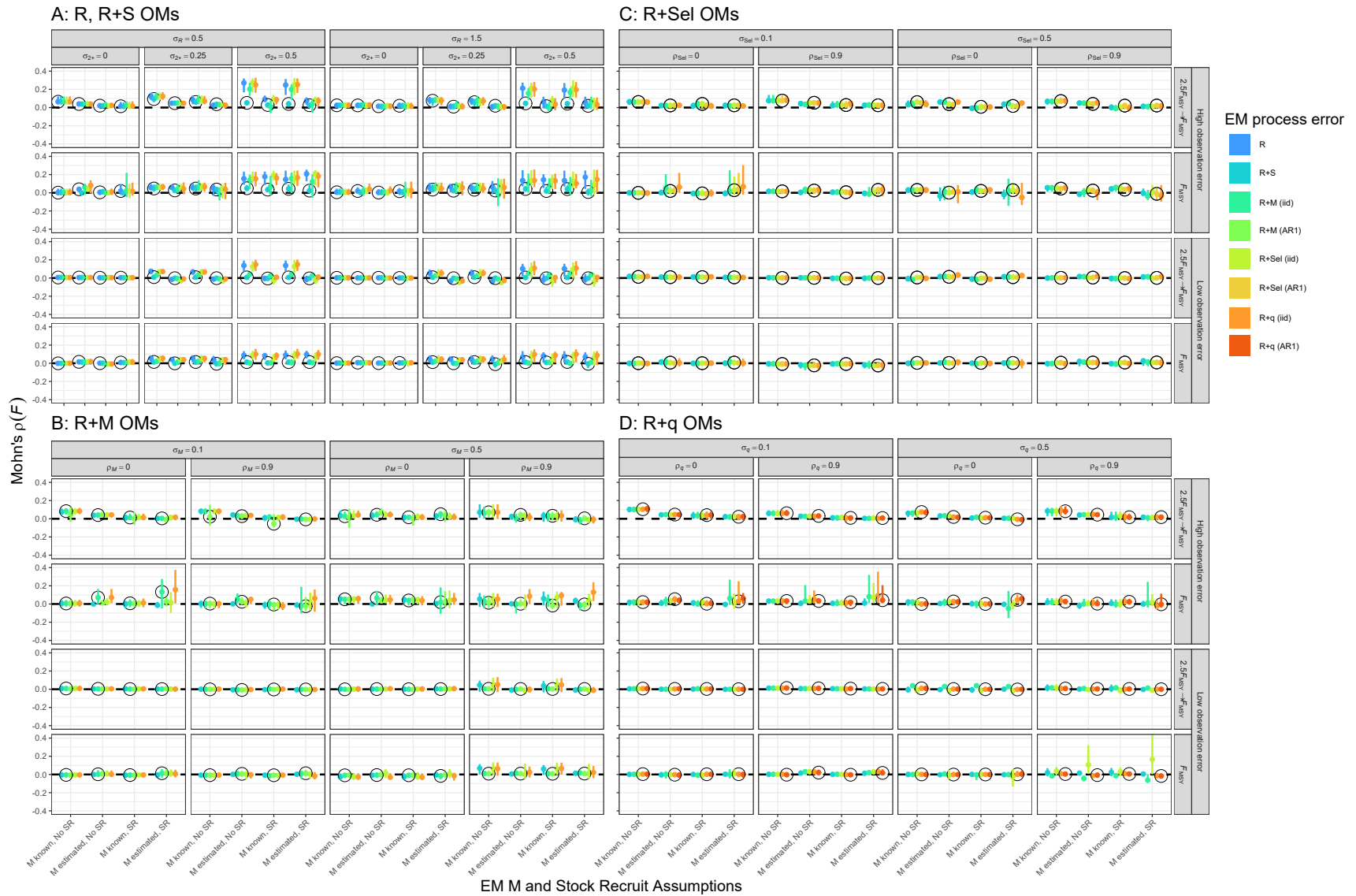


Fig. S8. Median Mohn's  $\rho$  of fishing mortality averaged over all age classes for estimating models fitted to data sets simulated with alternative process error structures: R and R+S (A), R+Sel (B), R+M (C), or R+q (D). Circled values indicate results where the EM process error structure matches that of the operating model and vertical lines represent 95% confidence intervals.



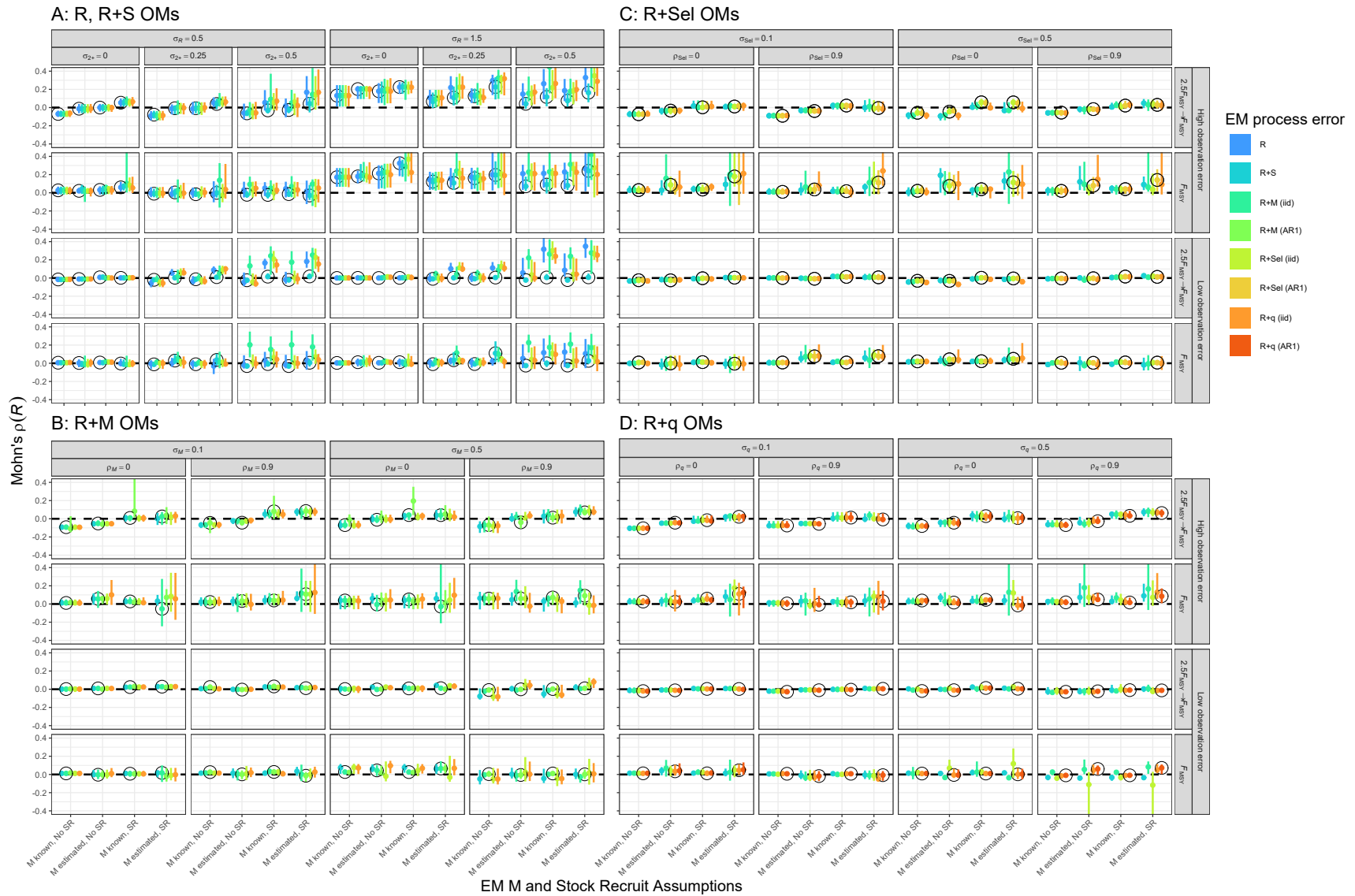


Fig. S9. Median Mohn's  $\rho$  of recruitment for estimating models fitted to data sets simulated with alternative process error structures: R and R+S (A), R+Sel (B), R+M (C), or R+q (D). Circled values indicate results where the EM process error structure matches that of the operating model and vertical lines represent 95% confidence intervals.